

THESE

présentée pour l'obtention du grade de
DOCTEUR DE L'UNIVERSITE DE STRASBOURG

par

Sébastien GALLIEN

Nouvelles méthodologies protéomiques d'aide à l'annotation des génomes et à la validation des séquences protéiques

Soutenue le 07 décembre 2009 devant la commission d'examen :

Dr. Alain VAN DORSSELAER	Directeur de thèse
Dr. Bruno DOMON	Rapporteur
Prof. Joël VANDEKERCKHOVE	Rapporteur
Dr. Philippe BULET	Examineur
Dr. Odile LECOMPTE	Examineur
Prof. Laurence SABATIER	Examineur

A mes parents

A Emilie

A mes amis

« Le meilleur moyen d'avoir une bonne idée est d'en avoir beaucoup.»

Linus Carl Pauling

Merci !

Ce travail de thèse a été réalisé au sein du Laboratoire de Spectrométrie de Masse Bio-Organique de l'Institut Pluridisciplinaire Hubert Curien de Strasbourg (UMR 7178, CNRS-ULP).

Mes plus sincères remerciements s'adressent tout d'abord à Alain Van Dorsselaer pour m'avoir accueilli, soutenu, encouragé et donné les moyens de réaliser cette thèse dans un cadre de travail idéal. Je tiens également à remercier Christine Schaeffer pour son encadrement au quotidien, pour avoir cru en mes idées et m'avoir donné les moyens de les réaliser.

Je remercie vivement M. Philippe Bulet qui a accepté de présider mon jury de thèse ainsi que Mmes Laurence Sabatier et Odile Lecompte et Messieurs Bruno Domon et Joël Vandekerckhove pour avoir accepté d'évaluer mon travail de thèse.

Je remercie également l'ensemble des personnes avec qui j'ai eu un immense plaisir à collaborer : O. Poch, O. Lecompte, E. Perrodou, J.M. Reyrat, C. Deshayes, G. Fuchs, W. Ismail, Y.R. Chiang, P. Bertin, F. Arsène-Ploetze, F. Goulhen-Chollet, A. Heinrich-Salmeron, J. Weissenbach, D. Le Paslier, C. Medigue, T. Bach, H. Schaller, et M. Boll, ...

Un grand merci à tous ceux avec qui j'ai partagé ces quelques années au LSMBO :

Les « anciens » partis : Audrey Bednarczyk (arrivés en même temps, tu nous a « abandonné » depuis un certain temps déjà, bonne réussite à toi dans ta carrière « suisse »), Haiko Herschbach (dit « la rigueur »), Flavie Robert (une grande « protectrice » de la planète s'il en est), Laetitia Fouillen (un modèle de goût musical, enfin pas pour tout...), Laurent Miguet (parti et revenu pas très loin, beaucoup de discussions scientifiques et footballistiques, et oui bientôt le Racing en national...), Dimitri Heintz (parti pas très loin, de nombreuses collaborations très intéressantes menées ensemble, je garderai de très très bons souvenirs de nos projets « off » du samedi, voire du dimanche...), ceux que je n'ai malheureusement pas vraiment eut le temps de connaître (Guillaume Chevreux, Raymond Hueber, Andy High, Marie-Laure Lortz, Julien Roeser, Vincent Bonifay...)

Les « anciens » encore là : Fabrice Varrier (désolé pour le « squattage » des ressources informatiques du labo et merci pour ta réactivité face à certains de mes besoins inopinés), Fabrice Bertile (compagnon d'un peu « tout » : de cigarette, de voyage, de réflexion, parfois très philosophiques, et même d'étude de rat !!!), Hélène Diemer et Jennyfer Jund (les grandes prêtresses de cette nouvelle religion qu'est l'ISO), Danièle thierse (qui nous apprend à tous la bonne manière de travailler, la manière « propre »), Jean-Marc Strub (un modèle de goût cinématographique avec qui discuter « technique instrumentale » est

toujours un plaisir même si nous ne sommes pas toujours d'accord), Sarah Sanglier (j'espère que mon passage t'a permis de parfaire tes connaissances culinaires), Agnès Hovasse (et ses « fameuses bottes », un grand plaisir de décortiquer avec toi ces trappes ioniques parfois très capricieuses), Cyril Colas (viens et pénètre du côté obscur de la protéomique et tu verras, tu ne pourras plus t'en passer !), Véronique Delval-Dubois (« de la vallée-de la forêt » pour les intimes, une « extérieure-intérieure » qui ne peut décidément pas se passer du LSMBO), Véronique Trimbou (un soutien administratif à toute épreuve), François Delalande (et sa technique d'enseignement toute particulière (et plus ou moins consciente) : provoquer des pannes pour faciliter la formation des nouveaux arrivants...) et Christine Carapito (un peu mon « mentor » dans le domaine de la protéomique, ce fut plus qu'un plaisir de travailler à tes côtés).

Les « neufs » (plus ou moins) :

Le bureau des « précieux » : Guillaume Béchade (peut-être à très bientôt un peu plus au nord de l'Europe, ce serait en tout cas un réel plaisir...), Christelle Husser (la relève de l'électrophorèse, beaucoup de boulot pour se sortir des griffes et de son chat et d'un fameux FV,...), Jean-Michel Saliou (de l'équipe de nuit du LSMBO, à qui on ne sait jamais s'il faut dire bonjour ou bonne nuit...), les tout tout « neufs » Alexandre Burel et Elie Alayi (bon courage à vous et profitez bien de votre passage au labo).

La « dream-team », j'ai nommé mes chers disciples « pouilleux » : Cédric Atmanene (un fervent supporter de l'OM, le digne héritier de la couronne « pouilleuse » mais déjà parti avant le début de son règne, ce fut un plaisir de t'avoir côtoyé au quotidien, à bientôt pour de nouveaux loisirs « on-line »...), Nicolas Barthelemy (plein d'idées et d'enthousiasme, tu verras, tu seras bientôt récompensé...), Thierry Wasselin (le spécialiste du dépouillement gel 2D, n'en n'oublie pas pour autant de manger !), Daniel Ayoub (le successeur pour le TMPP, j'ai confiance en toi pour reprendre en main tous ces projets et je suis sûr que tu auras plein de résultats...) et Stéphanie Petiot (la première demoiselle du bureau des pouilleux, bon courage à toi pour cette thèse qui débute, vue ton dynamisme tout devrait très bien se passer !).

Un grand merci à mon entourage et mes plus proches amis (même si l'éloignement géographique limite les contacts, c'est toujours un immense plaisir de vous retrouver !).

Un immense merci à ma famille et surtout mes parents et à Emilie pour m'avoir toujours soutenu et crût en moi...

Et à tous ceux que j'ai oubliés, MERCI !

Liste des principales abbréviations

ACN :	Acétonitrile
ADN :	Acide désoxyribonucléique
ARN :	Acide ribonucléique
BPC :	Base Peak Chromatogram
BLAST :	Basic Local alignment Search Tool
CDS :	CoDing Sequence
CID :	Collision-Induced Dissociation
Da :	Dalton
DTT :	Dithiothreitol
ECD :	Electron Capture Dissociation
ESI :	Electrospray Ionisation
ETD :	Electron Transfer Dissociation
HPLC :	High Performance Liquid Chromatography
ICDS :	Interrupted CoDing Sequence
IT :	Ion Trap
LC :	Liquid Chromatography
MALDI :	Matrix-Assisted Laser Desorption/Ionisation
MS :	Spectrométrie de Masse
MS/MS :	Spectrométrie de masse en tandem
m/z :	mass to charge ratio
Nano-ESI :	Nano-Electrospray
NanoLC-MS/MS :	Nano liquid chromatography tandem mass spectrometry
N-TOP :	N-Terminal Oriented Proteomic
PCR :	Polymerase Chain Reaction
pI :	Isoelectric point
ppm :	part per million
PTM :	Post-translational Modification
Q :	Quadripôle
qN-TOP :	quantitative N-Terminal Oriented Proteomic
RP :	Reverse Phase
SCX :	Strong Cation Exchange
SDS-PAGE :	Sodium DodecylSulfate PolyAcrylamide Gel Electrophoresis
TBP :	Tributylphosphine
TIGR :	Institute for Genomic Research
TOF :	Time of Flight

INTRODUCTION GENERALE.....	1
PARTIE BIBLIOGRAPHIQUE	5
Introduction.....	5
Chapitre 1 : Les banques de séquences protéiques.....	7
Chapitre 2 : Le séquençage et la problématique de l'annotation des génomes	17
Chapitre 3 : Les outils de l'analyse protéomique	29
Chapitre 4 : Les stratégies d'identification en analyse protéomique.....	53
Conclusion	75
Bibliographie	77
RESULTATS.....	89
Partie I : Développement et applications de méthodologies protéomiques d'aide à l'annotation génomique	89
Chapitre 1 : Mise en place et application d'une stratégie de protéogénomique	89
Chapitre 2 : Nouvelle méthode de protéogénomique axée sur la détermination des codons d'initiation des protéines	101
Chapitre 3 : Application de la stratégie N-TOP pour l'étude d'une enzyme issue d'un organisme dont le génome n'est pas séquencé	127
Chapitre 4 : Etude méta-protéo-génomique d'un écosystème microbien riche en arsenic.....	135
Conclusion	143
Bibliographie	145
Partie II : La dérivation chimique pour améliorer l'exploration des phosphoprotéomes.....	149
Chapitre 1 : Analyse des phosphorylations par spectrométrie de masse.....	149
Chapitre 2 : Le marquage TMPP pour l'analyse phosphoprotéomique	157
Conclusion	185
Bibliographie	187
Partie III : Protéomique Quantitative.....	193
Chapitre 1 : La spectrométrie de masse quantitative en analyse protéomique.....	193
Chapitre 2 : Méthode de quantification protéomique par comptage de spectres ; application à <i>Geobacter metallireducens</i>	207
Chapitre 3 : Validation et application d'une méthode de quantification protéomique avec marquage au $^{13}\text{C}_9$-TMPP.....	217
Conclusion	241
Bibliographie	243
CONCLUSION GENERALE.....	249
PARTIE EXPERIMENTALE.....	253

TRAVAUX SUPPLEMENTAIRES263

INTRODUCTION GENERALE	1
PARTIE BIBLIOGRAPHIQUE	5
Introduction.....	5
Chapitre 1 : Les banques de séquences protéiques.....	7
1. Origine, historique et alimentation des banques protéiques.....	7
2. Les différentes banques protéiques	10
2.1. Les banques de dépôt.....	10
2.1.1. La banque GenPept	10
2.1.2. La banque NCBI's Entrez Protein.....	10
2.1.3. La banque Refseq.....	10
2.2. Les banques soignées et le consortium UniProt	11
2.2.1. Les principales banques « soignées » à l'origine d'Uniprot.....	12
2.2.1.1. La banque PIR-PSD (Protein Information Resource Protein Sequence Database)	12
2.2.1.2. La banque Swiss-Prot, appelée UniProtKB/Swiss-Prot dans UniProtKB	12
2.2.1.3. La banque TrEMBL (« Translation from EMBL) appelée UniProtKB/TrEMBL dans UniProtKB.....	13
2.2.2. Les composantes d'UniProt	13
2.2.2.1. UniProt Archive (UniParc) [Leinonen et al., 2004]	13
2.2.2.2. UniProt Knowledgebase (UniProtKB) [Wu et al., 2006]	14
2.2.2.3. UniProt reference clusters (UniRef) [Suzek et al., 2007]	14
2.2.2.4. UniProt Metagenomic and Environmental Sequences (UniMES) [\"The_UniProt_Consortium\", 2009]	14
Chapitre 2 : Le séquençage et la problématique de l'annotation des génomes	17
1. Le séquençage des génomes.....	17
1.1. Le séquençage de l'ADN.....	17
1.2. Les stratégies de séquençage des génomes	18
2. L'annotation des génomes	20
2.1. Méthodes intrinsèques (ou <i>ab initio</i>) :	21
2.2. Méthodes extrinsèques :	22
3. Les erreurs d'annotation, leurs conséquences et les modes de correction.....	23
3.1. Les erreurs.....	23
3.1.1. Séquençage du génome	23
3.1.2. Annotation du génome	24
3.2. Les conséquences des erreurs	25
3.3. Les méthodes de correction.....	25
3.3.1. Corrections in silico	26
3.3.2. Correction expérimentale	26
Chapitre 3 : Les outils de l'analyse protéomique.....	29
1. Les outils de la spectrométrie de masse.....	30
1.1. Les sources d'ionisation.....	30
1.1.1. La source MALDI (Matrix assisted laser desorption ionization)	30
1.1.2. La source electrospray (ESI).....	31
1.1.2.1. Processus d'ionisation-désorption ESI	32
1.1.2.1.1. Production des gouttelettes chargées.....	32
1.1.2.1.2. Fission des gouttelettes chargées, l'explosion coulombienne	32
1.1.2.1.3. Emission des ions désolvatés en phase gazeuse	33
1.1.2.2. La source nano-électrospray.....	34
1.1.2.3. Ionisation des analytes.....	35
1.1.2.3.1. Ionisation par séparation des charges.....	35
1.1.2.3.2. Ionisation par formation d'adduits.....	35
1.1.2.3.3. Ionisation par réactions en phase gazeuse	35
1.1.2.4. Efficacité d'ionisation des analytes	36

1.1.2.4.1. Affinité pour la surface des gouttes.....	36
1.1.2.4.2. Prédiction de la réponse ESI des analytes	37
1.2. Les analyseurs	37
1.2.1. L'analyseur quadripolaire et le système quadripôle-temps de vol	37
1.2.1.1. L'analyseur quadripolaire.....	37
1.2.1.2. Le système quadripôle-temps de vol : des analyseurs en tandem.....	38
1.2.2. L'analyseur trappe ionique (« Ion Trap », IT).....	38
1.2.2.1. Le piégeage des ions.....	39
1.2.2.2. L'éjection des ions.....	39
1.2.2.3. L'isolation des ions	39
1.2.2.4. Excitation et fragmentation des ions	39
1.2.2.5. Séquence d'analyse de peptides en MS/MS	39
1.3. La fragmentation peptidique	40
1.3.1. Les étapes de la fragmentation	41
1.3.2. La nomenclature des fragmentations peptidiques	42
1.3.3. Mécanismes de fragmentation.....	42
1.3.3.1. Le modèle du proton mobile.....	42
1.3.3.2. Les coupures préférentielles	44
2. Les techniques séparatives de protéines et de peptides	45
2.1. Séparation des protéines.....	45
2.1.1. Le gel d'électrophorèse bidimensionnel (gel 2D)	45
2.1.2. Le gel monodimensionnel (gel 1D ou gel SDS [sodium DodecylSulfate] PAGE [PolyAcrylamide Gel Electrophoresis])	46
2.1.3. Les méthodes alternatives sans gel (« gel free »).....	46
2.2. Séparation des peptides	47
2.2.1. Le couplage RPLC-MS/MS (chromatographie liquide en phase inverse couplée à la spectrométrie de masse en tandem).....	47
2.2.2. La chromatographie liquide multidimensionnelle	48
2.2.2.1. Le principe clé de la chromatographie multidimensionnelle : l'orthogonalité	48
2.2.2.2. Les combinaisons standards	49
2.2.2.3. Configurations « off-line » et « on-line »	49
Chapitre 4 : Les stratégies d'identification en analyse protéomique.....	53
1. L'empreinte peptidique massique (« Peptide Mass fingerprinting », PMF)	53
2. Les approches par LC-MS/MS	54
2.1. L'approche par recherche dans les banques protéiques	55
2.1.1. Attribution d'un « score » à l'identification	57
2.1.1.1. Le score de corrélation	57
2.1.1.2. Le score probabiliste	58
2.1.2. Comparaison et combinaison de moteurs de recherche	59
2.1.2.1. Comparaison de moteurs de recherche	60
2.1.2.2. Utilisation combinée de plusieurs moteurs de recherche.....	60
2.1.3. Evaluation des identifications	61
2.1.3.1. Les stratégies « target-decoy »	62
2.1.3.1.1. Principes et mise en place	62
2.1.3.1.2. Etablissement des seuils.....	63
2.1.3.1.3. Des informations auxiliaires pour améliorer le processus d'évaluation des identifications	64
2.1.3.1.4. Déduction de l'identification des protéines (et de la confiance associée) à partir des identifications peptidiques	66
2.1.3.2. Les approches empiriques de Bayes.....	66
2.1.3.3. Les approches alternatives.....	68
2.1.3.4. Importance de l'évaluation des identifications en analyse protéomique	69
2.1.4. Les limitations de l'approche par recherche dans les banques protéiques	70
2.2. L'approche par séquençage de novo.....	70
2.3. Les approches alternatives	71
2.3.1. L'approche par recherche dans les banques de spectres.....	71
2.3.2. L'approche hybride	72

2.4. Les banques publiques de dépôt des données de protéomique	72
Conclusion	75
Bibliographie	77
RESULTATS.....	89
Partie I : Développement et applications de méthodologies protéomiques d'aide à l'annotation génomique	89
Chapitre 1 : Mise en place et application d'une stratégie de protéogénomique.....	89
1. La recherche de données MS/MS dans un génome complet	89
1.1. Origine de l'approche.....	89
1.2. Mise en place de la stratégie.....	90
1.2.1. Le découpage du génome.....	90
1.2.2. Intégration dans Mascot	90
1.2.3. Interrogation Mascot et identification des peptides.....	90
1.2.4. Identification de la fonction de la protéine par MS-BLAST	91
2. Application de la stratégie de recherche dans les génomes pour la correction d'erreurs de séquençage.....	93
2.1. Contexte de l'étude	93
2.2. Fréquence et origine des « ICDSs ».....	94
2.3. Stratégie d'analyse.....	94
2.3.1. Stratégie générale	94
2.3.2. Détection in silico et re-séquençage génomique des ICDSs	95
2.3.3. Stratégie protéogénomique.....	95
2.3.3.1. Préparation et analyse des échantillons	95
2.3.3.2. Stratégie de recherche dans le génome.....	96
2.4. Détection expérimentale des ICDS	96
2.5. Résultats publiés	99
2.6. Conclusion	99
Chapitre 2 : Nouvelle méthode de protéogénomique axée sur la détermination des codons d'initiation des protéines.....	101
1. Développement d'une nouvelle stratégie protéomique « orientée » N- terminale (« N-Terminal Oriented Proteomic », « N-TOP »).....	102
1.1. Les stratégies de détermination expérimentale des codons d'initiation des protéines.....	102
1.1.1. Les stratégies d'identification des codons d'initiation	102
1.1.2. Les stratégies d'enrichissement en peptides N-terminaux	103
1.2. Développement d'une nouvelle méthode pour améliorer la caractérisation des extrémités N- terminales des protéines (N-TOP)	106
1.2.1. L'étape clé de la stratégie N-TOP : le marquage N-terminal des protéines avec le TMPP-Ac- OSu	106
1.2.2. Elimination de l'excès de réactif et des composés de dégradation.....	107
1.2.3. Application de la méthode sur des protéines modèles.....	109
1.2.3.1. Impact du marquage TMPP sur l'efficacité d'ionisation et la rétention chromatographique des peptides N-terminaux	109
1.2.3.2. La spécificité du marquage TMPP	110
1.2.3.3. Impact du marquage TMPP sur la fragmentation des peptides N-terminaux	111
2. Application de la stratégie N-TOP comme aide à l'annotation génomique	113
2.1. Contexte de l'étude	113
2.2. Stratégie d'analyse protéogénomique	113
2.3. Corrections à l'annotation génomique de <i>M. smegmatis</i>	114
2.4. Propagation des déterminations expérimentales des codons d'initiation aux autres mycobactéries	115

2.5. Résultats publiés	116
2.6. Conclusion	116
3. Evolution de la stratégie N-TOP	117
3.1. Au niveau instrumental : une mesure de masse plus juste	117
3.2. Au niveau « traitement des données » : une stratégie de recherche dans les banques optimisée	117
3.3. Stratégie d'analyse protéogénomique	118
3.3.1. Peptides internes	119
3.3.2. Peptides N-terminaux marqués au TMPP	121
3.4. Résultats.....	123
3.4.1. Peptides internes	123
3.4.2. Peptides N-terminaux marqués au TMPP	123
3.4.2.1. Banque protéique.....	124
3.4.2.2. Sous-banque génomique.....	124
3.4.2.3. Banque génomique.....	124
3.4.2.4. Réaction aspécifique du TMPP	124
3.5. Conclusion	125
Chapitre 3 : Application de la stratégie N-TOP pour l'étude d'une enzyme issue d'un organisme dont le génome n'est pas séquencé	127
1. Contexte de l'étude.....	127
1.1. Métabolisme anaérobie du cholestérol.....	127
1.2. <i>Sterolibacterium denitrificans</i>	127
1.3. Objectif de l'étude.....	128
2. Stratégie d'analyse et résultats intermédiaires.....	128
2.1. Préparation des échantillons et purification partielle de l'enzyme d'intérêt.....	129
2.2. Analyse protéomique et identifications des peptides	130
2.3. Amplification et séquençage de la séquence génomique codant pour l'enzyme Acm B.	131
2.4. Approche N-TOP	131
2.5. Surexpression et purification de la protéine AcmB	132
2.6. Caractérisation de la protéine AcmB recombinante.....	132
3. Résultats publiés	133
4. Conclusion	133
Chapitre 4 : Etude méta-protéo-génomique d'un écosystème microbien riche en arsenic	135
1. Contexte de l'étude.....	135
1.1. Les communautés microbiennes et la métagénomique	135
1.2. Le drainage minier acide de Carnoulès (France).....	136
1.3. Objectif de l'étude.....	137
2. Stratégie d'analyse et résultats intermédiaires.....	137
2.1. Préparation des échantillons	137
2.2. Reconstitution des génomes individuels	138
2.3. Analyse Protéogénomique	138
2.3.1. Analyse par spectrométrie de masse	138
2.3.2. Stratégie d'identification des protéines	139
2.3.3. Distribution des identifications des protéines	140
3. Résultats	141
4. Conclusion et perspectives.....	142
Conclusion	143
Bibliographie	145

Partie II : La dérivation chimique pour améliorer l'exploration des phosphoprotéomes	149
Chapitre 1 : Analyse des phosphorylations par spectrométrie de masse	149
1. Les modifications post-traductionnelles des protéines	149
2. La phosphorylation des protéines	149
3. La nature des phosphorylations	151
4. L'analyse des phosphorylations	151
4.1. Enrichissement des phosphopeptides	151
4.1.1. Immunoprécipitation	151
4.1.2. Chromatographie d'affinité	151
4.1.2.1. IMAC (« Immobilized Metal-ion Affinity Chromatography »)	152
4.1.2.2. MOAC (« Metal Oxide Affinity Chromatography »)	152
4.1.3. Chromatographie d'échange de cations (« Strong Cation Exchange », SCX)	152
4.1.4. Modification chimique	153
4.2. Séquençage des phosphopeptides par spectrométrie de masse en tandem	154
4.2.1. Fragmentation MS ² induite par collision (« Collision-Induced Dissociation », CID)	154
4.2.2. Fragmentation MS ³ et pseudo-MS ³ induite par collision	155
4.2.3. Dissociation par capture d'électron (« Electron Capture Dissociation », ECD) et dissociation par transfert d'électron (« Electron Transfer Dissociation », ETD)	155
Chapitre 2 : Le marquage TMPP pour l'analyse phosphoprotéomique	157
1. Contexte de l'étude	157
1.1. Le marquage TMPP	157
1.2. Application en biochimie métabolique sur un modèle végétal : la biosynthèse d'isoprénoïdes chez <i>Arabidopsis thaliana</i>	157
1.3. Objectif de l'étude	159
2. Evaluation de la dérivation chimique au TMPP pour l'analyse phosphoprotéomique	160
2.1. Impact du marquage TMPP sur l'identification des phosphopeptides par LC-MS/MS	160
2.2. Impact du marquage TMPP sur la fragmentation des phosphopeptides	162
3. Développement de protocoles analytiques pour l'analyse phosphoprotéomique d'<i>A. thaliana</i>	167
3.1. Protocole analytique A : IMAC et IMAC - TiO ₂	168
3.2. Protocole analytique B : IMAC – fractionnement HPLC en phase inverse à pH basique	168
3.3. Protocole analytique C : TiO ₂ et Précipitation calcium - TiO ₂	170
4. Apport du marquage TMPP pour la phosphoprotéomique comparative exploratoire d'<i>A. thaliana</i> de génotypes sauvage et mutant <i>hmgr1-1</i>	171
4.1. Bilan des identifications	171
4.2. Une « sous-population » de phosphopeptides révélée par le marquage TMPP	173
4.2.1. Comparaison des sites de phosphorylation identifiés avec la base de données « PhosPhAt »	173
4.2.2. Impact du marquage TMPP sur l'identification des phosphopeptides par les moteurs de recherche	173
4.2.3. Comparaison des propriétés physicochimiques des phosphopeptides identifiés sous forme native et marquée au TMPP	177
4.2.3.1. Masse moléculaire	177
4.2.3.2. Etats de charge des ions précurseurs	177
4.2.3.3. Composition en acides aminés, nombre de phosphorylations et charge « globale »	178
4.2.3.4. Hydrophilicité	179

5. Signification biologique des différences observées entre les deux phosphoprotéomes.....	181
5.1. Localisation cellulaire prédite des phosphoprotéines d'abondance différente.....	182
5.2. Fonction biologique prédite des phosphoprotéines d'abondance différente.....	182
Conclusion	185
Bibliographie	187
Partie III : Protéomique Quantitative	193
Chapitre 1 : La spectrométrie de masse quantitative en analyse protéomique.....	193
1. La quantification relative en analyse protéomique.....	193
1.1. Les méthodes de marquage par isotopes stables	194
1.1.1. Marquage métabolique.....	195
1.1.2. Marquage chimique.....	196
1.1.2.1. Stratégies de marquage isotopique	196
1.1.2.2. Stratégies de marquage isobarique	197
1.1.3. Marquage par réaction enzymatique	197
1.2. Les méthodes sans marquage (« label-free »)	198
1.2.1. Comparaison des intensités des signaux MS.....	198
1.2.2. Comparaison des nombres de spectres (« spectral counting »).....	199
2. La quantification absolue en analyse protéomique.....	201
2.1.1. Les stratégies de dilution isotopique	201
2.1.2. Les méthodes alternatives	203
Chapitre 2 : Méthode de quantification protéomique par comptage de spectres ; application à <i>Geobacter metallireducens</i>	207
1. Contexte de l'étude.....	207
1.1. Métabolisme anaérobie des composés aromatiques	207
1.2. <i>Geobacter metallireducens</i>	208
1.3. Objectif de l'étude	209
2. Stratégie d'analyse protéomique.....	209
2.1. Préparation des échantillons	210
2.2. Analyses par spectrométrie de masse	210
2.3. Quantification des protéines identifiées.....	211
3. Résultats publiés	211
4. Etude complémentaire	213
4.1. Objectif.....	213
4.2. Mise en œuvre	213
4.3. Résultats publiés.....	214
5. Conclusion et perspectives.....	215
Chapitre 3 : Validation et application d'une méthode de quantification protéomique avec marquage au $^{13}\text{C}_9$-TMPP. ...	217
1. Développement et validation de la méthode de quantification (« quantitative N-terminal Oriented Proteomics », « qN-TOP »)	217
1.1. Principe	218
1.2. Validation de la méthode de quantification qN-TOP sur des protéines modèles	219
1.2.1. Comparaison de plusieurs modes d'analyse et de traitement pour la quantification relative des peptides N-terminaux.....	219
1.2.1.1. Mode d'acquisition automatique standard.....	219
1.2.1.2. Mode d'acquisition pDRE.....	222

1.2.2. Linéarité, justesse et reproductibilité de la méthode de quantification qN-TOP	223
2. Application de la stratégie qN-TOP à l'étude différentielle de processus protéolytiques	226
2.1. Contexte de l'étude.....	226
2.1.1. Les processus protéolytiques.....	226
2.1.2. L'étude des processus protéolytiques.....	227
2.1.3. Le modèle du jeûne prolongé : adaptations métaboliques.....	227
2.1.4. Objectif de l'étude.....	227
2.2. Stratégie d'analyse	228
2.2.1. Préparation des échantillons.....	228
2.2.2. Etude différentielle des processomes hépatiques chez le rat dénutri.....	228
2.2.3. Etude différentielle des protéomes hépatiques chez le rat dénutri	230
2.3. Résultats préliminaires	232
2.3.1. Etude différentielle des processomes hépatiques chez le rat dénutri.....	232
2.3.1.1. Identification des protéines.....	232
2.3.1.2. Identification des processus protéolytiques.....	232
2.3.1.2.1. Les peptides marqués en N-terminal au TMPP.....	232
2.3.1.2.2. Les peptides marqués aspécifiquement au TMPP.....	233
2.3.1.2.3. Perspectives	234
2.3.1.3. Quantification relative des peptides N-terminaux	235
2.3.2. Etude différentielle des protéomes hépatiques chez le rat dénutri	236
2.3.2.1. Identification des protéines.....	236
2.3.2.2. Quantification relative des protéines	238
2.4. Conclusion et perspectives.....	239
Conclusion	241
Bibliographie	243
CONCLUSION GENERALE	249
PARTIE EXPERIMENTALE	253
1. Partie I des résultats.....	253
1.1. Chapitre 1.....	253
1.2. Chapitre 2.....	253
1.3. Chapitre 3.....	254
1.4. Chapitre 4.....	254
2. Partie II des résultats	255
2.1. Chapitre 2.....	255
2.1.1. Préparation des fractions protéiques microsomales	255
2.1.2. Solubilisation et digestion des fractions protéiques microsomales.....	255
2.1.3. Dessalage des peptides.....	255
2.1.4. Marquage TMPP.....	256
2.1.5. Enrichissement des phosphopeptides par IMAC fer.....	256
2.1.6. Enrichissement des phosphopeptides par TiO ₂	256
2.1.7. Précipitation des phosphopeptides au phosphate de calcium.....	257
2.1.8. Analyses nanoLC-MS/MS.....	257
2.1.9. Fractionnement RP-HPLC à pH basique	258
2.1.10. Identification des phosphopeptides.....	258
2.1.11. Quantification relative des phosphopeptides entre les extraits protéiques d' <i>A. thaliana</i> sauvage et mutant <i>hmgr1-1</i> par approche spectral counting	258
3. Partie III des résultats.....	260
3.1. Chapitre 2.....	260
3.2. Chapitre 3.....	260
3.2.1. Expériences nanoLC-MS/MS dans l'étude différentielle des protéomes hépatiques chez le rat dénutri.....	260

3.2.2. Expériences nanoLC-MS/MS dans l'étude différentielle des processomes hépatiques
chez le rat dénutri..... 261

TRAVAUX SUPPLEMENTAIRES 263

INTRODUCTION GENERALE

INTRODUCTION GENERALE

Depuis la première annotation génomique complète d'un organisme cellulaire (la bactérie *Haemophilus influenzae* in 1995), plus de 950 séquences génomiques complètes ont pu être établies à ce jour (juillet 2009, <http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>). En revanche, le travail de prédiction des séquences protéiques réellement exprimées à l'aide de logiciels bioinformatiques est encore inachevé et demande à être validé. Les travaux de séquençage et d'annotation des génomes sont encore en pleine expansion avec plus de 2000 espèces actuellement à l'étude. L'important volume de données issues de ces projets, et particulièrement les banques de séquences protéiques, fournit aux biologistes une base de travail très précieuse. Toutefois, l'identification des séquences codantes contenues dans ces génomes réalisée *in silico*, par prédiction informatique automatisée, n'est pas exempte d'erreurs. L'introduction des erreurs au niveau des banques de séquences protéiques est parfois sous-estimée, voire ignorée et peut être très dommageable d'une part pour d'autres études *in silico* (effet domino sur des études phylogénétiques, des analyses de domaines fonctionnels,...) et d'autre part pour des études expérimentales (expression de protéines avec étude de fonctionnalité,...).

En parallèle, ces dernières années, l'analyse protéomique par spectrométrie de masse a considérablement progressé grâce à des développements technologiques jusqu'à devenir la méthode de choix pour l'étude des protéines. Les domaines d'application de la protéomique sont aujourd'hui très vastes et parmi ceux-ci un nouveau domaine est en pleine émergence : la protéogénomique dans laquelle la protéomique sert directement d'outil d'aide à l'annotation génomique. En effet, si le plus souvent l'étude du protéome repose sur la connaissance préalable du génome, l'analyse des protéines peut, à l'inverse, aider à l'analyse structurale du génome.

La partie bibliographique de ce manuscrit est consacrée à la description et la constitution (séquençage et annotation des génomes) des banques de séquences protéiques et à la présentation des outils et stratégies d'identifications utilisées en analyse protéomique.

Dans ce contexte mon travail de thèse a principalement consisté à développer de nouvelles approches d'aide à l'annotation génomique, avec un intérêt particulier porté sur la détermination exacte des codons d'initiation des protéines, basées sur les méthodologies de la protéomique.

La **première partie des résultats** est consacrée à cet objectif :

- le **premier chapitre** est un chapitre méthodologique présentant la mise en place et l'application d'une stratégie de protéogénomique destinée à la correction d'erreur de séquençage du génome de *Mycobacterium smegmatis*.
- Le **deuxième chapitre** présente le développement d'une nouvelle méthode de protéogénomique basée sur la détermination des codons d'initiation des protéines (stratégie dénommée « N-Terminal Oriented Proteomics », N-TOP) grâce à une

stratégie de marquage chimique originale des protéines et une stratégie d'identification optimisée des peptides dans les banques. La stratégie N-TOP a pu être appliquée à la correction/validation des annotations du génome de *M. smegmatis* (et par extension du génome de l'ensemble des mycobactéries).

- Le **troisième chapitre** porte sur l'étude d'une enzyme issue d'un organisme dont le génome n'est pas séquencé. La combinaison de la stratégie N-TOP avec une stratégie de séquençage *de novo* des peptides analysés par nanoLC-MS/MS et d'analyse RT-PCR a rendu possible l'annotation exacte de la séquence de l'enzyme cholest-4-en-3-one- Δ^1 -dehydrogenase impliquée dans une étape-clé du métabolisme anoxique du cholestérol chez *Sterolibacterium denitrificans*.
- Le **quatrième chapitre** aborde la problématique de la métaprotéogénomique. L'utilisation de la stratégie d'identification optimisée des peptides dans les banques a permis l'identification de protéines exprimées dans une communauté bactérienne présente sur un site pollué à l'arsenic et a démontré l'intérêt de la protéogénomique pour aider à l'assemblage et à la mise en évidence de zones manquantes d'un métagénome (ensembles des génomes des différents organismes présents dans une communauté, ici 7).

Les approches développées dans le premier chapitre, et particulièrement la stratégie N-TOP, ont permis d'améliorer la qualité des banques protéiques. Cette stratégie a également permis d'ouvrir des perspectives dans deux autres champs très importants de la protéomique qui sont la caractérisation et la quantification des protéines et qui sont explorés respectivement dans la deuxième et la troisième partie des résultats.

La **deuxième partie des résultats** est consacrée au développement d'une nouvelle stratégie d'analyse phosphoprotéomique employant le marquage utilisé dans la stratégie N-TOP :

- le **premier chapitre** décrit la problématique de l'analyse des phosphorylations par spectrométrie de masse et présente les difficultés de ce type d'analyse et les principales stratégies développées pour y remédier.
- Le **deuxième chapitre** présente le développement d'une nouvelle stratégie d'analyse phosphoprotéomique employant le marquage utilisé dans la stratégie N-TOP. La stratégie a pu être appliquée à l'analyse phosphoprotéomique d'*Arabidopsis thaliana* et a permis d'augmenter la couverture du phosphoprotéome étudié.

La **troisième partie des résultats** est consacrée à la protéomique quantitative :

- le **premier chapitre** est consacré à la description des méthodologies de quantification relative et absolue en analyse protéomique.
- Le **deuxième chapitre** présente l'application d'une stratégie de quantification sans marquage, par comptage des spectres (« spectral counting ») à l'analyse du protéome de

Geobacter metallireducens cultivé en condition anaérobie avec ou sans source de carbone aromatique. Cette quantification qui a permis d'améliorer la compréhension des mécanismes impliqués dans le métabolisme des composés aromatiques chez les bactéries anaérobies obligatoires. Elle a également donné lieu à des analyses protéomiques complémentaires qui ont participé à la caractérisation complète d'une nouvelle enzyme clé de cette voie métabolique.

- Le **troisième chapitre** décrit le développement d'une nouvelle méthode de quantification dérivée de l'approche N-TOP que nous avons baptisée approche qN-TOP (« quantitative N-terminal Oriented Proteomics »). Cette méthode a pu être appliquée dans l'étude différentielle des processomes hépatiques chez le rat dénutri. Elle a été couplée dans ce projet à une étude différentielle des protéomes hépatiques chez le rat dénutri réalisée par une approche spectral counting. Les résultats présentés dans ce chapitre sont des résultats préliminaires.

La réalisation de l'ensemble de ces travaux repose sur une série d'améliorations portant sur la réalisation des expériences de LC-MS/MS et sur l'exploitation par des moyens informatiques des spectres MS/MS générés.

Nous avons tout au long de ce travail du réaliser une série d'optimisations à tous les niveaux :

- Préparation de l'échantillon.
- Micro- et nano-chromatographie avec système de préconcentration.
- Paramétrage de l'acquisition des spectres MS et MS/MS.
- Utilisation des moteurs de recherche et conceptions d'outils pour l'exploitation plus sûre et plus efficace des données MS/MS dans l'identification des protéines avec des taux de recouvrement plus élevés.

Nous avons choisi de décrire toutes ces optimisations et améliorations non pas dans un chapitre unique, mais de façon « dispersée » dans les différents chapitres, au moment où nous avons été amené à les réaliser. Ce sont ces optimisations et améliorations qui contribuent à faire progresser l'art de l'analyse protéomique qui est au cœur de notre travail.

PARTIE BIBLIOGRAPHIQUE

Chapitre 1 : Les banques de séquences protéiques

Chapitre 2 : Le séquençage et la problématique de l'annotation des génomes

Chapitre 3 : Les outils de l'analyse protéomique

Chapitre 4 : Les stratégies d'identification en analyse protéomique

Introduction

De nos jours, en biologie, les banques de séquences protéiques jouent un rôle vital de centre de ressource pour le stockage des informations liées aux protéines comme leur séquence primaire, la nature de leurs modifications post-traductionnelles, leurs différents domaines ou encore leur fonction. L'information minimale nécessaire à l'incorporation d'une protéine dans une banque est la connaissance de sa séquence primaire. Les séquences protéiques se sont lentement accumulées dans les banques depuis le séquençage de la première protéine, l'insuline, par Sanger dans les années 50 [Sanger, 1959] qui lui valut le prix Nobel en 1958. A partir de 1975 et l'apparition de la méthode de séquençage de l'ADN [Sanger, 1988] qui valut un deuxième prix Nobel à Frederick Sanger, la détermination directe de la séquence protéique a été supplantée par la traduction des séquences d'ADN en protéines [Boguski, 1999]. Par la suite, les projets de séquençage massif des génomes et le développement de la bioinformatique ont alimenté les banques protéiques qui ont subi une croissance exponentielle jusqu'à aujourd'hui [Apweiler et al., 2004; Domon et al., 2006].

En parallèle, ces dernières années, l'analyse protéomique par spectrométrie de masse couplée aux méthodes de séparation des protéines et/ou des peptides a considérablement progressé grâce à des développements technologiques jusqu'à devenir la méthode de choix pour la caractérisation des protéines [Mann et al., 2001; Domon et al., 2006; Cravatt et al., 2007]. L'analyse protéomique permet une analyse directe de la séquence primaire de la protéine contrairement aux méthodes de traduction de l'ADN. De plus, grâce à l'élaboration de stratégies finement adaptées et faisant appel à la chimie ou non, la protéomique permet également d'accéder à des informations non disponibles au niveau de l'ADN, comme l'identification de modifications post-traductionnelles des protéines, leurs interactions, leur localisation ou encore la vérification de leur expression voire leur degré d'expression [Patterson et al., 2003; Sickmann et al., 2003; Domon et al., 2006; Cravatt et al., 2007].

Une revue bibliographique sur les banques de séquences protéiques et leur constitution sera présentée dans les deux premiers chapitres de cette partie bibliographique. Les outils et les stratégies d'identification en analyse protéomique ainsi que leurs méthodes de valorisation associées seront approfondies dans le troisième et le quatrième chapitre.

Définitions

Le protéome et la protéomique :

Le terme « protéome », introduit pour la première fois en 1994 par Mark Wilkins lors d'une conférence en Italie, définit l'ensemble des protéines présentes dans un milieu biologique [Wilkins et al., 1996].

Le terme « protéomique », dont la définition a été longuement discutée, signifie aujourd'hui : « l'étude de l'ensemble des protéines exprimées dans une cellule incluant leurs isoformes, les modifications qu'elles peuvent subir, leurs interactions avec d'autres protéines, leur description structurale et les complexes qu'elles peuvent former ; donc de ce fait tout ce qui est post-génomique. » [Tyers et al., 2003].

Chapitre 1 : Les banques de séquences protéiques

Les banques de séquences biologiques sont de deux types : les banques nucléotidiques et les banques protéiques ; les banques protéiques découlant des banques nucléotidiques. Ce chapitre sera focalisé sur les banques de séquences protéiques. Historiquement, le but de ces banques protéiques n'était pas d'établir une collection complète des séquences disponibles, comme dans les cas des grandes banques nucléotidiques, mais de constituer une ressource de connaissance centrée sur les protéines et leurs propriétés biologiques. Pour ce faire, il est nécessaire d'appliquer des critères de qualité telle que la non-redondance, l'homogénéité de l'information ou la valeur scientifique de l'annotation. Nous verrons au cours de ce chapitre comment l'explosion des projets de séquençage massifs de l'ADN a pu mettre à mal cet objectif de « qualité » au moins pour certaines banques.

1. Origine, historique et alimentation des banques protéiques

Frederick Sanger en séquençant la première protéine (l'insuline) dans les années 50 fournit la première pierre de l'énorme édifice que sont les banques protéiques de nos jours. A partir de cette date, les séquences s'accumulèrent lentement dans la littérature. Au début des années 60, Margaret Dayhoff créa la première banque de séquences protéiques qui n'était pas informatisée mais publiée dans *Atlas of Protein Sequence and Structure* [Dayhoff et al., 1965].

A partir de 1975 et l'apparition de la méthode de séquençage de l'ADN, la détermination directe de la séquence protéique a été supplantée par la traduction conceptuelle des séquences d'ADN en protéines en utilisant le code génétique. Pendant les quinze années qui ont suivi, les séquences protéiques déterminées étaient si nouvelles et significatives que chaque nouvel exemple faisait l'objet d'une publication accompagnée d'une interprétation très détaillée issue de l'hypothèse biologique ou du contexte qui avaient conduit à l'isolation et au clonage du gène d'intérêt. Cette période correspondit à l'ère du clonage fonctionnel pendant laquelle les banques protéiques étaient restreintes mais de très grande qualité. Au début des années 90, avec l'annonce du projet de séquençage du génome humain et le développement de stratégies pour le séquençage de gènes rapide, on a pu assister à un accroissement très rapide du nombre de séquences recensées. Des études qui rapportaient le séquençage de centaines ou de milliers de gènes ont commencé à apparaître dans la littérature. Toutefois la qualité d'annotation des nouvelles protéines a commencé à fortement décliner, ces protéines étant qualifiées de « nouvelles » ou « homologues à » en comparaison des séquences déjà présentes dans les banques. Le début des années 90 est également marqué par les premières productions massives de séquences d'EST (« Expressed sequence Tags ») [Adams et al., 1991] suivies par les premiers chromosomes [Oliver et al., 1992]. En 1995, la première annotation génomique complète d'un organisme cellulaire

(la bactérie *Haemophilus influenzae*) fut publiée [Fleischmann et al., 1995]. Cet accomplissement annonçait la disponibilité croissante de séquences génomiques complètes à venir contenant plusieurs milliers voire plusieurs dizaines de milliers de gènes prédits qui seront responsables du taux de croissance exponentiel du volume des banques. Les banques de séquences ont commencé à « accumuler » des protéines sous la dénomination « ORF » (Open Reading Frame, conceptual translation) ou « hypothetical protein » la plupart du temps prédites par annotation automatique *in silico*. Bien que ces nouvelles données aient été précieuses pour la biologie et pour l'émergence de la génomique comparative, elles ont aussi constitué un immense défi en termes d'annotation et d'expériences biologiques.

Un point de non-retour fut atteint quand la quantité de séquences génomiques connues ne permit plus d'espérer que les méthodes expérimentales traditionnelles puissent répondre à ce défi. Face à cet état de fait, beaucoup de chercheurs eurent recours aux approches bioinformatiques seules pour « valoriser » ces données. Dans ce cas, il est prudent de voir la plupart des annotations comme des hypothèses pouvant être incomplètes, trompeuses voire incorrectes. De même, l'annotation fonctionnelle des nouvelles protéines est réalisée par comparaison des séquences de celles-ci avec les séquences des protéines déjà présentes dans les banques et par transfert d'annotation entre protéines homologues [Frishman, 2007]. En juillet 2009, on compte plus de 950 séquences génomiques complètes disponibles (<http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>). Les travaux de séquençage et d'annotation des génomes sont encore en pleine expansion avec plus de 2000 espèces actuellement à l'étude. L'augmentation de la vitesse de séquençage et la baisse des coûts de séquençage (on annonce le séquençage du génome d'un individu pour 100 dollars prochainement) devraient encore favoriser l'augmentation exponentielle des génomes complets disponibles. L'apparition récente de la métagénomique, l'étude à large échelle des génomes de communautés bactériennes existant naturellement (plutôt que l'étude d'organismes isolés cultivés en laboratoire), devrait aussi participer à cette augmentation [Bertin et al., 2008].

Actuellement, la quasi-totalité des séquences protéiques qui sont intégrées dans les banques sont issues de traductions conceptuelles de séquences nucléotidiques obtenues expérimentalement qui sont la plupart du temps issues des projets de séquençage massifs. Ces séquences sont la plupart du temps intégrées automatiquement, sans vérification ni annotation fonctionnelle. Toutefois, quelques banques, comme Swiss-Prot [Boeckmann et al., 2003], fournissent un gros effort de validation, correction, annotation de ces séquences pour leur apporter une haute valeur ajoutée.

Pour exploiter pleinement les différentes banques, il est nécessaire de connaître le type d'information qu'elles contiennent. Le paragraphe suivant a pour objectif de présenter les principales banques protéiques.

Figure 1 : Les grandes étapes de l'évolution des techniques qui ont conduit aux banques protéiques d'aujourd'hui

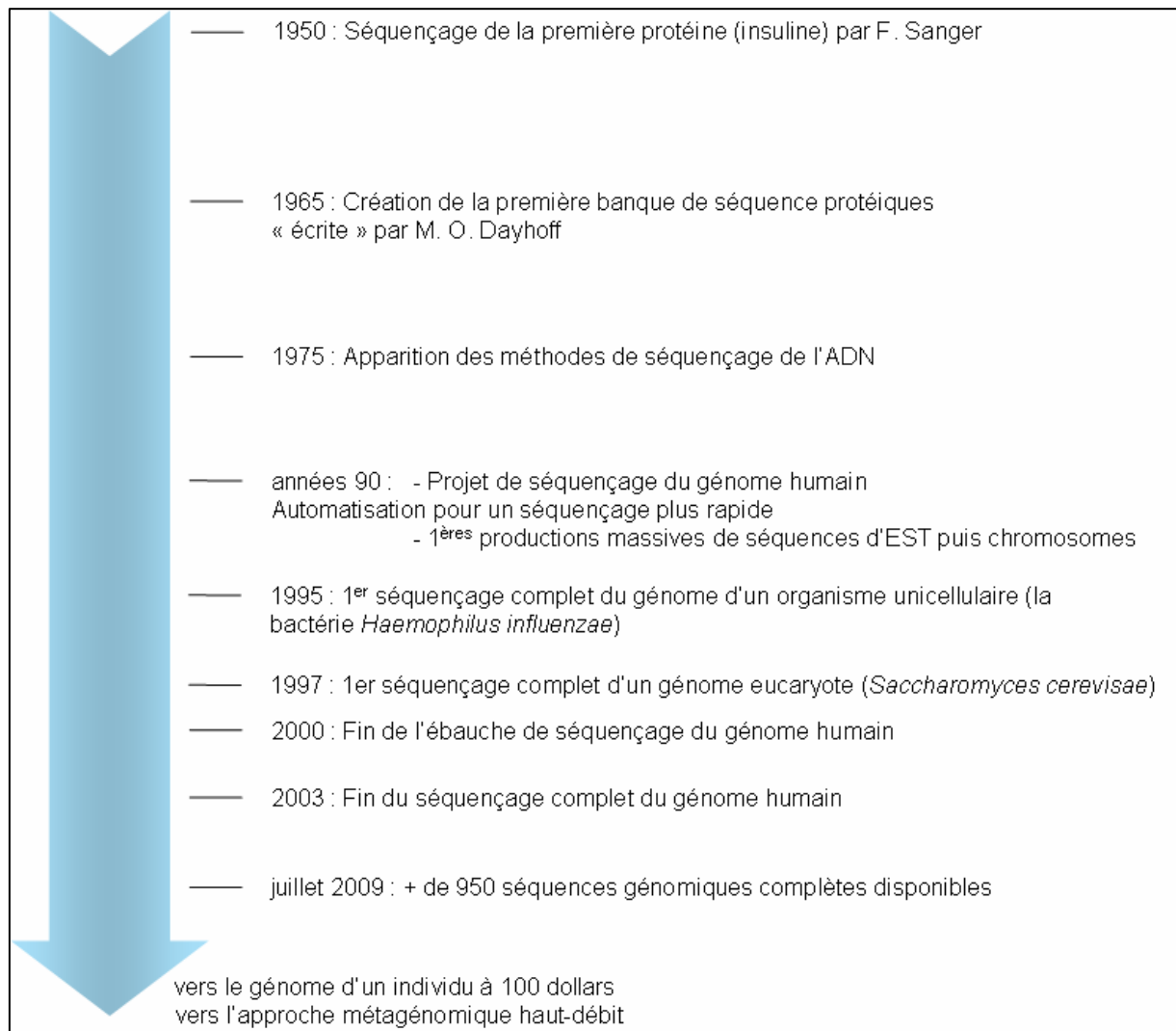
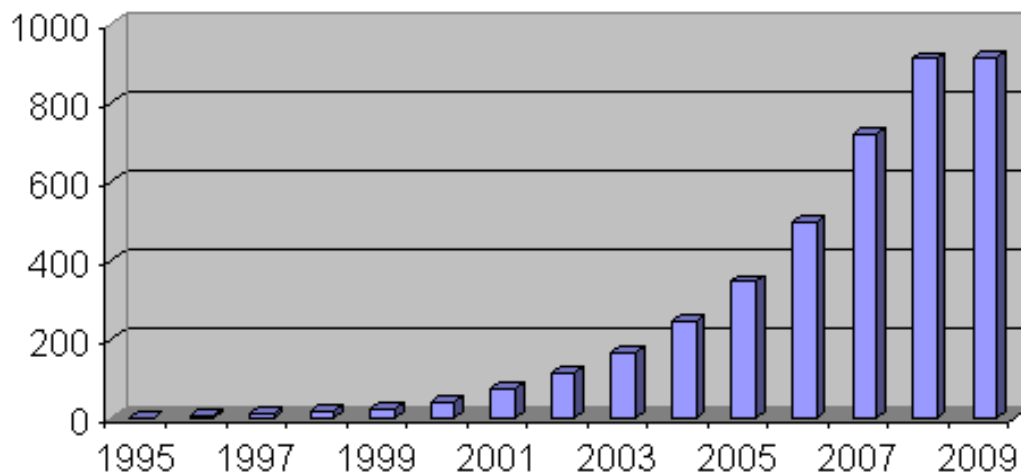


Figure 2 : Augmentation exponentielle du nombre de génomes séquencés entre 1995 et 2009, d'après les statistiques du site internet GOLD (<http://www.genomesonline.org>)



2. Les différentes banques protéiques

Il existe une grande variété de banques protéiques qui diffèrent selon leur redondance, leur exhaustivité, l'homogénéité des informations et la valeur scientifique des informations qui s'y trouvent [Apweiler et al., 2004]. C'est ainsi qu'on va pouvoir distinguer les banques de dépôt et les banques dites « soignées ».

2.1. Les banques de dépôt

Ces banques stockent les séquences protéiques, c'est-à-dire la séquence primaire de la protéine issue de la traduction du gène prédit sans vérification. Elles ne contiennent que très peu d'informations additionnelles (annotations fonctionnelles par exemple) et présentent un haut degré de redondance.

2.1.1. La banque GenPept

La banque GenPept (GenBank Gene Products Data Bank) produite par le National Center of Biotechnology Information (NCBI) [Wheeler et al., 2003] est sans doute l'exemple de la banque de dépôt la plus basique. Les séquences des protéines répertoriées découlent directement de la traduction des séquences présentes dans la banque nucléotidique gérée par GenBank [Benson et al., 2003], la European Molecular Biology Laboratory (EMBL) Nucleotide Sequence Database [Stoesser et al., 2003] et la DNA Data Bank of Japan (DDBJ) [Miyazaki et al., 2003] et contiennent des annotations minimales directement extraites des banques nucléotidiques. Les protéines manquent d'annotations additionnelles et la banque ne contient pas de séquences issues d'analyses directes de protéines. Enfin, le degré de redondance y est élevé, les protéines peuvent être répertoriées plusieurs fois. Aucun travail de regroupement des enregistrements multiples n'est effectué.

2.1.2. La banque NCBI's Entrez Protein

La banque NCBI's Entrez Protein est un autre exemple de banque de dépôt. Elle rassemble les séquences de la banque GenPept mais aussi les séquences annotées issues des banques Swiss-Prot [Boeckmann et al., 2003], Protein Information Ressource (PIR) [Wu et al., 2003], Protein Data Bank (PDB)[Westbrook et al., 2003] et Refseq [Pruitt et al., 2007]. Pour ces dernières, les annotations additionnelles sont également extraites des banques soignées. Le degré de redondance de cette banque est élevé mais elle présente l'avantage d'être relativement complète.

2.1.3. La banque Refseq

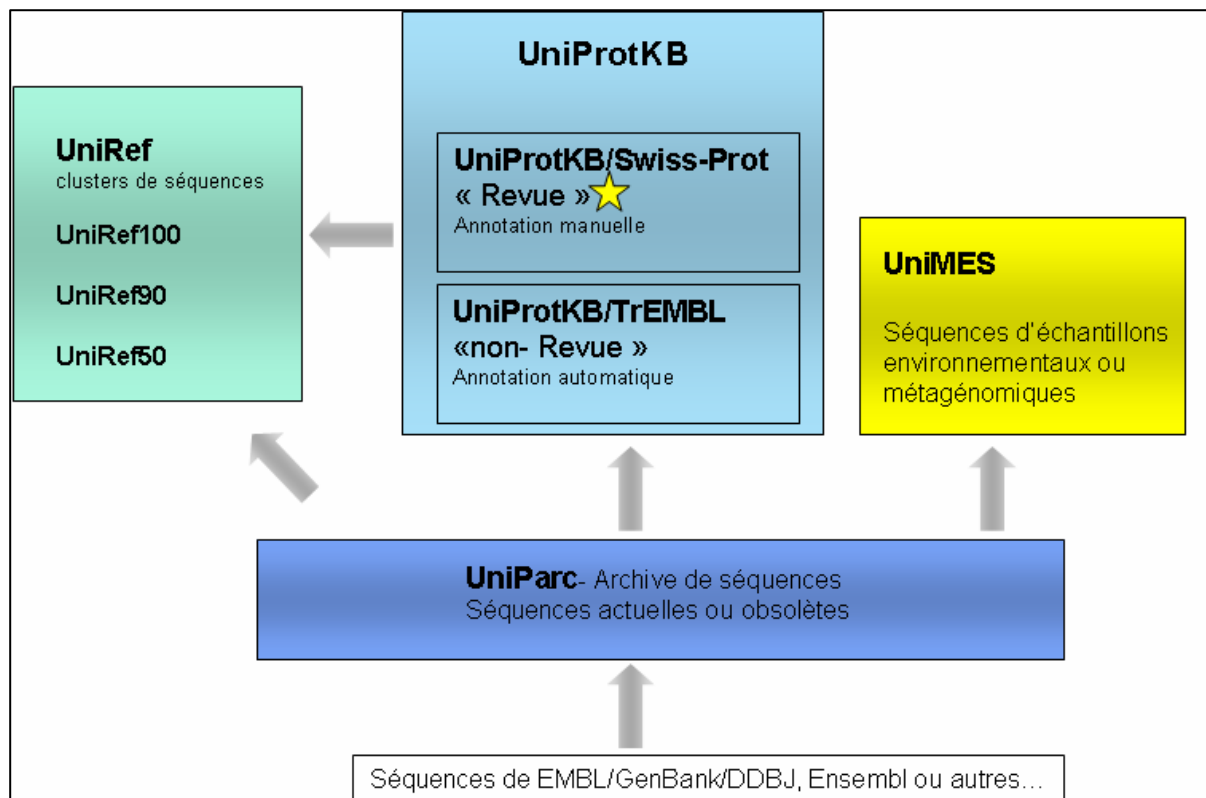
Cette banque produite par le NCBI se trouve à mi-chemin entre les banques de dépôt et les banques soignées [Pruitt et al., 2009]. Cette banque fut créée pour fournir un ensemble non redondant de séquences protéiques de référence. Elle répertorie les séquences d'un nombre limité d'espèces (plus de 6000 fin 2008). Cette banque est non redondante, fait le lien entre séquences protéiques et séquences nucléotidiques et est mise à jour régulièrement pour refléter l'état actuel de la connaissance en biologie. Les entrées présentent un numéro d'accès distinct et un format uniformisé. Les données sont validées et l'état d'avancement de cette validation par l'équipe du NCBI est indiqué. Toutefois, la majorité des enregistrements est générée automatiquement avec une intervention manuelle minimale. En juin 2008, la banque contenait 5 590 364 entrées dont 265 002 entrées revues manuellement (4.7 %). Ce très faible taux de séquences revues rend cette banque plus proche d'une banque de dépôt que d'une banque soignée.

2.2. Les banques soignées et le consortium UniProt

Bien que les banques de dépôt constituent un moyen de mise à disposition très rapide des séquences protéiques dès leur obtention (génomé annoté), il est clair que l'ajout d'informations additionnelles aux séquences primaires apporte une valeur ajoutée incontestable aux données. Dans les banques soignées, cet ajout d'informations additionnelles, qui ont été validées par les biologistes, est un préalable avant la mise à disposition de la séquence primaire pour s'assurer que les données présentes sont fiables. D'importants efforts sont également réalisés pour maintenir la non-redondance de ces banques.

UniProt symbolise la mise en commun de ces efforts réalisés par les responsables des banques « soignées » pour créer une banque à haute valeur ajoutée. En 2003, le Swiss Institute of Bioinformatics (SIB), l'European Bioinformatics Institute (EBI) et l'institut Protein Information Resource (PIR) ont regroupé leurs données et leurs compétences pour créer le consortium Uniprot. Ce consortium est le fruit du regroupement des banques Swiss-Prot, TrEMBL et PIR [Apweiler et al., 2004] et présente 3 composantes : la banque « UniProt knowledgebase » (UniProtKB), la banque « UniProt Archive » (UniParc) et la banque « UniProt reference clusters » (UniRef). Récemment, une quatrième composante s'est rajoutée au trois composantes d'origine, la banque « UniProt Metagenomic and Environmental Sequences » (UniMES) (Figure 3).

Figure 3 : Organisation d'UniProt



2.2.1. Les principales banques « soignées » à l'origine d'UniProt

2.2.1.1. La banque PIR-PSD (Protein Information Resource Protein Sequence Database)

Il s'agit de la plus ancienne des banques « soignées » établie en 1984 pour succéder à la National Biomedical Research Foundation Protein Sequence Database développée par Margareth O Dayhof pendant 20 ans et qui était publiée en format papier comme l' « Atlas of Protein Sequence and Structure » de 1965 à 1978 [Dayhoff et al., 1965]. Cette banque non-redondante est organisée selon une classification par familles et super-familles de protéines et annotée avec des données fonctionnelles, structurales, bibliographiques (liens vers Pubmed et Medline) et génétiques. La classification de cette banque aide à détecter et à corriger les erreurs d'annotation du génome et permet une meilleure compréhension des relations séquence-fonction-structure.

2.2.1.2. La banque Swiss-Prot, appelée UniProtKB/Swiss-Prot dans UniProtKB

Il s'agit de la banque « soignée » de référence. Créée en 1986 par Amos Bairoch, cette banque est la moins redondante, la mieux annotée et intégrée avec les autres banques [Gasteiger et al., 2001].

Chaque nouvelle entrée dans Swiss-Prot fait l'objet d'une analyse et d'une annotation poussée par les annotateurs pour assurer le maintien d'une banque de haute qualité. Les données relatives aux séquences sont extraites de la littérature et contrôlées par des programmes d'analyse de séquences. Les annotations vont des propriétés physicochimiques de la protéine jusqu'aux maladies associées avec une sous/sur-expression de cette protéine en passant par la fonction, les modifications post-traductionnelles, les sites et les domaines les structures secondaires et quaternaires de la protéine, les similarités avec d'autres protéines, le stade de développement durant lequel la protéine est exprimée, les tissus dans lesquelles la protéine est identifiée, les voies métaboliques dans lesquelles la protéine est impliquée, les conflits et les variants de séquences. La création d'une entrée totalement annotée dans Swiss-Prot est un processus assez long car nécessitant une intervention manuelle poussée. Les annotateurs de Swiss-Prot ne pouvant pas suivre le rythme d'accroissement de la quantité de séquences protéiques issues de la traduction des séquences nucléotidiques et maintenir le haut degré de « qualité » de la banque, la banque TrEMBL fut créée en complément de Swiss-Prot.

2.2.1.3. La banque TrEMBL (« Translation from EMBL) appelée UniProtKB/TrEMBL dans UniProtKB

Cette banque contient les séquences issues de la traduction des séquences présentes dans les banques nucléotidiques GenBank/EMBL/DDBJ qui n'ont pas encore été intégrées dans Swiss-Prot mais aussi des séquences protéiques issues de publications ou directement déposées par la communauté scientifique. Cette banque permet de rendre accessible les séquences protéiques le plus rapidement possible. Des programmes informatiques permettent d'éliminer les redondances entre les entrées en regroupant les enregistrements multiples des mêmes protéines et d'apporter une amélioration des annotations en comparant les séquences à celles des protéines bien caractérisées dans Swiss-Prot [Fleischmann et al., 1999]. Cette annotation était réalisée à l'origine par l'utilisation d'InterPro [Mulder et al., 2003], une ressource intégrée de familles, domaines et sites fonctionnels de protéines. Plus récemment, d'autres systèmes ont été mis en place comme par exemple HAMAP (High-quality Automated and Manual Annotation of Microbial Proteomes) [Lima et al., 2009] qui permet la propagation d'annotations à l'intérieur de familles de protéines ou le système d'annotation génomique par la communauté scientifique en utilisant le logiciel Wiki [Mons et al., 2008]. Ce travail d'annotation automatique permet à la banque TrEMBL de présenter des standards se rapprochant de ceux de Swiss-Prot, ce qui permet d'accélérer la vitesse d'intégration des protéines d'UniProtKB/TrEMBL à UniProtKB/Swiss-Prot.

2.2.2. Les composantes d'UniProt

2.2.2.1. UniProt Archive (UniParc) [Leinonen et al., 2004]

Cette banque est la plus complète des banques protéiques non redondantes avec accès publique. Elle contient les séquences protéiques présentes dans les banques généralistes et dans les banques « soignées » (Swiss-Prot, PIR-PSD, TrEMBL, EMBL, Ensembl [Hubbard et al., 2009], International Protein Index [Kersey et al., 2004], PDB, RefSeq, FlyBase [Wilson et al., 2008], WormBase [Harris et al., 2003]). Les protéines présentes dans plusieurs banques ou plusieurs fois dans la même banque sont regroupées sous le même identifiant UniParc en conservant les références aux banques sources. Au 28 juillet 2009, la banque Uniprot Archive comptait 20 070 606 séquences.

2.2.2.2. UniProt Knowledgebase (UniProtKB) [Wu et al., 2006]

Cette banque regroupe les séquences présentes dans Swiss-Prot, TrEMBL et PIR-PSD et est composée de deux parties : UniProtKB/Swiss-Prot et UniProtKB/TrEMBL. Au 28 juillet 2009, la banque UniprotKB comptait 9 232 223 entrées dont 8 760 751 entrées dans UniProtKB/TrEMBL et 471 472 entrées dans UniProtKB/Swiss-Prot (~5 %).

2.2.2.3. UniProt reference clusters (UniRef) [Suzek et al., 2007]

Cette banque fournit des clusters de toutes les séquences d'UniProtKB (en incluant des entrées séparées pour les épissages alternatifs) et de séquences sélectionnées d'UniParc pour obtenir une couverture complète des séquences existantes aux résolutions de 100, 90 et 50 % d'identité sans redondance. Les clusters d'UniRef fournissent un ensemble hiérarchique de clusters de séquences où chaque membre peut exister dans un seul cluster à chaque résolution et avoir un seul cluster « parent » (ou « enfant ») dans une autre résolution. La banque UniRef100 combine les séquences identiques, et les séquences représentant des fragments de séquences plus grandes, en une seule entrée avec les numéros d'accessions d'origine fusionnés. Les banques UniRef90 et UniRef50 sont construites respectivement à partir des clusters d'UniRef100 et des clusters d'UniRef90. Elles ont pour but de fournir un ensemble de séquences non-redondantes à la communauté scientifique pour effectuer des recherches d'homologies plus rapides (utiles pour l'annotation génomique, la génomique structurale, les analyses phylogénétiques, etc.). Les séquences de toutes les espèces qui ont une homologie de séquence mutuelle supérieure à 90 % ou 50 %, respectivement sont regroupées en une seule entrée (cluster) qui pointe les enregistrements des différents membres dans UniProt (banque source, nom de la protéine, informations taxonomiques) mais qui ne présente qu'une seule séquence et qu'un seul nom.

2.2.2.4. UniProt Metagenomic and Environmental Sequences (UniMES) ["The_UniProt_Consortium", 2009]

Cette banque fut créée pour répondre au développement de la métagénomique. Cette banque contient notamment les données issues du projet « Global Ocean Sampling Expedition » qui contient 6 millions de protéines provenant de bactéries océaniques. Dans cette banque, les séquences n'ont pas de taxonomie indiquée.

Chapitre 2 : Le séquençage et la problématique de l'annotation des génomes

Comme nous l'avons décrit précédemment, le séquençage des génomes à haut-débit est à l'origine de l'explosion du volume de séquences protéiques disponibles dans les banques. Dans ce paragraphe nous allons décrire sommairement ces méthodes de séquençage. Nous nous intéresserons également au processus d'annotation des génomes qui génèrent les séquences primaires protéiques alimentant les banques.

1. Le séquençage des génomes

1.1. Le séquençage de l'ADN

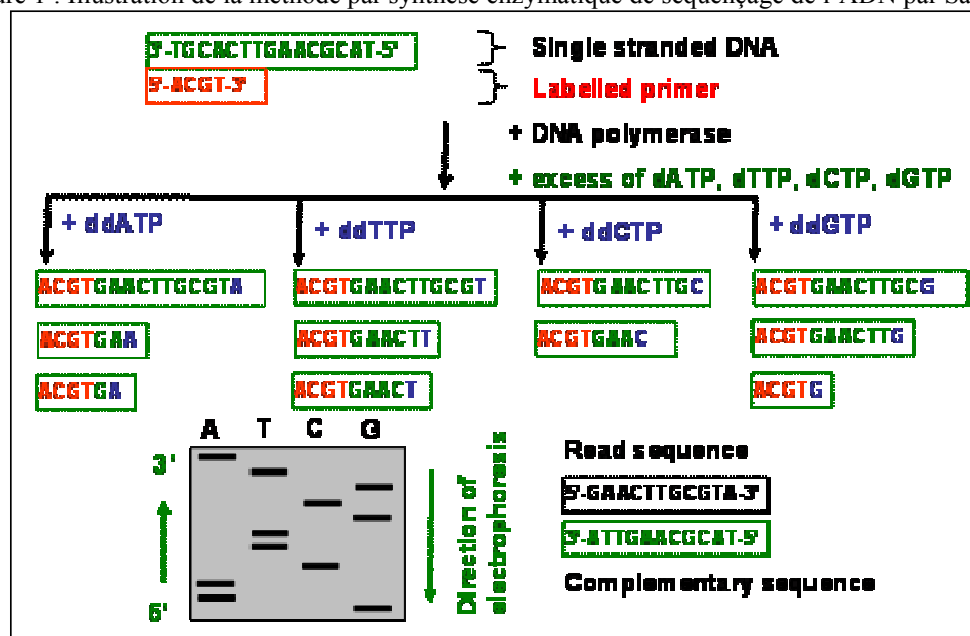
Le séquençage de l'ADN a été inventé en 1977. Deux méthodes ont été développées indépendamment et ont valu à leur inventeur le prix Nobel de chimie en 1980.

Méthode par dégradation chimique [Maxam et al., 1977] : la méthode de Maxam et Gilbert est basée sur une dégradation chimique de l'ADN et qui utilise les réactivités différentes des quatre bases A, T, G et C pour réaliser des coupures sélectives. Cette méthode nécessite des réactifs chimiques toxiques et ne permet pas de séquencer des fragments supérieurs à 250 nucléotides.

Méthode par synthèse enzymatique [Sanger et al., 1977] : la méthode de Sanger consiste à initier la polymérisation de l'ADN à l'aide d'un petit nucléotide (amorce) complémentaire à une partie du fragment d'ADN à séquencer. L'élongation de l'amorce est réalisée par le fragment de Klenow (une ADN polymérase I dépourvue d'activité exonucléase 5'→3') et maintenant par des ADN polymérases thermostables, celles qui sont utilisées pour la PCR. Les quatre désoxyribonucléotides (dATP, dCTP, dGTP, dTTP) sont ajoutés, ainsi qu'une faible concentration de l'un des quatre didésoxyribonucléotides (ddATP, ddCTP, ddGTP ou ddTTP). Ces didésoxyribonucléotides permettent de terminer les chaînes : une fois incorporés dans le nouveau brin synthétisé, ils empêchent la poursuite de l'élongation. Cette terminaison se fait spécifiquement au niveau des nucléotides correspondant au didésoxyribonucléotide incorporé dans la réaction. Pour le séquençage complet d'un même fragment d'ADN, on répète cette réaction quatre fois en parallèle, avec les quatre didésoxyribonucléotides différents (Figure 1). L'analyse des fragments résultants par gel de polyacrylamide permet d'obtenir la séquence en nucléotides de la molécule d'ADN étudiée.

De nos jours, la méthode de Sanger, en bénéficiant de développements technologiques permettant son automatisation et la baisse de son coût, reste la base des procédures de séquençage modernes excepté pour les nouvelles technologies comme le pyroséquençage [Bertin et al., 2008].

Figure 1 : Illustration de la méthode par synthèse enzymatique de séquençage de l'ADN par Sanger.



1.2. Les stratégies de séquençage des génomes

La connaissance de la structure d'un génome dans son intégralité passe par son séquençage. Cependant, la taille des génomes étant de plusieurs millions de bases (ou mégabases), il est nécessaire de coupler les approches de biologie moléculaire avec celles de l'informatique pour pouvoir traiter un nombre aussi important de données. Deux grands principes de séquençage de génomes entiers sont utilisés (Figure 2). Dans les deux cas, l'ADN génomique est préalablement fragmenté par des méthodes enzymatiques (enzymes de restriction) ou physiques (ultrasons).

La stratégie de séquençage globale (« whole genome shotgun sequencing »). Cette technique repose sur un principe simple : découper un génome en un grand nombre de fragments de petite taille. Les extrémités d'un certain nombre de ces fragments sont ensuite séquencées. Les séquences obtenues sont lues, alignées et progressivement assemblées sur la base de leurs chevauchements en des fragments plus grands (les contigs) grâce à des programmes bioinformatiques appropriés. Les contigs sont ensuite également assemblés pour aboutir à une réduction de leur nombre [Green, 2001]. Les régions pas encore séquencées à cette étape constituent des trous dans le génome qui seront remplis lors de la dernière étape du projet de séquençage. Cette étape de finition peut être réalisée par différentes méthodes notamment par le séquençage ciblé des extrémités des fragments d'ADN de plus grande taille. La difficulté d'une telle stratégie est d'obtenir suffisamment de fragments pour couvrir le

génomique dans sa totalité et réussir l'assemblage des fragments. Ainsi, plus le nombre d'extrémités de fragments séquencés sera important, plus la longueur des blocs de séquence pouvant être assemblés sera grande et plus la fraction du génome couverte augmentera ainsi que l'exactitude de la séquence (la redondance des séquences permettra de corriger d'éventuelles erreurs de séquençage). On parle de profondeur pour signifier combien de fois la taille du génome a été obtenue si on somme la totalité des séquences des fragments séquencés.

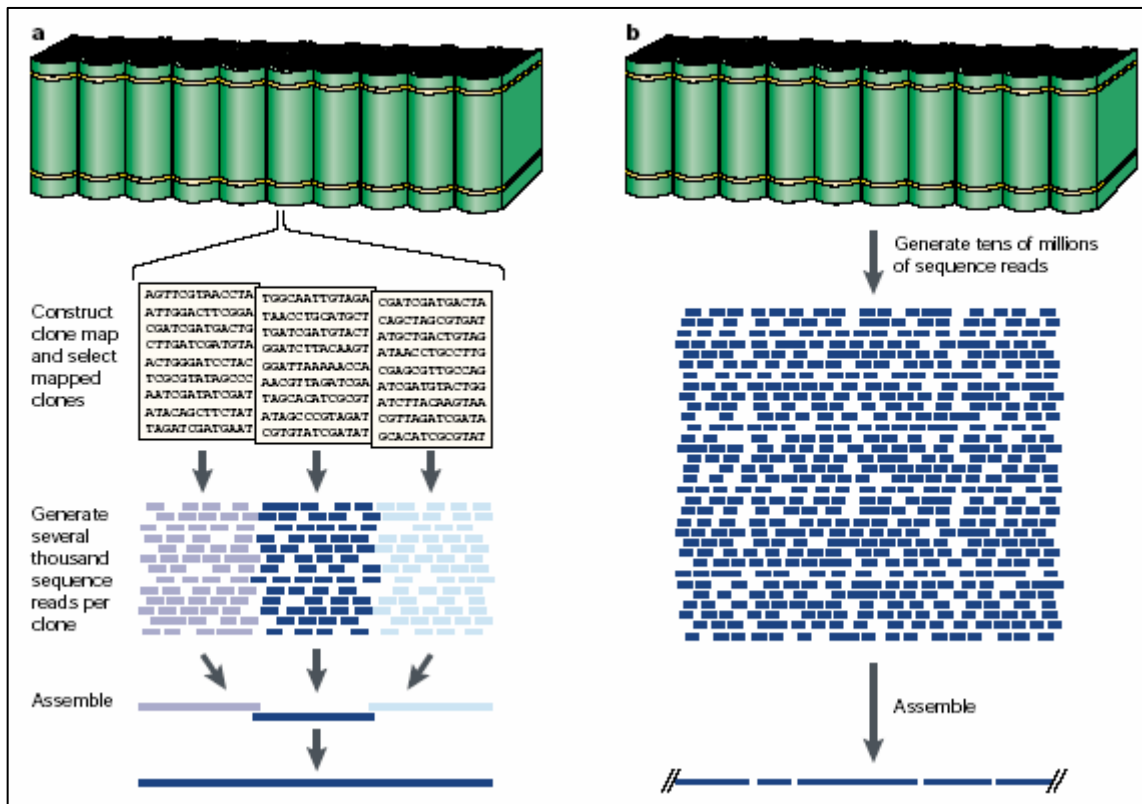
La stratégie de séquençage globale a été appliquée lors du premier séquençage complet du génome d'un organisme cellulaire, celui de la bactérie *Haemophilus influenzae* [Fleischmann et al., 1995]. Cette stratégie s'est imposée dans le cas des génomes bactériens [Fraser et al., 2000] mais son application sur les génomes eucaryotes est beaucoup plus difficile à cause de leur plus grande taille et de leur contenu plus répétitif.

La stratégie par ordonnancement hiérarchique (« clone by clone »). Cette stratégie a été adoptée par le consortium international pour le séquençage du génome humain (HGP : Human Genome Project). Il s'agit d'une démarche en deux temps : établissement d'une carte physique, ordonnant des clones de grande taille dans le génome humain, puis séquençage (de type "shotgun") de ces clones. La carte peut aussi être construite en même temps que le séquençage progresse. Une aide essentielle dans la construction d'une carte physique est apportée par les cartes de liaison. Après extraction, l'ADN génomique est découpé en fragments (de 50 à 200 kb) qui sont clonés dans un vecteur adapté comme les chromosomes artificiels bactériens (BAC). Le nombre de clones doit permettre une couverture de 5 à 10 fois la longueur totale du génome étudié. La stratégie de séquençage par ordonnancement hiérarchique a été appliquée lors du premier séquençage complet du génome d'un organisme eucaryote, celui de la levure *Saccharomyces cerevisiae* [1997] et est devenue la méthode de choix pour le séquençage de génomes eucaryotes [1998; 2000].

Aujourd'hui, pour le séquençage des génomes très complexes, une stratégie mixte combinant les deux approches décrites (séquençage aléatoire global et séquençage par ordonnancement hiérarchique) est généralement utilisée [Adams et al., 2000; Hoskins et al., 2000; Myers et al., 2000].

Figure 5 : Illustration des deux grandes stratégies de séquençage des génomes. D'après [Green, 2001]

- Séquençage par ordonnancement hiérarchique (« clone by clone »).
- Séquençage global (« whole genome shotgun sequencing »).



2. L'annotation des génomes

Une fois le processus de séquençage et d'assemblage du génome terminé, celui-ci peut ensuite être annoté. « Annoter un génome » consiste à lui rattacher les informations nécessaires à son utilisation. Dans les faits, cela se traduit par une annotation structurale (localisation des éléments génétiques), une annotation fonctionnelle (recherche de la fonction des gènes) et une intégration biologique (reconstitution de voies métaboliques, des interactions, etc.). Dans ce paragraphe nous nous intéresserons uniquement à l'annotation structurale du génome et plus particulièrement à l'annotation des gènes puisque les séquences alimentant les banques protéiques découlent directement de cette information. Dans un but de simplification, nous nous intéresserons essentiellement au cas des organismes procaryotes.

L'étape initiale de la prédiction des gènes consiste à extraire l'ensemble des cadres de lecture ouverts (« Open Reading Frames », ORFs) dans les six cadres de lecture de la séquence génomique complète. Les ORFs sont des morceaux de séquences compris entre un codon d'initiation (codon start) potentiel et un codon de terminaison (codon stop) potentiel. L'ensemble des ORFs d'une séquence génomique donnée peut être obtenu en extrayant dans les 6 cadres de lecture toutes les sous-séquences qui se terminent par un codon stop, dont le nombre de base est divisible par trois, qui ont un codon

start en première position et qui ne contiennent pas de codon stop interne (Figure 3). Alors que cette extraction des ORFs est relativement triviale, la sélection des ORFs qui codent réellement pour une protéine est plus difficile. Souvent, comme premier filtre, on commence par fixer une longueur minimale de séquence pour qu'une ORF soit extraite (par exemple 90 bases). A ce stade, on estime qu'environ 5 % des ORFs extraites codent réellement pour une protéine. Bien qu'une procédure manuelle pour la détermination de ce sous-ensemble d'intérêt puisse être appliquée, celle-ci serait très fastidieuse. Des programmes bioinformatiques ont donc été mis au point pour réaliser cette dernière partie du processus de prédiction de gènes. Deux approches de prédiction de gènes automatisées existent : les méthodes intrinsèques, aussi appelées prédiction *ab initio*, et les méthodes extrinsèques, basées sur la similarité.

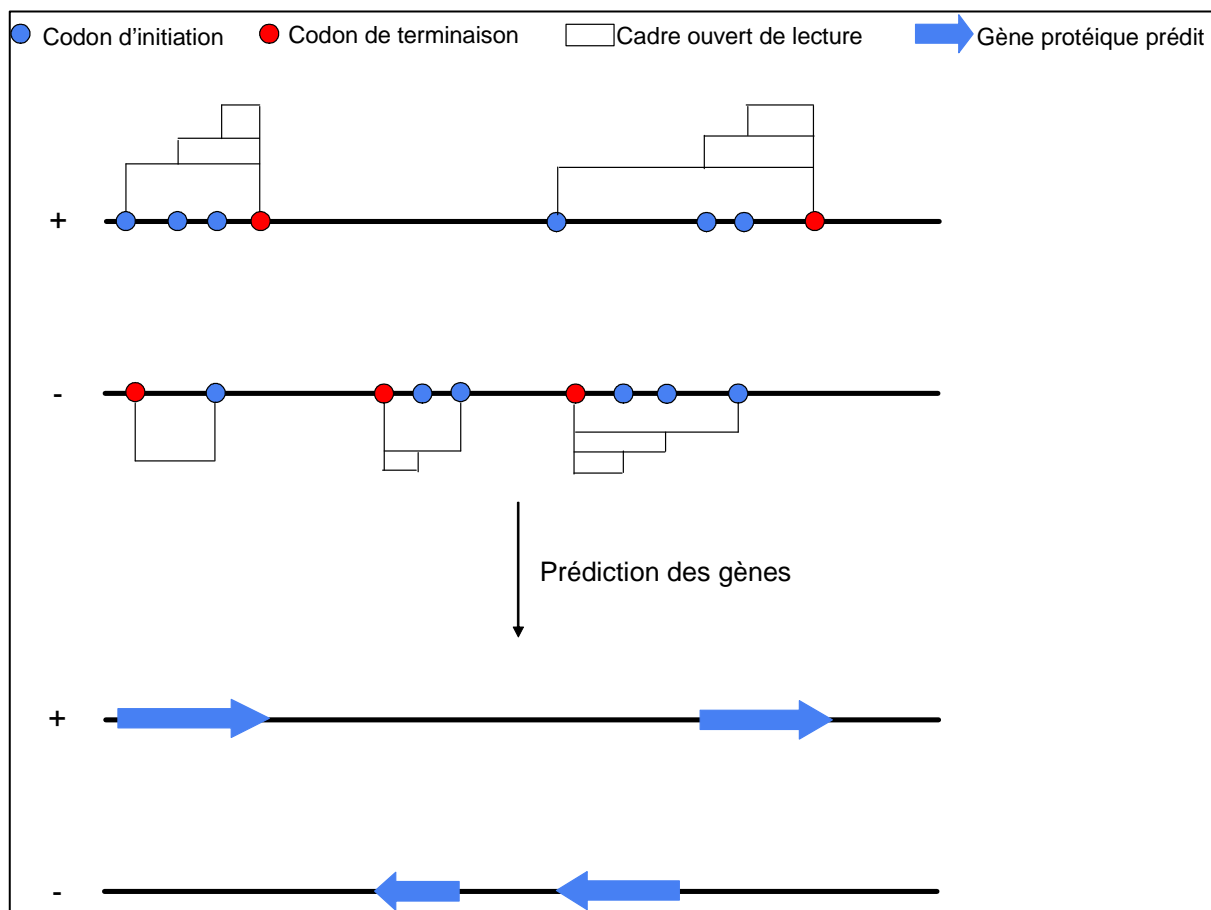
2.1. Méthodes intrinsèques (ou *ab initio*) :

Les séquences codantes présentent un certain nombre de propriétés qui les distinguent des séquences non-codantes [Fickett, 1982; Gribskov et al., 1984; Staden, 1984]. La composition en bases (régions riches en bases G et C par exemple), la composition des codons (usage des codons) ou encore la composition en acides aminés des régions codantes peuvent par exemple être utilisées pour les distinguer des ORFs non codantes.

Pour réaliser la détection automatique des séquences codantes, les techniques d'apprentissage automatique (« machine learning ») sont utilisées. Ces techniques consistent à fournir au logiciel un ensemble d'ORFs extraites du génome considérées codantes de manière certaine (ensemble de référence). Cet ensemble de référence positif (« positive training set ») est constitué de gènes connus expérimentalement ou de longues ORFs (par exemple comportant plus de 500 bases) extraites du génome qui ne présentent pas de chevauchement entre elles. Quelque fois on fournit aussi un ensemble de référence négatif au logiciel (ORFs qui sont considérées comme non-codantes de manière fiable). Typiquement, à partir des ensembles de référence, le logiciel de prédiction automatique de gènes construit des modèles définissant les séquences codantes et non codantes. L'ensemble des ORFs du génome est évalué en fonction du modèle pour les classer comme codantes ou non codantes pour des protéines. Les modèles construits par les programmes sont généralement de deux types : les modèles de Markov interpolés (« Interplated Markov Models », IMM) [Durbin et al., 1999] avec les logiciels GLIMMER 1.0 [Salzberg et al., 1998], GeneMark [Besemer et al., 2001] ou EasyGene [Nielsen et al., 2005] ou les modèles de contextes interpolés (« Interpolated Context Models », ICM) avec le logiciel GLIMMER 2.0 [Delcher et al., 1999] et GLIMMER 3.0 [Delcher et al., 2007].

L'ensemble d'apprentissage a un impact très important sur l'efficacité des prédictions *ab initio*. Les résultats seront meilleurs si l'apprentissage est réalisé directement sur l'espèce dont le génome est à annoter et si un grand nombre de gènes sont déjà connus expérimentalement. Parmi les défauts majeurs de ce type d'approche on peut citer la faible spécificité, qui se caractérise par une sur-prédiction de gènes (faux-positifs), la difficulté de prédiction des « petits » gènes [Skovgaard et al., 2001] ou encore les problèmes de prédiction des sites d'initiation [Overbeek et al., 2007].

Figure 3 : Prédiction des gènes. A partir d'un large ensemble d'ORFs candidates, un petit sous-ensemble de gènes est prédit, d'après [Overbeek et al., 2007]. L'ensemble des ORFs sont extraites (partie supérieure de la figure) puis la sélection des ORFs qui codent réellement pour une protéine est réalisée par les méthodes intrinsèques ou extrinsèques (partie inférieure de la figure).



2.2. Méthodes extrinsèques :

Ces méthodes reposent sur la comparaison des ORFs avec les protéines présentes dans les banques. Les ORFs présentant de fortes similarités avec des protéines connues sont considérées comme codantes. Les logiciels utilisant ces approches présentent tout de même toujours une composante de prédiction *ab initio*. Ainsi, les logiciels ORPHEUS [Frishman et al., 1998], CRITICA [Badger et al., 1999] ou Reganor [Linke et al., 2006] utilisent les recherches de similarités pour constituer leur ensemble de gènes de référence qui permettra de construire les modèles d'évaluation de l'ensemble des ORFs du génome.

A l'inverse des méthodes intrinsèques, ce type d'approche pour la prédiction de gènes présente une bonne spécificité (peu de faux-positifs) mais a aussi tendance à induire une sous-prédiction (faible

sensibilité) particulièrement pour la prédiction de « nouvelles » protéines qui n'ont pas d'homologue suffisamment proche dans les protéines présentes dans les banques.

Particularités des génomes eucaryotes :

Globalement le principe des méthodes appliquées à la prédiction de gènes dans les génomes eucaryotes est le même mais quelques adaptations doivent être réalisées pour tenir compte des particularités de ce type de génome qui viennent un peu compliquer l'annotation. Ainsi, pour ce type de génome, dans les méthodes de prédiction *ab initio*, il sera tenu compte de la présence de signaux de transcription (TATA-box, signal de polyadénylation, site de fixation de facteur de transcription, etc.) et des signaux de « splicing » (séquences consensus de sites d'épissage). Les séquences transcrites (EST) de l'organisme pourront aussi être utilisées lors du processus d'annotation. Pour l'annotation des génomes eucaryotes, il est nécessaire que les programmes prédisent les zones correspondant aux introns et aux exons [Burge et al., 1997; Burge et al., 1998; Parra et al., 2000].

3. Les erreurs d'annotation, leurs conséquences et les modes de correction

L'ensemble des protéines prédites lors de l'annotation d'un génome est ensuite intégré dans les banques de séquences protéiques. Les protéines issues de l'annotation d'un génome complet sont la principale source, voire quasiment la seule source, d'alimentation des banques protéiques à l'heure actuelle. Or, ces protéines ne sont que le fruit d'une prédiction automatique *in silico* qui n'est pas exempte d'erreurs. Même si le séquençage du génome lui-même est une source non-négligeable d'erreurs [Weinstock, 2000], la plupart de celles-ci proviennent de l'annotation du génome [Galperin et al., 1998]. Ces erreurs ont des conséquences néfastes pour la biologie. Toutefois, des méthodes existent pour détecter, limiter et corriger ces erreurs.

3.1. Les erreurs

3.1.1. Séquençage du génome

La fréquence des erreurs dans une séquence génomique dont le séquençage est fini est estimée à 1 erreur (délétion, substitution, addition d'une base) sur 10^3 à 10^5 bases [Weinstock, 2000]. Ces erreurs peuvent entraîner la substitution d'un acide aminé par un autre dans la séquence protéique ou un changement de cadre de lecture artificiel pour la séquence protéique («artificial frameshift») ou encore l'introduction d'un codon stop [Perrodou et al., 2006]. En plus, il n'est jamais impossible d'exclure totalement la possibilité d'une région manquante plus conséquente dans le génome.

3.1.2. Annotation du génome

Le premier type d'erreur concerne l'existence biologique même de la protéine prédite. En absence de la caractérisation expérimentale de la protéine, aucune preuve ne permet de valider son existence et un certain nombre de protéines présentes dans les banques ne sont effectivement pas exprimées, on parle de problème de sur-prédiction. Les protéines susceptibles de correspondre à ce cas de figure portent généralement dans les banques la dénomination « putative », « hypothetical », ou « uncharacterized protein ». A l'inverse, certaines protéines réellement exprimées dans l'organisme ne sont pas prédites lors du processus d'annotation, on parle de problème de sous-prédiction [Overbeek et al., 2007].

Le deuxième type d'erreur porte sur l'exactitude de la détermination des bornes de la protéine, ses extrémités N-terminale ou C-terminale. Lorsque l'on utilise les outils de la génomique comparative pour comparer plusieurs génomes, il devient évident que pour des génomes proches, l'assignation des codons d'initiation de certains gènes a été réalisée différemment par les équipes d'annotateurs ou les logiciels d'annotation (Figure 4). Si le problème est beaucoup moins fréquent pour les codons stop, il faut toutefois garder à l'esprit que le 21^{ème} acide aminé, la selenocysteine, est codé dans le génome par un codon TGA qui code habituellement pour un codon stop. Comme les programmes de prédiction attribuent toujours ce triplet à un codon stop, une protéine qui contient réellement la selenocysteine sera invariablement mal prédite voire pas prédite du tout (ORFs trop petites) (Figure 5).

Un troisième type d'erreur rencontré concerne les décalages de cadre de lecture réels (vrais « frameshifts ») dans le génome [Groisman et al., 1995; Gurvich et al., 2003]. Lors de l'annotation d'un génome présentant ce type de phénomène, la protéine réelle ne sera pas prédite sous sa vraie forme mais sous une forme tronquée ou sous la forme de deux protéines distinctes.

Enfin, un quatrième type d'erreur concerne les protéines isoformes qui présentent des épissages alternatifs. Ce type de problème concerne quasi exclusivement les eucaryotes. On estime que plus de 70 % des gènes humains subissent de l'épissage alternatif [Kriventseva et al., 2003; Stamm et al., 2005]. Les méthodes actuelles de prédiction *ab initio* et par recherche de similarité ont des difficultés à prédire précisément les structures de gènes épissés alternativement. En conséquence, certaines protéines isoformes étant réellement exprimées ne sont pas présentes dans les banques.

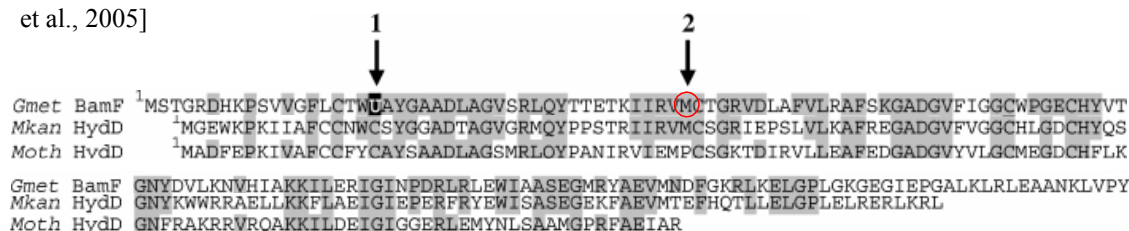
Figure 4 : Illustration des difficultés d'assignation des codons d'initiation. Alignement de séquences de l'« Hydrogenase expression/formation protein » chez des organismes proches, d'après cours Odile Lecompte, (<http://lbg.igbmc.fr/~lecompte/>)

```

058470-1 -----MKIKLEHGAGGELMEEIIRD 20
026307-1 MFIHTYPGFSDTDKIQTEVRERFYTISKIPYRQEIMQGESMKIGMSHGAGGEVMDLISD 60
Q58089-1 -----MKITRMHGAGGKVMQELIKD 20
067132-1 -----MKKILLSHGGGGEETQKLIK 21
                    ** **.* : .** :
058470-1 VILKNLTLN-SAGGIGLEALDDGATIPFEGKHLVFTIDGHTVRPLFFPGGDIGRLAVSGT 79
026307-1 IILSNIHNTRVNGGVGLEDLDDGASIPLDGYEIVISTDGHTIDPLFFPGGDIGRIAVAGT 120
Q58089-1 VILKNLEITSVNGGIGLESLLDSATIPIGDKIEIVFTVDGHTVKPIFFPGGDIGRLAVSGT 80
067132-1 LFLKYFSNEILN-----KLEDASLLNLSG-KVAFTTDAFTVSPIFFRGGDIGKLAVAGT 74
:.* . : *.* : : . . . . : *.* : *.* *****:.*.*

```


Figure 5 : Illustration des difficultés de prédiction des selenocystéines. Alignement de séquences de plusieurs F₄₂₀ non reducing hydrogenase d-subunit. La flèche 2 pointe le codon d'initiation prédit pour la séquence BamF. La flèche 1 pointe une selenocysteine (U) codée par un codon de terminaison TGA. D'après [Wischgoll et al., 2005]



3.2. Les conséquences des erreurs

Les erreurs de séquençage et d'annotations des génomes sont directement répercutées dans les banques protéiques dans lesquelles on trouve des séquences de protéines sans existence biologique, des protéines dont la séquence n'est pas exacte et dans lesquelles des protéines réellement exprimées sont absentes. Ces erreurs dans les banques protéiques ont des conséquences multiples pour la biologie. Par exemple, l'absence dans les banques d'une protéine dont le rôle est très important pour un processus biologique va pénaliser très fortement la compréhension de ce processus. Une protéine présente dans les banques mais sans existence biologique peut masquer un autre gène dans un cadre de lecture différent. Une protéine prédite avec des bornes inexactes, par exemple un codon d'initiation trop en amont ou en aval dans la séquence peut gêner la détection de motifs de régulation ou de promotion [Salgado et al., 2000; Edwards et al., 2005], conduire à des erreurs dans l'annotation fonctionnelle de la protéine par bioinformatique, ou à de grosses difficultés pour les expériences d'expression de protéines (protéines surexprimées qui n'ont pas d'activité, protéines non-cristallisables) [Trivedi et al., 2004; Horie et al., 2007]. Les erreurs introduites dans les banques protéiques pénalisent également toutes les expériences biologiques qui passent par l'analyse de peptides (ou des protéines) par spectrométrie de masse. Comme les approches extrinsèques de prédiction utilisent les informations des protéines présentes dans les banques, les erreurs peuvent donc être amplifiées, c'est ce qu'on appellera un effet domino négatif. La détection, la limitation et la correction des erreurs sont donc indispensables pour ne pas introduire de « bruit » dans les banques protéiques.

3.3. Les méthodes de correction

Les corrections des banques protéiques peuvent être réalisées à 2 niveaux. D'une part, des développements bioinformatiques sont réalisés pour combattre ces problèmes mais les corrections apportées restent des prédictions et ne sont en aucun cas issues d'analyses directes des protéines. Ces méthodes consistent plutôt à limiter les erreurs au maximum lors de l'annotation. Le deuxième mode

de correction repose sur l'analyse directe de la protéine principalement grâce aux outils de la protéomique.

3.3.1. Corrections *in silico*

La limitation des erreurs dans la séquence du génome passe par un séquençage de profondeur maximale : plus les fragments de génomes séquencés auront de zones de chevauchement, moins on constatera d'erreurs dans la séquence finale du génome. Des programmes de détection des erreurs de séquençage ont été développés et permettent de cibler spécifiquement des zones du génome à re-séquencer [Brown et al., 1998; Medigue et al., 1999; Perrodou et al., 2006].

Le problème des erreurs d'annotation est plus critique. Les phénomènes de sous- et sur-prédiction de gènes sont inhérents à la méthode d'annotation utilisée. Toutefois, en général, les systèmes d'annotation de génome (« pipelines ») associent les deux types de méthode dans une approche hybride qui combine les avantages des deux méthodes [Peterson et al., 2001; Meyer et al., 2003; Van Domselaar et al., 2005] pour limiter ces phénomènes. L'assignation correcte des codons d'initiations des gènes a fait l'objet de développements logiciels basés essentiellement sur la recherche du site de fixation du ribosome [Suzek et al., 2001; Tech et al., 2005], généralement caractérisé par la présence d'une séquence consensus complémentaire de l'extrémité 3' de l'ARNr 16S, la séquence Shine-Dalgarno [Shine et al., 1974]. Cette séquence est toutefois variable d'un organisme à l'autre et certains gènes sont dépourvus de séquence Shine-Dalgarno [Kozak, 1999] ce qui complique ces méthodologies de correction. Les approches de génomique comparative peuvent également participer à la correction des différentes erreurs d'annotation [Overbeek et al., 2007] en comparant les annotations des génomes de multiples organismes. Toutefois, ces approches doivent être appliquées précautionneusement pour ne pas conduire à la propagation des erreurs à la place des corrections. Enfin, une étape d'annotation experte manuelle réalisée suite aux prédictions automatiques pourrait permettre de corriger un certain nombre d'erreurs de prédiction. Il existe des initiatives de ce type, par exemple pour la banque protéique à haute valeur ajoutée Swiss-Prot mais face à l'ensemble du flot de données prédites quotidiennement, cette tâche paraît insurmontable [Gattiker et al., 2003].

3.3.2. Correction expérimentale

Les méthodes de correction *in silico* permettent de détecter et limiter les erreurs d'annotation mais restent basées sur des principes de prédiction. Il apparaît comme indispensable de vérifier exactement la prédiction des gènes.

Il existe des méthodes de vérification de l'existence et de la structure génomique des gènes prédits par utilisation de la réaction en chaîne par polymérase après transcription inverse (« Reverse Transcriptase- Polymerase Chain Reaction », RT-PCR) et le séquençage direct des produits de la RT-PCR [Wu et al., 2004]. Ces techniques de validation basées sur l'expression d'ARNm ne permettent

tout de même pas de fournir l'évidence et la structure exacte de la traduction de cet ARNm en protéine et peuvent être biaisées par l'annotation initiale du génome.

La meilleure option pour identifier indépendamment et de manière non-ambigue au moins une partie des gènes prédits dans un génome consiste à réaliser une analyse systématique des protéines exprimées naturellement par l'organisme, le protéome. Ces résultats expérimentaux permettent de valider ou de corriger les prédictions et donc de participer à l'annotation du génome. Aujourd'hui, l'analyse protéomique par spectrométrie de masse couplée aux méthodes de séparation des protéines et/ou des peptides est la méthode de choix pour la caractérisation des protéines et pourra donc être utilisée dans le but de participer à l'annotation des génomes. Cette nouvelle approche d'aide à l'annotation génomique représente un champs d'application très récent dans le domaine protéomique et porte le nom de « protéogénomique » [Jaffe et al., 2004].

Chapitre 3 : Les outils de l'analyse protéomique

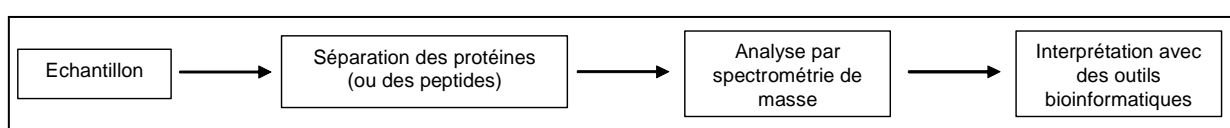
En biologie, la première information requise pour l'étude des protéines est la connaissance de leur séquence primaire (enchaînement d'acides aminés). Des années 50 jusque dans les années 1980 la méthode de choix pour accéder à cette information était le séquençage par dégradation d'Edman [Edman, 1949]. Cette méthode, nécessitant une bonne expertise, repose sur l'identification des acides aminés qui ont été coupés chimiquement et de manière séquentielle à partir de l'acide aminé N-terminal de la protéine. Pour mener à bien ce séquençage, qui est un processus assez lent, il était nécessaire de disposer de la protéine purifiée en quantité non négligeable (de l'ordre de la nanomole puis de la picomole avec les instruments de dernière génération).

Dans les années 90, la spectrométrie de masse a progressivement supplanté la dégradation d'Edman. Jusque-là, la spectrométrie de masse n'était utilisée que pour l'analyse de petites molécules thermostables parce qu'il n'existait pas de techniques « douces » pour ioniser et transférer les biomolécules ionisées en phase gazeuse sans induire trop de fragmentation. Toutefois, l'apparition des techniques d'ionisations douces, l'ionisation Electrospray (ESI) [Fenn et al., 1989] et le MALDI (Matrix Assisted Laser Desorption Ionisation) [Karas et al., 1988; Tanaka et al., 1988], a considérablement changé la donne en rendant les peptides et les protéines accessibles à l'analyse par spectrométrie de masse. Aujourd'hui, la spectrométrie de masse qui présente une très bonne sensibilité et qui permet de fragmenter les peptides en quelques secondes au lieu de quelques heures voire quelques jours pour la dégradation d'Edman [Wilm et al., 1996], est sans conteste la méthode de choix pour la caractérisation des biomolécules [Mann et al., 2001; Domon et al., 2006; Cravatt et al., 2007].

En parallèle, les développements des techniques séparatives des protéines et des peptides (gels d'électrophorèse, chromatographie liquide) en amont de la spectrométrie de masse ainsi que le développement des banques protéiques (vu dans le chapitre I) ont contribué à l'essor de l'analyse protéomique par spectrométrie de masse.

L'analyse protéomique est généralement composée de trois étapes (Figure 1). Après l'étape de préparation de l'échantillon protéique (ou peptidique si la digestion a déjà été réalisée), celui-ci est séparé (ou fractionné) en utilisant des méthodes séparatives (gels d'électrophorèse 1D/2D, chromatographie liquide). Les mélanges de peptides ainsi obtenus sont analysés par spectrométrie de masse. Finalement, l'identification des protéines par interprétation des résultats des analyses de spectrométrie de masse est réalisée grâce à des outils bioinformatiques.

Figure 1 : Les grandes étapes de l'analyse protéomique



Ce chapitre a pour objet de présenter les différents outils de la spectrométrie de masse et les techniques de séparation des protéines et/ou des peptides.

1. Les outils de la spectrométrie de masse

En général, un spectromètre de masse est composé de trois parties : une source d'ionisation qui volatilise et ionise les molécules, un analyseur qui sépare les ions et mesure leur rapport masse sur charge (m/z) et un détecteur qui mesure le courant induit par les ions.

Les sources d'ionisation généralement utilisées en protéomique sont la source MALDI (Matrix Assisted Laser Desorption Ionisation) [Karas et al., 1988; Tanaka et al., 1988] et l'ionisation Electrospray (ESI) [Fenn et al., 1989].

Une fois les ions produits, ils sont transférés jusqu'à l'analyseur à l'aide de gradients de pression et de champs électriques. Dans l'analyseur, les ions sont séparés en fonction de leur rapport m/z . Plusieurs analyseurs peuvent être combinés, on parle alors de spectrométrie de masse en tandem. Ce type de configuration permet d'obtenir des informations structurales par exemple par fragmentation induite par collision (CID). Dans ce cas, dans un deuxième temps, les fragments générés sont aussi séparés en fonction de leur rapport m/z .

Finalement, les ions atteignent le détecteur où un signal est généré puis traduit en spectre de masse dans lequel l'intensité du signal est mise en relation avec le rapport m/z .

1.1. Les sources d'ionisation

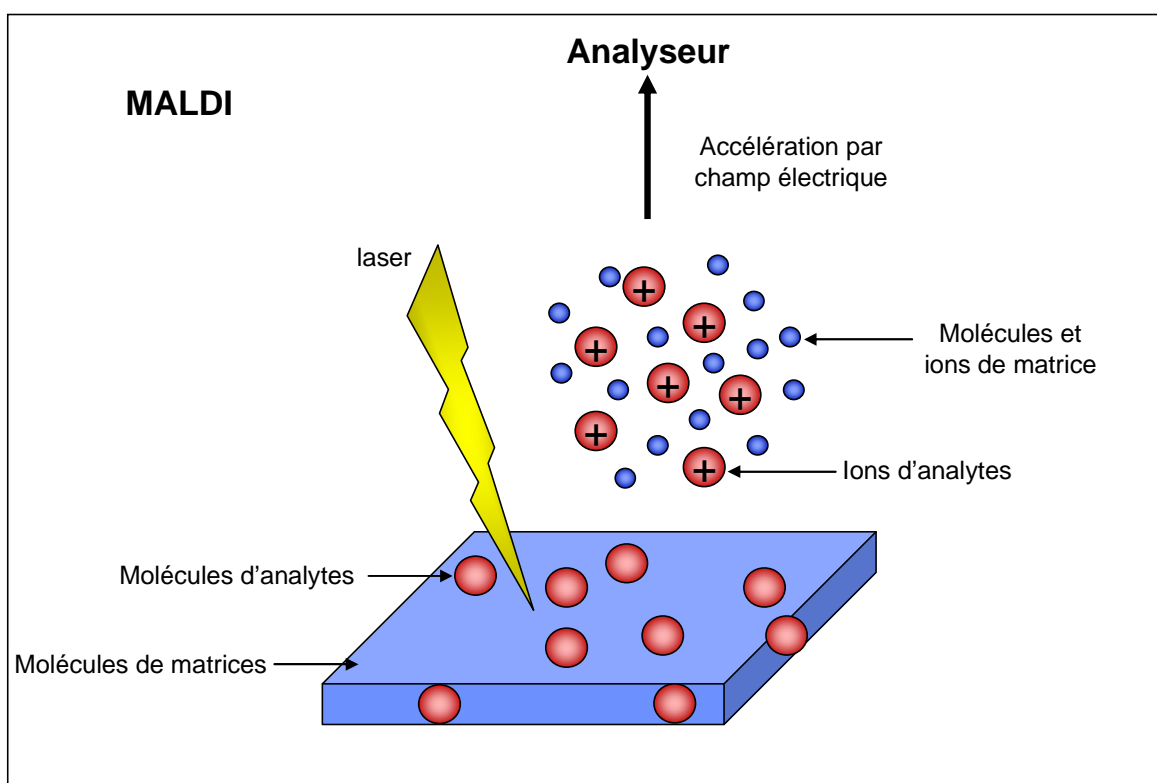
L'ionisation MALDI (Matrix Assisted Laser Desorption Ionisation) [Karas et al., 1988; Tanaka et al., 1988] et électrospray (ESI) [Fenn et al., 1989] sont les deux techniques les plus couramment utilisées pour volatiliser et ioniser les peptides et les protéines lors des analyses par spectrométrie de masse. Ces deux techniques sont qualifiées de techniques d'ionisation douce puisqu'elles permettent le transfert de biomolécules intactes en phase gazeuse pour l'analyse par spectrométrie de masse. John Fenn et Koichi Tanaka ont reçu le prix Nobel de chimie en 2002 pour leurs travaux sur ces nouvelles techniques d'ionisation. Le processus MALDI sera présenté ici très brièvement car l'essentiel des travaux de cette thèse a été réalisé avec des instruments de type ESI.

1.1.1. La source MALDI (Matrix assisted laser desorption ionization)

La technique d'ionisation MALDI consiste à irradier sous vide ($\sim 10^{-7}$ mbar) avec un faisceau laser pulsé de longueur d'onde donnée (dans la gamme UV) un dépôt cristallin contenant un mélange de matrice organique (généralement un composé de faible masse moléculaire, aromatique et acide) et d'échantillon à analyser. Le faisceau laser excite la matrice qui se dissocie, se sublime et entraîne

l'analyte dans ce plasma d'expansion. C'est au niveau de ce plasma que se produira l'ionisation des molécules. Les ions d'analytes ainsi obtenus sont accélérés par une différence de potentiel jusqu'à l'analyseur choisi [Steen et al., 2004]. Généralement, la technique MALDI produit des ions simplement chargés et donc des spectres de masse relativement simples. Cette technique a l'avantage d'être relativement tolérante aux sels, contrairement à l'ESI.

Figure 2 : L'ionisation MALDI



Un analyseur temps de vol (« Time Of Flight », TOF) est souvent combiné avec une source MALDI parce qu'il est extrêmement bien adapté pour analyser les « paquets d'ions » générés par l'impulsion du faisceau laser. Les ions générés sont accélérés avec la même énergie cinétique dans le tube de vol libre de champ. Le temps nécessaire à chaque ion pour atteindre le détecteur est mesuré. Comme l'énergie cinétique (zeE) est égale à $\frac{1}{2}mv^2$, la vitesse des ions est inversement proportionnelle à la racine carrée de leur masse et proportionnelle à la racine carrée de leur charge. En conséquence, des ions de rapport m/z différent auront des temps de vol différents. Un spectre de masse sera généré après avoir mesuré les temps de vol de chaque ion. Généralement, un réflectron est utilisé pour améliorer la résolution de cet analyseur. Ce réflectron refocalise les ions de même rapport m/z en augmentant la longueur de parcours des ions les plus énergétiques et en diminuant celui des ions les moins énergétiques. Dans ce cas, le détecteur est placé au point focal du réflectron.

1.1.2. La source electrospray (ESI)

Le mode d'ionisation électrospray (ESI) permet de générer des ions en phase gazeuse à pression atmosphérique et sous l'effet d'un fort champ électrique à partir d'analytes en solution. Cette technique a donc pu être naturellement couplée à la chromatographie liquide (« Liquid Chromatography », LC) [Whitehouse et al., 1985]. Le processus d'ionisation-désorption ESI peut être décomposé en trois étapes : i) production de gouttelettes chargées, ii) fission des gouttelettes chargées en gouttelettes filles et iii) émission des ions désolvatés en phase gazeuse (Figure 3).

1.1.2.1. Processus d'ionisation-désorption ESI

1.1.2.1.1. Production des gouttelettes chargées

Le processus électrospray est obtenu en appliquant un champ électrique intense (10^6 V/m) sur le capillaire contenant la solution d'analytes. Ce champ électrique provoque la polarisation du liquide et la séparation des charges positives et négatives. En mode d'ionisation positif (le plus fréquemment utilisé pour les études protéomiques), le capillaire métallique chargé positivement va jouer le rôle d'anode en attirant et neutralisant les charges négatives. En conséquence, l'excès de charges positives va se concentrer à la pointe du capillaire entraînant un « allongement » du liquide qui va former ce qu'on appelle le « cône de Taylor ». Ce cône va s'étirer jusqu'à éclater en minces gouttelettes enrichies en cations (Figure 3A). C'est pour cette raison que l'électrospray peut être comparé à une cellule électrochimique [Blades et al., 1991]. En pratique, l'électrospray est obtenu par l'infusion d'électrolytes à des débits compris entre 1 et 100 $\mu\text{L}/\text{min}$ qui nécessitent l'assistance d'un flux d'azote coaxial (appelé gaz de nébulisation) pour stabiliser la formation des gouttelettes.

1.1.2.1.2. Fission des gouttelettes chargées, l'explosion coulombienne

Au fur et à mesure que le solvant contenu dans les gouttelettes s'évapore, la taille de celles-ci diminue et la densité de charge augmente au sein de la gouttelette. Lorsque le rayon de la gouttelette devient inférieur au rayon critique de Rayleigh [Rayleigh, 1882] (Equation 1), limite à laquelle les forces de répulsion électrostatique sont égales aux forces de tension de surface, la gouttelette devient instable et subit une explosion coulombienne qui génère des gouttelettes filles [Kearle, 2000]. Ces gouttelettes filles subiront à leur tour le même processus, et ceci sur plusieurs générations jusqu'à donner des ions complètement désolvatés.

Equation 1 : Equation régissant le rayon de Rayleigh

$$Q = 8 \pi \sqrt{\epsilon_0 \gamma R_R^3}$$

avec

Q : charge de la gouttelette

R_R : rayon critique de Rayleigh

ϵ_0 : permittivité du vide

γ : tension de surface du liquide

1.1.2.1.3. Emission des ions désolvatés en phase gazeuse

Il existe principalement deux théories, aujourd'hui encore assez controversées, qui décrivent la production des ions désolvatés en phase gazeuse : le modèle de Dole [Dole et al., 1968] et le modèle d'Iribarne et Thomson [Thomson et al., 1976].

Modèle de Dole : le modèle de la charge résiduelle

Ce modèle propose que l'évaporation du solvant et le processus d'explosions coulombiennes se répète jusqu'à ce qu'il n'y ait plus qu'une charge par gouttelette (Figure 3B1).

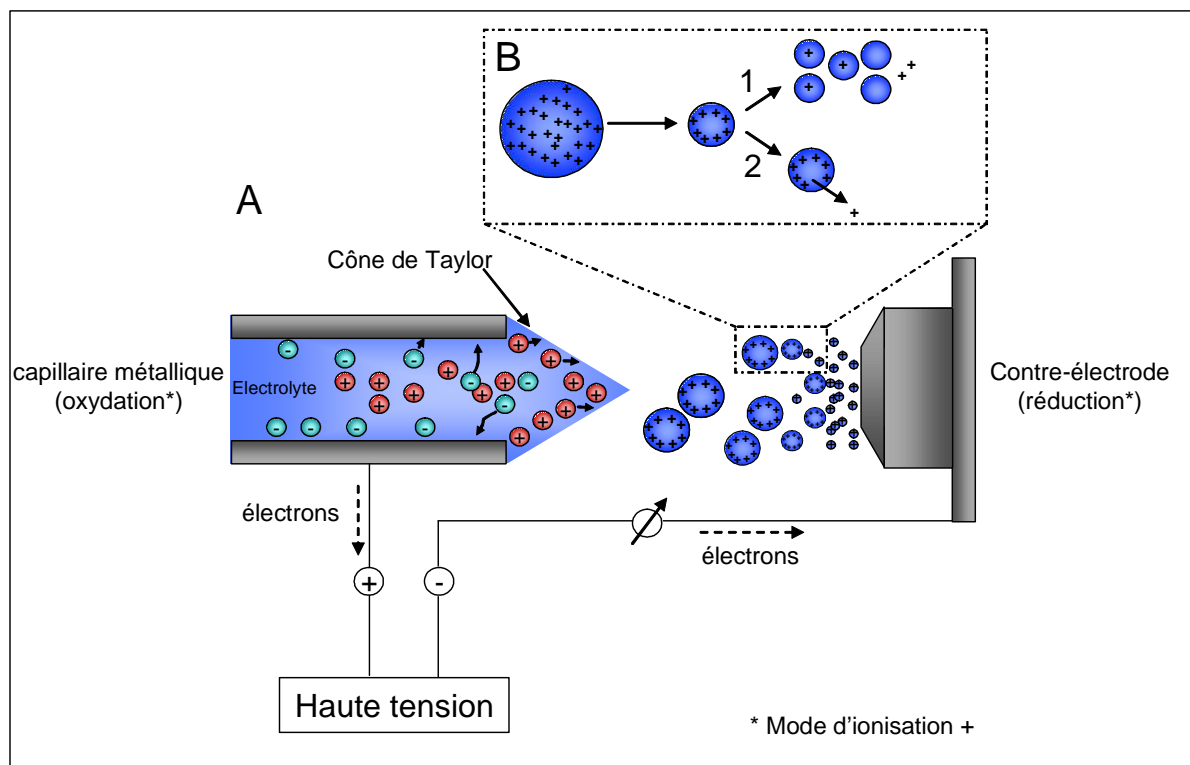
Modèle d'Iribarne et Thomson : le modèle de l'évaporation ionique

Ce modèle propose que lors du processus d'évaporation du solvant, la densité de charge devient suffisante pour émettre directement les ions en phase gazeuse (Figure 3B2).

Le processus de transfert des ions en phase gazeuse est toujours discuté, et bien que le modèle de l'évaporation ionique semble être le plus reconnu pour les petites molécules, il semblerait que le mécanisme décrit par le modèle de la charge résiduelle intervienne pour les macromolécules, type protéine [Kearle, 2000].

Figure 3 : L'ionisation électrospray. A) Production et fission des gouttes chargées. B) Transfert des ions en phase gazeuse selon le modèle de Dole (1) et le modèle d'Iribarne et Thomson (2).

D'après [Kearle, 2000] et [Steen et al., 2004].



1.1.2.2. La source nano-électrospray

En mode d'ionisation électrospray, le spectromètre de masse mesure un courant d'ion qui est fonction de la concentration des ions présents dans l'échantillon et pas du débit avec lequel l'échantillon est infusé [Ikonomou et al., 1990]. Pour les faibles quantités d'échantillon, ce qui est souvent le cas pour les échantillons biologiques, afin de gagner en sensibilité, il est préférable de travailler à des débits plus faibles et avec des solutions plus concentrées.

Ces observations ont conduit au développement des sources micro-électrospray (débits 300-800 nL/min) [Emmet et al., 1994] et nano-électrospray (débit ~20 nL/min) [Wilm et al., 1996; Wilm et al., 1996] qui utilisent des capillaires de diamètre interne de quelques μm (~100 μm pour le mode électrospray) et qui ne nécessitent plus d'assistance gazeuse. L'utilisation des débits plus faibles conduit à l'émission de gouttelettes plus petites (environ 10 fois) possédant une plus grande densité de charge et donc à une augmentation du rendement d'ionisation-désorption des analytes [Fernandez de la mora et al., 1994; Wilm et al., 1996; Karas et al., 2000]. De plus, ces sources miniaturisées vont permettre de réduire la consommation d'échantillon.

Pour tous ces avantages, ces sources d'ionisations couplées aux techniques chromatographiques compatibles (micro- ou nano-chromatographie) sont particulièrement bien adaptées aux analyses protéomiques.

1.1.2.3. Ionisation des analytes

Les 3 principaux mécanismes impliqués dans le processus électrospray sont les suivants :

1.1.2.3.1. Ionisation par séparation des charges

La séparation des charges en solution est la méthode de base par laquelle l'ionisation est réalisée pour les espèces inorganiques, pour les molécules biologiques et organiques avec des groupements fonctionnels basiques ou acides et pour les espèces contenant un groupe phosphonium, ammonium ou oxonium. En fonction de ces caractéristiques, les analytes sont analysés en mode positif ou négatif. Les protéines et les peptides qui contiennent des acides aminés basiques s'ionisent facilement par protonation pour être analysés en mode positif.

1.1.2.3.2. Ionisation par formation d'adduits.

Les analytes polaires qui n'ont pas de groupements basiques ou acides peuvent être ionisés en ESI par la formation d'adduits avec d'autres ions présents en solution. La formation d'adduits se produit en solution avant que le processus de séparation de charge soit réalisé.

1.1.2.3.3. Ionisation par réactions en phase gazeuse

L'ionisation des analytes peut également être influencée par des interactions en phase gazeuse d'une part parce que ce processus est réalisé à pression atmosphérique et d'autre part parce que de grandes quantités de molécules de solvant chargées sont générées en plus des analytes. Les interactions en phase gazeuse se produisent une fois que les analytes ne sont plus en solution donc durant la dernière phase du processus électrospray. Cette ionisation en phase gazeuse est généralement réalisée par des réactions de transfert de protons en phase gazeuse [Kearle et al., 1999]. Les molécules protonées en solution peuvent ainsi, en phase gazeuse, céder leurs protons à d'autres molécules (solvant ou analytes) qui ont une basicité en phase gazeuse supérieure. Par ce processus, des analytes neutres dans les gouttelettes du spray peuvent se charger au moment de leur transfert en phase gazeuse. En ESI, il est donc important d'utiliser un solvant avec une affinité protonique en phase gazeuse plus faible que celui des analytes d'intérêt. Cela n'est généralement pas un problème pour l'analyse des peptides et des protéines puisqu'ils ont généralement une basicité en phase gazeuse très élevée.

La possibilité d'ioniser les analytes par un des mécanismes décrits ci-dessus est une condition préalable à leur analyse en mode électrospray. Toutefois, même s'ils sont ionisables par un de ces mécanismes, on observe une grande variabilité d'efficacité d'ionisation entre les différents analytes

même si ceux-ci sont à la même concentration dans la solution analysée [Cech et al., 2001; Zhou et al., 2001]. Les caractéristiques chimiques de ces analytes ont donc une grande influence sur l'efficacité d'ionisation électrospray.

1.1.2.4. Efficacité d'ionisation des analytes

1.1.2.4.1. Affinité pour la surface des gouttes

Les analytes qui ont une grande affinité pour la surface des gouttes du spray ont généralement les meilleurs efficacités d'ionisation [Cech et al., 2000; Cech et al., 2001; Zhou et al., 2001]. Cette observation avait été initialement réalisée par les pionniers de la théorie de l'évaporation ionique [Iribarne et al., 1983] qui avaient observé une meilleure réponse électrospray pour les analytes non-polaires. Les auteurs ont suggéré que celle-ci était due au fait que les ions non-polaires préféreraient l'interface air-liquide et donc que ces ions se localiseraient préférentiellement à la surface de la goutte (Figure 4). Les raisons qui expliquent ce phénomène sont multiples :

- Les ions à la surface de la goutte se désolvateraient plus facilement en phase gazeuse que ceux de l'intérieur de la goutte et auraient donc une meilleure réponse en ESI [Tang et al., 1991].

- Le processus électrospray produit une quantité fixe d'excès de charges qui résident à la surface des gouttes du spray. L'intérieur de la goutte est lui constitué d'un mélange de cations et d'anions qui compensent leurs charges mutuellement pour obtenir un milieu électriquement neutre. Les analytes neutres et les analytes chargés formant des paires d'ions au centre de la goutte formeront des sels neutres durant leur désorption et ne seront pas détectables par le spectromètre de masse contrairement aux ions qui sont à la surface de la goutte [Enke, 1997; Sjoberg et al., 2001].

- Lors de l'explosion coulombienne des gouttes du spray, les ions à la surface ont davantage tendance à se retrouver dans les gouttelettes filles que ceux à l'intérieur de la goutte [Tang et al., 2001]. Ainsi, en mode nanospray, les gouttelettes étant de plus petites tailles, on observe moins la perte des ions qui sont localisés à l'intérieur de la goutte [Karas et al., 2000].

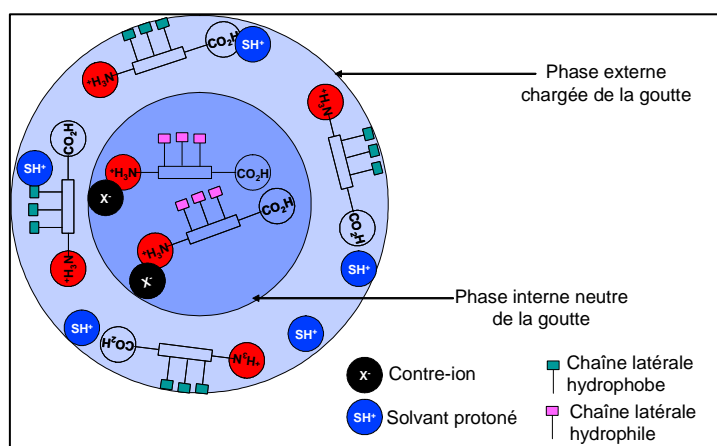


Figure 4 : Illustration de la localisation de peptides dans une goutte de spray. Les peptides polaires ont tendance à être localisés à l'intérieur de la goutte où ils sont neutralisés par des contre-ions. Les peptides non-polaires sont préférentiellement localisés à la surface de la goutte.

1.1.2.4.2. Prédiction de la réponse ESI des analytes

Le fait que l'affinité des analytes pour la surface des gouttes du spray influence la réponse ESI signifie que différents paramètres liés à la structure de l'analyte permettent de prédire son efficacité d'ionisation. Il a ainsi pu être montré que les peptides constitués d'acides aminés avec des chaînes latérales hydrophobes présentaient une meilleure réponse [Cech et al., 2000; Zhou et al., 2001]. Différents paramètres liés à la polarité des analytes peuvent être corrélés avec l'efficacité d'ionisation électrospray. Ainsi, on observe une relation relativement linéaire entre la réponse ESI et :

- L'aire de la surface non-polaire des peptides [Karplus, 1996; Cech et al., 2000].
- L'énergie libre de Gibbs de transfert des peptides [Cech et al., 2000].
- La composition de la phase mobile de la chromatographie liquide en phase inverse lors de l'éluion des peptides de la colonne [Cech et al., 2001].
- Le nombre de résidus non-polaires/polaires des peptides [Cech et al., 2001].

1.2. Les analyseurs

La qualité des résultats obtenus lors d'une analyse protéomique est principalement fonction de l'analyseur de masse utilisé. Les paramètres qui servent à juger des performances de l'analyseur sont :

- sa résolution
- sa justesse de mesure de masse
- sa gamme de masse
- sa capacité à réaliser de la spectrométrie de masse en tandem
- sa sensibilité

Dans ce paragraphe, je ne décrirai que les analyseurs combinés au mode d'ionisation électrospray qui ont servi à réaliser l'essentiel des travaux présentés dans cette thèse à savoir les trappes ioniques HCT Ultra (Bruker Daltonics) et le système quadripôle-temps de vol Synapt (Waters).

1.2.1. L'analyseur quadripolaire et le système quadripôle-temps de vol

1.2.1.1. L'analyseur quadripolaire

Le quadripôle est un filtre de masse constitué de 2 paires de barres métalliques parallèles connectées entre elles de manière à pouvoir appliquer un champ électrique en opposition de phase entre 2 barreaux adjacents. Ce champ électrique ne laisse passer que les ions qui ont un rapport m/z donné. Les ions qui ont un rapport m/z différent ont des trajectoires instables et ne continuent pas leur

parcours en sortie de l'analyseur. En balayant en amplitude le champ électrique, les ions avec différents rapports m/z pourront traverser l'analyseur. Le principe de fonctionnement détaillé de l'analyseur quadripolaire a été décrit en détail par Campana [Campana, 1980]. Un quadripôle peut être combiné à un autre quadripôle ou à un analyseur à temps de vol pour réaliser des analyses de spectrométrie de masse en tandem.

1.2.1.2. Le système quadripôle-temps de vol : des analyseurs en tandem

Le système quadripôle-temps de vol (« Quadrupole-Time Of Flight », Q-TOF) est constitué de l'assemblage de deux analyseurs, un premier analyseur quadripolaire suivi d'un tube de vol (décrit brièvement en 1.1.1), les deux étant séparés par une cellule de collision. Dans cette configuration, le quadripôle peut fonctionner comme un analyseur ou juste servir de guide d'ions.

Lorsque le quadripôle fonctionne comme guide d'ions (« RF-only »), les ions sont analysés par l'analyseur TOF.

Lorsque le quadripôle fonctionne comme analyseur, un ion peut être sélectionné dans le quadripôle, transmis et fragmenté dans la cellule de collision pour générer des fragments qui sont ensuite analysés dans l'analyseur TOF. C'est dans ce mode de fonctionnement qu'on réalise la spectrométrie de masse en tandem qui permet de faire de la MS/MS sur les ions sélectionnés. Dans cette configuration, le Q-TOF combiné à la chromatographie liquide est un outil de choix pour réaliser les analyses protéomiques. Durant ces analyses, tout au long de l'élution chromatographique des différents peptides, chaque cycle MS permet de sélectionner des peptides qui seront séquencés par analyse MS/MS. C'est dans cette configuration qu'a été utilisé le spectromètre Synapt de Waters.

Les spectres obtenus avec cet instrument présentent une haute résolution et une bonne justesse de mesure de masse (particulièrement avec le système de correction « lock mass ») aussi bien en MS qu'en MS/MS. Sur ces spectres, la détermination des états de charge et l'attribution des pics mono-isotopiques se trouvent facilitées. Ces avantages combinés à la justesse de mesure de masse facilitent grandement l'identification des peptides fragmentés par recherche dans les banques de données protéiques.

1.2.2. L'analyseur trappe ionique (« Ion Trap », IT)

La trappe ionique est constituée de 3 électrodes, une électrode annulaire centrale et 2 électrodes chapeaux quasi-hyperboliques [March, 2009]. La trappe présente un volume d'environ 1 cm³ dans lequel un gaz tampon, l'hélium est maintenu à une pression d'environ 5.10⁻³ mbar. Nous avons vu précédemment que pour le système Q-TOF, l'isolation des précurseurs, leur fragmentation et l'analyse des fragments étaient réalisés séquentiellement dans l'espace (à des endroits distincts). Pour la trappe ionique, les différentes étapes de l'analyse sont réalisées séquentiellement dans le temps mais au même endroit.

1.2.2.1. Le piégeage des ions

Le piégeage des ions est assuré par la présence du gaz tampon et par l'application d'une radiofréquence sur l'électrode annulaire.

1.2.2.2. L'éjection des ions

L'éjection des ions est réalisée par deux méthodes :

- par balayage croissant de l'amplitude la radiofréquence (éjection simple)

et/ou

- par application d'une tension alternative supplémentaire sur les électrodes chapeaux dont on fait varier la fréquence pour faire entrer en résonance les ions de différents m/z (éjection résonante).

1.2.2.3. L'isolation des ions

L'isolation des ions consiste à éjecter de la trappe tous les ions qui n'ont pas le rapport m/z souhaité. Celle-ci est réalisée par éjection simple des ions de m/z inférieur au m/z de l'ion à isoler et par éjection résonante de tous les ions de m/z supérieur au m/z de l'ion à isoler. Dans les faits, l'éjection résonante peut aussi être utilisée pour éjecter les ions de m/z inférieur.

1.2.2.4. Excitation et fragmentation des ions

La fragmentation des ions est réalisée par ajustement de la fréquence de la tension alternative sur l'électrode chapeau de manière à ce qu'elle corresponde à la fréquence de résonance de l'ion à fragmenter. L'amplitude de cette tension est réglée pour éviter d'éjecter les ions à fragmenter de la trappe. Les ions ainsi excités entrent en collision avec les molécules d'hélium et voient leur énergie cinétique transformée en énergie interne vibrationnelle, ce qui va entraîner leur fragmentation.


1.2.2.5. Séquence d'analyse de peptides en MS/MS

Pour les analyses protéomiques, les différentes opérations décrites précédemment seront réalisées successivement. Ainsi, durant ces analyses, tout au long de l'élution chromatographique des différents peptides analysés, des cycles MS et MS/MS seront alternés selon la séquence suivante :

- Piégeage des ions dans la trappe
- Ejection des ions

} MS

Sélection des ions d'intérêt

- Piégeage des ions
 - Isolations des ions d'intérêt
 - Excitation et fragmentations des ions isolés
 - Ejection des ions fragments.
- 

La trappe ionique permet également de réaliser des analyses MSⁿ. En effet les différentes étapes décrites ci-dessus pour des analyses MS/MS peuvent être répétées plusieurs fois. Ce type d'analyse peut fournir des informations très importantes d'un point de vue structural ou pour l'analyse de modifications post-traductionnelles par exemple.

Les trappes ioniques qui ont été utilisées pour réaliser les travaux présentés durant cette thèse sont des trappes HCT Ultra (Bruker Daltonics) qui présentent de très grandes capacités de piégeage d'ions pour des trappes 3D. Les trappes ioniques sont des outils de choix pour l'analyse protéomique de part leur très grande vitesse de balayage ou plus globalement par leur très grande vitesse pour enchaîner les cycles MS et MS/MS, ce qui permet de fragmenter un nombre très important de peptides en un temps limité. Cette particularité est d'un avantage précieux pour les analyses en couplage avec des systèmes chromatographiques pour lesquels les pics des peptides ne durent que quelques secondes (par exemple les systèmes de puces microfluidiques ou les systèmes de chromatographie à ultra haute pression) ou pour les stratégies de quantification par les approches « spectral counting » (décrites dans la partie III des résultats). La principale limite de cet analyseur est sa résolution limitée et sa faible justesse de mesure de masse. On peut également mentionner la gamme de masse limitée de piégeage des ions dans la trappe. L'efficacité de piégeage des ions d'un rapport m/z donné est fonction de l'amplitude de la radiofréquence appliquée sur l'électrode annulaire de la trappe. L'efficacité du piégeage ne peut pas être optimal à la fois pour les ions de rapport m/z faible et pour les ions de rapport m/z élevé. Ce phénomène est notamment à l'origine de la non-observation des ions immonium dans les spectres MS/MS

1.3. La fragmentation peptidique

De nos jours, la spectrométrie de masse en tandem est devenue un outil incontournable pour réaliser l'analyse MS/MS des peptides et donc déterminer leur séquence en acides aminés. L'identification des séquences peptidiques à partir de l'interprétation des spectres MS/MS est souvent très complexe. Même si en analyse protéomique, cette identification est généralement réalisée par des algorithmes de recherche dans les banques de séquences protéiques, il est nécessaire de connaître les mécanismes de fragmentation des peptides en phase gazeuse. Cette connaissance est indispensable pour le bon paramétrage des moteurs de recherche ou le développement de nouveaux moteurs de recherche, pour l'interprétation manuelle des spectres identifiés à partir des moteurs de recherche voire pour l'interprétation *de novo* de ces spectres.

Les deux modes de fragmentation les plus courants pour séquencer les peptides sont la fragmentation induite par collision (« Collision-Induced Dissociation », CID) et la fragmentation par capture d'électron (« Electron Capture Dissociation », ECD) ou par transfert d'électrons (« Electron Transfer Dissociation », ETD). Pour les fragmentations CID, on peut encore distinguer les fragmentations basse et haute énergie. Comme les spectres MS/MS obtenus sur les trappes ioniques et le Q-TOF utilisés pendant le travail de thèse mettent en jeu des énergies de fragmentation inférieures à une centaine d'eV, nous ne présenterons dans ce paragraphe que les mécanismes de fragmentations CID basse énergie.

Ce type de fragmentation est activé par collision avec un gaz inerte (par exemple de l'hélium sur les trappes HCT Ultra et de l'argon pour le Q-TOF Synapt).

1.3.1. Les étapes de la fragmentation

Classiquement, en couplage « chromatographie en phase inverse-spectrométrie de masse », quelque soit le ou les analyseurs de masse utilisés, pour obtenir un spectre de fragmentation d'un peptide informatif, on procède en trois étapes (Figure 5) :

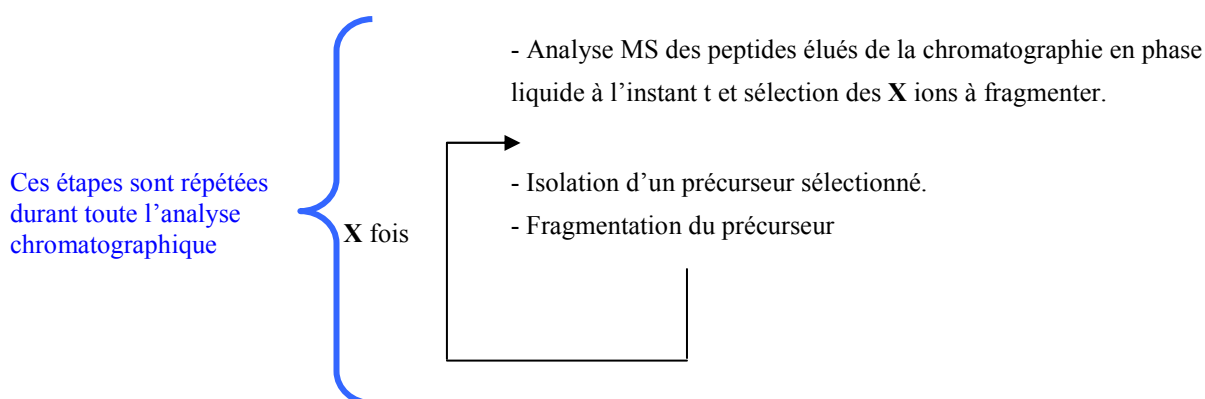
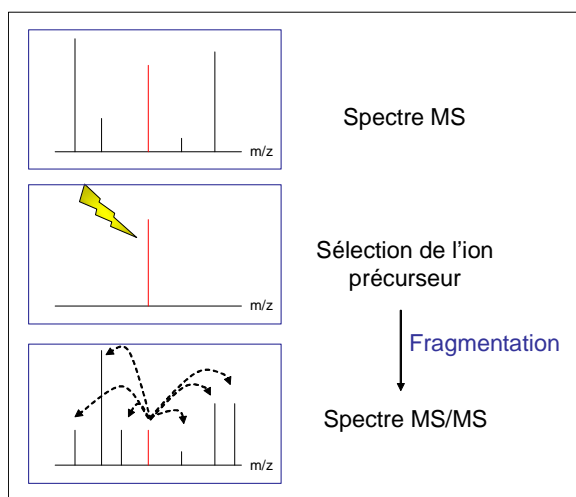


Figure 5 : Illustration des différentes étapes de la fragmentation d'un ion



1.3.2. La nomenclature des fragmentations peptidiques

La nomenclature acceptée pour les ions fragments fut proposée pour la première fois par [Roepstorff et al., 1984] puis modifiée par [Johnson et al., 1987] et [Biemann, 1988]. La nomenclature proposée par Biemann reste toujours la nomenclature officielle utilisée (Figure 6).

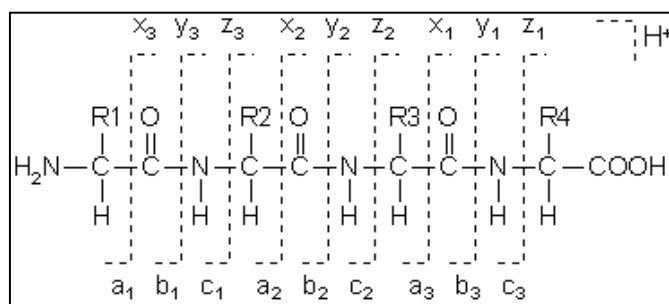
Globalement, dans cette nomenclature, pour les fragmentations basse énergie, on va distinguer 2 types d'ions fragments :

- Les ions qui contiennent la partie N-terminale du peptide précurseur qui sont les fragments a, b et c.

- Les ions qui contiennent la partie C-terminale du peptide précurseur qui sont les fragments x, y et z.

De plus, dans certains cas, on va observer des phénomènes de double fragmentation. Dans ce cas, on parle de fragments internes. Généralement, ils sont issus d'une fragmentation simultanée de type b- et y-. Parfois, les fragments internes proviennent d'une fragmentation concomitante de type a- et y- et dans ce cas on obtient un ion amino-immonium. Si en plus cet ion ne contient qu'une seule chaîne latérale, on parle alors d'ions immonium. Ces ions détectables dans la zone de faible m/z du spectre MS/MS peuvent être très utiles pour l'interprétation du spectre notamment avec une stratégie de séquençage *de novo*.

Figure 6 : Nomenclature de Biemann



1.3.3. Mécanismes de fragmentation

1.3.3.1. Le modèle du proton mobile

Le modèle le plus complet actuellement disponible pour décrire comment les peptides protonés sont fragmentés est le modèle du « proton mobile ». Ce modèle fut établi à la suite de nombreuses études réalisées par les groupes de Wysocki [Jones et al., 1994; Dongre et al., 1996; Tsaprailis et al., 1999; Wysocki et al., 2000], Harrison [Harrison et al., 1997], Gaskell [Cox et al., 1996] et d'autres. Les peptides protonés activés avec de faibles énergies de collision fragmentent

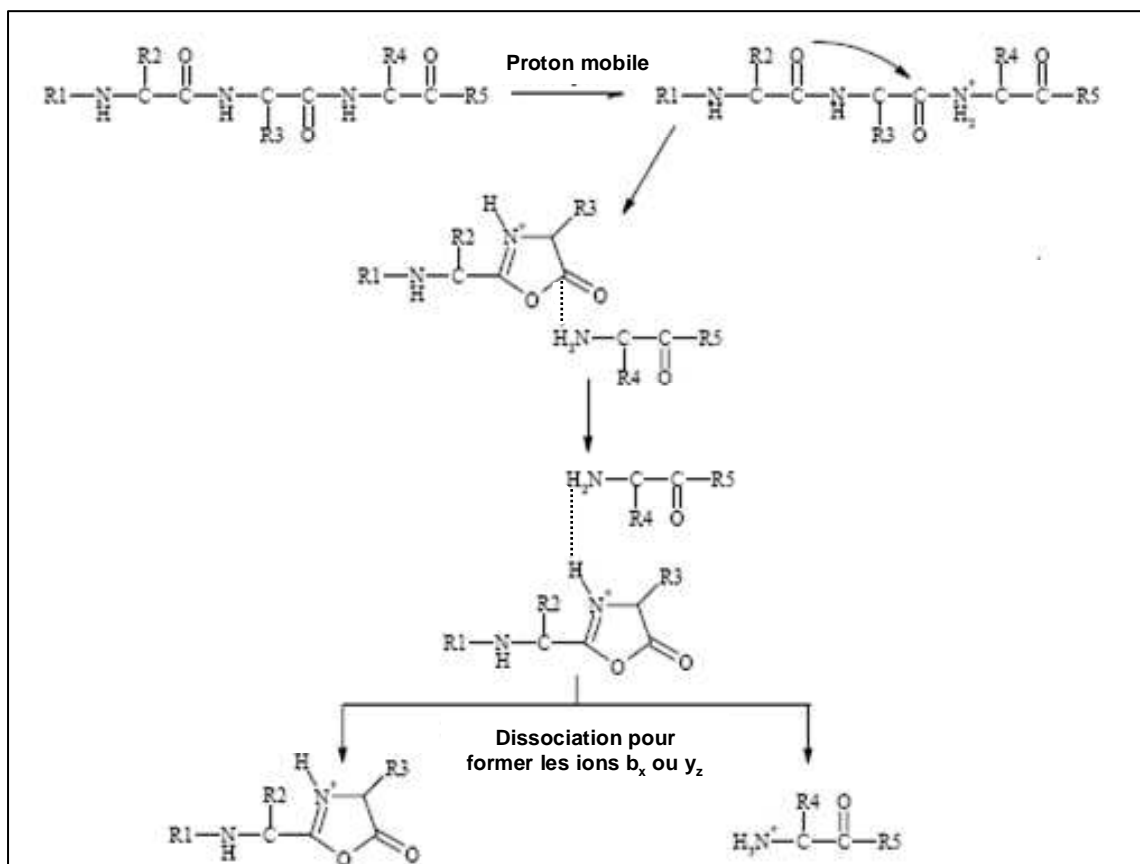
essentiellement par des réactions dirigées par la charge (« charge-directed fragmentation ») [Cox et al., 1996].

Les peptides étant des composés multifonctionnels, ils peuvent être protonés sur différents sites qui peuvent être :

- Très favorables : l'acide aminé N-terminal et les groupements basiques des chaînes latérales.
- Moins favorables : les atomes d'oxygène des liaisons amide.
- Très peu favorables : les atomes d'azote des liaisons amide.

Toutefois, l'activation des peptides par collision avec des molécules de gaz va provoquer l'excitation de la population initiale (d'abord protonée sur des sites favorables) et provoquer la migration des protons vers des sites moins favorables, comme les atomes d'azote des liaisons amides. Or, cette protonation conduit à l'affaiblissement de la liaison amide, fait du carbone du groupement carbonyle adjacent une possible cible pour une attaque nucléophile et initie la fragmentation au niveau des liaisons peptidiques, ce qui entraîne la formation d'ions fragments b- et y- qui retiennent la charge en fonction de leur affinité protonique [Paizs et al., 2004] (Figure 7).

Figure 7: Fragmentation du peptide selon le modèle du proton mobile. Attaque nucléophile de l'oxygène du carbonyle de la liaison peptidique en N-terminal de la liaison peptidique protonée sur le carbone de cette liaison peptidique protonée puis dissociation du peptide.



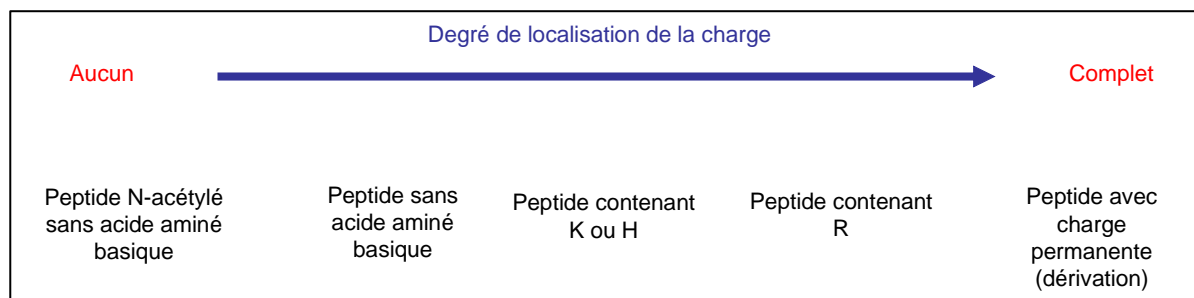
L'énergie nécessaire pour mobiliser le(s) proton(s) à délocaliser est fonction de la composition en acides aminés (particulièrement en acides aminés basiques), de la séquence et de la taille du peptide étudié [Wysocki et al., 2000]. Ainsi la mobilisation d'un proton sera d'autant plus difficile et demandera d'autant plus d'énergie qu'il est fortement séquestré sur un groupement. Ce niveau d'énergie nécessaire augmente à mesure qu'augmente la basicité en phase gazeuse du peptide (Figure 8).

Par exemple pour fragmenter un peptide tryptique monochargé qui a une arginine en C-terminal, il faudra beaucoup plus d'énergie que si ce peptide possède une lysine en C-terminal. Si les protons sont très fortement séquestrés, les fragmentations distantes de la charge deviennent très compétitives [Roth et al., 1998].

Pour des peptides tryptiques doublement chargés, les deux charges sont en général localisées sur l'amine libre N-terminale et sur la chaîne latérale de la lysine ou l'arginine C-terminale. La mobilisation du proton en N-terminal requiert peu d'énergie et on observe en général des fragments y- et b-. Les ions y- sont souvent plus majoritaires car plus résistant aux décompositions qui peuvent suivre. En effet, dans notre cas, l'ion y- retient sa charge sur la chaîne latérale de l'acide aminé C-ter alors que le proton initiant la fragmentation est retenu sur le squelette et peut engendrer de nouvelles fragmentations [Cox et al., 1996].

Malgré la prédominance des fragments y- et b-, on peut observer d'autres types de fragments dans les spectres MS/MS notamment les fragments a- (qui proviennent par exemple de perte de CO sur les fragments b- [Johnson et al., 1987]), les fragments internes ou les pertes de neutres. Les différents schémas de fragmentation proposés sont décrits en détail dans [Paizs et al., 2004].

Figure 8 : Degré de localisation de la charge en fonction des groupements basiques du peptide



1.3.3.2. Les coupures préférentielles

Plusieurs auteurs ont souligné les spectres MS/MS inhabituels de certains peptides qui présentent des coupures préférentielles à certains acides aminés ou certaines combinaisons d'acides aminés [Tsapraïlis et al., 1999; Wysocki et al., 2000; Breci et al., 2003; Huang et al., 2005]. Dans ces cas de figure, les spectres MS/MS ne permettent pas de fournir une séquence complète du squelette peptidique. On peut mentionner :

- Les résidus proline qui présentent une fragmentation préférentielle en N-terminal (l'azote plus basique de la proline favorise la séquestration du proton mobile) et difficile en C-terminal (encombrement stérique) quand le peptide présente un proton mobile.
- Les résidus acides qui présentent une fragmentation préférentielle en C-terminal quand le peptide n'a pas de proton mobile (fragmentation distante de la charge)
- Les résidus histidine qui présentent une fragmentation préférentielle en C-terminal quand ils sont protonés.
- Les résidus aliphatiques (valine, leucine, isoleucine) qui présentent :
 - ✓ une fragmentation préférentielle en C-terminal avec observation des fragments b- et y- quand le peptide a un proton mobile, avec observation des fragments y- quand le peptide a un proton partiellement mobile.
 - ✓ une fragmentation préférentielle en N-terminal avec observation des fragments b- quand le peptide a un proton partiellement mobile.

2. Les techniques séparatives de protéines et de peptides

En parallèle des développements des instruments de spectrométrie de masse, le développement des techniques séparatives, pour décomplexifier ou fractionner l'échantillon protéique avant l'analyse par spectrométrie de masse, ont contribué à l'essor de l'analyse protéomique par spectrométrie de masse.

La séparation peut se faire à 2 niveaux : au niveau protéique et au niveau peptidique. La méthode de séparation historique consistait à séparer les protéines par gel bidimensionnel (gel 2D) puis à analyser le digest peptidique directement ou par couplage de la chromatographie en phase inverse avec le spectromètre de masse. Des limitations inhérentes aux gels 2D ont conduit au développement de techniques séparatives additionnelles (gel monodimensionnel, chromatographie multidimensionnelle des peptides, etc...). Les stratégies de séparation les plus usuelles avant analyse par spectrométrie de masse seront présentées dans ce chapitre et résumées en Figure 10.

2.1. Séparation des protéines

2.1.1. Le gel d'électrophorèse bidimensionnel (gel 2D)

Cette technique consiste à séparer les protéines en fonction de leur point isoélectrique (pI) et de leur masse moléculaire. Après visualisation du gel, les spots d'intérêt sont découpés, digérés avec une protéase (généralement la trypsine) et les peptides ainsi générés sont analysés par MS (MALDI-

TOF) pour générer une empreinte peptidique massique (PMF) ou par MS/MS pour obtenir les informations de séquence (couplage LC-MS/MS).

Un gel 2D peut permettre de séparer plusieurs milliers de protéines et de détecter et quantifier des protéines à moins de 1ng par spot [Gorg et al., 2004]. Le gel 2D est également la méthode de choix pour différencier et quantifier les variants post-traductionnels et les protéines isoformes.

Toutefois, la réalisation des expériences de gel 2D (réalisation des gels, visualisation des spots, digestion, extraction et analyse MS ou MS/MS de chaque spot) est difficile à automatiser, laborieuse et ne peut pas être réalisée à haut débit. De plus, la gamme dynamique visible sur le gel est restreinte, et certaines catégories de protéines (membranaires, pI extrêmes) sont exclues du gel [Santoni et al., 2000; Braun et al., 2007].

2.1.2. Le gel monodimensionnel (gel 1D ou gel SDS [sodium DodecylSulfate] PAGE [PolyAcrylamide Gel Electrophoresis])

Cette technique consiste à séparer les protéines en fonction de leur masse moléculaire. De même que pour les gels 2D, les bandes sont ensuite découpées, digérées avec une protéase (généralement la trypsine) et les peptides ainsi générés sont analysés par MS (MALDI-TOF) ou MS/MS (couplage LC-MS/MS).

Cette technique, bien que non automatisable non plus, est plus rapide et plus facile à mettre en œuvre que le gel 2D et permet d'analyser les protéines basiques et hydrophobes [Xiong et al., 2005].

Toutefois, le pouvoir résolutif du gel 1D est très limité en comparaison des gels 2D. Sur un extrait biologique complexe, plusieurs centaines de protéines pourront avoir co-migré dans la même zone du gel. L'analyse directe MS des digests peptidiques ne sera applicable qu'à des études de sous-protéomes qui ne présentent en général qu'un faible nombre de protéines. Sinon, et ce sera le cas le plus souvent, l'analyse par couplage LC-MS/MS sera la règle.

2.1.3. Les méthodes alternatives sans gel (« gel free »)

En plus des méthodes qui passent par l'utilisation de gels d'électrophorèse, d'autres méthodes pour fractionner les protéines existent.

La méthode de séparation des protéines par isoélectrofocalisation, l'IEF off-gel, permet de recouvrer les protéines fractionnées en solution selon le principe de séparation en fonction du pI (comme la première dimension d'un gel 2D) [Arnaud et al., 2002; Ros et al., 2002]. Cette méthode peut être couplée à une deuxième séparation par gel 2D ou par chromatographie liquide avant

digestion et analyse LC-MS/MS. Cette méthode présente l'avantage de conserver les protéines dans un milieu liquide pour un meilleur recouvrement de l'échantillon de départ mais ne s'affranchit pas des problèmes liés à la première dimension du gel 2D (pI extrêmes, protéines membranaires). L'IEF off-gel peut aussi être utilisé pour séparer les peptides [Lam et al., 2007] par exemple comme première dimension de séparation.

Les méthodes de chromatographie liquide (LC) constituent une autre alternative pour la séparation des protéines. Les types de chromatographies qui ont déjà été appliqués à cette fin sont assez variés comme la chromatographie d'exclusion stérique [Zhang et al., 2001], la chromatographie en phase inverse [Le Coutre et al., 2000] ou la chromatographie d'échange d'ions [Schluesener et al., 2005] en couplage ou non avec une méthode chromatographique de sélectivité différente [McDonald et al., 2006]. Les méthodes de LC sont assez peu utilisées en première dimension de séparation des protéines car la solubilisation des protéines nécessite souvent l'utilisation de détergents ou de chaotropes qui ne sont pas toujours compatibles avec la LC utilisée et la résolution des séparations est très médiocre.

2.2. Séparation des peptides

La diversité physico-chimique des peptides (charge, point isoélectrique, hydrophobicité, taille) fait qu'ils peuvent être séparés par presque tous les modes de séparation en LC [Sandra et al., 2008]. Par conséquent, en général, la séparation des peptides est réalisée par chromatographie liquide et le plus souvent par chromatographie liquide en phase inverse.

2.2.1. Le couplage RPLC-MS/MS (chromatographie liquide en phase inverse couplée à la spectrométrie de masse en tandem)

Ce couplage direct est la technique d'analyse principale en protéomique. Elle est utilisée pour séparer et analyser un mélange de peptides provenant de la digestion d'un spot de gel 2D, d'une bande de gel 1D voire de la digestion d'un mélange plus complexe.

En RPLC (généralement sur colonne C₁₈), les peptides sont le plus souvent séparés en mode « gradient » grâce à leur différence de facteur de capacité (qui reflète directement l'affinité du composé pour la phase mobile et la phase stationnaire). L'acétonitrile est le modificateur organique de choix et la phase mobile contient un réactif de paire d'ions tel que l'acide trifluoroacétique (TFA) ou formique.

Comme l'intensité du signal en ESI-MS est concentration dépendant, la recherche d'une sensibilité accrue a conduit au développement de systèmes permettant de travailler à faibles (quelques $\mu\text{L}/\text{min}$) ou très faibles ($\sim 100\text{-}300\text{ nL}/\text{min}$) débits (nanoLC) et au développement de colonnes de très faible diamètre interne ($75\text{-}300\ \mu\text{m}$). Ces avancées permettent au couplage nanoLC-MS/MS de détecter quelques femtomoles de matériel.

Les développements technologiques de ces dernières années ont visé à améliorer la qualité de séparation des composés et donc à augmenter la capacité de pic [Sandra et al., 2008]. Pour cette raison, on a assisté au développement de :

- Colonnes contenant des particules de faible diamètre et qui sont utilisées en UHPLC (« Ultra High Pressure Liquid Chromatography ») (système nanoAcquity UPLC Waters par exemple) pour des séparations de peptides plus efficaces [Shen et al., 2005; Shen et al., 2005].
- Colonnes monolithiques (qui peuvent être de très grande taille, jusqu'à 140 cm [Ikegami et al., 2004]) pour des séparations de peptides plus efficaces [Premstaller et al., 2001; Rieux et al., 2005].
- Systèmes de puce microfluidique intégrant colonne de chargement, colonne analytique et pointe nanospray pour la réduction des volumes morts (système nanoHPLC-Chip cube Agilent par exemple) [Fortier et al., 2005].

Les analyses RPLC-MS ou RPLC-MS/MS réalisées dans les études présentées dans cette thèse ont été effectuées sur des systèmes Agilent microLC (débit de 4 $\mu\text{L}/\text{min}$, colonne de 300 μm de diamètre interne, particules de 3.5 μm) et nanoLC (nanoHPLC-Chip Agilent, débit de 300 nL/min, colonne de 75 μm diamètre interne, particules de 3.5 μm) et sur un système nanoLC Waters (nanoAcquity Waters, débit de 400nL/min, colonne de 75 μm diamètre interne, particules de 1.7 μm).

2.2.2. La chromatographie liquide multidimensionnelle

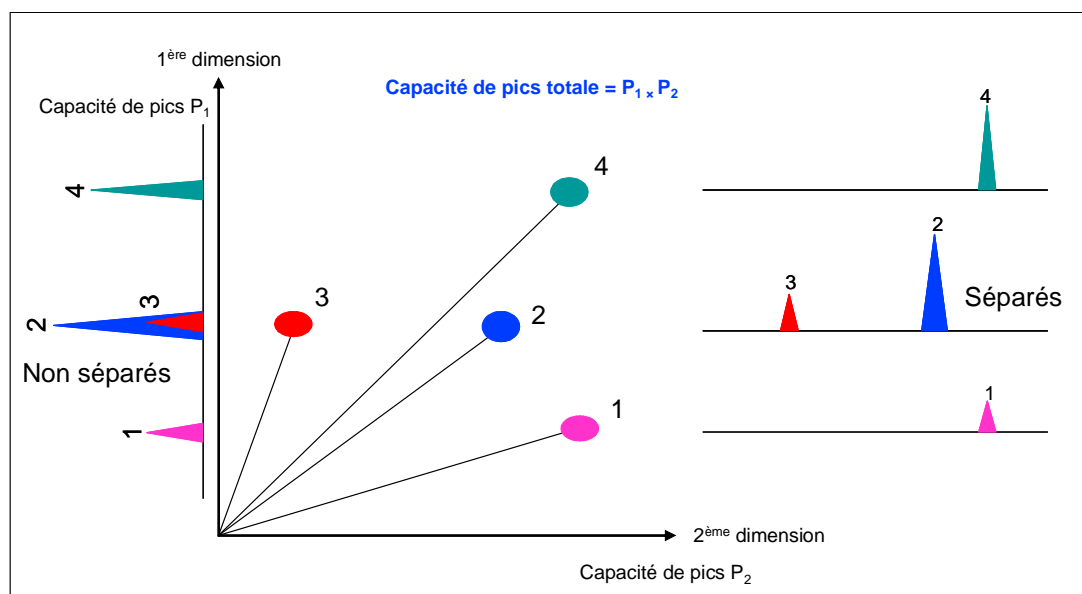
Si le mélange de peptides à analyser en RPLC-MS/MS est très complexe (par exemple issu de la digestion d'un extrait protéique total), il pourra être nécessaire de réaliser une première étape de séparation en amont de la RPLC, c'est ce que l'on appelle la chromatographie multidimensionnelle (dont le concept fut décrit la première fois en 1984 par Giddings [Giddings, 1984]) ou encore MudPIT (« Multi-Dimensional Protein Identification Technology ») [Washburn et al., 2001; Wolters et al., 2001].

Cette stratégie a ouvert la voie à ce que l'on appelle la protéomique « shotgun » (par analogie aux approches « shotgun » de séquençage des génomes). Cette stratégie consiste à analyser directement par MS et MS/MS les peptides générés après digestion de mélanges complexes de protéines et chromatographies multiples. On obtient ainsi un profil global des protéines présentes dans le mélange.

2.2.2.1. Le principe clé de la chromatographie multidimensionnelle : l'orthogonalité

Le principe de la chromatographie multidimensionnelle consiste à coupler des systèmes de chromatographie basés sur des mécanismes de rétention différents pour obtenir des séparations orthogonales entre les 2 dimensions du système (Figure 9). Dans un système de chromatographie multidimensionnelle idéal, la capacité de pics de l'ensemble du système devient le produit des capacités de pics de chaque dimension. Cependant, en réalité, les méthodes de séparation couplées ne sont jamais basées sur des mécanismes complètement différents et l'orthogonalité des méthodes n'est donc jamais parfaite [Motoyama et al., 2008].

Figure 9 : Illustration du concept de séparation orthogonale à 2 dimensions. D'après [Motoyama et al., 2008]



2.2.2.2. Les combinaisons standards

Une très grande variété de combinaisons entre les dimensions de la chromatographie multidimensionnelle a été décrite [Opiteck et al., 1997; Link et al., 1999; Mawuenyega et al., 2003; Delmotte et al., 2007]. Ces méthodes utilisent différentes techniques chromatographiques et un nombre de dimensions différent. Toutefois, la plupart du temps, la dernière étape de séparation interfacée directement avec le spectromètre de masse reste la RPLC. Par contre, la première étape peut être la chromatographie d'exclusion stérique [Opiteck et al., 1997], la chromatographie d'échange de cations (méthode la plus utilisée) [Washburn et al., 2001; Wolters et al., 2001; Wagner et al., 2003] ou d'échange d'anions [Motoyama et al., 2007], la chromatographie en phase inverse à pH basique [Delmotte et al., 2007] ou encore la chromatographie d'interaction hydrophile [Boersema et al., 2007].

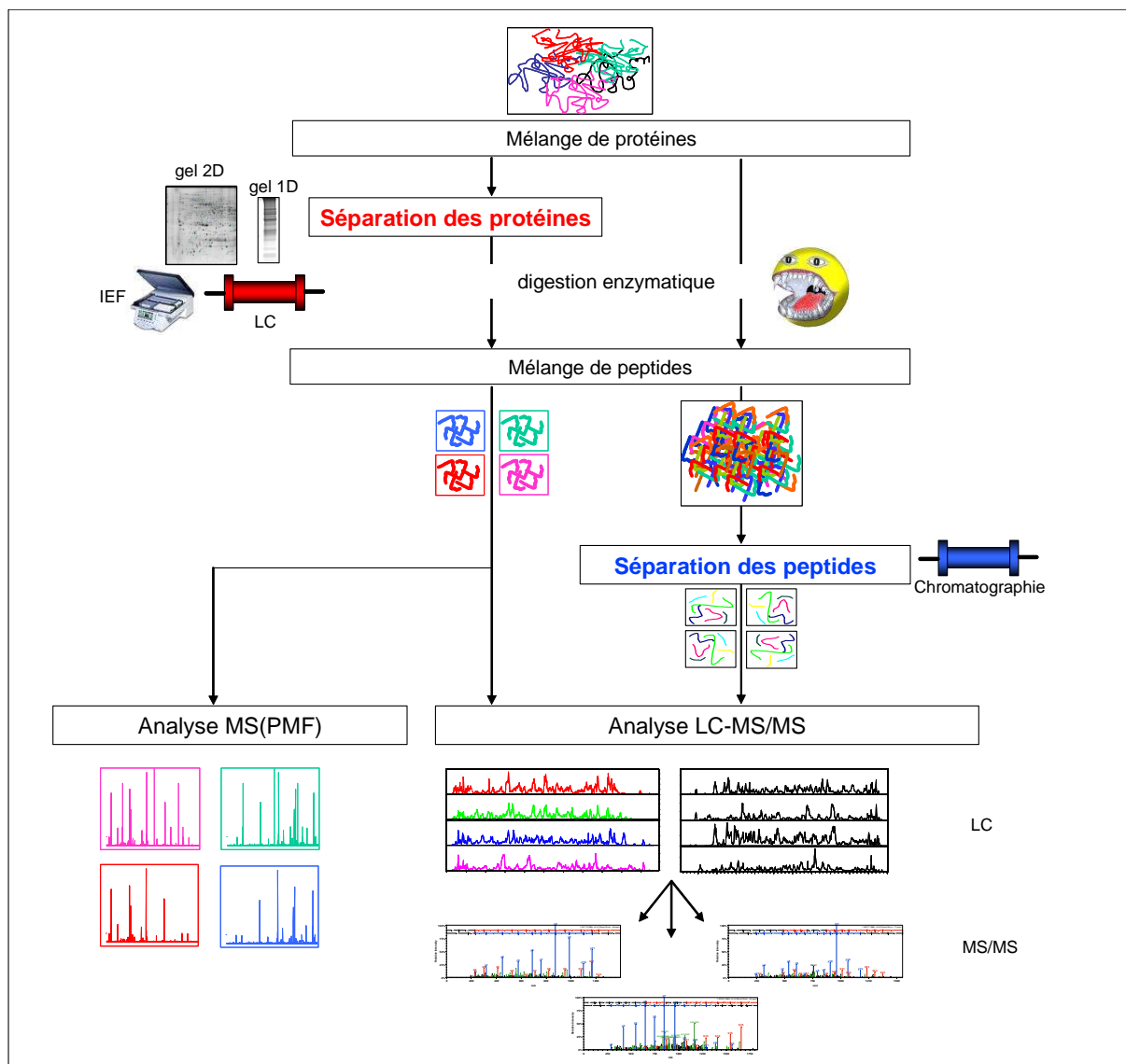
2.2.2.3. Configurations « off-line » et « on-line »

L'enchaînement des étapes de séparation peut être réalisé en configuration « off-line » (discontinue) ou « on-line » (continue). Les deux configurations présentent des avantages et des inconvénients.

La configuration « off-line » nécessite l'utilisation d'un collecteur de fractions dans la première dimension et les fractions ainsi collectées sont injectées dans la deuxième dimension. Cette configuration est bien adaptée pour les échantillons extrêmement complexes mais disponibles en relativement grande quantité. Cette configuration offre également une plus grande flexibilité et peut permettre une qualité de séparation globale supérieure parce que chaque dimension de séparation peut être optimisée indépendamment [Peng et al., 2003] et l'efficacité de la séparation dans la première dimension peut être contrôlée grâce à un système de détection. En plus, la configuration « off-line » permet une utilisation plus facile de sels non-volatils en chromatographie d'échange de cations. Enfin, cette configuration offre la possibilité de réanalyser les fractions collectées et de collecter plus de fractions sur la première dimension. Le désavantage majeur est qu'on ne peut pas automatiser totalement l'approche et qu'il existe un risque de perte d'échantillon ou de contamination.

Avec la configuration « on-line », les analytes sont transférés entre les deux dimensions de manière automatisée. Les avantages majeurs de cette configuration sont l'automatisation facile et complète du système et un moins grand risque de perte d'échantillon ou de contamination. Par contre cette configuration est moins adaptée aux échantillons extrêmement complexes.

Figure 10 : Les stratégies de séparation de protéines/peptides « classiques » en analyse protéomique



Chapitre 4 : Les stratégies d'identification en analyse protéomique

L'analyse protéomique par spectrométrie de masse couplée aux méthodes de séparation des protéines et des peptides fournit un ensemble de spectres MS et MS/MS reflétant l'ensemble des protéines détectables présentes dans l'échantillon. L'interprétation de ces spectres est indispensable pour effectuer l'identification des protéines correspondantes. Celle-ci passe par la mise en relation des spectres de masse avec les séquences protéiques correspondantes qui est réalisée grâce à des outils bioinformatiques.

Pour juger de la pertinence des conclusions de l'étude ou pour comparer significativement les résultats de différentes études, il est indispensable pour la protéomique de développer des outils d'évaluation et de diffusion des identifications pour valoriser les résultats.

Les principales stratégies d'identification des protéines obtenues avec les méthodes décrites dans le chapitre précédent ainsi que leurs méthodes de valorisation associées seront décrites dans le chapitre suivant.

1. L'empreinte peptidique massique (« Peptide Mass fingerprinting », PMF)

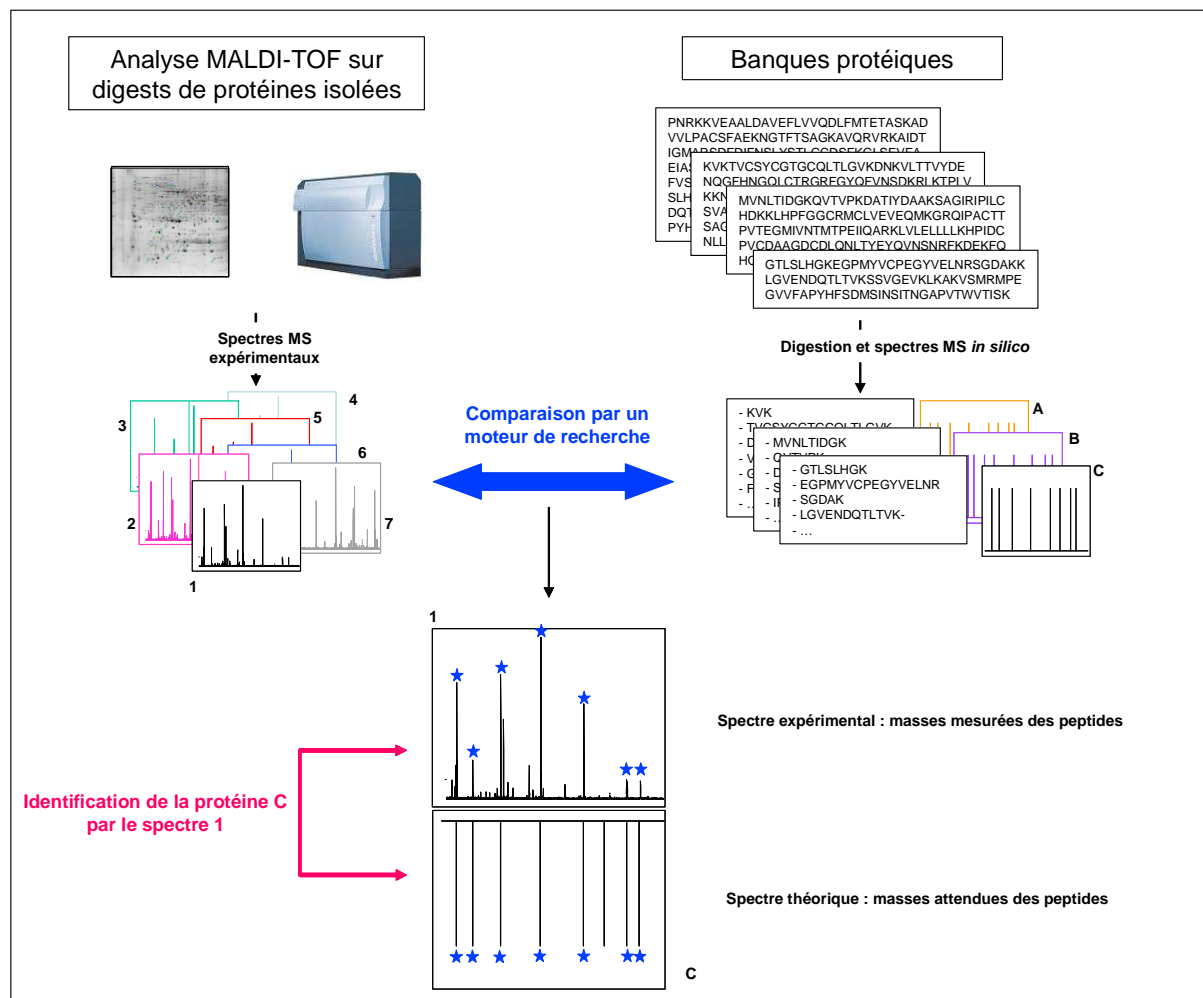
Cette stratégie consiste à comparer les masses expérimentales des peptides générés par digestion enzymatique d'une protéine aux masses théoriques des peptides obtenus par digestion *in silico* de l'ensemble des séquences d'une banque protéique (Figure 1). Cette stratégie est apparue en 1993 et est l'œuvre simultanée de 5 équipes [Henzel et al., 1993; James et al., 1993; Mann et al., 1993; Pappin et al., 1993; Yates et al., 1993].

Généralement, la stratégie PMF est utilisée pour réaliser l'identification des protéines séparées par gel 2D puis digérées et analysées par MS avec un instrument présentant une bonne précision et une bonne justesse de mesure de masse. En effet, pour une identification efficace des protéines, il est nécessaire de limiter au maximum la superposition d'empreintes peptidiques de plusieurs protéines et de réduire l'espace de recherche dans les banques. De nombreux programmes d'identification par PMF ont été développés ; parmi les plus connus on trouve MASCOT [Perkins et al., 1999] ou encore MS-Fit [Clauser et al., 1999].

Les analyses MS réalisées sur l'instrument MALDI-TOF-MS sont particulièrement bien adaptées pour le PMF car elles sont rapides, automatisables et présentent une bonne résolution et une bonne justesse de mesure de masse (~10 ppm). Les instruments de dernière génération combinant des sources ESI ou MALDI avec des analyseurs à haute voire très haute résolution (TOF, Q-TOF, FTICR, Orbitrap) peuvent aussi réaliser du PMF.

Bien que la stratégie d'identification par PMF soit efficace dans bien des cas, le séquençage MS/MS des peptides est une méthode d'identification bien plus sensible et spécifique [Steen et al., 2004] et qui ne nécessite pas une décomplexification préalable très poussée du mélange des protéines comme c'est le cas en PMF.

Figure 1 : Illustration de la stratégie d'identification par PMF sur des analyses MALDI-TOF-MS



2. Les approches par LC-MS/MS

Cette stratégie consiste à utiliser les données des spectres MS/MS pour réaliser l'identification des séquences des peptides présents dans l'échantillon.

A partir d'un échantillon protéique plus ou moins complexe, on réalise une digestion enzymatique qui génère des peptides qui sont ensuite analysés par LC-MS/MS. Lors des analyses LC-MS/MS, la sélection des ions parents qui seront fragmentés au cours de l'analyse chromatographique est réalisée de manière automatique sur la base de leur intensité, leur rapport m/z ou encore leur état de charge.

Au cours de l'analyse, on alterne :

- ✓ les spectres MS qui servent à mesurer les rapports m/z de l'ensemble des peptides arrivant au détecteur de l'instrument à l'instant t et à en sélectionner un certain nombre qui seront fragmentés

Et

- ✓ les spectres MS/MS qui servent à mesurer la masse de l'ensemble des fragments issus de la fragmentation du peptide sélectionné.

Les analyses LC-MS/MS ne nécessitent pas une décomplexification préalable très poussée du mélange des protéines comme c'est le cas en PMF.

Les spectres MS/MS obtenus correspondant à la fragmentation des peptides dont on connaît la masse sont ensuite interprétés pour obtenir des informations de séquence sur les peptides qui pourront ensuite permettre de remonter à l'identité des protéines dont ils sont issus.

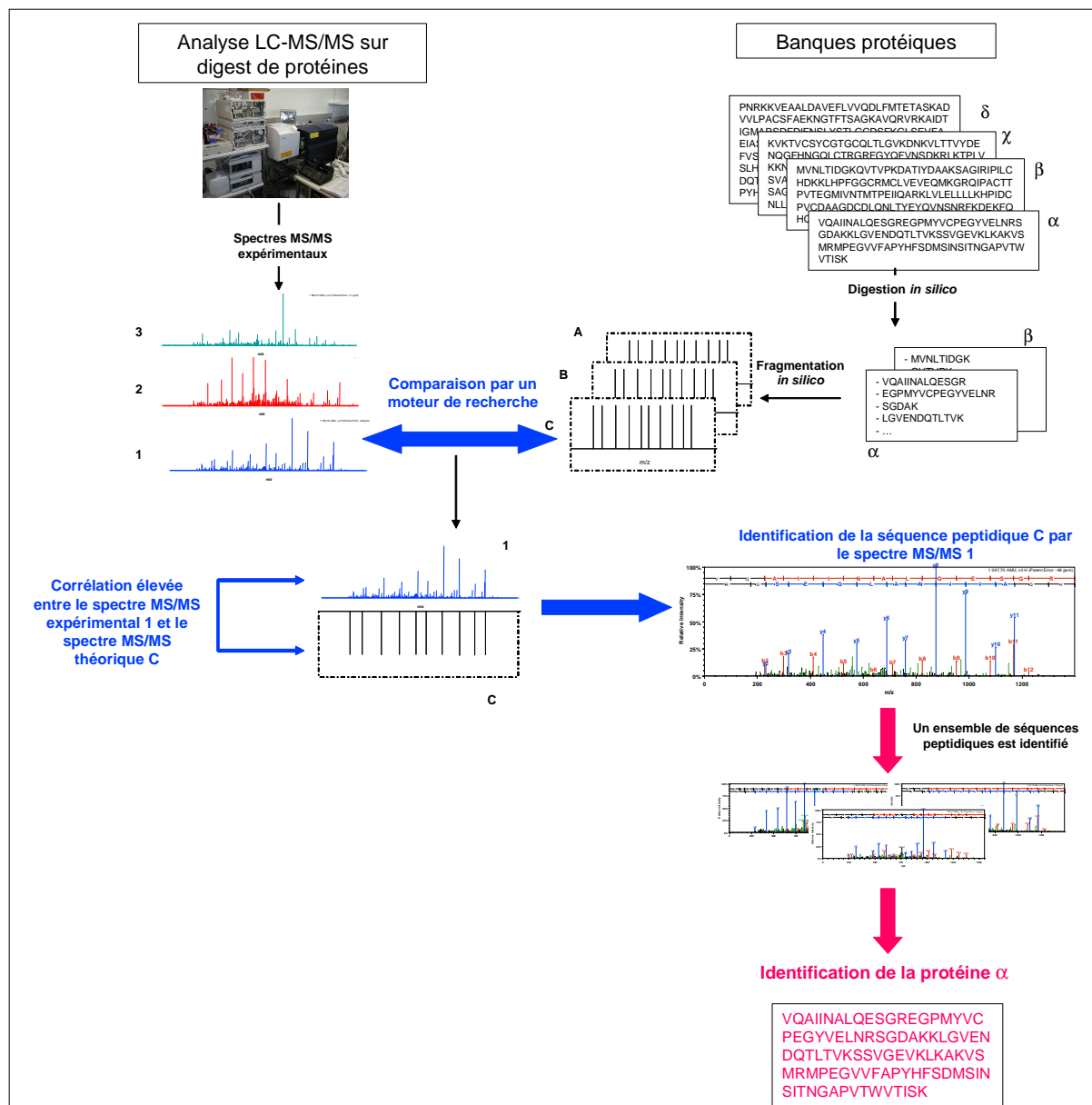
Il existe plusieurs approches d'interprétation de ces informations qui sont assistées par des outils bioinformatiques. Les deux principales approches qui seront décrites par la suite sont :

- ✓ L'approche par recherche dans les banques protéiques.
- ✓ L'approche par séquençage *de novo*.

2.1. L'approche par recherche dans les banques protéiques

Ce type d'approche, parfois appelée « Peptide Fragment Fingerprinting » (PFF) par analogie au PMF [Blueggel et al., 2004], est la méthode d'identification la plus fréquemment utilisée dans les études protéomiques à grande échelle. Dans cette approche, les séquences des peptides sont identifiées en corrélant les spectres de fragmentation expérimentaux avec les spectres théoriques prédits pour chaque peptide issus de la digestion *in silico* de l'ensemble des protéines présentes dans une banque. L'ensemble des peptides identifiés dans cette approche permet de remonter à l'identité des protéines présentes dans l'échantillon (Figure 2).

Figure 2 : Identification des peptides et des protéines par l'approche MS/MS avec recherche dans les banques protéiques



Il existe un grand nombre de programmes permettant de réaliser cette approche. Parmi les plus connus, on trouve :

- Mascot (<http://matrixscience.com>) [Perkins et al., 1999]
- Sequest (<http://fields.scripps.edu/sequest/>) [Eng et al., 1994]
- ProteinProspector (<http://prospector.ucsf.edu>) [Clauser et al., 1999]
- X!Tandem (<http://www.thegpm.org>) [Craig et al., 2004]
- SpectrumMill (<http://www.chem.agilent.com>)
- Phenyx (<http://www.phenyx-ms.com>) [Colinge et al., 2003]
- OMSSA (<http://pubchem.ncbi.nlm.nih.gov/omssa>) [Geer et al., 2004]

Même si l'ensemble de ces programmes ont un système de fonctionnement assez proche, ils divergent pour certains points du processus d'identification. Je présenterai dans la suite du chapitre les trois programmes utilisés pendant ma thèse : Mascot et les deux programmes « open-source » X!Tandem et OMSSA.

Ces programmes effectuent la corrélation des spectres expérimentaux avec les spectres théoriques calculés pour les peptides de la banque protéique considérée en tenant compte de plusieurs paramètres réglés par l'utilisateur :

- ✓ la tolérance sur la masse du précurseur
- ✓ l'enzyme de digestion utilisée (et les coupures enzymatiques non réalisées autorisées)
- ✓ les modifications des acides aminés autorisées
- ✓ la tolérance sur la masse des ions fragments
- ✓ les types de fragments recherchés

2.1.1. Attribution d'un « score » à l'identification

2.1.1.1. Le score de corrélation

Après avoir traité les données, les programmes fournissent une liste de séquences peptidiques attribuées à des spectres MS/MS qui sont classées en fonction d'un score de corrélation. Ce score de corrélation mesure le degré de similarité entre le spectre expérimental et le spectre théorique et sert donc de premier critère discriminatoire entre les différentes séquences candidates pour le spectre MS/MS. Généralement, seul le peptide qui a le meilleur score de corrélation est conservé pour la suite de l'évaluation de l'identification. Le score de corrélation est appelé différemment selon les moteurs (par exemple « Ion score » pour Mascot, « HyperScore » pour X!Tandem), calculé de manière propre à chaque programme et n'est pas toujours visible dans la fenêtre de résultats du programme.

Le processus de calcul du score de corrélation est différent selon les programmes. De plus, certains programmes comme Mascot ne rendent pas accessible leur modèle de calcul. Toutefois, nous pouvons constater que de manière générale, dans leur calcul de score de corrélation, les programmes tiennent compte :

- du nombre de fragments attribués
- du nombre de fragments non attribués
- du rapport (nombre de fragments attribués/ nombre de fragments non attribués)
- de l'intensité des fragments attribués et non attribués
- de la distribution des fragments sur toute la gamme de masse d'enregistrement
- des séries de fragments de même type consécutifs
- de l'erreur sur la masse des fragments
- du nombre d'acides aminés de la séquence identifiée

Prenons l'exemple le plus simple avec X!Tandem dont le score de corrélation (HyperScore) est calculé par la formule suivante :

$$\text{HyperScore} = \text{Score}_{by} \times N_y! \times N_b!$$

Avec

Score_{by} = somme des intensités des pics attribués pour les ions b- ou y-

$N_y!$ = nombre de fragments y attribués factoriel

$N_b!$ = nombre de fragments b attribués factoriel

2.1.1.2. Le score probabiliste

Le score de corrélation est ensuite converti en une mesure statistique, appelée « valeur E » (« Expectation value », « E-value ») qui correspond au nombre de peptides attendus avec un score supérieur ou égal au score observé sous l'hypothèse que ces peptides sont attribués au spectre expérimental par hasard. Cette « E-value » est calculée soit en considérant que la distribution des scores de corrélation suit une certaine distribution (par exemple une distribution de Poisson pour OMSSA) soit de manière empirique en observant la distribution de scores de corrélation de l'ensemble des différentes séquences candidates pour le spectre MS/MS (X!Tandem, Figure 3).

La présentation directe de la « E-value » comme valeur d'évaluation de l'identification peut être troublante parce que non seulement elle couvre une très large gamme de valeurs mais aussi parce qu'un score « élevé » correspond à une « faible » probabilité ce qui peut être ambigu. Pour ces raisons, et également par question d'habitude, la « E-value » est généralement convertie en score probabiliste. Cette conversion en score peut être réalisée de manière simple, par exemple en la convertissant en « $-\log(\text{E-value})$ », (X!Tandem ou OMSSA). Avec Mascot, la présentation de la valeur de score (score de corrélation) et de la valeur d'« identity score » (score identité) est préférée. L'« identity score » est un paramètre indicatif correspondant à la valeur de score pour laquelle l'identification a une « E value » de 0.05 (par défaut, mais il est possible de fixer la valeur d'« E-value » souhaitée et d'obtenir l'« identity score » correspondant). Avec Mascot, on a donc tendance à présenter le score probabiliste sous la forme d'un « difference score » qui correspond à la soustraction « score-identity score ». Plus cette différence est grande et plus la « E value » correspondante diminue. Par exemple, si cette différence est de 10, la « E value » est divisée par 10.

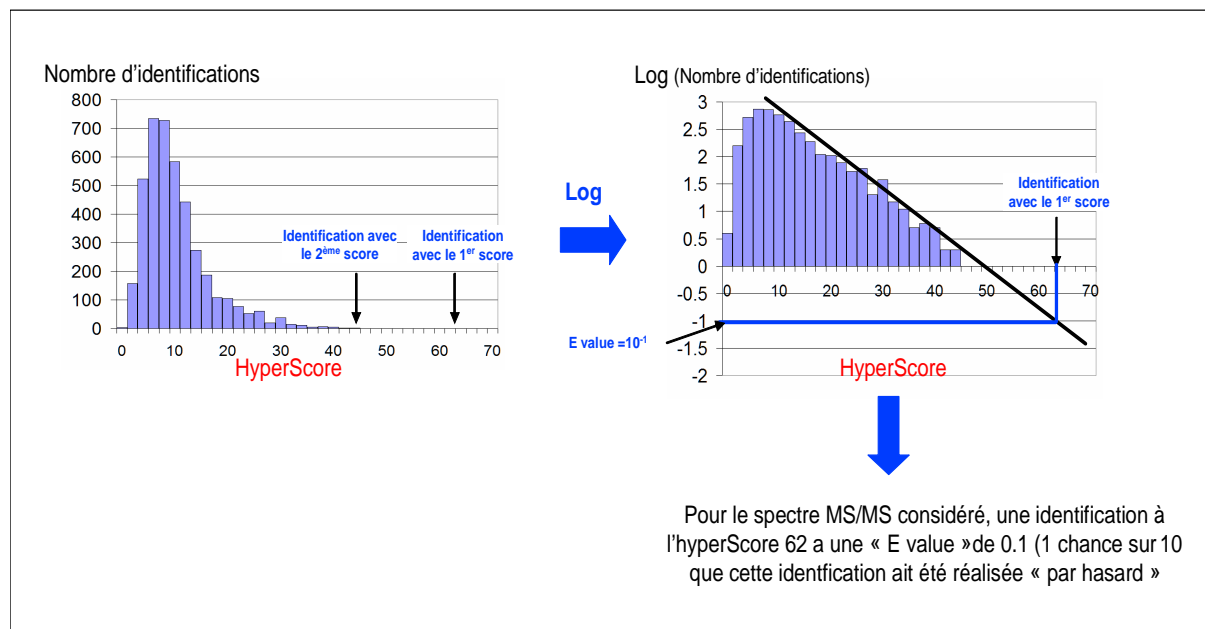
Le score probabiliste des moteurs de recherche contrairement au score de corrélation peut être comparé entre programmes. Les paramètres influençant cette valeur, hormis le score de corrélation, sont liés à l'espace de recherche dans lequel les identifications ont été effectuées, c'est-à-dire :

- La taille de la banque protéique.
- Les enzymes définies.
- La tolérance sur la masse du précurseur.
- Les modifications des acides aminés autorisées.

Par exemple, une recherche effectuée dans une très grande banque protéique (NCBI sans restriction taxonomique) avec une enzyme aspécifique et en tolérant une grande erreur sur la masse des précurseurs va conduire à des scores probabilistes beaucoup plus mauvais (« E value » très grande) que si cette recherche est effectuée dans un espace de recherche très réduit.

Il faut noter que même si le score probabiliste pour une identification paraît très bon, cela n'assure pas la fiabilité complète de l'identification. Des stratégies d'évaluation des identifications ont été développées et seront exposées au cours de ce chapitre. Pour la suite, la dénomination « score » correspondra au score probabiliste.

Figure 3 : Illustration du calcul d'une « E value » X!Tandem pour une identification à partir de la distribution des scores de corrélation de l'ensemble des différentes séquences candidates pour le spectre MS/MS



2.1.2. Comparaison et combinaison de moteurs de recherche

En général, avec l'approche par recherche dans les banques protéiques, seule une fraction de l'ensemble des spectres MS/MS acquis conduit à l'identification de peptides. Il existe plusieurs raisons pour expliquer cela. Elles peuvent être liées à l'échantillon : les spectres peuvent être de très mauvaise qualité, peuvent correspondre à des molécules de l'analyse qui ne sont pas des peptides, ou encore peuvent correspondre à des peptides modifiés ou qui ne sont pas dans les banques. Toutefois, cette importante fraction de spectres non identifiés peut également s'expliquer en partie par un problème lié au moteur de recherche utilisé pour l'identification de certains peptides. Comme chaque moteur de recherche utilise des méthodes différentes pour établir les scores (de corrélation et probabilistes) des identifications, les résultats obtenus diffèrent au moins en partie entre les différents programmes

[Searle et al., 2008]. Par conséquent, l'utilisation combinée de plusieurs moteurs de recherche devrait être utile pour fournir des résultats complémentaires et une validation croisée des résultats.

2.1.2.1. Comparaison de moteurs de recherche

Quelques études ont été réalisées sur la comparaison de différents moteurs de recherche (OMSSA, Mascot, Sequest, X!Tandem) [Bakalarski et al., 2007; Balgley et al., 2007] et ont permis d'observer que les programmes présentaient des performances différentes dépendantes du type de données utilisées (justesse de mesure de masse sur les précurseurs, modifications post-traductionnelles labiles considérées) qui découle de l'échantillon et du spectromètre de masse. La comparaison des résultats a pu être réalisée en utilisant une stratégie « target-decoy » (qui sera détaillée dans la suite du chapitre). Bien qu'il soit difficile d'établir des règles à partir d'un si petit nombre d'études, quelques tendances semblent se dégager :

- ❖ Sur des données MS/MS pour lesquelles la tolérance en masse sur le précurseur est élevée (2-3 Da, données acquises sur une trappe ThermoFinnigan LTQ), le moteur OMSSA a de meilleures performances en terme de sensibilité par rapport aux autres moteurs avec la même précision (~+30 % de peptides identifiés).
- ❖ Sur des données MS/MS pour lesquelles la tolérance en masse sur le précurseur est faible (< 10ppm, données acquises sur un instrument LTQ-FT ThermoFinnigan), les moteurs Mascot et Sequest permettent d'identifier un nombre légèrement plus important de peptides que les moteurs OMSSA et X!Tandem (~+10-20 %).
- ❖ Sur des données MS/MS de phosphopeptides (qui présentent généralement des pertes de neutres majoritaires dans les spectres de fragmentation qui nuisent à l'identification) acquises sur LTQ, Mascot, Sequest et OMSSA ont des performances supérieures à X!Tandem (~+50 %).
- ❖ Sur des données MS/MS de phosphopeptides acquises sur LTQ-FT, Mascot et Sequest ont des performances supérieures à X!Tandem et OMSSA (~+50 %).

Ces résultats soulignent que le choix du moteur utilisé pour réaliser les identifications pourrait être guidé par l'instrument utilisé et la nature des peptides étudiés. Toutefois, globalement, quelque soient les données MS/MS de départ, chaque moteur va pouvoir fournir des données d'identification qui lui sont uniques. Par conséquent, l'utilisation combinée de plusieurs moteurs de recherche semblent être une stratégie prometteuse pour une identification la plus exhaustive possible des peptides d'un échantillon.

2.1.2.2. Utilisation combinée de plusieurs moteurs de recherche

L'utilisation combinée de plusieurs moteurs de recherche pour bénéficier de leurs résultats complémentaires est une approche en développement. Plusieurs études ont rapporté que l'utilisation de plusieurs moteurs de recherche conduisait à l'identification d'un nombre plus important de peptides et par conséquent à l'identification de plus de protéines et à des identifications de protéines plus fiables [Resing et al., 2004; Elias et al., 2005; Kapp et al., 2005; Price et al., 2007]. De plus, les identifications identiques réalisées avec plusieurs moteurs de recherche peuvent également être considérées comme fiables même si les scores correspondants sont faibles (validation croisée) [Kapp et al., 2005].

Toutefois, pour pouvoir pleinement bénéficier des avantages de cette approche, il est nécessaire de pouvoir combiner les scores des identifications par les différents moteurs pour obtenir un score probabiliste final. Des procédures de calcul de ce score d'identification final ont été établies dans plusieurs études [Alves et al., 2008; Searle et al., 2008; Jones et al., 2009]. La procédure développée par l'équipe de Brian C. Searle [Searle et al., 2008] est intégrée dans le logiciel de gestion des données et des résultats d'analyses protéomiques Scaffold (<http://www.proteomesoftware.com>) et permet une gestion fiable et aisée des résultats d'identification protéomique issus de l'utilisation de plusieurs moteurs de recherche.

2.1.3. Evaluation des identifications

Comme nous l'avons vu précédemment, les moteurs de recherche fournissent une liste de correspondance « spectre MS/MS – peptide » avec un score probabiliste d'identification (typiquement une séquence peptidique candidate pour chaque spectre MS/MS). A ce stade, il convient d'évaluer la confiance à apporter aux résultats car certaines identifications sont vraies et d'autres fausses. Plusieurs méthodes d'évaluation existent pour juger de la fiabilité des identifications.

Au début de la protéomique, il était habituel de générer une liste d'identification « très sûre » en appliquant un seuil minimal que devait dépasser le score de l'identification pour qu'elle soit « validée ». Le plus souvent, ce seuil de score était accompagné d'une inspection manuelle de l'attribution des ions fragments sur le spectre MS/MS par un expert.

Cependant, les distributions des scores générés par un moteur de recherche dépendent de beaucoup de facteurs comme la performance du spectromètre de masse, la qualité des spectres MS/MS, l'espace de recherche pour l'identification. L'application d'un seuil de score identique sur des expériences totalement différentes n'est donc pas justifiée. L'inspection manuelle par l'expert n'apparaît pas non plus comme la solution adéquate puisque cette procédure prend énormément de temps et n'est pas compatible avec les études protéomiques à grande échelle. De plus, cette procédure est très subjective et dépend du niveau d'expertise de l'expert en question (expérience, connaissance des mécanismes de fragmentation). La protéomique moderne s'est donc progressivement orientée vers des approches probabilistes pour l'évaluation des identifications. Ces approches sont de deux types : i) les stratégies « target-decoy » et ii) les approches empiriques de Bayes.

2.1.3.1. Les stratégies « target-decoy »

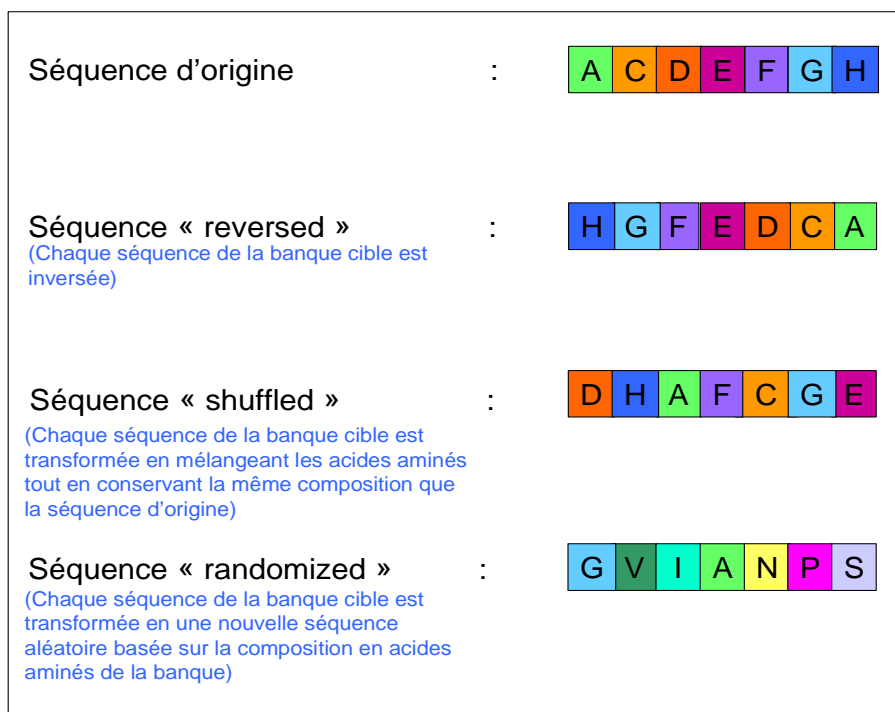
Ces stratégies consistent à réaliser les interrogations par les moteurs de recherche dans des banques protéiques « target-decoy » et à calculer un score seuil optimal pour chaque ensemble de données MS/MS.

2.1.3.1.1. Principes et mise en place

La stratégie « target-decoy » dont les premiers principes sont apparus en 2002 [Moore et al., 2002] et qui fut développée davantage par l'équipe de Steven P. Gygi [Peng et al., 2003; Elias et al., 2005; Elias et al., 2007] est réalisée en plusieurs étapes :

- ❖ Une banque protéique « target-decoy » est créée tout d'abord en obtenant l'ensemble des séquences de la banque protéique appropriée pour l'échantillon analysé (banque cible, « target »). Ensuite, une banque « leurre » ou « decoy » est créée en préservant la composition en acides aminés de la banque target et en limitant le nombre de peptides en commun avec la banque target. La banque target-decoy est obtenue en compilant les deux banques. Les différentes manières de créer les banques decoy sont illustrées en Figure 4.

Figure 4 : Illustration des différents modes de création des banques decoy



- ❖ Les identifications par le moteur de recherche sont réalisées dans la banque target-decoy. L'approche suppose que les identifications peptidiques dans la composante decoy de la banque et les identifications « fausses » dans la composante target suivent la même

distribution. Cette hypothèse est vérifiée quand la banque decoy est construite avec des séquences « reversed » mais nécessite l'application de facteurs correctifs pour les autres types de banques decoy [Elias et al., 2007].

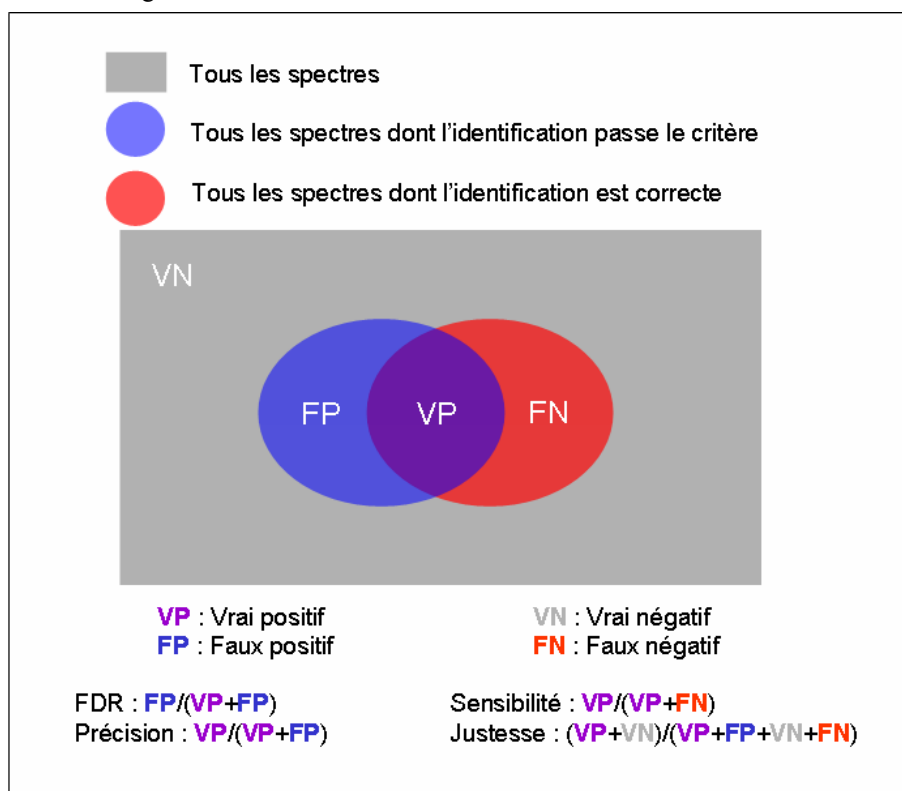
- ❖ Les identifications sont filtrées en utilisant différents seuils de score et le taux de faux-positifs (FDR) associé à chaque seuil de score est évalué par le calcul $2 N_d/N$ où N est le nombre d'identifications peptidiques dépassant le seuil de score dans la banque target-decoy et N_d le nombre d'identifications peptidiques dépassant le seuil de score dans la composante decoy de la banque.

La stratégie target-decoy ne permet pas de décider exactement quelles identifications sont correctes ou incorrectes mais elle évalue le taux de faux-positifs (FDR) d'un grand ensemble d'identifications. Cette stratégie permet d'estimer la probabilité qu'une identification soit correcte en considérant qu'elle fait partie d'un ensemble d'identifications avec un FDR mesuré.

2.1.3.1.2. Etablissement des seuils

L'établissement des seuils de score est réglé en fonction des objectifs de l'étude par la mesure de différents indices associés aux résultats et découlant directement des seuils. Ces indices de mesure (présentés et illustrés en Figure 5) sont le FDR, la précision (directement liée au FDR car la précision est égale à 1-FDR), mais aussi la sensibilité ou la justesse des identifications [Elias et al., 2007].

Figure 5 : Les indices de mesures associés aux identifications



Dans l'idéal, les seuils doivent être fixés pour maximiser la justesse, la sensibilité et la précision des identifications. Toutefois, la précision et la sensibilité des identifications ont tendance à évoluer dans des sens opposés.

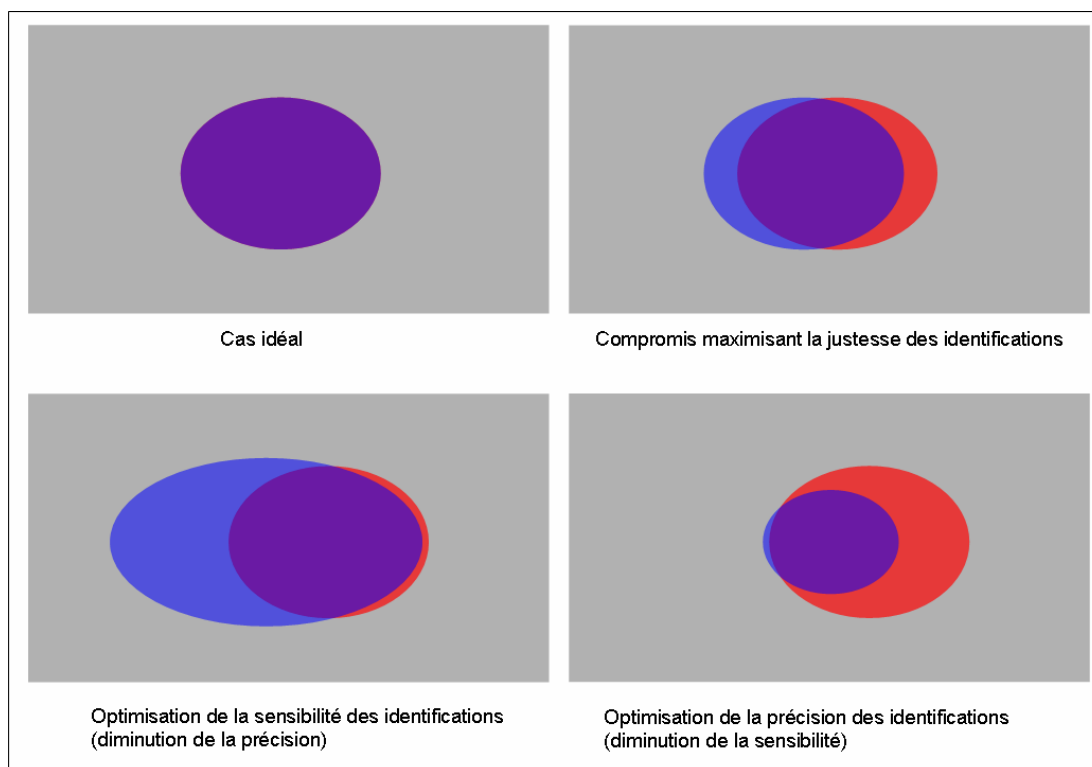
Ainsi, un compromis va consister à obtenir la meilleure justesse des identifications, ce qui nécessite de fixer des seuils moyens qui à la fois maximisent la sensibilité et la précision.

Par contre, s'il est nécessaire que l'étude soit la plus exhaustive possible en terme d'identification, les seuils de score seront fixés de manière moins stricte, ce qui aura pour effet d'augmenter le FDR et la sensibilité des identifications tout en diminuant la précision et la justesse de celles-ci. Ce choix est adapté par exemple pour des études de type exploratoire où les données protéomiques peuvent notamment être corrélées avec des données issues d'autres méthodes d'analyse.

S'il est nécessaire dans l'étude que chaque identification « validée » soit correcte et cela sans aucune ambiguïté, les seuils de score seront fixés de manière plus stricte, ce qui aura pour effet d'augmenter la précision des identifications tout en diminuant le FDR, la sensibilité et la justesse de celles-ci. Ce choix est adapté par exemple pour des études où seules des données protéomiques sont accessibles ou pour des études où l'identification d'une protéine incorrecte peut conduire à des interprétations biologiques erronées.

Les différents scénarios sont présentés en Figure 6.

Figure 6 : Illustration des différents scénarios possibles selon les valeurs de seuils choisies



2.1.3.1.3. Des informations auxiliaires pour améliorer le processus d'évaluation des identifications

Les stratégies target decoy pour l'évaluation des identifications sont basées sur les résultats des algorithmes de recherche. Toutefois, pour affiner l'évaluation, il est possible de tirer partie d'autres informations générées au cours des analyses protéomiques qui peuvent être les suivantes :

- **Informations de fragmentation.** Par exemple la plateforme de validation « EPIR » [Kristensen et al., 2004] va tirer partie de l'observation de coupures préférentielles connues (par exemple la « coupure proline ») ou encore de l'identification de séries de fragments consécutifs pour renforcer les identifications réalisées par le moteur de recherche Mascot. Il est également possible de réaliser une étape supplémentaire MS³ consécutive à la MS/MS et de calculer un score d'identification issu des deux types de fragmentation [Olsen et al., 2004].
- **Temps de rétention** en LC associés aux peptides identifiés. Des programmes de prédiction des temps de rétention des peptides éduqués à partir d'un ensemble de peptides modèles ont été développés. L'adéquation des temps de rétention prédits avec les temps de rétention expérimentaux permet de renforcer les identifications [Strittmatter et al., 2004; Qian et al., 2005; Pfeifer et al., 2009].
- **pI** associés aux peptides identifiés. Si une séparation électrophorétique des peptides a été réalisée en amont de l'analyse LC-MS/MS, il est également possible de vérifier que la séquence du peptide identifié et donc son pI théorique est en accord avec le pI observé expérimentalement [Cargile et al., 2004; Cargile et al., 2004; Malmstrom et al., 2006; Xie et al., 2006].
- **Coupures enzymatiques manquées ou justesse de la masse mesurée du précurseur.** On peut considérer que les séquences peptidiques identifiées comportant plusieurs sites de coupures manquées ou présentant une erreur importante sur leur masse doivent demander plus d'attention pour être validées (par exemple un seuil de score d'identification plus stringent) [Elias et al., 2007]
- **Contexte de l'étude.** La présence de certains acides aminés ou certains motifs dans la séquence peut être fortement attendue comme la présence du motif N-X-S/T pour des peptides N-glycosylés [Zhang et al., 2005].

Toutes ces informations peuvent être utilisées pour pénaliser ou renforcer l'identification réalisée par le moteur de recherche voire même pour pondérer le score du moteur de recherche et utiliser le nouveau score dans la stratégie target-decoy. Pour ce faire, un logiciel « open-source » comme « peptizer » développé par l'équipe de Joel Vandekerckhove [Helsens et al., 2008] est très intéressant. En effet, il permet de réaliser le traitement à posteriori des identifications réalisées par le moteur de recherche de manière très flexible car étant « adaptable » aux besoins de chacun. Tout utilisateur peut ainsi construire ses propres « filtres » d'aide à l'évaluation des identifications adaptés à l'étude protéomique en question et les intégrer dans une stratégie de type target-decoy.

2.1.3.1.4. Dédution de l'identification des protéines (et de la confiance associée) à partir des identifications peptidiques

Dans la plupart des études protéomiques, le but final n'est pas d'identifier les peptides mais les protéines présentes dans l'échantillon. Les différentes séquences peptidiques identifiées sont donc regroupées en fonction des protéines auxquelles elles appartiennent et la confiance accordée à l'identification doit être ré-évaluée.

Le regroupement des peptides en protéines est un processus délicat notamment parce que de nombreux peptides sont partagés par plusieurs protéines. Généralement, le regroupement des peptides en protéines est réalisé en déterminant le plus petit ensemble de protéines qui peuvent expliquer l'ensemble des peptides identifiés [Nesvizhskii et al., 2005].

L'estimation de la confiance dans les identifications au niveau protéique peut être réalisée de différentes manières [Nesvizhskii et al., 2007] et notamment par une stratégie target-decoy. Dans ce cas, pour l'identification d'une protéine, on peut considérer des critères sur l'identification des peptides associés à un nombre minimum de peptides passant les critères. Par exemple, une stratégie serait de fixer un FDR souhaité pour l'identification des protéines et d'en déduire :

- Un ensemble de critères A que doit satisfaire le peptide s'il est seul à participer à l'identification de la protéine.
- Un ensemble de critères B que doivent satisfaire les peptides s'ils sont plusieurs à participer à l'identification des protéines.

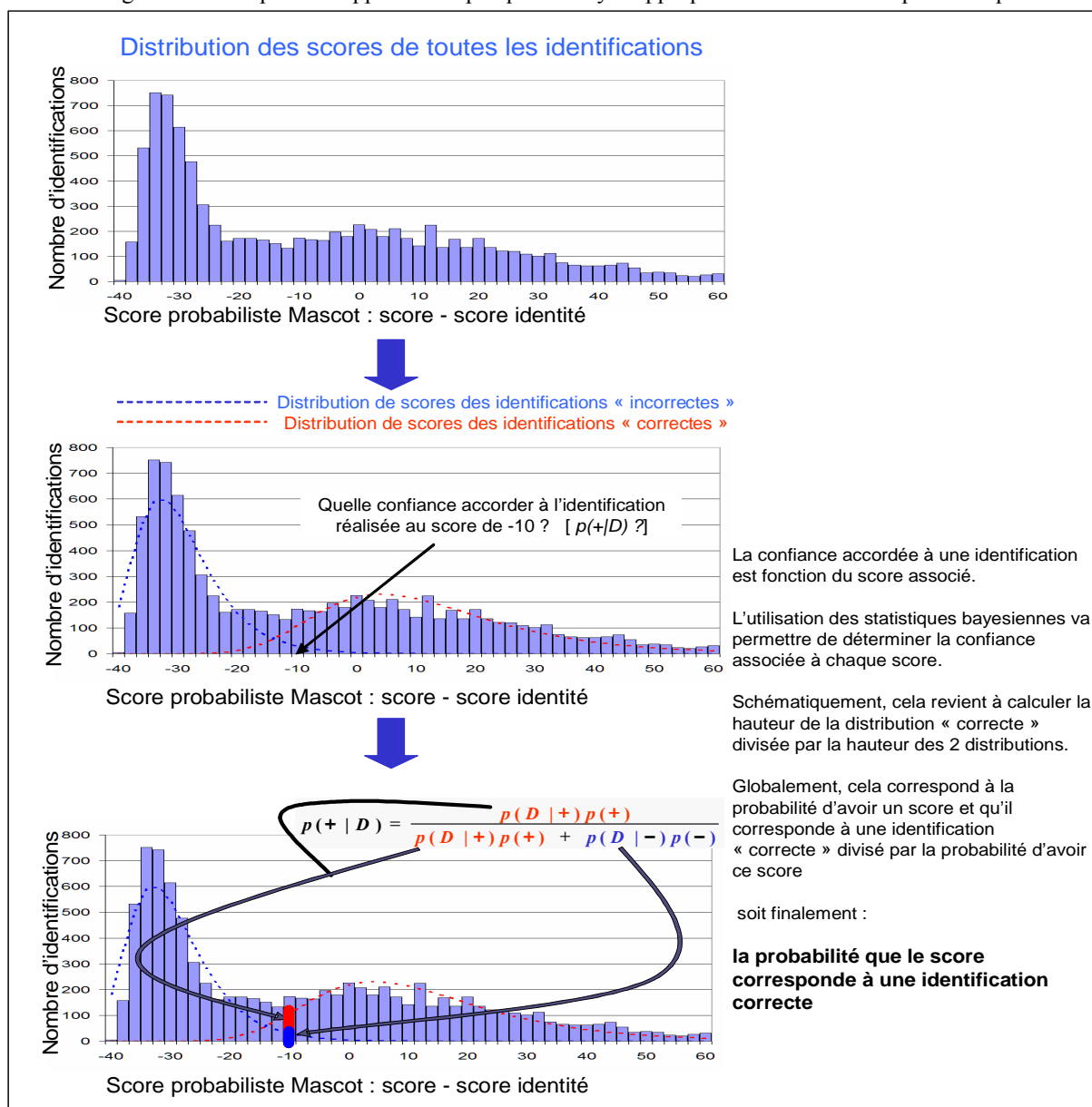
(B étant moins stringent que A)

Les stratégies target-decoy présentent l'avantage d'être simples à implémenter dans une stratégie d'analyse et faciles à utiliser dans de nombreuses situations. L'inconvénient est que le temps d'interrogation du moteur de recherche dans les banques protéiques est doublé.

2.1.3.2. Les approches empiriques de Bayes

Ce type d'approche consiste à déterminer ce que l'on appellerait actuellement la distribution *a posteriori* de la probabilité *p* d'une loi binomiale [Bayes et al., 1763]. En protéomique, l'application de cette approche est réalisée par le logiciel « PeptideProphet » [Keller et al., 2002] qui permet de modéliser la distribution des scores (probabilistes) des identifications peptidiques (et des informations auxiliaires comme vu en 2.1.3.1.3.) observées dans toute l'étude protéomique. Cette distribution est modélisée comme une bimodale dont l'une des composantes est issue de la distribution des scores des identifications peptidiques correctes et l'autre composante est issue de la distribution des scores des identifications peptidiques incorrectes. Cette modélisation va permettre de déduire des probabilités *a posteriori*, en utilisant les statistiques bayésiennes, sur les identifications en fonction du score qui leur est associé (Figure 7).

Figure 7 : Principes de l'approche empirique de Bayes appliquée à l'identification protéomique



De la même manière que pour les stratégies target-decoy, après avoir évaluée la confiance accordée aux identifications peptidiques, il faut finalement étendre l'évaluation au niveau protéique. Celle-ci peut être réalisée là aussi par les approches empiriques de Bayes par exemple par l'utilisation du programme « ProteinProphet » qui combine les probabilités de justesse d'identification des différents peptides participant à l'identification de la même protéine (« peptides frères » ou « sibling peptides ») pour calculer la probabilité que l'identification de la protéine soit juste (Figure 8). De plus, le programme va aussi recalculer la confiance accordée aux identifications peptidiques en tenant compte de l'existence des « peptides frères » [Nesvizhskii et al., 2003].

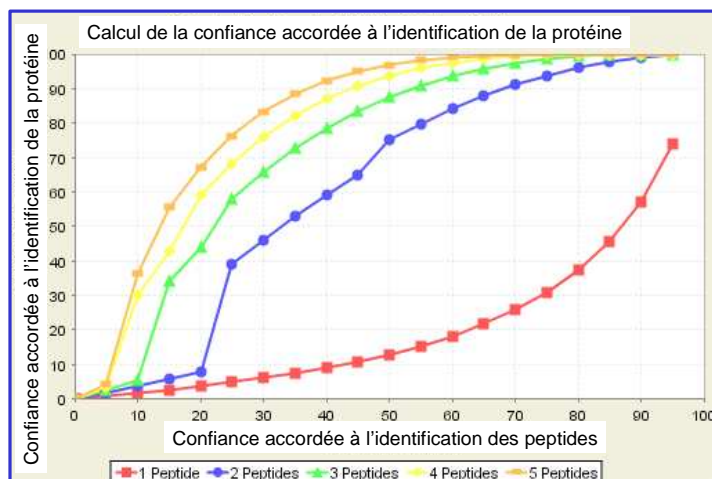


Figure 8 : Evaluation de la confiance accordée à l'identification des protéines en fonction de la confiance accordée à l'identification des peptides par un algorithme de type ProteinProphet. L'ensemble des données utilisées est issu de l'étude présentée dans le Chapitre 2, Partie III des résultats.

2.1.3.3. Les approches alternatives

Les deux types d'approches probabilistes pour l'évaluation des identifications (stratégie target-decoy et approche empirique de Bayes) sont utilisées de manière très courante maintenant en protéomique. Toutefois, il faut garder à l'esprit que pour que ces stratégies d'évaluation soient significatives, il faut qu'elles portent sur un très grand nombre de données. Or, certaines études protéomiques ne nécessitent pas l'analyse de milliers de protéines mais portent sur une centaine ou une dizaine de protéines voire moins.

Pour ces cas de figure, il est toujours possible par exemple d'appliquer une stratégie target-decoy. Toutefois, l'évaluation des identifications sera beaucoup moins précise [Elias et al., 2007] mais tout de même informative.

Pour ces études, la stratégie d'inspection manuelle des spectres reste une alternative très intéressante [Chen et al., 2005] malgré les critiques qui lui sont couramment adressées (vu en Chapitre 3. 2.1.3.). L'évaluation des identifications peptidiques sera très dépendante de l'expert qui la réalisera mais on peut tout de même souligner les critères les plus généralement considérés par l'évaluation manuelle :

- Présence d'une suite continue d'ions fragments y- ou b-.
- Présence d'ions fragments intenses qui ne sont pas de type y- ou b-.
- Présence d'ions intenses non identifiés.
- Pour les séquences contenant une proline, le pic le plus intense correspond à la coupure proline (ou autres coupures préférentielles explicables comme vu en Chapitre 2. 1.3.3).
- Intensité et état de charge des fragments explicables par le contenu en acides aminés basiques du peptide et son état de charge.
- Importance de l'erreur sur la masse du précurseur.
- Nombre important de modifications sur le peptide.
- Présence d'un bruit de fond important sur le spectres.

(et d'autres...)

De la même manière que pour les stratégies automatiques, le nombre de peptides participant à l'identification d'une protéine sera également pris en compte. Par exemple, si l'étude protéomique porte sur l'analyse de spots de gel 2D et sur un organisme dont le génome est séquencé (pas une approche de type « recherche inter-espèce »), on pourra s'attendre à observer l'identification d'au moins 2 ou 3 peptides de la protéine. Si le mélange analysé est un peu plus complexe ou très peu abondant, il faudra sans doute considérer les identifications protéiques réalisées avec un seul peptide. Pourtant, de nombreuses études éliminent systématiquement les identifications protéiques à un seul peptide. S'il est nécessaire de traiter prudemment ce type d'identification, leur élimination systématique n'est pas justifiée puisque d'après les études de l'équipe de Pavel Pevzner, l'identification de protéines avec deux peptides peut dans certains cas (selon les critères associés) être moins fiable que l'identification à un seul peptide [Gupta et al., 2009]. Cette équipe a également montré que 80 % des protéines identifiées avec un seul peptide étaient réellement exprimées dans leur étude et que leur élimination conduisait à la perte de 20-25 % de toutes les protéines identifiées [Gupta et al., 2007].

Enfin, une autre stratégie peut encore être utilisée pour l'évaluation des identifications peptidiques qui ne peuvent pas être traitées par les méthodes automatiques ou qui nécessitent encore un degré de confiance supplémentaire. Cette stratégie va consister à synthétiser un peptide de référence reflétant la séquence candidate à l'identification d'un spectre. Ce peptide de référence est ensuite analysé dans les conditions les plus proches possibles de l'analyse qui a généré le spectre dont l'identification est à évaluer. Enfin, les deux spectres sont ensuite comparés avec un programme de comparaison de spectres (tel que décrit en Chapitre 3. 2.3.1.) qui va évaluer le degré de similarités des deux spectres par l'estimation d'un coefficient de corrélation spectrale. Cette stratégie très lourde à mettre en place et très coûteuse ne sera utilisée que très ponctuellement. Par exemple, elle fut utilisée dans l'étude protéomique qui a alimenté l'actualité scientifique, et plus particulièrement protéomique, en 2007-2008 (et même encore maintenant) et qui a permis d'identifier des séquences protéomiques de Mammouth et de Tyrannosaure Rex [Asara et al., 2007]. Cette stratégie a été appliquée sur des peptides dont l'identification était la plus douteuse dans un premier temps puis étendue à d'autres peptides dans un deuxième temps pour répondre au scepticisme de la communauté scientifique qui a considéré, sans doute à raison, que « la science extraordinaire nécessite des preuves extraordinaires » [Buckley et al., 2008; Pevzner et al., 2008].

2.1.3.4. Importance de l'évaluation des identifications en analyse protéomique

Il est essentiel pour la protéomique de toujours associer une évaluation de la confiance accordée aux résultats d'identifications. Cela est indispensable pour pouvoir évaluer les conclusions biologiques (ou même méthodologiques) d'une étude et comparer les résultats de différentes études. Si ce n'est pas le cas, des informations erronées risquent de se propager dans la littérature et dans les

bases de données qui contiennent les données protéomiques. C'est pourquoi, les journaux spécialisés dans le domaine comme « Molecular and cellular Proteomics » ou « Proteomics » ont établi des directives sur les informations et la documentation à associer à l'analyse et l'identification des peptides et des protéines qui doivent être suivies pour pouvoir prétendre à publier dans ces journaux. L'objectif des directives est de formuler des critères standards pour que les éditeurs, les « reviewers » et les lecteurs puissent évaluer plus aisément les données issues de l'analyse protéomique.

Toutes les mesures et les outils décrits précédemment pour l'évaluation des identifications assurent une certaine qualité et un contrôle évident des études protéomiques. Toutefois, en biologie, il est dans l'absolu préféré d'établir un faisceau d'évidence à partir de méthodes d'analyse différentes qui se confirment mutuellement. Les identifications protéomiques qui présentent le plus d'intérêt dans le contexte d'une étude pourront donc être complétées par des méthodes d'analyse biochimique par exemple (Western blot, ELISA, etc...) pour apporter une « validation définitive » de l'identification d'une protéine.

2.1.4. Les limitations de l'approche par recherche dans les banques protéiques

L'approche par recherche dans les banques protéiques est une stratégie non tolérante aux erreurs. Avec cette approche, pour réaliser l'identification d'un peptide analysé, il est indispensable que celui-ci soit présent dans la banque de données interrogée. Ce type de stratégie n'est donc pas adapté si les protéines analysées ne sont pas dans les banques ou si elles n'ont pas d'homologue à très forte similarité dans les banques. Dans ces cas défavorables, on préférera l'approche alternative utilisant le séquençage *de novo* des peptides fragmentés pour identifier les protéines de l'échantillon.

2.2. L'approche par séquençage *de novo*

Dans cette approche, la séquence en acides aminés des peptides est directement « lue » sur les spectres MS/MS. A l'origine, cette approche était réalisée manuellement en tenant compte des règles de fragmentation des peptides (vu en Chapitre 2. 1.3.3). Plus récemment, une série d'outils a été développée pour assister le séquençage *de novo*. Parmi les plus connus on trouve Lutefisk [Johnson et al., 2002], Pepnovo [Frank et al., 2005] ou encore Peaks [Ma et al., 2003]. Seul ce dernier logiciel a été utilisé dans les travaux présentés dans cette thèse.

L'approche *de novo* manuelle ou assistée par les logiciels permet donc d'obtenir des fragments (« tags ») de séquences en acides aminés. Généralement, ceux-ci sont ensuite soumis à des programmes de recherche de similarités de séquence (par exemple BLAST) pour obtenir l'identité des protéines correspondantes. Comme l'algorithme BLAST n'est pas prévu pour fonctionner avec des

« petits tags » de séquence [Altschul et al., 1996; Pearson et al., 1997], une adaptation du programme a du être réalisée pour donner naissance au programme « MS-BLAST » [Shevchenko et al., 2001].

Le principal avantage de l'approche par séquençage *de novo* sur l'approche par recherche dans les banques protéiques est qu'elle permet l'identification de spectres pour lesquelles la séquence peptidique exacte correspondante n'est pas présente dans les banques de données. Ainsi, avec cette approche, on pourra identifier des peptides modifiés (mutation, inversion, délétion, insertion d'acides aminés) ou provenant de protéines d'organismes dont le génome n'a pas été séquencé par exemple.

Toutefois, cette approche est assez lourde à mettre en œuvre, surtout pour une étude protéomique à grande échelle, et requiert des spectres MS/MS de grande qualité. Cette approche sera généralement utilisée en complément d'une approche par recherche dans les banques protéiques qui n'aurait pas donnée des résultats satisfaisants.

2.3. Les approches alternatives

Les deux approches présentées précédemment, et particulièrement l'approche par recherche dans les banques protéiques, sont les approches les plus courantes pour l'identification des peptides et des protéines dans une analyse protéomique. Toutefois, on peut mentionner deux approches alternatives qui seront décrites très brièvement par la suite:

- ✓ L'approche par recherche dans les banques de spectres
- ✓ L'approche hybride

2.3.1. L'approche par recherche dans les banques de spectres

Cette approche très récente en protéomique tire partie du fait que les études protéomiques appliquées au même organisme conduisent à la répétition de l'identification de certains peptides et que de grandes quantités de données d'identifications protéomiques commencent à être disponibles dans des banques de données spécialisées comme PRIDE [Martens et al., 2005] ou PeptideAtlas [Desiere et al., 2005] par exemple. Par conséquent, une nouvelle étude protéomique sur un organisme déjà étudié peut conduire à l'identification de peptides qui avaient déjà été détectés auparavant, ce qui est particulièrement le cas sur les organismes modèles (homme, rat, etc..).

Dans cette approche, les spectres de fragmentation expérimentaux sont comparés à une librairie de spectres constituée d'un très grand nombre de spectres MS/MS de peptides ayant déjà été correctement identifiés. Le corrélation entre les spectres expérimentaux et les spectres de la librairie est réalisée par des outils tels que SpectraST [Lam et al., 2007], X! P3 [Craig et al., 2005] ou Biblispec [Frewen et al., 2006].

Cette approche est très performante en termes de vitesse, spécificité et sensibilité [Lam et al., 2007]. Toutefois, les librairies de spectres restent limitées à quelques organismes et cette approche

sera obligatoirement complétée par une approche plus classique de recherche dans les banques protéiques pour être assuré de ne pas avoir raté l'identification de peptides non détectés jusque là.

2.3.2. L'approche hybride

Cette approche combine des aspects de l'approche par recherche dans les banques protéiques et de l'approche par séquençage *de novo*. Le processus d'identification commence par l'établissement de petits « tags » de séquence à partir des spectres MS/MS, puis une recherche tolérante aux erreurs dans les banques protéiques est réalisée. Les outils permettant de réaliser une telle approche sont Peaks [Ma et al., 2003], Inspect [Tanner et al., 2005] ou encore GutenTag [Tabb et al., 2003].

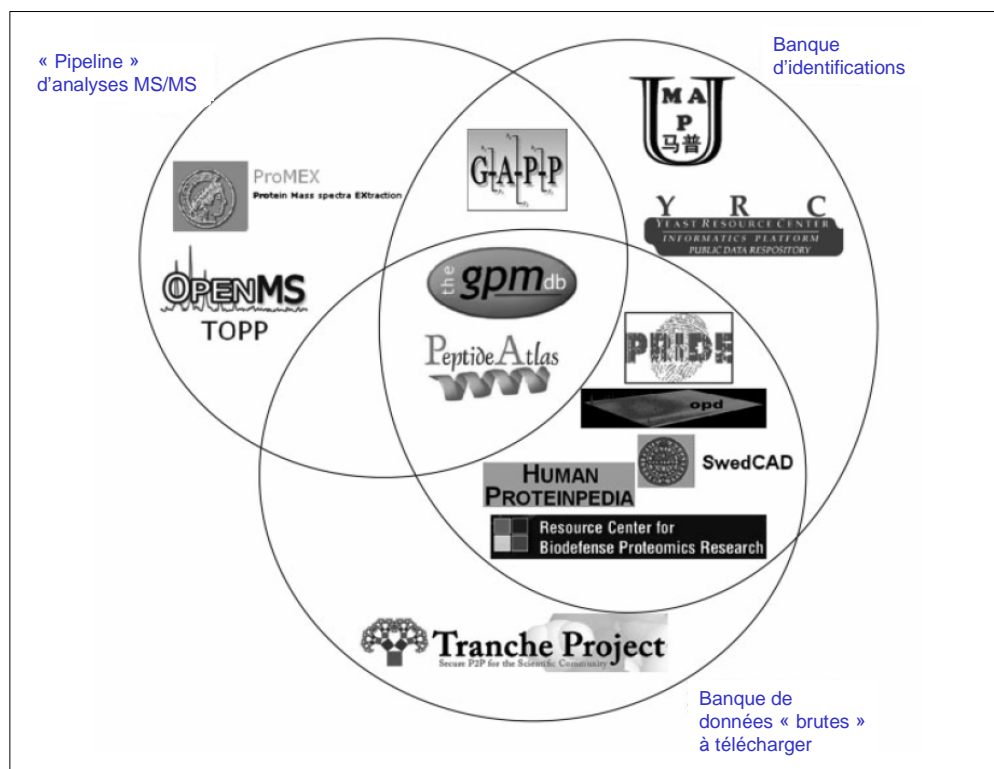
Cette approche permet une réduction du temps de recherche pour l'identification et une analyse systématique des peptides portant des modifications post-traductionnelles.

2.4. Les banques publiques de dépôt des données de protéomique

Le volume des données issues de l'analyse protéomique par spectrométrie de masse ayant explosé au cours des dernières années, des espaces de dépôt de ces données ont du être créés pour en faciliter l'accès à la communauté scientifique. Comme dans les autres domaines de l'« omique », l'accessibilité aux données apporte de nombreux bénéfices scientifiques comme la possibilité d'échanger, de comparer, de réanalyser les résultats, de développer des moteurs de recherche et des méthodologies d'analyses statistiques encore améliorés pour finalement obtenir des conclusions plus fiables et améliorer la connaissance biologique [Mead et al., 2007]. Des journaux comme ceux du groupe « Nature » préconisent également la soumission des données de protéomique à ces espaces de dépôt en complément des publications.

Il existe de multiples banques publiques de dépôt dont les objectifs et les fonctionnalités sont différentes puisque certaines banques de dépôt vont stocker les identifications peptidiques/protéomiques, certaines vont stocker les données d'analyses brutes et d'autres vont permettre de réanalyser les données (Figure 9).

Figure 9 : Panorama des principaux espaces de dépôts et « pipeline » publiques. D'après [Mead et al., 2009]



Ici, nous nous intéresserons seulement aux 3 espaces de dépôt fondateurs du consortium « ProteomExchange » et qui permettent de balayer les possibilités offertes par les espaces de dépôt.

Les 3 espaces de dépôt fondateurs du consortium « ProteomExchange » sont :

- PRIDE (<http://www.ebi.ac.uk/pride/>) [Jones et al., 2006]
- Tranche (<http://tranche.proteomecommons.org/>)
- PeptideAtlas (<http://www.peptideatlas.org>) [Desiere et al., 2006]

Ce consortium a été créé pour fournir un seul point d'accès pour la soumission de données de protéomique pour faciliter la tâche de l'utilisateur. Une fois soumises, les données peuvent être automatiquement distribuées à tous les espaces de dépôts du consortium.

Toutes les soumissions doivent contenir les 3 composantes principales suivantes :

- les données brutes des analyses par spectrométrie de masse
- les métadonnées liées aux analyses MS/MS
- les données d'identifications peptidiques et protéiques

Chaque espace de dépôt présentera une ou plusieurs parties de l'ensemble des données soumises :

- La banque **PRIDE** est axée sur le **stockage des identifications** et utilisée en support d'une publication. La présence des fichiers individuels MS ou MS/MS est optionnelle mais recommandée. Un lien permet de pointer sur les données brutes des analyses stockées dans l'espace de dépôt Tranche mais ces données ne sont pas stockées dans la banque PRIDE. Les métadonnées satisfaisant les directives « Minimal Information About a Proteomics Experiment » (« MIAPE ») pour la spectrométrie de masse [Taylor et al., 2007] sont requises.

- L'espace de dépôt **Tranche** peut stocker tout type de fichier lié à la protéomique. Le **stockage des données brutes** des analyses est recommandé.
- **PeptideAtlas** accepte seulement les fichiers de sortie des spectromètres de masse ainsi que les métadonnées associées mais pas les identifications. Les fichiers soumis sont **retraités en utilisant différentes stratégies de recherche** et le « Trans Proteomic Pipeline » (TPP, qui inclut les logiciels PeptideProphet et ProteinProphet notamment) [Keller et al., 2005].

Les trois espaces de dépôt de données de protéomique fondateurs du consortium « ProteomExchange » symbolisent les possibilités ouvertes aujourd'hui par ce type d'espace. Ces espaces de dépôt commencent à arriver à maturité et à jouer un rôle actif dans les projets de recherche.

Le contrôle de la qualité des données est soutenu notamment par le retraitement et l'évaluation des données dans « PeptideAtlas » mais aussi par la mise à disposition des fichiers de sortie d'analyse qui laisse la possibilité à tout utilisateur de retraiter les données selon sa propre méthodologie. Ces retraitements sont possibles aujourd'hui pour tout un chacun grâce à la mise en place de formats de fichiers standardisés, d'abord « mzData » et « mzXML » qui sont aujourd'hui en cours de remplacement par un format de fichier unique, le format « mzML » [Orchard et al., 2009].

Conclusion

Actuellement, les banques de séquences protéiques jouent un rôle vital de centre de ressource pour la biologie. Les protéines présentes dans les banques sont aujourd'hui quasi-exclusivement issues de prédictions automatiques *in silico* réalisées par des programmes bioinformatiques et peuvent présenter des erreurs. Pour faire face à ces erreurs très préjudiciables pour la biologie, les annotateurs ont amélioré leurs méthodes de prédiction et développé des outils de correction. Toutefois, ces annotations « affinées » ne restent que des prédictions dont l'efficacité est limitée. La seule manière d'obtenir des validations/corrections indiscutables des séquences protéiques est d'obtenir la séquence protéique expérimentale. Or, de nos jours, la technique de choix pour la caractérisation des protéines est l'analyse protéomique par spectrométrie de masse. L'analyse protéomique a subi un essor très important depuis 20 ans grâce au développement conjoint de la spectrométrie de masse, des techniques séparatives, des outils de valorisation des analyses et à la croissance exponentielle des banques protéiques. Banques protéiques et analyses protéomiques sont donc étroitement liées puisque le plus souvent l'étude du protéome repose sur la connaissance préalable du génome. Mais à l'inverse, l'analyse des protéines va pouvoir aider à l'analyse structurale du génome. Cette nouvelle approche d'aide à l'annotation génomique représente un champ d'application très récent dans le domaine protéomique et porte le nom de « protéogénomique ». La protéogénomique étant une approche assez récente elle nécessite encore des développements méthodologiques pour tirer le meilleur parti de son potentiel.

Au cours de ce travail de thèse, nous avons principalement développé une nouvelle approche d'aide à l'annotation génomique, basée sur les méthodologies protéomiques et notamment axée sur la détermination exacte des codons d'initiation des protéines, une des difficultés majeures dans les annotations génomiques. Ce travail a débouché sur d'autres applications de la méthodologie développée dans le domaine de la caractérisation de modifications post-traductionnelles (ici les phosphorylations) et dans le domaine de la quantification.

Bibliographie

- "The yeast genome directory." *Nature*, **1997**, 387 (6632 Suppl), 5.
- "Genome sequence of the nematode *C. elegans*: a platform for investigating biology." *Science*, **1998**, 282 (5396), 2012-8.
- "Analysis of the genome sequence of the flowering plant *Arabidopsis thaliana*." *Nature*, **2000**, 408 (6814), 796-815.
- "The UniProt Consortium"**
- "The Universal Protein Resource (UniProt) 2009." *Nucleic Acids Res*, **2009**, 37 (Database issue), D169-74.
- Adams, M. D., S. E. Celniker, R. A. Holt, C. A. Evans, J. D. Gocayne, P. G. Amanatides, S. E. Scherer, P. W. Li, R. A. Hoskins, R. F. Galle, R. A. George, S. E. Lewis, S. Richards, M. Ashburner, S. N. Henderson, G. G. Sutton, J. R. Wortman, M. D. Yandell, Q. Zhang, L. X. Chen, R. C. Brandon, Y. H. Rogers, R. G. Blazej, M. Champe, B. D. Pfeiffer, K. H. Wan, C. Doyle, E. G. Baxter, G. Helt, C. R. Nelson, G. L. Gabor, J. F. Abril, A. Agbayani, H. J. An, C. Andrews-Pfannkoch, D. Baldwin, R. M. Ballew, A. Basu, J. Baxendale, L. Bayraktaroglu, E. M. Beasley, K. Y. Beeson, P. V. Benos, B. P. Berman, D. Bhandari, S. Bolshakov, D. Borkova, M. R. Botchan, J. Bouck, P. Brokstein, P. Brottier, K. C. Burtis, D. A. Busam, H. Butler, E. Cadieu, A. Center, I. Chandra, J. M. Cherry, S. Cawley, C. Dahlke, L. B. Davenport, P. Davies, B. de Pablos, A. Delcher, Z. Deng, A. D. Mays, I. Dew, S. M. Dietz, K. Dodson, L. E. Doup, M. Downes, S. Dugan-Rocha, B. C. Dunkov, P. Dunn, K. J. Durbin, C. C. Evangelista, C. Ferraz, S. Ferreira, W. Fleischmann, C. Fosler, A. E. Gabrielian, N. S. Garg, W. M. Gelbart, K. Glasser, A. Glodek, F. Gong, J. H. Gorrell, Z. Gu, P. Guan, M. Harris, N. L. Harris, D. Harvey, T. J. Heiman, J. R. Hernandez, J. Houck, D. Hostin, K. A. Houston, T. J. Howland, M. H. Wei, C. Ibegwam, M. Jalali, F. Kalush, G. H. Karpen, Z. Ke, J. A. Kennison, K. A. Ketchum, B. E. Kimmel, C. D. Kodira, C. Kraft, S. Kravitz, D. Kulp, Z. Lai, P. Lasko, Y. Lei, A. A. Levitsky, J. Li, Z. Li, Y. Liang, X. Lin, X. Liu, B. Mattei, T. C. McIntosh, M. P. McLeod, D. McPherson, G. Merkulov, N. V. Milshina, C. Mobarry, J. Morris, A. Moshrefi, S. M. Mount, M. Moy, B. Murphy, L. Murphy, D. M. Muzny, D. L. Nelson, D. R. Nelson, K. A. Nelson, K. Nixon, D. R. Nusskern, J. M. Pacleb, M. Palazzolo, G. S. Pittman, S. Pan, J. Pollard, V. Puri, M. G. Reese, K. Reinert, K. Remington, R. D. Saunders, F. Scheeler, H. Shen, B. C. Shue, I. Siden-Kiamos, M. Simpson, M. P. Skupski, T. Smith, E. Spier, A. C. Spradling, M. Stapleton, R. Strong, E. Sun, R. Svirskas, C. Tector, R. Turner, E. Venter, A. H. Wang, X. Wang, Z. Y. Wang, D. A. Wassarman, G. M. Weinstock, J. Weissenbach, S. M. Williams, Woodage T, K. C. Worley, D. Wu, S. Yang, Q. A. Yao, J. Ye, R. F. Yeh, J. S. Zaveri, M. Zhan, G. Zhang, Q. Zhao, L. Zheng, X. H. Zheng, F. N. Zhong, W. Zhong, X. Zhou, S. Zhu, X. Zhu, H. O. Smith, R. A. Gibbs, E. W. Myers, G. M. Rubin and J. C. Venter**
- "The genome sequence of *Drosophila melanogaster*." *Science*, **2000**, 287 (5461), 2185-95.
- Adams, M. D., J. M. Kelley, J. D. Gocayne, M. Dubnick, M. H. Polymeropoulos, H. Xiao, C. R. Merril, A. Wu, B. Olde, R. F. Moreno and et al.**
- "Complementary DNA sequencing: expressed sequence tags and human genome project." *Science*, **1991**, 252 (5013), 1651-6.
- Altschul, S. F. and W. Gish**
- "Local alignment statistics." *Methods Enzymol*, **1996**, 266 460-80.
- Alves, G., W. W. Wu, G. Wang, R. F. Shen and Y. K. Yu**
- "Enhancing peptide identification confidence by combining search methods." *J Proteome Res*, **2008**, 7 (8), 3102-13.
- Apweiler, R., A. Bairoch and C. H. Wu**
- "Protein sequence databases." *Curr Opin Chem Biol*, **2004**, 8 (1), 76-80.
- Apweiler, R., A. Bairoch, C. H. Wu, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, D. A. Natale, C. O'Donovan, N. Redaschi and L. S. Yeh**
- "UniProt: the Universal Protein knowledgebase." *Nucleic Acids Res*, **2004**, 32 (Database issue), D115-9.
- Arnaud, I. L., J. Jossierand, J. S. Rossier and H. H. Girault**
- "Finite element simulation of Off-Gel trade mark buffering." *Electrophoresis*, **2002**, 23 (19), 3253-61.
- Asara, J. M., M. H. Schweitzer, L. M. Freemark, M. Phillips and L. C. Cantley**
- "Protein sequences from mastodon and *Tyrannosaurus rex* revealed by mass spectrometry." *Science*, **2007**, 316 (5822), 280-5.
- Badger, J. H. and G. J. Olsen**
- "CRITICA: coding region identification tool invoking comparative analysis." *Mol Biol Evol*, **1999**, 16 (4), 512-24.
- Bakalarski, C. E., W. Haas, N. E. Dephoure and S. P. Gygi**

"The effects of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics." *Anal Bioanal Chem*, **2007**, 389 (5), 1409-19.

Balgley, B. M., T. Laudeman, L. Yang, T. Song and C. S. Lee
 "Comparative evaluation of tandem MS search algorithms using a target-decoy search strategy." *Mol Cell Proteomics*, **2007**, 6 (9), 1599-608.

Bayes, T. and R. Price
 "An Essay towards solving a Problem in the Doctrine of Chances " *Philosophical Transactions of the Royal Society of London*, **1763**, 53 370-418.

Benson, D. A., I. Karsch-Mizrachi, D. J. Lipman, J. Ostell and D. L. Wheeler
 "GenBank." *Nucleic Acids Res*, **2003**, 31 (1), 23-7.

Bertin, P. N., C. Medigue and P. Normand
 "Advances in environmental genomics: towards an integrated view of micro-organisms and ecosystems." *Microbiology*, **2008**, 154 (Pt 2), 347-59.

Besemer, J., A. Lomsadze and M. Borodovsky
 "GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions." *Nucleic Acids Res*, **2001**, 29 (12), 2607-18.

Biemann, K.
 "Contributions of mass spectrometry to peptide and protein structure." *Biomed Environ Mass Spectrom*, **1988**, 16 (1-12), 99-111.

Blades, A. T., M. G. Ikononou and P. Kebarle
 "Mechanism of electrospray mass spectrometry. Electrospray as an electrolysis cell." *Analytical chemistry*, **1991**, 63 2109-2114.

Blueggel, M., D. Chamrad and H. E. Meyer
 "Bioinformatics in proteomics." *Curr Pharm Biotechnol*, **2004**, 5 (1), 79-88.

Boeckmann, B., A. Bairoch, R. Apweiler, M. C. Blatter, A. Estreicher, E. Gasteiger, M. J. Martin, K. Michoud, C. O'Donovan, I. Phan, S. Pilbout and M. Schneider
 "The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003." *Nucleic Acids Res*, **2003**, 31 (1), 365-70.

Boersema, P. J., N. Divecha, A. J. Heck and S. Mohammed
 "Evaluation and optimization of ZIC-HILIC-RP as an alternative MudPIT strategy." *J Proteome Res*, **2007**, 6 (3), 937-46.

Boguski, M. S.
 "Biosequence exegesis." *Science*, **1999**, 286 (5439), 453-5.

Braun, R. J., N. Kinkl, M. Beer and M. Ueffing
 "Two-dimensional electrophoresis of membrane proteins." *Anal Bioanal Chem*, **2007**, 389 (4), 1033-45.

Breci, L. A., D. L. Tabb, J. R. Yates, 3rd and V. H. Wysocki
 "Cleavage N-terminal to proline: analysis of a database of peptide tandem mass spectra." *Anal Chem*, **2003**, 75 (9), 1963-71.

Brown, N. P., C. Sander and P. Bork
 "Frame: detection of genomic sequencing errors." *Bioinformatics*, **1998**, 14 (4), 367-71.

Buckley, M., A. Walker, S. Y. Ho, Y. Yang, C. Smith, P. Ashton, J. T. Oates, E. Cappellini, H. Koon, K. Penkman, B. Elsworth, D. Ashford, C. Solazzo, P. Andrews, J. Strahler, B. Shapiro, P. Ostrom, H. Gandhi, W. Miller, B. Raney, M. I. Zylber, M. T. Gilbert, R. V. Prigodich, M. Ryan, K. F. Rijdsdijk, A. Janoo and M. J. Collins
 "Comment on "Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry"." *Science*, **2008**, 319 (5859), 33; author reply 33.

Burge, C. and S. Karlin
 "Prediction of complete gene structures in human genomic DNA." *J Mol Biol*, **1997**, 268 (1), 78-94.

Burge, C. B. and S. Karlin
 "Finding the genes in genomic DNA." *Curr Opin Struct Biol*, **1998**, 8 (3), 346-54.

Campana, J. E.
 "Elemental theory of the quadrupole mass filter." *International journal of mass spectrometry and ion processes*, **1980**, 33 101-117.

Cargile, B. J., J. L. Bundy, T. W. Freeman and J. L. Stephenson, Jr.
 "Gel based isoelectric focusing of peptides and the utility of isoelectric point in protein identification." *J Proteome Res*, **2004**, 3 (1), 112-9.

Cargile, B. J., D. L. Talley and J. L. Stephenson, Jr.
 "Immobilized pH gradients as a first dimension in shotgun proteomics and analysis of the accuracy of pI predictability of peptides." *Electrophoresis*, **2004**, 25 (6), 936-45.

Cech, N. B. and C. G. Enke

"Relating electrospray ionization response to nonpolar character of small peptides." *Anal Chem*, **2000**, 72 (13), 2717-23.

Cech, N. B., J. R. Krone and C. G. Enke
 "Electrospray ionization detection of inherently nonresponsive epoxides by peptide binding." *Rapid Commun Mass Spectrom*, **2001**, 15 (13), 1040-4.

Cech, N. B., J. R. Krone and C. G. Enke
 "Predicting electrospray response from chromatographic retention time." *Anal Chem*, **2001**, 73 (2), 208-13.

Chen, Y., S. W. Kwon, S. C. Kim and Y. Zhao
 "Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra." *J Proteome Res*, **2005**, 4 (3), 998-1005.

Clauser, K. R., P. Baker and A. L. Burlingame
 "Role of accurate mass measurement (+/- 10 ppm) in protein identification strategies employing MS or MS/MS and database searching." *Anal Chem*, **1999**, 71 (14), 2871-82.

Colinge, J., A. Masselot, M. Giron, T. Dessingy and J. Magnin
 "OLAV: towards high-throughput tandem mass spectrometry data identification." *Proteomics*, **2003**, 3 (8), 1454-63.

Cox, K. A. and S. J. Gaskell
 "Role of the site of protonation in the low-energy decompositions of gas-phase peptide ions." *Journal of the American Society for Mass Spectrometry*, **1996**, 7 522-531.

Craig, R. and R. C. Beavis
 "TANDEM: matching proteins with tandem mass spectra." *Bioinformatics*, **2004**, 20 (9), 1466-7.

Craig, R., J. P. Cortens and R. C. Beavis
 "The use of proteotypic peptide libraries for protein identification." *Rapid Commun Mass Spectrom*, **2005**, 19 (13), 1844-50.

Cravatt, B. F., G. M. Simon and J. R. Yates, 3rd
 "The biological impact of mass-spectrometry-based proteomics." *Nature*, **2007**, 450 (7172), 991-1000.

Dayhoff, M. O., R. V. Eck, M. A. Chang and M. R. Sochard
 "Atlas of Protein Sequence and Structure." *National biomedical research foundation*, **1965**,

Delcher, A. L., K. A. Bratke, E. C. Powers and S. L. Salzberg
 "Identifying bacterial genes and endosymbiont DNA with Glimmer." *Bioinformatics*, **2007**, 23 (6), 673-9.

Delcher, A. L., D. Harmon, S. Kasif, O. White and S. L. Salzberg
 "Improved microbial gene identification with GLIMMER." *Nucleic Acids Res*, **1999**, 27 (23), 4636-41.

Delmotte, N., M. Lasaosa, A. Tholey, E. Heinzle and C. G. Huber
 "Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis." *J Proteome Res*, **2007**, 6 (11), 4363-73.

Desiere, F., E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich and R. Aebersold
 "The PeptideAtlas project." *Nucleic Acids Res*, **2006**, 34 (Database issue), D655-8.

Desiere, F., E. W. Deutsch, A. I. Nesvizhskii, P. Mallick, N. L. King, J. K. Eng, A. Aderem, R. Boyle, E. Brunner, S. Donohoe, N. Fausto, E. Hafen, L. Hood, M. G. Katze, K. A. Kennedy, F. Kregenow, H. Lee, B. Lin, D. Martin, J. A. Ranish, D. J. Rawlings, L. E. Samelson, Y. Shiio, J. D. Watts, B. Wollscheid, M. E. Wright, W. Yan, L. Yang, E. C. Yi, H. Zhang and R. Aebersold
 "Integration with the human genome of peptide sequences obtained by high-throughput mass spectrometry." *Genome Biol*, **2005**, 6 (1), R9.

Dole, M., L. L. Mack, R. L. Himes, R. C. Mobley, L. D. Ferguson and M. D. Alice
 "Molecular beams of macroions." *The Journal of Chemical Physics*, **1968**, 49 2240-2249.

Domon, B. and R. Aebersold
 "Mass spectrometry and protein analysis." *Science*, **2006**, 312 (5771), 212-7.

Dongre, A. R., J. L. Jones, A. Somogyi and V. H. Wysocki
 "Influence of Peptide Composition, Gas-Phase Basicity, and Chemical Modification on Fragmentation Efficiency: Evidence for the Mobile Proton Model." *Journal of the American Chemical Society*, **1996**, 118 8365-8374.

Durbin, R., S. R. Eddy, A. Krogh and G. Mitchison
 "Biological Sequence Analysis : Probabilistic Models of Proteins and Nucleic Acids", **1999**, Cambridge, Cambridge University Press.

Edman, P.
 "A method for the determination of amino acid sequence in peptides." *Arch Biochem*, **1949**, 22 (3), 475.

Edwards, M. T., S. C. Rison, N. G. Stoker and L. Wernisch
 "A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context." *Nucleic Acids Res*, **2005**, 33 (10), 3253-62.

Elias, J. E. and S. P. Gygi

"Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry." *Nat Methods*, **2007**, 4 (3), 207-14.

Elias, J. E., W. Haas, B. K. Faherty and S. P. Gygi

"Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations." *Nat Methods*, **2005**, 2 (9), 667-75.

Emmet, M. R. and R. M. Caprioli

"Micro-electrospray mass spectrometry: ultra-high-sensitivity analysis of peptides and proteins." *Journal of the American Society for Mass Spectrometry*, **1994**, 5 (7), 605-613.

Eng, J. K., A. L. McCormack and J. R. Yates, 3rd

"An approach to correlate tandem mass-spectral data of peptides with amino-acid-sequences in a protein database." *Journal of the American Society for Mass Spectrometry*, **1994**, 5 976-989.

Enke, C. G.

"A predictive model for matrix and analyte effects in electrospray ionization of singly-charged ionic analytes." *Anal Chem*, **1997**, 69 (23), 4885-93.

Fenn, J. B., M. Mann, C. K. Meng, S. F. Wong and C. M. Whitehouse

"Electrospray ionization for mass spectrometry of large biomolecules." *Science*, **1989**, 246 (4926), 64-71.

Fernandez de la mora, J. and I. G. Loscertales

"The current emitted by highly conducting Taylor cones." *Journal of Fluid Mechanics Digital Archive*, **1994**, 260 155-184.

Fickett, J. W.

"Recognition of protein coding regions in DNA sequences." *Nucleic Acids Res*, **1982**, 10 (17), 5303-18.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and et al.

"Whole-genome random sequencing and assembly of *Haemophilus influenzae* Rd." *Science*, **1995**, 269 (5223), 496-512.

Fleischmann, W., S. Moller, A. Gateau and R. Apweiler

"A novel method for automatic functional annotation of proteins." *Bioinformatics*, **1999**, 15 (3), 228-33.

Fortier, M. H., E. Bonneil, P. Goodley and P. Thibault

"Integrated microfluidic device for mass spectrometry-based proteomics and its application to biomarker discovery programs." *Anal Chem*, **2005**, 77 (6), 1631-40.

Frank, A. and P. Pevzner

"PepNovo: de novo peptide sequencing via probabilistic network modeling." *Anal Chem*, **2005**, 77 (4), 964-73.

Fraser, C. M., J. A. Eisen and S. L. Salzberg

"Microbial genome sequencing." *Nature*, **2000**, 406 (6797), 799-803.

Frewen, B. E., G. E. Merrihew, C. C. Wu, W. S. Noble and M. J. MacCoss

"Analysis of peptide MS/MS spectra from large-scale proteomics experiments using spectrum libraries." *Anal Chem*, **2006**, 78 (16), 5678-84.

Frishman, D.

"Protein annotation at genomic scale: the current status." *Chem Rev*, **2007**, 107 (8), 3448-66.

Frishman, D., A. Mironov, H. W. Mewes and M. Gelfand

"Combining diverse evidence for gene recognition in completely sequenced bacterial genomes." *Nucleic Acids Res*, **1998**, 26 (12), 2941-7.

Galperin, M. Y., D. R. Walker and E. V. Koonin

"Analogous enzymes: independent inventions in enzyme evolution." *Genome Res*, **1998**, 8 (8), 779-90.

Gasteiger, E., E. Jung and A. Bairoch

"SWISS-PROT: connecting biomolecular knowledge via a protein database." *Curr Issues Mol Biol*, **2001**, 3 (3), 47-55.

Gattiker, A., K. Michoud, C. Rivoire, A. H. Auchincloss, E. Coudert, T. Lima, P. Kersey, M. Pagni, C. J. Sigrist, C. Lachaize, A. L. Veuthey, E. Gasteiger and A. Bairoch

"Automated annotation of microbial proteomes in SWISS-PROT." *Comput Biol Chem*, **2003**, 27 (1), 49-58.

Geer, L. Y., S. P. Markey, J. A. Kowalak, L. Wagner, M. Xu, D. M. Maynard, X. Yang, W. Shi and S. H. Bryant

"Open mass spectrometry search algorithm." *J Proteome Res*, **2004**, 3 (5), 958-64.

Giddings, J. C.

"Two-dimensional separations: concept and promise." *Anal Chem*, **1984**, 56 (12), 1258A-1260A, 1262A, 1264A passim.

Gorg, A., W. Weiss and M. J. Dunn

"Current two-dimensional electrophoresis technology for proteomics." *Proteomics*, **2004**, 4 (12), 3665-85.

Green, E. D.

"Strategies for the systematic sequencing of complex genomes." *Nat Rev Genet*, **2001**, 2 (8), 573-83.

Gribkov, M., J. Devereux and R. R. Burgess
 "The codon preference plot: graphic analysis of protein coding sequences and prediction of gene expression." *Nucleic Acids Res*, **1984**, 12 (1 Pt 2), 539-49.

Groisman, I. and H. Engelberg-Kulka
 "Translational bypassing: a new reading alternative of the genetic code." *Biochem Cell Biol*, **1995**, 73 (11-12), 1055-9.

Gupta, N. and P. Pevzner
 "False discovery rates of protein identifications: a strike against the two-peptide rule." *J Proteome Res*, **2009**,
Gupta, N., S. Tanner, N. Jaitly, J. N. Adkins, M. Lipton, R. Edwards, M. Romine, A. Osterman, V. Bafna, R. D. Smith and P. A. Pevzner
 "Whole proteome analysis of post-translational modifications: applications of mass-spectrometry for proteogenomic annotation." *Genome Res*, **2007**, 17 (9), 1362-77.

Gurvich, O. L., P. V. Baranov, J. Zhou, A. W. Hammer, R. F. Gesteland and J. F. Atkins
 "Sequences that direct significant levels of frameshifting are frequent in coding regions of Escherichia coli." *Embo J*, **2003**, 22 (21), 5941-50.

Harrison, A. G. and T. Yalcin
 "Proton mobility in protonated amino acids and peptides." *International journal of mass spectrometry and ion processes*, **1997**, 165-166 339-347.

Helsens, K., E. Timmerman, J. Vandekerckhove, K. Gevaert and L. Martens
 "Peptizer, a tool for assessing false positive peptide identifications and manually validating selected results." *Mol Cell Proteomics*, **2008**, 7 (12), 2364-72.

Henzel, W. J., T. M. Billeci, J. T. Stults, S. C. Wong, C. Grimley and C. Watanabe
 "Identifying proteins from two-dimensional gels by molecular mass searching of peptide fragments in protein sequence databases." *Proc Natl Acad Sci U S A*, **1993**, 90 (11), 5011-5.

Horie, M., K. Fukui, M. Xie, Y. Kageyama, K. Hamada, Y. Sakihama, K. Sugimori and K. Matsumoto
 "The N-terminal region is important for the nuclease activity and thermostability of the flap endonuclease-1 from *Sulfolobus tokodaii*." *Biosci Biotechnol Biochem*, **2007**, 71 (4), 855-65.

Hoskins, R. A., C. R. Nelson, B. P. Berman, T. R. Lavery, R. A. George, L. Ciesiolka, M. Naemuddin, A. D. Arenson, J. Durbin, R. G. David, P. E. Tabor, M. R. Bailey, D. R. DeShazo, J. Catanese, A. Mammoser, K. Osoegawa, P. J. de Jong, S. E. Celniker, R. A. Gibbs, G. M. Rubin and S. E. Scherer
 "A BAC-based physical map of the major autosomes of *Drosophila melanogaster*." *Science*, **2000**, 287 (5461), 2271-4.

Huang, Y., J. M. Triscari, G. C. Tseng, L. Pasa-Tolic, M. S. Lipton, R. D. Smith and V. H. Wysocki
 "Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns." *Anal Chem*, **2005**, 77 (18), 5800-13.

Hubbard, T. J., B. L. Aken, S. Ayling, B. Ballester, K. Beal, E. Bragin, S. Brent, Y. Chen, P. Clapham, L. Clarke, G. Coates, S. Fairley, S. Fitzgerald, J. Fernandez-Banet, L. Gordon, S. Graf, S. Haider, M. Hammond, R. Holland, K. Howe, A. Jenkinson, N. Johnson, A. Kahari, D. Keefe, S. Keenan, R. Kinsella, F. Kokocinski, E. Kulesha, D. Lawson, I. Longden, K. Megy, P. Meidl, B. Overduin, A. Parker, B. Pritchard, D. Rios, M. Schuster, G. Slater, D. Smedley, W. Spooner, G. Spudich, S. Trevanion, A. Vilella, J. Vogel, S. White, S. Wilder, A. Zadissa, E. Birney, F. Cunningham, V. Curwen, R. Durbin, X. M. Fernandez-Suarez, J. Herrero, A. Kasprzyk, G. Proctor, J. Smith, S. Searle and P. Flicek
 "Ensembl 2009." *Nucleic Acids Res*, **2009**, 37 (Database issue), D690-7.

Ikegami, T., E. Dicks, H. Kobayashi, H. Morisaka, D. Tokuda, K. Cabrera, K. Hosoya and N. Tanaka
 "How to utilize the true performance of monolithic silica columns." *J Sep Sci*, **2004**, 27 (15-16), 1292-302.

Ikonomou, M. G., A. T. Blades and P. Kebarle
 "Investigations of the electrospray interface for liquid chromatography/mass spectrometry." *Analytical chemistry*, **1990**, 62 (9), 957-967.

Iribarne, J. V., P. J. Dziedzic and B. A. Thomson
 "Atmospheric pressure ion evaporation-mass spectrometry." *International journal of mass spectrometry and ion physics*, **1983**, 50 331-347.

Jaffe, J. D., H. C. Berg and G. M. Church
 "Proteogenomic mapping as a complementary method to perform genome annotation." *Proteomics*, **2004**, 4 (1), 59-77.

James, P., M. Quadroni, E. Carafoli and G. Gonnet
 "Protein identification by mass profile fingerprinting." *Biochem Biophys Res Commun*, **1993**, 195 (1), 58-64.

Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults and J. T. Watson
 "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine." *Analytical chemistry*, **1987**, 59 (21), 2321-2625.

- Johnson, R. S., S. A. Martin, K. Biemann, J. T. Stults and J. T. Watson**
 "Novel fragmentation process of peptides by collision-induced decomposition in a tandem mass spectrometer: differentiation of leucine and isoleucine." *Anal Chem*, **1987**, 59 (21), 2621-5.
- Johnson, R. S. and J. A. Taylor**
 "Searching sequence databases via de novo peptide sequencing by tandem mass spectrometry." *Mol Biotechnol*, **2002**, 22 (3), 301-15.
- Jones, A. R., J. A. Siepen, S. J. Hubbard and N. W. Paton**
 "Improving sensitivity in proteome studies by analysis of false discovery rates for multiple search engines." *Proteomics*, **2009**, 9 (5), 1220-9.
- Jones, J. L., A. R. Dongre, A. Somogyi and V. H. Wysocki**
 "Sequence Dependence of Peptide Fragmentation Efficiency Curves Determined by Electrospray Ionization/Surface-Induced Dissociation Mass Spectrometry." *Journal of the American Chemical Society*, **1994**, 116 (18), 8368-8369.
- Jones, P., R. G. Cote, L. Martens, A. F. Quinn, C. F. Taylor, W. Derache, H. Hermjakob and R. Apweiler**
 "PRIDE: a public repository of protein and peptide identifications for the proteomics community." *Nucleic Acids Res*, **2006**, 34 (Database issue), D659-63.
- Kapp, E. A., F. Schutz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn and R. J. Simpson**
 "An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis." *Proteomics*, **2005**, 5 (13), 3475-90.
- Karas, M., U. Bahr and T. Dulcks**
 "Nano-electrospray ionization mass spectrometry: addressing analytical problems beyond routine." *Fresenius J Anal Chem*, **2000**, 366 (6-7), 669-76.
- Karas, M. and F. Hillenkamp**
 "Laser desorption ionization of proteins with molecular masses exceeding 10,000 daltons." *Anal Chem*, **1988**, 60 (20), 2299-301.
- Karplus, P. A.**
 "Hydrophobicity regained." *Protein Science*, **1996**, 6 1302-1307.
- Kebarle, P.**
 "A brief overview of the present status of the mechanisms involved in electrospray mass spectrometry." *J Mass Spectrom*, **2000**, 35 (7), 804-17.
- Kebarle, P. and M. Peschke**
 "On the mechanisms by which charge droplets produced by electrospray lead to gas phase ions." *Analytica Chimica Acta*, **1999**, 20070 1-25.
- Keller, A., J. Eng, N. Zhang, X. J. Li and R. Aebersold**
 "A uniform proteomics MS/MS analysis platform utilizing open XML file formats." *Mol Syst Biol*, **2005**, 1 2005 0017.
- Keller, A., A. I. Nesvizhskii, E. Kolker and R. Aebersold**
 "Empirical statistical model to estimate the accuracy of peptide identifications made by MS/MS and database search." *Anal Chem*, **2002**, 74 (20), 5383-92.
- Kersey, P. J., J. Duarte, A. Williams, Y. Karavidopoulou, E. Birney and R. Apweiler**
 "The International Protein Index: an integrated database for proteomics experiments." *Proteomics*, **2004**, 4 (7), 1985-8.
- Kozak, M.**
 "Initiation of translation in prokaryotes and eukaryotes." *Gene*, **1999**, 234 (2), 187-208.
- Kristensen, D. B., J. C. Brond, P. A. Nielsen, J. R. Andersen, O. T. Sorensen, V. Jorgensen, K. Budin, J. Matthiesen, P. Venø, H. M. Jespersen, C. H. Ahrens, S. Schandorff, P. T. Ruhoff, J. R. Wisniewski, K. L. Bennett and A. V. Podtelejnikov**
 "Experimental Peptide Identification Repository (EPIR): an integrated peptide-centric platform for validation and mining of tandem mass spectrometry data." *Mol Cell Proteomics*, **2004**, 3 (10), 1023-38.
- Kriventseva, E. V., I. Koch, R. Apweiler, M. Vingron, P. Bork, M. S. Gelfand and S. Sunyaev**
 "Increase of functional diversity by alternative splicing." *Trends Genet*, **2003**, 19 (3), 124-8.
- Lam, H., E. W. Deutsch, J. S. Eddes, J. K. Eng, N. King, S. E. Stein and R. Aebersold**
 "Development and validation of a spectral library searching method for peptide identification from MS/MS." *Proteomics*, **2007**, 7 (5), 655-67.
- Lam, H. T., J. Jossierand, N. Lion and H. H. Girault**
 "Modeling the isoelectric focusing of peptides in an OFFGEL multicompartiment cell." *J Proteome Res*, **2007**, 6 (5), 1666-76.
- le Coutre, J., J. P. Whitelegge, A. Gross, E. Turk, E. M. Wright, H. R. Kaback and K. F. Faull**
 "Proteomics on full-length membrane proteins using mass spectrometry." *Biochemistry*, **2000**, 39 (15), 4237-42.

Leinonen, R., F. G. Diez, D. Binns, W. Fleischmann, R. Lopez and R. Apweiler
 "UniProt archive." *Bioinformatics*, **2004**, 20 (17), 3236-7.

Lima, T., A. H. Auchincloss, E. Coudert, G. Keller, K. Michoud, C. Rivoire, V. Bulliard, E. de Castro, C. Lachaize, D. Baratin, I. Phan, L. Bougueleret and A. Bairoch
 "HAMAP: a database of completely sequenced microbial proteome sets and manually curated microbial protein families in UniProtKB/Swiss-Prot." *Nucleic Acids Res*, **2009**, 37 (Database issue), D471-8.

Link, A. J., J. Eng, D. M. Schieltz, E. Carmack, G. J. Mize, D. R. Morris, B. M. Garvik and J. R. Yates, 3rd
 "Direct analysis of protein complexes using mass spectrometry." *Nat Biotechnol*, **1999**, 17 (7), 676-82.

Linke, B., A. C. McHardy, H. Neuweger, L. Krause and F. Meyer
 "REGANOR: a gene prediction server for prokaryotic genomes and a database of high quality gene predictions for prokaryotes." *Appl Bioinformatics*, **2006**, 5 (3), 193-8.

Ma, B., K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie
 "PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry." *Rapid Commun Mass Spectrom*, **2003**, 17 (20), 2337-42.

Malmstrom, J., H. Lee, A. I. Nesvizhskii, D. Shteynberg, S. Mohanty, E. Brunner, M. Ye, G. Weber, C. Eckerskorn and R. Aebersold
 "Optimized peptide separation and identification for mass spectrometry based proteomics via free-flow electrophoresis." *J Proteome Res*, **2006**, 5 (9), 2241-9.

Mann, M., R. C. Hendrickson and A. Pandey
 "Analysis of proteins and proteomes by mass spectrometry." *Annu Rev Biochem*, **2001**, 70 437-73.

Mann, M., P. Hojrup and P. Roepstorff
 "Use of mass spectrometric molecular weight information to identify proteins in sequence databases." *Biol Mass Spectrom*, **1993**, 22 (6), 338-45.

March, R. E.
 "Quadrupole ion traps." *Mass Spectrom Rev*, **2009**,

Martens, L., H. Hermjakob, P. Jones, M. Adamski, C. Taylor, D. States, K. Gevaert, J. Vandekerckhove and R. Apweiler
 "PRIDE: the proteomics identifications database." *Proteomics*, **2005**, 5 (13), 3537-45.

Mawuenyega, K. G., H. Kaji, Y. Yamuchi, T. Shinkawa, H. Saito, M. Taoka, N. Takahashi and T. Isobe
 "Large-scale identification of *Caenorhabditis elegans* proteins by multidimensional liquid chromatography-tandem mass spectrometry." *J Proteome Res*, **2003**, 2 (1), 23-35.

Maxam, A. M. and W. Gilbert
 "A new method for sequencing DNA." *Proc Natl Acad Sci U S A*, **1977**, 74 (2), 560-4.

McDonald, T., S. Sheng, B. Stanley, D. Chen, Y. Ko, R. N. Cole, P. Pedersen and J. E. Van Eyk
 "Expanding the subproteome of the inner mitochondria using protein separation technologies: one- and two-dimensional liquid chromatography and two-dimensional gel electrophoresis." *Mol Cell Proteomics*, **2006**, 5 (12), 2392-411.

Mead, J. A., L. Bianco and C. Bessant
 "Recent developments in public proteomic MS repositories and pipelines." *Proteomics*, **2009**, 9 (4), 861-81.

Mead, J. A., I. P. Shadforth and C. Bessant
 "Public proteomic MS repositories and pipelines: available tools and biological applications." *Proteomics*, **2007**, 7 (16), 2769-86.

Medigue, C., M. Rose, A. Viari and A. Danchin
 "Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence." *Genome Res*, **1999**, 9 (11), 1116-27.

Meyer, F., A. Goesmann, A. C. McHardy, D. Bartels, T. Bekel, J. Clausen, J. Kalinowski, B. Linke, O. Rupp, R. Giegerich and A. Puhler
 "GenDB--an open source genome annotation system for prokaryote genomes." *Nucleic Acids Res*, **2003**, 31 (8), 2187-95.

Miyazaki, S., H. Sugawara, T. Gojobori and Y. Tateno
 "DNA Data Bank of Japan (DDBJ) in XML." *Nucleic Acids Res*, **2003**, 31 (1), 13-6.

Mons, B., M. Ashburner, C. Chichester, E. van Mulligen, M. Weeber, J. den Dunnen, G. J. van Ommen, M. Musen, M. Cockerill, H. Hermjakob, A. Mons, A. Packer, R. Pacheco, S. Lewis, A. Berkeley, W. Melton, N. Barris, J. Wales, G. Meijssen, E. Moeller, P. J. Roes, K. Borner and A. Bairoch
 "Calling on a million minds for community annotation in WikiProteins." *Genome Biol*, **2008**, 9 (5), R89.

Moore, R. E., M. K. Young and T. D. Lee
 "Qscore: an algorithm for evaluating SEQUEST database search results." *J Am Soc Mass Spectrom*, **2002**, 13 (4), 378-86.

Motoyama, A., T. Xu, C. I. Ruse, J. A. Wohlschlegel and J. R. Yates, 3rd

"Anion and cation mixed-bed ion exchange for enhanced multidimensional separations of peptides and phosphopeptides." *Anal Chem*, **2007**, 79 (10), 3623-34.

Motoyama, A. and J. R. Yates, 3rd
 "Multidimensional LC separations in shotgun proteomics." *Anal Chem*, **2008**, 80 (19), 7187-93.

Mulder, N. J., R. Apweiler, T. K. Attwood, A. Bairoch, D. Barrell, A. Bateman, D. Binns, M. Biswas, P. Bradley, P. Bork, P. Bucher, R. R. Copley, E. Courcelle, U. Das, R. Durbin, L. Falquet, W. Fleischmann, S. Griffiths-Jones, D. Haft, N. Harte, N. Hulo, D. Kahn, A. Kanapin, M. Krestyaninova, R. Lopez, I. Letunic, D. Lonsdale, V. Silventoinen, S. E. Orchard, M. Pagni, D. Peyruc, C. P. Ponting, J. D. Selengut, F. Servant, C. J. Sigrist, R. Vaughan and E. M. Zdobnov
 "The InterPro Database, 2003 brings increased coverage and new features." *Nucleic Acids Res*, **2003**, 31 (1), 315-8.

Myers, E. W., G. G. Sutton, A. L. Delcher, I. M. Dew, D. P. Fasulo, M. J. Flanigan, S. A. Kravitz, C. M. Mobarry, K. H. Reinert, K. A. Remington, E. L. Anson, R. A. Bolanos, H. H. Chou, C. M. Jordan, A. L. Halpern, S. Lonardi, E. M. Beasley, R. C. Brandon, L. Chen, P. J. Dunn, Z. Lai, Y. Liang, D. R. Nusskern, M. Zhan, Q. Zhang, X. Zheng, G. M. Rubin, M. D. Adams and J. C. Venter
 "A whole-genome assembly of *Drosophila*." *Science*, **2000**, 287 (5461), 2196-204.

Nesvizhskii, A. I. and R. Aebersold
 "Interpretation of shotgun proteomic data: the protein inference problem." *Mol Cell Proteomics*, **2005**, 4 (10), 1419-40.

Nesvizhskii, A. I., A. Keller, E. Kolker and R. Aebersold
 "A statistical model for identifying proteins by tandem mass spectrometry." *Anal Chem*, **2003**, 75 (17), 4646-58.

Nesvizhskii, A. I., O. Vitek and R. Aebersold
 "Analysis and validation of proteomic data generated by tandem mass spectrometry." *Nat Methods*, **2007**, 4 (10), 787-97.

Nielsen, P. and A. Krogh
 "Large-scale prokaryotic gene prediction and comparison to genome annotation." *Bioinformatics*, **2005**, 21 (24), 4322-9.

Oliver, S. G., Q. J. van der Aart, M. L. Agostoni-Carbone, M. Aigle, L. Alberghina, D. Alexandraki, G. Antoine, R. Anwar, J. P. Ballesta, P. Benit and et al.
 "The complete DNA sequence of yeast chromosome III." *Nature*, **1992**, 357 (6373), 38-46.

Olsen, J. V. and M. Mann
 "Improved peptide identification in proteomics by two consecutive stages of mass spectrometric fragmentation." *Proc Natl Acad Sci U S A*, **2004**, 101 (37), 13417-22.

Opiteck, G. J. and J. W. Jorgenson
 "Two-dimensional SEC/RPLC coupled to mass spectrometry for the analysis of peptides." *Anal Chem*, **1997**, 69 (13), 2283-91.

Orchard, S., C. Hoogland, A. Bairoch, M. Eisenacher, H. J. Kraus and P. A. Binz
 "Managing the data explosion. A report on the HUPO-PSI Workshop. August 2008, Amsterdam, The Netherlands." *Proteomics*, **2009**, 9 (3), 499-501.

Overbeek, R., D. Bartels, V. Vonstein and F. Meyer
 "Annotation of bacterial and archaeal genomes: improving accuracy and consistency." *Chem Rev*, **2007**, 107 (8), 3431-47.

Paizs, B. and S. Suhai
 "Towards understanding the tandem mass spectra of protonated oligopeptides. 1: mechanism of amide bond cleavage." *J Am Soc Mass Spectrom*, **2004**, 15 (1), 103-13.

Pappin, D. J., P. Hojrup and A. J. Bleasby
 "Rapid identification of proteins by peptide-mass fingerprinting." *Curr Biol*, **1993**, 3 (6), 327-32.

Parra, G., E. Blanco and R. Guigo
 "GeneID in *Drosophila*." *Genome Res*, **2000**, 10 (4), 511-5.

Patterson, S. D. and R. H. Aebersold
 "Proteomics: the first decade and beyond." *Nat Genet*, **2003**, 33 Suppl 311-23.

Pearson, W. R., T. Wood, Z. Zhang and W. Miller
 "Comparison of DNA sequences with protein sequences." *Genomics*, **1997**, 46 (1), 24-36.

Peng, J., J. E. Elias, C. C. Thoreen, L. J. Licklider and S. P. Gygi
 "Evaluation of multidimensional chromatography coupled with tandem mass spectrometry (LC/LC-MS/MS) for large-scale protein analysis: the yeast proteome." *J Proteome Res*, **2003**, 2 (1), 43-50.

Perkins, D. N., D. J. Pappin, D. M. Creasy and J. S. Cottrell
 "Probability-based protein identification by searching sequence databases using mass spectrometry data." *Electrophoresis*, **1999**, 20 (18), 3551-67.

Perrodou, E., C. Deshayes, J. Muller, C. Schaeffer, A. Van Dorsselaer, R. Ripp, O. Poch, J. M. Reytrat and O. Lecompte
 "ICDS database: interrupted CoDing sequences in prokaryotic genomes." *Nucleic Acids Res*, **2006**, 34 (Database issue), D338-43.

Peterson, J. D., L. A. Umayam, T. Dickinson, E. K. Hickey and O. White
 "The Comprehensive Microbial Resource." *Nucleic Acids Res*, **2001**, 29 (1), 123-5.

Pevzner, P. A., S. Kim and J. Ng
 "Comment on "Protein sequences from mastodon and Tyrannosaurus rex revealed by mass spectrometry"." *Science*, **2008**, 321 (5892), 1040; author reply 1040.

Pfeifer, N., A. Leinenbach, C. G. Huber and O. Kohlbacher
 "Improving Peptide identification in proteome analysis by a two-dimensional retention time filtering approach." *J Proteome Res*, **2009**, 8 (8), 4109-15.

Premstaller, A., H. Oberacher, W. Walcher, A. M. Timperio, L. Zolla, J. P. Chervet, N. Cavusoglu, A. van Dorsselaer and C. G. Huber
 "High-performance liquid chromatography-electrospray ionization mass spectrometry using monolithic capillary columns for proteomic studies." *Anal Chem*, **2001**, 73 (11), 2390-6.

Price, T. S., M. B. Lucitt, W. Wu, D. J. Austin, A. Pizarro, A. K. Yocum, I. A. Blair, G. A. FitzGerald and T. Grosser
 "EBP, a program for protein identification using multiple tandem mass spectrometry datasets." *Mol Cell Proteomics*, **2007**, 6 (3), 527-36.

Pruitt, K. D., T. Tatusova, W. Klimke and D. R. Maglott
 "NCBI Reference Sequences: current status, policy and new initiatives." *Nucleic Acids Res*, **2009**, 37 (Database issue), D32-6.

Pruitt, K. D., T. Tatusova and D. R. Maglott
 "NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins." *Nucleic Acids Res*, **2007**, 35 (Database issue), D61-5.

Qian, W. J., T. Liu, M. E. Monroe, E. F. Strittmatter, J. M. Jacobs, L. J. Kangas, K. Petritis, D. G. Camp, 2nd and R. D. Smith
 "Probability-based evaluation of peptide and protein identifications from tandem mass spectrometry and SEQUEST analysis: the human proteome." *J Proteome Res*, **2005**, 4 (1), 53-62.

Rayleigh, L.
 "On the equilibrium of liquid conducting masses charged with electricity." *Philosophical Magazine*, **1882**, 14 184-186.

Resing, K. A., K. Meyer-Arendt, A. M. Mendoza, L. D. Aveline-Wolf, K. R. Jonscher, K. G. Pierce, W. M. Old, H. T. Cheung, S. Russell, J. L. Wattawa, G. R. Goehle, R. D. Knight and N. G. Ahn
 "Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics." *Anal Chem*, **2004**, 76 (13), 3556-68.

Rieux, L., H. Niederlander, E. Verpoorte and R. Bischoff
 "Silica monolithic columns: synthesis, characterisation and applications to the analysis of biological molecules." *J Sep Sci*, **2005**, 28 (14), 1628-41.

Roepstorff, P. and J. Fohlman
 "Proposal for a common nomenclature for sequence ions in mass spectra of peptides." *Biomed Mass Spectrom*, **1984**, 11 (11), 601.

Ros, A., M. Faupel, H. Mees, J. Oostrum, R. Ferrigno, F. Reymond, P. Michel, J. S. Rossier and H. H. Girault
 "Protein purification by Off-Gel electrophoresis." *Proteomics*, **2002**, 2 (2), 151-6.

Roth, K. D., Z. H. Huang, N. Sadagopan and J. T. Watson
 "Charge derivatization of peptides for analysis by mass spectrometry." *Mass Spectrom Rev*, **1998**, 17 (4), 255-74.

Salgado, H., G. Moreno-Hagelsieb, T. F. Smith and J. Collado-Vides
 "Operons in Escherichia coli: genomic analyses and predictions." *Proc Natl Acad Sci U S A*, **2000**, 97 (12), 6652-7.

Salzberg, S. L., A. L. Delcher, S. Kasif and O. White
 "Microbial gene identification using interpolated Markov models." *Nucleic Acids Res*, **1998**, 26 (2), 544-8.

Sandra, K., M. Moshir, F. D'Hondt, K. Verleysen, K. Kas and P. Sandra
 "Highly efficient peptide separations in proteomics Part 1. Unidimensional high performance liquid chromatography." *J Chromatogr B Analyt Technol Biomed Life Sci*, **2008**, 866 (1-2), 48-63.

Sanger, F.
 "Chemistry of insulin; determination of the structure of insulin opens the way to greater understanding of life processes." *Science*, **1959**, 129 (3359), 1340-4.

Sanger, F.
 "Sequences, sequences, and sequences." *Annu Rev Biochem*, **1988**, 57 1-28.

Sanger, F., G. M. Air, B. G. Barrell, N. L. Brown, A. R. Coulson, C. A. Fiddes, C. A. Hutchison, P. M. Slocombe and M. Smith
 "Nucleotide sequence of bacteriophage phi X174 DNA." *Nature*, **1977**, 265 (5596), 687-95.

Santoni, V., M. Molloy and T. Rabilloud
 "Membrane proteins and proteomics: un amour impossible?" *Electrophoresis*, **2000**, 21 (6), 1054-70.

Schluesener, D., F. Fischer, J. Kruip, M. Rogner and A. Poetsch
 "Mapping the membrane proteome of *Corynebacterium glutamicum*." *Proteomics*, **2005**, 5 (5), 1317-30.

Searle, B. C., M. Turner and A. I. Nesvizhskii
 "Improving sensitivity by probabilistically combining results from multiple MS/MS search methodologies." *J Proteome Res*, **2008**, 7 (1), 245-53.

Shen, Y., R. D. Smith, K. K. Unger, D. Kumar and D. Lubda
 "Ultrahigh-throughput proteomics using fast RPLC separations with ESI-MS/MS." *Anal Chem*, **2005**, 77 (20), 6692-701.

Shen, Y., R. Zhang, R. J. Moore, J. Kim, T. O. Metz, K. K. Hixson, R. Zhao, E. A. Livesay, H. R. Udseth and R. D. Smith
 "Automated 20 kpsi RPLC-MS and MS/MS with chromatographic peak capacities of 1000-1500 and capabilities in proteomics and metabolomics." *Anal Chem*, **2005**, 77 (10), 3090-100.

Shevchenko, A., S. Sunyaev, A. Loboda, A. Shevchenko, P. Bork, W. Ens and K. G. Standing
 "Charting the proteomes of organisms with unsequenced genomes by MALDI-quadrupole time-of-flight mass spectrometry and BLAST homology searching." *Anal Chem*, **2001**, 73 (9), 1917-26.

Shine, J. and L. Dalgarno
 "The 3'-terminal sequence of *Escherichia coli* 16S ribosomal RNA: complementarity to nonsense triplets and ribosome binding sites." *Proc Natl Acad Sci U S A*, **1974**, 71 (4), 1342-6.

Sickmann, A., M. Mreyen and H. E. Meyer
 "Mass spectrometry--a key technology in proteome research." *Adv Biochem Eng Biotechnol*, **2003**, 83 141-76.

Sjoberg, P. J., C. F. Bokman, D. Bylund and K. E. Markides
 "A method for determination of ion distribution within electrosprayed droplets." *Anal Chem*, **2001**, 73 (1), 23-8.

Skovgaard, M., L. J. Jensen, S. Brunak, D. Ussery and A. Krogh
 "On the total number of genes and their length distribution in complete microbial genomes." *Trends Genet*, **2001**, 17 (8), 425-8.

Staden, R.
 "Measurements of the effects that coding for a protein has on a DNA sequence and their use for finding genes." *Nucleic Acids Res*, **1984**, 12 (1 Pt 2), 551-67.

Stamm, S., S. Ben-Ari, I. Rafalska, Y. Tang, Z. Zhang, D. Toiber, T. A. Thanaraj and H. Soreq
 "Function of alternative splicing." *Gene*, **2005**, 344 1-20.

Steen, H. and M. Mann
 "The ABC's (and XYZ's) of peptide sequencing." *Nat Rev Mol Cell Biol*, **2004**, 5 (9), 699-711.

Stoesser, G., W. Baker, A. van den Broek, M. Garcia-Pastor, C. Kanz, T. Kulikova, R. Leinonen, Q. Lin, V. Lombard, R. Lopez, R. Mancuso, F. Nardone, P. Stoehr, M. A. Tuli, K. Tzouvara and R. Vaughan
 "The EMBL Nucleotide Sequence Database: major new developments." *Nucleic Acids Res*, **2003**, 31 (1), 17-22.

Strittmatter, E. F., L. J. Kangas, K. Petritis, H. M. Mottaz, G. A. Anderson, Y. Shen, J. M. Jacobs, D. G. Camp, 2nd and R. D. Smith
 "Application of peptide LC retention time information in a discriminant function for peptide identification by tandem mass spectrometry." *J Proteome Res*, **2004**, 3 (4), 760-9.

Suzek, B. E., M. D. Ermolaeva, M. Schreiber and S. L. Salzberg
 "A probabilistic method for identifying start codons in bacterial genomes." *Bioinformatics*, **2001**, 17 (12), 1123-30.

Suzek, B. E., H. Huang, P. McGarvey, R. Mazumder and C. H. Wu
 "UniRef: comprehensive and non-redundant UniProt reference clusters." *Bioinformatics*, **2007**, 23 (10), 1282-8.

Tabb, D. L., A. Saraf and J. R. Yates, 3rd
 "GutenTag: high-throughput sequence tagging via an empirically derived fragmentation model." *Anal Chem*, **2003**, 75 (23), 6415-21.

Tanaka, T., D. J. Slamon, H. Shimoda, C. Waki, Y. Kawaguchi, Y. Tanaka and N. Ida
 "Expression of Ha-ras oncogene products in human neuroblastomas and the significant correlation with a patient's prognosis." *Cancer Res*, **1988**, 48 (4), 1030-4.

Tang, K. and R. D. Smith
 "Physical/chemical separations in the break-up of highly charged droplets from electrosprays." *J Am Soc Mass Spectrom*, **2001**, 12 (3), 343-7.

Tang, L. and P. Kebarle

"Effect of the conductivity of the electrosprayed solution on the electrospray current. Factors determining analyte sensitivity in electrospray mass spectrometry." *Analytical chemistry*, **1991**, 63 2709-2715.

Tanner, S., H. Shu, A. Frank, L. C. Wang, E. Zandi, M. Mumby, P. A. Pevzner and V. Bafna

"InsPecT: identification of posttranslationally modified peptides from tandem mass spectra." *Anal Chem*, **2005**, 77 (14), 4626-39.

Taylor, C. F., N. W. Paton, K. S. Lilley, P. A. Binz, R. K. Julian, Jr., A. R. Jones, W. Zhu, R. Apweiler, R. Aebersold, E. W. Deutsch, M. J. Dunn, A. J. Heck, A. Leitner, M. Macht, M. Mann, L. Martens, T. A. Neubert, S. D. Patterson, P. Ping, S. L. Seymour, P. Souda, A. Tsugita, J. Vandekerckhove, T. M.

Vondriska, J. P. Whitelegge, M. R. Wilkins, I. Xenarios, J. R. Yates, 3rd and H. Hermjakob

"The minimum information about a proteomics experiment (MIAPE)." *Nat Biotechnol*, **2007**, 25 (8), 887-93.

Tech, M., N. Pfeifer, B. Morgenstern and P. Meinicke

"TICO: a tool for improving predictions of prokaryotic translation initiation sites." *Bioinformatics*, **2005**, 21 (17), 3568-9.

Thomson, B. A. and J. V. Iribarne

"Field induced ion evaporation from liquid surfaces at atmospheric pressure." *The Journal of Chemical Physics*, **1976**, 71 (11), 4451-4463.

Trivedi, O. A., P. Arora, V. Sridharan, R. Tickoo, D. Mohanty and R. S. Gokhale

"Enzymic activation and transfer of fatty acids as acyl-adenylates in mycobacteria." *Nature*, **2004**, 428 (6981), 441-5.

Tsaprailis, G., H. Nair, A. Somogyi, V. H. Wysocki, W. Zhong, J. H. Futrell, S. G. Summerfield and S. J. Gaskell

"Influence of secondary structure on the fragmentation of protonated peptides." *Journal of the American Chemical Society*, **1999**, 121 (22), 5142-5154.

Tyers, M. and M. Mann

"From genomics to proteomics." *Nature*, **2003**, 422 (6928), 193-7.

Van Domselaar, G. H., P. Stothard, S. Shrivastava, J. A. Cruz, A. Guo, X. Dong, P. Lu, D. Szafron, R. Greiner and D. S. Wishart

"BASys: a web server for automated bacterial genome annotation." *Nucleic Acids Res*, **2005**, 33 (Web Server issue), W455-9.

Wagner, Y., A. Sickmann, H. E. Meyer and G. Daum

"Multidimensional nano-HPLC for analysis of protein complexes." *J Am Soc Mass Spectrom*, **2003**, 14 (9), 1003-11.

Washburn, M. P., D. Wolters and J. R. Yates, 3rd

"Large-scale analysis of the yeast proteome by multidimensional protein identification technology." *Nat Biotechnol*, **2001**, 19 (3), 242-7.

Weinstock, G. M.

"Genomics and bacterial pathogenesis." *Emerg Infect Dis*, **2000**, 6 (5), 496-504.

Westbrook, J., Z. Feng, L. Chen, H. Yang and H. M. Berman

"The Protein Data Bank and structural genomics." *Nucleic Acids Res*, **2003**, 31 (1), 489-91.

Wheeler, D. L., D. M. Church, S. Federhen, A. E. Lash, T. L. Madden, J. U. Pontius, G. D. Schuler, L. M. Schriml, E. Sequeira, T. A. Tatusova and L. Wagner

"Database resources of the National Center for Biotechnology." *Nucleic Acids Res*, **2003**, 31 (1), 28-33.

Whitehouse, C. M., R. N. Dreyer, M. Yamashita and J. B. Fenn

"Electrospray interface for liquid chromatographs and mass spectrometers." *Analytical chemistry*, **1985**, 57 (3), 675-679.

Wilkins, M. R., C. Pasquali, R. D. Appel, K. Ou, O. Golaz, J. C. Sanchez, J. X. Yan, A. A. Gooley, G.

Hughes, I. Humphery-Smith, K. L. Williams and D. F. Hochstrasser

"From proteins to proteomes: large scale protein identification by two-dimensional electrophoresis and amino acid analysis." *Biotechnology (N Y)*, **1996**, 14 (1), 61-5.

Wilm, M. and M. Mann

"Analytical properties of the nanoelectrospray ion source." *Anal Chem*, **1996**, 68 (1), 1-8.

Wilm, M., A. Shevchenko, T. Houthaeve, S. Breit, L. Schweigerer, T. Fotsis and M. Mann

"Femtomole sequencing of proteins from polyacrylamide gels by nano-electrospray mass spectrometry." *Nature*, **1996**, 379 (6564), 466-9.

Wilson, R. J., J. L. Goodman and V. B. Strelets

"FlyBase: integration and improvements to query tools." *Nucleic Acids Res*, **2008**, 36 (Database issue), D588-93.

Wischgoll, S., D. Heintz, F. Peters, A. Erxleben, E. Sarnighausen, R. Reski, A. Van Dorsselaer and M. Boll

"Gene clusters involved in anaerobic benzoate degradation of *Geobacter metallireducens*." *Mol Microbiol*, **2005**, 58 (5), 1238-52.

- Wolters, D. A., M. P. Washburn and J. R. Yates, 3rd**
 "An automated multidimensional protein identification technology for shotgun proteomics." *Anal Chem*, **2001**, 73 (23), 5683-90.
- Wu, C. H., R. Apweiler, A. Bairoch, D. A. Natale, W. C. Barker, B. Boeckmann, S. Ferro, E. Gasteiger, H. Huang, R. Lopez, M. Magrane, M. J. Martin, R. Mazumder, C. O'Donovan, N. Redaschi and B. Suzek**
 "The Universal Protein Resource (UniProt): an expanding universe of protein information." *Nucleic Acids Res*, **2006**, 34 (Database issue), D187-91.
- Wu, C. H., L. S. Yeh, H. Huang, L. Arminski, J. Castro-Alvear, Y. Chen, Z. Hu, P. Kourtesis, R. S. Ledley, B. E. Suzek, C. R. Vinayaka, J. Zhang and W. C. Barker**
 "The Protein Information Resource." *Nucleic Acids Res*, **2003**, 31 (1), 345-7.
- Wu, J. Q., D. Shteynberg, M. Arumugam, R. A. Gibbs and M. R. Brent**
 "Identification of rat genes by TWINSCAN gene prediction, RT-PCR, and direct sequencing." *Genome Res*, **2004**, 14 (4), 665-71.
- Wysocki, V. H., G. Tsaprailis, L. L. Smith and L. A. Breci**
 "Mobile and localized protons: a framework for understanding peptide dissociation." *J Mass Spectrom*, **2000**, 35 (12), 1399-406.
- Xie, H. and T. J. Griffin**
 "Trade-off between high sensitivity and increased potential for false positive peptide sequence matches using a two-dimensional linear ion trap for tandem mass spectrometry-based proteomics." *J Proteome Res*, **2006**, 5 (4), 1003-9.
- Xiong, Y., M. J. Chalmers, F. P. Gao, T. A. Cross and A. G. Marshall**
 "Identification of Mycobacterium tuberculosis H37Rv integral membrane proteins by one-dimensional gel electrophoresis and liquid chromatography electrospray ionization tandem mass spectrometry." *J Proteome Res*, **2005**, 4 (3), 855-61.
- Yates, J. R., 3rd, S. Speicher, P. R. Griffin and T. Hunkapiller**
 "Peptide mass maps: a highly informative approach to protein identification." *Anal Biochem*, **1993**, 214 (2), 397-408.
- Zhang, H., E. C. Yi, X. J. Li, P. Mallick, K. S. Kelly-Spratt, C. D. Masselon, D. G. Camp, 2nd, R. D. Smith, C. J. Kemp and R. Aebersold**
 "High throughput quantitative analysis of serum proteins using glycopeptide capture and liquid chromatography mass spectrometry." *Mol Cell Proteomics*, **2005**, 4 (2), 144-55.
- Zhang, Z., D. L. Smith and J. B. Smith**
 "Multiple separations facilitate identification of protein variants by mass spectrometry." *Proteomics*, **2001**, 1 (8), 1001-9.
- Zhou, S. and K. D. Cook**
 "A mechanistic study of electrospray mass spectrometry: charge gradients within electrospray droplets and their influence on ion response." *J Am Soc Mass Spectrom*, **2001**, 12 (2), 206-14.

RESULTATS

Partie I : Développement et applications de méthodologies protéomiques d'aide à l'annotation génomique

Chapitre 1 : Mise en place et application d'une stratégie de protéogénomique

Chapitre 2 : Nouvelle méthode de protéogénomique axée sur la détermination des codons d'initiation des protéines

Chapitre 3 : Application de la stratégie N-TOP pour l'étude d'une enzyme issue d'un organisme dont le génome n'est pas séquencé

Chapitre 4 : Etude méta-protéo-génomique d'un écosystème microbien riche en arsenic

Chapitre 1 : Mise en place et application d'une stratégie de protéogénomique

Depuis la première annotation génomique complète d'un organisme cellulaire (la bactérie *Haemophilus influenzae* en 1995 [Fleischmann et al., 1995]), plus de 950 séquences génomiques ont pu être établies à ce jour (<http://www.ncbi.nlm.nih.gov/genomes/static/gpstat.html>). En revanche, le travail de prédiction des séquences protéiques à l'aide de logiciels bioinformatiques est encore inachevé et demande à être validé car ce processus est très délicat et n'est pas exempt d'erreurs (voir Partie bibliographique, Chapitre 2. 3.).

Les identifications issues de l'analyse protéomique par spectrométrie de masse sont fortement dépendantes de la qualité de l'annotation du génome de l'organisme étudié. En effet, en général, ces identifications sont réalisées par recherche dans la banque de données protéiques issue de l'annotation du génome de l'organisme en question (voir Partie bibliographique. Chapitre 4. 2.1.).

Toutefois, cette connaissance préalable du génome nécessaire pour l'étude du protéome ne nécessite pas obligatoirement que le génome soit annoté. Ainsi, de manière semblable à la recherche de données MS/MS dans un ensemble de séquences protéiques, il est possible d'identifier les gènes codant pour des protéines dans un génome en recherchant les données MS/MS dans une traduction des six cadres de lecture de la séquence génomique. Avec ce type d'approche, les résultats de l'analyse protéomique ne sont plus dépendants de la qualité de l'annotation du génome et au contraire, vont participer à cette annotation. Cette stratégie d'aide à l'annotation génomique par l'utilisation des données de protéomique porte le nom de protéogénomique [Jaffe et al., 2004].

La mise en place de l'approche de recherche des données MS/MS dans le génome complet de l'organisme étudié, point clé d'une stratégie de protéogénomique, sera décrite dans la première partie de ce chapitre. Un premier exemple de stratégie protéogénomique utilisant cette stratégie de recherche originale sera également présenté en deuxième partie du chapitre. La stratégie protéogénomique développée ici a pour objectif d'aider à l'annotation génomique en participant à la correction d'erreurs de séquençage d'un génome.

1. La recherche de données MS/MS dans un génome complet

1.1. Origine de l'approche

L'idée de la recherche des données MS/MS dans des séquences nucléiques est apparue la première fois dans une étude de l'équipe de John R. Yates [Yates et al., 1995] qui a tenté d'intégrer les données issues du séquençage d'ADNc et de génome avec des études biochimiques. Dans cette

approche, la séquence nucléique représentant la banque « génomique » est traduite dans les six cadres de lecture et utilisée pour la recherche des données MS/MS et donc l'identification des gènes. La première analyse protéomique de la bactérie *H. influenzae* par la même équipe [Link et al., 1997] représente la première étude rapportée où une séquence génomique bactérienne complète est utilisée pour la recherche de données MS/MS en vue d'identifier les gènes. La possibilité d'appliquer le même genre d'approche sur des séquences génomiques eucaryotes de grande taille fut démontrée plus tard par Kuster [Kuster et al., 2001] pour le génome d'*Arabidopsis thaliana* et Choudhary [Choudhary et al., 2001] pour le génome humain. Ces études, basées sur les données expérimentales issues des analyses protéomiques, ont permis une annotation plus fiable des génomes correspondants grâce à la confirmation d'un ensemble de gènes prédits, l'identification de nouveaux gènes, la validation d'ORFs hypothétiques et la correction de prédictions de gènes erronées.

1.2. Mise en place de la stratégie

La mise en place de la stratégie de recherche dans un génome complet a été grandement initiée au laboratoire par Christine Carapito et détaillée dans sa thèse [Carapito, 2006]. J'en rappellerai ici brièvement les principaux points qui seront illustrés dans la Figure 1.

1.2.1. Le découpage du génome

Selon l'état d'avancement du projet de séquençage du génome de l'organisme étudié, la séquence génomique peut être disponible sous forme de BACs (Bacterial Artificial Chromosome), de contigs ou de chromosomes partiellement ou complètement assemblés. L'interface graphique de Mascot, le moteur de recherche principalement utilisé au laboratoire, étant limité en nombre d'acides aminés affichables (<10000 acides aminés), les séquences très longues ne sont pas visualisables dans les fenêtres d'affichage des résultats. Par conséquent, un script de découpage des génomes en langage TcL (Tool Command Language) a été développé au laboratoire pour transformer les séquences entières en segments plus courts et ainsi visualiser les résultats. La taille des segments ainsi que les longueurs de recouvrements de séquences sont paramétrables. Le fichier de sortie rassemblant les différents segments nucléotidiques est au format FASTA.

1.2.2. Intégration dans Mascot

La banque de segments générée comme décrit précédemment est intégrée dans Mascot et traduite dans les 6 cadres de lecture automatiquement

1.2.3. Interrogation Mascot et identification des peptides

Les requêtes Mascot sont lancées de manière standard. Les peptides identifiés permettent de mettre en évidence des régions codant pour des protéines sur le génome. Avec une telle banque, les peptides qui sont localisés dans le même cadre de lecture d'une zone génomique commune délimitée par deux codons stop appartiennent à la même protéine.

1.2.4. Identification de la fonction de la protéine par MS-BLAST

Il est ensuite possible de déterminer la fonction de la protéine identifiée. Pour cela, les peptides peuvent être soumis à une interrogation MS-BLAST qui permet d'identifier la protéine analysée si le génome de l'organisme étudié est annoté ou des protéines homologues si ce n'est pas le cas. Si le génome de l'organisme étudié est annoté, la comparaison des séquences peptidiques expérimentales avec les protéines prédites permet de mettre aisément en évidence des erreurs de prédiction ou de valider les prédictions. Un script permettant d'extraire les peptides identifiés et d'automatiser les interrogations MS-BLAST a également été développé au laboratoire.

2. Application de la stratégie de recherche dans les génomes pour la correction d'erreurs de séquençage

Cette étude a été réalisée en collaboration avec l'équipe de bioinformatique du Docteur Olivier Poch du Laboratoire de Biologie et Génomique Structurale à l'Institut de Génétique et de Biologie Moléculaire et Cellulaire (CNRS/INSERM/ULP) de Strasbourg et l'équipe du groupe Avenir du Dr Jean-Marc Reyrat de l'Unité de Pathogénie des Infections Systémiques (INSERM-UMR 7512) de l'Université Paris Descartes.

2.1. Contexte de l'étude

De nos jours, la mise à disposition d'un très grand nombre de génomes complets donne la possibilité aux biologistes d'examiner les fonctions des gènes et des protéines à un niveau sans précédent comme le démontre le grand nombre de projets en protéomique et génomique structurale. Ces études « haut-débit » peuvent être entravées par des limitations techniques, en particulier d'un point de vue automatisation, mais elles dépendent aussi très fortement de la fiabilité des séquences des gènes et des protéines. L'introduction d'erreurs au premier stade de l'analyse génomique, c'est-à-dire le séquençage et la prédiction des gènes peut avoir de lourdes conséquences pour les études ultérieures. Par exemple l'attribution correcte des gènes et des séquences protéiques est cruciale pour la production de protéines fonctionnelles ou pour l'identification des protéines dans les analyses protéomiques par spectrométrie de masse. Si dans les génomes eucaryotes, la qualité de l'annotation est fortement améliorée par les données de transcriptomiques (projets « EST », analyse en série de l'expression des gènes [SAGE], détermination d'épissage alternatif, etc...) [Bianchetti et al., 2005], l'annotation des génomes procaryotes repose essentiellement sur des programmes de prédiction *ab initio*. Donc, soigner la détermination des séquences protéiques prédites (« CoDing Sequence », « CDS ») chez les procaryotes est un investissement très important pour maintenir et améliorer l'utilisation des informations génomiques dans l'ère post-génomique.

Durant l'annotation génomique, la première source d'erreur est la séquence elle-même. Si le séquençage du génome n'est pas à l'origine de la majorité des erreurs, il représente tout de même une source non-négligeable d'erreurs (estimé à 1 erreur pour 10^3 - 10^5 bases [Weinstock, 2000]).

Ces erreurs de séquençage peuvent consister en :

- La substitution d'une base par une autre, ce qui a des conséquences limitées sur la prédiction de gènes si cette substitution n'introduit pas un codon de terminaison (codon stop).
- L'insertion ou la délétion de base(s) produisant un décalage de cadre de lecture dans la région codante (« artificial frameshift »).

L'introduction de codons stop ou de « frameshifts » artificiels engendrent des séquences codantes interrompues (« Interrupted CoDing Sequences », « ICDSs ») et conduisent donc à des erreurs d'annotation.

Dans ce contexte, nous avons développé une stratégie qui combine des techniques de bioinformatique, de re-séquençage génomique et d'analyse protéogénomique pour détecter les « ICDSs », vérifier leur origine et corriger les erreurs qu'ils induisent dans les banques protéiques. Pour cette étude, nous avons utilisé *Mycobacterium smegmatis* mc²155 comme espèce modèle car cette bactérie saprophyte dont le génome est séquencé et annoté est souvent utilisée comme organisme modèle pour l'étude des fonctions de *Mycobacterium tuberculosis*, responsable de la tuberculose.

2.2. Fréquence et origine des « ICDSs »

La plupart des génomes bactériens publiés contiennent des « ICDS » qui représentent entre 1 et 5 % de toutes les séquences codantes prédites dans le génome [Brown et al., 1998; Medigue et al., 1999; Liu et al., 2004; Perrodou et al., 2006]. Les « ICDS » peuvent être présents dans des gènes de fonction connue ou inconnue.

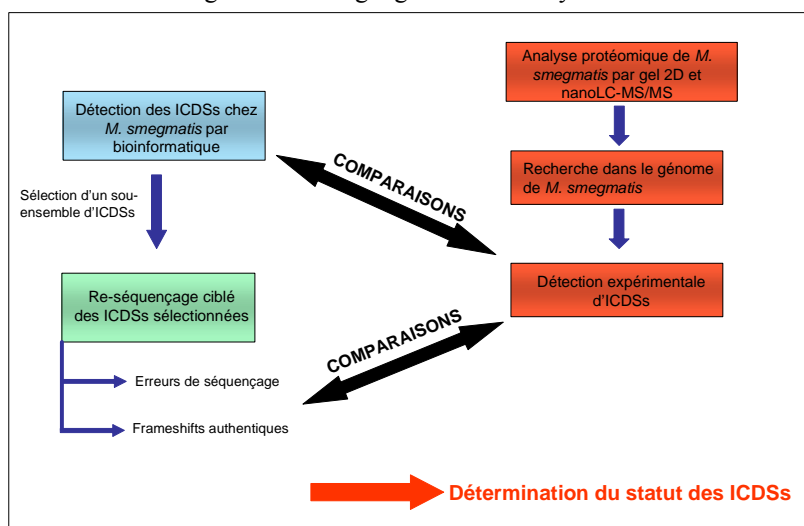
Les « ICDS » peuvent résulter de « frameshifts » artificiels (erreurs de séquençage) mais peuvent aussi être issus de « frameshifts » authentiques. Ce dernier cas correspond à des mutations réelles du génome qui entraînent par exemple la perte d'une fonction protéique [Constant et al., 2002; Perez et al., 2004] ou qui sont compensées par des mécanismes de contournement du « frameshift » et de restauration du cadre de lecture correct (par exemple par glissement du ribosome [Groisman et al., 1995; Wang et al., 2000]).

2.3. Stratégie d'analyse

2.3.1. Stratégie générale

La stratégie générale mise en place pour détecter les « ICDSs », vérifier leur origine et corriger les erreurs de séquençage combine des techniques de bioinformatique, de re-séquençage génomique et d'analyse protéogénomique (Figure 2). Le détail des expériences est décrit dans la publication des résultats.

Figure 2 : Stratégie générale d'analyse

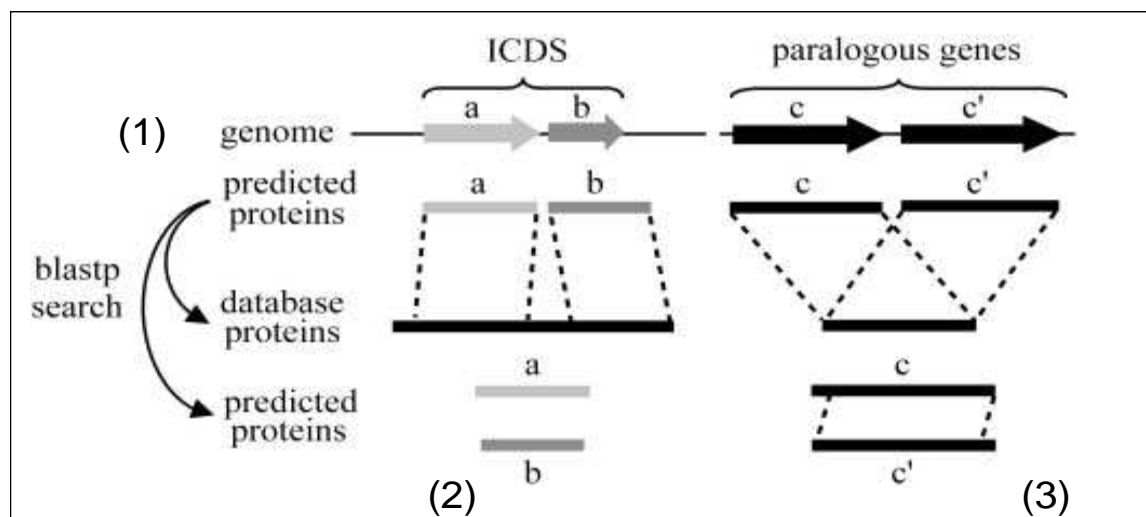


2.3.2. Détection *in silico* et re-séquençage génomique des ICDSs

Une analyse *in silico* du génome de *M. smegmatis* mc²155 a été réalisée pour détecter les ICDSs sur le principe décrit dans [Perrodou et al., 2006] et résumé brièvement en Figure 3.

A partir de l'analyse bioinformatique du génome, un sous-ensemble d'ICDSs est sélectionné (ceux qui ne correspondent pas à des duplications d'ORFs ou qui ne présentent pas de trop haut degré de paralogie) pour re-séquençage ciblé et déduction de la nature des ICDSs (erreurs de séquençage ou « frameshift » authentique).

Figure 3 : Principe de détection bioinformatique des ICDSs. (1) Les couples de protéines présentant au moins un homologue commun sont retenus. Les couples peuvent correspondre à des ICDSs ou à des gènes paralogues adjacents. (2) Si les 2 composants du couple ne présentent pas de similarité significative ($E < 10^{-3}$), ces gènes sont considérés comme des ICDSs. (3) S'il existe une similarité significative, les gènes sont considérés comme paralogues et écartés dans la suite de l'analyse. D'après [Perrodou et al., 2006]



2.3.3. Stratégie protéogénomique

2.3.3.1. Préparation et analyse des échantillons

Les extraits protéiques issus de la culture de *M. smegmatis* mc²155 ont été séparés par gel d'électrophorèse 2D. Les 120 spots majoritaires ont été découpés et digérés à la trypsine. Les peptides issus de la digestion ont ensuite été analysés par nanoLC-MS/MS sur un système nanoHPLC (Agilent Series 1100) couplé à une trappe ionique HCT Ultra (Bruker Daltonics). Les séparations chromatographiques ont été réalisées sur une colonne chromatographique de phase inverse de 75 μm de diamètre interne (colonne Zorbax 300 SB-C18, 15 cm \times 75 μm , particules de 3.5 μm , Agilent Technologies). Le dessalage des échantillons a été réalisé sur une pré-colonne d'enrichissement

(Zorbax 300SB-C18, 5 mm × 0.3 mm, particules de 5 µm, Agilent Technologies). Dans la configuration instrumentale utilisée ici, la colonne analytique est connectée à l'aiguille de nanospray de 20 µm de diamètre interne et 360 µm de diamètre externe (PicoTip Emitter, New Objective, FS360-20-10-CE-20). Le gradient d'élution chromatographique des peptides a été paramétré en tenant compte de la quantité et de la complexité des échantillons à analyser qui sont relativement faibles ici (spots de gel 2D). De même, les paramètres de sélection des ions précurseurs à fragmenter et les cycles d'acquisition des spectres MS et MS/MS ont été finement optimisés. Le paramétrage de l'exclusion dynamique des précurseurs a également été réglé en tenant compte du type d'échantillon analysé. En effet, il s'est avéré sur ce système instrumental que pour l'analyse d'échantillons relativement peu complexes, les meilleurs résultats d'identification étaient obtenus en laissant la possibilité de sélectionner deux fois consécutivement le même ion précurseur pour le fragmenter puis en l'excluant de la sélection automatique MS/MS sur toute la durée de son pic chromatographique.

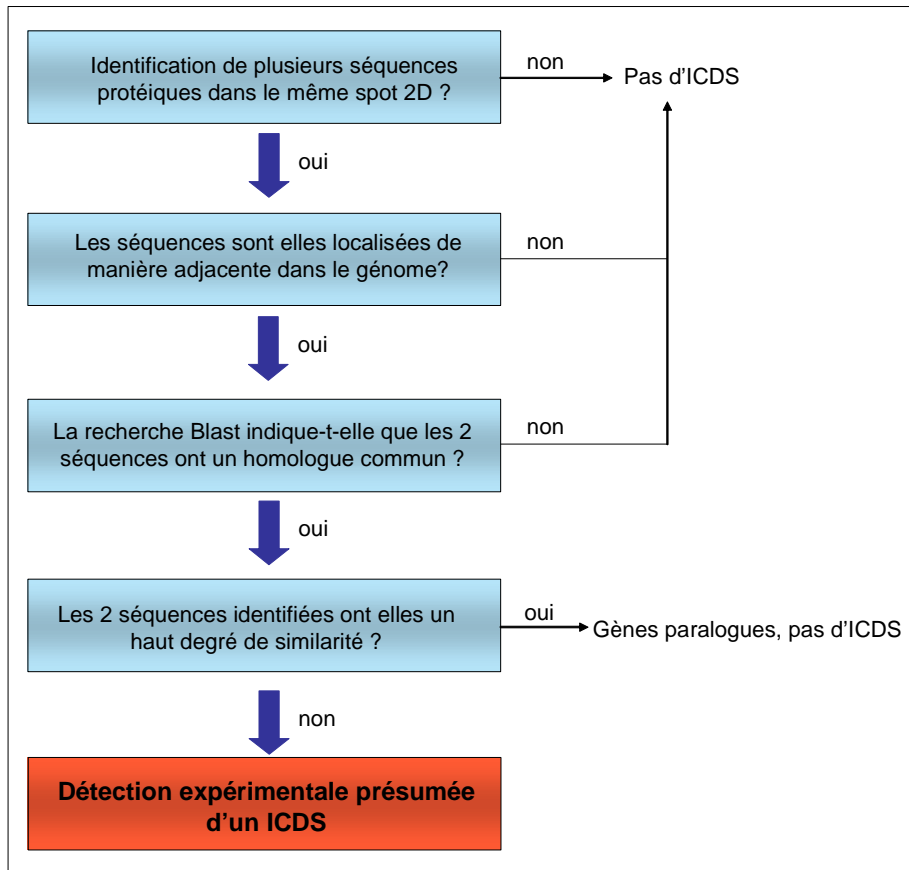
2.3.3.2. Stratégie de recherche dans le génome

La séquence du génome de *M. smegmatis* téléchargée sur le site de TIGR (<http://www.tigr.org/>) a une taille de 6.98 Mbps. Le génome a été découpé en segments de 9480 acides nucléiques soit 3160 acides aminés avec des recouvrements de séquence de 2500 acides nucléiques de part et d'autre. La taille des segments et des recouvrements a été optimisée de façon à éviter les identifications de protéines à la jonction de deux segments et pour pouvoir visualiser les résultats dans Mascot (voir Chapitre 1. 1.2.1.). Cette banque de segments ainsi générée a été importée dans Mascot où elle a été traduite dans les 6 cadres de lecture. Les interrogations Mascot avec les données de nanoLC-MS/MS ont ensuite été réalisées dans cette banque génomique.

2.4. Détection expérimentale des ICDS

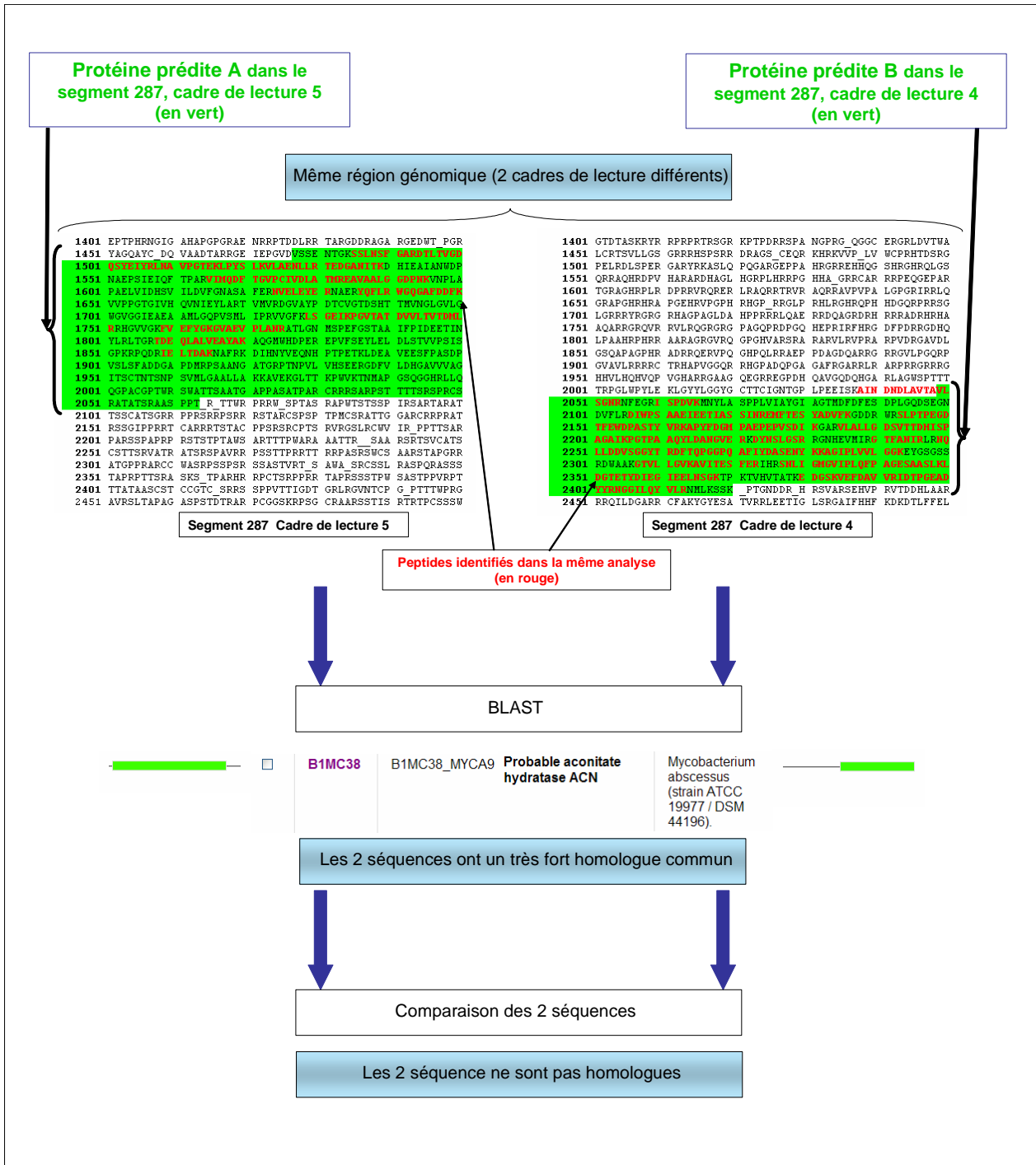
Nous avons choisi d'utiliser une stratégie d'analyse protéomique utilisant le gel 2D pour identifier plus facilement les ICDSs. En effet, comme les protéines ont été séparées par gel 2D, dans l'absolu, une seule séquence protéique (comme définie en Chapitre 1. 1.2.3.) doit être identifiée par spot. Les spots analysés pour lesquels plusieurs séquences protéiques sont identifiées contiennent potentiellement un ICDS. Toutefois, comme l'identification de plusieurs protéines dans le même spot a été largement rapportée en analyse protéomique [Campostrini et al., 2005], ces cas ont été traités prudemment selon la stratégie décrite dans la Figure 4.

Figure 4 : Stratégie de détection expérimentale des ICDSs



Un exemple de détection expérimentale d'un ICDS est illustré en Figure 5.

Figure 5 : Exemple de détection expérimentale d'un ICDS



Dans cet exemple, la recherche de données MS/MS issues de l'analyse d'un spot de gel 2D permet la mise en évidence d'un ICDS potentiel. La requête BLAST sur les 2 séquences protéiques indique un homologue commun et les 2 séquences ne sont pas homologues. L'ICDS déterminé expérimentalement par protéogénomique a aussi été détecté *in silico*. Le re-séquençage génomique ciblé de cet ICDS indique qu'il correspond à une erreur de séquençage du génome. Une nouvelle interrogation Mascot des mêmes données de MS/MS a été réalisée dans la séquence protéique prédite

après correction du génome. Elle a permis d'identifier sur une seule protéine l'ensemble des peptides qui étaient localisés dans deux cadres de lecture différents avant correction. De plus, 2 peptides supplémentaires ont été identifiés ici (Figure 6).

Figure 6 : Résultats de l'interrogation Mascot après correction de la séquence génomique.
(En vert, les nouveaux peptides identifiés)

1	MSSENTGKSS	LNSFGARDTL	TVGDQSYEII	RLNAVPGTEK	LPYSLKVLAE
51	NLLRTEDGAN	ITKDHEAIA	NWDPNAEPSI	EIQFTPARVI	MQDFTGVPCI
101	VDLATMREAV	AALGGDPNKV	NPLAPAEELVI	DHSVILDVFG	NASAFERNVE
151	LEYERNAERY	QFLRWGQGF	DDFKVVPVPGT	GIVHQVNIIEY	LARTVMVRDG
201	VAYPDTCVGT	DSHTTMVNGL	GVLGQVGGI	EAEAAMLGQP	VSMLIPRVVG
251	FKLSGEIKPG	VTATDVVLT	TDMLRRHGTV	GKFEVEYKGG	VAEVPLANRA
301	TLGNMSPEFG	STAAIFPIDE	ETINYLRRTG	RTDEQLALVE	AYAKAQGMWH
351	DPEREPVFESE	YLELDLSTVV	PSISGPKRPQ	DRIELTDAKN	AFRKDIHNYV
401	EQNHPTPETK	LDEAVEESFP	ASDPVSLSFA	DDGAPDMRPS	AANGATGRPT
451	NPVLVHSEER	GDFVLDHGAV	VVAGITSCTN	TSMPSVMLGA	ALLAKKAVEK
501	GLTTKPWVKI	MMAPGSQVVT	DYDQAGLWP	YLERLGYLGG	GYGCTTCIGN
551	TGPLPEEISK	AINDNDLAVT	AVLSGNNRNF	GRISPDVKMN	YLASPPLVIA
601	YGIAGTDFD	FESDPLGQDS	EGNDVFLRDI	WPSAAEIEET	IASSINREMF
651	TESYADVFKG	DDRWRSLPTP	EGDTFEWDP	STYVRKAPYF	DGMPAEPEPV
701	SDIKGARVLA	LLGDSVTTDH	ISPAGAIKPG	TPAAQYLDAN	GVERKDYNL
751	GSRRGHEVM	IRGTFANIRL	RNQLLDDVSG	GYTRDFTQPG	GPQAFIYDAS
801	ENYKKAGIPL	VVLGGKEYGS	GSSRDWAAKG	TVLLGVKAVI	TESFERIHR
851	NLIGMGVIPL	QFPAGESAAS	LKLDGTETYD	IEGIEELNSG	KTPKTVHVT
901	TKEDGSKVEF	DAVVRIDTPG	EADYYRNGGI	LQYVLRNMLK	SSK

2.5. Résultats publiés

Les résultats obtenus dans ce travail ont fait l'objet d'une publication acceptée dans le journal *Genome biology* en février 2007. L'analyse bioinformatique du génome de *M. smegmatis* a permis de détecter en tout 94 ICDSs mais seulement 73 ont été conservés, les 21 autres correspondant à des gènes paralogues ou à des duplications d'ORFs.

Les 73 ICDSs conservés ont pu être re-séquencés. Dans 28 cas, il s'est avéré que les ICDSs correspondaient à des erreurs de séquençage, les 45 ICDS restant devant correspondre à des « frameshifts » authentiques.

L'analyse protéogénomique a permis d'identifier 4 ICDSs expérimentalement. Ces 4 ICDSs correspondaient à des erreurs de séquençage et avaient été détectés *in silico*. L'analyse protéogénomique n'a pas permis d'identifier d'authentiques « frameshifts ».

2.6. Conclusion

Cette étude a permis de montrer que les ICDSs représentaient une source non-négligeable d'erreurs pour les banques protéiques (1.4 % de l'ensemble des protéines prédites chez *M. smegmatis*). Les corrections des erreurs de séquençage ont pu être répercutées sur les séquences protéiques disponibles à la communauté scientifique puisque nous avons soumis ces corrections aux gestionnaires des banques (NCBI, Uniprot). Nous n'avons toutefois pas identifié expérimentalement d'ICDSs

correspondant à des « frameshifts » authentiques. Cela aurait pu nous permettre de déterminer la séquence protéique correspondant réellement exprimée et si des phénomènes de « contournement » de « frameshifts » étaient observés dans notre modèle d'étude.

Interrupted coding sequences in *Mycobacterium smegmatis*: authentic mutations or sequencing errors?

Caroline Deshayes^{*†}, Emmanuel Perrodou[‡], Sebastien Gallien[§], Daniel Euphrasie^{*}, Christine Schaeffer[§], Alain Van-Dorsselaer[§], Olivier Poch[‡], Odile Lecompte[‡] and Jean-Marc Reytrat^{*†}

Addresses: ^{*}Université Paris Descartes, Faculté de Médecine René Descartes, Paris Cedex 15, F-75730, France. [†]Inserm, U570, Unité de Pathogénie des Infections Systémiques-Groupe AVENIR, Paris Cedex 15, F-75730, France. [‡]Laboratoire de Biologie et Génomique Structurales, IGBMC CNRS/INSERM/ULP, BP 163, 67404 Illkirch Cedex, France. [§]Laboratoire de Spectrométrie de Masse Bio-Organique, UMR7178, ECPM, rue Becquerel, Strasbourg, F-67087 cedex 2, France.

Correspondence: Jean-Marc Reytrat. Email: jmreytrat@necker.fr

Published: 12 February 2007

Genome **Biology** 2007, **8**:R20 (doi:10.1186/gb-2007-8-2-r20)

The electronic version of this article is the complete one and can be found online at <http://genomebiology.com/2007/8/2/R20>

Received: 7 September 2006

Revised: 20 November 2006

Accepted: 12 February 2007

© 2007 Deshayes et al.; licensee BioMed Central Ltd.

This is an open access article distributed under the terms of the Creative Commons Attribution License (<http://creativecommons.org/licenses/by/2.0>), which permits unrestricted use, distribution, and reproduction in any medium, provided the original work is properly cited.

Abstract

Background: *In silico* analysis has shown that all bacterial genomes contain a low percentage of ORFs with undetected frameshifts and in-frame stop codons. These interrupted coding sequences (ICDSs) may really be present in the organism or may result from misannotation based on sequencing errors. The reality or otherwise of these sequences has major implications for all subsequent functional characterization steps, including module prediction, comparative genomics and high-throughput proteomic projects.

Results: We show here, using *Mycobacterium smegmatis* as a model species, that a significant proportion of these ICDSs result from sequencing errors. We used a resequencing procedure and mass spectrometry analysis to determine the nature of a number of ICDSs in this organism. We found that 28 of the 73 ICDSs investigated correspond to sequencing errors.

Conclusion: The correction of these errors results in modification of the predicted amino acid sequences of the corresponding proteins and changes in annotation. We suggest that each bacterial ICDS should be investigated individually, to determine its true status and to ensure that the genome sequence is appropriate for comparative genomics analyses.

Background

More than 250 complete bacterial genome sequences are now available, providing unprecedented opportunities for investigating gene and protein functions [1]. The introduction of errors at the first stage of genome sequencing and gene prediction has a major impact on all subsequent studies. One source of errors in genome annotation is the sequence itself.

The development of programs identifying position-specific errors has considerably increased the quality of genomic sequences [2-4]. These errors may introduce stop codons or 'artificial' frameshifts in the coding region that are easily detected by computer-assisted methods [5-7]. Such sequence errors lead to errors in annotation and comparison. An *in silico* survey of the published bacterial genomes shows that

most contain interrupted coding sequences (ICDSs) [5-7]. They occur at low frequency, between 2 and 258 per Mb, not correlated with the size or GC content of the genome. A mean of 74 ICDSs were identified per prokaryotic genome tested [5]. If this is translated into ICDSs per total coding sequences, a figure of 1% to 5% is obtained, with similar figures reported by various independent studies [5,8]. The only notable exception is *Mycobacterium leprae*, which has 30% ICDSs, frequently described as pseudogenes [8]. ICDSs may be present in genes of known or unknown function. A number of bacterial species are known to have developed sophisticated mechanisms for bypassing frameshifts and restoring the correct reading frame, but such mechanisms are unlikely to be general [9,10]. Moreover, the frameshifts bypassed by the ribosome are generally preceded by a unique sequence that can be identified [11]. Thus, the detected ICDSs may either reflect the real genome sequence of the organism, with all the ensuing consequences for the composition of the encoded protein, or they may result from sequencing errors.

We used *M. smegmatis* mc²155 as the model species for this study. This saprophytic bacterium, which is often used as a model organism for studies of *M. tuberculosis* functions, has recently been sequenced [12]. By resequencing the ICDSs of this strain, we show that the genome sequence of this organism contains multiple errors. We systematically corrected the errors, and in all cases, these corrections rendered the predicted protein more similar to its ortholog. We also confirm, by a combined proteome and mass spectrometry analysis, that the sequences of some proteins have been incorrectly predicted due to sequencing errors. However, several ICDSs do correspond to true frameshifts. Authentic frameshifts provide a positive addition to our knowledge and make it possible to investigate gene and protein function, whereas sequencing errors generate false knowledge and confound comparative analyses. We show here that the individual analysis of ICDSs can lead to re-evaluation of the annotation of the genome and the proteome. We suggest that each bacterial ICDS should be investigated individually to ascertain its status and to produce a genome sequence suitable for productive comparative genomics.

Results

ICDSs in *M. smegmatis* mc²155: a resequencing analysis

An *in silico* analysis of the genome of *M. smegmatis* mc²155 revealed that it contains 94 ICDSs [5]. The ICDS database was created using a program based on the analysis of physically adjacent genes to predict putative ICDSs in complete genomes. Briefly, pairs of adjacent genes with at least one common homolog are defined as 'coding sequences (CDSs) containing common hits' and may correspond to a pair of adjacent paralogs or ICDSs. We excluded paralogs from the analysis by searching for sequence similarity between the two 'CDSs containing common hits'. The remaining CDSs are considered to be ICDSs, indicating frameshifts or in-frame stop

codon insertion, due to sequencing errors or authentic events. These 94 ICDSs account for 1.4% of the total coding capacity of this organism. They may result from mutations acquired during evolution or from errors in genome sequencing.

We resequenced the genome of this strain to determine the status of these ICDSs. We did not resequence 21 ICDSs due to the duplication of some open reading frames (ORFs) or high levels of paralogy. The remaining 73 ICDSs were amplified and sequenced on both strands. We compared the nucleotide sequences obtained with the publicly available genome sequence of *M. smegmatis* mc²155. We found that 28 of the 73 ICDSs investigated correspond to sequencing errors (Table 1). These 28 genes containing sequencing errors correspond to 4 errors per megabase in the complete genome. In most cases, correction of the error reunified two adjacent ORFs, resulting in a single ORF rather than the two small ORFs of the original sequence (Figure 1).

Three types of error can be distinguished: miscall, overcall and undercall (Table 1) [2-4]. However, no miscalls (incorrect prediction of a specific nucleotide at a given position) were observed within the 28 sequences containing errors, due to the nature of the program used. The predicted amino acid sequences derived from the corrected nucleotide sequences differed greatly from the original predicted sequences and, in all cases, were systematically more similar to their orthologs. In one case (ICDS0089), the ORF containing the frameshift was not even predicted; the frameshift was probably responsible for the non-assignment of this ORF. The genes affected by the sequencing errors encode proteins of several classes, including 'unknown', 'intermediary metabolism', 'regulation' and 'lipid metabolism' (Table 1). The genes containing frameshifts encode proteins of several classes, including all of those cited above (Table 2). No particular pattern of nucleotides was associated with the 28 sequences containing errors or with the 45 sequences containing frameshifts.

As *M. smegmatis* mc²155 was derived from strain ATCC607, we carried out a comparative analysis of the ICDSs in these two strains. The mc²155 strain was generated from ATCC607 by selection for adaptation to genetic manipulation [13]. The mc²155 strain differs phenotypically from its progenitor (ATCC607) in several ways [13,14]. The frameshifts in mc²155 may well have been acquired recently in the laboratory, due either to counter-selection of pathways of little utility or selection for genetic manipulability. We therefore investigated whether the genes containing frameshifts were acquired before or after the divergence of the two strains. The genome of the ATCC607 strain has not been sequenced, but as both strains belong to the same species (*M. smegmatis*), the sequencing primers originally designed for the mc²155 strain could also be used for the ATCC607 strain. We resequenced the 45 genes containing a frameshift of mc²155 strain in ATCC607 (Table 2). All these genes but one (ICDS0020) also contain a frameshift in the progenitor (ATCC607), suggesting

Table 1**ICDSs shown by resequencing to correspond to sequencing errors in *M. smegmatis* mc²155**

ICDS number	5' position	ORF number	Putative function	Functional classification	Accession number	Type of event
0012	1639371	1547	Hypothetical	Unknown	DQ866846	U
0019	1918521	1842-1843	Adenosylhomocysteinase	Intermediary metabolism	DQ866847	U
0022	1930746	1854-1855	Sodium/proton antiporter	Cell wall, process	DQ866848	U
0024	2055797	1975-1976	Methane/phenol/toluene hydroxylase	Intermediary metabolism	DQ866849	O
0026	2119141	2042	Conserved hypothetical	Unknown	DQ866850	O
0027	2162020	2086-2087	Ferredoxin-NADP reductase	Intermediary metabolism	DQ866851	O
0028	2221312	2149-2150	Hypothetical	Unknown	DQ866852	U
0030	2290855	2215-2216	CoA-transferase	Intermediary metabolism	DQ866853	O
0035	2799279	2732-2733	Conserved hypothetical	Unknown	DQ866854	U
0039	3216877	3151	Aconitate hydratase	Intermediary metabolism	DQ866855	O (× 2)
0040	3262835	3192-3193	Maltooligosyltrehalose synthase	Intermediary metabolism	DQ866856	U
0041	3313327	3240	ABC transporter (CydC)	Intermediary metabolism	DQ866857	O
0051	3902349	3837	Dephospho-CoA kinase	Intermediary metabolism	DQ866858	O (× 2)
0053	3961899	3892-3893	Transcriptional regulator	Regulation	DQ866859	O
0054	4017126	3952-3953	Hypothetical	Unknown	DQ866860	O
0057	4255762	4183	Pyruvate dehydrogenase	Intermediary metabolism	DQ866861	U
0058	4288648	4211-4212	Nitrate reductase	Intermediary metabolism	DQ866862	U
0061	4637174	4539-4540	Oxidoreductase	Intermediary metabolism	DQ866863	O
0072	5644787	5533-5534	Hypothetical	Unknown	DQ866864	U
0073	5855980	5754	Acetyltransferase	Intermediary metabolism	DQ866865	O
0076	6078397	5970-5971	Fatty-acid CoA synthetase	Lipid metabolism	DQ866866	U
0080	6600510	6504-6505	Conserved hypothetical	Unknown	DQ866867	U
0082	6670969	6579	Helicase	DNA metabolism	DQ866868	O
0083	6673489	6581	Hypothetical	Unknown	DQ866869	U
0089	342400	*	Methyltransferase	Intermediary metabolism	DQ866870	U
0091	601272	0511-0512	Hypothetical	Unknown	DQ866871	U
0092	809979	0716-0717	Transcriptional regulator	Regulation	DQ866872	U
0093	428949	1395-1396	Elongation factor G	Translation	DQ866873	O

The nucleotide position, the affected ORF (according to the TIGR website), its putative function computed after the correction of the sequencing errors, its functional classification and its accession number are indicated for each ICDS. The asterisk indicates an ORF not predicted by TIGR. Two types of error were observed: overcall (O), an extra nucleotide not present in the target sequence was initially predicted at a given position; and undercall (U), a nucleotide corresponding to a true target sequence was not predicted at a given position.

that these mutations were acquired before the divergence of the two strains. Thus, the selection of the mc²155 strain and its repeated culture in laboratory conditions had no major impact on frameshift acquisition and pseudogene formation.

Our analysis shows that the genome sequence of *M. smegmatis* mc²155 contains ICDSs, some of which correspond to authentic mutations acquired during evolution, with others resulting entirely from sequencing errors. Our results show that 18 predicted genes do not actually exist in this species (due to fusion of the two ORFs following the correction of the errors) and that one gene was even not predicted in the former sequence, presumably due to these sequencing errors. In all cases, the new predicted genes are actually more similar than previously thought to orthologs in other species.

ICDSs in *M. smegmatis* mc²155: a proteome analysis

As ICDSs (corresponding to authentic events or to sequencing errors) accounted for 1.4% of the ORF content of *M. smegmatis* mc²155, we surveyed a fraction of the proteome to determine the percentage of proteins originating from ORFs not predicted due to misannotations. We carried out two-dimensional electrophoresis of a soluble protein extract. The major spots (120) were excised, digested and analyzed by nano-LC-MS-MS (nanoflow liquid chromatography coupled to tandem mass spectrometry). We were able to identify about 250 proteins unambiguously by comparing the MS-MS data obtained from the tryptic peptides. We compared these MS-MS data directly with public nucleotide sequences, rather than using the classic comparison of MS-MS data with protein sequences [15,16] to prevent the introduction of bias. The identification of several proteins for a single spot is not surprising and has been widely reported in proteomic analysis

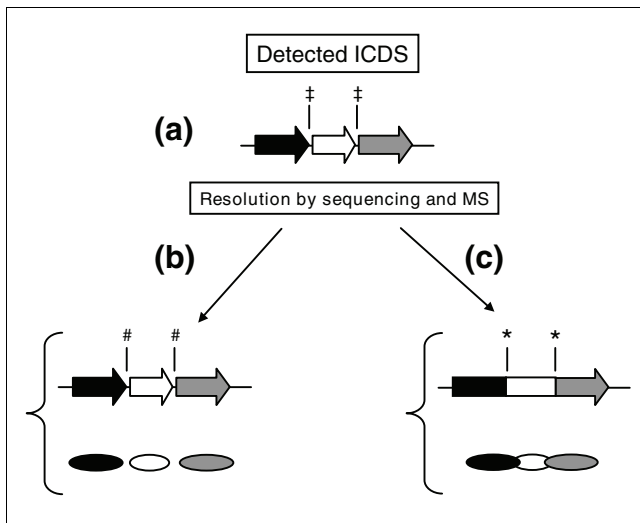


Figure 1

Scheme for ICDS detection and resolution strategy. **(a)** ICDSs are detected within the genome by *in silico* analysis. The double daggers (‡) indicate the regions containing the identified frameshift. Upon resolution by sequencing and mass spectrometry analysis, the ICDSs can be classified as **(b)** true frameshifts or **(c)** sequencing errors. The hash symbol (#) indicates the region of the ORF containing the frameshift. The asterisks (*) indicate sites of corrected sequencing errors resulting in the reconstitution of a full-length ORF. The ORFs are depicted with arrows. The ORF may or may not be in the same frame. Proteins are represented by ellipses.

[17]. For four spots the tryptic peptides identified by nano-LC-MS-MS analysis matched two contiguous hypothetical ORFs each (Table 3, Figure 2). There are two possible explanations for this finding. Firstly, two different proteins, encoded by two different frames in the same genome region, may be present in the same two-dimensional gel electrophoresis spot. This is unlikely, due to differences in molecular masses (Table 3), but cannot be entirely excluded. Secondly, these peptides may be derived from the same protein. In this case, a bypassed stop codon or a sequencing error could account for such an observation.

For the four proteins concerned, MS-BLAST showed that all the tryptic peptides identified matched the same protein on the basis of sequence similarity with other organisms. We carried out a new search with the MS-MS data obtained for the four two-dimensional gel electrophoresis spots using the corrected sequences obtained after resequencing of all the ICDSs. For all four spots the peptides were found to match in the same frame and new peptides from the proteins were detected (Table 3, Figure 2). We can conclude, therefore, that the four ICDSs detected were due to sequencing errors. These ICDSs are ICDS0019, ICDS0039, ICDS0040 and ICDS0093. We show ICDS0040 as an example in Figure 2.

Thus, proteome analysis identified errors in sequences that were not predicted to correspond to an ORF. All four cases detected in this way were found to correspond to sequencing errors (Table 1). There is, therefore, strong congruence between *in silico* data and nucleotide and proteomic analyses.

Discussion

Previous *in silico* analyses have shown that all bacterial species contain ICDSs in their genome [5]. Here, using *M. smegmatis* and two experimentally independent approaches, we show that these ICDSs correspond to authentic mutations and to sequencing errors. By contrast, a recent large-scale proteome analysis (more than 900 proteins) of *M. smegmatis* mc²155 provided no evidence of sequencing errors [18]. Statistically, 16 sequencing errors should have been detected. Possible explanations for this discrepancy are that, by chance, no protein corresponding to an ICDS was extracted, or that proteins in conflict with genomic data were excluded from the analysis.

True frameshifts provide positive information, useful for characterization of the variation of amino acid sequences between various orthologs, whereas sequencing errors introduce noise and create artifactual genetic differences between strains and species. These sequencing errors may result from under-representation of the region in the genomic library or structures making sequencing difficult. Although most genomes have been sequenced with eight-fold coverage (each nucleotide being sequenced eight times), the sequences generated remain a statistical estimation and many regions of low coverage (less than three-fold) still exist in genome sequences [19]. No assembly data are available for the *M. smegmatis* genome project, but the sequencing errors are probably located in such low-coverage regions. In *M. smegmatis* mc²155, 28 of the 73 re-sequenced ICDSs were shown to result from errors. The correction of these errors modified the predicted amino acid sequences of the corresponding proteins. These changes in amino acid sequence increased similarity to orthologs, with consequences for comparative genomics. Unfortunately, it was not possible to associate a particular sequence or stretch of nucleotides with sequence errors. It is, therefore, not possible to predict whether a given ICDS corresponds to an authentic event or to a sequence error. The nature of each ICDS must, therefore, be investigated individually.

Modern biology approaches based on massive sequence comparisons need accurate sequences for meaningful analyses of genetic differences and similarities. Re-sequencing and the correction of errors in genomic sequences are likely to lead to the identification of new protein sequences. For instance, in *M. leprae*, which has a large number of ICDSs in its genome (845), even a small proportion of sequencing errors will provide researchers with substantial numbers of new protein

Table 2**ICDSs shown by resequencing to correspond to authentic mutations in both *M. smegmatis* mc²155 and ATCC607**

ICDS number	5' position	ORF number	Putative function	Functional classification
0003	1169121	1094-1095	Oxidoreductase	Intermediary metabolism
0004	1232918	1164-1165	Arsenic resistance protein	Cell wall, process
0005	1277324	1200-1201	Glycosyltransferase	Intermediary metabolism
0006	1304141	1226-1227	ABC transporter (permease)	Cell wall, process
0007	1508649	1403-1404	Sodium/proton antiporter	Cell wall, process
0008	1510156	1405-1406	Arginine/ornithine antiporter	Cell wall, process
0009	1510156	1405-1407	Arginine/ornithine antiporter	Cell wall, process
0010	1510315	1406-1407	Arginine/ornithine antiporter	Cell wall, process
0011	1545509	1447	Secreted immunogenic protein (Mpt70)	Cell wall, process
0013	1645546	1552-1553	Conserved hypothetical	Unknown
0014	1650143	1557-1558	Hypothetical	Unknown
0015	1669043	1575-1576	Hypothetical	Unknown
0020	1922875	1848-1849	Formate dehydrogenase, alpha subunit	Intermediary metabolism
0021	1924487	1849	Formate dehydrogenase, alpha subunit	Intermediary metabolism
0023	2026072	1949-1950	Hypothetical	Unknown
0025	2097821	2019-2020	Cytochrome P450	Intermediary metabolism
0029	2234814	2164-2165	Substrate-CoA ligase	Lipid metabolism
0033	2557504	2472-2473	Sugar transporter	Cell wall, process
0036	2877071	2816-2817	Two-component system regulator	Cell wall, process
0038	3161135	3097-3098	O-methyltransferase	Intermediary metabolism
0042	3351460	3281-3282	Sugar ABC transporter	Cell wall, process
0043	3410192	3341	Fatty acid desaturase (DesA3)	Lipid metabolism
0044	3442071	3378	Dehydrogenase/reductase	Intermediary metabolism
0045	3471038	3405-3406	Hypothetical	Unknown
0046	3506575	3443-3344	Hypothetical	Unknown
0049	3849109	3785	Conserved hypothetical	Unknown
0052	3930423	3862-3863	Polyprenol-monophosphomannose synthase (Ppm1)	Cell wall, process
0055	4172910	4102-4103	Dehydrogenase	Intermediary metabolism
0059	4551995	4464-4465	Hypothetical	Unknown
0063	5113475	5001	Transporter	Cell wall, process
0064	5127828	5017-5018	Multidrug resistance efflux protein (Tap)	Cell wall, process
0067	5238606	5122-5123	Nitrate reductase (NarX)	Intermediary metabolism
0070	5596138	5488	Conserved hypothetical	Unknown
0071	5639815	5527-5528	Protein-glutamate methylesterase	Intermediary metabolism
0074	6014123	5909-5910	Hypothetical	Unknown
0075	6071755	5963-5964	Integral membrane protein	Unknown
0078	6147983	6046	AraC-family transcriptional regulator	Regulation
0079	6260084	6152-6153	Anion transporter	Cell wall, process
0084	6846273	6761	Oxidoreductase	Intermediary metabolism
0085	6862121	6775	Major facilitator transporter	Cell wall, process
0086	6955671	6870-6871	Glutamine transporter	Cell wall, process
0087	6977889	6889-6890	Thioredoxin	Intermediary metabolism
0088	17247	0017-0018	Hypothetical	Unknown
0094	3456823	*	Dihydrolipoamide dehydrogenase	Intermediary metabolism

The nucleotide position, the affected ORF (according to the TIGR website), its putative function and its functional classification are indicated for each ICDS. The asterisk indicates an ORF not predicted by TIGR.

sequences, making it possible to identify new functional genes, or to develop new serological tests.

Other mycobacterial species also contain ICDSs in their

Table 3

ICDSs shown by nano-LC-MS-MS analysis to correspond to sequencing errors in *M. smegmatis* mc²155

ICDS number	Affected ORF	Calculated mass before correction	Calculated mass after correction
0019	1842-1843	45,980-7,370	53,460
0039	3151	64,570	101,200
0040	3192-3193	48,730-33,880	83,490
0093	1395-1396	21,560-63,800	77,220

The affected ORFs (according to the TIGR website) and their predicted molecular weights before and after genomic correction are indicated.

genome, some of which have been shown to correspond to authentic mutations acquired during evolution. For instance, the genomes of *M. tuberculosis* H37Rv, *M. tuberculosis* CDC1551 and *M. bovis* contain 96, 123 and 111 ICDSs, respectively, corresponding to about 2% of total gene content in each case [5]. Interestingly, a number of ICDSs corresponding to authentic events have been fortuitously characterized. In several cases it has been shown that these events inactivate the gene. For instance, ICDS0066 of *M. tuberculosis* H37Rv, corresponding to a gene encoding a polyketide synthase (*pkc1*), includes a frameshift, generating two distinct ORFs, *pkc1* and *pkc15*. In contrast, *M. bovis* and *M. leprae* carry a *pkc1* gene with no frameshift. The complementation of *M. tuberculosis* with the *pkc1* of *M. bovis* leads to the synthesis of a new metabolite, phenolphthiocerol [20]. Thus, *M. tuberculosis* has clearly lost the ability to synthesize phenolphthiocerol due to a frameshift within the *pkc1* gene. Another example is ICDS0067 in *M. bovis*, which occurs in a sequence encoding a putative glycosyltransferase. The ortholog of this gene has no frameshift in *M. tuberculosis* (Rv2958) [21]. The complementation of *M. bovis* BCG with Rv2958 from *M. tuberculosis* leads to the accumulation of a new product in this strain: diglycosylated phenolglycolipid [21]. Thus, *M. bovis* has lost the ability to metabolize the diglycosylated phenolglycolipid due to the frameshift within the glycosyltransferase gene.

These two examples, taken from published work, illustrate that, as expected, a frameshift within ORF may lead to a loss of function. It should be noted that the genes for which function has been lost (such as *pkc1* or Rv2958) have been split into only two pieces and could, therefore, theoretically revert to the wild-type allele with ease. These genes containing frameshifts are in the process of becoming pseudogenes (pseudogenization) but need to acquire additional mutations before they are fixed, leading to an almost irreversible loss of function.

The conclusion of this work may be extended to most, if not all, bacterial genomes sequenced to date. These findings have major implications for comparative genomics. Firstly, the resolution of sequencing errors reduces protein variability,

facilitating the precise definition of module composition and function. Secondly, as ICDSs corresponding to authentic mutations probably lead to a loss of protein function, the choice of strain or species is of particular importance for investigations of the function of a particular gene. Researchers should carefully consider their investment before creating mutants in these ORFs or producing the corresponding polypeptides. It should be noted that a small number of ORFs containing frameshifts may retain their function or even lead to the acquisition of a new function. It would be interesting to re-frame these ORFs to evaluate the impact on protein function.

We have shown here that 28 of the 73 ICDSs resulted from sequencing errors. It seems highly likely that all sequenced genomes contain ICDSs resulting from sequencing errors. The current ICDS database contains more than 6,600 ICDSs (in 120 genomes) awaiting characterization. In this study, we detected sequencing errors at a rate of 4 per megabase. The calculated number of ICDSs is obviously an underestimate of the reality as some events such as fusion or fission that maintain the correct frame are not detected by the algorithm used [5].

Very few articles have dealt with sequence fidelity. TIGR has reported an error rate for finished genomes of 1 in 88,000 nucleotides [22,23] whereas Weinstock [19] estimated that the frequency of error was between 10⁻³ and 10⁻⁵. The frequency of errors clearly depends on the chemical system used and the research centers carrying out the sequencing work [24]. The development of error prediction programs has greatly helped to reduce the error rate [2-4]. However, as shown in this study, sequencing errors are clearly a persistent problem in genomic databases. The major problem is that the bioinformaticians who assemble genomes have, for years, discarded precious information about how all the individual sequence fragments align on the assembled chromosome. The only way to test the nature of the ICDSs is to re-sequence the fragment. The NCBI has recently developed the 'Assembly Archive', which stores records of both the way in which a particular assembly was constructed and alignments of any set of traces to a reference genome [25]. This resource makes it pos-

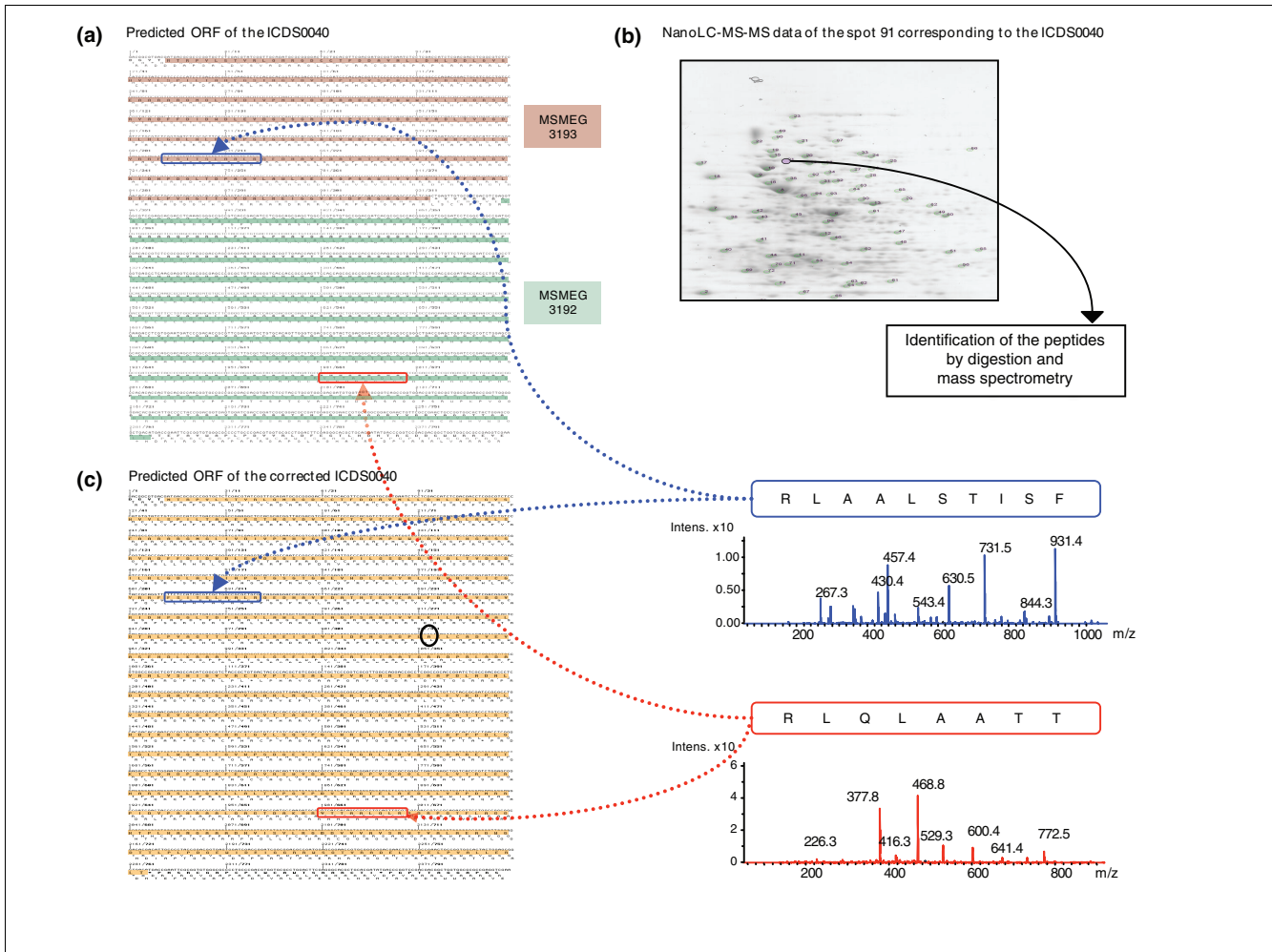


Figure 2
 Comparison of genomic prediction with proteomic results (example of ICDS0040). **(a)** Representation of the DNA region and its predicted ORFs (in color). **(b)** Detailed view of the two-dimensional gel. Nano-LC-MS-MS data are obtained after extraction and digestion of the protein. The matching peptides are boxed in the translated genomic sequence (a,c). **(c)** Representation of the DNA region and its predicted ORF upon correction of the sequencing errors (depicted in the ellipse). Correction of the sequencing errors reassociates the two peptides to give a single protein, accounting for their appearance at a single spot.

sible to determine whether an ICDS corresponds to a region of low coverage and to evaluate the quality of the raw data. It would clearly be easier to resolve the ICDSs in various genomes if all the sequencing centers made complete assembly data available.

Materials and methods

Bacterial strains

M. smegmatis mc²155 (ATCC700084) and *M. smegmatis* NRRL B-692 (Trevisan) Lehman and Neumann (ATCC607) were purchased from the American Type Culture Collection (Manassas, Virginia, USA).

ICDS detection in *M. smegmatis* mc²155

The genome sequence of *M. smegmatis* mc²155 was taken from the TIGR website [12]. The ICDSs were detected using the method developed by Perrodou *et al.* [5].

Primer design and sequence analysis

The primers used to sequence frameshifts were designed as previously described [5] using an optimized version of the CADO4MI program (Computed Assisted Design of Oligonucleotides for Microarray). It is a freeware (GNU General Public License) accessible online [26]. For each genome, sequencing primers are available online [27]. The chromosomal DNA of the mc²155 and ATCC607 strains of *M. smegmatis* used for PCR amplification was purified as previously described [28]. Pairs of primers were used for amplification with Pfu Turbo DNA polymerase (Stratagene, La Jolla, CA, USA). PCR samples were run on a 0.8% agarose gel and the

fragments were excised from the gel and purified using the QIAquick Gel purification kit (Qiagen Chatsworth, CA, USA). The PCR fragments had a mean length of 300 base-pairs. Purified PCR fragments were used as templates in sequencing reactions with each primer used for PCR amplification. The nucleotide and inferred amino acid sequences were analyzed with DNA Strider [29]. Three independent amplicons were sequenced for each ICDS.

Protein extraction and two-dimensional gel electrophoresis

M. smegmatis strain mc²155 (1 liter) was grown in M9 minimal medium (Difco, Detroit, USA) for 5 days and then centrifuged. Bacterial pellets were used for two-dimensional electrophoresis. Unless otherwise specified, all chemicals were obtained from Sigma (St Louis, MO, USA). Dithiothreitol (DTT) and iodoacetamide were obtained from Fluka (Buchs, Switzerland). The pellet fraction was incubated with extraction buffer (50 mM Tris, pH 7.5, 1 mM phenylmethylsulfonyl fluoride, 1 mM EDTA, 1 mM DTT, protease inhibitor mixture (complete from Roche, Basel, Switzerland)) for 45 minutes at 4°C. The mixture was sonicated for a few seconds and its protein concentration determined by Bradford assay. The solvent of the protein extract was evaporated off and the protein residue was suspended in rehydration buffer (8 M urea, 2 M thiourea, 4% 3-[(3-cholamidopropyl)dimethylammonio]-1-propanesulfonic acid, 0.5% Triton X-100, 1% DTT, 20 mM spermine, 2% Pharmalyte (Amersham Pharmacia Biotech, Piscataway, NJ, USA)). The sample was incubated for 30 minutes at 20°C and centrifuged at 15,000 rpm at 20°C.

Protein extract was run on a strip of gel of pH range 3 to 10 (Bio-Rad Laboratories, Hercules, CA, USA) for 15 h at 20°C under 50 V in a PROTEAN isoelectric focusing cell (Bio-Rad). Isoelectric focusing was carried out with several voltage steps: 1 h at 200 V, then 4 h at 1,000 V followed by 16 h at 5,000 V and finally 7 h at 500 V at 20°C. The strips were incubated for 30 minutes at 20°C in electrophoresis buffer (50 mM Tris-HCl, pH 8.8, 6 M urea, 30% (v/v) glycerol, 2% (w/v) SDS, and 1% DTT), followed by 30 minutes in the same buffer supplemented with 2.5% iodoacetamide. Electrophoresis in a gradient gel (5% to 20% acrylamide) on a PROTEAN II (Bio-Rad) apparatus at 5 mA for 1 h and 10 mA overnight was used as the second dimension. The gel was stained with Colloidal blue (G260, Sigma); 120 spots were selected by visual inspection and gel slices were excised with a Proteineer SP automated spot picker (Bruker Daltonics, Bremen, Germany) according to the manufacturer's instructions.

Mass spectrometry

The two-dimensional gel spots were excised, washed, destained, reduced, alkylated and dehydrated for in-gel digestion of the proteins with an automated protein digestion system, MassPREP Station (Waters, Milford, MA, USA). The proteins were digested overnight at room temperature with

trypsin. They were then extracted with 60% (v/v) acetonitrile in 5% (v/v) formic acid and then with 100% acetonitrile. The resulting peptide extracts were analyzed directly by nano-LC-MS-MS on an Agilent 1100 Series capillary LC system (Agilent Technologies, Palo Alto, USA) coupled to an HCT Ultra ion trap (Bruker Daltonics). This instrument was equipped with a nanospray ion source and chromatographic separation was carried out on reverse phase (RP) capillary columns (C18, 75 µm id, 15 cm length, Agilent Technologies) with a flow rate of 200 nl/minute. The voltage applied to the capillary cap was optimized to -2,000 V. MS-MS scanning mode was performed in the Ultra Scan resolution mode at a scan rate of 26,000 m/z per second. Eight scans were averaged to obtain an MS-MS mass spectrum. The complete system was fully controlled by Agilent ChemStation and EsquireControl (Bruker Daltonics) software. The generated peak-lists of fragments were used for public *M. smegmatis* genome database searches.

Acknowledgements

Data were obtained from TIGR from their website [30]. We thank INSERM for funding this project through an Avenir program grant to JMR, Chargé de Recherches at INSERM. This work was also funded by a 'Protéomique et Génie des Protéines' grant (project no. PGP 04-013), the RNG (Réseau National de Génopoles) Strasbourg Bioinformatics Platform infrastructures and EVI-GENORET (LSHG-CT-2005-512036). CD is funded by a doctoral grant from INSERM - Région Ile de France. We thank E Stewart for critical reading and correcting the English of this manuscript.

References

- Bernal A, Ear U, Kyrpides N: **Genomes OnLine Database (GOLD): a monitor of genome projects world-wide.** *Nucleic Acids Res* 2001, **29**:126-127.
- Lawrence CB, Solovyev VV: **Assignment of position-specific error probability to primary DNA sequence data.** *Nucleic Acids Res* 1994, **22**:1272-1280.
- Ewing B, Green P: **Base-calling of automated sequencer traces using phred. II. Error probabilities.** *Genome Res* 1998, **8**:186-194.
- Ewing B, Hillier L, Wendl MC, Green P: **Base-calling of automated sequencer traces using phred. I. Accuracy assessment.** *Genome Res* 1998, **8**:175-185.
- Perrodou E, Deshayes C, Muller J, Schaeffer C, Van Dorsselaer A, Ripp R, Poch O, Reyrat JM, Lecompte O: **ICDS database: interrupted CoDing sequences in prokaryotic genomes.** *Nucleic Acids Res* 2006, **34**:D338-343.
- Brown NP, Sander C, Bork P: **Frame: detection of genomic sequencing errors.** *Bioinformatics* 1998, **14**:367-371.
- Medigue C, Rose M, Viari A, Danchin A: **Detecting and analyzing DNA sequencing errors: toward a higher quality of the *Bacillus subtilis* genome sequence.** *Genome Res* 1999, **9**:1116-1127.
- Liu Y, Harrison PM, Kunin V, Gerstein M: **Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes.** *Genome Biol* 2004, **5**:R64.
- Wang G, Ge Z, Rasko DA, Taylor DE: **Lewis antigens in *Helicobacter pylori*: biosynthesis and phase variation.** *Mol Microbiol* 2000, **36**:1187-1196.
- Groisman I, Engelberg-Kulka H: **Translational bypassing: a new reading alternative of the genetic code.** *Biochem Cell Biol* 1995, **73**:1055-1059.
- Gurvich OL, Baranov PV, Zhou J, Hammer AW, Gesteland RF, Atkins JF: **Sequences that direct significant levels of frameshifting are frequent in coding regions of *Escherichia coli*.** *EMBO J* 2003, **22**:5941-5950.
- Mycobacterium smegmatis* mc² 155 Genome Page** [<http://>

- cmr.tigr.org/tigr-scripts/CMR/GenomePage.cgi?database=gms]
13. Snapper SB, Melton RE, Mustafa S, Kieser T, Jacobs WR Jr: **Isolation and characterization of efficient plasmid transformation mutants of *Mycobacterium smegmatis***. *Mol Microbiol* 1990, **4**:1911-1919.
 14. Etienne G, Villeneuve C, Billman-Jacobe H, Astarie-Dequeker C, Dupont MA, Daffe M: **The impact of the absence of glycopeptidolipids on the ultrastructure, cell surface and cell wall properties, and phagocytosis of *Mycobacterium smegmatis***. *Microbiology* 2002, **148**:3089-3100.
 15. Bradshaw RA: **Revised draft guidelines for proteomic data publication**. *Mol Cell Proteomics* 2005, **4**:1223-1225.
 16. Steen H, Mann M: **The ABC's (and XYZ's) of peptide sequencing**. *Nat Rev Mol Cell Biol* 2004, **5**:699-711.
 17. Camprostrini N, Areces LB, Rappsilber J, Pietrogrande MC, Dondi F, Pastorino F, Ponzoni M, Righetti PG: **Spot overlapping in two-dimensional maps: a serious problem ignored for much too long**. *Proteomics* 2005, **5**:2385-2395.
 18. Wang R, Prince JT, Marcotte EM: **Mass spectrometry of the *M. smegmatis* proteome: protein expression levels correlate with function, operons, and codon bias**. *Genome Res* 2005, **15**:1118-1126.
 19. Weinstock GM: **Genomics and bacterial pathogenesis**. *Emerg Infect Dis* 2000, **6**:496-504.
 20. Constant P, Perez E, Malaga W, Laneelle MA, Saurel O, Daffe M, Guilhot C: **Role of the *pks15/1* gene in the biosynthesis of phenolglycolipids in the *Mycobacterium tuberculosis* complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the *pks15/1* gene**. *J Biol Chem* 2002, **277**:38148-38158.
 21. Perez E, Constant P, Lemassu A, Laval F, Daffe M, Guilhot C: **Characterization of three glycosyltransferases involved in the biosynthesis of the phenolic glycolipid antigens from the *Mycobacterium tuberculosis* complex**. *J Biol Chem* 2004, **279**:42574-42583.
 22. Read TD, Salzberg SL, Pop M, Shumway M, Umayam L, Jiang L, Holtzapple E, Busch JD, Smith KL, Schupp JM, et al.: **Comparative genome sequencing for discovery of novel polymorphisms in *Bacillus anthracis***. *Science* 2002, **296**:2028-2033.
 23. Fleischmann R: **Single nucleotide polymorphisms in *Mycobacterium tuberculosis* structural genes**. *Emerg Infect Dis* 2001, **7**:487-488.
 24. Richterich P: **Estimation of errors in "raw" DNA sequences: a validation study**. *Genome Res* 1998, **8**:251-259.
 25. Salzberg SL, Church D, DiCuccio M, Yaschenko E, Ostell J: **The genome Assembly Archive: a new public resource**. *PLoS Biol* 2004, **2**:E285.
 26. **Computed Assisted Design of Oligonucleotides for Microarray** [<http://bips.u-strasbg.fr/CADO4MI/>]
 27. **ICDS Database** [<http://alnitak.u-strasbg.fr/ICDS/>]
 28. Pelicic V, Reyrat JM, Gicquel B: **Generation of unmarked directed mutations in mycobacteria, using sucrose counterselectable suicide vectors**. *Mol Microbiol* 1996, **20**:919-925.
 29. Marck C: **'DNA Strider': a 'C' program for the fast analysis of DNA and protein sequences on the Apple Macintosh family of computers**. *Nucleic Acids Res* 1988, **16**:1829-1836.
 30. **The Institute for Genomic Research** [<http://www.tigr.org>]

Chapitre 2 : Nouvelle méthode de protéogénomique axée sur la détermination des codons d'initiation des protéines

Le séquençage du génome d'un organisme est une source d'erreur non-négligeable lors de la constitution des banques protéiques dédiées à cet organisme. Toutefois, ce type d'erreur reste mineur en comparaison aux erreurs liées à la prédiction des gènes [Galperin et al., 1998]. Ces prédictions *in silico* peuvent présenter différents types d'erreurs (détaillé dans la Partie bibliographique, Chapitre 2.3.1.2.).

Un premier type d'erreur concerne l'existence biologique de la protéine prédite. Les gènes « hypothétiques » représentent habituellement environ 30 à 50 % de tous les gènes prédits pour un génome. Or, en l'absence de la caractérisation expérimentale de la protéine, aucune preuve ne permet de valider son existence. Certaines protéines présentes dans les banques ne sont en réalité pas exprimées et leur présence dans les banques résulte donc du phénomène de « sur-prédiction ». L'identification par analyse protéomique de protéines dénommées « hypothetical » dans les banques permet de valider leur expression. Plusieurs études protéogénomiques ont permis de valider l'expression de ce type de protéines [Lipton et al., 2002; Kolker et al., 2004; Hixson et al., 2006]. A l'inverse, certaines protéines réellement exprimées dans l'organisme ne sont pas prédites lors du processus d'annotation, on parle de phénomène de « sous-prédiction » [Overbeek et al., 2007]. Plusieurs études protéogénomiques ont permis d'identifier des gènes qui avaient été « ratés » lors du processus d'annotation [Jungblut et al., 2001; Jaffe et al., 2004].

Un deuxième type d'erreur, la prédiction incorrecte des codons d'initiation des protéines dans les génomes procaryotes, est sans aucun doute l'erreur la plus répandue dans les banques protéiques puisqu'une étude publiée durant ma thèse a montré que le taux d'erreur dans la prédiction des codons d'initiation pouvait varier de 10 % à 44 % chez *Halobacterium salinarum* et *Natromonas pharaonis* en fonction du programme de prédiction de gènes utilisé [Aivaliotis et al., 2007].

Ce chapitre sera consacré au développement et à l'application d'une stratégie de protéogénomique visant à corriger les erreurs de prédiction des gènes avec une attention particulière portée sur les codons d'initiations de ces gènes.

1. Développement d'une nouvelle stratégie protéomique « orientée » N-terminale (« N-Terminal Oriented Proteomic », « N-TOP »)

L'utilisation d'une stratégie de protéogénomique basée sur la recherche de données MS/MS dans le génome permet de détecter des peptides localisés dans des régions génomiques en amont des gènes prédits et donc de corriger les extrémités N-terminales des gènes concernés.

Toutefois, la faible couverture obtenue sur les séquences protéiques dans les études haut-débit résulte généralement en une faible couverture des peptides N-terminaux. En plus, si cette stratégie peut permettre de détecter les séquences prédites trop courtes, elle ne permet pas facilement de localiser exactement le réel codon d'initiation des protéines ni de mettre en évidence les séquences prédites trop longues. Une stratégie de détermination expérimentale efficace des codons d'initiation des protéines implique l'utilisation d'une technique permettant la mise en évidence des peptides N-terminaux des protéines et donc la caractérisation de l'extrémité N-terminale des protéines.

1.1. Les stratégies de détermination expérimentale des codons d'initiation des protéines

1.1.1. Les stratégies d'identification des codons d'initiation

Plusieurs études protéogénomiques visant à caractériser les extrémités N-terminales des protéines pour valider/corriger les codons d'initiation correspondant ont été rapportées dans la littérature au cours de ma thèse. La plupart de ces études [Rison et al., 2007; de Souza et al., 2008; Gupta et al., 2008] utilisaient les techniques dites « classiques » de la protéomique. Dans ce cas, l'extrémité N-terminale des protéines ne subit pas de marquage chimique particulier et est identifiée après digestion trypsique par un peptide « semi-trypsique » (l'acide aminé N-terminal du peptide ne suit pas un acide aminé arginine ou lysine comme c'est le cas pour les peptides internes de digestion trypsique).

Ce critère n'est pas suffisant pour affirmer que ce peptide est bien le peptide N-terminal de la protéine car bien que la trypsine soit décrite comme une enzyme hautement spécifique [Olsen et al., 2004], l'identification de peptides semi-trypsiques est courante dans les études protéomiques [Picotti et al., 2007]. Même si généralement lors de l'analyse LC-MS/MS ces peptides semi-trypsiques correspondent à des signaux très peu intenses, ils peuvent tout de même être séquencés et identifiés, en particulier pour les protéines les plus abondantes. Un peptide semi-trypsique peut également être issu de la digestion endogène d'une protéine (par exemple la coupure d'un peptide signal) et donc correspondre au peptide N-terminal de la protéine tronquée et non à celui de la protéine native. Dans ce cas, il ne permet pas de corriger l'attribution du codon d'initiation. Seul le peptide N-terminal de la protéine « native » est utile à cette fin.

En général, pour s'affranchir de ces différents problèmes qui pourraient engendrer des conclusions erronées, les peptides semi-trypsiques considérés comme permettant de valider/corriger les codons d'initiation des protéines sont ceux :

- Pour lesquels aucun autre peptide identifié n'est situé en amont dans la séquence génomique jusqu'au codon stop le précédent dans le même cadre de lecture.
- Dont le premier acide aminé est codé par un codon d'initiation, ATG, GTG ou TTG, ou par le premier codon suivant un codon d'initiation (dans le cas d'une coupure de la méthionine N-terminale de la protéine).

Même si le risque de considérer un peptide par erreur comme le peptide N-terminal de la forme entière d'une protéine n'est pas totalement exclu, il est tout de même fortement limité par l'utilisation de ces critères.

Toutefois, cette stratégie ne permet pas une couverture satisfaisante des peptides N-terminaux. Pour améliorer l'identification des peptides N-terminaux, plusieurs méthodes ont été développées.

1.1.2. Les stratégies d'enrichissement en peptides N-terminaux

Plusieurs stratégies d'enrichissement des peptides N-terminaux ont été rapportées dans la littérature avant et au cours de ma thèse. Je présenterai ici les stratégies principales [Gevaert et al., 2003; McDonald et al., 2005; Dormeyer et al., 2007; Shen et al., 2007; Yamaguchi et al., 2007; Staes et al., 2008; Yamaguchi et al., 2008].

Une première méthode pour l'enrichissement des peptides N-terminaux naturellement « bloqués » (le plus généralement par N-acétylation, qui est fréquente chez les eucaryotes mais très rare chez les procaryotes [Falb et al., 2006]) ne passant pas par la dérivation chimique est la chromatographie liquide d'échange de cations. En effet, comme les peptides N-terminaux n'ont pas (ou peu, en fonction de l'enzyme de digestion utilisée) de charges positives en condition acide, ils sont élués dans la fraction chromatographique à faible concentration en sels [Dormeyer et al., 2007].

Un panel non exhaustif des autres stratégies principales d'enrichissement des peptides N-terminaux est présenté en Figure 1 et brièvement décrit ici :

- La méthode développée au laboratoire de J. Vandekerckhove est sans doute la méthode de référence dans ce domaine. Cette méthode consiste à isoler les peptides N-terminaux par une chromatographie diagonale des fractions collectées lors d'une première étape chromatographique [Gevaert et al., 2003; Staes et al., 2008]. Cette méthode appelée COFRADIC (COmbined FRActional Diagonal Chromatography) comprend 4 étapes : 2 réactions chimiques et 2 chromatographies. Dans la première étape les amines libres des protéines (N-terminale et chaîne latérale des lysines) sont bloquées par acétylation et les protéines sont digérées à la trypsine. Les peptides générés par la digestion sont ensuite séparés et collectés par RP-HPLC. Chaque fraction collectée est traitée avec le « 2,4,6-trinitrobenzenesulfonic acid » (TNBS) qui réagit avec l'amine libre N-terminal des peptides internes. Comme cette modification induit une augmentation de l'hydrophobicité des peptides, ils sont

ensuite séparés des peptides N-terminaux acétylés ou naturellement bloqués lors de la deuxième étape de chromatographie. Des améliorations de cette méthode ont été introduites récemment avec de nouveaux traitements enzymatiques [Staes et al., 2008]. Cette méthode a prouvé son efficacité sur une étude à grande échelle en combinaison avec des méthodes de protéomique plus classiques [Aivaliotis et al., 2007]. Les autres méthodes exposées par la suite n'ont pas été appliquées à notre connaissance pour la détermination à grande échelle des peptides N-terminaux de protéines natives.

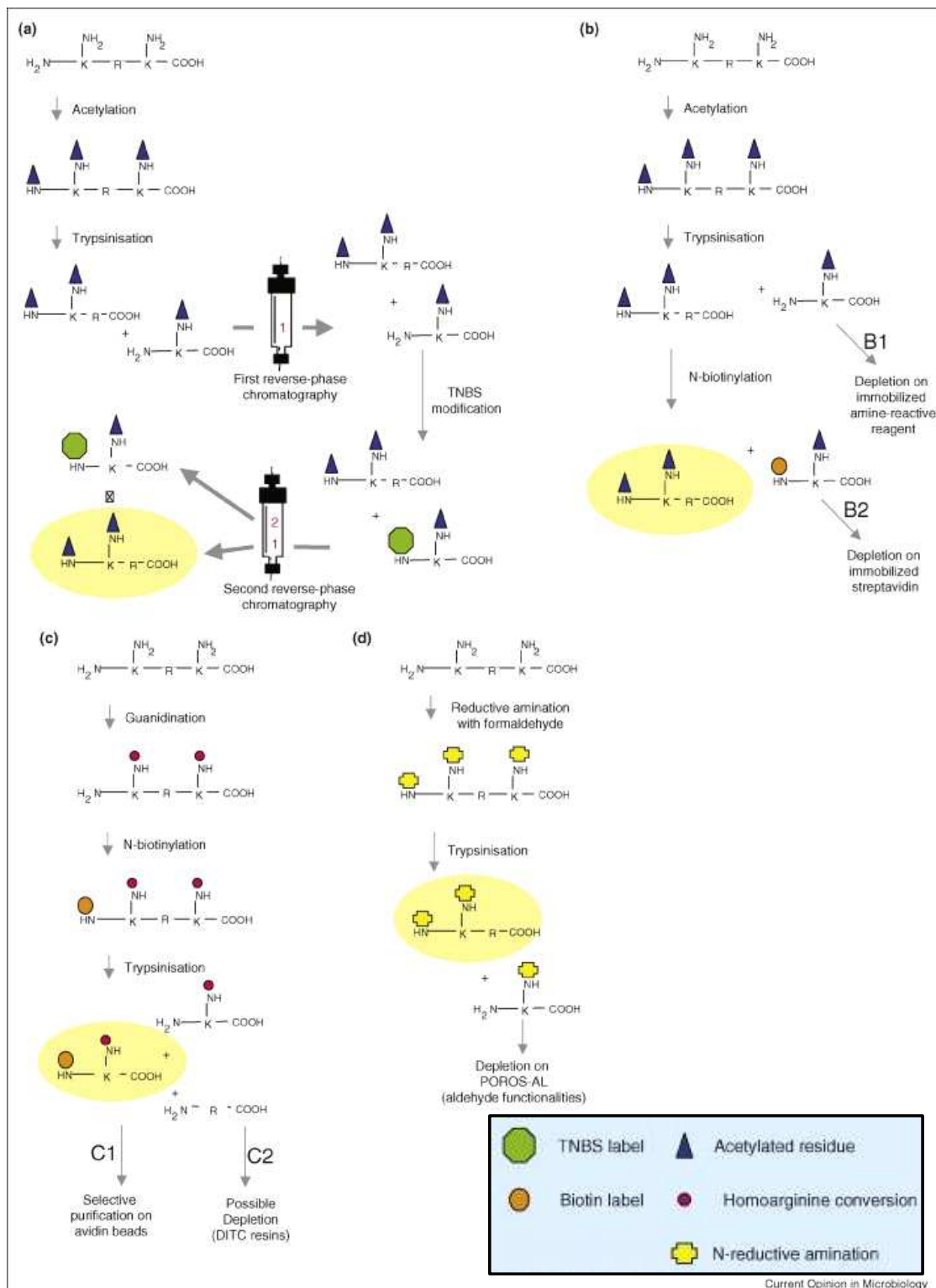
- Une deuxième méthode consiste à acétyler les amines libres des protéines avant digestion à la trypsine puis à biotinyler les amines libres des peptides internes et à capturer les peptides internes portant un groupement biotine [McDonald et al., 2005].

- Une autre méthode repose sur la guanidination des lysines suivie de la N-biotinylation des protéines puis de la digestion tryptique. La purification sélective des peptides N-terminaux est réalisée grâce à des billes d'avidine [Yamaguchi et al., 2007]. La déplétion des peptides internes est une alternative possible [Yamaguchi et al., 2008].

- Une quatrième méthode dénommée « Dimethyl Isotope Coded Affinity Selection (DICAS) » consiste à modifier les amines libres des protéines avec le formaldéhyde. Les peptides internes obtenus après digestion sont capturés sur des supports solides présentant des fonctions aldehyde. Les peptides restant en solution sont principalement des peptides N-terminaux.

Toutes ces stratégies d'enrichissement passent par un marquage N-terminal des protéines qui pourra être utilisé comme un indice supplémentaire pour l'identification des codons d'initiation des protéines.

Figure 1 : Les stratégies principales d'enrichissement des peptides N-terminaux. (a) COFRADIC, (b) acétylation avec ou sans biotinylation, (c) guanidination et biotinylation, (d) DICAS. D'après [Armengaud, 2009]



1.2. Développement d'une nouvelle méthode pour améliorer la caractérisation des extrémités N-terminales des protéines (N-TOP)

Nous avons développé une nouvelle méthode visant à améliorer la caractérisation des peptides N-terminaux tout en laissant la possibilité d'identifier les peptides internes dans la même analyse pour une correction plus globale des erreurs de prédiction des gènes. Pour atteindre cet objectif, la stratégie utilisée ne doit pas impliquer la modification des peptides internes. Or, dans les stratégies d'enrichissement décrites précédemment, la modification des peptides internes faisait partie intégrante des protocoles.

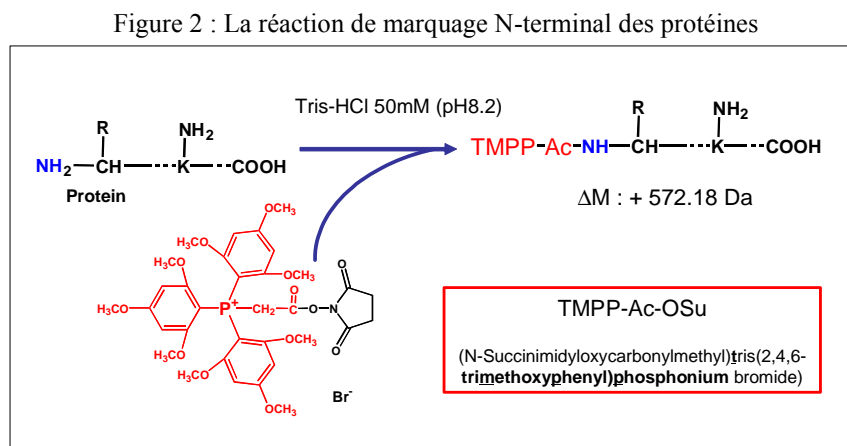
Pour mener à bien notre objectif, nous avons décidé de modifier sélectivement le peptide N-terminal des protéines de manière à améliorer sa détection dans les analyses LC-MS/MS après digestion enzymatique en laissant intact les peptides internes. Nous avons appelé cette approche la méthode « N-Terminal Oriented Proteomic » (N-TOP). Lors de son développement, nous avons sélectionné un réactif de marquage qui permet :

- D'améliorer l'efficacité d'ionisation des peptides N-terminaux en mode electrospray.
- De simplifier leur fragmentation en MS/MS.
- De modifier leur temps de rétention en chromatographie liquide pour qu'ils soient séparés de la plupart des peptides internes.

Notre choix s'est porté sur le N-Succinimidyl(oxycarbonylmethyl)tris(2,4,6-triméthoxyphényl) phosphonium bromide (TMPP-Ac-OSu) qui remplit ces conditions.

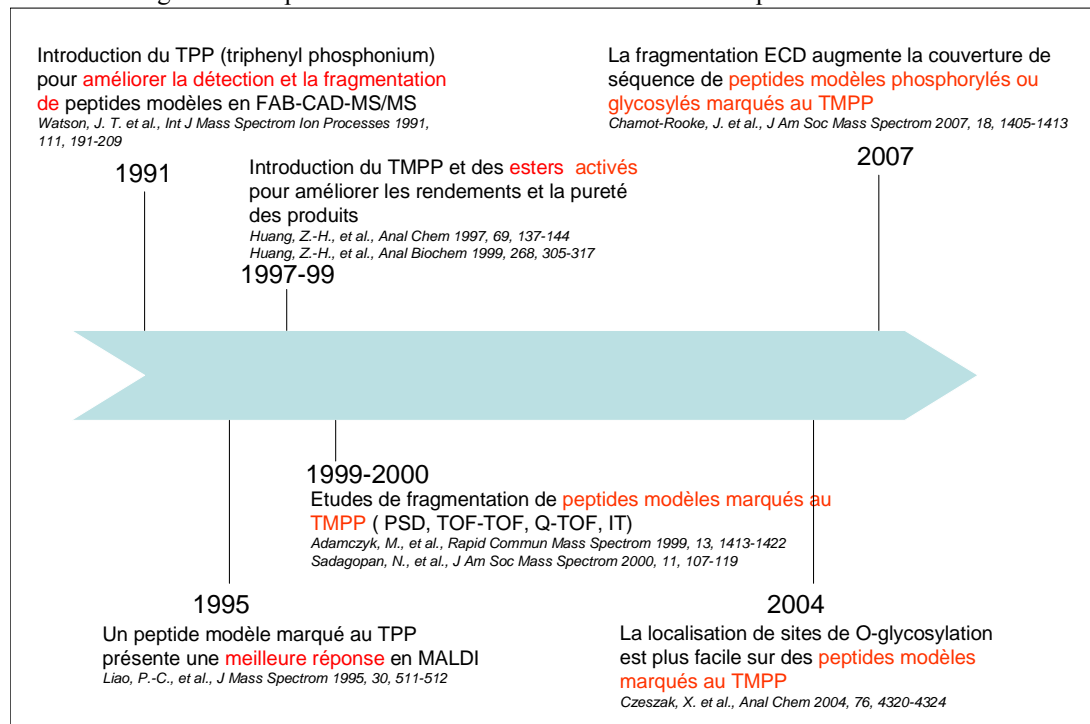
1.2.1. L'étape clé de la stratégie N-TOP : le marquage N-terminal des protéines avec le TMPP-Ac-OSu

Le marquage N-terminal des protéines avec le TMPP-Ac-OSu est réalisé dans des conditions strictes de pH (pH 8.2) pour conserver intactes les amines libres des chaînes latérales des lysines (Figure 2).



Le réactif TMPP-Ac-OSu a déjà été utilisé dans l'histoire de la spectrométrie de masse. Un bref historique de son utilisation est présenté en Figure 3.

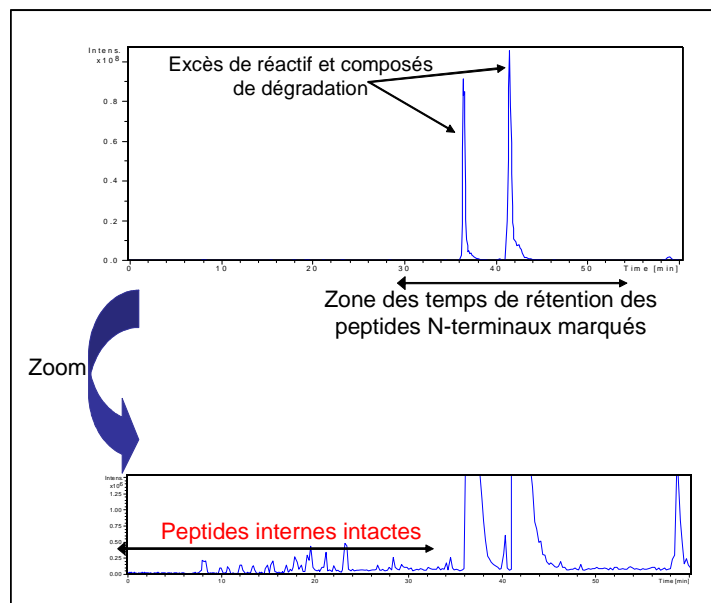
Figure 3 : Un petit résumé de l' « histoire » du TMPP en spectrométrie de masse



Les conditions de réaction de ce réactif sur des peptides modèles ont donc déjà été décrites [Roth et al., 1998; Adamczyk et al., 1999; Huang et al., 1999; Sadagopan et al., 2000; Czeszak et al., 2004; Chamot-Rooke et al., 2007] mais elles ne sont pas adaptées aux extraits protéiques complexes en présence de détergents. Nous avons donc développé un nouveau mode opératoire et déterminer des conditions expérimentales pour réaliser le marquage N-terminal des protéines à partir d'un extrait biologique total avec les détergents, les agents chaotropes et les conditions de réduction habituels utilisés pour l'extraction des protéines. Il fut par exemple nécessaire de remplacer le dithiothreitol (DTT) par la tributylphosphine (TBP) dans le tampon de réaction car le DTT gêne la réaction au TMPP. De plus, dans notre cas un grand excès de TMPP doit être utilisé pour marquer un extrait très complexe (200 équivalents). Avant les analyses LC-MS/MS qui suivent la digestion enzymatique, tout l'excès de TMPP et de ses composés de dégradation doivent être éliminés car leur temps de rétention chromatographique est proche de celui des peptides marqués au TMPP et il va donc gêner leur détection par spectrométrie de masse.

1.2.2. Elimination de l'excès de réactif et des composés de dégradation

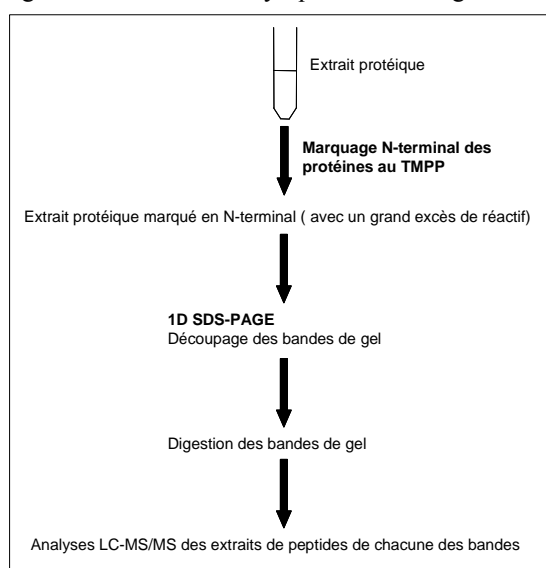
Plusieurs méthodes pour éliminer l'excès de réactif et des composés de dégradation ont été testées, par exemple des systèmes de filtration sur membrane, mais ils entraînaient des pertes de matériel et l'élimination restait peu satisfaisante (Figure 4)



Finalement, l'utilisation d'une étape de gel d'électrophorèse monodimensionnel (gel 1D) après marquage N-terminal des protéines s'est avérée être idéale pour éliminer l'excès de réactif et les composés de dégradation. Le gel 1D présente également l'avantage d'être compatible avec les détergents très puissants (SDS) et de réduire la complexité des extraits protéiques avant digestion et analyse LC-MS/MS.

Le protocole analytique final qui sera d'abord testé sur des mélanges de protéines modèles est présenté en Figure 5. Le mode opératoire est détaillé dans la publication des résultats.

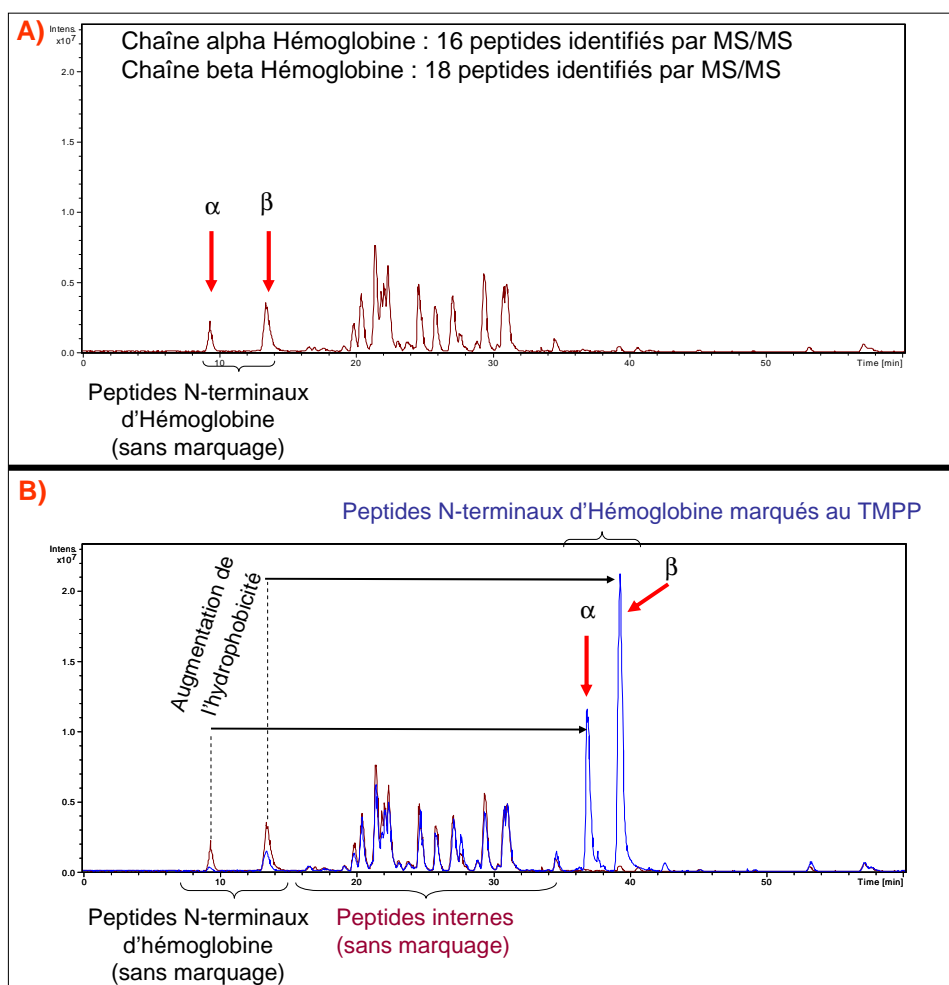
Figure 5 : Protocole analytique de la stratégie N-TOP



1.2.3. Application de la méthode sur des protéines modèles

Dans un premier temps, la stratégie N-TOP a été appliquée à un mélange de protéines modèles constitué des chaînes alpha et beta d'hémoglobine solubilisées dans le tampon de réaction. Les protéines modèles marquées au TMPP en N-terminal d'une part et sans marquage d'autre part ont été séparées sur un gel 1D puis digérées à la trypsine. Les peptides générés ont été extraits et analysés par LC-MS/MS (Figure 6). La détection des peptides N-terminaux est basée sur la modification de masse engendrée par l'ajout du groupement TMPP (+572.18 Da).

Figure 6 : Comparaison des chromatogrammes BPC à partir des analyses LC-MS/MS des digests d'hémoglobines. A) sans marquage ; B) en superposant le signal sans marquage et le signal avec marquage



1.2.3.1. Impact du marquage TMPP sur l'efficacité d'ionisation et la rétention chromatographique des peptides N-terminaux

Sur la Figure 6, on observe que le marquage au TMPP augmente de manière très significative l'efficacité d'ionisation des peptides en mode électrospray. Les 2 pics correspondant aux peptides N-terminaux sont mineurs dans le chromatogramme de l'analyse sans marquage alors qu'ils deviennent

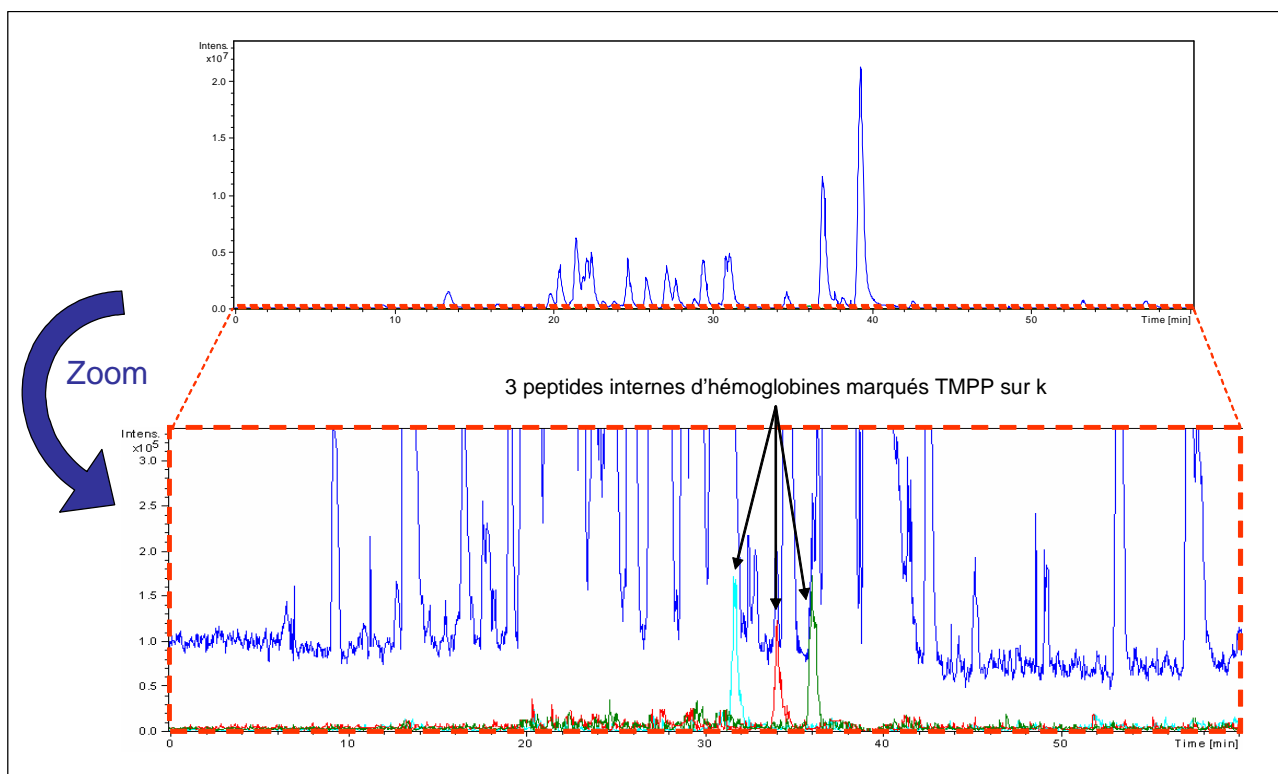
les pics les plus intenses du chromatogramme dans l'analyse avec marquage. La modification TMPP introduit une charge permanente positive et un groupement hydrophobe qui conduisent à l'amélioration de l'efficacité d'ionisation des peptides. La charge permanente assure au moins une charge à tous les peptides marqués et l'hydrophobicité augmente l'affinité des peptides marqués pour la surface des gouttes du spray et donc leur efficacité d'ionisation (comme vu en Partie bibliographique Chapitre 3. 1.1.2.4.)

Sur la Figure 6, on observe également que les peptides marqués au TMPP voient leur temps de rétention augmenté en RP-HPLC grâce à l'addition du groupement TMPP très hydrophobe. Pour l'analyse de mélanges biologiques plus complexes, cette modification des temps de rétention permettra une meilleure détection des peptides marqués au TMPP car ils seront élués dans une partie moins complexe du chromatogramme (après l'ensemble des peptides internes).

1.2.3.2. La spécificité du marquage TMPP

Sur la figure 6, il apparaît que les peptides internes ne subissent pas de modifications, ce qui signifie que l'amine libre de la chaîne latérale des lysines reste intacte grâce au contrôle du pH à 8.2 durant le marquage. En effet, une régiosélectivité du marquage TMPP de 95 % sur les amines libres N-terminales ($pK_a \sim 7.8$) vis à vis des amines libres des lysines ($pK_a \sim 11$) à un pH de 8.2 est rapportée dans la littérature [Huang et al., 1999]. Toutefois, si les protéines sont en très grandes quantités, il sera possible de détecter quelques peptides modifiés sur la chaîne latérale de leur lysine. Néanmoins, ces réactions indésirables sont extrêmement mineures comme le montre la Figure 7. Sur cette figure, on observe que lorsqu'on réalise un agrandissement important sur le bruit de fond du chromatogramme BPC, seuls 3 pics de très faible intensité (inférieure au seuil de sélection MS/MS) sont détectés. Ils correspondent à des peptides tryptiques internes modifiés sur leur lysine. Ces réactions parasites extrêmement mineures ne sont pas réhibitoires pour notre stratégie car elles génèrent des peptides tryptiques modifiés dont l'intensité ne représente environ qu'1 % de celle des peptides N-terminaux modifiés.

Figure 7 : Chromatogramme BPC à partir de l'analyse LC-MS/MS du digest d'hémoglobine avec marquage. En zoom, les chromatogrammes d'ions extraits des 3 peptides tryptiques internes d'hémoglobines qui ont subi une réaction parasite sur la chaîne latérale de leur lysine.



1.2.3.3. Impact du marquage TMPP sur la fragmentation des peptides N-terminaux

En plus d'améliorer l'efficacité d'ionisation en mode électrospray et d'augmenter très significativement le temps de rétention en RP-HPLC des peptides N-terminaux, le marquage TMPP modifie également leur fragmentation. La comparaison des spectres de fragmentation CID du peptide N-terminal doublement chargé de la chaîne alpha d'hémoglobine (VLSPDAK) avec et sans marquage TMPP illustre cette modification de fragmentation (Figure 8).

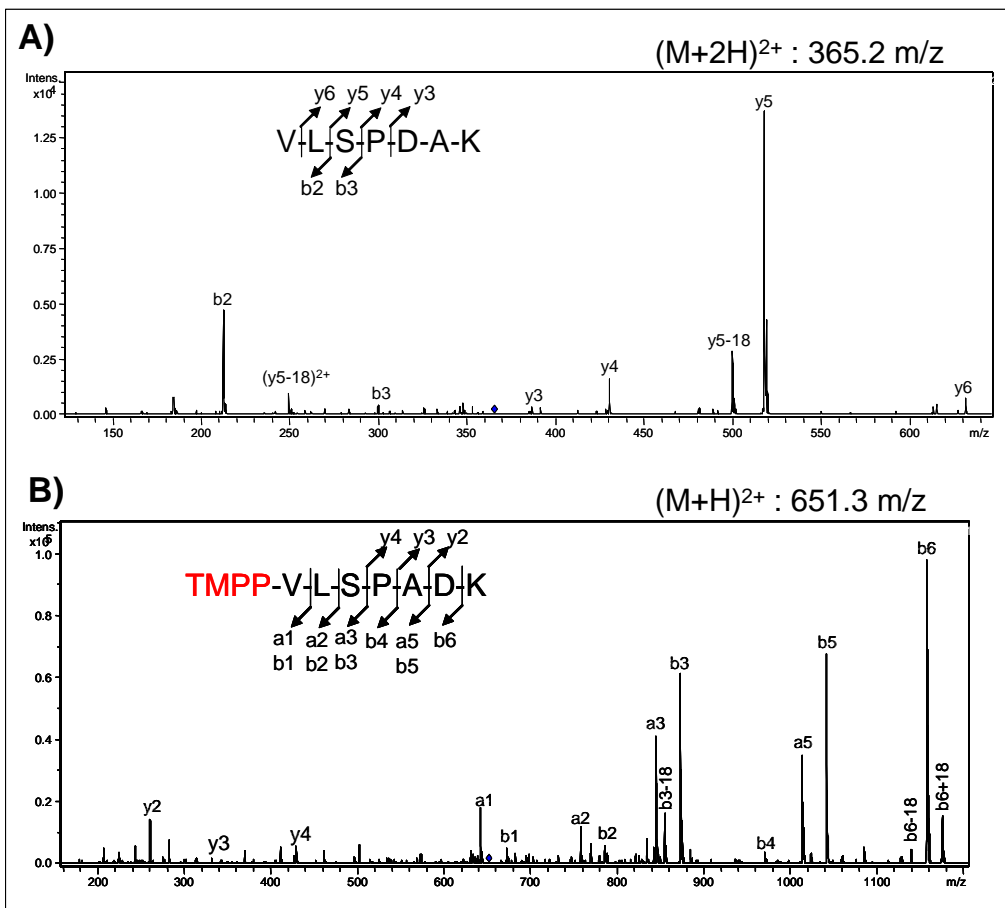
Sur la Figure 8, on constate que le spectre MS/MS du peptide non marqué présente essentiellement des fragments y - et quelques fragments b -. Ce spectre est assez représentatif de la fragmentation CID des peptides tryptiques doublement chargés. En effet, les deux charges sont probablement localisées sur l'amine libre N-terminale et sur la chaîne latérale de la lysine. La mobilisation du proton en N-terminal nécessite peu d'énergie et le peptide se fragmente en suivant le modèle du proton mobile (Partie bibliographique, Chapitre 3. 1.3.3.) et en générant principalement des fragments y - et b -. Les ions y -, retenant la charge sur la chaîne latérale de l'acide aminé C-terminal, sont majoritaires car plus résistants aux décompositions qui peuvent suivre.

Le spectre MS/MS du peptide marqué au TMPP présente essentiellement des fragments N-terminaux (a - et b -). Les quelques fragments y - sont très minoritaires. Ce spectre illustre bien le type

de fragmentation subi par les peptides doublement chargés marqués au TMPP. Ici, les deux charges sont la charge permanente du phosphonium quaternaire et la charge due au proton localisé sur la chaîne latérale de la lysine. La mobilisation et la délocalisation du proton sur le squelette peptidique permet au peptide de se fragmenter en générant des ions y- et b- avec des ions b- majoritaires à cause de la charge retenue en N-terminal. Toutefois, ici, la mobilisation du proton nécessite plus d'énergie et par conséquent, le mode de fragmentation distant de la charge devient très compétitif en terme d'énergie requise. C'est la raison qui semble expliquer la présence assez importante des fragments a-. Ce peptide doit vraisemblablement se fragmenter par les deux modes de fragmentation. Les peptides N-terminaux marqués au TMPP doublement chargés présenteront donc généralement cette zone sans fragments dans les basses masses du spectre CID (peu de fragments y- et accroissement de la masse des fragments N-terminaux par le TMPP).

Les spectres de fragmentation du peptide N-terminal de la chaîne beta d'hémoglobine ne sont pas présentés ici mais ils sont très similaires.

Figure 8 : Fragmentation CID du peptide N-terminal de la chaîne alpha d'Hémoglobine doublement chargé avec et sans marquage TMPP



Les résultats obtenus par application de la stratégie N-TOP sur les protéines modèles montrent que le protocole analytique développé permet d'améliorer de façon significative l'identification des peptides N-terminaux. Il peut donc être appliqué à l'étude de mélanges très complexes de protéines.

2. Application de la stratégie N-TOP comme aide à l'annotation génomique

Cette étude a été réalisée en collaboration avec l'équipe de bioinformatique du Dr Olivier Poch du Laboratoire de Biologie et Génomique Structurale à l'Institut de Génétique et de Biologie Moléculaire et Cellulaire (CNRS/INSERM/ULP) de Strasbourg et l'équipe du groupe Avenir du Dr Jean-Marc Reyrat de l'Unité de Pathogénie des Infections Systémiques (INSERM-UMR 7512) de l'université Paris Descartes.

2.1. Contexte de l'étude

La prédiction incorrecte des codons d'initiation des protéines dans les génomes procaryotes est sans aucun doute le type d'erreur le plus répandu lors des processus d'annotation des génomes des organismes procaryotes. Les erreurs d'annotation sont directement répercutées dans les banques protéiques et ont des conséquences multiples pour les études biologiques liées à l'organisme en question (Partie bibliographique, Chapitre 2. 3.2.). De plus, les approches extrinsèques d'annotation génomiques utilisent les informations des protéines présentes dans les banques. Donc, les erreurs peuvent encore être amplifiées, c'est ce qu'on appelle un « effet domino négatif ».

Dans ce contexte, pour briser ce cercle vicieux, les stratégies de protéogénomiques sont prometteuses. Nous avons donc couplé la stratégie de recherche de données MS/MS dans le génome avec la nouvelle stratégie N-TOP développée précédemment pour aider à la détermination des codons d'initiation des protéines. En plus, nous avons combiné cette stratégie protéogénomique avec une stratégie de génomique comparative. Cette approche que nous avons appelée « ortho-proteogenomic » utilise les déterminations expérimentales des codons d'initiation des protéines d'un organisme de référence pour les propager aux protéines orthologues des espèces proches.

Le genre *Mycobacterium* a été choisi comme premier exemple d'étude pour cette stratégie car 17 génomes complets de souches ou d'espèces de mycobactéries sont disponibles et annotés dont plusieurs espèces pathogènes importantes comme *M. tuberculosis*, *M. leprae* ou encore *M. ulcerans*. Parmi les mycobactéries, *M. smegmatis* est particulièrement adapté pour tester notre approche globale en tant qu'espèce modèle qui présente un grand nombre de gènes.

2.2. Stratégie d'analyse protéogénomique

Le protocole analytique de la stratégie N-TOP, tel que défini en 1.2. de ce chapitre, a été appliqué sur un extrait protéique total de *M. smegmatis*. Le détail des expériences est décrit dans la publication des résultats. Je présenterai ici les quelques points principaux.

Après réduction/alkylation des cystéines et marquage N-terminal au TMPP de l'extrait protéique total, celui-ci a été séparé sur gel 1D. Le gel a été découpé systématiquement en 100 bandes

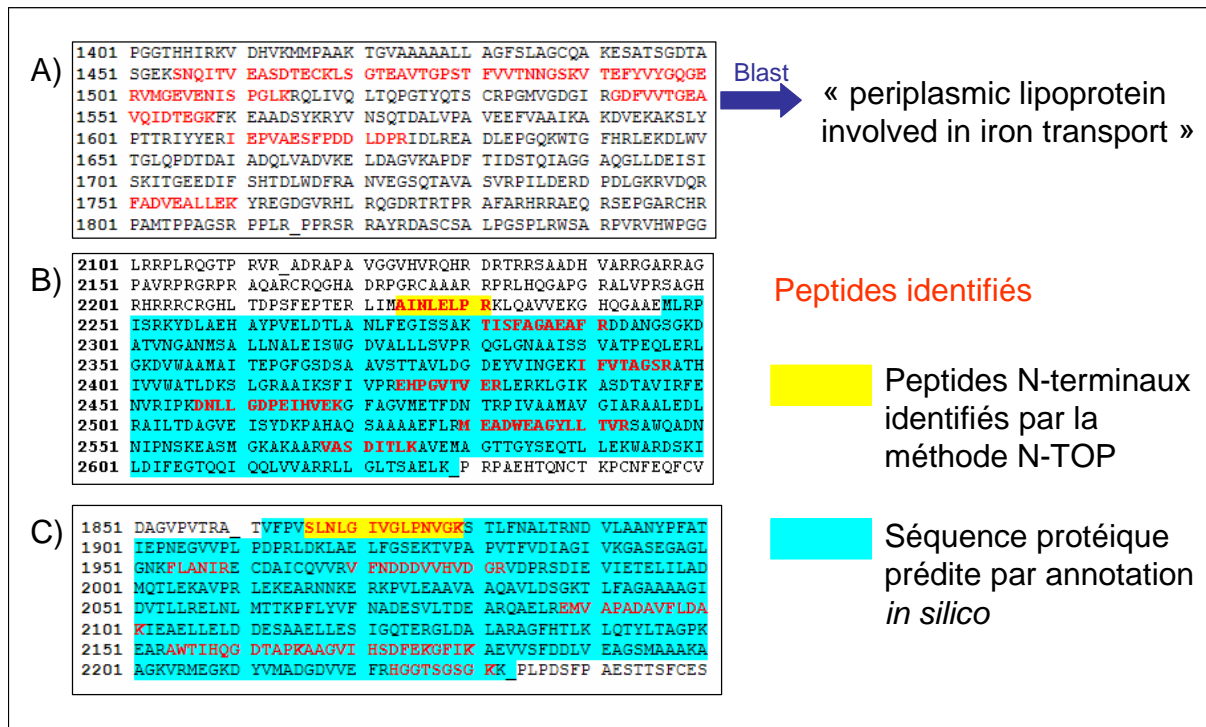
horizontales et chaque bande a été divisée en 2 morceaux pour réaliser deux digestions enzymatiques en parallèle à la trypsine et à l'endoprotéinase AspN. La taille de la piste de gel et la quantité de protéines estimée qui y a été déposée ont été adaptées à la sensibilité du système instrumental utilisé pour les analyses LC-MS/MS. En effet, les peptides de digestion ont été analysés par LC-MS/MS sur un système microHPLC (Agilent Series 1100) couplé à une trappe ionique HCT Ultra (Bruker Daltonics). Ce système utilise une source électrospray et une colonne chromatographique de 300 µm de diamètre interne, ce qui rend le système moins sensible (d'un facteur ~8 d'après notre expérience sur des protéines modèles) qu'un système employant une colonne de 75 µm de diamètre interne et une source nano-électrospray (comme le système de puce microfluidique HPLC-Chip d'Agilent). De plus, la complexité des mélanges peptidiques et le comportement singulier des peptides N-terminaux marqués au TMPP en RP-HPLC nous ont conduit à réaliser les séparations chromatographiques avec un gradient de modification de la composition de la phase mobile de pente très douce (10-70 % acétonitrile sur 170 minutes) pour permettre la fragmentation d'un maximum de composés. Le paramétrage de l'acquisition du spectromètre de masse a également été réglé pour permettre l'identification d'un nombre maximum de peptides distincts dans les mélanges complexes analysés (cycles d'acquisition des spectres MS et MS/MS, seuil de sélection des ions à fragmenter). Le paramétrage de l'exclusion dynamique des précurseurs a également été réglé en tenant compte de la grande complexité des échantillons analysés. Il s'est avéré pour notre étude que les meilleurs résultats d'identification étaient obtenus en ne sélectionnant et en ne fragmentant qu'une seule fois les ions précurseurs au cours d'une analyse LC-MS/MS (exclusion du m/z du précurseur de la sélection sur toute la durée du pic chromatographique, ~1min)

Les peptides N-terminaux marqués au TMPP et les peptides internes non modifiés ont été identifiés par soumission des données de LC-MS/MS à des requêtes Mascot dans une version « target-decoy » (voir Partie bibliographique, Chapitre 4 2.1.3.1) de la banque génomique de *M. smegmatis* avec un paramétrage du moteur de recherche prenant en compte les particularités de fragmentation des peptides N-terminaux marqués au TMPP (décrit dans la publication des résultats).

2.3. Corrections à l'annotation génomique de *M. smegmatis*

La stratégie protéogénomique appliquée dans notre étude a permis de participer à l'annotation génomique de *M. smegmatis* à différents niveaux. En plus d'avoir permis la validation de l'expression de nombreuses protéines et l'attribution de nombreux codons d'initiations, elle a aussi apporté des corrections à cette annotation. Ainsi, la stratégie protéogénomique a permis d'identifier plusieurs gènes non prédits (Figure 9 A), d'étendre la séquence protéique quand elle était prédite trop « courte » (Figure 9 B) ou de la raccourcir quand elle était prédite trop « longue » (Figure 9 C).

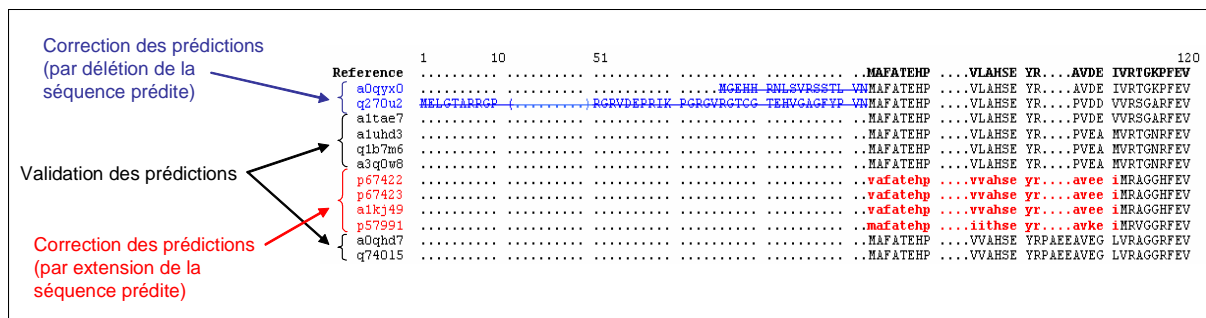
Figure 9 : Exemples de corrections expérimentales à l'annotation génomique de *M. smegmatis*. A) Identification d'un nouveau gène. B) Séquence protéique prédite trop « courte ». C) Séquence protéique prédite trop « longue ».



2.4. Propagation des déterminations expérimentales des codons d'initiation aux autres mycobactéries

La stratégie de génomique comparative appliquée dans notre étude a permis de participer à l'annotation génomique des autres mycobactéries par propagation des déterminations expérimentales des codons d'initiation chez *M. smegmatis*. Cette propagation a permis de valider ou de corriger les codons d'initiations prédits des protéines orthologues des autres mycobactéries (Figure 10).

Figure 10 : Propagation des déterminations expérimentales des codons des protéines de *M. smegmatis* aux autres mycobactéries par validation, extension ou délétion des séquences protéiques prédites



2.5. Résultats publiés

Les résultats obtenus dans cette étude ont fait l'objet d'une publication acceptée dans le journal *Genome Research* en octobre 2008.

La stratégie protéogénomique appliquée à l'étude de *M. smegmatis* a permis grâce à l'analyse simultanée des peptides N-terminaux marqués et des peptides internes de valider l'expression de 917 protéines et de détecter 29 nouveaux gènes pour un total de 946 protéines identifiées (FDR<1 %). Elle a aussi permis d'identifier exactement le codon d'initiation de 443 protéines. Les peptides N-terminaux identifiés ont mis en évidence un taux d'erreur de 19 % sur la prédiction des codons d'initiation. Toutes ces validations/corrections des prédictions ont été soumises à la banque UniprotKB pour intégration des résultats. L'analyse comparative appliquée aux séquences protéiques des 16 autres mycobactéries a conduit au traitement de 4328 séquences pour lesquelles le codon d'initiation a été validé dans 3727 cas (86 %) et corrigé dans 601 cas (14 %).

2.6. Conclusion

L'étude réalisée grâce à notre approche « ortho-proteogenomic » constitue la première évaluation de l'exactitude de l'annotation génomique des codons d'initiation des protéines au niveau d'un genre taxonomique, en l'occurrence le genre *Mycobacterium*. Les taux d'erreurs dans les banques protéiques publiques mis en évidence par cette évaluation sont relativement élevés (entre 9 et 21 % selon l'espèce) sachant que certaines espèces de mycobactéries (par exemple *M. tuberculosis* H37Rv) sont des bactéries pathogènes très étudiées. Ces résultats soulignent l'intérêt d'utiliser les données expérimentales de protéomique comme aide à l'annotation génomique.

La propagation des déterminations des codons d'initiation des protéines obtenus pour *M. smegmatis* aux autres mycobactéries constitue un nouveau mode extrinsèque d'annotation génomique. En effet, ici, les données de référence propagées à l'annotation génomique des 16 espèces de mycobactéries dont le génome est séquencé à ce jour sont purement expérimentales. L'« effet domino négatif » de la propagation d'erreurs de prédiction parfois obtenu par les approches extrinsèques d'annotation génomique peut ainsi être converti en un « effet domino positif ». Notre approche, utilisant des techniques standards de protéomique et de bioinformatique, pourrait être associée de manière beaucoup plus systématique aux projets d'annotation des génomes.

Toutefois, l'approche telle qu'elle a été appliquée dans cette étude reste limitée en terme d'automatisation pour la partie protéogénomique, notamment en ce qui concerne la validation des identifications des peptides N-terminaux qui est réalisée manuellement. Le développement d'un protocole de validation plus automatisé des identifications des peptides N-terminaux marqués passe par l'utilisation unique des seuils de score Mascot et l'évaluation du FDR associé (par une stratégie target-decoy par exemple). Nous avons donc fait évoluer la méthode pour améliorer son efficacité et son automatisation. Cette évolution sera présentée dans la suite du chapitre.



Ortho-proteogenomics: Multiple proteomes investigation through orthology and a new MS-based protocol

Sébastien Gallien, Emmanuel Perrodou, Christine Carapito, et al.

Genome Res. 2009 19: 128-135 originally published online October 27, 2008

Access the most recent version at doi:[10.1101/gr.081901.108](https://doi.org/10.1101/gr.081901.108)

Supplemental Material <http://genome.cshlp.org/content/suppl/2008/12/05/gr.081901.108.DC1.html>

References This article cites 41 articles, 16 of which can be accessed free at:
<http://genome.cshlp.org/content/19/1/128.full.html#ref-list-1>

Email alerting service Receive free email alerts when new articles cite this article - sign up in the box at the top right corner of the article or [click here](#)

To subscribe to *Genome Research* go to:
<http://genome.cshlp.org/subscriptions>

Methods

Ortho-proteogenomics: Multiple proteomes investigation through orthology and a new MS-based protocol

Sébastien Gallien,^{1,8} Emmanuel Perrodou,^{2,3,4,5} Christine Carapito,¹ Caroline Deshayes,^{6,7} Jean-Marc Reyrat,^{6,7} Alain Van Dorselaer,¹ Olivier Poch,^{2,3,4,5} Christine Schaeffer,¹ and Odile Lecompte^{2,3,4,5}

¹Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC-DSA, ULP, CNRS, UMR7178, 67 087 Strasbourg, France;

²Department of Structural Biology and Genomics, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), F-67400 Illkirch, France; ³INSERM, U596, F-67400 Illkirch, France; ⁴CNRS, UMR7104, F-67400 Illkirch, France; ⁵Faculté des Sciences de la Vie, Université Louis Pasteur, F-67000 Strasbourg, France; ⁶Faculté de Médecine René Descartes, Université Paris Descartes, Paris Cedex 15, F-75730, France; ⁷INSERM, U570, Unité de Pathogénie des Infections Systémiques, Paris Cedex 15, F-75730, France

The progress in sequencing technologies irrigates biology with an ever-increasing number of genome sequences. In most cases, the gene repertoire is predicted *in silico* and conceptually translated into proteins. As recently highlighted, the predicted genes exhibit frequent errors, particularly in start codons, with a serious impact on subsequent biological studies. A new “ortho-proteogenomic” approach is presented here for the annotation refinement of multiple genomes at once. It combines comparative genomics with an original proteomic protocol that allows the characterization of both N-terminal and internal peptides in a single experiment. This strategy was applied to the *Mycobacterium* genus with *Mycobacterium smegmatis* as the reference, and identified 946 distinct proteins, including 443 characterized N termini. These experimental data allowed the correction of 19% of the characterized start codons, the identification of 29 proteins missed during the annotation process, and the curation, thanks to comparative genomics, of 4328 sequences of 16 other *Mycobacterium* proteomes.

[Supplemental material is available online at www.genome.org.]

The increasing availability of data from multiple genome sequencing projects provides biologists with an invaluable framework to integrate experimental results and design new experiments at different scales. However, several recent studies have highlighted the prevalence of gene prediction errors, even in the “simple” prokaryotic genomes. Genome sequencing itself represents a non-negligible source of errors (Weinstock 2000), but despite major advances, most inconsistencies result from *in silico* predictions (Galperin et al. 1998). Among these errors, the incorrect prediction of initiation codons in prokaryotic genomes is particularly widespread (Aivaliotis et al. 2007). For instance, error rates in start codon prediction vary from 10% to 44% in *Halobacterium salinarum* and *Natromonas pharaonis* (Aivaliotis et al. 2007), depending on the gene prediction program used. This reality is often underestimated or even ignored by biologists, even though the correct definition of genes is determinant for subsequent *in silico* and experimental studies. For example, by altering the definition of the coding sequence of a gene, an erroneous start codon can hamper the detection of regulatory motifs on the genome or even mask another gene in a compact genome (Salgado et al. 2000; Edwards et al. 2005). Moreover, the protein sequence itself can be either truncated or extended, leading to errors in bioinformatics protein characterization (func-

tion, localization, etc.) and, obviously, to major difficulties in protein expression experiments (Trivedi et al. 2004; Horie et al. 2007). The second highly prejudicial error encountered in prokaryotic genome annotation is under-prediction of small genes or genes exhibiting an unusual composition. The accumulation of erroneous information in genomic and protein databases will continue to grow since features are frequently transferred from annotated to unknown sequences (Doerks et al. 1998), which only amplifies the errors.

To break this vicious circle and to cope with the multiplication of prokaryotic genome data, including many projects aimed at exploring genetic diversity within a genus or a species by multiple-strain sequencing (Liolios et al. 2008), one cannot rely solely on manual curation. In this context, the proteogenomic approach, i.e., annotation refinement through proteomics, is promising and has already been used to investigate several bacterial genomes (Jaffe et al. 2004a,b; Wang et al. 2005; Gupta et al. 2007, 2008), revealing the expression of genes annotated as pseudogenes as well as some completely missed genes or some errors in start codon annotation. However, these high-throughput studies do not focus on the N-terminal identification of proteins, limiting the correction of gene boundaries. In contrast, other methods have aimed at the specific identification of N-terminal peptides from the digest of a protein extract (Gevaert et al. 2003; McDonald et al. 2005; McDonald and Beynon 2006), but these methods imply the loss of all internal peptides, which is a major drawback both for protein and proteome coverage.

Here, we report an original strategy coupling a new N-

⁸Corresponding author.

E-mail sgallien@chimie.u-strasbg.fr; fax 33-3-90-24-27-81.

Article published online before print. Article and publication date are at <http://www.genome.org/cgi/doi/10.1101/gr.081901.108>.

terminal-oriented proteomic (N-TOP) method with comparative genomics. The concept underlying this “ortho-proteogenomic” approach is to characterize large sets of N-terminal and internal peptides in a single run for a reference organism and to propagate the obtained experimental information to orthologs of closely related species. We have developed two original strategies: (1) a new straightforward mass spectrometry (MS)-based workflow relying on a single specific N-terminal labeling of the intact proteins, preserving internal peptides, and (2) a conservative comparative approach to curate the annotation of multiple closely related microbial genomes simultaneously. *Mycobacterium* was chosen as a first study example for this strategy since 17 complete genomes of *Mycobacterium* strains or species are available, including important pathogens such as *M. tuberculosis*, *M. leprae*, and *M. ulcerans*. Within the *Mycobacterium* genus, *M. smegmatis* is ideally suited to test our combined approach since it is a model species for experiments (a fast-growing and nonpathogenic species) that exhibits a large repertoire of genes. Our proteogenomic approach allowed the experimental identification of 946 proteins from *M. smegmatis*, revealing 29 new proteins missed during the annotation process. In the same experiment, 443 N-terminal peptides of the 946 proteins were characterized. These N termini sequences revealed an error rate of 19% in the prediction of the initiation codon. Comparative analysis applied to the sequences of the 16 other mycobacteria resulted in 4328 curated protein sequences. Besides the immediate value of these data to the whole scientific community working on *Mycobacterium*, the ortho-proteogenomic method presented here should initiate a new step in genome annotation and sequence database curation. The data used in this study are available at <ftp://ftp-igbmc.u-strasbg.fr/pub/lecompte/Msmegmatis>.

Results

TMPP labeling of protein N termini: Workflow establishment

N-succinimidylloxycarbonylmethyl)tris(2,4,6-trimethoxyphenyl) phosphonium bromide (TMPP-Ac-OSu) was selected as the labeling reagent in order to obtain a better ionization efficiency, simplified fragmentation, and retention times allowing a better separation of N-terminal peptides analyzed by liquid chromatography-coupled mass spectrometry (LC-MS/MS) after enzymatic digestion. A new workflow was developed and experimental conditions were setup for N-terminal protein labeling from a total biological extract with all the usual detergents, chaotropic agents, and reduction conditions used for protein extraction, including membrane proteins (Xiong et al. 2005). It was necessary to replace dithiothreitol (DTT) by tributylphosphine (TBP) in the labeling buffer since DTT hindered the TMPP reaction.

Derivatization conditions on model peptides have been described (Roth et al. 1998; Adamczyk et al. 1999; Huang et al. 1999; Sadagopan and Watson 2000; Czeszak et al. 2004; Chamot-Rooke et al. 2007), but they were totally inadequate for complex protein extracts in the presence of detergents. In our case, a large excess of TMPP must be used for complex extract labeling. Prior to LC-MS/MS analysis, all traces of TMPP must be removed, as TMPP retention time is close to that of TMPP-derivatized peptides and thus interferes with their MS detection. Several methods to remove the excess were tested, including different membrane filtration devices, but they induced dramatic peptide material losses and did not allow complete elimination of TMPP.

Finally, a one-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis (1D SDS-PAGE) step was shown to be ideal for removing excess reagent and had the additional advantages of being compatible with strong detergents and of reducing the complexity of protein extracts prior to LC-MS/MS.

Application of the workflow to model proteins

Purified hemoglobin alpha and beta chains solubilized in the labeling buffer used for biological samples (see labeling buffer in Methods) were employed as the model system. TMPP-derivatized and -nonderivatized hemoglobin chains were separated on 1D SDS-PAGE and digested in-gel, and the peptide extracts were analyzed by LC-MS/MS (see Supplemental Methods S1 for more details). Recognition of N-terminal peptides was based on the characteristic mass shift caused by the TMPP labeling (+572.18 Da). Figure 1 shows that the N-terminal derivatized peptides have an increased retention time due to the addition of the hydrophobic TMPP group, in contrast to internal peptides. This implies that the lysine side chain ϵ -amines were preserved fully intact by a careful control of pH at 8.2 during the labeling. For studies of very complex biological samples, this retention time shift allows a better LC-MS/MS analysis of TMPP-derivatized peptides since it shifts the elution times toward a less complex part of the chromatogram.

Figure 1 also shows that the TMPP labeling significantly increases the ionization efficiency. The two peaks of N-terminal peptides that were minor in the native form become major after derivatization, whereas all internal peptides remain unchanged. So, TMPP derivatization introduces a permanent positive charge and a hydrophobic group resulting in an enhancement of the ionization efficiency. These results show that the established workflow provides efficient N-terminal peptide identification in a complex mixture of proteins.

Application of the workflow to the proteome of *M. smegmatis*

The workflow established for the *M. smegmatis* proteome is summarized in Figure 2. After N-terminal labeling, the total protein extract was loaded on a 1D SDS-PAGE. The gel was systematically cut into 100 horizontal bands, and each band was divided into

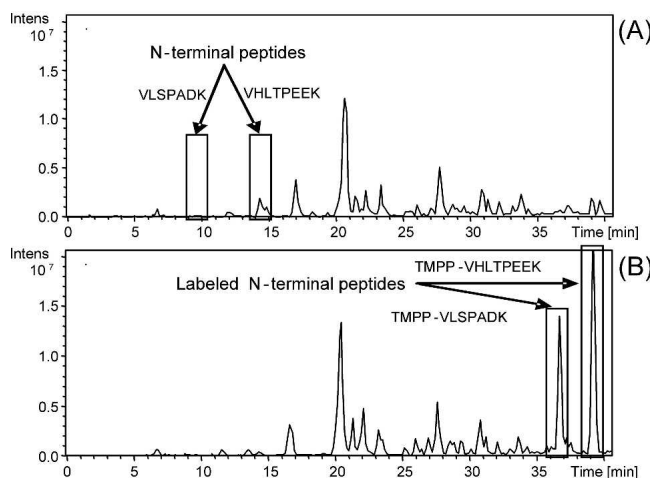


Figure 1. Comparison of base peak chromatograms from LC-MS/MS of hemoglobin digests. (A) Without N-terminal protein labeling; (B) with N-terminal protein labeling.

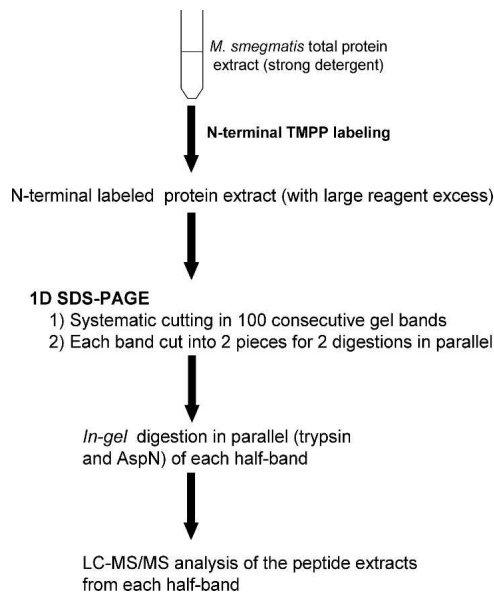


Figure 2. Analytical workflow.

two equal gel slices for enzymatic digestion with trypsin and endoproteinase AspN. Finally, after in-gel digestion, the peptide extracts were analyzed by LC-MS/MS. Due to their high complexity, a 170-min LC gradient was optimized from 10% to 70% CH₃CN with a slope of 0.35% per minute. N-terminal-labeled and internal native peptides were identified using Mascot MS/MS data searches against the complete genomic sequence rather than protein sequence databases, avoiding problems associated with computational predictions and annotations (Choudhary et al. 2001; Kuster et al. 2001; Oshiro et al. 2002; Jaffe et al. 2004a,b; Fermin et al. 2006; Gupta et al. 2007; Tanner et al. 2007).

The N-TOP strategy led to the identification of 443 unique N-terminal peptides (from 591 N-terminal sequences: 361 tryptic and 230 AspN sequences) and to a total of 946 validated proteins (Table 1). The two digestion modes appeared to be complementary but also allowed the cross-validation of the determined N termini in 148 cases (Supplemental Fig. S2). Trypsin digestion was the most efficient since it led to the identification of >80% of all identified start codons. All experimentally determined N-terminal sequences, internal sequences, protein lists, and N-terminal spectra are available at <ftp://ftp-igbmc.u-strasbg.fr/pub/lecompte/Msmegmatis> and the MGF peak lists are available at <http://tranche.proteomecommons.org/> using the following hash:

```
zmBZ3tBKxJNci5SHJLwbUz3S3wSwvKabSuDzsbSVI0GgC1iXRY
af9iYmnuYwVTdOmSt9/fVCIONBtN/bfkWwAGadbOQAAAA
AAAADGw==.
```

The improved ionization efficiency of the labeled peptides on model proteins could be generalized at the proteome level. The impact on reversed phase chromatography was also of major importance since the analysis of the chromatograms revealed two distinct elution zones: the labeled N-terminal peptides (Fig. 3A) and the native internal peptides (Fig. 3B). These two zones are partially overlapping because, despite derivatization, very hydrophilic-labeled peptides can be eluted before very hydrophobic

native peptides. Nevertheless, this overall shift in chromatographic behavior prevented ionization suppression due to internal peptides and improved the characterization of labeled peptides.

In addition, the developed workflow had the advantage of not modifying internal peptides, allowing their simultaneous identification in a single classical LC-MS/MS analysis, even in the case of post-translational modifications. These internal peptides allowed the identification of 503 additional proteins and confirmed 92% of the protein identifications based on the N-terminal peptide.

Definition of rules for N-terminal peptide identification

Protein identification was commonly performed for internal peptides (see Methods). However, for N-terminal peptide identification, the Mascot search parameter settings had to be carefully adapted, since the TMPP labeling has a strong influence on the N-terminal labeled peptide fragmentation behavior, especially for doubly charged precursors. Considering that peptide scores from classical search engines are based mainly on the fragmentation behavior of native fully tryptic peptides, each fragmentation spectrum of a potential derivatized N-terminal peptide (whose first amino acid of the sequence is coded by a start codon or by the first codon following a start codon, ATG, GTG, or TTG) was manually validated using several criteria. The first criterion was the retention time shift described above. The second criterion was the fragmentation pattern observed for a large set of derivatized peptides, which allowed us to improve previous observations on a few model peptides (Adamczyk et al. 1999; Sadagopan and Watson 2000, 2001) and to establish fragmentation “rules.” Generally, due to the permanent positive charge, N-terminal fragment ions (a_n and b_n) are predominantly observed from enzymatic TMPP-derivatized peptides, with intensities depending on the peptide sequences (Fig. 4). As an example, the CID spectrum of a triply charged labeled peptide (Fig. 4A) displays a series of singly and doubly charged b_n ions plus a few y-type ions. Fragmentation of doubly charged labeled peptides generates fewer peaks (mostly N-terminal fragments) than unlabeled ones, especially in the lower mass range (Fig. 4B). This is expected because only C-terminal fragments can be present below 630 m/z (the mass of the a-type fragment corresponding to the TMPP-labeled glycine). This particular pattern of fragmentation for doubly charged peptides is specific to the TMPP-labeled peptides. For these peptides, in spite of high-quality fragmentation spectra, Mascot ion scores are generally lower than for unlabeled peptides because few fragments remain in the lower mass

Table 1. Summary of the results obtained in the current work

	Results
N-TOP strategy on <i>M. smegmatis</i>	
Unique N-terminal peptides identified	443
Start codon errors	86
Missed during annotation	15
Validated N termini	342
Additional proteins identified by internal peptides	503
Missed during annotation	14
Sequencing errors	3
Total number of proteins	946
Comparative approach in mycobacteria	
Start codon errors	601
Validated N termini	3727

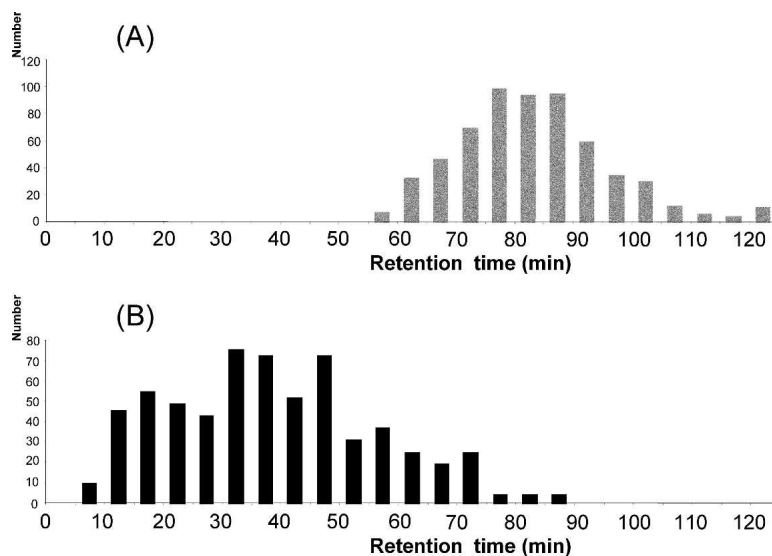


Figure 3. Comparison of the peptide retention times. (A) Retention time of the 591 N-terminal-labeled sequences. (B) Retention time of a random selection of 600 identified internal peptides.

range (Supplemental Table S5). In addition, the presence of very intense a-type ions is not common with ESI-Ion Trap mass spectrometers; they are observed with MALDI-TOF/TOF spectrometers, for example. So it was necessary to add these a-type fragments into the searched ESI-IT fragments used in Mascot searches to prevent false negative identifications. Finally, the last criteria taken into account were the length of the peptide and the identification of additional internal peptides. In the case of single-peptide assignments, all peptide sequences shorter than seven amino acids were excluded.

All N-terminal sequences collected in this study begin with, or begin immediately after, a start codon. Other labeled peptides (whose first amino acid of the sequence is not coded by a start codon or by the first codon following a start codon), for which N-terminal labeling was probably made after endogenous digestion of proteins (for example, signal peptide cleavage), were discarded at the initial step of the validation process (a total of 32 peptides), and no suggestions have been made on the position of any start codon not determined exactly. In this context, it was also interesting to observe the N-terminal methionine cleavages in the experimental sequences. Indeed, the final list of validated N-terminal sequences can be divided into two subsets: N-terminal sequences with and without methionine removal. The N-terminal methionine excision process occurs in all organisms and involves methionine aminopeptidase (MAP), whose activity has been reported to be linked mainly to the side chain size of the penultimate amino acid of the protein (Frottin et al. 2006). In our experimental data, the occurrence of methionine removals correlates well with the penultimate amino acids (Supplemental Table S4), in agreement with previously determined rules (Hirel et al. 1989; Link et al. 1997). This suggests that our start site data set does not contain sequences corresponding to endogenously digested proteins; otherwise, we would have observed random methionine cleavages. The manual validation has improved our ability to interpret TMPP-labeled peptide fragmentation spectra and allowed us to establish precise criteria and threshold scores that will facilitate the automation of the workflow in subsequent studies (Supplemental Table S5) and that allow estimation of the false discovery rate in our start site data set.

Comparison of the experimental proteome versus the predicted proteome

The experimentally determined sequences were compared with the predicted protein sequences (Table 1) from The Institute for Genomic Research (<http://www.tigr.org/>, released in 2005), revealing 86 errors (19%) in start codon prediction. 19% of the wrongly predicted sequences were too short, while 81% were too long. This error rate is in agreement with previous estimates from studies of other G+C-rich prokaryotes (Aivaliotis et al. 2007).

In addition, we detected three sequencing errors in the genome that had been reported previously (Deshayes et al. 2007), and we identified 29 proteins that were missed in the initial annotation. Start codons were identified for 15 of these proteins. *M. smegmatis* protein se-

quences with experimental start codon validation or correction and proteins missing in the annotation have been submitted to the Swiss-Prot database (<http://expasy.org/sprot/>) and are available at <ftp://ftp-igbmc.u-strasbg.fr/pub/lecompte/Msmegmatis>.

These results demonstrate that proteomics data can allow the identification of new proteins that would otherwise have been missed by classical computational annotation. They also confirm the utility of preserving internal peptides and clearly establish the need to work with genomic sequences rather than predicted proteins in order to identify new coding regions and start codons upstream of the predicted start sites.

Comparative genomic approach to correct mycobacteria sequences

We have capitalized on our large set of experimentally determined N-terminal sequences in order to correct predicted sequences in 16 other *Mycobacterium* genomes using a comparative genomic approach. Each reference sequence of *M. smegmatis* with an experimentally determined N terminus was aligned to its orthologs in mycobacteria, and the N-terminal region of the multiple alignment was analyzed (see example in Supplemental Fig. S3). A total of 4648 protein sequences were processed: 3727 N-terminal sequences (80% of the initial set) were validated and 601 (13%) were corrected. Since we adopted stringent parameters to avoid propagation errors, the remaining 320 sequences (7%) were not modified despite some discrepancy with the reference sequence of *M. smegmatis*. We thus obtained a conservative estimate of error rates in start prediction in *Mycobacterium* ranging from 9% to 21% (14% on average; see Supplemental Table S6). The validated and corrected sequences of mycobacteria are available at <ftp://ftp-igbmc.u-strasbg.fr/pub/lecompte/Mycobacteria>.

As observed in the comparison of the predicted proteome to experimental data, the large majority (84%) of detected errors are due to an erroneous 5' extension of the gene (Fig. 5). Sixty percent of deletions are short (at most five amino acids), while 75% of artificial extensions are longer than five amino acids, with 61 encompassing more than 29 amino acids. Such major errors will

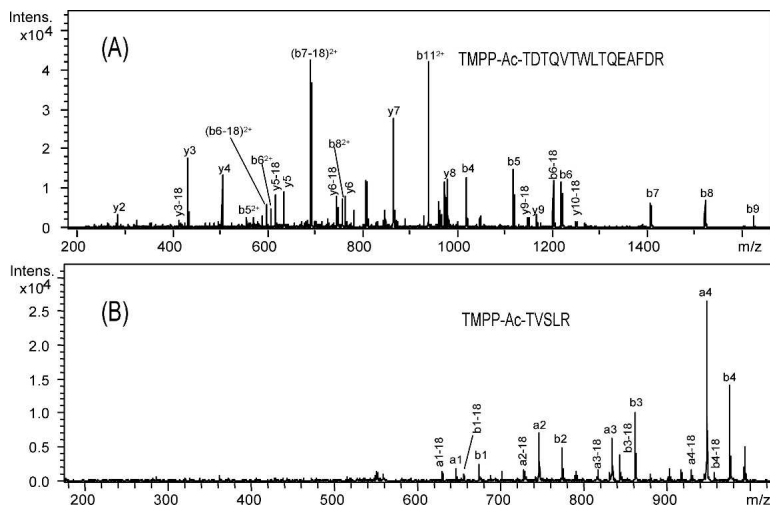


Figure 4. CID spectra of N-terminal-labeled peptides. (A) Triply charged labeled peptide. (B) Doubly charged labeled peptide.

seriously compromise subsequent *in silico* and/or experimental studies. Interestingly, minor initiation codons (GTG and TTG) are overrepresented in the corrected start codons (48% and 8%, respectively) compared with 35% and 4% reported for *M. tuberculosis* protein genes, for example (Cole et al. 1998). This suggests an overprediction of ATG as the translational start codon, either by gene prediction programs or during manual annotation.

Discussion

Recently, several proteomic initiatives have highlighted some inherent limits of *in silico* gene predictions in genomes, especially concerning exon/intron boundaries and start codon predictions. Our “ortho-proteogenomic” study represents the first quantitative evaluation of start codon prediction in public prokaryotic genome annotations at a genus level. The prediction error rates from 9% to 21% are surprisingly high considering that we are working on well studied pathogenic bacteria, such as *M. tuberculosis* H37Rv, whose genome has been annotated twice (Cole et al. 1998; Camus et al. 2002) and has benefited from the input of numerous experimental studies. The relative homogeneity of error rates within the *Mycobacterium* genus clearly demonstrates that errors cannot be attributed to a particular *in silico* annotation pipeline. Although the quality of genome annotation can be improved by combining different methods during the reannotation process, gene prediction tools are still faced with intrinsic limits. In the case of *M. smegmatis*, new annotations have been released since we began this work. In the latest version of the protein sequence database downloaded in 2007 from <http://www.uniprot.org/>, some annotation errors have been corrected: The error rate decreased from 19% to 13% for start codon prediction, and 29 proteins were missed in the initial annotation versus four in the latest one. However, the comparison between the two annotation releases (Supplemental Table S7) reveals that some new errors have been introduced in the latest version, emphasizing the limits of the reannotation processes.

In this context, there is an urgent need to promote genome annotation refinement and protein database curation by large-scale proteomic analysis (the so-called proteogenomics ap-

proach). The N-TOP workflow established for the *M. smegmatis* proteome is an optimized process involving standard proteomic techniques coupled with a fast one-step specific N-terminal protein labeling of a complete proteome. It combines the advantages of the usual proteogenomic approach and of N-terminal-specific techniques since it allows the characterization of both internal and N-terminal peptides in the same experiment. In comparison with other complete proteome analyses, our method results in a similar level of proteome coverage (946 proteins identified). For example, in 2005, Wang and colleagues (Wang et al. 2005) used multidimensional chromatography and tandem mass spectrometry to analyze the *M. smegmatis* proteome under 25 different growth conditions and obtained 901 distinct proteins. In order to compare our

method with another widespread proteomic approach, we performed a classical 2D-gel-based analysis of the *M. smegmatis* proteome (see Supplemental Methods S1) and identified 846 proteins with the same identification protocol. Thus, the N-TOP workflow provides satisfactory coverage and, in addition, offers the invaluable identification of a large set of N-terminal peptides. Indeed, we obtained only 173 N-terminal peptides with the 2D-gel-based analysis (data not shown) versus 443 with the N-TOP strategy. When comparing with other N-terminal specific approaches, the N-TOP strategy raised results similar to the largest published data set for prokaryotes (Aivaliotis et al. 2007) in a single experiment, using the extract corresponding to one culture condition and with a reduced number of LC-MS/MS analyses. Applying the same experimental workflow to extracts obtained under different growth conditions (rich media, different additives) and phase cultures would allow the identification of additional N-terminal peptides from proteins expressed specifically during these phases or in these media (Wang et al. 2005; Gupta et al. 2007) and would thus allow an increase of the overall proteome coverage. With an established and optimized protocol, the 4328-curated sequences obtained in this study can be produced with two person weeks of work using three instrument-weeks, and thus can constitute a reproducible part of a genomics pipeline.

An efficient and synergic integration of high-throughput experimental data with *in silico* predictions is a challenging task requiring standardization, automation, and portability. Both the proteomic and bioinformatic communities are becoming aware of this problem as attested by several recent initiatives, such as the development of standards for MS data representation (Orchard and Hermjakob 2008) or new software such as PepLine (Ferro et al. 2008) that allow easier mapping of MS/MS data on eukaryotic genomic sequences. The straightforward ortho-proteogenomic workflow presented here uses standard proteomic and bioinformatic techniques and thus can be applied routinely in any proteomic laboratory. In the case of the *Mycobacterium* genus, our conservative comparative approach, the “ortho” part of ortho-proteogenomic, was particularly fruitful since one experimentally determined N terminus allows us to correct/validate roughly 10 genes simultaneously. With 10 bacterial ge-

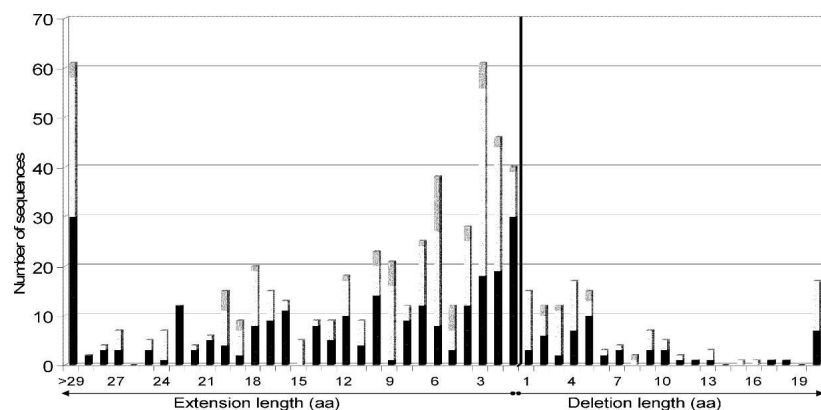


Figure 5. Distribution of false extension and deletion lengths in mycobacteria (including *M. smegmatis*). Bars are shaded proportionally to the number of each actual initiation codon after correction: (black) ATG, (white) GTG, (gray) TTG.

nuses totaling 172 complete genome sequences at the time of writing, our strategy represents a cost-effective and promising means to curate the huge amount of incoming genomic data. More generally, it would be interesting to extend our strategy of propagation to more distantly related organisms (within a class, for instance) by sampling the extreme representative species at the experimental level and integrating the data in the evolutionary context of ortholog alignment.

Methods

N-terminal derivatization of *M. smegmatis* protein extract

Unless otherwise specified, all chemicals were obtained from Sigma. A solution of 0.1 M of (N-succinimidyl-oxycarbonyl-methyl)tris(2,4,6-trimethoxyphenyl) phosphonium bromide (TMPP-Ac-OSu) in CH₃CN:water (2:8, v/v) was added at a molar ratio of 200:1 to 500 µg of *M. smegmatis* protein extract (see Supplemental Methods S1 for bacterial strains, growth conditions, and protein preparation) solubilized in labeling buffer (50 mM Tris-HCl, 8 M urea, 2 M thiourea, pH 8.2, 1 mM phenyl-methylsulfonyl fluoride, 1 mM EDTA, 5 mM TBP [Bio-Rad Laboratories], protease inhibitor mixture [Roche], 10% CH₃CN, 1% SDS). After a quick mix, the reaction was maintained at room temperature for 1 h. Residual derivatizing reagent was quenched by adding a solution of 0.1 M hydroxylamine at room temperature for 1 h.

ID SDS-PAGE separation

M. smegmatis N-terminal labeled protein extract was finally supplemented with glycerol at a concentration of 10%. Proteins were then separated on a 5%–20% 1D SDS-PAGE (20 cm × 20 cm) on a PROTEAN II (Bio-Rad) apparatus at 5 mA for 10 min and 10 mA overnight. The gel was stained with Colloidal blue. The whole lane was systematically cut into 100 bands of ~2 mm, which were processed for enzymatic digestion. Each band was cut into two pieces for further digestion with two different enzymes (trypsin and AspN) and mass spectrometry analysis with an Agilent 1100 Series capillary LC system (Agilent Technologies) coupled to an HCT Ultra ion trap (Bruker Daltonics) (see Supplemental Methods S1 for more details about in-gel digestion and mass spectrometry analysis).

M. smegmatis genome database construction

The complete genome sequence of *M. smegmatis* was downloaded from the TIGR (<http://www.tigr.org/>). Using an in-house script, the 6.98-Mbp sequence was fragmented into regular segments to generate a nucleic acid database, which was imported into a local Mascot server and translated into six reading frames on the fly.

Identification validation

The MS/MS data were analyzed using the Mascot 2. 2. 0. algorithm (Matrix Science) to search against the constructed *M. smegmatis* genome database. For protein identification from internal peptides, the searches were performed with carbamidomethylation of cysteines and

oxidation of methionines specified as variable modifications. For the two digestion modes, trypsin and AspN_ambic, a maximum of one missed cleavage was tolerated; 0.5 Da error in MS and MS/MS search modes was tolerated. Proteins identified with at least three internal peptides with a Mascot ion score greater than 25 were validated. For the estimation of the protein identification false positive rate, a target-decoy database search was performed (for review, see Elias and Gygi 2007). In this approach, peptides are matched against a concatenated database composed of the target database and a decoy database consisting of sequence-reversed entries. Applying the same criteria, we estimated the false positive rate to be <1%.

For N-terminal peptide identification, the searches were performed on genome database subsets under the same conditions as for internal peptides, except that N-terminal modification with TMPP was setup in Mascot (+572.18 Da) as a variable modification, and semi-trypsin or semi-AspN_ambic were used as digestion enzymes. The fragmentation spectra of the putative labeled N-terminal peptides were manually inspected, taking into account the contribution of the N-terminal TMPP group (chromatographic retention time shift, particular fragmentation pathways, and charge states of the peptides).

Comparative genomic approach to correct mycobacteria sequences

We used 16 strains of the *Mycobacterium* genus having a complete genome sequence (see Supplemental Methods S1 for the list of strains of the *Mycobacterium* genus used). A nucleic acid database of the *Mycobacterium* genomic sequences extracted from GenBank was constructed, and a database of the corresponding protein sequences was retrieved from Uniprot. The 443 proteins of *M. smegmatis* with an experimentally determined N terminus were compared with this protein database using BLASTP (Altschul et al. 1997). For each protein, the detected homologs were included in a clustered multiple alignment of complete sequences constructed using the PipeAlign program suite (Plewniak et al. 2003). Sequences sharing at least 70% identity with the *M. smegmatis* reference sequence or belonging to the same cluster within the multiple alignments were selected for N-terminal comparison. If the first amino acids of the sequence to be validated were aligned with the first amino acids of the reference sequence, the sequence was assumed to be correct. Otherwise, the protein sequence was localized on its genomic sequence by a TBLASTN search, and the genomic sequence upstream of or

downstream from the current start codon was searched for alternative initiation codons. The N terminus of the sequence was corrected if an initiation codon was found within ± 9 bp (three amino acids) around the reference start codon. The complete comparative process was performed automatically by TCL/TK scripts, available on request.

Acknowledgments

We thank Raymond Ripp for his assistance during this work and Julie Thompson for a critical reading of the manuscript. This work was supported by institutional funds from INSERM, CNRS, and ULP, by "Protéomique et génie des protéines" (project no. PGP 04-013), ANR-05-BLAN-0407-02, and ANR no. 2007 PFTV 018 01 grants. The "Fondation pour la Recherche Médicale" is also acknowledged for the acquisition of a high-resolution mass spectrometer.

References

- Adamczyk, M., Gebler, J.C., and Wu, J. 1999. Charge derivatization of peptides to simplify their sequencing with an ion trap mass spectrometer. *Rapid Commun. Mass Spectrom.* **13**: 1413–1422.
- Aivaliotis, M., Gevaert, K., Falb, M., Tebbe, A., Konstantinidis, K., Bisle, B., Klein, C., Martens, L., Staes, A., Timmerman, E., et al. 2007. Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*. *J. Proteome Res.* **6**: 2195–2204.
- Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: A new generation of protein database search programs. *Nucleic Acids Res.* **25**: 3389–3402.
- Camus, J.C., Pryor, M.J., Medigue, C., and Cole, S.T. 2002. Re-annotation of the genome sequence of *Mycobacterium tuberculosis* H37Rv. *Microbiology* **148**: 2967–2973.
- Chamot-Rooke, J., van der Rest, G., Dalleu, A., Bay, S., and Lemoine, J. 2007. The combination of electron capture dissociation and fixed charge derivatization increases sequence coverage for O-glycosylated and O-phosphorylated peptides. *J. Am. Soc. Mass Spectrom.* **18**: 1405–1413.
- Choudhary, J.S., Blackstock, W.P., Creasy, D.M., and Cottrell, J.S. 2001. Interrogating the human genome using uninterpreted mass spectrometry data. *Proteomics* **1**: 651–667.
- Cole, S.T., Brosch, R., Parkhill, J., Garnier, T., Churcher, C., Harris, D., Gordon, S.V., Eiglmeier, K., Gas, S., Barry III, C.E., et al. 1998. Deciphering the biology of *Mycobacterium tuberculosis* from the complete genome sequence. *Nature* **393**: 537–544.
- Czeszak, X., Morelle, W., Ricart, G., Tetaert, D., and Lemoine, J. 2004. Localization of the O-glycosylated sites in peptides by fixed-charge derivatization with a phosphonium group. *Anal. Chem.* **76**: 4320–4324.
- Deshayes, C., Perrodou, E., Gallien, S., Euphrasie, D., Schaeffer, C., Van-Dorsselaer, A., Poch, O., Lecompte, O., and Reyrat, J.M. 2007. Interrupted coding sequences in *Mycobacterium smegmatis*: Authentic mutations or sequencing errors? *Genome Biol.* **8**: R20.
- Doerks, T., Bairoch, A., and Bork, P. 1998. Protein annotation: Detective work for function prediction. *Trends Genet.* **14**: 248–250.
- Edwards, M.T., Rison, S.C., Stoker, N.G., and Wernisch, L. 2005. A universally applicable method of operon map prediction on minimally annotated genomes using conserved genomic context. *Nucleic Acids Res.* **33**: 3253–3262.
- Elias, J.E. and Gygi, S.P. 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**: 207–214.
- Fermin, D., Allen, B.B., Blackwell, T.W., Menon, R., Adamski, M., Xu, Y., Ulintz, P., Omenn, G.S., and States, D.J. 2006. Novel gene and gene model detection using a whole genome open reading frame analysis in proteomics. *Genome Biol.* **7**: R35. doi: 10.1186/gb-2006-7-4-r35.
- Ferro, M., Tardif, M., Reguer, E., Cahuzac, R., Bruley, C., Verdat, T., Nugues, E., Vigouroux, M., Vandenbrouck, Y., Garin, J., et al. 2008. Pepline: A software pipeline for high-throughput direct mapping of tandem mass spectrometry data on genomic sequences. *J. Proteome Res.* **7**: 1873–1883.
- Frottin, F., Martinez, A., Peynot, P., Mitra, S., Holz, R.C., Giglione, C., and Meinnel, T. 2006. The proteomics of N-terminal methionine cleavage. *Mol. Cell. Proteomics* **5**: 2336–2349.
- Galperin, M.Y., Walker, D.R., and Koonin, E.V. 1998. Analogous enzymes: Independent inventions in enzyme evolution. *Genome Res.* **8**: 779–790.
- Gevaert, K., Goethals, M., Martens, L., Van Damme, J., Staes, A., Thomas, G.R., and Vandekerckhove, J. 2003. Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides. *Nat. Biotechnol.* **21**: 566–569.
- Gupta, N., Tanner, S., Jaitly, N., Adkins, J.N., Lipton, M., Edwards, R., Romine, M., Osterman, A., Bafna, V., Smith, R.D., et al. 2007. Whole proteome analysis of post-translational modifications: Applications of mass-spectrometry for proteogenomic annotation. *Genome Res.* **17**: 1362–1377.
- Gupta, N., Benhamida, J., Bhargava, V., Goodman, D., Kain, E., Kerman, I., Nguyen, N., Ollikainen, N., Rodriguez, J., Wang, J., et al. 2008. Comparative proteogenomics: Combining mass spectrometry and comparative genomics to analyze multiple genomes. *Genome Res.* **18**: 1133–1142.
- Hirel, P.H., Schmitter, M.J., Dessen, P., Fayat, G., and Blanquet, S. 1989. Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. *Proc. Natl. Acad. Sci.* **86**: 8247–8251.
- Horie, M., Fukui, K., Xie, M., Kageyama, Y., Hamada, K., Sakihama, Y., Sugimori, K., and Matsumoto, K. 2007. The N-terminal region is important for the nuclease activity and thermostability of the flap endonuclease-1 from *Sulfolobus tokodaii*. *Biosci. Biotechnol. Biochem.* **71**: 855–865.
- Huang, Z.H., Shen, T., Wu, J., Gage, D.A., and Watson, J.T. 1999. Protein sequencing by matrix-assisted laser desorption/ionization-postsourc decay-mass spectrometry analysis of the N-Tris(2,4,6-trimethoxyphenyl)phosphine-acetylated tryptic digests. *Anal. Biochem.* **268**: 305–317.
- Jaffe, J.D., Berg, H.C., and Church, G.M. 2004a. Proteogenomic mapping as a complementary method to perform genome annotation. *Proteomics* **4**: 59–77.
- Jaffe, J.D., Stange-Thomann, N., Smith, C., DeCaprio, D., Fisher, S., Butler, J., Calvo, S., Elkins, T., FitzGerald, M.G., Hafez, N., et al. 2004b. The complete genome and proteome of *Mycoplasma mobile*. *Genome Res.* **14**: 1447–1461.
- Kuster, B., Mortensen, P., Andersen, J.S., and Mann, M. 2001. Mass spectrometry allows direct identification of proteins in large genomes. *Proteomics* **1**: 641–650.
- Link, A.J., Robison, K., and Church, G.M. 1997. Comparing the predicted and observed properties of proteins encoded in the genome of *Escherichia coli* K-12. *Electrophoresis* **18**: 1259–1313.
- Lioliou, K., Mavromatis, K., Tavernarakis, N., and Kyrpides, N.C. 2008. The Genomes On Line Database (GOLD) in 2007: Status of genomic and metagenomic projects and their associated metadata. *Nucleic Acids Res.* **36**: D475–D479.
- McDonald, L. and Beynon, R.J. 2006. Positional proteomics: Preparation of amino-terminal peptides as a strategy for proteome simplification and characterization. *Nat. Protocols* **1**: 1790–1798.
- McDonald, L., Robertson, D.H., Hurst, J.L., and Beynon, R.J. 2005. Positional proteomics: Selective recovery and analysis of N-terminal proteolytic peptides. *Nat. Methods* **2**: 955–957.
- Orchard, S. and Hermjakob, H. 2008. The HUPO proteomics standards initiative—easing communication and minimizing data loss in a changing world. *Brief. Bioinform.* **9**: 166–173.
- Oshiro, G., Wodicka, L.M., Washburn, M.P., Yates 3rd, J.R., Lockhart, D.J., and Winzler, E.A. 2002. Parallel identification of new genes in *Saccharomyces cerevisiae*. *Genome Res.* **12**: 1210–1220.
- Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J., et al. 2003. PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Res.* **31**: 3829–3832.
- Roth, K.D., Huang, Z.H., Sadagopan, N., and Watson, J.T. 1998. Charge derivatization of peptides for analysis by mass spectrometry. *Mass Spectrom. Rev.* **17**: 255–274.
- Sadagopan, N. and Watson, J.T. 2000. Investigation of the tris(trimethoxyphenyl)phosphonium acetyl charged derivatives of peptides by electrospray ionization mass spectrometry and tandem mass spectrometry. *J. Am. Soc. Mass Spectrom.* **11**: 107–119.
- Sadagopan, N. and Watson, J.T. 2001. Mass spectrometric evidence for mechanisms of fragmentation of charge-derivatized peptides. *J. Am. Soc. Mass Spectrom.* **12**: 399–409.
- Salgado, H., Moreno-Hagelsieb, G., Smith, T.F., and Collado-Vides, J. 2000. Operons in *Escherichia coli*: Genomic analyses and predictions. *Proc. Natl. Acad. Sci.* **97**: 6652–6657.

- Tanner, S., Shen, Z., Ng, J., Florea, L., Guigo, R., Briggs, S.P., and Bafna, V. 2007. Improving gene annotation using peptide mass spectrometry. *Genome Res.* **17**: 231–239.
- Trivedi, O.A., Arora, P., Sridharan, V., Tickoo, R., Mohanty, D., and Gokhale, R.S. 2004. Enzymic activation and transfer of fatty acids as acyl-adenylates in mycobacteria. *Nature* **428**: 441–445.
- Wang, R., Prince, J.T., and Marcotte, E.M. 2005. Mass spectrometry of the *M. smegmatis* proteome: Protein expression levels correlate with function, operons, and codon bias. *Genome Res.* **15**: 1118–1126.
- Weinstock, G.M. 2000. Genomics and bacterial pathogenesis. *Emerg. Infect. Dis.* **6**: 496–504.
- Xiong, Y., Chalmers, M.J., Gao, F.P., Cross, T.A., and Marshall, A.G. 2005. Identification of *Mycobacterium tuberculosis* H37Rv integral membrane proteins by one-dimensional gel electrophoresis and liquid chromatography electrospray ionization tandem mass spectrometry. *J. Proteome Res.* **4**: 855–861.

Received June 5, 2008; accepted in revised form October 2, 2008.

3. Evolution de la stratégie N-TOP

Fort de notre expérience sur une première application à grande échelle de la stratégie N-TOP, nous avons décidé de faire évoluer la stratégie pour améliorer son efficacité et son automatisation. Pour cela, nous avons modifié deux composantes dans la stratégie :

- La composante instrumentale : le système LC-MS/MS utilisé pour les analyses protéomiques.
- La composante « traitement des données » : la stratégie de recherche des données MS/MS dans les banques de données.

3.1. Au niveau instrumental : une mesure de masse plus juste

Pour améliorer l'identification des peptides, notamment des peptides N-terminaux marqués au TMPP qui ont tendance à présenter un score Mascot plus faible que les peptides internes tryptiques non modifiés, nous avons décidé de réduire l'espace de recherche des données MS/MS. Ceci aura une conséquence directe sur les scores d'identification probabilistes (Partie bibliographique, Chapitre 4. 2.1.1.2.). Pour cela, une stratégie consiste à utiliser un instrument permettant une mesure plus juste de la masse des peptides. Cette mesure plus juste permet lors de l'interrogation Mascot de réduire la tolérance autorisée sur les masses expérimentales des peptides. Ainsi, le nombre de séquences peptidiques candidates considérées par le moteur de recherche est fortement diminué, ce qui permet d'obtenir des scores probabilistes supérieurs sur les identifications [Domon et al., 2006; Bakalarski et al., 2007; Liu et al., 2007].

Les analyses qui avaient été jusque là réalisées sur un système microHPLC (Agilent Series 1100) couplé à une trappe ionique HCT Ultra (Bruker Daltonics) ont ici été réalisées sur un système nanoAcquity UPLC couplé à un Q-TOF Synapt (Waters). De manière générale, les spectromètres de masse de type Q-TOF présentent de meilleures performances résolutive et permettent une mesure de masse plus juste que les trappes ioniques. L'instrument Q-TOF Synapt utilisé est équipé d'un système « lock-mass » qui permet d'infuser un composé de référence à intervalle de temps défini pour une limitation de la dérive d'étalonnage et une justesse de la mesure de masse des composés accrue.

3.2. Au niveau « traitement des données » : une stratégie de recherche dans les banques optimisée

En plus de dépendre de la tolérance autorisée sur la masse du précurseur, l'espace de recherche des données MS/MS dans les banques protéiques est également fonction de :

- La taille de la banque protéique.

- Les enzymes considérées.
- Les modifications des acides aminés autorisées.

Dans la première application à grande échelle de la stratégie N-TOP, l'ensemble des paramètres de recherche utilisé tendait à augmenter significativement l'espace de recherche par rapport à une recherche de données « classique » en analyse protéomique. En effet, le fait d'effectuer la recherche :

- dans une banque génomique plutôt que dans une banque protéique,
- avec une enzyme semi-spécifique plutôt que spécifique,
- avec la modification TMPP potentielle sur chaque peptide plutôt que sans modification TMPP augmentait beaucoup l'espace de recherche considéré.

Dans notre stratégie de recherche optimisée dans les banques, nous nous sommes efforcés de limiter l'espace de recherche à chaque fois que cela était possible.

Pour faciliter l'automatisation du traitement des identifications, la construction de la banque génomique a été modifiée. Ici, le génome complet de *M. smegmatis* a d'abord été traduit dans les 6 cadres de lecture (sans découpage du génome préalable) en utilisant l'application « Transeq » (<http://www.ebi.ac.uk/Tools/emboss/transeq/index.html>) puis toutes les séquences en acides aminés contenues entre 2 codons stop dans un cadre de lecture ont été extraites en séquences protéiques individuelles potentielles (grâce à un script développé au laboratoire). Finalement, une version target-decoy de la banque génomique contenant toutes ces séquences a été créée.

3.3. Stratégie d'analyse protéogénomique

La première partie du protocole analytique de la stratégie N-TOP est restée inchangée, de la préparation des échantillons jusqu'à la séparation des protéines sur gel 1D. Toutefois, ici, les quantités de protéines déposées sur le gel ont pu être diminuées en adéquation avec la sensibilité du système instrumental utilisé pour les analyses nanoLC-MS/MS. Le gel a été coupé systématiquement en 64 bandes horizontales et chaque bande a été digérée à la trypsine seulement. Les peptides de digestion ont été analysés par nanoLC-MS/MS sur un système nanoAcquity UPLC couplé à un Q-TOF Synapt (Waters). Il n'a pas été nécessaire ici d'utiliser un gradient chromatographique de pente aussi douce que pour les analyses issues de la première application de la stratégie N-TOP à l'étude de *M. smegmatis* pour laquelle la séparation chromatographique avait été réalisée sur un système microHPLC (Agilent Technologies) équipé d'une colonne chromatographique de 300 µm de diamètre interne avec des particules de 3.5 µm. En effet, le système nanoAcquity UPLC utilisé ici est équipé d'une colonne utilisant des particules de phase stationnaire plus petites (1.7 µm) et est capable de délivrer et supporter les pressions (résultant des débits de phase mobile) requises avec ce type de colonne (UPLC). Cette diminution de la taille des particules de phase stationnaire permet de diminuer

la Hauteur Equivalente à un Plateau Théorique (HEPT) d'une séparation chromatographique et donc d'améliorer l'efficacité de séparation des peptides et notamment la résolution. L'utilisation d'un gradient 10-70 % acétonitrile délivré sur 120 minutes s'est avérée adéquate pour l'analyse de nos échantillons sur ce système. Les paramètres d'acquisition du spectromètre ont également du être optimisés (sélection des 3 ions précurseurs les plus intenses du spectre MS, durée d'analyse MS de 0.5 s, durée d'analyse MS/MS de 0.6 s) pour être adaptés à la finesse des pics chromatographiques des peptides analysés (~15 s). Des détails supplémentaires sur les expériences de nanoLC-MS/MS sont décrits dans la partie expérimentale générale. Les peptides N-terminaux marqués au TMPP et les peptides internes non modifiés ont été identifiés par interrogation Mascot avec une tolérance de 30 ppm sur la masse des précurseurs et 0.1 Da sur la masse des ions fragments, avec la carbamidomethylation obligatoire des cystéines, en autorisant l'oxydation des méthionines et avec le paramétrage du moteur de recherche prenant en compte les particularités de fragmentation des peptides N-terminaux marqués au TMPP.

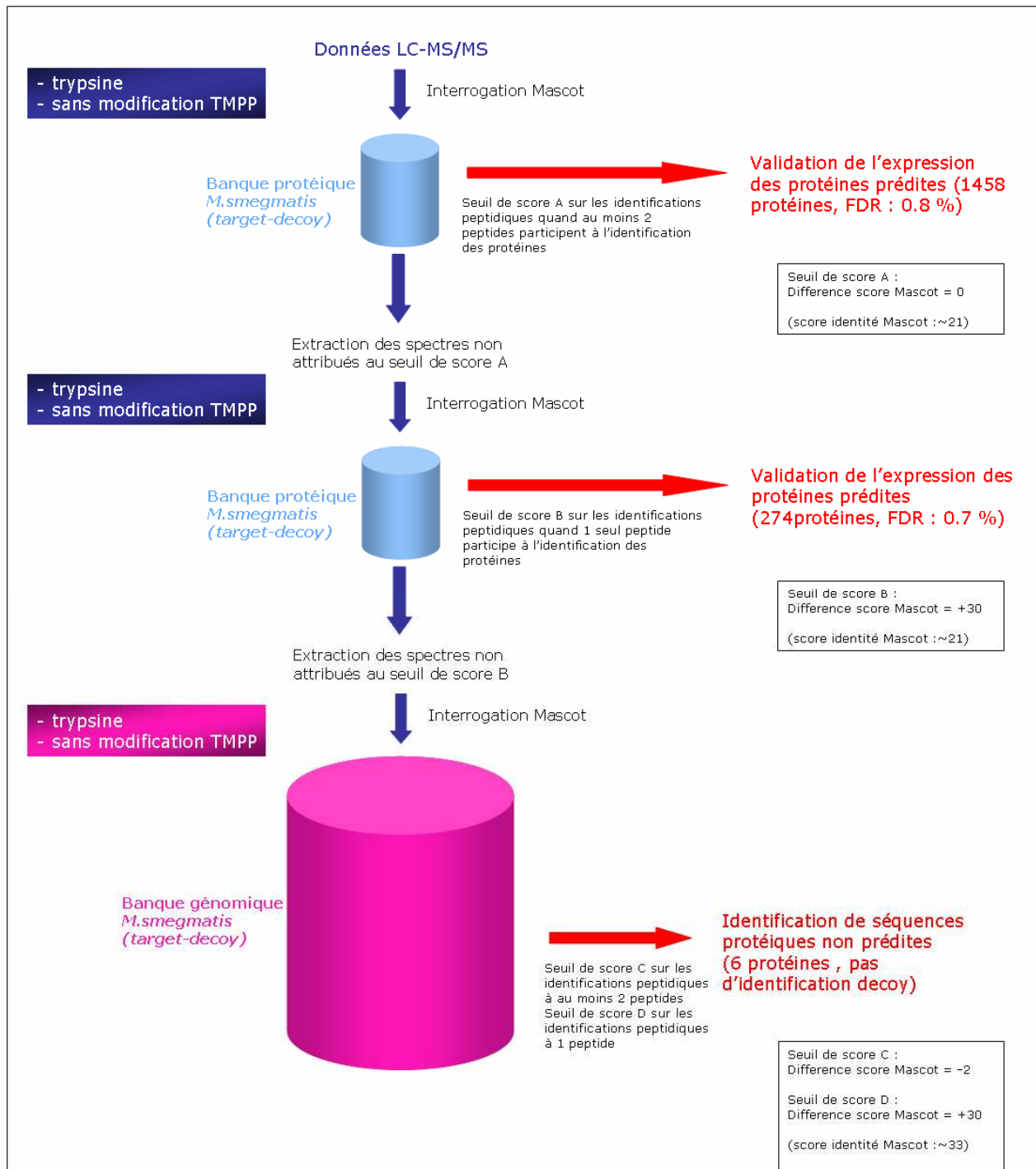
3.3.1. Peptides internes

Pour l'identification des protéines à partir des peptides internes, nous avons donc établi un protocole de recherche séquentiel et d'extraction des spectres non-attribués détaillé en Figure 11. Brièvement,

- Dans une **première étape**, les données MS/MS ont d'abord été soumises à une interrogation Mascot dans la **banque protéique prédite de *M. smegmatis*** (avec intégration des données de correction établies dans la première application de la stratégie N-TOP). Des seuils de scores ont été établis sur les identifications peptidiques quand les protéines étaient identifiées avec au moins 2 peptides (seuil de score A) ou avec 1 seul peptide (seuil de score B, plus stringent que le seuil de score A). Les spectres MS/MS non attribués en utilisant ces seuils de score d'identification ont été extraits.
- Dans une **deuxième étape**, ces spectres MS/MS extraits ont été soumis à une nouvelle interrogation Mascot dans la **banque génomique** de *M. smegmatis*. Des seuils de scores adaptés à cette nouvelle interrogation Mascot ont été fixés (seuil de score C pour les identifications à 2 peptides ou plus, seuil de score D pour les identifications à 1 peptide). Les peptides identifiés dans cette deuxième étape permettent la mise en évidence d'erreurs d'annotation du génome.

Tous ces seuils de score ont été établis pour obtenir un FDR sur les identifications de protéines d'environ 1 %. Les valeurs de seuils de score, les nombres de protéines identifiées ainsi que les FDR associés sont indiqués en Figure 11.

Figure 11 : Protocole de recherche séquentiel des peptides internes. (Difference score Mascot : Score de corrélation Mascot – Score identité Mascot).



3.3.2. Peptides N-terminaux marqués au TMPP

Les identifications des peptides N-terminaux marqués au TMPP ont été réalisées en parallèle dans la banque protéique prédite de *M. smegmatis* (corrections incluses), dans la banque génomique de *M. smegmatis* et dans une sous-banque issue de la banque génomique (Figure 12).

La recherche des données dans la banque protéique a été utilisée pour valider les prédictions sur l'attribution des codons d'initiation des protéines ou les corrections issues de la première application de la stratégie N-TOP (Figure 12 A). Dans cette recherche, la modification TMPP n'a donc été autorisée qu'en N-terminal des protéines et seuls les peptides entièrement tryptiques ont été considérés.

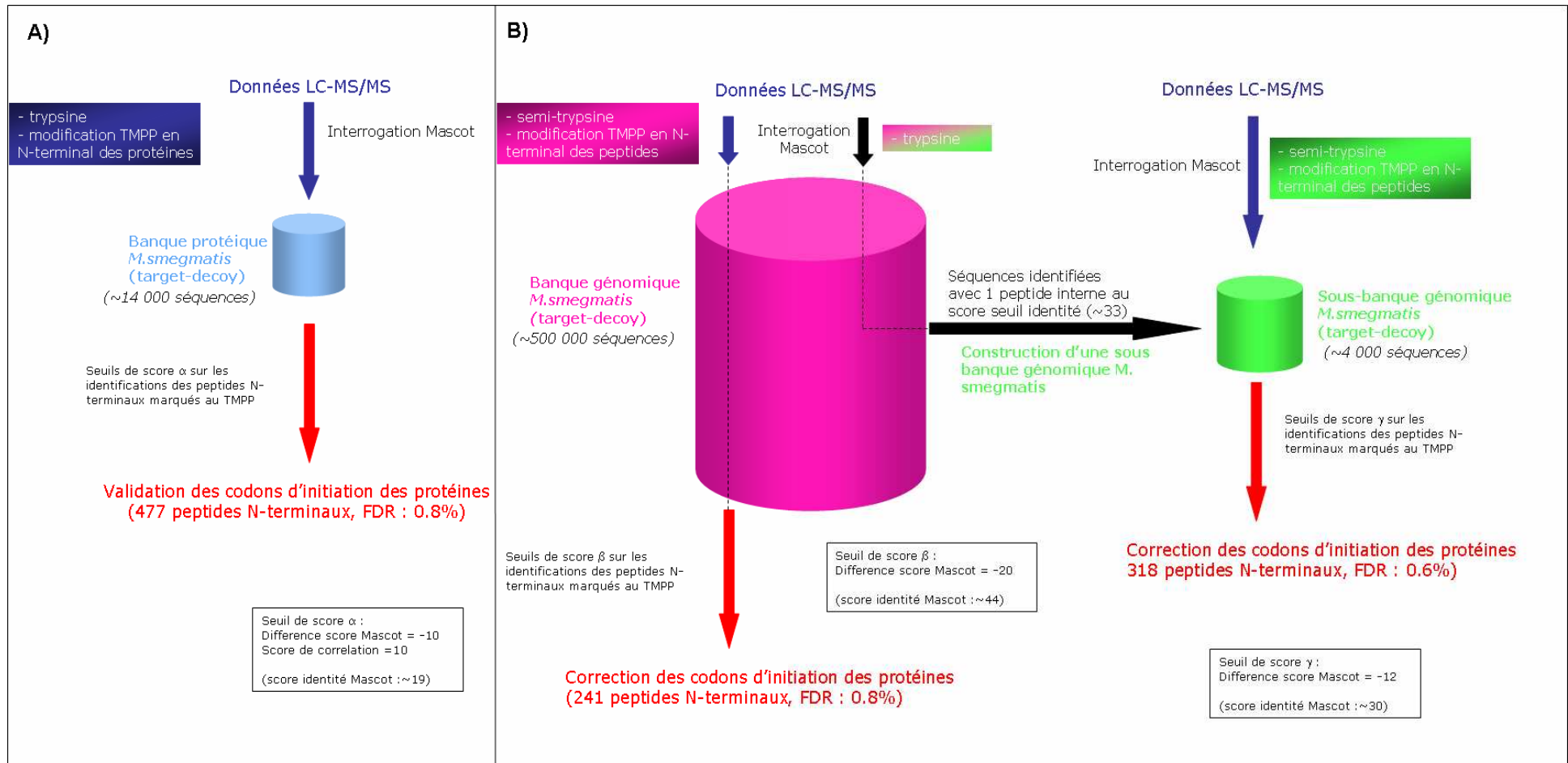
La recherche des données dans la banque génomique (ou la sous-banque génomique) a été utilisée pour déterminer les codons d'initiation des protéines mal prédits (Figure 12 B). Dans ce cas, la modification TMPP a été autorisée en N-terminal de tous les peptides semi-tryptiques. Seuls les peptides N-terminaux marqués au TMPP qui permettent de corriger l'attribution des codons d'initiation des protéines ont été considérés (c'est-à-dire les peptides dont le premier acide aminé est codé par un codon d'initiation, ATG, GTG ou TTG, ou par le premier codon suivant un codon d'initiation et pour lesquels aucun autre peptide n'est identifié en amont dans la même séquence protéique).

La sous-banque génomique est constituée de toutes les séquences issues de la banque génomique pour lesquelles au moins 1 peptide interne a été identifié au score Mascot « identité ».

Pour assurer la spécificité des identifications, les peptides N-terminaux de moins de 6 acides aminés ont du être identifiés avec au moins un peptide interne de la même protéine pour être conservés. Enfin, pour vérifier que les identifications des peptides N-terminaux potentiels marqués au TMPP ne sont pas la conséquence d'un marquage aspécifique plutôt qu'un marquage N-terminal, une nouvelle interrogation Mascot a été réalisée. Cette nouvelle interrogation Mascot a été réalisée en autorisant le marquage TMPP sur la lysine (vu en 1.2.3.2. du chapitre 2) ou sur la tyrosine (reporté dans la littérature [Roth et al., 1998]) sur les données MS/MS correspondant aux peptides N-terminaux marqués potentiels.

Un score seuil a été établi pour chacune des recherches afin d'obtenir un FDR sur les identifications des peptides N-terminaux marqués au TMPP d'environ 1 %. Les valeurs des seuils de score, les nombres de peptides N-terminaux identifiés ainsi que les FDR associés sont indiqués en Figure 12.

Figure 12 : Protocole de recherche en parallèle des peptides N-terminaux marqués au TMPP. A) validation des prédictions sur l'attribution des codons d'initiation des protéines ou les corrections issues de la première application de la stratégie N-TOP. B) détermination des codons d'initiation des protéines mal prédits



3.4. Résultats

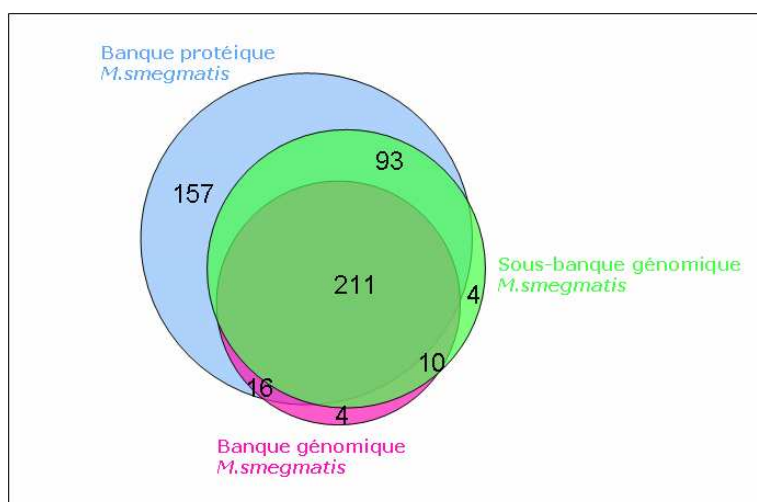
3.4.1. Peptides internes

La stratégie N-TOP optimisée appliquée à l'étude de *M. smegmatis* a permis grâce à l'analyse simultanée des peptides N-terminaux et des peptides internes de valider l'expression de **1732 protéines uniques** et de détecter **6 nouveaux gènes** pour un total de **1738 protéines identifiées** (FDR<1 %).

3.4.2. Peptides N-terminaux marqués au TMPP

Notre stratégie N-TOP optimisée a aussi permis de déterminer le codon d'initiation de **495 protéines** (FDR<1 %) (**477 validations, 18 corrections**) en utilisant une approche totalement automatisée pour la « validation » des identifications des peptides N-terminaux marqués au TMPP. Le protocole de recherche optimisé des données MS/MS dans les banques a permis de limiter l'espace de recherche quand cela était possible en tirant partie des identifications des peptides internes et des prédictions. Bien que ce protocole conduise à une correction des codons d'initiation sur des critères plus stringents que leur validation, il permet de déterminer le maximum de codons d'initiation. Les identifications dans les 3 banques utilisées sont très redondantes mais chacune des banques a fourni des résultats uniques (la distribution des identifications est détaillée dans la Figure 13).

Figure 13 : Distribution des identifications des peptides N-terminaux marqués au TMPP dans les différentes banques



3.4.2.1. Banque protéique

La stratégie de recherche dans la banque protéique prédite (avec les corrections de l'étude précédente incluses) a fourni la part la plus importante des identifications des peptides N-terminaux (**477**) grâce à un espace de recherche très restreint.

Pour s'assurer que la restriction importante de l'espace de recherche n'engendre pas un biais sur les identifications, nous avons procédé à une nouvelle recherche des données MS/MS dans cette banque avec les mêmes paramètres mais en remplaçant la modification TMPP (+572.18 Da) par une modification fictive de masse proche (+614.23 Da). En appliquant les mêmes seuils de score, 9 peptides N-terminaux porteurs de cette modification fictive ont été identifiés dans la composante target de la banque et 7 dans sa composante decoy (FDR : 88 %), ce qui indique qu'il ne semble pas y avoir de biais induit par l'espace de recherche restreint.

3.4.2.2. Sous-banque génomique

La stratégie de recherche dans la sous-banque génomique a fourni le deuxième ensemble le plus important d'identifications des peptides N-terminaux (**318**) grâce à un espace de recherche restreint (séquences identifiées avec au moins 1 peptide interne au score Mascot identité). Cette recherche a permis la correction de **14** codons d'initiation de protéines mal-prédits. Toutefois, le type de banque utilisé ici ne permet pas la correction des codons d'initiation de protéines pour lesquelles aucun peptide interne n'est identifié.

3.4.2.3. Banque génomique

La stratégie de recherche dans la banque génomique a fourni le plus faible nombre d'identifications de peptides N-terminaux (**241**) à cause de l'espace de recherche plus important. Toutefois, c'est la seule stratégie qui permette de mettre en évidence des erreurs de prédictions des codons d'initiation des protéines quand seul le peptide N-terminal marqué au TMPP est identifié (**4 cas**).

3.4.2.4. Réaction aspécifique du TMPP

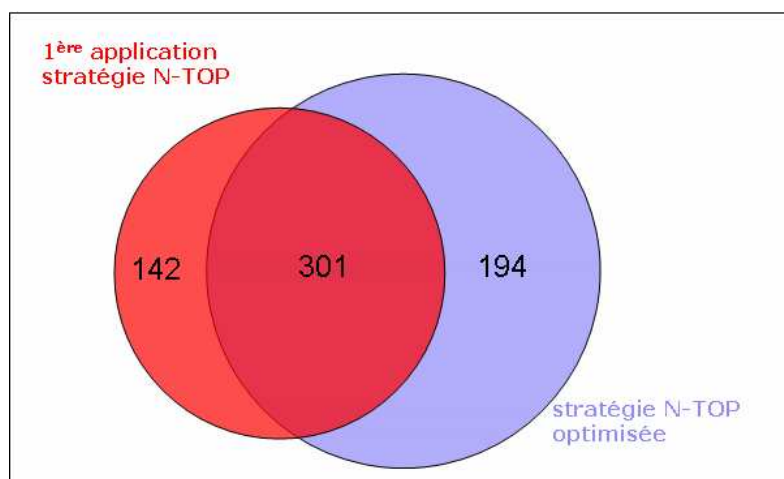
Nous avons retiré des résultats 4 potentiels peptides N-terminaux marqués au TMPP pour lesquels il apparaissait que la réaction était réalisée sur la lysine et pas sur l'acide aminé N-terminal. Aucun potentiel peptide N-terminal marqué au TMPP sur la tyrosine n'a été identifié. Les quelques réactions parasites observables sont donc quasi totalement éliminées de notre ensemble de résultats par notre mode de sélection des potentiels peptides N-terminaux permettant de corriger le codon d'initiation des protéines. Il est à noter que sur la première application de la stratégie N-TOP, aucun

peptide N-terminal potentiel marqué au TMPP n'avait été éliminé de notre ensemble final d'identifications à cause d'une réaction parasite.

3.5. Conclusion

Les optimisations apportées à la stratégie N-TOP ont permis d'identifier un plus grand nombre de peptides N-terminaux marqués au TMPP que lors de la première application de la stratégie N-TOP en analysant moins de bandes de gel 1D et avec une seule digestion enzymatique (Figure 14). Plus des 2/3 des codons d'initiation déterminés lors de la première application de la stratégie l'ont également été avec notre stratégie optimisée. Le 1/3 restant provient essentiellement des identifications issues de la digestion à l'endoprotéinase AspN. Notre nouvelle stratégie optimisée a permis d'identifier 194 peptides N-terminaux supplémentaires et porte le nombre total de peptides N-terminaux uniques identifiés à 637. Ces résultats constituent un des plus grands ensembles d'extrémités N-terminales de protéines déterminées expérimentalement pour un organisme procaryote à notre connaissance. Enfin, la validation des identifications des peptides N-terminaux marqués au TMPP est maintenant réalisée de manière totalement automatisée. Toutes ces nouvelles identifications sont intégrées dans une publication en cours de rédaction sur la ré-annotation complète du génome de *M. smegmatis* (également envoyée au NCBI sous le numéro d'accession CP001663).

Figure 14 : Comparaison des identifications des peptides N-terminaux marqués au TMPP issus des 2 applications de la stratégie N-TOP



La stratégie N-TOP et la stratégie de recherches optimisée de données MS/MS dans les banques développées dans ce chapitre ont également pu être appliquées à 2 autres études qui seront présentées dans les 2 chapitres suivants de cette partie I des résultats.

Chapitre 3 : Application de la stratégie N-TOP pour l'étude d'une enzyme issue d'un organisme dont le génome n'est pas séquencé

Cette étude a été réalisée en collaboration avec l'équipe du Professeur Georg Fuchs du Laboratoire de microbiologie à l'Institut Biologie II de l'Université Albert Ludwigs de Freiburg en Allemagne.

1. Contexte de l'étude

1.1. Métabolisme anaérobie du cholestérol

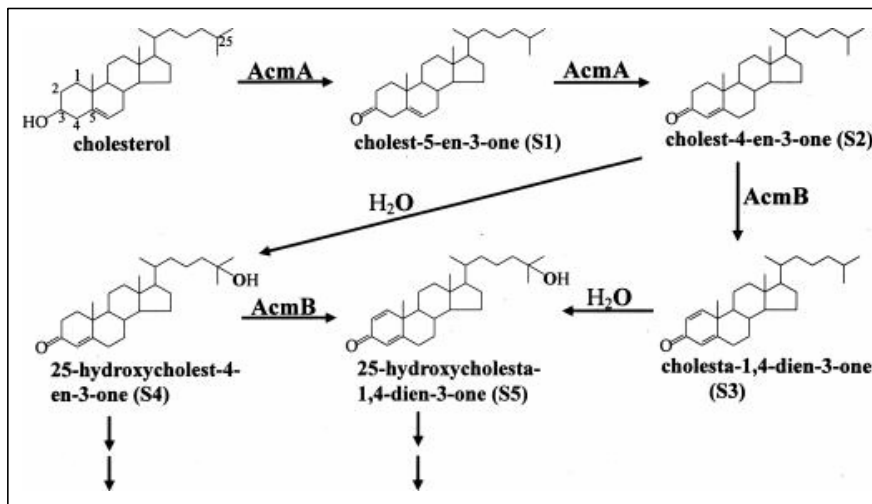
Le caractère ubiquitaire du cholestérol et des stérols apparentés dans l'environnement ont fait de lui une source habituelle de carbone pour les microorganismes. Cependant, la dégradation microbienne du cholestérol est difficile à cause de sa structure chimique complexe, de sa faible solubilité dans l'eau, du faible nombre de groupements fonctionnels et de la présence de quatre structures alicycliques et de deux atomes de carbone quaternaires [Harder et al., 1997]. Plusieurs genres de bactéries, comme *Mycobacterium*, *Corynebacterium* ou *Rhodococcus* par exemple, ont été rapportés comme minéralisant le cholestérol en présence d'oxygène [Kieslich, 1985]. En dépit de l'importance des enzymes métabolisant le cholestérol pour les applications pharmaceutiques ou cliniques, seulement quelques enzymes ont été étudiées en détail. Le métabolisme du cholestérol en l'absence d'oxygène est encore moins bien connu. En dehors de la transformation anaérobie du cholestérol en coprostanol par les bactéries fermentatives de l'intestin [Freier et al., 1994], le champ de connaissance de ce métabolisme anaérobie est très restreint et mérite une attention particulière.

1.2. *Sterolibacterium denitrificans*

La bactérie *Sterolibacterium denitrificans*, l'organisme modèle de cette étude, est une des quelques bactéries qui peuvent complètement minéraliser le cholestérol en CO₂ en l'absence d'oxygène [Harder et al., 1997; Tarlera et al., 2003]. L'équipe du professeur Fuchs a récemment identifié les produits initiaux du métabolisme anaérobie du cholestérol [Chiang et al., 2007]. La voie métabolique commence par l'oxydation du cholestérol en cholest-4-en-3-one (Figure 1 S1) par l'enzyme cholesterol dehydrogenase/isomerase. Cet intermédiaire est ensuite oxydé en cholesta-1,4-dien-3-one (Figure 1 S3) par une enzyme encore inconnue (appelée Cholest-4-en-3-one- Δ 1-

dehydrogenase ou AcmB) mais qui va participer à l'introduction d'une double-liaison entre C-1 et C-2. Dans une 3^{ème} réaction, l'atome C-25 de la chaîne latérale est hydroxylé par l'eau.

Figure 1: les étapes initiales de la voie métabolique anaérobie du cholestérol chez *Sterolibacterium denitrificans*. D'après [Chiang et al., 2008]



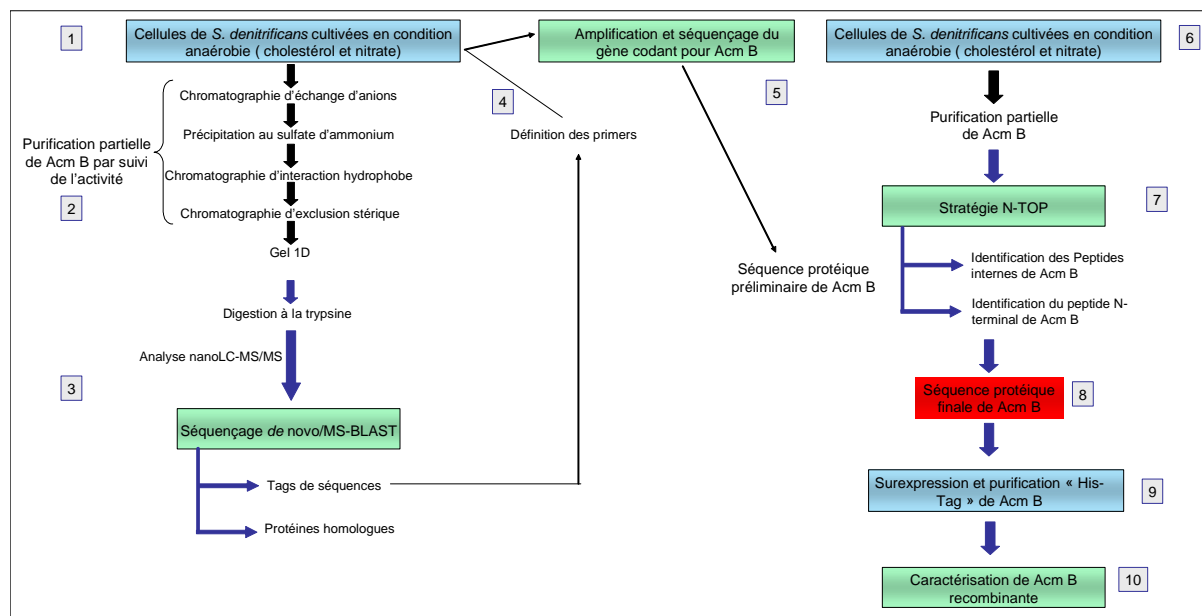
1.3. Objectif de l'étude

Dans ce contexte, pour améliorer la compréhension du métabolisme anaérobie du cholestérol chez *S. denitrificans*, nous nous sommes intéressés à l'étude de la 2^{ème} enzyme de la voie métabolique, AcmB. Pour cela, l'enzyme a été purifiée en suivant son activité. Ensuite, pour déterminer sa séquence, nous avons développé une stratégie combinant l'analyse nanoLC-MS/MS avec identification par approche de novo, le séquençage génomique ciblé et l'approche N-TOP.

2. Stratégie d'analyse et résultats intermédiaires

La stratégie générale d'analyse est présentée en Figure 2. Le détail des expériences est décrit dans la publication des résultats.

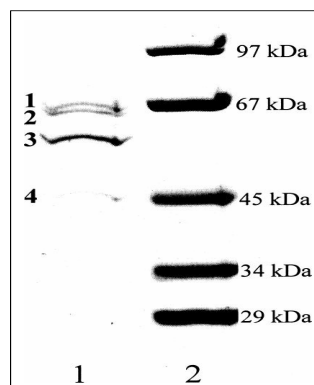
Figure 2 : stratégie générale d'analyse de l'enzyme AcmB



2.1. Préparation des échantillons et purification partielle de l'enzyme d'intérêt

A partir de cellules de *S. denitrificans* cultivées en condition anaérobie (cholestérol et nitrate), un extrait protéique soluble a été préparé. L'enzyme AcmB a été partiellement purifiée par plusieurs étapes successives de fractionnement en suivant la fraction protéique présentant l'activité enzymatique de AcmB. L'étape finale de séparation des protéines par gel 1D a montré la présence de 4 bandes (Figure 3). Les 4 bandes d'intérêt ont été découpées puis digérées à la trypsine.

Figure 3 : Gel 1D de la fraction protéique partiellement purifiée en AcmB. D'après [Chiang et al., 2008]



2.2. Analyse protéomique et identifications des peptides

Les peptides de digestion ont ensuite été analysés par nanoLC-MS/MS sur un système nanoHPLC-Chip (Agilent Series 1100) couplé à une trappe ionique HCT Ultra (Bruker Daltonics). La puce microfluidique (nanoHPLC-Chip, Agilent technologies) présente l'avantage d'intégrer colonne de chargement, colonne analytique et pointe nanospray avec une réduction importante des volumes morts de connection et donc un affinement des pics chromatographiques. Etant donné la très faible complexité des mélanges analysés ici, une puce microfluidique comprenant une colonne analytique « courte » (43 mm) a été utilisée. Le nombre de spectres MS/MS moyennés a été paramétré à une valeur relativement importante (8) pour obtenir des spectres MS/MS finaux présentant un rapport signal/bruit suffisant pour une analyse *de novo* des spectres MS/MS facilitée.

L'ensemble des données MS/MS a été traité par séquençage *de novo* à l'aide du logiciel PEAKS Studio (Bioinformatics Solutions) [Ma et al., 2003] qui fournit un ensemble de fragments de séquence en acides aminés. L'identification des peptides a été réalisée dans cette étude par une approche *de novo* car le génome de *S. denitrificans* n'est pas séquencé. Les fragments de séquence ont ensuite été soumis au programme MS-BLAST qui a permis d'identifier des protéines homologues à celles présentes dans l'échantillon. L'approche *de novo*/MS-BLAST a notamment mis en évidence que la protéine d'une des bandes (bande 3) présentait de fortes similarités avec la protéine 3-ketosteroid- Δ 1-dehydrogenase de *Comamonas testosteroni* (Tableau 1). Cette protéine catalyse le même genre de réaction que AcMB dans le métabolisme oxydatif de stéroïdes sans chaîne aliphatique en position C-17 (du type testostérone) [Choi et al., 1995; Morii et al., 1998; van der Geize et al., 2002]. La protéine de la bande 3 constitue donc un bon candidat pour AcMB.

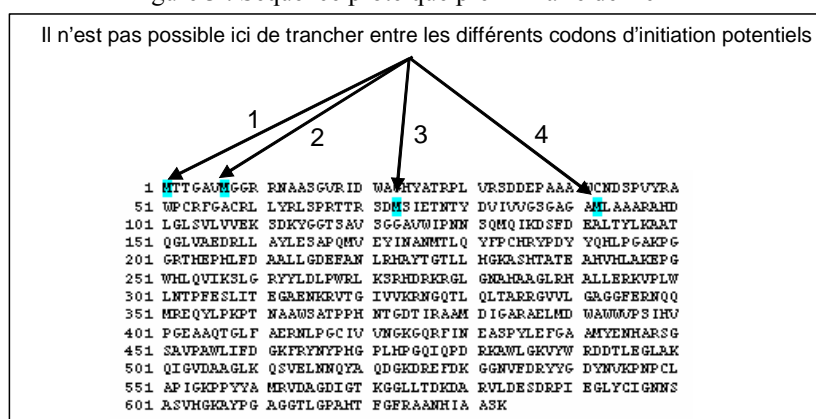
Tableau 1 : Résultats d'identification de la bande 3 par approche *de novo*/MS-BLAST

Bande	Protéine homologue identifiée	Numéro d'accèsion	Peptides identifiés	Total score MS-BLAST
4	3-ketosteroid- Δ 1-dehydrogenase (<i>Comamonas testosteroni</i>)	Q7WSH6	MAALQYDVLVVGSGAGAMLGAIK RGVVLGAGGFER GAVPAWLLFD AEWTATPVGGNTGDAHR VGSAGAMLAAR YFLDYPWR GAGGTLGVTVTFGR PMGPLMPG YAGAGSTLGPAMTFAFR	378

2.3. Amplification et séquençage de la séquence génomique codant pour l'enzyme Acm B.

Les amorces nécessaires à l'amplification et au séquençage de la séquence génomique codant pour AcmB ont été déduites à partir des séquences peptidiques identifiées (Figure 2). Finalement une séquence protéique préliminaire de AcmB a été obtenue par clonage et séquençage. La séquence protéique obtenue pour AcmB n'est que préliminaire puisque s'il n'y a pas d'ambiguïté sur le codon de terminaison de cette séquence, le codon d'initiation par contre n'est pas défini (Figure 3). A ce stade, il n'est pas possible de trancher sur les différents codons d'initiation potentiels.

Figure 3 : Séquence protéique préliminaire de AcmB

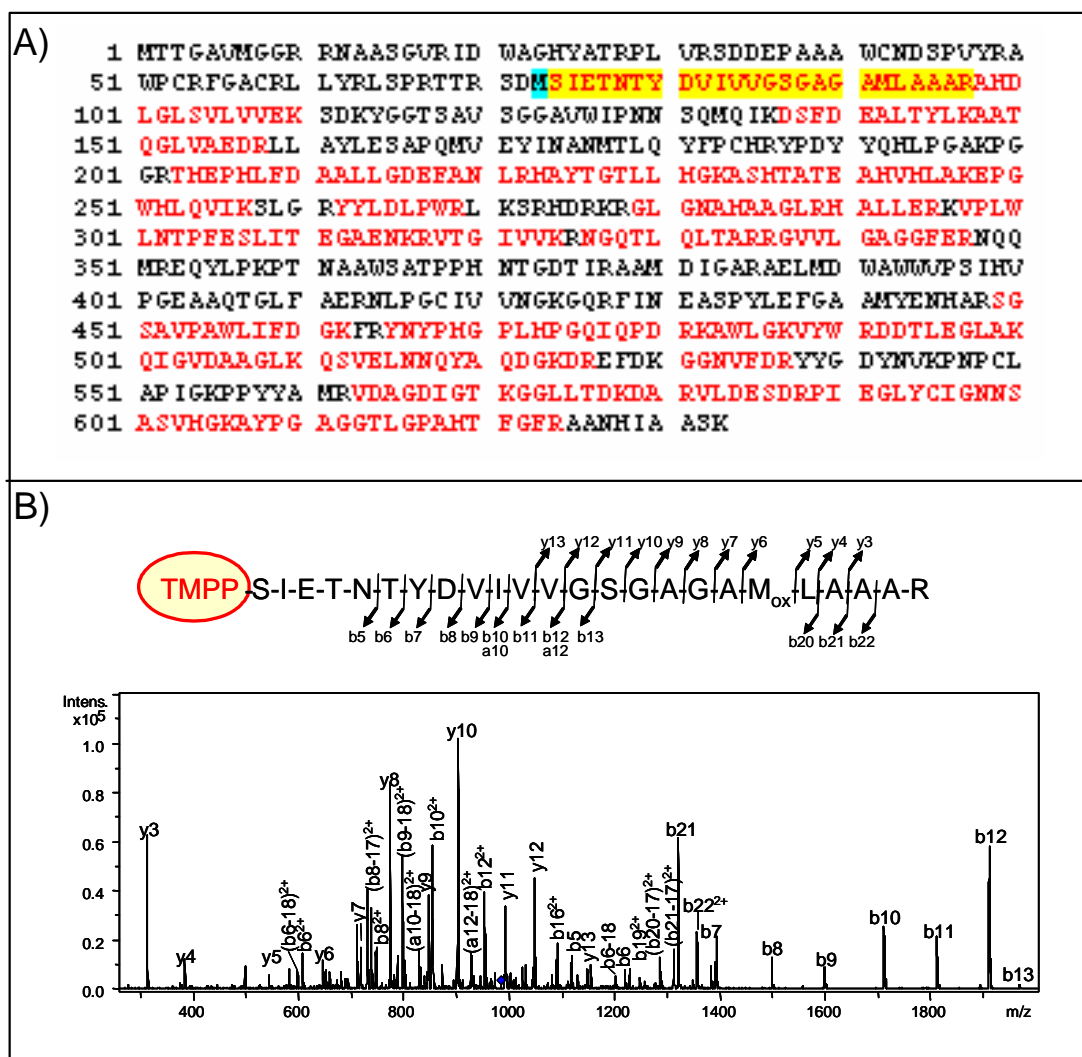


2.4. Approche N-TOP

Pour déterminer le codon d'initiation de la protéine AcmB et donc obtenir sa séquence exacte, nous avons appliqué l'approche N-TOP à l'étude de l'enzyme AcmB partiellement purifiée avant dépôt sur gel 1D (Figure 2). La bande d'intérêt a ensuite été découpée puis digérée à la trypsine. Les peptides de digestion ont été analysés sur le même système nanoLC-MS/MS à la différence qu'ici, une puce microfluidique comprenant une colonne analytique « longue » (150 mm) a été utilisée pour assurer la séparation chromatographique du peptide N-terminal marqué au TMPP de l'ensemble des peptides internes.

Les données MS/MS ont été soumises à une interrogation Mascot dans la séquence protéique préliminaire de AcmB (en utilisant les mêmes paramétrages que ceux utilisés dans la première application de la stratégie N-TOP sur *M. smegmatis*, Partie I des résultats, Chapitre 2. 2.). Nous avons ainsi pu identifier 28 peptides tryptiques internes de AcmB (avec un score Mascot>20) et le peptide N-terminal marqué au TMPP (avec un score Mascot de 71) (Figure 4).

Figure 4 : Résultats d'analyse de l'approche N-TOP. A) Identification Mascot des peptides d'AcMB (en rouge). Le peptide N-terminal identifié est surligné en jaune. Le codon d'initiation déterminé est surligné en bleu. B) Spectre MS/MS de l'ion triplement chargé de m/z 985.45 correspondant au peptide N-terminale de AcMB marqué au TMPP



2.5. Surexpression et purification de la protéine AcMB

Le gène codant pour AcMB a été cloné et exprimé dans *E. coli* sous la forme d'une protéine de fusion marquée en histidines à son extrémité C-terminale. La protéine recombinante a ensuite été purifiée par chromatographie d'affinité.

2.6. Caractérisation de la protéine AcMB recombinante

Finalement, la protéine recombinante purifiée AcMB a pu être caractérisée. Il fut procédé à son analyse fonctionnelle, à l'analyse de ses propriétés catalytiques, de sa stabilité, de sa spécificité pour des substrats et de ses propriétés moléculaires.

3. Résultats publiés

Les résultats obtenus sur la première partie de l'étude (jusqu'à l'obtention de la séquence protéique préliminaire de Acmb) ont été intégrés dans une publication dont je ne suis pas un des co-auteurs donc je ne l'intégrerai pas dans cette thèse [Chiang et al., 2008]. Les résultats obtenus dans la deuxième partie de l'étude (à partir de la stratégie N-TOP) ont fait l'objet d'une publication acceptée en octobre 2007 dans *Applied and Environmental microbiology*. Les étapes de surexpression, purification et caractérisation de la protéine recombinante Acmb sont détaillées dans la publication.

4. Conclusion

L'ensemble de la stratégie développée dans cette étude apparaît bien adaptée pour l'étude de protéines qui présentent un intérêt important mais qui appartiennent à des organismes plus ou moins « exotiques » et pour lesquels le génome n'est pas séquencé. L'approche « *de novo*/MS-BLAST » était indispensable ici. En effet la recherche des données nanoLC-MS/MS dans les banques protéiques généralistes par interrogation Mascot n'avait pas permis d'identifier de protéine. Cette étude illustre aussi la possibilité d'appliquer la stratégie N-TOP « à façon » dans le but de déterminer exactement les bornes d'une séquence protéique puisque tant que celles-ci ne sont pas connues, la caractérisation d'une enzyme telle que Cholest-4-en-3-one- Δ 1-dehydrogenase (Acmb) n'a pas de sens. Par contre, une fois la séquence protéique parfaitement connue, la caractérisation de Acmb recombinante a bien montré que cette enzyme catalysait l'oxydation attendue de cholest-4-en-3-one en cholesta-1, 4-dien-3-one.

Finalement, l'enzyme Cholest-4-en-3-one- Δ 1-dehydrogenase (Acmb) est potentiellement intéressante pour des applications pharmaceutiques car sa faculté à catalyser la formation d'une double-liaison entre C-1 et C-2 des 3-ketosteroides pourrait être utilisée pour la production d'analogues 1-dehydro de corticostéroïdes comme la prednisone ou la prednisolone

Cholest-4-En-3-One- Δ^1 -Dehydrogenase, a Flavoprotein Catalyzing the Second Step in Anoxic Cholesterol Metabolism[∇]

Yin-Ru Chiang,¹ Wael Ismail,¹ Sébastien Gallien,² Dimitri Heintz,²
Alain Van Dorselaer,² and Georg Fuchs^{1*}

Mikrobiologie, Institut Biologie II, Albert-Ludwigs-Universität Freiburg, Freiburg, Germany,¹ and Laboratoire de Spectrométrie de Masse Bio-Organique, CRNS, ECPM, Université Louis Pasteur, Strasbourg, France²

Received 28 August 2007/Accepted 31 October 2007

The anoxic metabolism of cholesterol was studied in the denitrifying bacterium *Sterolibacterium denitrificans*, which was grown with cholesterol and nitrate. Cholest-4-en-3-one was identified before as the product of cholesterol dehydrogenase/isomerase, the first enzyme of the pathway. The postulated second enzyme, cholest-4-en-3-one- Δ^1 -dehydrogenase, was partially purified, and its N-terminal amino acid sequence and tryptic peptide sequences were determined. Based on this information, the corresponding gene was amplified and cloned and the His-tagged recombinant protein was overproduced, purified, and characterized. The recombinant enzyme catalyzes the expected Δ^1 -desaturation (cholest-4-en-3-one to cholesta-1,4-dien-3-one) under anoxic conditions. It contains approximately one molecule of FAD per 62-kDa subunit and forms high molecular aggregates in the absence of detergents. The enzyme accepts various artificial electron acceptors, including dichlorophenol indophenol and methylene blue. It oxidizes not only cholest-4-en-3-one, but also progesterone (with highest catalytic efficiency, androst-4-en-3,17-dione, testosterone, 19-nortestosterone, and cholest-5-en-3-one. Two steroids, corticosterone and estrone, act as competitive inhibitors. The dehydrogenase resembles 3-ketosteroid- Δ^1 -dehydrogenases from other organisms (highest amino acid sequence identity with that from *Pseudoalteromonas haloplanktis*), with some interesting differences. Due to its catalytic properties, the enzyme may be useful in steroid transformations.

The microbial metabolism of cholesterol and related steroids has attracted great attention because of its potential impact on pharmaceutical and clinical applications (14, 26). Detailed studies have been conducted on the oxic metabolism of cholesterol, as shown in Fig. 1A (12). The pathway for oxic cholesterol metabolism starts with the transformation of cholesterol to cholest-4-en-3-one by a bifunctional cholesterol dehydrogenase (oxidase) (8, 14). The subsequent reaction involves the introduction of a double bond between C-1 and C-2 to produce cholesta-1,4-dien-3-one. So far the enzyme that catalyzes such a reaction in oxic cholesterol metabolism has not been reported. However, 3-ketosteroid- Δ^1 -dehydrogenase (KSTD), which catalyzes the same type of reaction, has been reported for oxic metabolism of other steroids lacking the aliphatic side chain on the C-17 position, e.g., testosterone (4, 10, 16, 24, 25). Further metabolism of cholesterol requires a monooxygenase catalyzed hydroxylation of the side chain to an alcohol group, followed by its oxidation to a carboxyl group, and subsequent repeated cycles of β -oxidation to degrade the side chain. Recently, genes involved in the oxic cholesterol metabolism by *Rhodococcus* sp. strain RHA1 were discovered (26).

The denitrifying bacterium *Sterolibacterium denitrificans*, our model organism in this study, is one of a few bacteria that can completely mineralize cholesterol to CO₂ in the absence of oxygen (7, 23). We have recently identified the initial products of anoxic cholesterol metabolism (3). The pathway starts with the oxidation of cholesterol to cholest-4-en-3-one (S2; Fig. 1B).

This intermediate is further oxidized to the corresponding 1,2-dehydro derivative, cholesta-1,4-dien-3-one (S3). In a third reaction, the tertiary C-25 atom of the side chain becomes hydroxylated with water as the oxygen donor, providing a first example for such unprecedented, but expected oxygen-independent hydroxylation reactions (Fig. 1B).

Based on these findings, we searched for the genes coding for the initial enzymes of the pathway by using a reverse genetics approach (Chiang et al., submitted for publication). A 43-kbp chromosomal DNA fragment was identified and sequenced, which carried one open reading frame (*acmB* [anoxic cholesterol metabolism enzyme B]) that showed high sequence similarity to KSTD from several bacteria. *AcmB* therefore may be a good candidate for the second enzyme in the anoxic pathway (transformation of S2 to S3; Fig. 1B). To substantiate the role of *acmB* in the anoxic cholesterol metabolism, we cloned *acmB* and expressed it in *Escherichia coli*. Here, we show that the recombinant His-tagged *AcmB* (*AcmB*_{his}) is a cholest-4-en-3-one- Δ^1 -dehydrogenase that catalyzes the Δ^1 desaturation under anoxic condition.

MATERIALS AND METHODS

Materials and bacterial strains. The chemicals used were of analytical grade and were purchased as reported before (3). *Sterolibacterium denitrificans* Chol-1S^T (DSMZ13999) (23) was obtained from the Deutsche Sammlung für Mikroorganismen und Zellkulturen (Braunschweig, Germany). Primers were purchased from Biomers (Ulm, Germany).

Bacterial cultures and preparation of cell extract. Large-scale cultivation was performed in a 200-liter fermentor. *S. denitrificans* was grown anaerobically as described before (3). *E. coli* cells that produced the recombinant *AcmB*_{his} were grown as described below. All steps used for preparation of cell extracts were performed at 4°C. Frozen cells were suspended in twice the volume of 20 mM 3-morpholinopropanesulfonic acid (MOPS)-K⁺ (pH 7.9) containing 0.1 mg of DNase I ml⁻¹. Cells were broken by passing the cell suspension through a French

* Corresponding author. Mailing address: Mikrobiologie, Institut Biologie II, Schänzlestr. 1, D-79104 Freiburg, Germany. Phone: (49) 761-2032649. Fax: (49) 761-2032626. E-mail: georg.fuchs@biologie.uni-freiburg.de.

[∇] Published ahead of print on 9 November 2007.

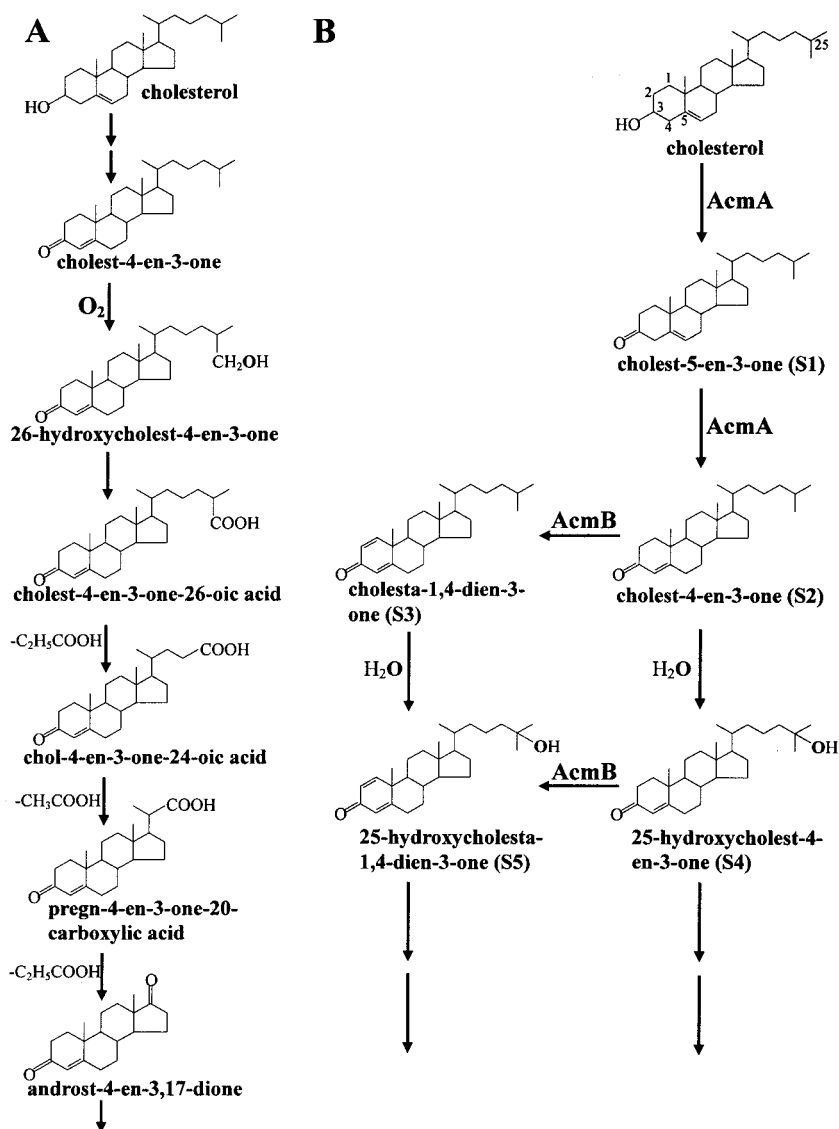


FIG. 1. The initial steps in the microbial cholesterol metabolism. (A) Established oxic pathway. (B) Proposed initial steps in the anoxic pathway, as studied in *S. denitrificans*.

pressure cell (American Instruments, Silver Spring, MD) twice at 137 MPa. The soluble protein fraction was obtained by centrifugation at $100,000 \times g$ for 1.5 h. Protein was determined by the Bradford method (2).

Cloning and expression of the *acmB* gene coding for cholest-4-en-3-one- Δ^1 -dehydrogenase. The *acmB* gene in *S. denitrificans* was cloned in *E. coli* with a pEXP5-CT/TOPO TA expression kit according to the manufacturer's instructions (Invitrogen, Karlsruhe, Germany). Primers were designed to remove the native stop codon and place the gene in frame with the DNA coding for a C-terminal peptide containing six histidines. The standard protocol was used to isolate chromosomal DNA from *S. denitrificans* cells (1). The gene was amplified from genomic DNA of *S. denitrificans* by using *Taq* DNA polymerase and the following primer pair: ATGAGCATCGAAACCAACACATATGACGT (*Acmb_for*, forward primer based on the N-terminal amino acid sequence of *Acmb*)/CTTGCTCGCGGATATGGTTGGCCGCG (*Acmb_rev*, reverse primer based on the C-terminal amino acid sequence of *Acmb*). The following PCR program was used: 95°C for 5 min followed by 40 cycles of 95°C for 45 s, 60°C for 45 s, and 72°C for 3 min, and then 72°C for 5 min. The recombinant plasmid (pYRE2) was transformed into One Shot TOP 10 chemically competent *E. coli* cells (Invitrogen, Karlsruhe, Germany) and *E. coli* K38/pGP1-2 cells (a kind gift from J. Heider, Darmstadt, Germany) (22) for maintenance and overexpression, respectively. The cloned PCR amplicon was verified by DNA se-

quencing and restriction analysis. For gene overexpression experiments, *E. coli* K38/pGP1-2 containing pYRE2 was incubated in LB medium containing ampicillin ($100 \text{ mg liter}^{-1}$) and kanamycin (50 mg liter^{-1}) at 30°C until the optical density at 578 nm (OD_{578}) reached 0.7. The culture was heat shocked at 42°C for 30 min and then incubated at 37°C for further 2 h. *E. coli* cells were harvested at OD_{578} of 1.4 and stored at -70°C . The soluble protein fraction from frozen cells was prepared as described above.

Purification of *Acmb_{his}* from *E. coli* K38/pGP1-2. A Ni^{2+} -chelating Sepharose affinity column (50 ml; Amersham Biosciences Europa, Freiburg, Germany) was equilibrated with 150 ml of 20 mM MOPS- K^+ buffer (pH 7.9) containing 0.25 M KCl (buffer A). The protein sample (100 ml of soluble protein fraction) was applied to the column followed by washing with buffer A containing 0.15 M imidazole. The fusion protein was subsequently eluted with buffer A containing 0.35 M imidazole. The flow rate was 1 ml min^{-1} , and the eluate was collected in 2-ml fractions. The salts were removed from the combined active fractions by using a prepacked PD-10 desalting column (8.3 ml) according to the manufacturer's instructions (Amersham Biosciences Europa).

Gel filtration chromatography. The native mass was estimated by using gel filtration chromatography and ferritin (450 kDa), catalase (240 kDa), alcohol dehydrogenase (150 kDa), bovine serum albumin (67 kDa), and ovalbumin (45 kDa) as the molecular mass standards. The recombinant *Acmb_{his}* purified by

Ni²⁺-chelating Sepharose affinity column was concentrated to 2 ml by ultrafiltration (30-kDa-cutoff membrane; Amicon, Witten, Germany) and applied to a 289-ml Superdex 200 HiLoad 26/60 column (Amersham Biosciences Europa). The column was equilibrated with 2 bed volumes of equilibration buffer containing 20 mM MOPS-K⁺ (pH 7.9) and 0.1 M KCl. The purified recombinant protein was loaded onto the column and eluted at a flow rate of 1.0 ml min⁻¹, and protein elution was monitored at 280 nm.

Partial purification of cholest-4-en-3-one- Δ^1 -dehydrogenase (AcmB) from *S. denitrificans*. AcmB was purified from 50 g of *S. denitrificans* cells grown anaerobically with cholesterol and nitrate. The basic buffer (buffer B) used in the following purification steps was 20 mM MOPS-K⁺ (pH 7.9). The enzyme activity was tested by the thin-layer chromatography (TLC) method (see below).

(i) **DEAE-Sepharose.** A column with a 70-ml bed volume (Amersham Biosciences Europa) was equilibrated with 2 volumes of buffer B. The soluble protein fraction (100 ml) was loaded onto the column, followed by washing with 2 volumes of buffer B. The flowthrough and the wash were collected in 3-ml fractions. The flow rate was 2 ml min⁻¹.

(ii) **Ammonium sulfate precipitation.** The active DEAE pool was further fractionated by ammonium sulfate precipitation. The protein sample was precipitated with ammonium sulfate at 50% saturation, followed by centrifugation at 20,000 \times g for 20 min. The supernatant was discarded, and the protein pellet was redissolved in 50 ml of buffer B containing 0.5 M ammonium sulfate.

(iii) **Hydrophobic interaction chromatography.** A 25-ml butyl-Sepharose 4 fast flow column (Amersham Biosciences Europa) was equilibrated with 2 volumes of buffer B containing 500 mM ammonium sulfate. The protein sample from the ammonium sulfate treatment was applied to the column, followed by washing with buffer B containing 0.5 M ammonium sulfate. The column was further washed sequentially with buffer B containing 0.1 M ammonium sulfate and buffer B. The active protein was eluted with buffer B containing 0.3% Tween 20. The flow rate was 2 ml min⁻¹, and the eluate was collected in 2-ml fractions. The active fractions were pooled (50 ml) and concentrated to 2.5 ml with ultrafiltration (30-kDa-cutoff membrane).

(iv) **Gel filtration chromatography.** A 289-ml Superdex 200 HiLoad 26/60 column was equilibrated with 2 bed volumes of buffer B containing 0.1 M KCl. The protein sample (2.5 ml) was loaded onto the column and was then eluted at a flow rate of 0.4 ml min⁻¹. Fractions of 3 ml were collected, and the dominant protein peak showing activity was concentrated to 1 ml by ultrafiltration (30-kDa-cutoff membrane).

Enzyme assays. AcmB activity was routinely measured at 37°C under anoxic conditions.

(i) **TLC assay.** AcmB dehydrogenase activity was measured by monitoring the formation of cholesta-1,4-dien-3-one. The assay mixture (0.6 ml) containing soluble protein fractions from *S. denitrificans* (0.2 to 5.0 mg), 100 mM MOPS-K⁺ (pH 7.9), 5 mM 2,6-dichlorophenolindophenol (DCPIP) and 0.5 mM cholest-4-en-3-one was incubated for 2 h with shaking. Assay mixtures were first extracted twice by an equal volume of ethyl acetate, and the ethyl acetate fraction was concentrated under vacuum. The extracted products were separated on silica gel aluminum TLC plates (silica gel 60 F₂₅₄; Merck, Darmstadt, Germany). The solvent system used was *n*-hexane-ethyl acetate (65:35 [vol/vol]). The product was visualized under UV light at 254 nm.

Electron acceptor specificity of purified recombinant AcmB_{his} was tested by TLC assay under anoxic conditions by incubating the following electron acceptors (5 mM) with 0.5 mM cholest-4-en-3-one, 90 μ g enzyme, and 100 mM MOPS-K⁺ (pH 7.9) in a final volume of 0.8 ml: NaNO₃, K₃[Fe(CN)₆], NAD⁺, NADP⁺, phenazine methosulfate, DCPIP, methylene blue, and methyl viologen. After 2 h of incubation, the steroids were extracted and separated on TLC plates as described above. The position corresponding to the substrate was scraped off carefully from the TLC plate and extracted with 0.6 ml ethyl acetate three times. The ethyl acetate fractions were combined and evaporated to dryness. Acetonitrile (0.5 ml) was used to redissolve the residual substrate. The amount of residual steroid substrate was determined spectrophotometrically at 238 nm with cholest-4-en-3-one (5 to 20 μ g) as the standard.

(ii) **Spectrophotometric assays.** The substrate-dependent reduction of DCPIP was followed at 600 nm (ϵ_{600} 11,000 M⁻¹ cm⁻¹ [pH 6.0]). The assay mixtures (0.5 ml) contained enzyme (1.2 to 5.6 μ g), 100 mM Na⁺-phosphate (pH 6.0), 100 μ M DCPIP, and 100 μ M steroid substrates dissolved in 2-propanol (10 μ l). The apparent K_m and V_{max} values of different steroid substrates were determined by varying the substrate concentration (1 to 500 μ M) and keeping a fixed concentration of DCPIP (100 μ M). Apparent K_m and V_{max} values were obtained by fitting Michaelis-Menten curve to the data using the Prism GraphPad software package (Graphpad Software, San Diego, CA). Buffers used to determine the pH optimum were 100 mM Na⁺-phosphate (pH 5.0 to 6.0), 100 mM MOPS-KOH (pH 6.5 to 7.9), and 100 mM glycine-NaOH (pH 8.5 to 10.0). To determine any

inhibitory effect, the assay was started by addition of 100 μ M cholest-4-en-3-one after incubation for 3 min with the tested compound. The type of inhibition was determined by incubating purified AcmB_{his} with the inhibitor (10 to 100 μ M) for 3 min before starting the reaction with the substrate (cholest-4-en-3-one; 25 to 200 μ M).

SDS-PAGE. Sodium dodecyl sulfate-polyacrylamide gel electrophoresis (SDS-PAGE; 10% polyacrylamide) was performed with a discontinuous buffer system (13). Protein bands were visualized by Coomassie blue staining. The following molecular mass protein standards were used: phosphorylase B, 97 kDa; bovine serum albumin, 67 kDa; ovalbumin, 45 kDa; lactate dehydrogenase, 34 kDa; carbonic anhydrase, 29 kDa; and lysozyme, 14 kDa.

N-terminal labeling of native cholest-4-en-3-one- Δ^1 -dehydrogenase. Partially purified native AcmB protein in equilibration buffer containing 50 mM Tris-HCl (pH 8.2) and 100 mM KCl (1 ml) was treated with 5*N*-(succinimidyl)oxycarbonylmethyltris(2,4,6-trimethoxyphenyl) phosphonium bromide (TMPP) under conditions allowing the labeling of only the N-terminal amino group of proteins. After N-terminal TMPP labeling, the sample volume was reduced to 20 μ l using Vivaspin microconcentrator (30-kDa-cutoff membrane) (Vivascience, Hannover, Germany). The sample was separated by SDS-PAGE (12% polyacrylamide). The proteins in the gel were precipitated by using 200 ml of the mixture containing 50% ethanol and 3% phosphoric acid for 2 h. After three washes with 500 ml distilled water, the gel was stained by the colloidal Coomassie method (17) and then cut into 2-mm-thick slices. The gel slices between 50 and 70 kDa were digested with trypsin, and the resulting tryptic peptides were extracted and analyzed by liquid chromatography-tandem mass spectrometry (LC-MS/MS).

LC-MS/MS. Nanoscale LC-MS/MS analysis of the tryptic peptides was performed using an Agilent 1100 series high-performance liquid chromatography (HPLC)-Chip/MS system (Agilent Technologies, Palo Alto, CA) coupled to an HCT Ultra ion trap (Bruker Daltonics, Bremen, Germany). The voltage applied to the capillary cap was optimized to -1,850 V. For MS/MS experiments, the system was operated with automatic switching between MS and MS/MS modes. The three most abundant peptides, preferring doubly charged ions, were selected on each MS spectrum for further isolation and fragmentation. The MS/MS scanning was performed in the ultrascan resolution mode at a scan rate of 26,000 m/z s⁻¹. A total of six scans were averaged to obtain an MS/MS spectrum. The complete system was fully controlled by ChemStation (Agilent Technologies) and EsquireControl (Bruker Daltonics) softwares.

Analysis of the flavin cofactor. The flavin cofactor from the purified AcmB_{his} (1.9 mg ml⁻¹) was extracted and identified as described previously (15). The eluate was monitored at 450 nm in this study. Under these conditions, the retention times were 6.3 min for flavin adenine dinucleotide (FAD), 10.2 min for flavin mononucleotide (FMN), and 28.0 min for riboflavin. The amount of extracted flavin cofactor was determined by comparison of the peak area with that of flavin standard (0.5 to 2.0 μ g of FAD).

GC-EI-MS. Gas chromatography-electron impact mass spectrometry (GC-EI-MS) analysis was performed as follows. MS analyses were performed on a Agilent Technologies 6890N gas chromatograph equipped with a split/splitless programmed temperature injector and an HP-5MS fused silica column (30 m by 0.25 mm; film thickness, 0.25 μ m) connected to a Agilent 5975 inert MSD quadrupole spectrometer. The mass spectrometer was operated in electron impact (EI) mode at 70 eV, and spectra were recovered over a mass range of m/z 50 to 500 with a cycle time of 1.6 scan s⁻¹. The oven temperature was programmed from 180°C to 300°C at 6°C min⁻¹ and then held isothermal for 10 min. The other conditions were as follows: helium split, 1:10; constant flow, 1.5 ml min⁻¹; transfer line, 280°C; and MS source temperature, 230°C. Samples were injected using the split mode with a ratio of 1:10 via an autoinjector, and the temperature of the injector was 280°C.

UV-visible spectroscopy. The UV-visible absorption spectrum of purified recombinant AcmB_{his} protein was obtained by using a CARY 100 Bio UV-visible spectrometer (Varian Deutschland, Darmstadt, Germany). The concentration of the recombinant enzyme was 2.4 μ M in 20 mM MOPS-K⁺ (pH 7.9).

Nucleotide sequence accession number. The sequence data reported in this study have been deposited in the GenBank database under accession no. EU004090.

RESULTS

Cholest-4-en-3-one- Δ^1 -dehydrogenase (AcmB) activity in cell extract from *S. denitrificans* and partial sequence determination of the partially purified protein. The soluble protein fraction from *S. denitrificans* grown anaerobically on cholest-

TABLE 1. Utilization of electron acceptors of the recombinant enzyme^a

Artificial electron acceptor	Residual cholest-4-en-3-one concn (μM)
Negative control.....	442 ± 13
NaNO ₃	383 ± 10
K ₃ [Fe(CN) ₆].....	99 ± 8
NAD ⁺	307 ± 15
NADP ⁺	326 ± 19
Phenazine methosulfate.....	263 ± 13
DCPIP.....	36 ± 5
Methylene blue.....	38 ± 11
Methyl viologen.....	340 ± 18

^a The assay mixtures containing 90 μg purified enzyme, 0.5 mM cholest-4-en-3-one, and 5 mM each of different electron acceptors were incubated at 37°C for 2 h under anoxic conditions. Samples were treated and measured as described in Materials and Methods. The negative control was without addition of any electron acceptor. Data are averages ± standard deviations of three experimental measurements.

terol transformed cholest-4-en-3-one (S2) to cholesta-1,4-dien-3-one (S3) (Fig. 1B). The reaction took place in the presence of DCPIP as an artificial electron acceptor. Based on this enzyme assay the enzyme was purified in four steps. SDS-PAGE analysis of the active enzyme pool after the final fractionation revealed four protein bands (data not shown). The amino acid sequences of tryptic peptides from the dominant protein in the active pool exhibited high sequence similarity (24% coverage) to 3-ketosterol-Δ¹-dehydrogenase from *Comamonas testosteroni* (Q7WSH6). Thus, Acmb likely represents the target enzyme. The N-terminal amino acid sequence (SIETNTYDVIVVVGSGAGAMLAAAR) of the putative cholest-4-en-3-one-Δ¹-dehydrogenase was determined by LC-MS/MS using a strategy of N-terminal protein derivatization as described in Materials and Methods.

Heterologous overexpression of the *acmB* gene. Recent studies on the genes involved in anoxic cholesterol metabolism by *S. denitrificans* revealed several ORFs that might play a role in that metabolic route (Chiang et al., submitted). A putative gene, *acmB*, encodes a protein of 61.4 kDa that is highly similar to KSTD from several bacteria. In addition, the determined N-terminal amino acid sequence and the amino acid sequences of the tryptic peptides of native Acmb purified from *S. denitrificans* were identical to those deduced from the *acmB* gene.

The *acmB* gene was cloned and expressed in *E. coli* as a C-terminal His-tagged fusion protein. The cholest-4-en-3-one-Δ¹-dehydrogenase activity was measured spectrophotometrically assay monitoring the reduction of DCPIP. The recombinant protein was purified from the soluble protein fraction of *E. coli* by one Ni²⁺-chelating affinity chromatography step. SDS-PAGE analysis of the active protein pool revealed a single protein band with an apparent molecular mass of ca. 60 kDa (data not shown). This molecular mass agrees well with that deduced from the *acmB* gene (61.4 kDa).

Functional analysis of the recombinant Acmb_{his}. The recombinant protein catalyzed the transformation of cholest-4-en-3-one to cholesta-1,4-dien-3-one in the presence of DCPIP under anoxic conditions. The identity of the reaction product was confirmed by TLC, HPLC, and GC-EI-MS. Its EI mass spectrum (data not shown) matched exactly with that of cho-

TABLE 2. Substrate spectrum of the recombinant Acmb_{his}^a

Substrate	<i>K_m</i> (μM)	<i>k_{cat}</i> (s ⁻¹)	<i>k_{cat}/K_m</i> (s ⁻¹ M ⁻¹)
Cholest-4-en-3-one	42 ± 6	69 ± 4	1.6 × 10 ⁶
Androst-4-en-3,17-dione	100 ± 12	82 ± 5	8.0 × 10 ⁵
Progesterone	4 ± 1	43 ± 2	1.2 × 10 ⁷
Testosterone	15 ± 3	33 ± 2	2.2 × 10 ⁶
19-Nortestosterone	50 ± 9	26 ± 1	5.2 × 10 ⁵
Cholest-5-en-3-one	9 ± 2	14 ± 0	1.6 × 10 ⁶
Androsta-1,4-dien-3,17-dione		<1	
5α-Cholestane		<1	
Cholesterol		<1	
Cholesta-3,5-diene		<1	
Cortisone		<1	
Corticosterone		<1	
Estrone		<1	

^a The apparent *K_m* values and *k_{cat}* values (which refer to monomers of 62 kDa) were measured in triplicate at 37°C. The assay mixture contained 100 μM DCPIP, 2.8 μg enzyme, 100 mM Na⁺-phosphate (pH 6.0), and different concentrations of the respective substrate (1 to 500 μM).

lest-1,4-dien-3-one in a spectral database (www.aist.go.jp/RIODB/SDBS).

Catalytic properties and stability. To facilitate the kinetic measurements with purified recombinant Acmb_{his}, cholest-4-en-3-one oxidation with artificial electron acceptors was followed by monitoring product formation. DCPIP and methylene blue were the best (Table 1). DCPIP was thus chosen for all further experiments and for spectrophotometric assays.

The recombinant enzyme had a pH optimum of 6.0 and a temperature optimum of 40°C (data not shown). Purified Acmb_{his} retained activity without significant loss when kept at 4°C for 2 weeks or frozen for several months in 20 mM MOPS-K⁺ (pH 7.9). In addition, no decrease in activity was observed after 20 h of exposure to air at 4°C, thus indicating that purified Acmb_{his} is not oxygen labile. The recombinant enzyme had a specific activity of 67 μmol min⁻¹ mg protein⁻¹ and an apparent *K_m* for the natural substrate cholest-4-en-3-one of 42 ± 6 μM (Table 2). The turnover rate per monomer was 69 ± 4 s⁻¹, and the catalytic efficiency, *k_{cat}/K_m*, was calculated to be 1.6 × 10⁶ s⁻¹ M⁻¹ (Table 2).

Substrate specificity. The substrate specificity of the recombinant Acmb_{his} was tested by screening various steroids (Table 2), whose structures are shown in Fig. 2. The enzyme showed activity with steroids having a carbonyl group at position C-3. 3-Hydroxysteroids such as cholesterol and estrone could not serve as substrates. Moreover, the enzyme showed no activity with cortisone and corticosterone, thus suggesting that a functional group (carbonyl or hydroxyl group) at the C-11 position of the steroid substrates may hinder the dehydrogenase activity. In addition, the enzyme was also inactive with steroids lacking a functional group at the C-3 position such as 5α-cholestane and cholesta-3,5-diene. However, a C = C structure at C4 position or a methyl group at C-19 position of steroid substrates was not necessary for the activity since the enzyme was active toward cholest-5-en-3-one and 19-nortestosterone (Table 2). In summary, progesterone is an excellent substrate for the enzyme, even much better than the natural substrate.

Inhibition and inactivation. The recombinant Acmb_{his} was extremely sensitive to Ag⁺; 1.5 μM of AgNO₃ completely

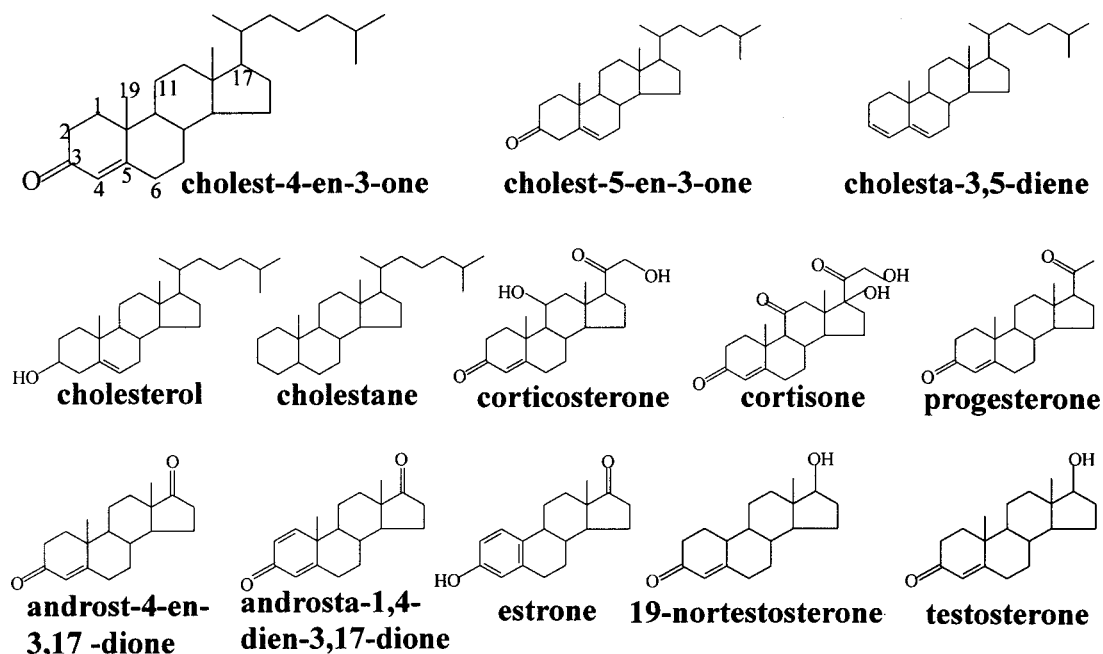


FIG. 2. Structures of different steroids that were tested by using the spectrophotometric assay with DCPIP. The carbon numbering system of steroids is as shown for cholest-4-en-3-one.

inactivated the dehydrogenase activity. Also, 100 μM of Cu^{2+} completely inactivated the enzyme. The enzyme was sensitive to neither thiol reagents such as iodoacetamide (500 μM) nor metal chelators such as EDTA (500 μM). Corticosterone and estrone strongly inhibited the enzyme activity. According to Dixon plot analysis, the two compounds act as competitive inhibitors. The K_i value for corticosterone was $28 \pm 10 \mu\text{M}$, and that for estrone was $68 \pm 18 \mu\text{M}$ by fitting the following modified Michaelis-Menton equation (5) to the data: $v = [E]_0[S]k_{\text{cat}}/[S] + K_m(1 + [I]/K_i)$.

Other steroids such as androsta-1,4-dien-3,17-dione, 5α -cholestane, cholesterol, cholesta-3,5-diene, and cortisone (50 μM) exhibited no inhibitory effects, when 100 μM substrate (cholest-4-en-3-one) was used (for the structures, see Fig. 2).

Molecular properties. The UV-visible spectrum of the oxidized AcmB_{his} showed an absorption spectrum typical for flavoproteins (10, 16). Three dominant peaks were observed at 270, 368, and 454 nm as well as a shoulder at around 470 nm (Fig. 3). The flavin cofactor was readily dissociated from the holoprotein by treatment with perchloric acid. The free flavin cofactor was identified by HPLC as FAD (data not shown). The amount of FAD was also determined by HPLC to be 0.65 mol FAD per mol of the recombinant enzyme based on the molecular mass of 62.2 kDa. In addition, taking into account the UV-visible spectrum of the oxidized AcmB_{his} (Fig. 3), 1 mol of enzyme contained 1.21 mol of FAD on the basis of the molar absorption coefficient for free FAD of $11.3 \text{ mM}^{-1} \text{ cm}^{-1}$ at 450 nm and the protein concentration (2.4 μM) determined by the Bradford method.

Gel filtration chromatography was applied to determine the molecular mass of the recombinant AcmB_{his} . The active enzyme resulting from the affinity chromatography appeared in the void volume, indicating that AcmB_{his} forms a massive ag-

gregate (>600 kDa). To find out whether the native cholest-4-en-3-one- Δ^1 -dehydrogenase also exists as aggregates in the parent strain *S. denitrificans*, the soluble cell protein was fractionated via the same gel filtration column. Here also, the enzyme activity was found exclusively in the void volume.

DISCUSSION

Catalytic properties. The purified cholest-4-en-3-one- Δ^1 -dehydrogenase from *S. denitrificans* catalyzes the expected ox-

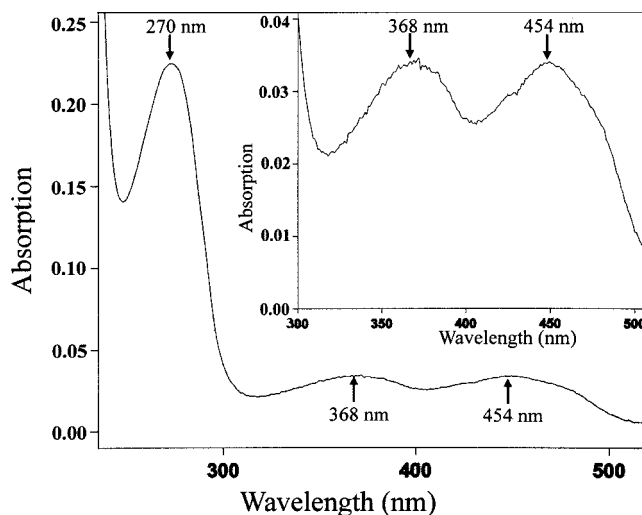


FIG. 3. UV-visible absorption spectrum of the purified recombinant enzyme. The concentration of the recombinant enzyme was 2.4 μM in 20 mM MOPS- K^+ (pH 7.9). The molar absorption coefficient for free FAD at 450 nm is $11.3 \text{ mM}^{-1} \text{ cm}^{-1}$ (pH 7).

<i>S. denitrificans</i>	2	IETNTYDVIIVGSGAGAMLAARAHDLGSLVVEKSDKYGGSAVSGGA
<i>C. testosteroni</i>	1	MAEQEYDLIVGSGAGAMLGAIRAQEQQLKTLVEKTELEGGGSAISGGG
<i>P. haloplanktis</i>	6	DVEKNYDVIIVGSGAGAMTAAALFAADQGNSTLVEKTDKYGGGSAISGGG
<i>S. pealeana</i>	12	DNQTFDIIIVGSGAGAMASSIVASDHGSLVVEKSGKFGGSAISGGG
<i>R. eutropha</i>	8	MQPSEFDVIVGSGAGMMLAACRAADRGLSVVLEKSSQYGGGSAVSGGA
<i>N. farcinica</i>	6	LDPHSYDVIIVGSGAGMATAALTAHHGLRIVVLEKAAHYGGSTARSGGG

FIG. 4. Partial sequence alignment of the deduced amino acid sequence of the *acmB* gene from *S. denitrificans* with FAD-dependent 3-KSTD from *Comamonas testosteroni* (BAC81692), *Pseudoalteromonas haloplanktis* (YP_340631), *Ralstonia eutropha* (YP_728801), *Nocardia farcinica* (YP_116669), and succinate dehydrogenase (flavoprotein subunit) from *Shewanella pealeana* (ZP_01603424). Conserved amino acids are highlighted by letters in black, dark gray, and light gray, depending on their similarity from high to low. A conserved consensus sequence for the proposed FAD-binding region, G-X-G-X-X-G(A)-(X)₁₇-E, is underlined.

dation of cholest-4-en-3-one to cholesta-1,4-dien-3-one. The enzyme was competitively inhibited by corticosterone and estrone. Interestingly, cortisone, which is highly similar to corticosterone, is neither a substrate nor an inhibitor for AcMB. It is unknown why a hydroxyl group at C-11 position (corticosterone in this study) allows binding of the steroid as inhibitor, whereas the corresponding carbonyl group on the same carbon (cortisone in this study) does not. The enzyme accepts various artificial electron acceptors. However, the natural electron acceptor remains to be identified.

In general, all steroids, which could serve as substrates for the enzyme, have in common a carbonyl group at the C-3 position. The importance of the 3-keto group is supported by the observation that cholesterol, 5 α -cholestane, and cholesta-3,5-diene are not substrates for AcMB. This was also reported for other KSTDs (4, 10). Therefore, it is tempting to postulate an important role of the carbonyl group and the corresponding binding sites of the enzyme (discussed by Kadode et al. [11]).

The enzyme was sensitive to thiol reagents such as AgNO₃ and CuCl₂, suggesting that thiol groups may play an important role in the catalysis by AcMB. On the contrary, iodoacetamide did not inhibit the enzyme activity, thus apparently excluding the involvement of thiol groups. In addition, no conserved cysteine residues were found in amino acid sequences of AcMB and related KSTDs.

Potential application. The double-bond formation between C-1 and C-2 of 3-ketosteroids is of pharmaceutical interest. Examples are the production of 1-dehydro analogues of some adrenal cortical steroids such as prednisone and prednisolone (4). The enzyme KSTD carrying out such a reaction has been characterized from a variety of microorganisms during the past two decades (4, 10, 16, 20, 24, 25). However, these studies focused only on steroids without an aliphatic side chain at the C-17 position. Although the same Δ^1 -desaturation reaction takes place during the (an)oxic metabolism of cholesterol (oxidation of cholest-4-en-3-one to cholesta-1,4-dien-3-one; Fig. 1A and B), the corresponding enzyme has not been characterized (3, 12).

Molecular properties. The enzyme is highly similar to KSTDs from several microorganisms (highest amino acid sequence identity [56%] with that from *Pseudoalteromonas haloplanktis* [YP_340631]). These enzymes are flavoproteins, and they possess near their N termini a highly conserved FAD-binding domain, which comprises the G-X-G-X-X-G(A) sequence as an adenine binding motif (6, 20, 24). Furthermore, a conserved glutamate residue is found near the FAD-binding motif and was reported also to be involved in the binding of FAD (18). In addition, this consensus FAD-binding domain of

AcMB is preceded (residues 9 to 12) and followed (residues 28 to 35) by many hydrophobic amino acids, resulting in a typical $\beta\alpha\beta$ structure (16, 21). A similar motif is found in other flavoproteins such as succinate dehydrogenase from *Shewanella pealeana* (Fig. 4). In many succinate dehydrogenases, FAD is covalently bound to a highly conserved histidine residue located about 20 residues downstream of the adenine-binding motif (19). Despite the high similarity (54% amino acid sequence identity) between the AcMB and succinate dehydrogenase from *S. pealeana*, the conserved histidine residue is lacking in AcMB and the KSTDs (Fig. 4).

Comparison with other KSTDs. Despite the overall sequence similarity and similar catalyzed reaction, AcMB differs from KSTDs in several aspects. First, cortisone serves as a substrate for many characterized KSTDs (4, 10), while it does not for AcMB. Second, these flavoproteins may utilize quinones as physiological electron acceptors. However, they react differently toward various artificial electron acceptors. For instance, phenazine methosulfate is an excellent electron acceptor for KSTD from *Nocardia corallina*, while K₃[Fe(CN)₆], NAD⁺, and NADP⁺ cannot be used at all (10). For AcMB, DCPIP, methylene blue, and K₃[Fe(CN)₆] are highly efficient electron acceptors, whereas NAD⁺, NADP⁺, and phenazine methosulfate are modest ones. Third, the pH optimum is around pH 6 for AcMB and about pH 10 for KSTDs from *N. corallina* and *Arthrobacter simplex* (4, 10). Fourth, KSTDs from *N. corallina* and *A. simplex* are monomeric proteins (4, 10), whereas AcMB forms soluble oligomeric aggregates, as is well known for some enzyme families: e.g., the members of the aldehyde dehydrogenase enzyme family (9). The aggregation of AcMB may result from hydrophobic interactions between the monomers. However, it cannot be excluded that formation of monomers may be triggered by a small amount of detergent or other compounds in vivo.

ACKNOWLEDGMENTS

This work was supported by the Deutscher Akademischer Austauschdienst (DAAD) by awarding a fellowship to Y.-R.C.

We thank Nasser Gad'on, Freiburg, for expert technical assistance and Christine Schaeffer, Strasbourg, for expert analysis in gas chromatography-electron impact mass spectrometry.

REFERENCES

1. Ausubel, F. M., R. Brent, R. E. Kingston, D. D. Moore, J. G. Seidman, J. A. Smith, and K. Struhl. 1987. Current protocols in molecular biology. John Wiley and Sons, New York, NY.
2. Bradford, M. 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Anal. Biochem.* 72:248-254.
3. Chiang, Y. R., W. Ismail, M. Müller, and G. Fuchs. 2007. Initial steps in the

- anoxic metabolism of cholesterol by the denitrifying *Sterolibacterium denitrificans*. *J. Biol. Chem.* **282**:13240–13249.
4. **Choi, K. P., I. Molnár, M. Yamashita, and Y. Murooka.** 1995. Purification and characterization of the 3-ketosteroid- Δ^1 -dehydrogenase of *Arthrobacter simplex* produced in *Streptomyces lividans*. *J. Biochem.* **117**:1043–1049.
 5. **Fersht, A.** 1999. The basic equation of enzyme kinetics, p. 103–131. *In* M. R. Julet (ed.), *Structure and mechanism in protein science*. W. H. Freeman and Company, New York, NY.
 6. **Florin, C., T. Köhler, M. Grandguillot, and P. Plesiat.** 1996. *Comamonas testosteroni* 3-ketosteroid- $\Delta^4(5\alpha)$ -dehydrogenase: gene and protein characterization. *J. Bacteriol.* **178**:3322–3330.
 7. **Harder, J., and C. Probian.** 1997. Anaerobic mineralization of cholesterol by a novel type of denitrifying bacterium. *Arch. Microbiol.* **167**:269–274.
 8. **Horinouchi, S., H. Ishizuka, and T. Beppu.** 1991. Cloning, nucleotide sequence, and transcriptional analysis of the NAD(P)-dependent cholesterol dehydrogenase gene from a *Nocardia* sp. and its hyperexpression in *Streptomyces* spp. *Appl. Environ. Microbiol.* **57**:1386–1393.
 9. **Ichihara, K., Y. Noda, C. Tanaka, and M. Kusunose.** 1986. Purification of aldehyde dehydrogenase reconstitutively active in fatty alcohol oxidation from rabbit intestinal microsomes. *Biochim. Biophys. Acta* **878**:419–425.
 10. **Itagaki, E., T. Wakabayashi, and T. Hatta.** 1990. Purification and characterization of 3-ketosteroid- Δ^1 -dehydrogenase from *Nocardia corallina*. *Biochim. Biophys. Acta* **1038**:60–67.
 11. **Kadode, M., H. Matsushita, K. Suzuki, and E. Itagaki.** 1994. Essential arginine residue(s) of 3-ketosteroid- Δ^1 -dehydrogenase, p. 339–342. *In* K. Yagi (ed.), *Flavins and flavoproteins*. Walter de Gruyter & Co., Berlin, Germany.
 12. **Kieslich, K.** 1985. Microbial side-chain degradation of sterols. *J. Basic Microbiol.* **25**:461–474.
 13. **Laemmli, U. K.** 1970. Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**:680–685.
 14. **MacLachlan, J., A. T. L. Wotherspoon, R. O. Ansell, and C. J. W. Brooks.** 2000. Cholesterol oxidase: sources, physical properties and analytical applications. *J. Steroid Biochem. Mol. Biol.* **72**:169–195.
 15. **Mohamed, M. E., A. Zaar, C. Ebenau-Jehle, and G. Fuchs.** 2001. Reinvestigation of a new type of aerobic benzoate metabolism in the proteobacterium *Azoarcus evansii*. *J. Bacteriol.* **183**:1899–1908.
 16. **Morii, S., C. Fujii, T. Miyoshi, M. Iwami, and E. Itagaki.** 1998. 3-Ketosteroid- Δ^1 -dehydrogenase of *Rhodococcus rhodochrous*: sequencing of the genomic DNA and hyperexpression, purification, and characterization of the recombinant enzyme. *J. Biochem.* **124**:1026–1032.
 17. **Neuhoff, V., R. Stamm, I. Pardowitz, N. Arold, W. Ehrhardt, and D. Taube.** 1990. Essential problems in quantification of proteins following colloidal staining with Coomassie brilliant blue dyes in polyacrylamide gels, and their solution. *Electrophoresis* **11**:101–117.
 18. **Nishiya, Y., and T. Imanaka.** 1996. Analysis of interaction of the *Arthrobacter* sarcosine oxidase and the coenzyme flavin adenine dinucleotide by site-directed mutagenesis. *Appl. Environ. Microbiol.* **62**:2405–2410.
 19. **Pearling, S. L., A. C. Black, F. D. C. Manson, B. Ward, S. K. Chapman, and G. A. Reid.** 1992. Sequence of the gene encoding flavocytochrome c from *Shewanella putrefaciens*: a tetraheme flavoenzyme that is a soluble fumarate reductase related to the membrane-bound enzymes from other bacteria. *Biochemistry* **31**:12132–12140.
 20. **Plesiat, P., M. Grandguillot, S. Harayama, S. Vragar, and Y. Michel-Briand.** 1991. Cloning, sequencing, and expression of the *Pseudomonas testosteroni* gene encoding 3-ketosteroid- Δ^1 -dehydrogenase. *J. Bacteriol.* **173**:7219–7227.
 21. **Schreuder, H. A., J. M. Van der Laan, W.-G. J. Hol, and J. Drenth.** 1991. The structure of *p*-hydroxybenzoate hydroxylase, p. 31–64. *In* F. Müller (ed.), *Chemistry and biochemistry of flavoproteins*, vol. II. CRC Press, Boca Raton, FL.
 22. **Tabor, S., and C. C. Richardson.** 1985. A bacteriophage T7 RNA polymerase/promoter system for controlled exclusive expression of specific genes. *Proc. Natl. Acad. Sci. USA* **82**:1074–1078.
 23. **Tarlera, S., and E. B. M. Denner.** 2003. *Sterolibacterium denitrificans* gen. nov., sp. nov., a novel cholesterol-oxidizing, denitrifying member of the β -*Proteobacteria*. *Int. J. Syst. Evol. Microbiol.* **53**:1085–1091.
 24. **Van der Geize, R., G. I. Hessels, and L. Dijkhuizen.** 2002. Molecular and functional characterization of the *kstD* gene of *Rhodococcus erythropolis* SQ1 encoding a second 3-ketosteroid- Δ^1 -dehydrogenase isoenzyme. *Microbiology* **148**:3285–3292.
 25. **Van der Geize, R., G. I. Hessels, R. Van Gerwen, J. W. Vrijbloed, P. Van der Meijden, and L. Dijkhuizen.** 2000. Targeted disruption of the *kstD* gene encoding a 3-ketosteroid- Δ^1 -dehydrogenase isoenzyme of *Rhodococcus erythropolis* strain SQ1. *Appl. Environ. Microbiol.* **66**:2029–2036.
 26. **Van der Geize, R., K. Yam, T. Heuser, M. H. Wilbrink, H. Hara, M. C. Anderton, E. Sim, L. Dijkhuizen, J. E. Davies, W. W. Mohn, and L. D. Eltis.** 2007. A gene cluster encoding cholesterol catabolism in a soil actinomycete provides insight into *Mycobacterium tuberculosis* survival in macrophages. *Proc. Natl. Acad. Sci. USA* **104**:1947–1952.

Chapitre 4 : Etude méta-protéo-génomique d'un écosystème microbien riche en arsenic

Cette étude réalisée en collaboration avec neuf autres laboratoires a été initiée par l'équipe de Génétique Moléculaire, Génomique et Microbiologie (CNRS/UDS-UMR7156) du Professeur Philippe Bertin à l'Institut de Botanique de l'Université de Strasbourg et l'équipe de Jean Weissenbach du Genoscope à Paris.

1. Contexte de l'étude

1.1. Les communautés microbiennes et la métagénomique

Les organismes microbiens jouent un rôle clé dans un grand nombre de processus, du maintien de l'équilibre biologique à la surface de la Terre jusqu'au combat contre les maladies. Dans leur environnement naturel, ces organismes fonctionnent rarement de manière isolée, mais plutôt dans le contexte de diverses communautés bactériennes. La compréhension de la structure et des activités dans les communautés bactériennes dépend de la capacité à extraire les informations génomiques de tous les membres de la communauté. Cependant, les micro-organismes actuellement non cultivables représentent la majorité des organismes dans la plupart des environnements sur Terre (estimé entre 80 et 99 %) [VerBerkmoes et al., 2009]. Pour cette raison et parce que des souches isolées pourraient se comporter différemment en culture que dans leur environnement naturel, un grand intérêt a été porté au développement de méthodes qui permettent d'étudier les communautés bactériennes dans leur milieu naturel et pas en culture [Handelsman, 2004; Allen et al., 2005].

La métagénomique a émergé relativement récemment comme un outil puissant pour l'analyse génomique en milieu naturel d'une population de micro-organismes. En métagénomique, l'ADN génomique est extrait directement des communautés microbiennes de l'échantillon. Il n'est donc pas nécessaire de cultiver les organismes étudiés et les interactions environnementales et dans la communauté sont maintenues [Handelsman, 2004]. En métagénomique, les communautés peuvent donc être explorées dans leur ensemble par le séquençage de leur contenu en ADN génomique (métagénome). En dépit des difficultés d'analyse, d'interprétation et de comparaison des données métagénomiques [Foerstner et al., 2006], et selon la diversité des micro-organismes présents dans l'environnement étudié, une reconstruction du génome de certains d'entre eux est envisageable. Des

questions telles que quels sont les organismes présents et quelles sont les fonctions métaboliques impliquées dans biogéochimiques peuvent maintenant être abordées au niveau moléculaire.

1.2. Le drainage minier acide de Carnoulès (France)

Les activités minières sont connues pour produire d'immenses quantités de déchets conduisant à une contamination importante en métaux lourds. Quand ils sont exposés à l'eau, ces déchets provoquent un drainage minier acide (DMA), généralement riche en métaux et métalloïdes comme l'arsenic. Ces solutions minérales sont souvent acides et hautement toxiques pour la biocénose [Johnson et al., 2005]. Récemment, des analyses de micro-organismes cultivables présents dans des drainages miniers acides ont révélé que certains micro-organismes sont capables de transformer les composés toxiques en des formes moins toxiques et d'exercer d'importantes fonctions métaboliques dans ces environnement particuliers [Amaral Zettler et al., 2003; Johnson et al., 2005; Ram et al., 2005; Hallberg et al., 2006; Tan et al., 2007; Souza-Egipsy et al., 2008]. Cependant, le fonctionnement de telles communautés reste largement inconnu car l'étude des micro-organismes cultivés n'est pas suffisante pour élucider le fonctionnement de l'ensemble de la communauté.

Le site de Carnoulès, dans le Gard, en France, fournit un exemple remarquable d'un environnement hautement contaminé par les déchets miniers, comme le zinc, le plomb et l'arsenic. L'arsenic présent dans l'arsénopyrite qu'on trouve dans les déchets est entraîné par l'eau de pluie et la nappe phréatique qui donnent naissance à un cours d'eau appelé « Reigous » et qui contient à sa source entre 100 et 300 mg/L d'arsenic principalement sous forme arsenite As(III) (Figure 1). Cette concentration, bien que restant très élevée, diminue de 95 % entre la source du « Reigous » et sa confluence avec la rivière « Amous », 1.5 km en aval. Ce processus naturel de décontamination semble résulter principalement des effets des micro-organismes présents qui contribuent à l'oxydation du Fe(II) en Fe(III) et de l'arsenite AS(III) en arsenate (V) conduisant à la co-précipitation de l'arsenate avec les ions ferriques [Casiot et al., 2003]. Cependant en dépit de plusieurs études publiées ces dernières années, seuls certains des micro-organismes ont été isolés jusque-là [Bruneel et al., 2003; Casiot et al., 2003; Duquesne et al., 2008], ce qui suggère que le fonctionnement de l'écosystème reste largement inconnu .

Figure 1 : le Reigous, ruisseau acide minier chargé en arsenic



1.3. Objectif de l'étude

Dans ce contexte, nous avons utilisé une approche multidisciplinaire combinant la géochimie, l'étude de taxonomie, la génomique et la protéomique pour obtenir une vision intégrée des micro-organismes présents et de leur rôle spécifique dans l'écosystème du site de Carnoulès.

2. Stratégie d'analyse et résultats intermédiaires

Le détail des expériences est décrit dans la publication en cours de soumission jointe au chapitre.

2.1. Préparation des échantillons

Les sédiments de la rivière « Reigous » ont été collectés sur 5 cm de profondeur à une position appelée COWG [Bruneel et al., 2003] et située 30 m en aval de la source. Pour maximiser le recouvrement en micro-organismes bactériens et pour limiter la possible contamination par le protiste eucaryote *Euglena sp* très largement présent sur le site [Casiot et al., 2004], les sédiments ont été

déposés sur un gradient de Nycodenz. La fraction enrichie en cellules bactériennes a été utilisée pour l'extraction de l'ADN et des protéines.

2.2. Reconstitution des génomes individuels

L'ADN extrait a été fragmenté puis séquencé pour produire un total de 400 Mb. 2/3 des séquences correspondantes ont ensuite pu être assemblées en 1124 contigs puis en 32 supercontigs. Finalement, en utilisant les séquences 16S rRNA, le pourcentage en GC, la couverture des différents contigs et les similarités avec les génomes déjà séquencés, les supercontigs ont pu être assemblés en 7 groupes de supercontigs qui ont été considérés comme correspondant chacun au génome d'un organisme majeur (dénommés CARN1 à CARN7).

Une approche phylogénomique a été utilisée pour évaluer les origines phylogénétiques des 7 génomes. Cette évaluation a montré que la plupart de ces génomes appartenaient à des lignées bactériennes pour l'instant non cultivées et que 2 génomes ne correspondaient à aucune classe taxonomique connue :

- CARN2 est phylogénétiquement proche de la β -proteobactérie *Thiomonas* sp. 3As et d'autres bactéries de l'ordre des Burkholdérales.
- CARN3 est phylogénétiquement proche de l'acidobactérie *Acidobacteria bacterium* Ellin345.
- CARN5 est phylogénétiquement proche de la γ -proteobactérie *Acidithiobacillus ferrooxidans* ATCC 23270.
- CARN7 est phylogénétiquement lié au genre au genre *Gallionella* des β -proteobactéries.
- CARN6 peut être relié à l'ordre *Hydrogenophilales* des β -proteobactéries.
- CARN1 et CARN4 ont une importante relation phylogénétique et peuvent représenter 2 sous-populations. Ils ne correspondent à aucune classe taxonomique connue et d'après leurs propriétés métaboliques identifiées dans l'étude, ce nouveau genre sera appelé « *Candidatus Fodinabacter aminivorans* ».

Enfin, chacun des 7 génomes a été annoté automatiquement puis manuellement pour générer une banque de protéines prédites globale.

2.3. Analyse Protéogénomique

2.3.1. Analyse par spectrométrie de masse

Nous avons fait le choix de réaliser une préparation employant le gel 1D suite à des essais préliminaires qui nous ont permis de comparer cette préparation avec une préparation sans gel. En effet, les analyses nanoLC-MS/MS des digests peptidiques issues des préparations sans gel présentaient de nombreuses substances interférentes en analyse protéomique (issues des sédiments), ce qui n'était pas le cas des analyses issues de la préparation par gel 1D. Les protéines extraites de la

fraction enrichie en cellules bactériennes ont donc été séparées sur gel 1D. Le gel a été découpé systématiquement en 35 bandes qui ont été digérées à la trypsine. Les peptides de digestion ont ensuite été analysés par nanoLC-MS/MS sur un système nanoAcquity UPLC couplé à un Q-TOF Synapt (Waters). Le gradient de modification de la phase mobile pour la séparation chromatographique utilisé ici a été optimisé pour une analyse nanoLC-MS/MS efficace d'un mélange relativement complexe de peptides tryptiques « standards » (1-50 % acétonitrile sur 35 minutes).

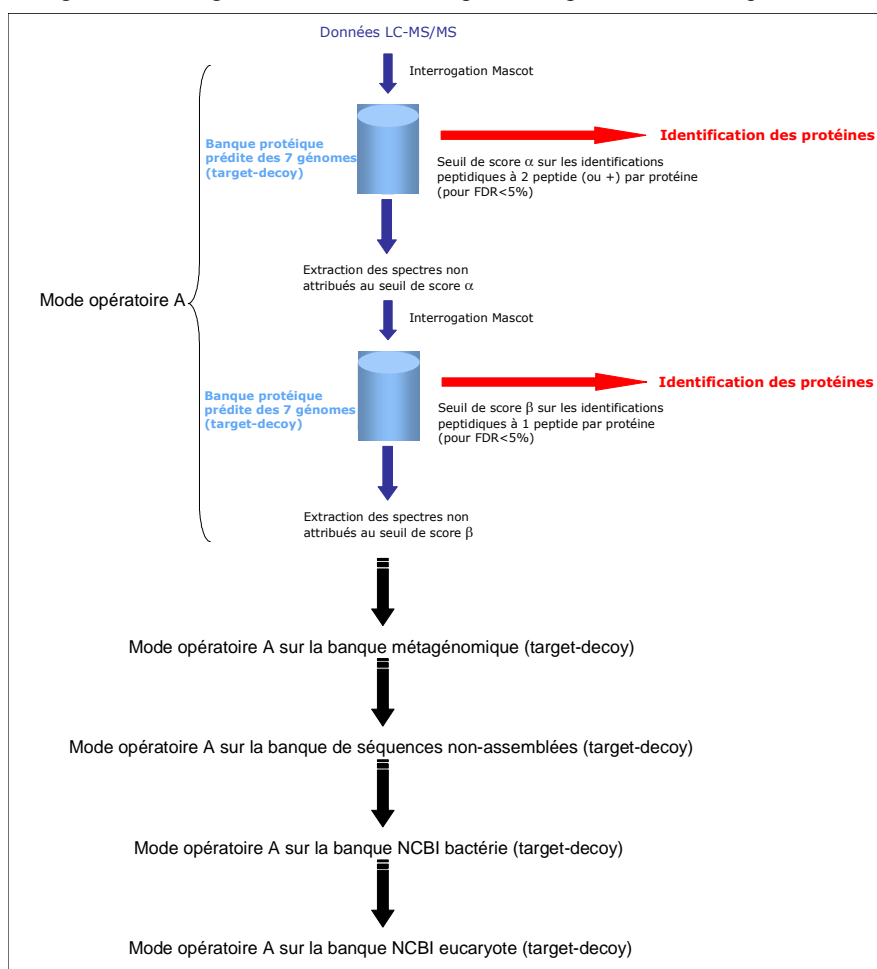
2.3.2. Stratégie d'identification des protéines

L'ensemble des données MS/MS a été soumis à une interrogation Mascot avec une tolérance de 30 ppm sur la masse des précurseurs et 0.1 Da sur la masse des ions fragments, en autorisant une coupure manquée par la trypsine et avec l'oxydation des méthionines et la carbamidométhylation des cystéines spécifiées comme modifications variables.

L'annotation des 7 génomes définis en 2.2. a permis de constituer une banque de séquences protéiques prédites des 7 organismes bactériens majeurs de l'écosystème. Toutefois, pour identifier de la manière la plus exhaustive possible l'ensemble des protéines présentes dans l'échantillon, il n'est pas suffisant de réaliser la recherche des données MS/MS dans la banque de séquences protéiques prédite à partir du génome des 7 organismes. Ainsi, les études réalisées dans le chapitre 1 et le chapitre 2 de cette partie ont montré que le processus d'annotation des génomes n'était pas exempt d'erreurs même sur le cas simple d'un génome séquencé à partir d'un organisme isolé en culture. Dans le contexte de l'étude de communautés microbiennes, en plus des difficultés habituelles de prédiction des gènes, de nouveaux problèmes se posent. Par exemple, ici, il est beaucoup plus difficile d'accéder à la séquence génomique complète des 7 organismes majeurs et le recouvrement de certains contigs est assez faible. De plus, un certain nombre de séquences génomiques obtenues n'a pas été assemblé dans les 7 organismes majeurs.

Par conséquent, nous avons établi un protocole de recherche séquentiel et d'extraction des spectres non-attribués sur le principe de celui développé dans le chapitre 2. 3.3.1. de cette partie. Ici, les données MS/MS ont été soumises à l'interrogation Mascot de manière séquentielle dans des versions target-decoy de la banque protéique prédite à partir des 7 génomes, de la banque métagénomique (construite sur le principe décrit dans le chapitre 2. 3.2. de cette partie), de la banque constituée des séquences métagénomiques non-assemblées, de la banque NCBI restreinte aux bactéries, de la banque NCBI restreinte aux eucaryotes. Le protocole est détaillé en Figure 2.

Figure 2 : Stratégie d'identification des protéines par recherche séquentielle.



2.3.3. Distribution des identifications des protéines

Les seuils de score des identifications peptidiques réalisées dans les différentes recherches ont été établis pour maintenir un FDR maximal de 5 %. Les seuils de score sur les identifications peptidiques, les résultats d'identification des protéines ainsi que les FDR associés sont indiqués dans le Tableau 1.

En tout, 749 protéines ont été identifiées avec un FDR moyen inférieur à 5 %. Comme attendu, la majorité des protéines identifiées (~70 %) sont présentes dans la banque protéique prédite à partir des 7 génomes. Globalement, une bonne corrélation a notamment pu être réalisée entre le nombre de protéines identifiées pour un organisme et la couverture des contigs le constituant. Par exemple, l'organisme pour lequel le plus grand nombre de protéines a été identifié est CARN1 qui est aussi l'organisme pour lequel les contigs présentent les meilleures couvertures. De même, seulement 2 protéines ont été identifiées chez CARN3 qui est l'organisme pour lequel les contigs ont la plus faible couverture.

On peut également noter dans cette étude que si seulement 4 protéines réellement exprimées n'ont pas été prédites lors de l'annotation des 7 génomes, 59 protéines ont été identifiées dans les

séquences non-assemblées. De même, 98 et 65 protéines ont été identifiées respectivement dans la banque NCBI restreinte aux bactéries et dans la banque NCBI restreinte aux eucaryotes (principalement des protéines du protiste eucaryote *Euglena sp* dans ce cas). Parmi les 98 protéines identifiées dans la banque protéique NCBI restreinte aux bactéries, 4 protéines d'une bactérie (*Acidithiobacillus ferrooxidans*) phylogénétiquement très proche d'un des 7 organismes majeurs de la communauté (CARN5) ont été identifiées. La comparaison des données génomiques de CARN5 avec celles des génomes déjà disponibles de *At ferrooxidans* a révélé l'existence d'un moins un « trou » important dans le génome de CARN5. Pour s'assurer que ces protéines sont bien codées dans le génome de CARN5, les séquences génomique comprenant les gènes codant pour ces protéines ont été isolées, amplifiées et séquencées de manière ciblée et ont pu finalement être intégrées au génome de CARN5.

Pour l'étude des propriétés métaboliques de chacun de 7 organismes, seules les protéines issues des 7 génomes plus les 4 protéines d'*Acidithiobacillus ferrooxidans* ont été considérées (et intégrées dans la publication des résultats). Les autres protéines issues d'autres organismes présents dans la communauté de manière plus minoritaire ou dont les gènes sont localisés dans des zones manquantes ou non assemblées des 7 génomes ont été exclues de cette analyse car elles ne peuvent pas être attribuée à un des 7 organismes majeurs à l'heure actuelle.

Tableau 1 : Résultats d'identification des protéines obtenus par le protocole de recherche séquentiel des données MS/MS

Banque target-decoy	Seuil de score sur les identifications peptidiques à 2 peptides (ou +) par protéine	Seuil de score sur les identifications peptidiques à 1 peptide par protéine	Protéines identifiées	
			Nombre	FDR
Banque protéique prédite (à partir des 7 génomes)	Score identité Mascot -12	Score identité Mascot +10	523	4 %
Banque métagénomique	Score identité Mascot-12	Score identité Mascot+18	4	Pas d'identification decoy
Banque de séquences non-assemblées	Score identité Mascot-14	Score identité Mascot +19	59	5 %
NCBI bactérie	Score identité Mascot -16	Score identité Mascot +10	98	5 %
NCBI eucaryotes	Score identité Mascot -20	Score identité Mascot +12	65	5 %

3. Résultats

Pour étudier les propriétés métaboliques possibles de chaque organisme, une exploration *in silico* du contenu en gènes de chacun des 7 génomes a été réalisée en parallèle de l'identification des

activités microbiennes majeures exercées *in situ* (issue de l'analyse protéogénomique). Ces résultats sont largement détaillés dans la publication en cours de soumission à *Science* jointe au chapitre.

Globalement, l'étude a permis de faire l'inventaire des multiples voies métaboliques et énergétiques de l'écosystème du site de Carnoulès et surtout de visualiser plusieurs activités majeures en action. L'étude a fourni un aperçu du rôle de chaque micro-organisme, notamment pour la fixation de l'azote et du carbone, pour l'oxydation du fer et de l'arsenic et pour le recyclage de la matière organique.

4. Conclusion et perspectives

L'étude réalisée a permis de mettre en évidence dans l'écosystème un équilibre entre autotrophie et hétérotrophie qui fournit à l'ensemble des partenaires de la communauté tous les nutriments essentiels à leur fonctionnement. Si deux organismes parmi la communauté (CARN2 et CARN5) apparaissent comme ayant un rôle primordial dans la détoxification naturelle du site contaminé de Carnoulès, il semblerait que l'ensemble des membres participent à leur niveau à celui-ci. Cela illustre bien la nécessité d'étudier la communauté dans son ensemble pour une compréhension fine des mécanismes impliqués qui puisse permettre à l'homme d'élaborer des processus de détoxification efficaces inspirés de ceux de la nature.

La stratégie protéogénomique utilisée dans l'étude s'est avérée efficace pour identifier les protéines exprimées par les organismes de la communauté mais aussi pour mettre en évidence des erreurs non seulement dans l'annotation des génomes mais aussi dans leur assemblage voire pour aider à « combler des trous » dans le génome de certains organismes. Dans cette étude, le perfectionnement de la construction du métagénome grâce à l'utilisation des données expérimentales de protéomique n'a été réalisé que pour 4 protéines surtout parce que l'une d'entre elles présentait une fonction extrêmement importante pour l'écosystème. Cependant, il est envisagé d'étendre ce processus aux autres protéines identifiées expérimentalement mais non codées dans la version actuellement disponible du métagénome. De même l'application de la stratégie N-TOP dans un cas comme celui-ci aurait également pu apporter sa contribution à l'annotation des 7 génomes.

Diversity of trophic interactions inside an arsenic-rich microbial ecosystem

Philippe N. Bertin¹, Audrey Heinrich-Salmeron¹, Eric Pelletier^{2,3,4}, Florence Goulhen-Chollet¹, Florence Arsène-Ploetze¹, Sébastien Gallien⁶, Alexandra Calteau^{3,4,5}, David Vallenet^{3,4,5}, Corinne Casiot⁷, Béatrice Chane-Woon-Ming^{3,4,5}, Ludovic Giloteaux⁸, Mohamed Barakat⁹, Violaine Bonnefoy¹⁰, Odile Bruneel⁷, Michael Chandler¹¹, Jessica Cleiss¹, Robert Duran⁸, Françoise Elbaz-Poulichet⁷, Nuria Fonknechten^{2,3,4}, Béatrice Lauga⁸, Damien Mornico^{3,4,5}, Philippe Ortet⁹, Christine Schaeffer⁶, Patricia Siguier¹¹, Adam Alexander Thil Smith^{3,4,5}, Alain Van Dorsselaer⁶, Jean Weissenbach^{2,3,4}, Claudine Médigue^{3,4,5}, Denis Le Paslier^{2,3,4}.

1 Génétique Moléculaire, Génomique et Microbiologie, UMR7156 CNRS & UdS, Strasbourg, France, **2** CEA, DSV, IG, Genoscope, Laboratoire de Métagénomique des Procaryotes, **3** CNRS UMR8030, Evry, France, **4** UEVE, Université d'Evry, **5** CEA, DSV, IG, Genoscope, Laboratoire de Génomique Comparative, **6** Laboratoire de Spectrométrie de Masse Bio-Organique, Institut Pluridisciplinaire Hubert Curien, UMR7178 CNRS & UdS, Strasbourg, France, **7** Laboratoire Hydrosociétés Montpellier, UMR 5569, Montpellier, France, **8** Environnement et Microbiologie, UMR5254 CNRS & UPPA, Institut Pluridisciplinaire de Recherche sur l'Environnement et les Matériaux, Pau, France, **9** Laboratoire d'Ecologie Microbienne de la Rhizosphère et d'Environnements Extrêmes, UMR6191 CNRS, CEA & Université Aix-Marseille II, Saint-Paul-lez-Durance, France, **10** Laboratoire de Chimie Bactérienne, UPR9043 CNRS et Université de la Méditerranée, Marseille, France, **11** Laboratoire de Microbiologie et Génétique Moléculaires, UMR5100 CNRS, Toulouse, France

Running title: Deciphering the Carnoulès metagenome

*To whom correspondence should be addressed. E-mail: philippe.bertin@unistra.fr

Until recently, the understanding of both the structure and the function of microbial communities has been hampered by a lack of genomic data on uncultivated microorganisms. Here, environmental genomics of an environment highly contaminated with arsenic led to the reconstruction of seven individual microbial genomes. Many of them are yet uncultivated microorganisms, including those belonging to a novel bacterial phylum. In depth analysis of the metagenomic data combined with a functional metaproteomic approach gave insight into the role of each microorganism, in particular in carbon and nitrogen fixation, in iron and arsenic oxidation, and organic matter recycling. These results provided an integrated picture of the metabolic interactions at work inside an arsenic-rich ecosystem.

Although microorganisms are by far the largest reservoir of genetic biodiversity on our planet, it is generally recognized that most of those present in the environment are not accessible by culture-dependent techniques. In addition, the existing phylogenetic inventories are not sufficient to understand how ecosystems function. The recent development of descriptive and functional genomics has given an unprecedented opportunity to gain insight into what could be considered until recently as unexplorable. Consequently, environmental genomics has extended the analysis of microbial communities far beyond the sole taxonomic studies, giving rise to an integrated picture of ecosystems [1, 2].

Microorganisms are involved in biogeochemical nutrient cycles and play therefore a crucial role in the biosphere. In the environments which harboured the first forms of life on earth, As(III), although toxic, may have been one of the main mineral substrates providing chemolithotrophic organisms with energy [3]. Nowadays, mine drainage waters, where the metalloid is usually associated with sulfur, iron and other metals, are amongst the surface fluids with the highest arsenic contents [4]. Although some forms of life are able to develop in these acid waters, the biodiversity is usually low in comparison with neutral waters [5].

The site of Carnoulès, Gard (France) provides an outstanding example of an environment highly contaminated by mine wastes. These sulfurous wastes contain As-rich pyrite and the leached waters are the source of a small stream called Reigous that contains between 100 and 300 mg/l of soluble arsenic, mainly in the form of arsenite As(III). However, although the arsenic levels remain still high, this concentration decreases by 95% between the source of the Reigous and its confluence with the river Amous, 1.5 km downstream. This natural process of remediation seems to result mainly from the effects of microorganisms contributing, leading to the co-precipitation of iron and arsenic [6, 7]. However, despite many works published these last years, only a few of these microbes have been isolated so far [8]; [6, 9]. In the present study, we used a multidisciplinary approach to get an integrated view of the microorganisms present in Carnoulès and their role in the complex metabolic processes at work.

The upper zone of white sediments was collected at the bottom of the Reigous creek 30 m downstream of the source. The physico-chemical conditions prevailing at the study site can be considered as extreme, i.e. a pH value roughly 3.5, and arsenic, iron and sulfate concentrations of 87, 625 and 3,209 mg.l⁻¹, respectively (Supplementary Table 1). To reduce a possible contamination by the eucaryotic protist *Euglena* sp. present on the study site [10], sediments were subjected to Nycodenz gradient. DNA was then sequenced using chain-terminator and pyrosequencing procedures to produce a total of more than 400 Mb. Two third of the corresponding sequences, i.e. 550,920 Sanger and 281,758 GS-FLEX reads, were successfully assembled into 1,124 contigs ranging from 287 pb to 546 kb and then organized in 32 supercontigs ranging from 1.74 kb to 1.26 Mb. Combining 16S rRNA gene sequences, GC% and mean coverage of the various contigs (Fig. S1), and similarity to already sequenced genomes, including *Acidithiobacillus ferrooxidans* [11] and *Thiomonas* sp. (Arsène-Ploetze *et al.*, submitted), made then possible the assembly of 7 major bins, herein called CARN1 to CARN7. The mean polymorphism frequency in the population was assessed using SNIPer, a ssaha-SNP based software suite [12], and ranged from 5.10⁻⁴ to 2.8.10⁻² event per kb (Fig. S2 and Supplementary Table 2). Therefore, even though the association of sequences from closely related organisms cannot be ruled out, each supercontig group was considered to correspond to the genome of one single major organism. The 7 bins were then integrated into the MicroScope platform (<http://www.genoscope.cns.fr/agc/mage/carnoulescope>) for prediction of coding sequences followed by automatic and expert annotation [13].

To evaluate the origin of the 7 bins, a phylogenetic analysis of a 16S rRNA sequences obtained from clone library and metagenome sequencing was performed (Fig. 1). Except for CARN4 whose the missing rRNA sequence precluded its 16S-based classification, clone sequences revealed a dominance of bacteria belonging to the Proteobacteria phylum (β - and γ -Proteobacteria with 38 and 23%, respectively). Remaining clones belong to Firmicutes (4%) and Acidobacteria (3%), and to a lesser extent to Spirochaetes (1%) and α -Proteobacteria (1%). This suggests that the microbial diversity at Carnoulès is probably higher than that represented by the 7 major genomes, although it seems limited to some bacterial phyla. In addition, a phylogenomic analysis was performed from metagenomic data using 27 universal marker genes (individually or combined) supposed to be sufficiently conserved to build a tree of life [14]. All sequences belonged to the bacterial domain, further supporting the low prevalence of Archaea inside the ecosystem under study [15]. Except for CARN6 where no marker was found, and in agreement with the results obtained by 16S rRNA sequence analysis (Fig. 1), the candidates CARN2, CARN3 and CARN5 were affiliated to the β -Proteobacterium *Thiomonas* genus, the Acidobacteria clade and to the γ -Proteobacterium *Acidithiobacillus ferrooxidans*, respectively (Supplementary Table 3). The comparison of CARN7 16S rRNA gene with sequences in the RDP database suggests that this organism is phylogenetically related to the *Gallionella* β -Proteobacterium. All phylogenetic results for the bins were supported, at least in part, by an analysis of gene order conservation known to be correlated with evolutive distances [16]. Remarkably, the candidates corresponding to the CARN1 and CARN4 bins presented an important phylogenetic relationship and may represent 2 subpopulations [17]. They did not correspond however to any known taxonomic class and, according to the metabolic properties identified in the present study (see below), this new genus was herein named *Candidatus Fodinabacter aminivorans*. Interestingly, similar 16S rRNA sequences are frequently identified in acid mine environment, suggesting that these uncultivated bacteria are widespread in such ecosystems (data not shown).

The size of the 7 bins ranged from 1.5 Mb to more than 4.0 Mb and their GC content ranged approximately from 52 to 65% (Supplementary Table 4). DNA sequencing led to a low coverage for some contigs, in particular those of CARN3, which suggests that the size of the corresponding genome is underestimated. In addition, the comparison of the CARN5 data with those of the already available *At. ferrooxidans* genomes revealed the existence of at least one large gap in CARN5. In both ATCC reference strains, this region contains a duplication of most tRNA encoding genes, several genes involved in the synthesis of membrane sugars and multiple transposases, which suggests a probable acquisition by lateral gene transfer. Similar results were obtained by 454 pyrosequencing, which supports the lack of any major bias resulting from the cloning procedure for genomic library construction (Fig. S3).

To investigate the possible metabolic properties of each organism, we explored *in silico* the gene content of each bin. In parallel, we performed a metaproteome analysis to identify the major microbial activities exerted *in situ*, which allowed the reliable identification of more than 500 unique proteins with a very low false positive rate (below 5%) (Supplementary Table 5a). To accommodate with the seasonal fluctuations in the oxygen concentration of the ecosystem, a wide diversity of bioenergetic electron chains may be needed. Both metagenomic and/or metaproteomic data suggest several terminal oxidases might be operative in all strains of the Carnoulès microorganisms, e.g. the cytochrome oxidase *cox* and *cta* operons and the *cyo* and *cyd* operons encoding quinol oxidases. In addition, several operons involved in anaerobic respiration were present in the bin of *Thiomonas* sp. (CARN2), in particular those involved in the metabolism of nitrate and nitrite, such as *ntr*, *nar*, and *nas*, or the use of fumarate as an electron acceptor under anaerobic conditions (Fig. 2). On the other

hand, possible inorganic electron donors may comprise sulfur compounds and arsenite, as attested by the synthesis by the *Thiomonas* sp. of proteins such as *sor*, *sox*, *tetH* and *sqr* as well as *aox* involved in the oxidation of either reduced inorganic sulfur compounds (sulfide, sulfur, sulfite, thiosulfate and tetrathionate) or As(III). Similarly, the CARN7 bin harbours a *dsrABEFHCMKLJOPN* operon. The DsrA protein showed 72% similarity with the amino-acid sequence of the sulfur-oxidising *Thiobacillus denitrificans* corresponding protein. This suggests that the *Gallionella*-like strain is also able to use sulfur compounds as electron donor in its energy metabolism, as demonstrated in *Gallionella ferruginea* [18]. In addition, the CARN5 bin of *Acidithiobacillus* sp. expressed the sole Rus protein (Supplementary Table 5b), encoding a rusticyanin involved in electron transport with iron used as an energy source. This suggests a major role of this strain in the iron oxidation observed on the study site [6,19]. Finally, the oxidation of H₂ to protons may also be a source of energy due to the presence in CARN2, CARN5 and CARN6 of several hydrogenase encoding genes. In contrast, no respiratory arsenate reductase, which allows anaerobic respiration of As(V), was identified.

To maintain bacterial metabolic activities under changing nutritional conditions, the temporary accumulation of various compounds is often required. *Thiomonas* sp. CARN2 was supposed to store poly-beta-hydroxybutyrate due to the presence in its genome of genes such as *phbC* and *phaZ* encoding a polymerase and a depolymerase, respectively. In addition, both *Acidithiobacillus* sp. (CARN5) and *Gallionella*-like (CARN7) may be able to metabolize glycogen by the means of a set of genes such as *glg* and *pgm*. Like *Thiomonas* sp. CARN2 and *Acidithiobacillus* sp. CARN5, both *Candidatus* F. aminivorans CARN1 and CARN4 contain genes coding for polyphosphate kinase and exopolyphosphatase, suggesting that they are able to accumulate and metabolize polyphosphate. In addition, genes coding for phosphonate transport and hydrolysis of carbon-to-phosphorus bond were identified in *Thiomonas* sp., *Acidithiobacillus* sp. and *Gallionella*-like. The use of organophosphorus compounds as a phosphorus source may constitute an advantage for growth in a partly oligotrophic environment. Finally, due to the structural similarity between P and As(V), arsenic metabolizing strains may preferentially transport phosphate via the specific Pst phosphate transport system rather than the Pit general transport mechanism [20,21]. No Pit gene was identified in any bin while all strains contain a Pst system (Supplementary Table 5b), further supporting this hypothesis.

Arsenic toxicity results from various biological effects, including oxidative stress and DNA damage. With this respect, all genomes carry genes protecting against oxidative stress, including those coding for catalase, superoxide dismutase and thioredoxin-thioredoxin reductase. In addition, most of them contain genes coding for bacterioferritin comigratory protein, which protects cells against toxic hydroxyl radicals. In addition, multiple genes coding for enzymes involved in DNA recombination and repair were identified. Their inactivation, e.g. *radA* and *recQ*, has been recently shown to result in an important loss of viability in the presence of arsenic [20]. On the other hand, resistance to arsenic mainly depends on membrane transport system. All Carnoulès strains were shown to possess at least one *ars* operon conferring resistance to arsenic by the means of an ArsB efflux pump (Fig. 2). Surprisingly, the 3 operons identified in *Thiomonas* sp. CARN2 were located in the vicinity of 3 *aoxAB* operons, suggesting the existence of 3 distinct arsenic islands in this strain (Fig. S5). One of those islands was flanked by multiple transposases (Supplementary Table 6), suggesting an acquisition by lateral gene transfer. In addition, the presence of an *arsM* gene coding for an arsenite S-adenosylmethyltransferase was identified in the bins of CARN5 and CARN6, in agreement with the presence of methylated forms of arsenic, i.e. mono- and di-methylarsine, at Carnoulès (C. Stels, unpublished).

The rapid and efficient adaptation of natural isolates to stressful environmental conditions and their ability to efficiently colonize an ecosystem depend on multiple mechanisms. In addition to

various alternative sigma factors, all bins encode nucleoid-associated proteins such as HU, IHF, FIS or H-NS [22], which play a major role in both the structure and the function of chromosomal DNA. Multiple regulators, including those belonging to the two-component systems, were also identified in all strains, suggesting that these bacteria may respond to a wide panel of stimuli (Supplementary Tables 7a and 7b). For example, NarL and NarP suggest that most Carnoulès bacteria may respond to nitrate availability in anaerobic respiration. Similarly, the *fur* and *arsR* genes coding for the ferric uptake regulatory protein and the *ars* operon repressor, respectively, may illustrate the fitness of the Carnoulès community to the abundance of iron and arsenic in the AMD under study. Finally, LuxR-like regulator genes were found in most bins while the autoinducer synthesis *soll* gene was found in *Thiomonas* sp. (CARN2), *Acidithiobacillus* sp. (CARN5) and *Gallionella*-like (CARN7). These observations suggest that microorganisms of Carnoulès interact with each other by a quorum-sensing mechanism, which may be of importance in the maintenance of the community integrity. With this respect, motility and aggregation have been often demonstrated to be a major determinant in biofilm formation. Except for *Acidithiobacillus* sp. CARN5, most bins contain multiple genes required for flagellum biosynthesis and motility. In contrast, this strain contains several genes involved in capsular biosynthesis, e.g. *kdtA*, *crd* and *gmd* coding. *Thiomonas* sp., *Acidithiobacillus* sp. and *Gallionella*-like were also shown to contain Type 4 pilus operons, which play an important role in twitching motility and adhesion.

Finally, as the integrity of the Carnoulès microbial community may also rely on the existence of metabolic exchanges, we explored the bin capacities regarding carbon and nitrogen transport and metabolism (Fig. 2). As expected from an oligotrophic environment, we identified several enzymes responsible for autotrophic metabolism. Carbon fixation may depend on genes coding for ribulose 1,5-biphosphate carboxylase/oxygenase (involved in Calvin cycle), carboxysome structural proteins and carbon monoxide dehydrogenase (involved in acetyl-coenzyme A synthesis) that were identified in *Thiomonas* sp (CARN2), *Acidithiobacillus* sp. (CARN5) and/or *Gallionella*-like (CARN7). In addition, CARN5 and CARN7 were also able to fix nitrogen by means of *nif*, which encodes a nitrogenase. Nevertheless, mixotrophic or heterotrophic metabolism may also exist in Carnoulès, suggesting that microorganisms recycle some organic compounds released by others. Indeed, the *Thiomonas* sp. (CARN2) bin encodes the phosphoenolpyruvate (PEP)-dependent sugar phosphotransferase system (PTS), which suggests that this strain is able to catabolize carbohydrates. Like CARN2, both CARN5 and CARN7 encode lactate, formate or succinate dehydrogenases, while *fucP* and *exuT* genes coding for L-fucose and hexuronate transporters, respectively are present in CARN3. In addition, the CARN6 bin carries the cellulase encoding gene *bczS* and the alpha-amylase *amyM* gene, suggesting that this strain is able to metabolize more complex carbohydrates. The presence of a complete urease-encoding *ure* operon in *Thiomonas* sp. CARN2 strain further supports a metabolic interaction between the Carnoulès microorganisms. Such an interplay may also include eucaryotic organisms such as *Euglena* sp. present on the study site [10]. In addition, several enzymes required for amino acid transport and metabolism were identified in the bins lacking the carbon and nitrogen fixation genetic determinants. In particular, the CARN1 and CARN4 bins of *Candidatus* F. aminivorans, which were shown to expressed most of the identified proteins (Supplementary Table 5b), encode multiple peptidases and also contain *liv* and *opp* genes involved in branched amino-acid and oligopeptide transport, respectively. In addition, genes involved in lysine fermentation were also identified in both bins. This pathway is principally found in obligate anaerobic bacteria, which convert lysine into butyrate, acetate and ammonia [23]. These observations were further supported by a Multiple Factor Analysis (MFA) on a reaction frequency matrix. Factorial planes segregated CARN2 from the other bins (Fig. 3a) and separated CARN1-CARN4 from CARN5-CARN7. From the variable

classification results (Supplementary Table 8), several clusters of reactions were then associated to bin groups. These include, for example, Cluster 4 linked to CARN2 (energy metabolism, fatty acid by-products degradation, inorganic nutrient metabolism and arsenic detoxification (Supplementary Table 8); Cluster 1 groups reactions common to CARN2 and CARN5 (Calvin-Beson-Bassham cycle and urea degradation pathways); Cluster 5 includes reactions related to lysine fermentation and other amino acid degradation pathways in CARN1 and CARN4. In addition, (Fig. 3b), CARN3 was opposed to CARN6, revealing CARN6-specific reactions (Clusters 5 and 6) that include those involved in cellulose metabolism and purine degradation (Supplementary Table 8). Interestingly, no such correlation was observed between our data and those from the acidic AMD biofilm [24], which suggests that these two ecosystems differ markedly.

Taken together, these genomic approaches allowed us to inventory multiple metabolic and energy pathways in the ecosystem under study and, more importantly, to visualize several major activities at work (Figs. 2 and 3). These processes include not only the fixation of inorganic carbon and nitrogen but also the use or the recycling of other mineral and organic resources such as iron, sulfur, phosphorus, urea and amino acids. Such an equilibrium between auto- and heterotrophy may provide all partners with essential nutrients. With this respect, the role of *Thiomonas* sp. in arsenic oxidation associated with other metabolic activities, including iron oxidation by *At. ferrooxidans*, seems to be of prime importance in the partial but efficient natural remediation of the Carnoulès contaminated ecosystem [6,7]. In addition, other microorganisms, in particular *Candidatus* F. aminivorans that is the first representative of a novel bacterial phylum, may play an important role in recycling organic compounds released by others. Our observations therefore revealed the diversity of trophic interactions that may exist inside a microbial community, allowing microorganisms, including yet uncultivated bacteria, to efficiently colonize an arsenic-rich environment.

References and Notes

1. Allen EE, Banfield JF (2005) Community genomics in microbial ecology and evolution. *Nat Rev Microbiol* 3: 489-498.
2. Bertin PN, Medigue C, Normand P (2008) Advances in environmental genomics: towards an integrated view of microorganisms and ecosystems. *Microbiology* 154: 347-359.
3. Lebrun E, Brugna M, Baymann F, Muller D, Lievreumont D, et al. (2003) Arsenite oxidase, an ancient bioenergetic enzyme. *Mol Biol Evol* 20: 686-693.
4. Vaughan DJ (2006) Arsenic. *Elements* 2: 71-75.
5. Johnson DB, Hallberg KB (2003) The microbiology of acidic mine waters. *Res Microbiol* 154: 466-473.
6. Duquesne K, Lebrun S, Casiot C, Bruneel O, Personne JC, et al. (2003) Immobilization of arsenite and ferric iron by *Acidithiobacillus ferrooxidans* and its relevance to acid mine drainage. *Appl Environ Microbiol* 69: 6165-6173.
7. Casiot C, Morin G, Juillot F, Bruneel O, Personne JC, et al. (2003) Bacterial immobilization and oxidation of arsenic in acid mine drainage (Carnoules creek, France). *Water Res* 37: 2929-2936.
8. Bruneel O, Personne JC, Casiot C, Leblanc M, Elbaz-Poulichet F, et al. (2003) Mediation of arsenic oxidation by *Thiomonas* sp. in acid-mine drainage (Carnoules, France). *J Appl Microbiol* 95: 492-499.
9. Duquesne K, Lieutaud A, Ratouchniak J, Muller D, Lett MC, et al. (2008) Arsenite oxidation by a chemoautotrophic moderately acidophilic *Thiomonas* sp.: from the strain isolation to the gene study. *Environ Microbiol* 10: 228-237.
10. Casiot C, Bruneel O, Personne JC, Leblanc M, Elbaz-Poulichet F (2004) Arsenic oxidation and bioaccumulation by the acidophilic protozoan, *Euglena mutabilis*, in acid mine drainage (Carnoules, France). *Sci Total Environ* 320: 259-267.

11. Valdes J, Pedroso I, Quatrini R, Dodson RJ, Tettelin H, et al. (2008) *Acidithiobacillus ferrooxidans* metabolism: from genome sequence to industrial applications. BMC Genomics 9: 597.
12. Ning Z, Cox AJ, Mullikin JC (2001) SSAHA: a fast search method for large DNA databases. Genome Res 11: 1725-1729.
13. Vallenet D, Labarre L, Rouy Z, Barbe V, Bocs S, et al. (2006) MaGe: a microbial genome annotation system supported by synteny results. Nucleic Acids Res 34: 53-65.
14. Ciccarelli FD, Doerks T, von Mering C, Creevey CJ, Snel B, et al. (2006) Toward automatic reconstruction of a highly resolved tree of life. Science 311: 1283-1287.
15. Bruneel O, Pascault N, Egal M, Bancon-Montigny C, Goni-Urriza MS, et al. (2008) Archaeal diversity in a Fe-As rich acid mine drainage at Carnoulès (France). Extremophiles 12: 563-571.
16. Huynen MA, Bork P (1998) Measuring genome evolution. Proc Natl Acad Sci U S A 95: 5849-5856.
17. Wilmes P, Simmons SL, Denev VJ, Banfield JF (2009) The dynamic genetic repertoire of microbial communities. FEMS Microbiol Rev 33: 109-132.
18. Lutterszekalla S (1990) Lithoautotrophic Growth of the Iron Bacterium *Gallionella ferruginea* with Thiosulfate or Sulfide as Energy-Source. Archives of Microbiology 154: 417-421.
19. Morin G, Juillot F, Casiot C, Bruneel O, Personne JC, et al. (2003) Bacterial formation of tooeelite and mixed arsenic(III) or arsenic(V)-iron(III) gels in the Carnoules acid mine drainage, France. A XANES, XRD, and SEM study. Environ Sci Technol 37: 1705-1712.
20. Muller D, Medigue C, Koechler S, Barbe V, Barakat M, et al. (2007) A tale of two oxidation states: Bacterial colonization of arsenic-rich environments. Plos Genetics 3:
21. Weiss S, Carapito C, Cleiss J, Koechler S, Turlin E, et al. (2009) Enhanced structural and functional genome elucidation of the arsenite-oxidizing strain *Herminiimonas arsenicoxydans* by proteomics data. Biochimie 91: 192-203.
22. Tendeng C, Bertin PN (2003) H-NS in Gram-negative bacteria: a family of multifaceted proteins. Trends in Microbiology 11: 511-518.
23. Kreimeyer A, Perret A, Lechaplais C, Vallenet D, Medigue C, et al. (2007) Identification of the last unknown genes in the fermentation pathway of lysine. J Biol Chem 282: 7191-7197.
24. Tyson GW, Lo I, Baker BJ, Allen EE, Hugenholtz P, et al. (2005) Genome-directed isolation of the key nitrogen fixer *Leptospirillum ferrodiazotrophum* sp. nov. from an acidophilic microbial community. Appl Environ Microbiol 71: 6319-6324.

Acknowledgments

A.H.S. was supported by a grant from the French Ministry of Education and Research. Financial support came from the Université de Strasbourg (UdS), the Consortium National de Recherche en Génomique (CNRG), the Centre National de la Recherche Scientifique (CNRS) and the Agence Nationale de la Recherche (ANR) in the frame of the RARE and MicroScope projects. ISfinder is supported by the CNRS and has received some support from the ARC. This work was done in the frame of the « Groupement de Recherche - Métabolisme de l'Arsenic chez les Micro-organismes (GDR2909-CNRS) » (<http://gdr2909.u-strasbg.fr>).

Figure legends

Figure 1. Phylogenetic tree representing the taxonomic affiliation of the Carnoulès community microorganisms. The 16S rRNA gene sequences were obtained from DNA sediments after PCR amplification (clones CG-X) or metagenomic sequencing (CARN bins, except for CARN4, see text). A total of 759 positions was obtained. The scale bar corresponds to 0.05 substitutions per site. Percentages of 1000 bootstrap resamplings that supported the branching orders in each analysis are shown above or near the relevant nodes. Bootstrap values are shown for branches with more than 50% bootstrap support. These results were obtained by a Neighbor-Joining method and a Maximum-likelihood method gave rise to a similar phylogenetic reconstruction (data not shown).

Figure 2. Model of the Carnoulès bacterial community highlighting the major functions identified by metagenome sequencing or metaproteome characterization. These activities include carbon and nitrogen fixation, energy metabolism, flagellum and capsule, biosynthesis, amino acid transport and degradation, detoxification and stress response, arsenic and iron metabolism. The possible interactions between these microorganisms or with other chemical or biological components present on the study site are indicated by arrows. CARN bins are numbered from 1 to 7.

Figure 3. Multiple Factorial Analysis of the 7 Carnoulès bins, performed from a two dimensional-matrix combining bins and enzymatic reactions, respectively. To highlight possible metabolic distinctions between bins, 3 axes (F1 to F3) with the highest dispersion in the resulting dot cloud were selected; they represent more than half the total dispersion. Color lines represent the variable vectors corresponding to the enzymatic reaction frequencies, the external disk indicating those also identified or not in the metaproteomic data. The reaction variables were then hierarchically clustered, which led to 7 and 9 classes for the first and second factorial planes, respectively. The corresponding functions are listed in Suppl. Table 8ab.

Figure 1

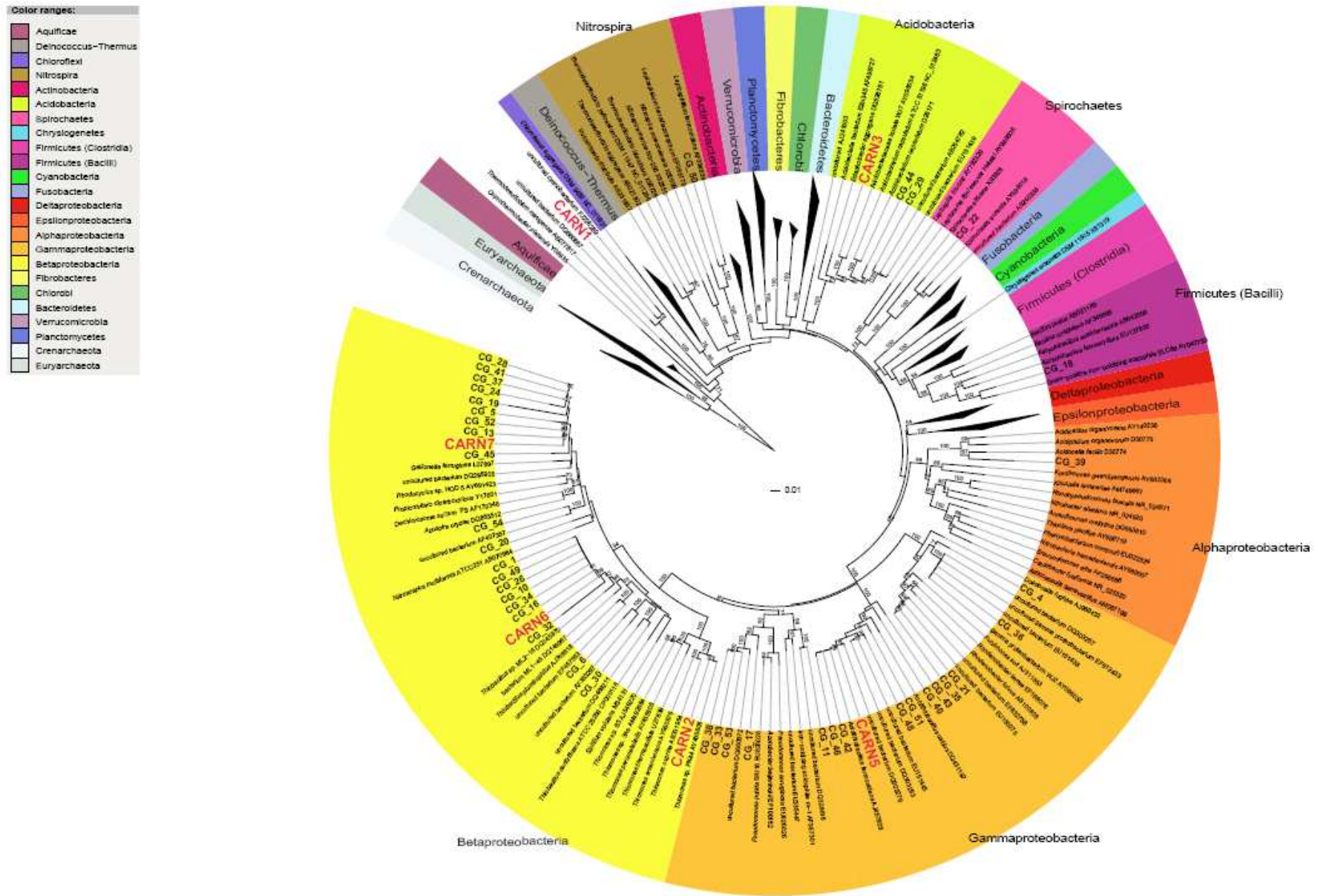


Figure 2

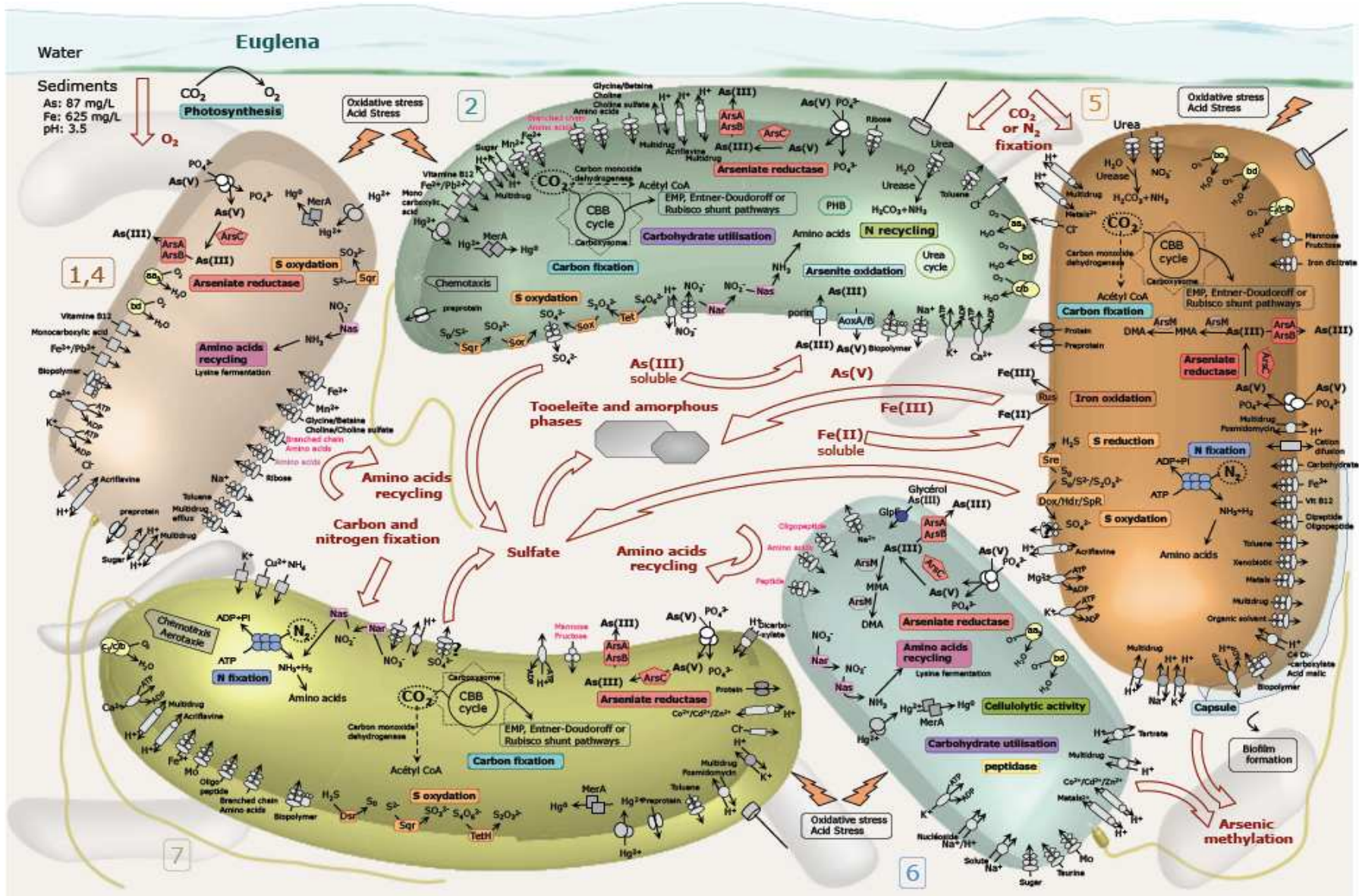
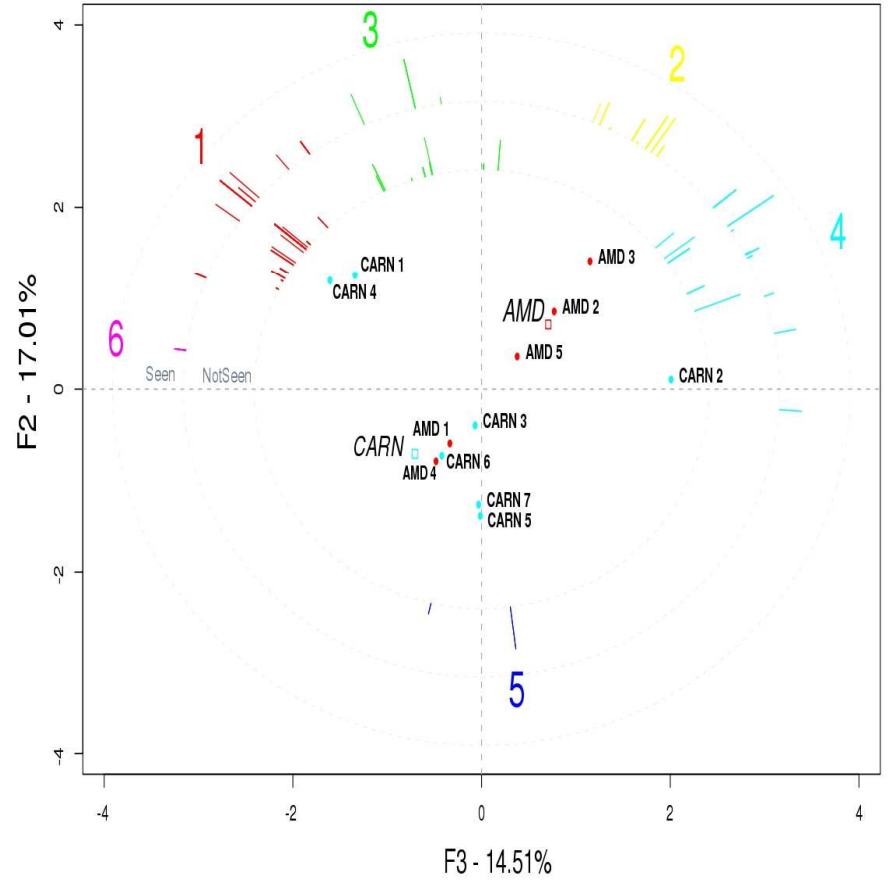
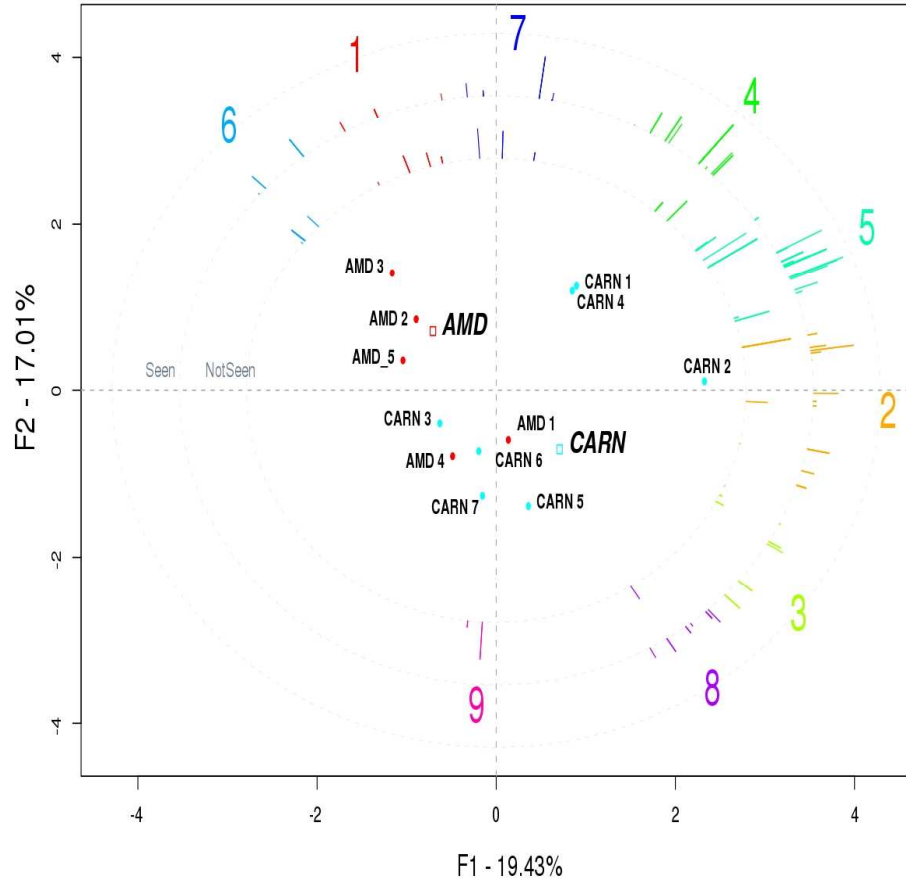


Figure 3



Materials and Methods

Sampling and chemical analysis. Samples were collected in the Reigous creek in May 2007 at the station called COWG (Bruneel *et al.*, 2003) and located 30 m downstream of the spring. Five cm deep white sediments covering the bottom of the creek were collected using a sterile glass, and the running water (i.e. a thin column, less than 10 cm) covering these sediments was filtered (300 mL) through sterile 0.22 μm Nuclepore filters. These filters were transferred into a collection tube, frozen in liquid nitrogen, and stored at -80°C until further analysis. This sampling was done in three replicates.

The main physico-chemical parameters (pH, T° , dissolved oxygen...) were determined in the field at the sampling station and arsenic speciation, Fe(II) and sulfate analyses were performed as previously described (Goulhen-Chollet *et al.*, submitted).

DNA isolation and sequencing. To recover microbial cells, ten grams of sediments were washed in 10 mL of solution 1 (in g.L^{-1} : $\text{Na}_2\text{SO}_4 \cdot 10\text{H}_2\text{O}$, 0.15 ; $(\text{NH}_4)_2\text{SO}_4$, 0.45 ; KCl, 0.05 ; $\text{MgSO}_4 \cdot 7\text{H}_2\text{O}$, 0.5 ; KH_2PO_4 , 0.05 ; $\text{Ca}(\text{NO}_3)_2 \cdot 4\text{H}_2\text{O}$, 0.014) and agitated one night at 4°C . After 10 minutes decantation, 7.5 mL of supernatant were added without mixing on 17.5 mL of Nycodenz solution (Axis-Shield, Dundee, Scotland), and then centrifuged 30 minutes at $10\,000 \times g$. The cellular fraction (upper fraction) was removed and washed by adding 2 volumes of solution 1 and centrifuge 15 min at $10\,000 \times g$ and 4°C . DNA was extracted from sediments either directly using the UltraClean Soil DNA Isolation Kit according to the recommendation of the manufacturer (MoBio Laboratories Inc., Carlsbad, CA, USA), or after separation of microbial cells by Nycodenz gradient, using the Wizard genomic DNA extraction kit (Promega, U.S.A.). All genomic DNA samples were stored at -20°C until further processing.

DNA extracted from bacterial cells was then fragmented by nebulisation and fragments ranging from 3 and 5 kb were used to construct a genomic library. DNA inserts were sequenced from both ends as previously described (Muller *et al.*, 2007). In parallel, 281,758 DNA reads were obtained by GS-FLEX pyrosequencing using standard procedures. Both procedures produced a total of 430.3 Mb.

Bioinformatics. Binning was performed as follows. First, seven scaffolds containing 16S rRNA gene sequences were characterized. The characterization of conserved gene families (Cicarelli, 2006) was also used for all the bins. Contigs were assigned to a bin if at least 2 contigs from a scaffold could be assigned to the same bin. Isolated contigs were not assigned to any contig as well as scaffolds with only one binned contig. Second, bins 1 and 4 have the same 16S rRNA gene sequence, similar GC% content but could be distinguished by their genomic coverage (6-13 versus 22-35) and bin 1 was used to complete bin 4 by contig sequence similarity. Bin 2 was obtained by similarity with *Thiomonas* sp., GC% and coverage. Bins 3 and 5 have a similar GC% and coverage were distinguished because of the similarity of bins 5 with *Acidithiobacillus ferrooxidans*. Finally bins 6 and 7 were separated by the presence of the 16S rRNA gene, conserved gene families and GC%.

Coding sequences were predicted using the AMIGene (Annotation of Microbial Genomes) software (Bocs *et al.*, 2003) and then submitted to automatic functional annotation using a set of tools (Vallenet *et al.*, 2006). Putative orthology relationships between two genomes were defined by gene pairs satisfying either the Bidirectional Best Hit criterion or an alignment threshold (at least 40% sequence identity over at least 80% of the length of the smallest protein). These relationships were subsequently used to search for conserved gene clusters (synteny groups) among several bacterial genomes using an algorithm based on an exact graph-theoretical approach (Boyer *et al.*, 2005). This method allowed multiple correspondences between genes, detection of paralogy relationships, gene fusions, and chromosomal rearrangements (inversion, insertion/deletion). The 'gap' parameter, representing the maximum number of consecutive genes that are not involved in a synteny group, was set to five. Manual validation of automatic annotations was performed using the MaGe web interface (Vallenet *et al.*, 2006) in the Arsenoscope relational database

(<http://www.genoscope.cns.fr/agc/mage/arsenoscope>).

Clone library and phylogenetic analyses. The bacterial diversity was analyzed by cloning PCR amplified 16S rRNA genes. Bacterial 16S rRNA genes were amplified with 8F and 1489R primers (Weisburg *et al.*, 1991). PCR products were cloned in *Escherichia coli* TOP10 using the pCR2.1 Topo TA cloning kit (Invitrogen, Inc.) following manufacturer's instructions. Cloned 16S rRNA gene fragments were then amplified using the primers M13F (5'- GTAAAACGACGGCCAG -3') and M13R (5'- CAGGAAACAGCTATGAC-3'), located on the vector. Clones (named CG) were then digested with the enzymes *HaeIII* and *HinfI* (New England Biolabs). Ninety-six clones from the library were analyzed on 3% agarose gel electrophoresis and grouped according to their RFLP patterns (*HaeIII* and *HinfI* digestion). Only one representative of each group was sequenced using the Big Dye[®] Terminator v3.1 cycle sequencing kit (Applied Biosystems). Sequences were compared with the RDP database (<http://rdp.cme.msu.edu>) by BLAST online searches (Altschul *et al.*, 1998). Multiple sequence alignment of clones and 16S rRNA sequences from CARN bins, except for CARN4, along with closely-related sequences of known phyla chosen on BLAST similarities was performed by using CLUSTALX (Thompson *et al.*, 1997). Phylogenies were constructed with the Molecular Evolutionary Genetics Analysis v4.0 program (Tamura *et al.*, 2007) using Nucleotide: Maximum Composite Likelihood model and Neighbour-joining algorithm. Trees were then annotated with Itol (Letunic and Bork, 2007). Significance of branching order was determined by bootstrap analysis with 1000 resampled data sets. PAST (PAleontological Statistics v1.60) software from <http://folk.uio.no/ohammer/past/> website was used to perform rarefaction analysis for the clone library with clone phenotype similarity defined at 97% 16S rRNA sequence similarity. Coverage value was calculated to determine how efficient our clone library described the complexity of a theoretical community such as original bacterial community. The coverage (Good, 1953) value is given as $C = 1/(n1/N)$ where $n1$ is the number of clones which occurred only once in the library. The sequences of clones CG determined in this study have been submitted to the EMBL database and assigned Accession Nos. FN391809 to FN391849.

Phylogenomic approach. Molecular phylogenies were inferred using 27 marker genes (Supplementary Table 2) chosen from the reference gene set proposed to reconstruct the tree of life (Cicarelli *et al.*, 2006). For each marker, the corresponding family of homologous genes from the HOGENOM database (Perrière *et al.*, 2000; release 4) was identified. Eukaryotic sequences were removed from each family. Moreover, we completed these families with sequences from 64 new prokaryotic genomes of potential interest in the framework of this study (Supplementary Table 3). Then, each family dataset was aligned with the program MUSCLE (Edgar, 2004) and filtered using the program GBLOCKS (Castresana, 2000; Talavera and Castresana, 2007) in order to select unambiguously aligned positions. Maximum-likelihood phylogenies were reconstructed with PhyML (v2.4.4, Guindon and Gascuel, 2003) using the Jones–Taylor–Thornton (JTT) model of amino-acid substitution (Jones *et al.* 1992) and performing 1000 bootstrap replicates. Heterogeneities between sites were estimated under a gamma law based model of substitution, with estimation of the alpha parameter by PHYML.

A super-alignment was also performed by concatenating the individual alignments of the 9 marker genes that were most commonly found in the bins (Supplementary Table 4). This approach allows increasing the phylogenetic signal by summing the information contained in each individual dataset. Missing data corresponding to taxonomic sampling differences were coded by gaps. Then, a maximum-likelihood phylogeny was reconstructed using the same options used for individual phylogenies.

Protein extraction and mass spectrometry identification. Cells recovered using Nycodenz gradient were resuspended in a buffer containing SDS concentration (1%, w/v) and β -mercaptoethanol (2.5%)

and boiled 5 min before loading. Proteins were separated by SDS-PAGE using 12 % gradient slab gels (PROTEAN II, Bio-Rad laboratories). Electrophoresis was carried out at 50 mA per gel. Proteins were stained with coomassie brilliant blue R-250, systematically excised from the gel and stored at -20 °C until analysis. *In gel* digestion of gel slices was performed as previously described (Weiss et al., 2008). The resulting peptide extracts were analyzed by nanoLC-MS/MS on a nanoACQUITY Ultra-Performance-LC (UPLC, Waters, Milford, MA) coupled to SYNAPT hybrid quadrupole orthogonal acceleration time-of-flight tandem mass spectrometer (Waters, Milford, MA). Peptide mixtures were loaded on a Symmetry C18 (180 µm inner diameter × 20 mm, particle size 5 µm; Waters) trap column using 0.1% formic acid at 5 µL min⁻¹. After washing, the peptides were eluted with a gradient 1-50% acetonitrile in 0.1% formic acid delivered over 35 min at a flow rate of 400 nL min⁻¹ through the BEH130 C18 (75 µm inner diameter × 200 mm, particle size 1.7 µm; Waters) analytical column. The general mass spectrometric parameters were as follows: the capillary voltage was set at 3,500V and the cone voltage at 35V. For tandem MS experiments, the system was operated with automatic switching between MS and MS/MS modes. The 3 most abundant peptides, preferably doubly and triply charged ions, were selected on each MS spectrum for further isolation and CID fragmentation with 2 energies set using collision energy profile. The complete system was fully controlled by MassLynx 4.1 (SCN 566, Waters, Milford, MA). Raw data collected during nanoLC-MS/MS analyses were processed and converted with ProteinLynx Browser 2.3 (23, Waters, Milford, MA) into .pkl peak list format.

The MS/MS data were analyzed using the MASCOT 2.2.0. algorithm (Matrix Science, London, UK) to search against a target-decoy protein database composed of predicted protein sequences from the 7 bin-genomes of the Carnoulès microbial community concatenated with reversed copies of all sequences and with common contaminants such as keratins and trypsin. Spectra were searched with a mass tolerance of 30 ppm for MS and 0.1 Da for MS/MS data, allowing a maximum of one missed cleavage with trypsin and with carbamidomethylation of cysteines and oxidation of methionines specified as variable modifications. Protein identifications were validated when at least two peptides with high quality MS/MS spectra (less than 12 points below Mascot's threshold score of identity at 95% confidence level) were detected. In the case of one-peptide hits, the score of the unique peptide had to be greater (minimal "difference score" of 10) than the 95% significance Mascot threshold. After removal of keratins and trypsin, these thresholds allowed identifying 516 unique proteins with an estimated false discovery rate less than 4.4 % searching the target-decoy database. An additional database search was performed against a target-decoy protein database composed of sequences downloaded from the NCBI database restricted to the closest organisms of the 7 bin-genomes: *Acidobacteria bacterium* Ellin345, *Acidithiobacillus ferrooxidans* ATCC 23270, *Thiomonas* and *Gallionella*, concatenated with reversed copies of all sequences. Protein identifications were validated when at least two peptides with high quality MS/MS spectra (Mascot ion score greater than 35) were detected. In the case of one-peptide hits, the score of the unique peptide had to be greater than 100. These thresholds allowed identifying 4 additional unique proteins with no protein identified in the reversed sequences.

Statistical analysis. The metabolic network of each bin was predicted by the "Pathway Tools" software (Karp *et al.*, 2002) using MetaCyc (Caspi *et al.*, 2008) as a reference pathway database (version 12.0). This software applies selection rules to infer possible metabolic pathways and builds a special database for each bin, called a PGDB (Pathway/Genome Database). From those data, a two dimensional-matrix was built, wherein each line represents a bin (CARN1 to CARN7), each column corresponding to a specific reaction; each value counts how often genes linked to the given enzymatic reaction were seen within the given bin. This frequency matrix is the starting-point of an exploratory

factorial analysis, in order to highlight possible metabolic distinctions between bins. All bins were taken into account in this statistical analysis, but only reactions with non-constant counts could be analysed. The particular method implemented here was a Multiple Factor Analysis (MFA), the specificity of which is to group variables (in this case, normalized quantitative variables) and assign to each group the same weight in the global analysis. In our case, reactions were grouped into the MetaCyc pathways they belonged to (a same reaction belonging to several pathways is thus repeated as many times, one for each pathway/group). This approach favours the discovery of factors which are common to several groups.

In total, 715 distinct reactions were included in the analysis, for a total of 359 pathways. After examination of the amount of inertia captured by the method's resulting factors, 3 were kept for further analysis. For graphical representations, the variable (reaction) plot and the individual (bins) plot were combined, restricting plotted variables to those with a quality of representation greater than 0.75, in order to conserve interpretability.

Solely as an aid to interpretation and to listing readability, for each factorial plane, the reaction variables were hierarchically clustered according to the *local* angles between their vectors. The clustering method used an Euclidean distance and ward's criterion; the number of classes was chosen after manual examination of the cluster tree (Figure S1), and led to 7 and 9 classes for the first and second factorial planes, respectively. As a further aid to interpretation, the variables were plotted in two separate disks, according to the observation (or lack of) of the proteins corresponding to the reactions in the metaproteomic data.

Supplementary figure legends

Figure S1. Distribution plot showing the repartition of the seven major contigs based on their respective GC% and mean coverage. Colors indicate the different bin assignment of supercontigs (CARN1, white, CARN2, red; CARN3, purple; CARN4, green, CARN5, blue; CARN6, yellow and CARN7, light red).

Figure S2. Amount of polymorphism in the sequenced population regarding the CARN bin reference sequences. **A:** SNPs+indel events frequencies relative to each bin sequence length; **B:** distribution of the polymorphism along the respective sequences of CARN1 and CARN4 bins.

Figure S3. Comparison between CARN5 genomic data and *Acidithiobacillus ferrooxidans* ATCC 23270. Both genomes were compared using MUMmer 3.0 (Kurtz *et al.*, 2004), using the nucleic mode, with *Acidithiobacillus ferrooxidans* as a reference, and individual contigs of the CARN5 bin as a query. X-axis represents the reference genome, Y-axis the % of similarity between the two sequences, the bottom red line representing the projection of the different segments mapped onto the reference genome. Similar results were obtained with the *Acidithiobacillus ferrooxidans* reference strain ATCC 53993.

Figure S4. Comparison between the experimental protein pattern obtained by MS/MS identification of the proteins expressed *in situ* (CARN1, green, CARN2, yellow; CARN3, black; CARN4, pink, CARN5, light blue; CARN6, brown and CARN7, blue) and the theoretical distribution predicted from metagenomic data in grey.

Figure S5. Genetic environment of the 3 arsenic islands identified in the genome of the *Thiomonas* sp. CARN2 bin. The various CDS are indicated in the 6 phases according to their position along the chromosome on the MaGe interface (Vallenet *et al.*, 2006). A red color indicates that an expert annotation has been performed. A vertical black box indicates the limit of a contig (<http://www.genoscope.cns.fr/agc/mage/carnoulescope>).

Figure S1

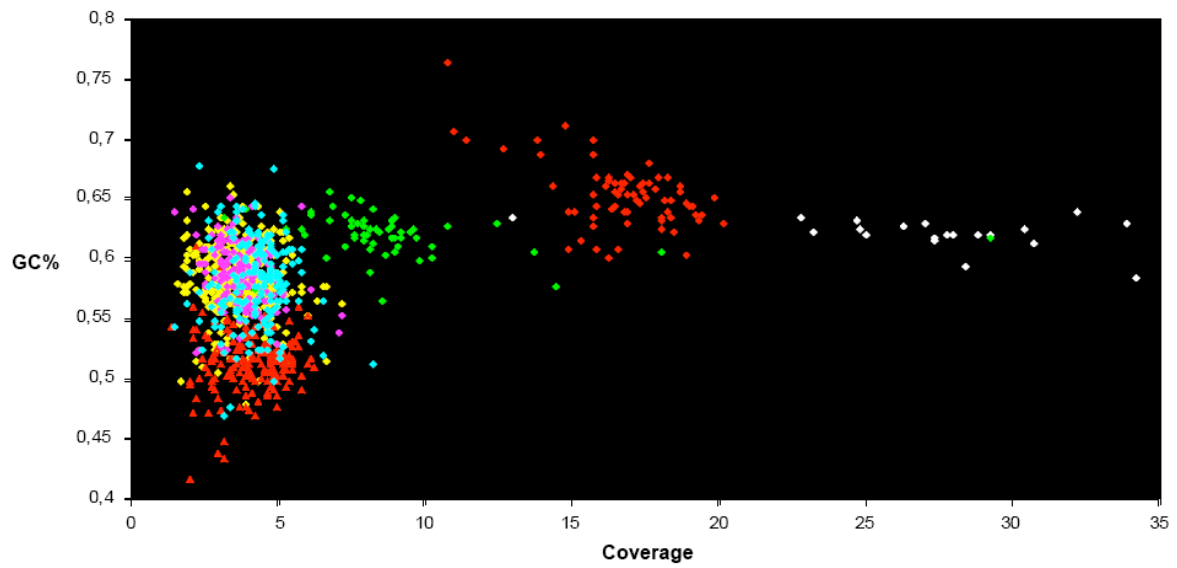


Figure S2

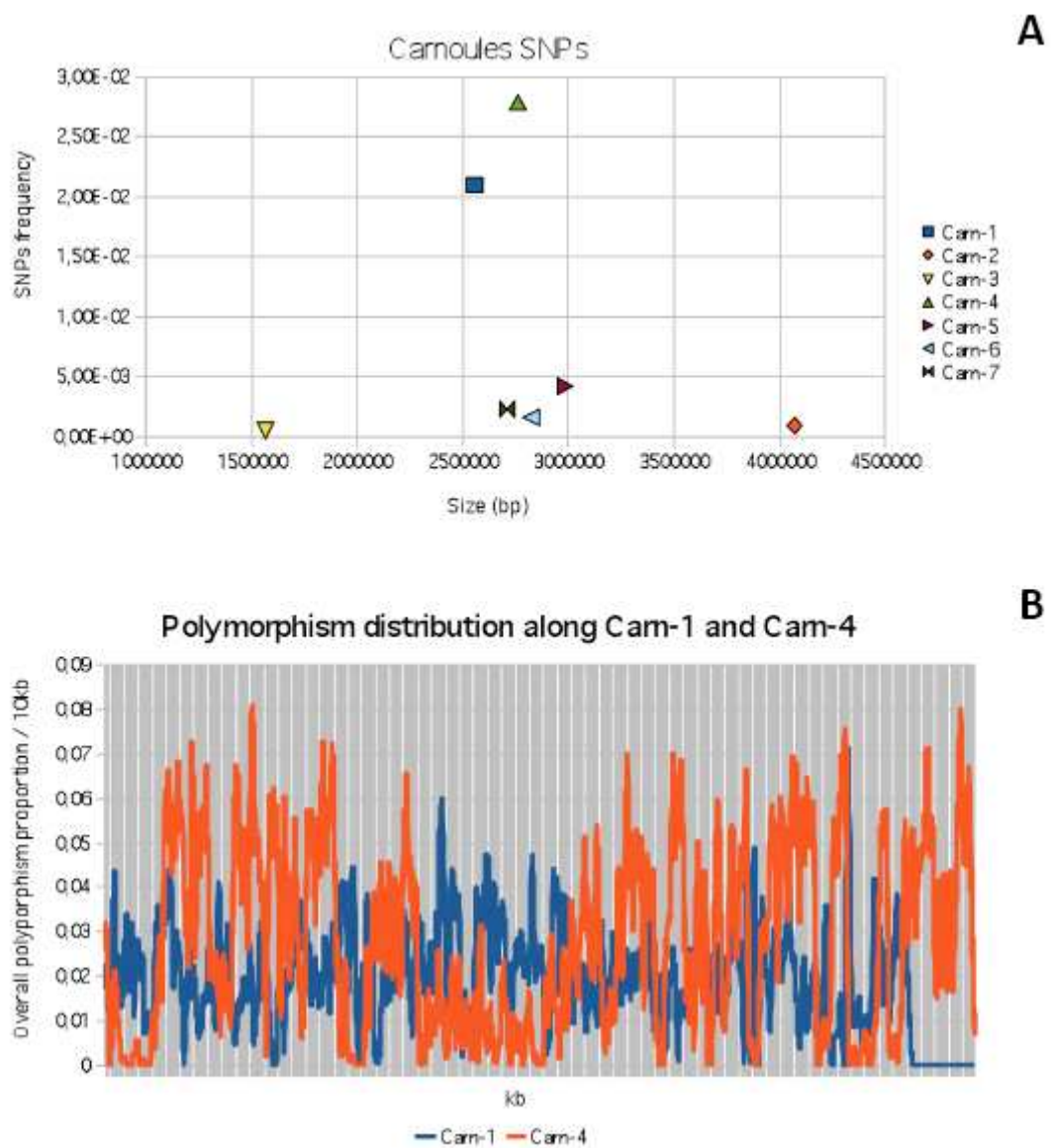


Figure S3

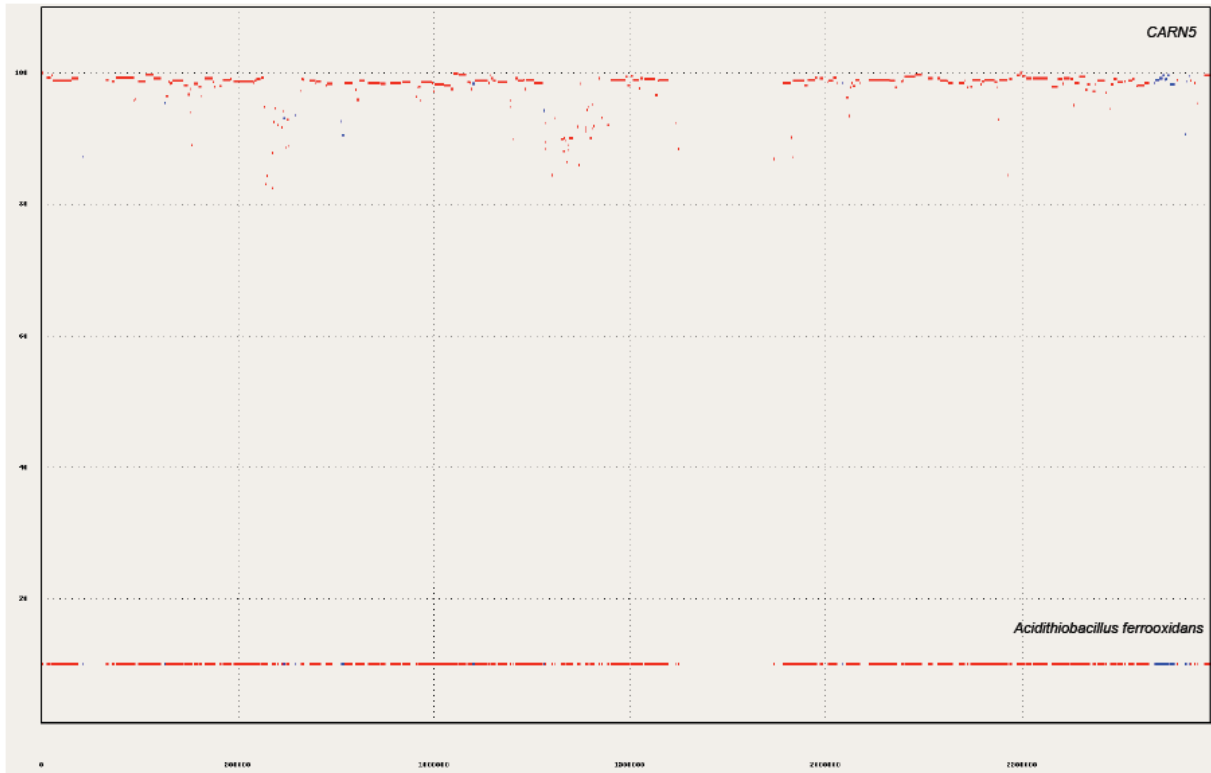
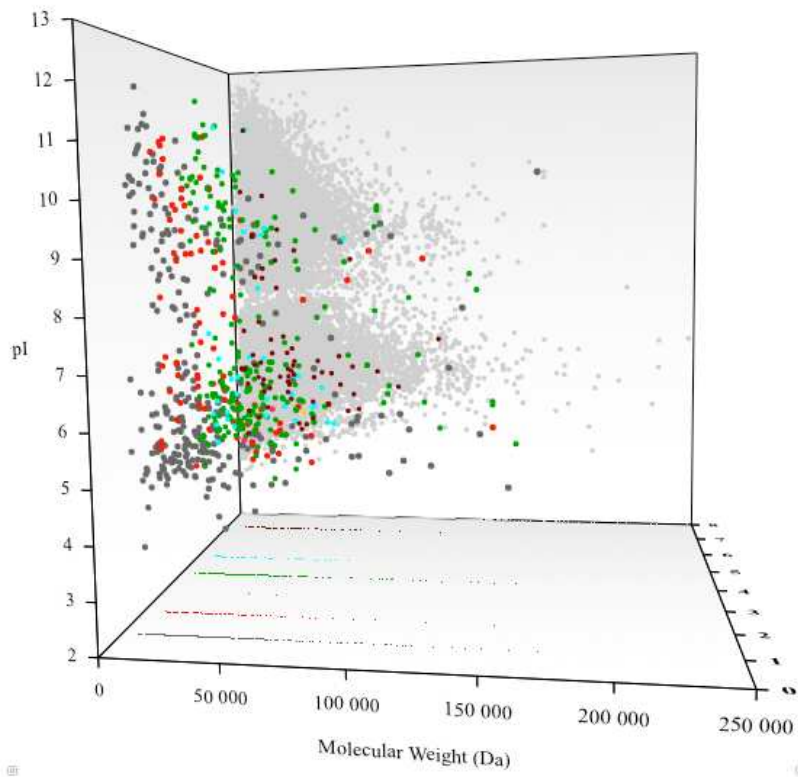


Figure S4



Conclusion

Dans cette première partie des résultats, nous avons présenté les développements méthodologiques nécessaires pour atteindre l'objectif principal de ce travail de thèse : l'utilisation des techniques de l'analyse protéomique comme aide à l'annotation génomique. Pour cela, il a notamment fallu développer :

- Une nouvelle méthode pour améliorer la caractérisation des extrémités N-terminales des protéines (stratégie N-TOP) pour une détermination facilitée des codons d'initiation des protéines.
- Une stratégie optimisée de recherche des données MS/MS dans les banques (génomiques ou protéiques) pour faciliter la mise en évidence des erreurs dans les banques protéiques.

Ces stratégies ont pu être appliquées à différentes études et ont permis d'améliorer significativement la qualité des banques protéiques par la correction d'erreurs de séquençage et surtout d'erreurs d'annotation des génomes, démontrant ainsi toute l'importance de la protéogénomique dans la constitution de banques protéiques de qualité.

Notre nouvelle stratégie N-TOP appliquée à l'étude du protéome de *M. smegmatis* a permis de répertorier un des plus grands ensembles d'extrémités N-terminales de protéines déterminées expérimentalement pour un organisme procaryote (637) avec une méthodologie simple et tout à fait compatible avec la plupart des projets de protéomique. Ces peptides N-terminaux ainsi que les peptides internes également identifiés par l'approche N-TOP ont permis de corriger les annotations du génome de l'organisme étudié. Cet ensemble de données a aussi permis d'évaluer la qualité de l'annotation génomique *in silico* des codons d'initiation des protéines pour la première fois au niveau d'un genre taxonomique et de corriger ces prédictions grâce à un nouveau mode extrinsèque d'annotation génomique basé sur les données protéomiques expérimentales.

Les stratégies développées dans cette partie ont également permis de mettre en évidence l'existence de fragments de génomes manquant (non séquencés ou non assemblés), et de participer à leur détermination, dans le contexte particulier de l'étude de communautés microbiennes par méta-protéo-génomique. Finalement, elles ont permis d'obtenir des données protéomiques expérimentales participant à la détermination de la séquence protéique d'une enzyme issue d'un organisme dont le génome n'est pas séquencé.

La stratégie N-TOP a également ouvert des perspectives dans deux autres champs de la protéomique relativement récents que sont la caractérisation (notamment la détermination des modifications post-traductionnelles) et la quantification des protéines. En effet, les résultats obtenus dans cette première partie laissent penser que le marquage N-terminal au TMPP pourrait être très utile pour améliorer la détection des peptides phosphorylés. De plus cette stratégie de marquage pourrait constituer un point de départ intéressant pour le développement d'une méthode de quantification protéomique. Ces deux nouvelles applications de la stratégie de marquage au TMPP seront développées dans les parties II et III de la partie « Résultats » de cette thèse.

Bibliographie

- Adamczyk, M., J. C. Gebler and J. Wu**
"Charge derivatization of peptides to simplify their sequencing with an ion trap mass spectrometer." *Rapid Commun Mass Spectrom*, **1999**, 13 (14), 1413-22.
- Aivaliotis, M., K. Gevaert, M. Falb, A. Tebbe, K. Konstantinidis, B. Bisle, C. Klein, L. Martens, A. Staes, E. Timmerman, J. Van Damme, F. Siedler, F. Pfeiffer, J. Vandekerckhove and D. Oesterhelt**
"Large-scale identification of N-terminal peptides in the halophilic archaea *Halobacterium salinarum* and *Natronomonas pharaonis*." *J Proteome Res*, **2007**, 6 (6), 2195-204.
- Allen, E. E. and J. F. Banfield**
"Community genomics in microbial ecology and evolution." *Nat Rev Microbiol*, **2005**, 3 (6), 489-98.
- Amaral Zettler, L. A., M. A. Messerli, A. D. Laatsch, P. J. Smith and M. L. Sogin**
"From genes to genomes: beyond biodiversity in Spain's Rio Tinto." *Biol Bull*, **2003**, 204 (2), 205-9.
- Armengaud, J.**
"A perfect genome annotation is within reach with the proteomics and genomics alliance." *Curr Opin Microbiol*, **2009**, 12 (3), 292-300.
- Bakalarski, C. E., W. Haas, N. E. Dephoure and S. P. Gygi**
"The effects of mass accuracy, data acquisition speed, and search algorithm choice on peptide identification rates in phosphoproteomics." *Anal Bioanal Chem*, **2007**, 389 (5), 1409-19.
- Bianchetti, L., J. D. Thompson, O. Lecompte, F. Plewniak and O. Poch**
"vALId: validation of protein sequence quality based on multiple alignment data." *J Bioinform Comput Biol*, **2005**, 3 (4), 929-47.
- Brown, N. P., C. Sander and P. Bork**
"Frame: detection of genomic sequencing errors." *Bioinformatics*, **1998**, 14 (4), 367-71.
- Bruneel, O., J. C. Personne, C. Casiot, M. Leblanc, F. Elbaz-Poulichet, B. J. Mahler, A. Le Fleche and P. A. Grimont**
"Mediation of arsenic oxidation by *Thiomonas* sp. in acid-mine drainage (Carnoules, France)." *J Appl Microbiol*, **2003**, 95 (3), 492-9.
- Campostrini, N., L. B. Areces, J. Rappsilber, M. C. Pietrogrande, F. Dondi, F. Pastorino, M. Ponzoni and P. G. Righetti**
"Spot overlapping in two-dimensional maps: a serious problem ignored for much too long." *Proteomics*, **2005**, 5 (9), 2385-95.
- Carapito, C.**
"Vers une meilleure utilisation des données de spectrométrie de masse en analyse protéomique." *Thèse de l'université Louis Pasteur de Strasbourg*, **2006**,
- Casiot, C., O. Bruneel, J. C. Personne, M. Leblanc and F. Elbaz-Poulichet**
"Arsenic oxidation and bioaccumulation by the acidophilic protozoan, *Euglena mutabilis*, in acid mine drainage (Carnoules, France)." *Sci Total Environ*, **2004**, 320 (2-3), 259-67.
- Casiot, C., G. Morin, F. Juillot, O. Bruneel, J. C. Personne, M. Leblanc, K. Duquesne, V. Bonnefoy and F. Elbaz-Poulichet**
"Bacterial immobilization and oxidation of arsenic in acid mine drainage (Carnoules creek, France)." *Water Res*, **2003**, 37 (12), 2929-36.
- Chamot-Rooke, J., G. van der Rest, A. Dalleu, S. Bay and J. Lemoine**
"The combination of electron capture dissociation and fixed charge derivatization increases sequence coverage for o-glycosylated and o-phosphorylated peptides." *J Am Soc Mass Spectrom*, **2007**, 18 (8), 1405-13.
- Chiang, Y. R., W. Ismail, D. Heintz, C. Schaeffer, A. Van Dorselaer and G. Fuchs**
"Study of anoxic and oxic cholesterol metabolism by *Sterolibacterium denitrificans*." *J Bacteriol*, **2008**, 190 (3), 905-14.
- Chiang, Y. R., W. Ismail, M. Muller and G. Fuchs**
"Initial steps in the anoxic metabolism of cholesterol by the denitrifying *Sterolibacterium denitrificans*." *J Biol Chem*, **2007**, 282 (18), 13240-9.
- Choi, K. P., I. Molnar, M. Yamashita and Y. Murooka**
"Purification and characterization of the 3-ketosteroid-delta 1-dehydrogenase of *Arthrobacter simplex* produced in *Streptomyces lividans*." *J Biochem*, **1995**, 117 (5), 1043-9.
- Choudhary, J. S., W. P. Blackstock, D. M. Creasy and J. S. Cottrell**
"Interrogating the human genome using uninterpreted mass spectrometry data." *Proteomics*, **2001**, 1 (5), 651-67.
- Constant, P., E. Perez, W. Malaga, M. A. Lancelle, O. Saurel, M. Daffe and C. Guilhot**

"Role of the pks15/1 gene in the biosynthesis of phenolglycolipids in the Mycobacterium tuberculosis complex. Evidence that all strains synthesize glycosylated p-hydroxybenzoic methyl esters and that strains devoid of phenolglycolipids harbor a frameshift mutation in the pks15/1 gene." *J Biol Chem*, **2002**, 277 (41), 38148-58.

Czeszak, X., W. Morelle, G. Ricart, D. Tetaert and J. Lemoine
 "Localization of the O-glycosylated sites in peptides by fixed-charge derivatization with a phosphonium group." *Anal Chem*, **2004**, 76 (15), 4320-4.

de Souza, G. A., H. Malen, T. Softeland, G. Saelensminde, S. Prasad, I. Jonassen and H. G. Wiker
 "High accuracy mass spectrometry analysis as a tool to verify and improve gene annotation using Mycobacterium tuberculosis as an example." *BMC Genomics*, **2008**, 9 316.

Domon, B. and R. Aebersold
 "Mass spectrometry and protein analysis." *Science*, **2006**, 312 (5771), 212-7.

Dormeyer, W., S. Mohammed, B. Breukelen, J. Krijgsveld and A. J. Heck
 "Targeted analysis of protein termini." *J Proteome Res*, **2007**, 6 (12), 4634-45.

Duquesne, K., A. Lieutaud, J. Ratouchniak, D. Muller, M. C. Lett and V. Bonnefoy
 "Arsenite oxidation by a chemoautotrophic moderately acidophilic Thiomonas sp.: from the strain isolation to the gene study." *Environ Microbiol*, **2008**, 10 (1), 228-37.

Falb, M., M. Aivaliotis, C. Garcia-Rizo, B. Bisle, A. Tebbe, C. Klein, K. Konstantinidis, F. Siedler, F. Pfeiffer and D. Oesterheld
 "Archaeal N-terminal protein maturation commonly involves N-terminal acetylation: a large-scale proteomics survey." *J Mol Biol*, **2006**, 362 (5), 915-24.

Fleischmann, R. D., M. D. Adams, O. White, R. A. Clayton, E. F. Kirkness, A. R. Kerlavage, C. J. Bult, J. F. Tomb, B. A. Dougherty, J. M. Merrick and et al.
 "Whole-genome random sequencing and assembly of Haemophilus influenzae Rd." *Science*, **1995**, 269 (5223), 496-512.

Foerstner, K. U., C. von Mering and P. Bork
 "Comparative analysis of environmental sequences: potential and challenges." *Philos Trans R Soc Lond B Biol Sci*, **2006**, 361 (1467), 519-23.

Freier, T. A., D. C. Beitz, L. Li and P. A. Hartman
 "Characterization of Eubacterium coprostanoligenes sp. nov., a cholesterol-reducing anaerobe." *Int J Syst Bacteriol*, **1994**, 44 (1), 137-42.

Galperin, M. Y., D. R. Walker and E. V. Koonin
 "Analogous enzymes: independent inventions in enzyme evolution." *Genome Res*, **1998**, 8 (8), 779-90.

Gevaert, K., M. Goethals, L. Martens, J. Van Damme, A. Staes, G. R. Thomas and J. Vandekerckhove
 "Exploring proteomes and analyzing protein processing by mass spectrometric identification of sorted N-terminal peptides." *Nat Biotechnol*, **2003**, 21 (5), 566-9.

Groisman, I. and H. Engelberg-Kulka
 "Translational bypassing: a new reading alternative of the genetic code." *Biochem Cell Biol*, **1995**, 73 (11-12), 1055-9.

Gupta, N., J. Benhamida, V. Bhargava, D. Goodman, E. Kain, I. Kerman, N. Nguyen, N. Ollikainen, J. Rodriguez, J. Wang, M. S. Lipton, M. Romine, V. Bafna, R. D. Smith and P. A. Pevzner
 "Comparative proteogenomics: combining mass spectrometry and comparative genomics to analyze multiple genomes." *Genome Res*, **2008**, 18 (7), 1133-42.

Hallberg, K. B., K. Coupland, S. Kimura and D. B. Johnson
 "Macroscopic streamer growths in acidic, metal-rich mine waters in north wales consist of novel and remarkably simple bacterial communities." *Appl Environ Microbiol*, **2006**, 72 (3), 2022-30.

Handelsman, J.
 "Metagenomics: application of genomics to uncultured microorganisms." *Microbiol Mol Biol Rev*, **2004**, 68 (4), 669-85.

Harder, J. and C. Probian
 "Anaerobic mineralization of cholesterol by a novel type of denitrifying bacterium." *Arch Microbiol*, **1997**, 167 (5), 269-74.

Hixson, K. K., J. N. Adkins, S. E. Baker, R. J. Moore, B. A. Chromy, R. D. Smith, S. L. McCutchen-Maloney and M. S. Lipton
 "Biomarker candidate identification in Yersinia pestis using organism-wide semiquantitative proteomics." *J Proteome Res*, **2006**, 5 (11), 3008-17.

Huang, Z. H., T. Shen, J. Wu, D. A. Gage and J. T. Watson
 "Protein sequencing by matrix-assisted laser desorption ionization-postsource decay-mass spectrometry analysis of the N-Tris(2,4,6-trimethoxyphenyl)phosphine-acetylated tryptic digests." *Anal Biochem*, **1999**, 268 (2), 305-17.

Jaffe, J. D., H. C. Berg and G. M. Church

- "Proteogenomic mapping as a complementary method to perform genome annotation." *Proteomics*, **2004**, 4 (1), 59-77.
- Johnson, D. B. and K. B. Hallberg**
"Acid mine drainage remediation options: a review." *Sci Total Environ*, **2005**, 338 (1-2), 3-14.
- Jungblut, P. R., E. C. Muller, J. Mattow and S. H. Kaufmann**
"Proteomics reveals open reading frames in Mycobacterium tuberculosis H37Rv not predicted by genomics." *Infect Immun*, **2001**, 69 (9), 5905-7.
- Kieslich, K.**
"Microbial side-chain degradation of sterols." *J Basic Microbiol*, **1985**, 25 (7), 461-74.
- Kolker, E., K. S. Makarova, S. Shabalina, A. F. Picone, S. Purvine, T. Holzman, T. Cherny, D. Armbruster, R. S. Munson, Jr., G. Kolesov, D. Frishman and M. Y. Galperin**
"Identification and functional analysis of 'hypothetical' genes expressed in Haemophilus influenzae." *Nucleic Acids Res*, **2004**, 32 (8), 2353-61.
- Kuster, B., P. Mortensen, J. S. Andersen and M. Mann**
"Mass spectrometry allows direct identification of proteins in large genomes." *Proteomics*, **2001**, 1 (5), 641-50.
- Link, A. J., L. G. Hays, E. B. Carmack and J. R. Yates, 3rd**
"Identifying the major proteome components of Haemophilus influenzae type-strain NCTC 8143." *Electrophoresis*, **1997**, 18 (8), 1314-34.
- Lipton, M. S., L. Pasa-Tolic, G. A. Anderson, D. J. Anderson, D. L. Auberry, J. R. Battista, M. J. Daly, J. Fredrickson, K. K. Hixson, H. Kostandarithes, C. Masselon, L. M. Markillie, R. J. Moore, M. F. Romine, Y. Shen, E. Stritmatter, N. Tolic, H. R. Udseth, A. Venkateswaran, K. K. Wong, R. Zhao and R. D. Smith**
"Global analysis of the Deinococcus radiodurans proteome by using accurate mass tags." *Proc Natl Acad Sci U S A*, **2002**, 99 (17), 11049-54.
- Liu, T., M. E. Belov, N. Jaitly, W. J. Qian and R. D. Smith**
"Accurate mass measurements in proteomics." *Chem Rev*, **2007**, 107 (8), 3621-53.
- Liu, Y., P. M. Harrison, V. Kunin and M. Gerstein**
"Comprehensive analysis of pseudogenes in prokaryotes: widespread gene decay and failure of putative horizontally transferred genes." *Genome Biol*, **2004**, 5 (9), R64.
- Ma, B., K. Zhang, C. Hendrie, C. Liang, M. Li, A. Doherty-Kirby and G. Lajoie**
"PEAKS: powerful software for peptide de novo sequencing by tandem mass spectrometry." *Rapid Commun Mass Spectrom*, **2003**, 17 (20), 2337-42.
- McDonald, L., D. H. Robertson, J. L. Hurst and R. J. Beynon**
"Positional proteomics: selective recovery and analysis of N-terminal proteolytic peptides." *Nat Methods*, **2005**, 2 (12), 955-7.
- Medigue, C., M. Rose, A. Viari and A. Danchin**
"Detecting and analyzing DNA sequencing errors: toward a higher quality of the Bacillus subtilis genome sequence." *Genome Res*, **1999**, 9 (11), 1116-27.
- Morii, S., C. Fujii, T. Miyoshi, M. Iwami and E. Itagaki**
"3-Ketosteroid-delta1-dehydrogenase of Rhodococcus rhodochrous: sequencing of the genomic DNA and hyperexpression, purification, and characterization of the recombinant enzyme." *J Biochem*, **1998**, 124 (5), 1026-32.
- Olsen, J. V., S. E. Ong and M. Mann**
"Trypsin cleaves exclusively C-terminal to arginine and lysine residues." *Mol Cell Proteomics*, **2004**, 3 (6), 608-14.
- Overbeek, R., D. Bartels, V. Vonstein and F. Meyer**
"Annotation of bacterial and archaeal genomes: improving accuracy and consistency." *Chem Rev*, **2007**, 107 (8), 3431-47.
- Perez, E., P. Constant, A. Lemassu, F. Laval, M. Daffe and C. Guilhot**
"Characterization of three glycosyltransferases involved in the biosynthesis of the phenolic glycolipid antigens from the Mycobacterium tuberculosis complex." *J Biol Chem*, **2004**, 279 (41), 42574-83.
- Perrodou, E., C. Deshayes, J. Muller, C. Schaeffer, A. Van Dorselaer, R. Ripp, O. Poch, J. M. Reyrat and O. Lecompte**
"ICDS database: interrupted CoDing sequences in prokaryotic genomes." *Nucleic Acids Res*, **2006**, 34 (Database issue), D338-43.
- Picotti, P., R. Aebersold and B. Doman**
"The implications of proteolytic background for shotgun proteomics." *Mol Cell Proteomics*, **2007**, 6 (9), 1589-98.
- Ram, R. J., N. C. Verberkmoes, M. P. Thelen, G. W. Tyson, B. J. Baker, R. C. Blake, 2nd, M. Shah, R. L. Hettich and J. F. Banfield**
"Community proteomics of a natural microbial biofilm." *Science*, **2005**, 308 (5730), 1915-20.

Rison, S. C., J. Mattow, P. R. Jungblut and N. G. Stoker

"Experimental determination of translational starts using peptide mass mapping and tandem mass spectrometry within the proteome of *Mycobacterium tuberculosis*." *Microbiology*, **2007**, 153 (Pt 2), 521-8.

Roth, K. D., Z. H. Huang, N. Sadagopan and J. T. Watson

"Charge derivatization of peptides for analysis by mass spectrometry." *Mass Spectrom Rev*, **1998**, 17 (4), 255-74.

Sadagopan, N. and J. T. Watson

"Investigation of the tris(trimethoxyphenyl)phosphonium acetyl charged derivatives of peptides by electrospray ionization mass spectrometry and tandem mass spectrometry." *J Am Soc Mass Spectrom*, **2000**, 11 (2), 107-19.

Shen, P. T., J. L. Hsu and S. H. Chen

"Dimethyl isotope-coded affinity selection for the analysis of free and blocked N-termini of proteins using LC-MS/MS." *Anal Chem*, **2007**, 79 (24), 9520-30.

Souza-Egipsy, V., E. Gonzalez-Toril, E. Zettler, L. Amaral-Zettler, A. Aguilera and R. Amils

"Prokaryotic community structure in algal photosynthetic biofilms from extreme acidic streams in Rio Tinto (Huelva, Spain)." *Int Microbiol*, **2008**, 11 (4), 251-60.

Staes, A., P. Van Damme, K. Helsens, H. Demol, J. Vandekerckhove and K. Gevaert

"Improved recovery of proteome-informative, protein N-terminal peptides by combined fractional diagonal chromatography (COFRADIC)." *Proteomics*, **2008**, 8 (7), 1362-70.

Tan, G. L., W. S. Shu, K. B. Hallberg, F. Li, C. Y. Lan and L. N. Huang

"Cultivation-dependent and cultivation-independent characterization of the microbial community in acid mine drainage associated with acidic Pb/Zn mine tailings at Lechang, Guangdong, China." *FEMS Microbiol Ecol*, **2007**, 59 (1), 118-26.

Tarlera, S. and E. B. Denner

"*Sterolibacterium denitrificans* gen. nov., sp. nov., a novel cholesterol-oxidizing, denitrifying member of the beta-Proteobacteria." *Int J Syst Evol Microbiol*, **2003**, 53 (Pt 4), 1085-91.

van der Geize, R., G. I. Hessels and L. Dijkhuizen

"Molecular and functional characterization of the *kstD2* gene of *Rhodococcus erythropolis* SQ1 encoding a second 3-ketosteroid Delta(1)-dehydrogenase isoenzyme." *Microbiology*, **2002**, 148 (Pt 10), 3285-92.

VerBerkmoes, N. C., V. J. Denef, R. L. Hettich and J. F. Banfield

"Systems biology: Functional analysis of natural microbial consortia using community proteomics." *Nat Rev Microbiol*, **2009**, 7 (3), 196-205.

Wang, G., Z. Ge, D. A. Rasko and D. E. Taylor

"Lewis antigens in *Helicobacter pylori*: biosynthesis and phase variation." *Mol Microbiol*, **2000**, 36 (6), 1187-96.

Weinstock, G. M.

"Genomics and bacterial pathogenesis." *Emerg Infect Dis*, **2000**, 6 (5), 496-504.

Yamaguchi, M., D. Nakayama, K. Shima, H. Kuyama, E. Ando, T. A. Okamura, N. Ueyama, T. Nakazawa, S. Norioka, O. Nishimura and S. Tsunasawa

"Selective isolation of N-terminal peptides from proteins and their de novo sequencing by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry without regard to unblocking or blocking of N-terminal amino acids." *Rapid Commun Mass Spectrom*, **2008**, 22 (20), 3313-9.

Yamaguchi, M., T. Obama, H. Kuyama, D. Nakayama, E. Ando, T. A. Okamura, N. Ueyama, T. Nakazawa, S. Norioka, O. Nishimura and S. Tsunasawa

"Specific isolation of N-terminal fragments from proteins and their high-fidelity de novo sequencing." *Rapid Commun Mass Spectrom*, **2007**, 21 (20), 3329-36.

Yates, J. R., 3rd, J. K. Eng and A. L. McCormack

"Mining genomes: correlating tandem mass spectra of modified and unmodified peptides to sequences in nucleotide databases." *Anal Chem*, **1995**, 67 (18), 3202-10.

Partie II : La dérivation chimique pour améliorer l'exploration des phosphoprotéomes

Chapitre 1 : Analyse des phosphorylations par spectrométrie de masse

Chapitre 2 : Le marquage TMPP pour l'analyse phosphoprotéomique

Chapitre 1 : Analyse des phosphorylations par spectrométrie de masse

L'achèvement du décryptage du génome humain a révélé que le nombre de gènes présents était considérablement plus faible (20000-25000 gènes) que le nombre attendu. Il fut considéré rapidement que le nombre de protéines dans le protéome d'une espèce excède très largement le nombre de gènes dans le génome correspondant. Le protéome d'un organisme est donc plus complexe que son génome car un seul gène peut générer plusieurs protéines, résultant notamment d'épissages alternatifs ou de modifications post-traductionnelles comme la phosphorylation ou les processus protéolytiques par exemple, qui pourront avoir des fonctions ou des localisations différentes. Pour une compréhension plus poussée des fonctions biologiques des protéines, la détermination de leur séquence primaire en acides aminés n'est donc pas toujours suffisante et il peut être nécessaire de caractériser plus finement la protéine notamment par l'étude des modifications post-traductionnelles.

1. Les modifications post-traductionnelles des protéines

Plus de 300 modifications post-traductionnelles des protéines ont été identifiées. Elles influencent significativement la fonction, la localisation, l'activité et la structure des protéines. Les modifications post-traductionnelles les plus courantes sont la phosphorylation [Cohen, 2002], la glycosylation [Walsh et al., 2006], la conjugaison de petites protéines telle que l'ubiquitine [Nandi et al., 2006; Daviet et al., 2008] et ses analogues [Geiss-Friedlander et al., 2007], l'acétylation N-terminale [Polevoda et al., 2003] et des lysines [Sadoul et al., 2008] ainsi que la modification protéolytique des protéines [Lopez-Otin et al., 2002].

2. La phosphorylation des protéines

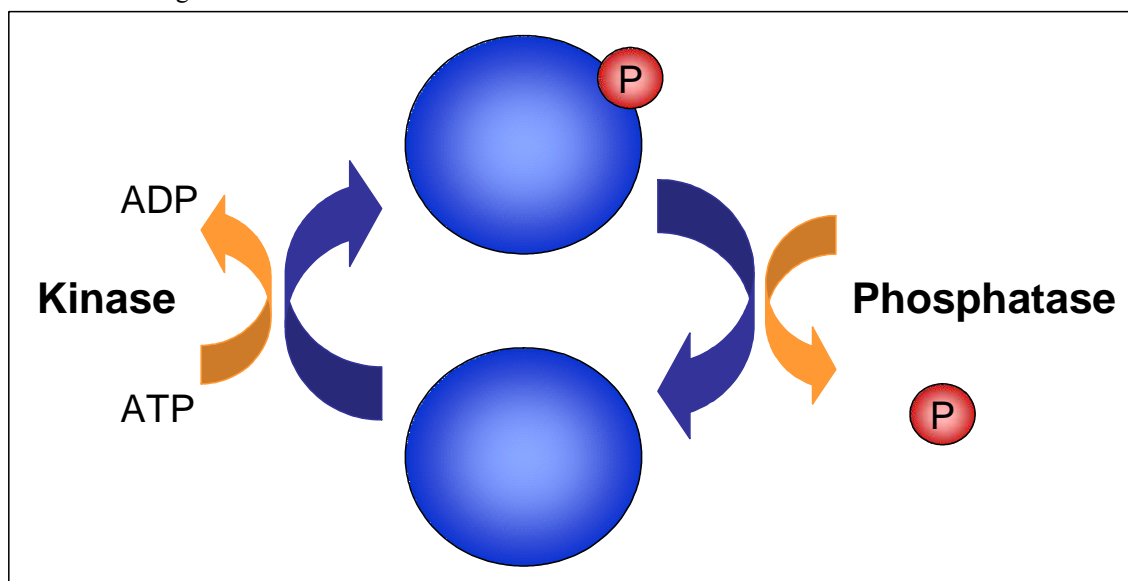
En 1955, la protéine « Glycogen phosphorylase » fut la première protéine cellulaire reconnue comme pouvant subir une phosphorylation et une déphosphorylation enzymatique [Fischer et al., 1955; Sutherland et al., 1955]. Dans une revue publiée par Krebs et Beavo en 1979, 20 enzymes ont été rapportées comme subissant des réactions de phosphorylation réversibles [Krebs et al., 1979]. 30 ans plus tard, de nombreuses avancées ont été réalisées et notre vision sur la transduction du signal a complètement changé. Aujourd'hui, on pense que la phosphorylation réversible est essentielle dans la régulation des activités cellulaires, ainsi que dans la prolifération et le développement des cellules. Il est estimé que chez l'homme 30 % des protéines sont phosphorylées, et autour de 500 kinases et 180

phosphatases sont codées par le génome [Hunter, 1995; Manning et al., 2002; Alonso et al., 2004; Wang et al., 2004; Arena et al., 2005]. De plus, puisque la phosphorylation des protéines peut altérer leur fonction, des maladies majeures comme le cancer, le diabète et la maladie d'Alzheimer ont été reconnues comme causées par, ou résultant en la dérégulation de la phosphorylation [Hunter, 2000; Blume-Jensen et al., 2001; Cohen, 2002; Arena et al., 2005].

Dans un organisme multicellulaire, des mécanismes de signalisation élaborés sont essentiels aux cellules pour communiquer les unes avec les autres. Les signaux intra- et inter-cellulaires sont interprétés et transférés par des réseaux complexes de protéines. Un des mécanismes par lequel les signaux sont transférés est la phosphorylation qui fournit un commutateur moléculaire pour l'activité de la protéine. Elle permet à la protéine d'exécuter différentes fonctions dans une cellule. La phosphorylation d'une protéine modifie sa conformation ou crée des sites de liaison pour d'autres protéines. Dans les voies de signalisation, des protéines peuvent être phosphorylées ou déphosphorylées et deviennent activées ou inactivées, conduisant à la modification d'une autre protéine qui devient activée ou inactivée, résultant en une cascade de signalisation à travers la cellule.

La phosphorylation des protéines cibles est catalysée par les kinases et leur déphosphorylation est catalysée par les phosphatases (Figure 1). Les cascades de signalisation commencent généralement avec la phosphorylation d'une protéine et se termine quand le groupement phosphate est enlevé par une phosphatase. Dans certains cas, la phosphorylation termine la cascade de signalisation et la déphosphorylation conduit à l'activation. L'équilibre entre l'action des kinases et des phosphatases est nécessaire pour le bon fonctionnement de la cellule [Hunter, 1995].

Figure 1 : Représentation schématique de l'action coordonnée des kinases et des phosphatases dans les voies de signalisation cellulaires



3. La nature des phosphorylations

Chez les eucaryotes, on rencontre le plus majoritairement des phosphorylations de type O-phosphorylation qui consistent en l'addition réversible d'un groupement phosphate labile sur la fonction alcool des sérines, thréonines et tyrosines. Comme au cours de ce travail de thèse, seules les O-phosphorylations des protéines ont été étudiées, nous ne porterons notre attention que sur ce type de phosphorylation.

4. L'analyse des phosphorylations

Etant donné l'importance des phosphorylations des protéines en biologie, leur analyse est évidemment primordiale. Comme pour l'analyse des séquences primaires des protéines, l'analyse protéomique par spectrométrie de masse est l'outil de choix pour l'analyse des phosphorylations [Mann et al., 2003].

Toutefois, l'analyse phosphoprotéomique est toujours un défi de part:

- La faible efficacité d'ionisation intrinsèque des phosphopeptides [Xu et al., 2008].
- La faible abondance des phosphoprotéines qui conduit à des effets de suppression d'ionisation sur les phosphopeptides dus à l'ensemble des peptides non-phosphorylés présents dans l'échantillon de manière plus abondante.
- Le caractère très hydrophile des phosphopeptides qui peut entraîner leur perte en chromatographie de phase inverse [Jalili et al., 2007].
- La faible efficacité de fragmentation des phosphopeptides en mode CID [Paradela et al., 2008]

Au cours des 10 dernières années, plusieurs stratégies ont été développées pour enrichir sélectivement les phosphopeptides et les fragmenter plus efficacement. Dans la suite de ce chapitre, je décrirai les techniques principalement utilisées dans ce but.

4.1. Enrichissement des phosphopeptides

4.1.1. Immunoprécipitation

L'enrichissement des phosphopeptides peut être réalisé par immunoprécipitation à l'aide d'anticorps dirigés contre les résidus phosphorylés. Généralement, cette technique est utilisée pour les résidus tyrosines phosphorylés [Steen et al., 2002] car les anticorps dirigés contre les résidus sérine et thréonine phosphorylés sont peu efficaces [Mann et al., 2002].

4.1.2. Chromatographie d'affinité

4.1.2.1. IMAC (« Immobilized Metal-ion Affinity Chromatography »)

La méthode IMAC basée sur l'affinité des phosphopeptides pour certains métaux (démontrée la première fois en 1986 [Andersson et al., 1986]) est sans doute la technique d'enrichissement des phosphopeptides la plus utilisée. Le traitement préalable des échantillons avec cette technique permet une analyse par spectrométrie de masse plus efficace des phosphopeptides car elle réduit les effets de suppression d'ionisation causés par l'ensemble des peptides non phosphorylés de l'échantillon [Corthals et al., 2005].

Avec cette technique, les peptides phosphorylés sont liés à la phase stationnaire IMAC par des interactions électrostatiques entre leur groupement phosphate chargé négativement et les ions métalliques. Les ions métalliques immobilisés sur la colonne sont le plus souvent Fe^{3+} , Ga^{3+} , Al^{3+} ou Zr^{4+} qui ont été décrits comme présentant une bonne affinité avec les groupements phosphate des phosphopeptides [Paradela et al., 2008]. Généralement, les phosphopeptides sont liés à la phase IMAC dans des conditions acides ou neutres et élués en utilisant des tampons phosphates et/ou basiques. Un désavantage de cette méthode est que les résidus acides comme les acides aspartiques ou glutamiques peuvent aussi porter une charge négative dans les conditions de chargement de la phase et se lier aussi à celle-ci, empêchant la liaison efficace des peptides phosphorylés. Ce problème de liaison aspécifique pour les peptides acides peut être réduit par une estérification préalable des acides carboxyliques [Ficarro et al., 2002]. La technique IMAC pour l'enrichissement des phosphopeptides est devenue très populaire rapidement grâce à sa bonne compatibilité avec les techniques d'analyse MALDI-TOF ou LC-MS/MS.

4.1.2.2. MOAC (« Metal Oxide Affinity Chromatography »)

Plus récemment, la technique d'enrichissement des phosphopeptides par affinité pour les oxydes de métaux, essentiellement le dioxyde de titane (TiO_2), a émergé [Pinkse et al., 2004]. Cette technique est basée sur l'interaction sélective des groupements phosphate avec les microsphères poreuses de TiO_2 via une liaison bidentée avec la surface des microsphères. Les phosphopeptides sont piégés sur une colonne TiO_2 en conditions acides et élués en conditions alcalines. La spécificité des colonnes TiO_2 pour les phosphopeptides est rapportée comme meilleure que celle de l'IMAC bien que ces colonnes retiennent également des peptides acides non phosphorylés. L'addition de l'acide 2,5-dihydroxybenzoïque (DHB) dans le mélange peptidique à enrichir permet de réduire les liaisons aspécifiques des peptides acides [Larsen et al., 2005].

Récemment, l'utilisation du dioxyde de zirconium (ZrO_2) à la place du TiO_2 a également été rapportée pour l'enrichissement des phosphopeptides avec des performances globales similaires à celles de TiO_2 [Kweon et al., 2006].

4.1.3. Chromatographie d'échange de cations (« Strong Cation Exchange », SCX)

La chromatographie d'échange de cations a été appliquée avec succès comme alternative aux chromatographies d'affinité pour séparer les phosphopeptides de l'ensemble de tous les peptides de digestion [Beausoleil et al., 2004; Gruhler et al., 2005]. Dans cette technique, la rétention sur la colonne dépend des interactions entre la résine chargée négativement de la colonne et les peptides chargés positivement. Si l'échantillon est chargé sur la colonne dans des conditions fortement acides (pH~2.7), les acides carboxyliques sont neutres alors que les groupements phosphates retiennent une charge négative. Par conséquent, la charge des phosphopeptides tryptiques est réduite de +2 à +1 et les forces d'interaction entre les phosphopeptides et la résine échangeuse de cations sont diminuées. L'élution avec un gradient de concentration en sels croissant permet donc d'éluer les phosphopeptides plus rapidement que les peptides non phosphorylés.

4.1.4. Modification chimique

La modification chimique est une autre méthode d'enrichissement des phosphopeptides. Par exemple, le groupement phosphate des sérines et des thréonines peut être enlevé par β -élimination et remplacé par un di-thiol par addition de Michael d'un groupement nucléophile porteur d'une « étiquette » chimique. Cette « étiquette » chimique sert ensuite de cible pour une chromatographie d'affinité qui sépare donc les peptides marqués chimiquement (auparavant phosphorylés) de tous les autres peptides [Oda et al., 2001; McLachlin et al., 2003; He et al., 2004]. L'inconvénient de cette méthode est son caractère aspécifique et l'impossibilité de l'appliquer sur les résidus tyrosine phosphorylés.

Une autre approche d'enrichissement des phosphopeptides par modification chimique consiste à activer les groupements phosphate et à établir une liaison phosphoramidate entre ceux-ci et un support solide en utilisant plusieurs étapes de dérivations chimiques (méthode « Phosphoramidate Chemistry », PAC) [Zhou et al., 2001]. Après élimination des peptides non-phosphorylés, les phosphopeptides sont décrochés du support dans des conditions acides douces sous leur forme native. Cette méthode peut être appliquée aux résidus sérine, thréonine et tyrosine phosphorylés. Elle reste toutefois fastidieuse notamment à cause du grand nombre d'étapes qui peut aussi entraîner des pertes d'échantillon. Récemment, des améliorations apportées à cette approche ont amélioré son efficacité et facilité son application, en particulier grâce à la réduction du nombre d'étapes [Tao et al., 2005; Bodenmiller et al., 2007].

En conclusion, toutes ces méthodes ont leurs propres avantages et inconvénients et notamment des spécificités différentes. Par exemple, il est généralement rapporté que la chromatographie TiO_2 enrichit plus efficacement les peptides monophosphorylés alors que l'IMAC tendrait à enrichir préférentiellement les peptides porteurs de plusieurs phosphorylations [Thingholm et al., 2009]. Une étude comparative des méthodes d'enrichissement de phosphopeptides les plus courantes (IMAC,

TiO₂, PAC) réalisée par l'équipe de R. Aebersold a montré qu'aucune des méthodes testées ne permettaient d'identifier le phosphoprotéome entier à l'étude mais que les différentes méthodes permettaient d'identifier des fractions distinctes et se chevauchant partiellement du phosphoprotéome [Bodenmiller et al., 2007].

Pour améliorer l'enrichissement sur les échantillons très complexes, il est également possible de coupler les méthodes entre elles comme dans la stratégie SIMAC qui combine IMAC et TiO₂ [Thingholm et al., 2008]. L'utilisation de la chromatographie échangeuse de cations, voire d'une autre technique chromatographique comme l'HILIC, est également utile comme méthode de fractionnement en combinaison avec une méthode de chromatographie d'affinité [McNulty et al., 2008].

L'utilisation de plusieurs méthodes d'enrichissement de phosphopeptides complémentaires semble être le meilleur moyen de maximiser la couverture du phosphoprotéome. Toutefois, les quantités de protéines disponibles dans les échantillons biologiques d'intérêt sont souvent limitées et ne permettent pas l'utilisation en parallèle de toutes ces méthodes d'enrichissement.

4.2. Séquençage des phosphopeptides par spectrométrie de masse en tandem

Les phosphoprotéines sont généralement phosphorylées sur différents sites et à des degrés différents. Il est important d'être capable d'attribuer les acides aminés spécifiques qui se phosphorylent en réponse à un événement biologique. Actuellement, dans la plupart des études, les sites de phosphorylation sont identifiés par spectrométrie de masse en tandem. Cependant, le séquençage des phosphopeptides par MS/MS n'est pas trivial à cause de leur faible efficacité d'ionisation et de la perte du groupement phosphate labile qui conduit à un faible nombre d'ions fragments pour l'identification des phosphopeptides.

4.2.1. Fragmentation MS² induite par collision (« Collision-Induced Dissociation », CID)

La phosphorylation sur les résidus sérine et thréonine est souvent labile et la fragmentation CID conduit à la perte partielle typique d'acide phosphorique (perte de neutre H₃PO₄, 98 Da) résultant de la β-élimination en phase gazeuse de la liaison phosphoester, générant la déhydroalanine et l'acide déhydroaminobutyrique. L'essentiel de l'énergie de collision est utilisé pour cette voie de fragmentation et il en résulte une fragmentation réduite du squelette peptidique qui limite l'information de séquence obtenue et compromet l'identification des sites de phosphorylation. Des pertes de neutre partielles sont aussi observées sur les résidus phosphotyrosine (perte de neutre HPO₃, 80 Da) bien que les groupements phosphate sur les résidus tyrosine soient beaucoup plus stables que sur les résidus sérine et thréonine. Un ion immonium phosphotyrosine caractéristique de m/z 216 dans

le spectre de fragmentation peut être utilisé comme indicateur de la présence d'un résidu phosphotyrosine dans la séquence peptidique [Steen et al., 2001; Steen et al., 2002].

4.2.2. Fragmentation MS³ et pseudo-MS³ induite par collision

Puisque les groupements phosphate sur les résidus sérine et thréonine sont labiles, l'ion issu de la perte de l'acide phosphorique peut être sélectionné pour une deuxième étape de fragmentation MS³ [Beausoleil et al., 2004]. Dans ce cas, au cours des analyses nanoLC-MS³, quand la perte de neutre est détectée dans le spectre MS/MS d'un ion précurseur, l'ion fragment correspondant est automatiquement isolé pour une nouvelle étape de fragmentation. Cette méthode fournit plus d'informations sur la séquence des phosphopeptides et aide à l'attribution du site exact de phosphorylation. Il faut toutefois noter que la fragmentation MS³ n'est efficace que si les ions phosphopeptides sont suffisamment abondants.

L'activation multi-niveaux (« Multiple-Stage Activation », MSA) ou pseudo-MS³ est une variation de la fragmentation MS³. En MSA, la fragmentation de l'ion précurseur est réalisée simultanément avec la fragmentation de l'ion résultant de la perte de neutre. Les spectres MS² et MS³ sont donc combinés dans un spectre « hybride » qui peut être plus informatif sur la séquence des phosphopeptides et faciliter l'attribution des sites de phosphorylation [Schroeder et al., 2004]. L'utilisation de ces modes MS³ conduit toutefois à un nombre de phosphopeptides fragmentés plus faible à cause du temps nécessaire à la réalisation de ces cycles.

4.2.3. Dissociation par capture d'électron (« Electron Capture Dissociation », ECD) et dissociation par transfert d'électron (« Electron Transfer Dissociation », ETD)

Les techniques de fragmentations ETD et ECD ont été introduites récemment pour limiter la perte prédominante du groupement phosphate observée dans les spectres CID. Avec ces 2 techniques de fragmentation, les modifications post-traductionnelles labiles restent intactes rendant l'attribution des sites de phosphorylation plus précise [Mikesh et al., 2006] et la détermination des séquences peptidiques plus facile. Le principe de l'ECD est de faire réagir des peptides multiples chargés avec des électrons faiblement énergétiques. Cela induit la fragmentation des liaisons amides (N-C α) pour produire des ions fragment de type c- et z-. Dans le cas de la fragmentation ETD, le transfert des électrons est réalisé entre des radicaux anions (par exemple anthracène ou fluoranthène monochargé) et les peptides multiples chargés qui se fragmentent ensuite selon un processus analogue à celui de l'ECD. L'utilisation du mode de fragmentation ETD dans une étude phosphoprotéomique à grande échelle a été rapportée comme permettant d'identifier plus de phosphopeptides [Molina et al., 2007] qu'en mode CID. Toutefois, le mode CID permet aussi d'identifier des phosphopeptides qui ne le sont pas en mode ETD. Finalement, la combinaison des modes CID et ETD permet d'obtenir les meilleurs

résultats car le chevauchement entre les 2 ensembles de phosphopeptides obtenus par chacun des modes de fragmentation est très faible [Molina et al., 2007]. Plusieurs exemples ont montré la possibilité d'utiliser le mode ECD pour l'identification des sites de phosphorylation [Stensballe et al., 2000; Shi et al., 2001; Chalmers et al., 2004] bien qu'aucune étude à grande échelle n'ait été encore publiée.

Chapitre 2 : Le marquage TMPP pour l'analyse phosphoprotéomique

Cette étude a été réalisée en collaboration avec l'équipe du Docteur Hubert Schaller et du Professeur Thomas J Bach du département « Réseau Métaboliques Végétaux » de l'Institut Biologie Moléculaire des Plantes (IBMP-CNRS, UPR2357) de l'Université de Strasbourg.

1. Contexte de l'étude

1.1. Le marquage TMPP

Comme nous l'avons vu dans le Chapitre 1, au cours des 10 dernières années, un grand nombre de stratégies a pu être développé pour enrichir sélectivement les phosphopeptides. Ces techniques ont été appliquées avec succès à des échantillons biologiques complexes préservant les phosphopeptides des effets de suppression ionique et de sous-échantillonnage dus aux peptides non-phosphorylés. Néanmoins, ces méthodes ne permettent pas en général d'améliorer l'efficacité d'ionisation intrinsèque, l'hydrophobicité et l'efficacité de fragmentation des phosphopeptides.

Or, nous avons démontré, lors du développement et de l'application de la stratégie N-TOP (Partie I des résultats, Chapitre 2), que le marquage N-terminal des protéines par le TMPP permettait aux peptides N-terminaux marqués de présenter une meilleure efficacité d'ionisation électrospray (particulièrement pour les peptides à faible efficacité d'ionisation), et une hydrophobicité plus importante en chromatographie de phase inverse, ce qui améliorerait leur détection en analyse LC-MS/MS. De plus, des études sur quelques phosphopeptides modèles ont montré que le marquage TMPP pouvait conduire à des spectres de fragmentation plus informatifs pour les phosphopeptides [Sadagopan et al., 1999; Chamot-Rooke et al., 2007]. Le marquage des phosphopeptides par le TMPP apparaît donc comme une stratégie potentiellement très intéressante pour améliorer leur caractérisation par LC-MS/MS.

1.2. Application en biochimie métabolique sur un modèle végétal : la biosynthèse d'isoprénoïdes chez *Arabidopsis thaliana*

Les isoprénoïdes représentent la plus grande classe de produits naturels comprenant plus de 35000 composés de structures chimiques très différentes. Tous les organismes utilisent les isoprénoïdes pour leur fonctionnement. C'est cependant chez les plantes que la palette d'isoprénoïdes

est la plus large. Les isoprénoïdes jouent chez les plantes des rôles majeurs dans des processus très divers tels que le fonctionnement des membranes cellulaires dont la fluidité est régulée par les stérols, la croissance et le développement de tissus et d'organes contrôlés par les hormones végétales, la transduction des signaux, la pollinisation, la défense contre les pathogènes [Harborne, 1991; Bouwmeester, 2006]. Les enzymes de biosynthèse des isoprénoïdes sont toutes très bien identifiées et assez largement caractérisées dans leur ensemble. Malgré l'importance des isoprénoïdes végétaux sur un plan fondamental et par ailleurs sur un plan pharmacologique, médical et industriel, les mécanismes qui régulent leur biosynthèse (flux de métabolites), leur transport et leur stockage au niveau cellulaire sont toujours très mal connus. Chacun des 35 000 isoprénoïdes décrits à ce jour est synthétisé à partir d'un précurseur commun, une molécule à cinq atomes de carbone, qui est l'isopentényl diphosphate (IPP). Cet IPP dérive de l'acétyl-coenzyme A via l'acide mévalonique chez les animaux, chez les champignons et chez les plantes supérieures : c'est la voie du mévalonate (MVA). Il existe une alternative à cette voie de biosynthèse de l'IPP : c'est la voie du méthylérythritol phosphate (MEP), propre aux eubactéries, et aux plastides des algues et des plantes supérieures [Rohmer, 1999].

La 3-Hydroxy-3-Methyl-Glutaryl-CoA Reductase (HMGR) est une enzyme-clé dans la voie du MVA chez les animaux et les plantes [Bach, 1986; Goldstein et al., 1990; Enjuto et al., 1994; Chappell et al., 1995; Ohyama et al., 2007]. Cette enzyme est indispensable à la synthèse des stérols, qu'elle contrôle (un excès de stérols inhibe l'activité enzymatique de HMGR [Harker et al., 2003]). Plusieurs copies du gène HMGR existent chez les plantes (par exemple, *HMGR1* et *HMGR2* chez *Arabidopsis thaliana* [Enjuto et al., 1995; Ohyama et al., 2007]). Chez cette dernière, l'inhibition génétique, c'est-à-dire la perte de fonction par mutation des deux gènes, ou l'inhibition chimique de l'activité des enzymes est létale. Si on ne mute que le gène *HMGR1* chez *A. thaliana*, la plante reste viable mais son développement est fortement perturbé et les quantités de stérols issus du métabolisme de l'HMG CoA sont réduites [Suzuki et al., 2004]. En fonction des conditions de l'environnement, c'est à dire des conditions de température et de photopériode, ce mutant *hmgr1-1* présente un développement faiblement perturbé consistant en un retard de développement (floraison) de 15 jours environ par rapport à la plante de génotype sauvage ainsi qu'une taille plus petite. La plante semble donc compenser l'inactivité d'un de ces gènes (*HMGR1*) en activant celle de l'autre gène (*HMGR2*), ou bien encore en mettant en œuvre un autre mécanisme pour le moment inconnu. Sachant que la régulation de l'activité de l'HMGR est réalisée au niveau transcriptionnel mais aussi au niveau post-traductionnel chez les animaux [Goldstein et al., 1990], on soupçonne légitimement dans le cas des plantes que le métabolisme de l'HMG CoA (synthèse d'isoprénoïdes et de stérols) et plus généralement la réponse de la plante, en terme de croissance, à l'effet de la déficience en *HMGR1*, puissent être régulées par phosphorylation / déphosphorylation. L'analyse différentielle des phosphoprotéomes du mutant *hmgr1-1* d'*Arabidopsis thaliana* et du sauvage de même fond génétique devrait permettre de mettre en évidence des protéines différenciellement phosphorylées qui seraient impliquées dans la régulation de la biosynthèse et du transport des stérols, leur adressage et leur stockage chez les plantes, par exemple. Enfin, comme les enzymes impliquées dans la synthèse de stérols sont membranaires, et que les localisations finales des stérols sont les membranes plasmiques et

du réticulum endoplasmique essentiellement, l'étude du phosphoprotéome membranaire ou associé à la membrane revêt donc une importance toute particulière.

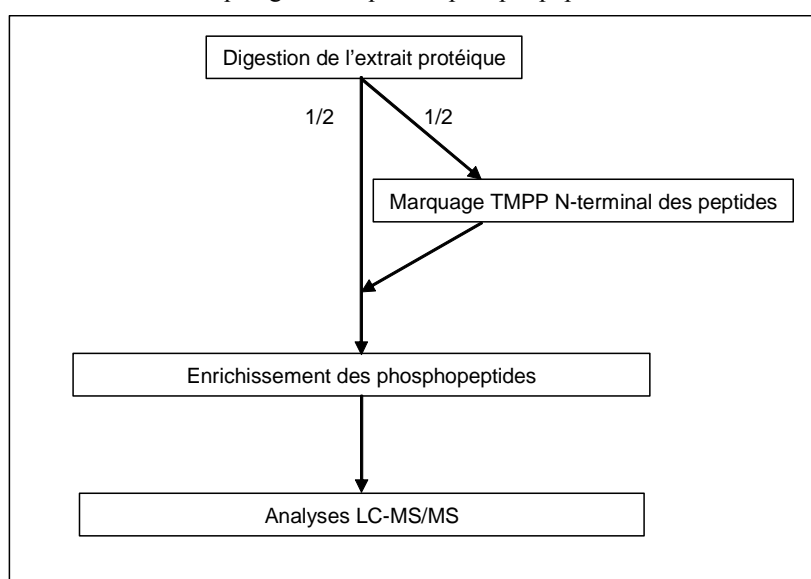
1.3. Objectif de l'étude

Nous avons développé une stratégie d'analyse phosphoprotéomique d'un extrait protéique utilisant le marquage N-terminal au TMPP de l'ensemble des peptides de digestion. Cibler l'extrémité N-terminale des peptides permet de modifier tous les phosphopeptides sélectivement en exploitant les différences de pKa entre l'extrémité N-terminale et l'amine libre de la chaîne latérale des lysines des peptides alors que l'extrémité C-terminale et la fonction acide carboxylique de la chaîne latérale des acides aspartiques et glutamiques ont des valeurs de pKa similaires.

La stratégie développée ici consiste donc à insérer, après digestion de l'échantillon, une étape supplémentaire de dérivation chimique de la moitié de l'échantillon dans un mode opératoire « classique » d'analyse phosphoprotéomique. Après cette étape, les deux mélanges (modifiés et non-modifiés) sont regroupés avant enrichissement des phosphopeptides et analyse LC-MS/MS (Figure 1).

Après évaluation du potentiel de cette stratégie sur des protéines modèles, elle a ensuite pu être appliquée à l'exploration des phosphoprotéomes de fractions microsomales d'*A. thaliana*. Dans notre étude, les fractions protéiques microsomales d'*A. thaliana* sauvage et du mutant *hmgr1-1* d'*A. thaliana* ont été analysées et comparées pour obtenir un premier aperçu des différences entre leur phosphoprotéome respectif. Les fractions protéiques microsomales ont été ciblées pour cette étude car elles présentent l'avantage d'être enrichies en protéines membranaires et associées à la membrane tout en étant simples à préparer.

Figure 1 : Principe de la stratégie d'analyse phosphoprotéomique utilisant le marquage chimique des phosphopeptides au TMPP



2. Evaluation de la dérivation chimique au TMPP pour l'analyse phosphoprotéomique

Afin d'évaluer le potentiel de la stratégie de dérivation chimique au TMPP pour améliorer l'analyse phosphoprotéomique, nous avons décidé de l'appliquer sur des digests de protéines modèles phosphorylées : alpha-casein S1 (α -S1), alpha-casein S2 (α -S2) et beta-casein (β -C).

2.1. Impact du marquage TMPP sur l'identification des phosphopeptides par LC-MS/MS

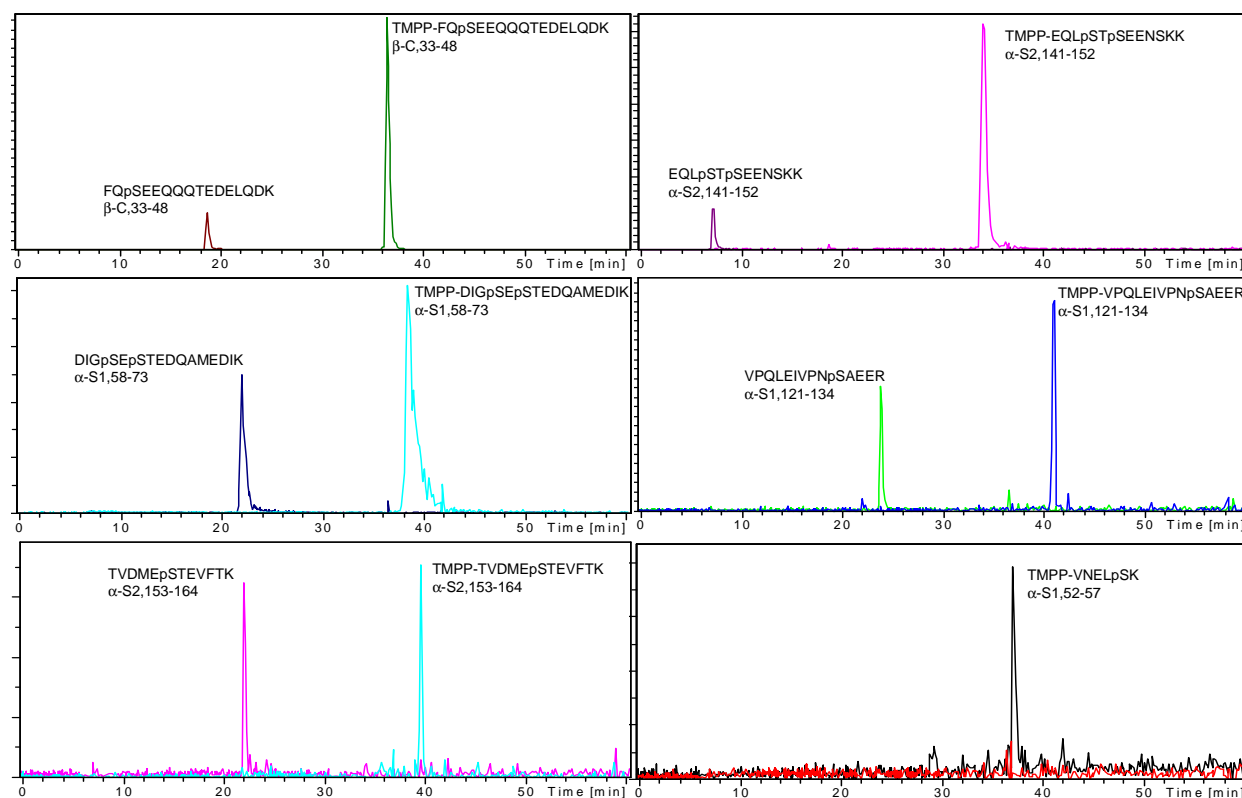
Dans les applications précédentes de la stratégie N-TOP (Partie I des résultats, Chapitre 2), il a pu être constaté que les peptides marqués au TMPP étaient séparés en chromatographie en phase inverse des peptides non marqués grâce à l'augmentation de leur hydrophobicité. Par conséquent, ces deux classes de peptides n'interfèrent pas mutuellement lors des analyses LC-MS/MS. Nous avons donc décidé dans notre mode opératoire de réaliser le marquage TMPP sur la moitié des digests de protéines modèles puis de regrouper les deux mélanges avant enrichissement par chromatographie d'affinité IMAC fer et analyses LC-MS/MS. Les analyses ont été réalisées sur un système microHPLC (Agilent Series 1100) couplé à une trappe ionique HCT Ultra (Bruker Daltonics). Les données MS/MS issues de l'analyse ont été soumises à une interrogation Mascot et ont permis l'identification, grâce aux phosphopeptides non-modifiés et modifiés au TMPP, des 8 sites de phosphorylation habituellement identifiés en mode d'ionisation électrospray [Larsen et al., 2005]. Les chromatogrammes d'extraits d'ions présentés en Figure 2 montrent que le marquage TMPP a augmenté le temps de rétention des phosphopeptides et amélioré dans l'ensemble leur efficacité d'ionisation. Tous les phosphopeptides identifiés sont présents sous forme « native » et sous forme modifiée au TMPP excepté le peptide VNELpSK (α -S1, 52-57) identifié uniquement sous forme modifiée. Les efficacités d'ionisation après marquage ont été améliorées d'un facteur 1 à 7 pour les autres phosphopeptides. Le phosphopeptide VNELpSK n'est habituellement pas identifié dans les études phosphoprotéomiques de digests de caséines mais le site de phosphorylation de ce peptide est parfois observés grâce à une coupure tryptique manquée dans le peptide VNELpSKDIGpSEpSTEDQAMEDIK (α -S1, 52-73) [Larsen et al., 2005]. Dans notre analyse, les temps de rétention en analyse chromatographique en phase inverse ont été inférieurs à 24 minutes pour les phosphopeptides natifs et supérieurs à 36 minutes pour les phosphopeptides modifiés au TMPP, ce qui a bien permis de les séparer en deux zones d'élution distinctes comme nous l'attendions. Par conséquent, ces deux classes de peptides ne devraient pas entraîner de suppression d'ionisation mutuelle et peuvent être analysés simultanément. De plus, l'augmentation du temps de rétention des phosphopeptides après marquage au TMPP permet d'empêcher leur perte lors de leur chargement sur colonne chromatographique de phase inverse, ce qui est probablement le cas pour VNELpSK. Dans notre expérience, 4 peptides mono-phosphorylés et 2 peptides doublement phosphorylés ont été

identifiés. Ce biais pour la détection des peptides mono-phosphorylés n'est pas étonnant car plusieurs études ont rapporté que pour les échantillons analysés en mode électrospray on pouvait constater :

- une détection préférentielle des peptides mono-phosphorylés
- que plusieurs peptides multiplement phosphorylés détectés en MADI-MS n'étaient pas observés en mode d'ionisation électrospray [Gruhler et al., 2005; Larsen et al., 2005].

Nous avons également réalisé l'enrichissement IMAC fer et l'analyse LC-MS/MS d'une part uniquement sur le digest de protéines modèles n'ayant pas subi le marquage TMPP et d'autre part uniquement sur le digest de protéines modèles ayant subi le marquage TMPP. Les résultats obtenus ont été très similaires à ceux issus de l'expérience où les digests étaient recombinés. Par conséquent, pour la suite de l'étude, il a donc été choisi de conserver la recombinaison des deux types d'échantillon avant enrichissement, ce qui permet de limiter le nombre d'analyses LC-MS/MS finales.

Figure 2 : Chromatogrammes d'ions extraits des phosphopeptides natifs et de leurs analogues modifiés TMPP issus des digests d'alpha-casein S1 (α -S1), alpha-casein S2 (α -S2) and beta-casein (β -C).



Pour augmenter le nombre de phosphopeptides identifiés, une stratégie alternative consiste à substituer la chromatographie d'enrichissement IMAC par la combinaison d'une précipitation au phosphate de calcium avec une chromatographie d'affinité au dioxyde de titane [Zhang et al., 2007]. L'utilisation de ce mode d'enrichissement sur nos digests de protéines modèles a permis d'identifier par analyse nano-LC-MS/MS 4 peptides porteurs de 4 phosphorylations qui n'avaient pas été détectés dans l'analyse précédente. Deux de ces phosphopeptides ont été identifiés sous forme native et 2 sous forme modifiée au TMPP. Il est intéressant d'observer que les 2 peptides portant 4 phosphorylations

identifiés sous forme native (KNTMEHVpSpSpSEEpSIISQETYK, α -S2, 16-36 et RELEELNVPGEIVEpSLpSpSpSEESITR, β -C, 16-40) contiennent une coupure trypsique « manquée » et donc un acide aminé basique en plus dans leur séquence. Cette augmentation du caractère basique, et donc du pI, des phosphopeptides semble être nécessaire pour compenser la diminution de leur charge globale causée par le nombre important de groupements phosphate portant une charge négative. Le groupement TMPP permet aussi de compenser cette diminution de charge grâce à sa charge permanente positive comme illustré par le phosphopeptide TMPP-NTMEHVpSpSpSEEpSIISQETYK (α -S2, 17-36) identifié sans coupure trypsique manquée contrairement à son analogue natif. De même le phosphopeptide TMPP-NANEEEEYSIGpSpSpSEEpSAEVATEEVK (α -S2, 61-85), porteur lui aussi de 4 groupements phosphate, a été identifié uniquement sous forme modifiée au TMPP.

Ces résultats soulignent que le marquage TMPP permet d'identifier des phosphopeptides qui ne le sont pas sous forme native mais aussi que certains phosphopeptides identifiés sous forme native ne le sont pas sous forme marquée (Tableau 1). Ces résultats préliminaires laissent entrevoir une complémentarité des deux populations de peptides (phosphopeptides marqués et phosphopeptides natifs) pour l'analyse phosphoprotéomique.

Tableau 1 : Phosphopeptides issus des protéines modèles identifiés sous forme native et/ou marquée au TMPP

Phosphopeptide	Protéine et position dans la séquence	Nombre de groupements phosphate	Identifié sous forme native	Identifié sous forme marquée au TMPP
FQpSEEQQTEDELQDK	β -C, 33-48	1	Oui	Oui
DIGpSEpSTEDQAMEDIK	α -S1, 58-73	2	Oui	Oui
TVDMEpSTEVFTK	α -S2, 153-164	1	Oui	Oui
EQLpSTpSEENSK	α -S2, 141-152	2	Oui	Oui
VPQLEIVPnPSAEER	α -S1, 121-134	1	Oui	Oui
VNELpSK	α -S1, 52-57	1	Non	Oui
KNTMEHVpSpSpSEEpSIISQETYK	α -S2, 16-36	4	Oui	Non
RELEELNVPGEIVEpSLpSpSpSEESITR	β -C, 16-40	4	Oui	Non
NTMEHVpSpSpSEEpSIISQETYK	α -S2, 17-36	4	Non	Oui
NANEEEEYSIGpSpSpSEEpSAEVATEEVK	α -S2, 61-85	4	Non	Oui

2.2. Impact du marquage TMPP sur la fragmentation des phosphopeptides

Le marquage TMPP modifie aussi le schéma de fragmentation des phosphopeptides. La charge permanente apportée par le groupement TMPP conduit généralement à l'observation

préférentielle d'ions fragments N-terminaux dans les spectres de fragmentation CID des phosphopeptides modifiés. La comparaison des spectres de fragmentation des phosphopeptides natifs et des phosphopeptides modifiés présentant le même état de charge est réalisée en Figure 3 avec différents exemples de spectres de fragmentation de peptides doublement et triplement chargés et porteurs de 1 ou 2 groupements phosphate. Pour chaque phosphopeptide examiné, nous avons évalué dans le Tableau 2 l'efficacité de la fragmentation avec :

- Le calcul d'un ratio : « nombre de groupements phosphate retenus par les ions fragments » / « nombre de groupements phosphate perdus par les ions fragments ».
- L'estimation qualitative de l'abondance des pics correspondant à la perte d'acide phosphorique du précurseur.

Nous avons observé que le marquage TMPP ne conduit pas systématiquement à des spectres MS/MS plus informatifs des phosphopeptides comme suggéré dans l'étude de [Sadagopan et al., 1999]. Ainsi les phosphopeptides doublement chargés EQLpSTpSEENSK et VPQLEIVNpSAEER (Figure 3 A à D) présentent une meilleure efficacité de fragmentation sous forme native. Par contre, les phosphopeptides triplement chargés FQpSEEQQTEDELQDK et DIGpSEpSTEDQAMEDIK (Figure 3 E à H) présentent une fragmentation d'efficacité voisine sous forme native ou marquée au TMPP voire légèrement plus informative sous forme marquée (si on se réfère au ratio d'évaluation de la perte de groupements phosphate). Plusieurs études ont décrit l'étendue de la perte de neutre dans la fragmentation des phosphopeptides comme dépendant du rapport entre l'état de charge et le nombre d'acides aminés basiques du phosphopeptide et que plus ce ratio est élevé plus la perte de neutre est limitée [Tholey et al., 1999; Palumbo et al., 2008; Boersema et al., 2009]. Dans nos exemples, pour que ce rapport soit positif, les phosphopeptides sélectionnés pour la fragmentation doivent présenter un état de charge minimale 2^+ sous forme native et 3^+ sous forme marquée (à cause de la charge permanente positive qui remplace l'amine libre N-terminale et qui augmente donc le degré de localisation de la charge [Roth et al., 1998]). La nécessité de fragmenter les phosphopeptides marqués à l'état de charge 3^+ n'est pas rédhibitoire dans notre approche puisque l'addition du groupement TMPP conduit généralement à l'observation de ces états de charge plus élevés [Adamczyk et al., 1999].

Le groupement TMPP peut également avoir une influence sur :

- ✓ Les différents mécanismes à l'origine de la perte d'acide phosphorique en fragmentation CID [Boersema et al., 2009].
- ✓ Les voies de fragmentation du squelette peptidique puisque le marquage semble induire une fragmentation à la fois dirigée par la charge et distante de la charge (Partie II des résultats, Chapitre 2. 1.2.3.3.).

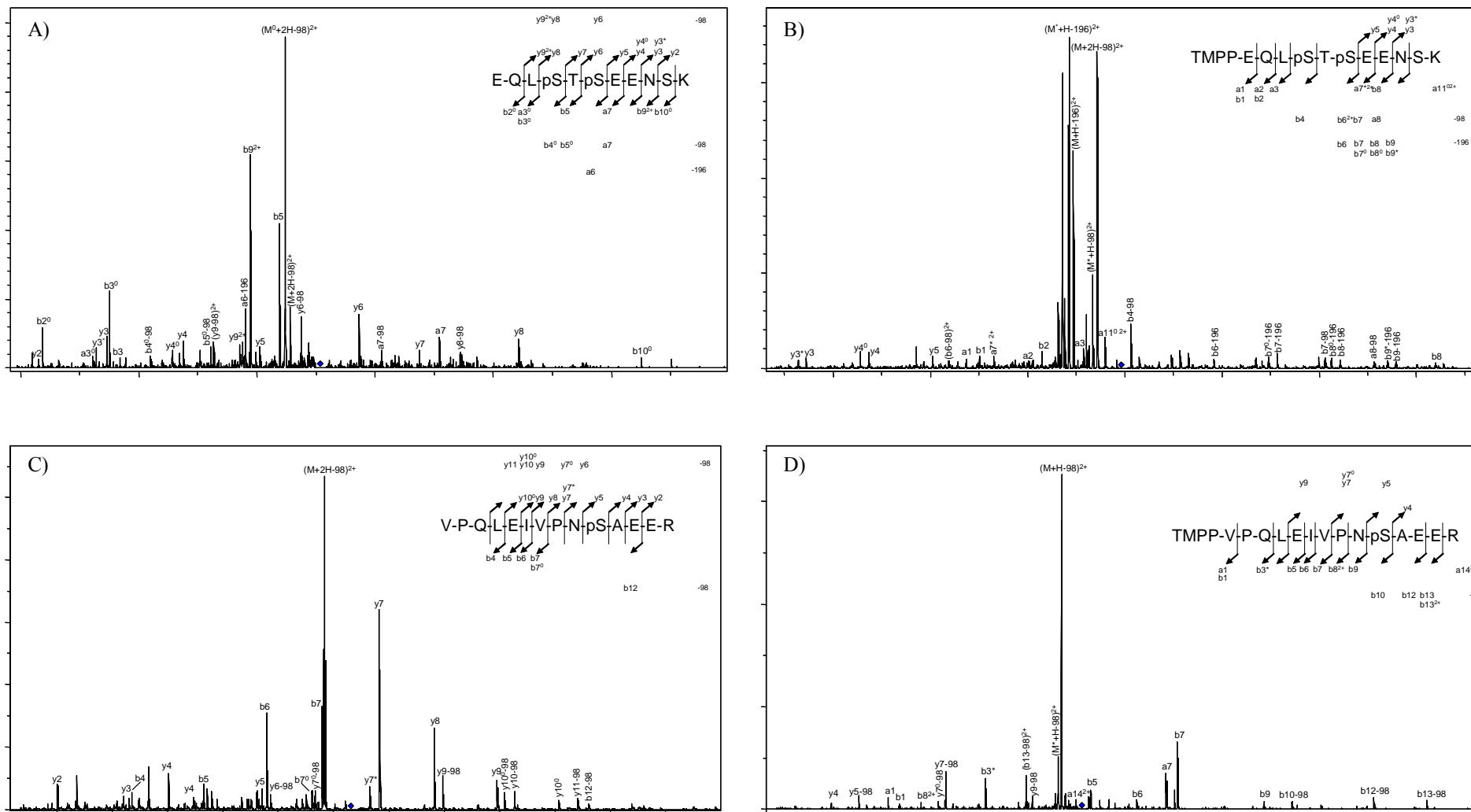
Toutefois, les données obtenues ici sont trop limitées pour pouvoir observer cette influence de manière significative. Néanmoins, ces données sont suffisantes pour laisser présager que le marquage TMPP devrait permettre d'améliorer la couverture du phosphoprotéome analysé grâce à des spectres MS/MS de phosphopeptides modifiés apportant des résultats complémentaires à ceux des phosphopeptides natifs. Des analyses supplémentaires ont été réalisées en modifiant l'amplitude de la tension

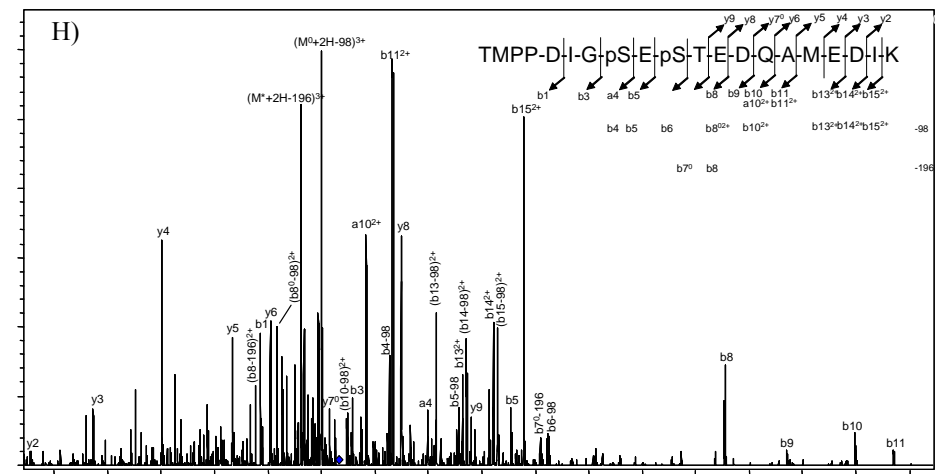
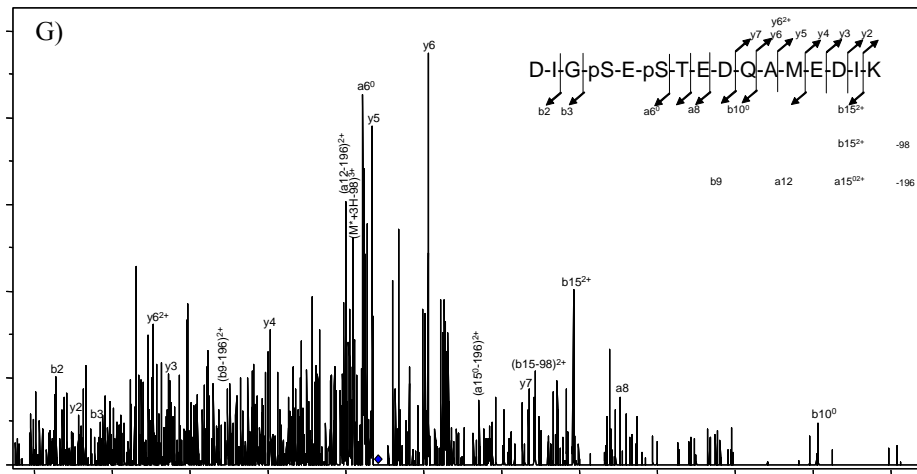
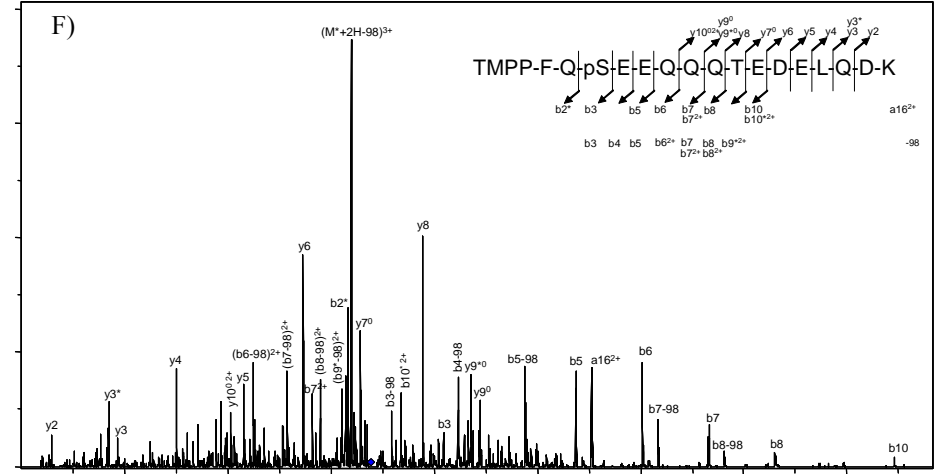
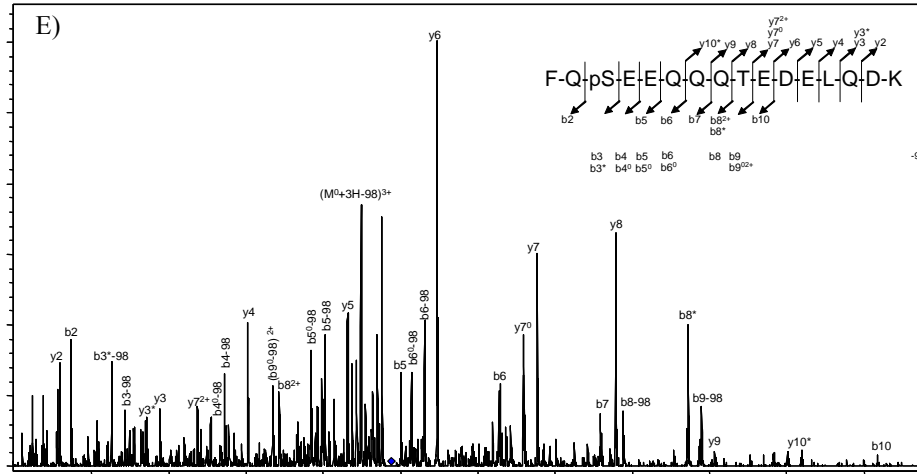
alternative appliquée sur l'électrode chapeau de la trappe pendant la fragmentation sans toutefois observer de modification significative des schémas de fragmentation.

Tableau 2 : Evaluation de l'efficacité de fragmentation des phosphopeptides sous forme native et marquée au TMPP. Le ratio d'évaluation de la perte de H_3PO_4 sur les ions fragments a été calculé en divisant le « nombre de groupements phosphate retenus par les ions fragments » par le « nombre de groupements phosphate perdus par les ions fragments ». L'évaluation de la perte de H_3PO_4 sur les ions précurseurs a été réalisée de manière qualitative en observant l'intensité des pics correspondant à ce phénomène comparativement à l'intensité des pics correspondant à la fragmentation du squelette peptidique

Phosphopeptide	Protéine et position dans la séquence	Nombre de groupements phosphate	Etat de charge	Forme native		Forme marquée au TMPP	
				Ratio d'évaluation de la perte de H_3PO_4 sur les ions fragments	Evaluation qualitative de la perte de H_3PO_4 sur le précurseur	Ratio d'évaluation de la perte de H_3PO_4 sur les ions fragments	Evaluation qualitative de la perte de H_3PO_4 sur le précurseur
EQLpSTpSEENSK	α -S2, 141-152	2	2 ⁺	2	+	0.5	++
VPQLEIVPNpSAEER	α -S1, 121-134	1	2 ⁺	0.9	+	0.1	++
FQpSEEQQTDELQDK	β -C, 33-48	1	3 ⁺	0.5	-	1	-
DIGpSEpSTEDQAMEDIK	α -S1, 58-73	2	3 ⁺	1.3	-	2.2	-

Figure 3 : Comparaison des spectres de fragmentation CID des phosphopeptides d'alpha-casein S1 et S2 et beta-casein natifs et marqués au TMPP.



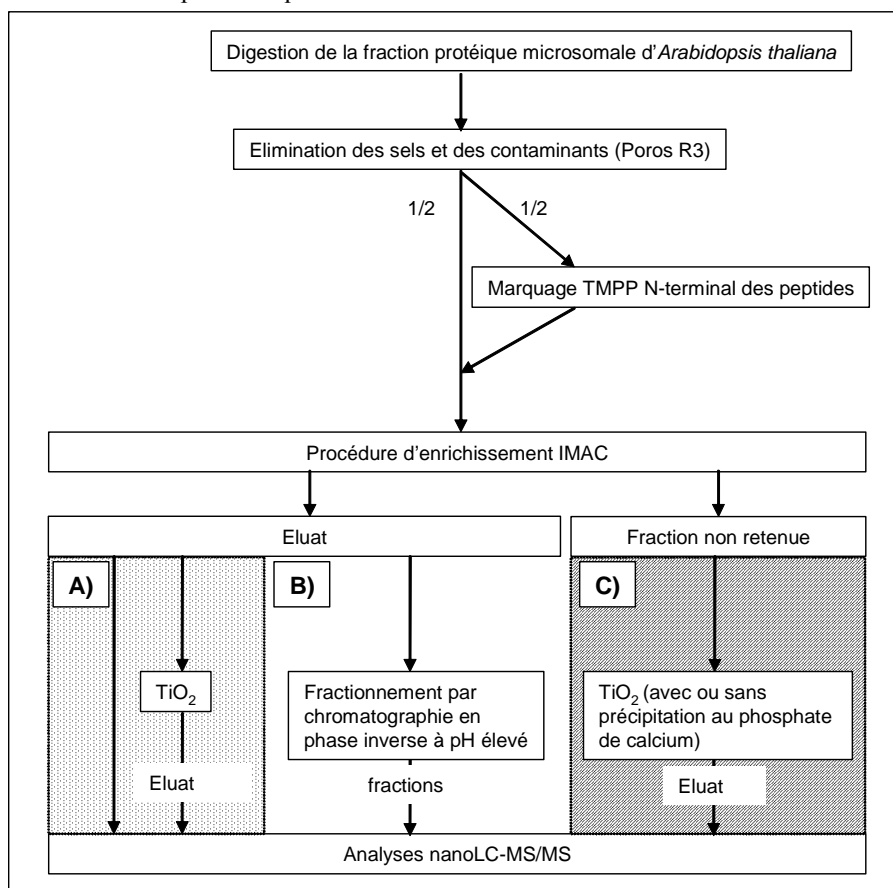


3. Développement de protocoles analytiques pour l'analyse phosphoprotéomique d'*A. thaliana*

L'utilisation du marquage TMPP a permis d'améliorer l'analyse phosphoprotéomique de protéines modèles grâce à des résultats complémentaires. Pour confirmer sur un système biologique plus complexe ces résultats prometteurs, nous avons dans un deuxième temps utilisé notre stratégie de marquage pour l'analyse phosphoprotéomique de fractions protéiques microsomaux d'*A. thaliana*. Cette préparation protéique est obtenue à partir de cellules de feuilles de plantes arrivées à maturité. Après broyage des tissus dans le tampon de lyse, les débris cellulaires et de la paroi sont éliminés par une première étape de filtration et de centrifugation (3000 g) et la fraction protéique microsomale est récupérée grâce à une deuxième étape de centrifugation (100000 g) du surnageant. Une étape finale de lavage des microsomes permet d'éliminer une grande partie des protéines solubles contaminant la préparation. On obtient au finale une fraction protéique enrichie en protéines membranaires ou associées à la membrane. Le détail de la préparation est décrit dans la partie expérimentale générale.

Pour obtenir la couverture de phosphoprotéome la plus complète, plusieurs protocoles analytiques ont été appliqués (Figure 4, A, B et C). En plus, pour limiter les problèmes de sous-échantillonnage, les expériences ont été répétées plusieurs fois sur les différents protocoles analytiques.

Figure 4 : Protocoles analytiques utilisés pour l'analyse phosphoprotéomique des fractions protéiques microsomaux d'*A. thaliana*



Comme point de départ, la moitié du digest dessalé de la préparation protéique microsomale a été soumis à la modification TMPP puis regroupé avec le reste du digest non modifié.

3.1. Protocole analytique A : IMAC et IMAC - TiO₂

Dans le premier protocole analytique (Figure 4 A), le mélange a été enrichi par chromatographie d'affinité IMAC fer ou par une combinaison des chromatographies d'affinité IMAC fer et TiO₂ puis analysé par nanoLC-MS/MS. Les données MS/MS issues de ces analyses ont été soumises à une interrogation Mascot dans une version target-decoy de la banque protéique d'*A. thaliana* et ont conduit à l'identification de 112 phosphopeptides natifs et de 61 phosphopeptides marqués au TMPP. Deux raisons semblent expliquer le plus faible nombre de phosphopeptides identifiés :

- Avec le gradient chromatographique optimisé utilisé pour ces analyses, les peptides non modifiés sont élués sur une plage temporelle de 30 minutes alors que les peptides modifiés au TMPP sont élués en 20 minutes, ce qui diminue le nombre de peptides échantillonnés par le spectromètre de masse en mode d'acquisition automatique.
- Des traces de TMPP restant dans l'échantillon malgré les chromatographies d'affinité interfèrent avec la détection par spectrométrie de masse des phosphopeptides marqués au TMPP. L'excès de réactif et de ses composés de dégradation, dont l'éluion chromatographique est réalisée dans la même zone que celle des phosphopeptides modifiés, est plus gênant dans le cas des échantillons biologiques car les peptides phosphorylés sont présents en plus faibles quantités.

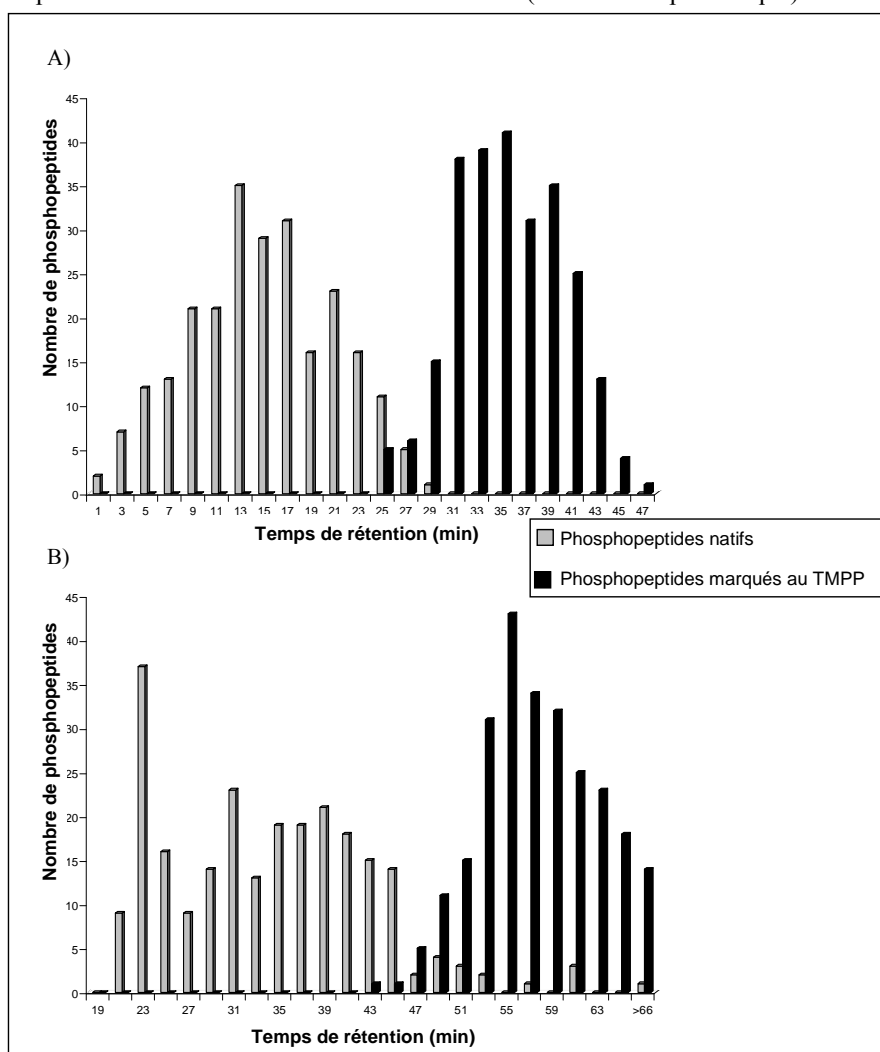
3.2. Protocole analytique B : IMAC – fractionnement HPLC en phase inverse à pH basique

Pour surmonter ces limitations, un deuxième protocole analytique employant la chromatographie à deux dimensions a été développé (Figure 4 B). L'ajout d'une première étape de chromatographie, généralement la chromatographie d'échange de cations forts (« Strong Cation Exchange », « SCX ») ou la chromatographie d'interaction hydrophile (Hydrophilic Interaction Chromatography », « HILIC »), est souvent utilisé pour fractionner les mélanges très complexes de phosphopeptides et obtenir une couverture plus complète du phosphoprotéome analysé [Beausoleil et al., 2004; Gruhler et al., 2005; Li et al., 2007; Albuquerque et al., 2008; McNulty et al., 2008]. Ici, comme première dimension de séparation, nous avons testé la chromatographie HPLC en phase inverse à pH élevé qui a déjà été appliquée avec succès pour l'analyse protéomique à grande échelle [Delmotte et al., 2007; Delmotte et al., 2009; Worner et al., 2009] mais qui n'a jamais été appliquée jusque-là pour l'analyse phosphoprotéomique.

Dans le protocole analytique développé ici (Figure 4 B), après enrichissement IMAC fer, le mélange peptidique a été séparé et collecté en 24 fractions par chromatographie HPLC en phase inverse à pH élevé. Les expériences ont été réalisées deux fois avec ce protocole et chaque fraction a ensuite été analysée par nanoLC-MS/MS. L'ensemble des analyses réalisées a permis d'identifier 496 peptides phosphorylés issus de 266 protéines. Le nombre de peptides phosphorylés identifiés a grandement été amélioré (environ 3 fois plus) grâce à l'utilisation de la chromatographie bi-dimensionnelle, particulièrement pour les phosphopeptides marqués au TMPP (environ 4 fois plus). Ce protocole a même permis de détecter plus de phosphopeptides sous forme marquée au TMPP (253 phosphopeptides) que sous forme native (243).

L'identification des phosphopeptides marqués au TMPP n'a pas été réalisée au détriment de l'identification des phosphopeptides natifs puisque les 2 classes de peptides sont séparées dans chacune des étapes de chromatographie en 2 zones d'élution présentant un chevauchement très limité comme illustré dans la Figure 5. Ainsi, avec les gradients chromatographiques utilisés dans l'étude, plus de 95 % des phosphopeptides natifs ont été élués avant 25 minutes en RP-HPLC à pH acide et avant 49 minutes en RP-HPLC à pH basique alors que 95 % des phosphopeptides marqués au TMPP ont été élués après 27 minutes en RP-HPLC à pH acide et après 47 minutes en RP-HPLC à pH basique. Le marquage TMPP permet donc de diminuer les phénomènes de sous-échantillonnage en donnant une seconde possibilité aux phosphopeptides d'être identifiés. De plus, les 2 classes de phosphopeptides ont fourni des résultats très complémentaires. En effet, sur le total des 509 sites de phosphorylation distincts identifiés, 179 sites ont été détectés uniquement grâce à un peptide marqué au TMPP (35 %), 171 sites uniquement grâce à un peptide natif (34 %) et seulement 159 sites grâce aux deux classes de peptides (31 %). Le nombre de sites de phosphorylation identifiés à la fois dans la première et la deuxième expérience avec le protocole B relativement plus important (44 %) que le nombre de sites identifiés par les 2 classes de phosphopeptides (31 %) suggère que l'augmentation de 55 % de la couverture de phosphoprotéome par le marquage TMPP ne peut pas être entièrement expliquée par les phénomènes de sous-échantillonnage. L'utilisation combinée de la RP-HPLC à pH basique et de la RP-HPLC à pH acide s'est avérée efficace pour l'analyse phosphoprotéomique des fractions protéomiques microsomaux d'*A. thaliana*, particulièrement avec l'approche par marquage chimique développé ici.

Figure 5 : Distribution des temps de rétention des phosphopeptides analysés avec le protocole analytique B. A) Temps de rétention dans les analyses nanoLC-MS/MS (RP-HPLC à pH acide). B) Temps de rétention lors de la collecte de fractions (RP-HPLC à pH basique).



3.3. Protocole analytique C : TiO_2 et Précipitation calcium - TiO_2

Finalement, pour atteindre une couverture du phosphoprotéome maximale, la fraction non retenue lors de la chromatographie d'affinité IMAC fer a été soumise à une nouvelle étape d'enrichissement (Figure 4 C). Si on considère que chaque méthode d'enrichissement permet d'isoler un ensemble différent de phosphopeptides [Bodenmiller et al., 2007], l'étape d'enrichissement par chromatographie d'affinité TiO_2 , combinée ou non avec la précipitation au phosphate de calcium, sur la fraction non retenue de l'enrichissement IMAC devrait permettre d'identifier des phosphopeptides supplémentaires. Les analyses nanoLC-MS/MS réalisées sur les éluats issus de la chromatographie d'affinité TiO_2 ont apporté un gain modéré sur la couverture du phosphoprotéome analysé avec

l'identification de 11 phosphopeptides natifs et 8 phosphopeptides marqués au TMPP qui n'avaient pu être détectés avec les autres protocoles analytiques.

Les résultats principaux obtenus avec les différents protocoles analytiques sur l'analyse phosphoprotéomique des fractions protéiques microsomales *d'A. thaliana* sont résumés dans le Tableau 3. Le protocole analytique B (Figure 4 B), utilisant la chromatographie bi-dimensionnelle sur les éluats issus de la chromatographie d'enrichissement IMAC fer, apparaît comme le plus efficace puisqu'il a fourni la majorité des identifications des phosphopeptides (~90 %). Les autres protocoles analytiques permettent cependant d'identifier des phosphopeptides supplémentaires et d'améliorer la couverture du phosphoprotéome.

Tableau 3 : Résultats principaux obtenus sur l'analyse phosphoprotéomique des fractions protéiques microsomales *d'A. thaliana*

		Catégorie de peptides	Fraction protéique microsomale <i>d'A. thaliana</i> sauvage				
			Protocole A	Protocole B	Protocole C	Total	
Nombre de phosphopeptides		Marqués au TMPP	61	253	28	269	
		Natifs	112	243	45	299	
		Total	173	496	73	568	
Sites de phosphorylation	Nombre	Phosphopeptides marqués au TMPP	32	179	19	165	
		Phosphopeptides natifs	Non-bloqués	89	135	26	168
			N-acétylés	16	36	5	45
			Total	105	171	31	213
		Communs au 2 catégories	56	159	11	187	
Total	193	509	61	565			

4. Apport du marquage TMPP pour la phosphoprotéomique comparative exploratoire d'*A. thaliana* de géotypes sauvage et mutant *hmgr1-1*

Pour apporter de nouvelles informations sur les mécanismes de régulation de la biosynthèse des stérols chez les plantes, nous avons réalisé l'analyse phosphoprotéomique du mutant *hmgr1-1 d'A. thaliana*. L'ensemble des protocoles analytiques utilisés dans le cadre de l'analyse du phosphoprotéome microsomal *d'A. thaliana* sauvage (Figure 4) ont donc également été appliqués au mutant.

4.1. Bilan des identifications

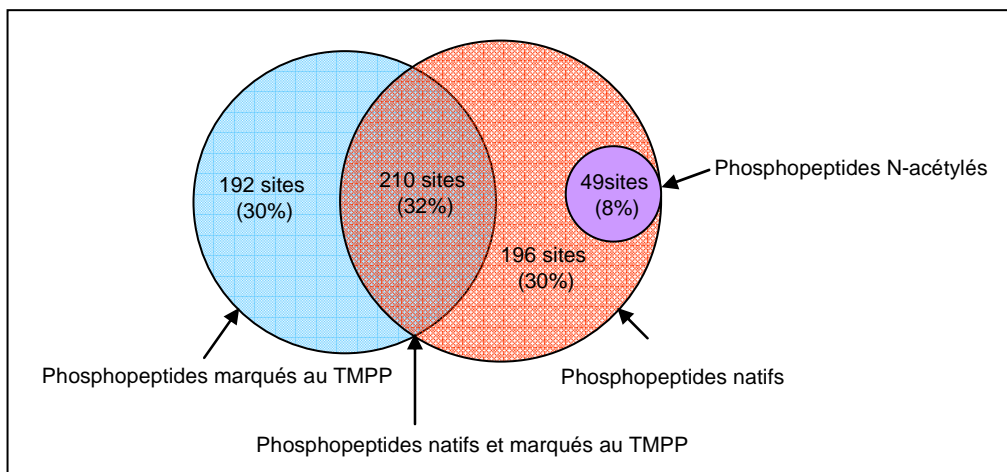
En combinant l'ensemble des résultats obtenus dans toute cette étude, 334 phosphopeptides natifs et 297 phosphopeptides marqués au TMPP ont pu être identifiés (Tableau 4). Les 647 sites de phosphorylations identifiés sur 323 protéines sont distribués comme suit : 192 sites ont été détectés uniquement grâce à un peptide marqué au TMPP (30 %), 245 uniquement grâce à un peptide natif (38

%) et 210 grâce aux deux catégories de peptides (32 %). Cette distribution atteste que le marquage TMPP a permis d'augmenter la couverture du phosphoprotéome analysé de 40 %. Il est également intéressant d'observer que 20 % des sites de phosphorylation identifiés uniquement grâce à un peptide natif sont localisés sur des peptides N-terminaux acétylés (49 sites) qui n'ont donc pu être marqués chimiquement (Figure 6).

Tableau 4 : Bilan des identifications des phosphopeptides obtenues sur l'analyse des fractions protéiques microsomales d'*A. thaliana* sauvage et mutant *hmgr1-1*.

		Catégorie de peptides	Résultats totaux (<i>A. thaliana</i> sauvage et <i>hmgr1-1</i>)	
			Total	
Nombre de phosphopeptides	Marqués au TMPP		297	
	Natifs		334	
	Total		631	
Sites de phosphorylation	Nombre	Phosphopeptides marqués au TMPP		192
		Phosphopeptides natifs	Non-bloqués	196
			N-acétylés	49
			Total	245
		Communs au 2 catégories		210
	Total		647	
	Localisation exacte		121	
		88,7 %		
		11 %		
		0,3 %		
		526		

Figure 6 : Distribution des sites de phosphorylation identifiés dans toute l'étude en fonction de la nature des peptides qui ont permis leur identification



La localisation exacte des sites de phosphorylation a pu être établie pour 81 % des sites (526 sur les 647 sites). Nous avons estimé que 88,7 % des phosphorylations étaient localisées sur une sérine, 11 % sur une thréonine et 0,3 % sur une Tyrosine. Cette distribution est identique à celle rapportée par l'étude de [Reiland et al., 2009] dans laquelle l'analyse phosphoprotéomique a également porté sur des tissus de plantes différenciés.

4.2. Une « sous-population » de phosphopeptides révélée par le marquage TMPP

4.2.1. Comparaison des sites de phosphorylation identifiés avec la base de données « PhosPhAt »

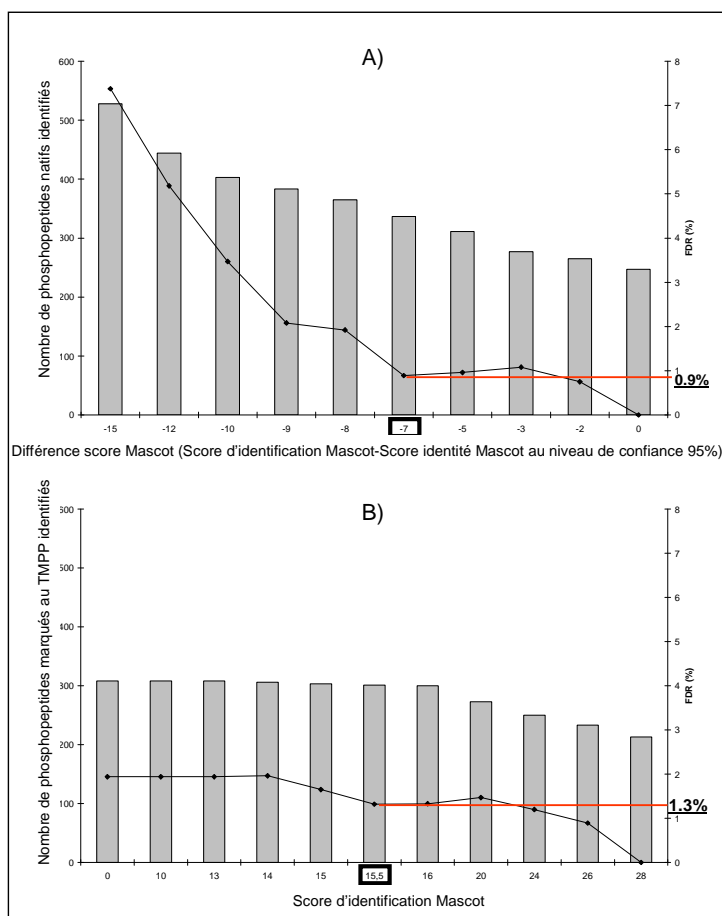
Les analyses réalisées dans notre étude ont conduit à l'identification d'un grand nombre de phosphopeptides. Toutefois, certaines études phosphoprotéomiques récentes sur *A. thaliana* ont permis d'identifier d'avantage de phosphopeptides (par exemple 3029 phosphopeptides dans l'étude de [Reiland et al., 2009] ou 2597 phosphopeptides dans l'étude de [Sugiyama et al., 2008]) mais ces études ont été réalisées sur des protéomes totaux de cultures cellulaires ou de plantes entières. Ici, nous avons ciblé la fraction protéique microsomale, c'est-à-dire enrichie en protéines membranaires ou associées à la membrane, obtenue uniquement à partir des feuilles de la plante. L'élimination d'une grande partie des protéines solubles doit donc diminuer le nombre de phosphopeptides identifiés mais permet aussi de simplifier le mélange pour une identification plus aisée des protéines d'intérêt pour l'étude. Nous avons comparé les sites de phosphorylation identifiés dans cette étude avec la base de données « PhosPhAt » [Heazlewood et al., 2008] qui répertorie les sites de phosphorylation identifiés dans le protéome d'*A. thaliana* à partir de plusieurs études (plus de 6000 phosphopeptides issus de 10 études enregistrés en juillet 2009). Nous avons constaté que 50 % des sites de phosphorylation identifiés dans notre étude (323 sites sur les 647) n'étaient pas répertoriés dans la base de données, ce qui atteste que notre mode opératoire permet d'accéder à des peptides phosphorylés non détectés avec des analyses plus globales. Si on porte notre attention uniquement sur les phosphopeptides marqués au TMPP, la proportion des sites de phosphorylation non répertoriés dans la base de données atteint 60.5 % (116 sites sur les 192). Ces données confirment que le marquage TMPP permet d'accéder à une population de phosphopeptides « ignorée » par les méthodes d'analyses habituelles.

4.2.2. Impact du marquage TMPP sur l'identification des phosphopeptides par les moteurs de recherche

L'effet du marquage TMPP sur la fragmentation des phosphopeptides a été décrit dans le paragraphe 2.2. de ce chapitre traitant des protéines modèles. Dans une étude précédente [Gallien et al., 2009], nous avons observé qu'en dépit de la qualité des spectres de fragmentation, les scores d'identification Mascot étaient généralement plus faibles pour les peptides marqués au TMPP que pour les peptides natifs. Les seuils de significativité sur les scores d'identification calculés *in silico* par Mascot n'étaient donc pas adaptés aux peptides marqués au TMPP analysés par une trappe d'ions. Cependant, nous avons aussi observé qu'une stratégie d'évaluation des identifications par approche target-decoy restait applicable pour contrôler le FDR à partir du moment où les seuils étaient adaptés

aux peptides marqués au TMPP. Ainsi, les seuils de scores Mascot pour l'identification des phosphopeptides natifs et des phosphopeptides marqués au TMPP ont été fixés indépendamment pour obtenir un FDR voisin de 1 % sur chacune des 2 catégories de peptides (Figure 7). Dans notre étude, pour être « validée », l'identification d'un spectre correspondant à la séquence d'un phosphopeptide natif a du être réalisée avec un score Mascot inférieur de 7 points au maximum au seuil identité et l'identification d'un spectre correspondant à la séquence d'un phosphopeptides marqué au TMPP a du être réalisée avec un score Mascot minimum de 15.5.

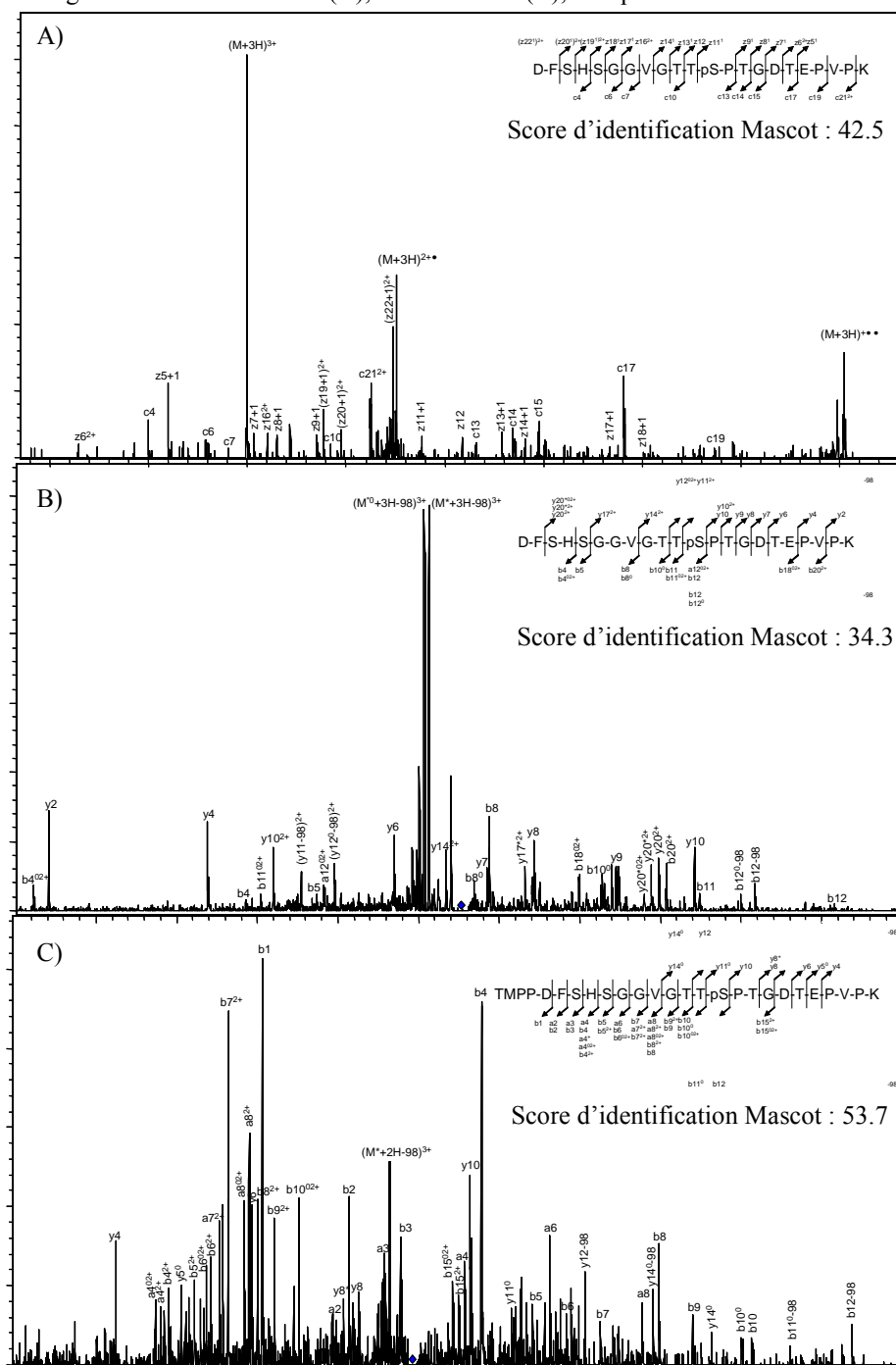
Figure 7 : Evolution du FDR et du nombre de phosphopeptides natifs (A) et marqués au TMPP (B) en fonction des seuils de scores d'identification Mascot



Nous avons décidé d'examiner la possibilité d'identifier des phosphopeptides supplémentaires grâce à la modification de fragmentation induite par le marquage TMPP. Pour cela, nous avons réalisé des analyses nanoLC-MS/MS sur les éluats issus de la chromatographie d'affinité IMAC fer des digests des fractions protéiques microsomales (protocole analytique A, Figure 4 A) en utilisant un mode de fragmentation ETD à la place du mode fragmentation CID. Le mode de fragmentation ETD a été décrit comme particulièrement utile pour l'analyse de peptides porteurs de modifications post-traductionnelles labiles et permettant d'accéder à une bonne couverture de séquence [Syka et al., 2004].

Le mode de fragmentation ETD a permis d'identifier 46 phosphopeptides natifs parmi lesquels 5 n'avaient pas été identifiés en mode CID malgré l'utilisation des 3 protocoles analytiques (Figure 4). Il est intéressant d'observer que 4 phosphopeptides sur les 5 avaient toutefois été identifiés en mode CID mais sous forme marquée au TMPP. Ainsi, malgré un volume de données issues des analyses ETD limité, les résultats suggèrent que, de manière similaire à la fragmentation ETD, le marquage TMPP conduit pour certains phosphopeptides à des spectres de fragmentation plus informatifs. Par exemple, sur la Figure 8 est réalisée la comparaison des spectres MS/MS du phosphopeptide DFSSGGVGTTpSPTGDTEPVPK triplement chargé obtenu en mode ETD (Figure 8 A), obtenu en mode CID mais ne passant le seuil de score Mascot (Figure 8 B) et de son analogue marqué au TMPP triplement chargé obtenu en mode CID (Figure 8 C). Nous avons observé que le spectre CID du phosphopeptide marqué au TMPP et que le spectre ETD du phosphopeptide natif permettaient d'obtenir les plus grands nombres de fragments interprétables (respectivement 17 liaisons peptidiques et 21 liaisons peptidiques fragmentées). Le spectre CID du phosphopeptide natif a été moins informatif car nous avons observé seulement 13 liaisons peptidiques fragmentées et des pics correspondant à des pertes de neutre prédominants.

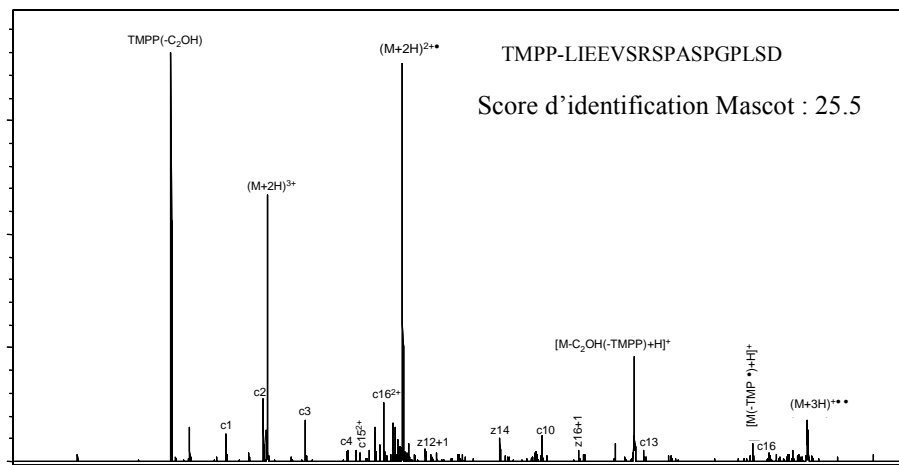
Figure 8 : Comparaison des spectres MS/MS du phosphopeptide DFSHSGGVGTTpSPTGDTEPVPK triplement chargé obtenu en mode ETD (A), en mode CID (B), marqué au TMPP obtenu en mode CID (C)



L'analyse ETD a aussi permis d'identifier 5 phosphopeptides marqués au TMPP pour lesquels les spectres de fragmentation étaient fortement dominés par des pics correspondant à la perte du radical triméthoxyphényle et au radical triméthoxyphényle lui-même en plus des pics correspondant aux états de charge réduit du précurseur (Spectre ETD du peptide TMPP-LIEEVSRSPASGPLSD triplement chargé en Figure 9). Une étude précédente sur quelques phosphopeptides modèles traitant de la fragmentation induite par le marquage TMPP en mode ECD a mentionné ce phénomène de fragmentation concurrentielle se déroulant sur le groupement phosphonium [Chamot-Rooke et al.,

2007]. Dans notre étude, ces fragments indésirables et très intenses ont conduit à des spectres de fragmentation moins informatifs sur la séquence du peptide et ont probablement pénalisé les identifications de Mascot, ce qui explique probablement le très faible nombre de phosphopeptides marqués au TMPP identifiés en mode ETD.

Figure 9 : Spectre ETD du peptide TMPP-LIEEVSRSPASGPLSD triplement chargé



4.2.3. Comparaison des propriétés physicochimiques des phosphopeptides identifiés sous forme native et marquée au TMPP

Nous avons réalisé l'analyse des propriétés physico-chimiques des phosphopeptides identifiés afin de déterminer de potentielles relations entre ces propriétés et la détection des phosphopeptides sous forme marquée ou native.

4.2.3.1. Masse moléculaire

Tout d'abord, nous avons examiné la distribution des masses moléculaires des phosphopeptides et nous n'avons pas observé de corrélation entre les masses moléculaires et l'identification préférentielle des phosphopeptides sous leur forme native ou sous leur forme marquée au TMPP (Figure 10 A).

4.2.3.2. Etats de charge des ions précurseurs

Nous avons ensuite observé la distribution des états de charge de tous les précurseurs pour lesquelles l'assignation des spectres MS/MS ont été « validés » avec les seuils établis. Nous avons pu noter que les phosphopeptides natifs étaient le plus fréquemment identifiés grâce à des spectres

MS/MS de précurseurs doublement chargés (~2/3 des phosphopeptides natifs) alors que la grande majorité des phosphopeptides marqués au TMPP ont été identifiés grâce à des spectres MS/MS de précurseurs triplement chargés (~3/4 des phosphopeptides marqués au TMPP) (Figure 10 B). Ces résultats confirment la détection préférentielle des précurseurs présentant de plus hauts états de charge, et conduisant donc à des spectres plus informatifs (vu dans le paragraphe 2.2. de ce chapitre), pour les phosphopeptides marqués au TMPP.

4.2.3.3. Composition en acides aminés, nombre de phosphorylations et charge « globale »

Nous avons également comparé la composition relative en acides aminés des 2 classes de phosphopeptides (Figure 10 C) et nous avons observé un biais pour :

- ✓ Les acides aminés acides, particulièrement pour l'acide glutamique, pour les phosphopeptides marqués au TMPP identifiés.
- ✓ Les acides aminés basiques pour les phosphopeptides natifs.

Ces acides aminés portent respectivement une charge positive et une charge négative, ce qui suggère que la propriété de charge joue un rôle important dans l'efficacité d'ionisation électrospray et donc dans la détectabilité des phosphopeptides.

Nous avons aussi pu noter une plus grande tendance à détecter les peptides multi-phosphorylés sous forme marquée (51 %) plutôt que sous forme native (45 %). Cette sur-représentation des peptides multi-phosphorylés est encore plus prononcée pour les phosphopeptides identifiés uniquement sous forme marquée (57 %) (Figure 10 D).

La comparaison de la distribution des « charges globales » des 2 classes de phosphopeptides (les acides aminés K, R et H portant 1 charge positive et les acides aminés D, E et les groupements phosphate portant 1 charge négative [Fauchere et al., 1988]) montre clairement que le marquage TMPP permet d'améliorer l'identification des phosphopeptides avec une « charge globale » faible au détriment des phosphopeptides avec une « charge globale » élevée (Figure 10 E). La capacité du groupement TMPP chargé positivement à compenser la diminution de « charge globale » causée par le grand nombre de groupements phosphate et d'acides aminés acides mentionné dans l'étude des phosphoprotéines modèle (paragraphe 2.1. de ce chapitre) est confirmé ici. Ces résultats soulignent la difficulté à détecter en mode électrospray les peptides de « charge globale » faible comme les peptides très acides et/ou phosphorylés. Au contraire, l'identification des phosphopeptides marqués au TMPP comprenant un nombre important d'acides aminés basiques, par exemple des peptides comportant des coupures enzymatiques manquées, est plutôt défavorable, notamment à cause de l'efficacité de marquage relativement limitée pour ce genre de peptides [Huang et al., 1999].

En plus, nous avons aussi observé une sur-représentation des acides aminés proline dans les séquences des phosphopeptides identifiés sous forme marquée au TMPP en comparaison des séquences des phosphopeptides natifs. La présence de cet acide aminé dans un peptide conduit souvent à des spectres CID peu informatifs à cause des coupures préférentielles induites [Huang et al., 2005].

Ici, nous postulons que le marquage TMPP peut améliorer la fragmentation des phosphopeptides contenant des prolines bien que les aspects mécanistiques impliqués restent incertains [Sadagopan et al., 2000].

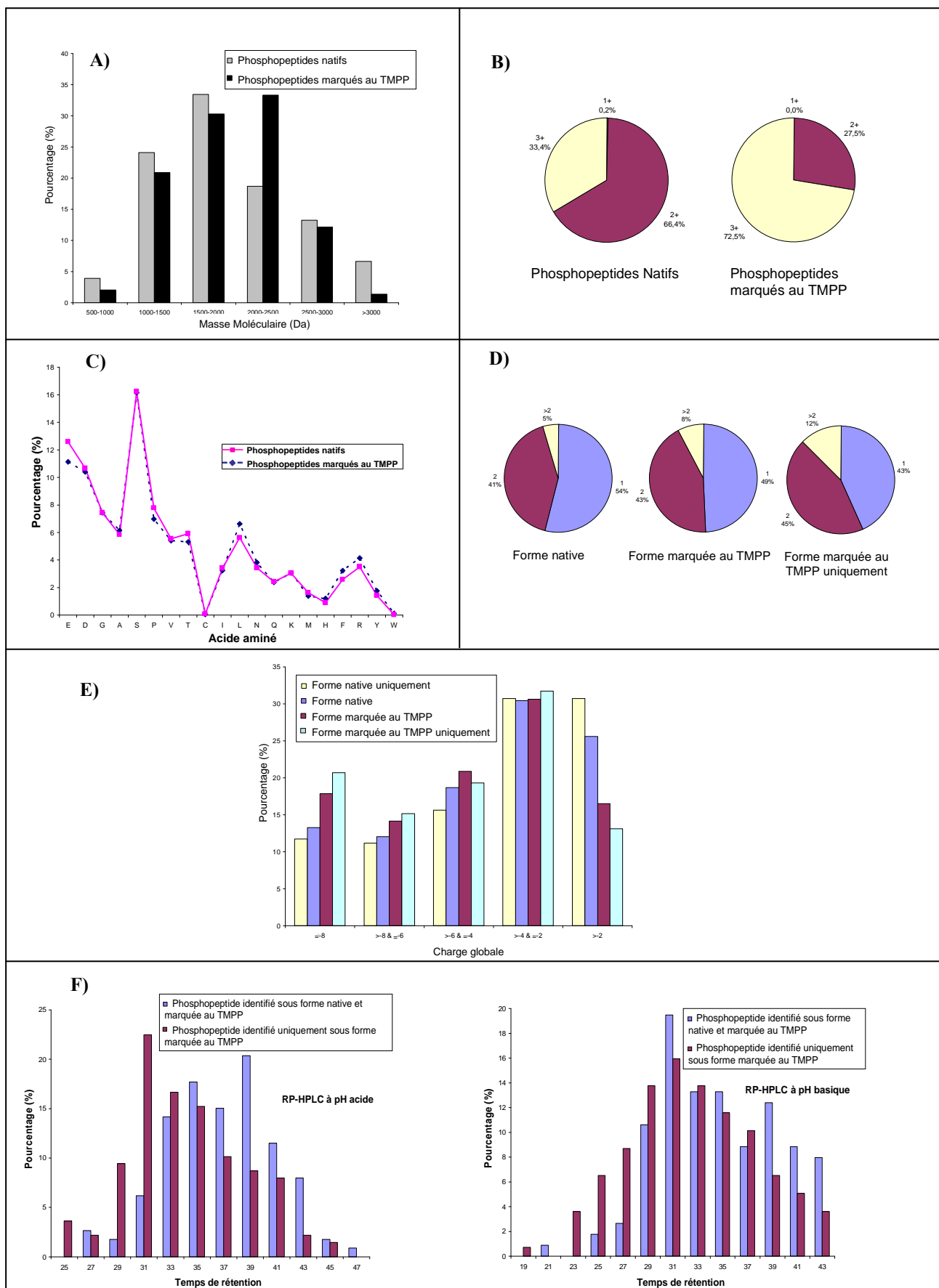
Enfin, la comparaison relative en acides aminés des 2 catégories de phosphopeptides a révélé un biais pour les deux acides aminés les plus hydrophobes, la phénylalanine et la leucine [Wilce et al., 1995], pour les phosphopeptides natifs identifiés en comparaison des phosphopeptides marqués au TMPP.

4.2.3.4. Hydrophilicité

Pour examiner d'avantage l'influence de l'hydrophilicité sur la détection des phosphopeptides, nous avons comparé la distribution des temps de rétention en RP-HPLC à pH acide et basique des phosphopeptides identifiés uniquement sous forme marquée avec celle des phosphopeptides identifiés sous les deux formes (Figure 10 F). Nous avons observé que les phosphopeptides identifiés uniquement sous forme marquée étaient généralement plus hydrophiles que les phosphopeptides identifiés sous les deux formes. Etant donné que de manière générale les peptides les plus hydrophiles sous forme native restent les peptides les plus hydrophiles sous forme marquée, il semblerait donc que le marquage TMPP améliore l'identification des peptides les plus hydrophiles notamment en limitant leur perte lors de leur chargement sur une colonne chromatographique de phase inverse.

La comparaison des propriétés physicochimiques des phosphopeptides identifiés sous forme native ou marquée au TMPP a permis de mettre en évidence que le marquage TMPP permet d'améliorer l'identification des phosphopeptides les plus acides, hydrophiles et multi-phosphorylés. Ce type de peptide apparaît donc comme très difficile à détecter dans les analyses phosphoprotéomiques « classiques ».

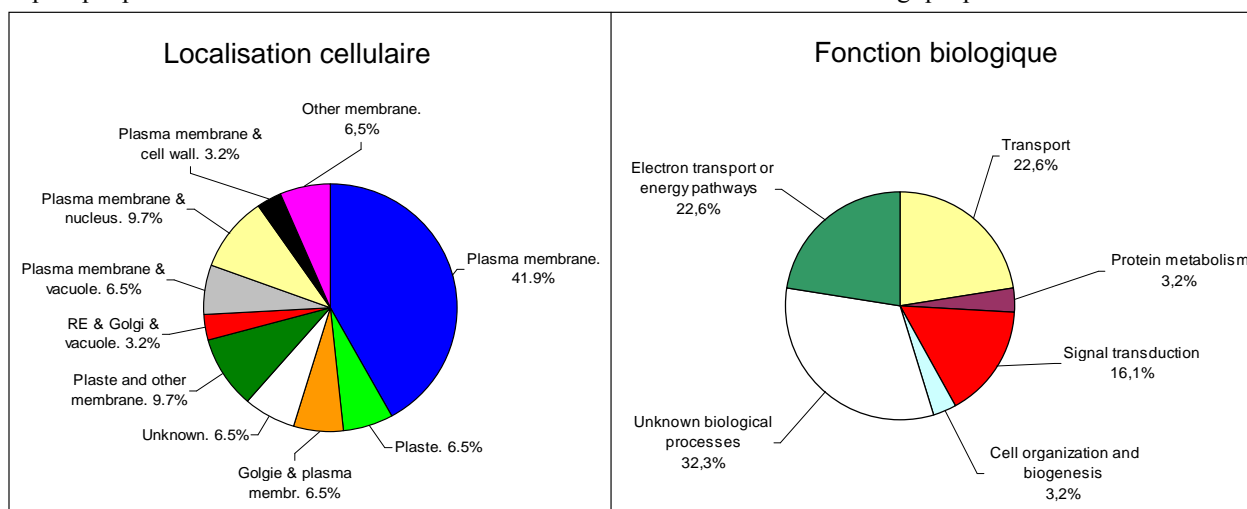
Figure 10 : Comparaison des propriétés physicochimiques des phosphopeptides identifiés sous forme native et marquée au TMPP. A) Masse moléculaire, B) état de charge des ions précurseurs, C) composition en acides aminés, D) nombre de phosphorylations portées, E) charge « globale », F) hydrophilicité.



5. Signification biologique des différences observées entre les deux phosphoprotéomes

Les résultats obtenus de l'analyse phosphoprotéomique du mutant *Arabidopsis thaliana hmgr1-1* ont été comparés à ceux obtenus de l'analyse du sauvage de même fond génétique afin d'obtenir un premier aperçu des différences entre les deux phosphoprotéomes. La détermination des niveaux d'expression relatifs des phosphopeptides entre le sauvage et le mutant ressort d'une approche par « spectral counting » (comptage des spectres) qui est décrite en détails dans la Partie III des résultats. Brièvement, cette approche est basée sur l'observation empirique que le taux d'échantillonnage MS/MS d'un peptide donné est directement relié à l'abondance de ce peptide dans l'échantillon [Gilchrist et al., 2006; Wang et al., 2008; Heintz et al., 2009]. L'approche par « spectral counting » est préférentiellement appliquée au niveau protéique afin de comptabiliser, pour chaque protéine, un grand nombre de spectres résultant de l'identification de plusieurs peptides. Cependant, dans notre étude, l'utilisation de plusieurs protocoles analytiques, la réalisation de plusieurs expériences avec ces protocoles et la possibilité d'additionner les spectres des phosphopeptides natifs et marqués au TMPP a permis d'augmenter le nombre de spectres considérés pour l'analyse différentielle. En moyenne, plus de 12 spectres ont pu être comptabilisés par phosphopeptide. Bien que ce mode de quantification relative doive être considéré avec précaution au niveau peptidique, il permet néanmoins de montrer les différences les plus significatives entre les deux phosphoprotéomes considérés. Après normalisation du nombre de spectres attribués à chaque peptide (en se basant sur le nombre total de spectres attribués à tous les phosphopeptides de chacun des échantillons), un test de proportion (Z-test) a permis de mettre en évidence 36 phosphopeptides (issus de 31 protéines) présentant des différences d'abondance (au risque $p < 0.05$). Nous avons utilisé l'annotation TAIR (« The Arabidopsis Information Resource, <http://www.arabidopsis.org/tools/bulk/go/>) pour classer ces 31 phosphoprotéines selon leur localisation cellulaire prédite en 10 groupes et selon leur fonction biologique prédite en 6 groupes (Figure 11).

Figure 11 : Classification des 31 protéines présentant des phosphopeptides d'abondance altérée entre les deux phosphoprotéomes en fonction de leur localisation cellulaire et leur fonction biologique prédites



Les interprétations biologiques des différences observées sur l'abondance des phosphopeptides entre l'extrait protéique microsomal d'*A. thaliana* sauvage et mutant *hmgr1-1* sont en cours. Toutefois quelques observations et hypothèses biologiques peuvent déjà être proposées.

5.1. Localisation cellulaire prédite des phosphoprotéines d'abondance différente

Les protéines pour lesquelles l'abondance des phosphopeptides varie sont majoritairement localisées dans la membrane plasmique (41.9 %). Ce résultat n'est pas surprenant puisque la membrane plasmique est un site majeur d'accumulation de stérols. On peut donc envisager l'hypothèse selon laquelle la mutation *hmgr1-1* a une incidence sur l'état de phosphorylation des protéines de la membrane plasmique impliquées dans les processus de transport ou de transduction des signaux décrits dans la Figure 11.

Un autre groupe important de protéines est localisé dans le plaste et d'autres membranes non identifiées précisément (9.7 %) ou uniquement dans le plaste (6.5 %). Ce résultat montre que la mutation d'un gène qui code pour une enzyme clé de la voie cytosolique du MVA a une incidence sur le phosphoprotéome du plaste. Ce lien entre plaste et la voie du MVA a été démontré à plusieurs reprises [Hemmerlin et al., 2003; Laule et al., 2003]. En effet, il a été décrit que lorsque la voie du MVA est altérée, il peut exister une compensation par la voie du MEP qui se trouve dans le chloroplaste et qui fournit l'IPP déficient dans la voie du MVA (mécanisme « cross-talk »). On peut donc envisager ici l'hypothèse que la mutation *hmgr1-1* de l'enzyme clé de la voie du MVA induise une participation accrue du plaste dans le métabolisme de plantes portant cette mutation *hmgr1-1*, par exemple par l'intervention de la voie du MEP pour la synthèse de stérols (à démontrer) ; participation dont le relai serait assuré par des phosphoprotéines impliquées dans la transduction de signaux.

Il est intéressant de noter que parmi ces protéines, aucune n'est strictement soluble, toutes ont au moins une localisation prédite dans un compartiment membranaire.

5.2. Fonction biologique prédite des phosphoprotéines d'abondance différente

Les protéines pour lesquelles l'abondance des phosphopeptides varie sont majoritairement des protéines de fonction inconnue (32.3 %).

Une grande proportion des protéines de fonction connue sont impliquées dans des mécanismes de transport (22.6 %). On peut notamment mentionner l'identification de 3 aquaporines, protéines membranaires qui jouent un rôle dans le maintien de la pression osmotique cellulaire en cas de stress [Javot et al., 2003; Nuhse et al., 2004; Hem et al., 2007]. On peut envisager ici l'hypothèse que la perturbation de la synthèse des stérols, et plus précisément la réduction de la quantité de stérols mesurée chez les plantes *hmgr1-1* [Suzuki et al., 2004], a une incidence sur les propriétés physicochimiques des membranes cellulaires régulées par les stérols, et que ceci entraîne un stress

osmotique que tente de juguler la cellule en activant les aquaporines. L'identification de phosphoprotéines d'abondance différente impliquées dans la synthèse de la paroi comme la protéine « Cellulose synthase A catalytic subunit 1 [UDP-forming] » vient appuyer cette hypothèse. En effet, le rôle de cette protéine est de synthétiser la cellulose pour rigidifier la paroi [Peng et al., 2000]. On peut donc imaginer que sous l'effet d'un stress osmotique du à la déficience en stérols, il ya modification de la paroi pour rigidifier l'architecture cellulaire.

Un autre groupe important est celui des phosphoprotéines impliquées dans la production d'énergie (22.6 %). Parmi ces protéines, on peut noter la présence de 4 ATPases membranaires, protéines responsables de la synthèse et du transport d'ATP de part et d'autre de la membrane, qui son souvent induites (ou dont la phosphorylation est induite) en réponse à des stress [Nuhse et al., 2004; Benschop et al., 2007; Niittyla et al., 2007; Nuhse et al., 2007].

Enfin, plusieurs kinases, récepteurs kinase et phosphatases présentant des phosphopeptides d'abondance altérée ont été identifiées et pourraient être des candidats potentiels intervenant dans les mécanismes de transduction de signaux dans le métabolisme des stérols de plante au cours de son développement.

Cette étude différentielle du phosphoprotéome d'*Arabidopsis thaliana* sauvage et mutant *hmgr1-1* a permis de proposer des interprétations biologiques préliminaires à la base de l'élaboration d'expériences biologiques complémentaires qui viendront étayer ces hypothèses.

Le détail de l'ensemble des protocoles expérimentaux (préparation des fractions protéiques microsomales, marquage TMPP, enrichissement IMAC, TiO₂, précipitation au phosphate de calcium, analyses nanoLC-MS/MS, fractionnement RP-HPLC à pH basique et identification des phosphopeptides) est décrit dans la partie expérimentale générale.

Conclusion

Cette deuxième partie des résultats a été consacrée au développement d'une méthode de dérivation chimique des phosphopeptides au TMPP permettant l'amélioration de l'étude par spectrométrie de masse des phosphorylations des protéines. Cette méthode a ensuite été appliquée dans le cadre d'une étude phosphoprotéomique sur des échantillons protéiques relativement complexes d'*Arabidopsis thaliana*. En tirant partie des propriétés des phosphopeptides marqués au TMPP (efficacité d'ionisation améliorée, modification de la fragmentation CID, augmentation de l'hydrophobicité en chromatographie en phase inverse), la méthode a permis d'améliorer d'environ 50 % la couverture du phosphoprotéome d'*A. thaliana*. Les résultats obtenus ont été complémentaires à ceux obtenus avec des conditions d'analyses classiques sans marquage. En effet, même si l'intégralité des phosphopeptides détectés sans marquage n'a pas été identifiée par l'approche TMPP, une « sous-population » de phosphopeptides a pu être révélée grâce au marquage. Cette étude sur un échantillon protéique complexe a ainsi permis de mettre en évidence des corrélations entre la nature physico-chimique des phosphopeptides et leur détection sous forme marquée TMPP ou native et donc de mieux appréhender les difficultés relatives aux analyses des phosphorylations. Comme notre stratégie consiste en une simple étape supplémentaire à insérer dès les premières étapes d'un protocole « standard » d'analyse phosphoprotéomique, elle ne conduit donc pas à la réalisation d'analyses LC-MS/MS additionnelles puisque les deux classes de phosphopeptides (marqués et non marqués) sont analysés dans le même « run » chromatographique sans interférence mutuelle. En ce sens, elle pourrait être utilisée assez largement dans le domaine de la phosphoprotéomique. Une publication rapportant ces résultats est actuellement en cours de rédaction.

La comparaison des phosphoprotéomes d'*A. thaliana* sauvage et du mutant *hmgr1-1* a aussi permis d'obtenir un premier aperçu de leurs différences et de formuler quelques hypothèses biologiques préliminaires sur certains mécanismes impliqués dans le métabolisme des stérols chez les plantes.

Bibliographie

- Adamczyk, M., J. C. Gebler and J. Wu**
"Charge derivatization of peptides to simplify their sequencing with an ion trap mass spectrometer." *Rapid Commun Mass Spectrom*, **1999**, 13 (14), 1413-22.
- Albuquerque, C. P., M. B. Smolka, S. H. Payne, V. Bafna, J. Eng and H. Zhou**
"A multidimensional chromatography technology for in-depth phosphoproteome analysis." *Mol Cell Proteomics*, **2008**, 7 (7), 1389-96.
- Alonso, A., J. Sasin, N. Bottini, I. Friedberg, I. Friedberg, A. Osterman, A. Godzik, T. Hunter, J. Dixon and T. Mustelin**
"Protein tyrosine phosphatases in the human genome." *Cell*, **2004**, 117 (6), 699-711.
- Andersson, L. and J. Porath**
"Isolation of phosphoproteins by immobilized metal (Fe³⁺) affinity chromatography." *Anal Biochem*, **1986**, 154 (1), 250-4.
- Arena, S., S. Benvenuti and A. Bardelli**
"Genetic analysis of the kinome and phosphatome in cancer." *Cell Mol Life Sci*, **2005**, 62 (18), 2092-9.
- Bach, T. J.**
"Hydroxymethylglutaryl-CoA reductase, a key enzyme in phytosterol synthesis?" *Lipids*, **1986**, 21 (1), 82-8.
- Beausoleil, S. A., M. Jedrychowski, D. Schwartz, J. E. Elias, J. Villen, J. Li, M. A. Cohn, L. C. Cantley and S. P. Gygi**
"Large-scale characterization of HeLa cell nuclear phosphoproteins." *Proc Natl Acad Sci U S A*, **2004**, 101 (33), 12130-5.
- Benschop, J. J., S. Mohammed, M. O'Flaherty, A. J. Heck, M. Slijper and F. L. Menke**
"Quantitative phosphoproteomics of early elicitor signaling in Arabidopsis." *Mol Cell Proteomics*, **2007**, 6 (7), 1198-214.
- Blume-Jensen, P. and T. Hunter**
"Oncogenic kinase signalling." *Nature*, **2001**, 411 (6835), 355-65.
- Bodenmiller, B., L. N. Mueller, M. Mueller, B. Domon and R. Aebersold**
"Reproducible isolation of distinct, overlapping segments of the phosphoproteome." *Nat Methods*, **2007**, 4 (3), 231-7.
- Bodenmiller, B., L. N. Mueller, P. G. Pedrioli, D. Pflieger, M. A. Junger, J. K. Eng, R. Aebersold and W. A. Tao**
"An integrated chemical, mass spectrometric and computational strategy for (quantitative) phosphoproteomics: application to Drosophila melanogaster Kc167 cells." *Mol Biosyst*, **2007**, 3 (4), 275-86.
- Boersema, P. J., S. Mohammed and A. J. Heck**
"Phosphopeptide fragmentation and analysis by mass spectrometry." *J Mass Spectrom*, **2009**, 44 (6), 861-78.
- Bouwmeester, H. J.**
"Engineering the essence of plants." *Nat Biotechnol*, **2006**, 24 (11), 1359-61.
- Chalmers, M. J., K. Hakansson, R. Johnson, R. Smith, J. Shen, M. R. Emmett and A. G. Marshall**
"Protein kinase A phosphorylation characterized by tandem Fourier transform ion cyclotron resonance mass spectrometry." *Proteomics*, **2004**, 4 (4), 970-81.
- Chamot-Rooke, J., G. van der Rest, A. Dalleu, S. Bay and J. Lemoine**
"The combination of electron capture dissociation and fixed charge derivatization increases sequence coverage for o-glycosylated and o-phosphorylated peptides." *J Am Soc Mass Spectrom*, **2007**, 18 (8), 1405-13.
- Chappell, J., F. Wolf, J. Proulx, R. Cuellar and C. Saunders**
"Is the Reaction Catalyzed by 3-Hydroxy-3-Methylglutaryl Coenzyme A Reductase a Rate-Limiting Step for Isoprenoid Biosynthesis in Plants?" *Plant Physiol*, **1995**, 109 (4), 1337-1343.
- Cohen, P.**
"The origins of protein phosphorylation." *Nat Cell Biol*, **2002**, 4 (5), E127-30.
- Cohen, P.**
"Protein kinases--the major drug targets of the twenty-first century?" *Nat Rev Drug Discov*, **2002**, 1 (4), 309-15.
- Corthals, G. L., R. Aebersold and D. R. Goodlett**
"Identification of phosphorylation sites using microimmobilized metal affinity chromatography." *Methods Enzymol*, **2005**, 405 66-81.
- Daviet, L. and F. Colland**
"Targeting ubiquitin specific proteases for drug discovery." *Biochimie*, **2008**, 90 (2), 270-83.
- Delmotte, N., M. Lasasa, A. Tholey, E. Heinzle and C. G. Huber**
"Two-dimensional reversed-phase x ion-pair reversed-phase HPLC: an alternative approach to high-resolution peptide separation for shotgun proteome analysis." *J Proteome Res*, **2007**, 6 (11), 4363-73.

Delmotte, N., M. Lasaosa, A. Tholey, E. Heinzle, A. van Dorsselaer and C. G. Huber
 "Repeatability of peptide identifications in shotgun proteome analysis employing off-line two-dimensional chromatographic separations and ion-trap MS." *J Sep Sci*, **2009**, 32 (8), 1156-64.

Enjuto, M., L. Balcells, N. Campos, C. Caelles, M. Arro and A. Boronat
 "Arabidopsis thaliana contains two differentially expressed 3-hydroxy-3-methylglutaryl-CoA reductase genes, which encode microsomal forms of the enzyme." *Proc Natl Acad Sci U S A*, **1994**, 91 (3), 927-31.

Enjuto, M., V. Lumbreras, C. Marin and A. Boronat
 "Expression of the Arabidopsis HMG2 gene, encoding 3-hydroxy-3-methylglutaryl coenzyme A reductase, is restricted to meristematic and floral tissues." *Plant Cell*, **1995**, 7 (5), 517-27.

Fauchere, J. L., M. Charton, L. B. Kier, A. Verloop and V. Pliska
 "Amino acid side chain parameters for correlation studies in biology and pharmacology." *Int J Pept Protein Res*, **1988**, 32 (4), 269-78.

Ficarro, S. B., M. L. McClelland, P. T. Stukenberg, D. J. Burke, M. M. Ross, J. Shabanowitz, D. F. Hunt and F. M. White
 "Phosphoproteome analysis by mass spectrometry and its application to *Saccharomyces cerevisiae*." *Nat Biotechnol*, **2002**, 20 (3), 301-5.

Fischer, E. H. and E. G. Krebs
 "Conversion of phosphorylase b to phosphorylase a in muscle extracts." *J Biol Chem*, **1955**, 216 (1), 121-32.

Gallien, S., E. Perrodou, C. Carapito, C. Deshayes, J. M. Reyrat, A. Van Dorsselaer, O. Poch, C. Schaeffer and O. Lecompte
 "Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol." *Genome Res*, **2009**, 19 (1), 128-35.

Geiss-Friedlander, R. and F. Melchior
 "Concepts in sumoylation: a decade on." *Nat Rev Mol Cell Biol*, **2007**, 8 (12), 947-56.

Gilchrist, A., C. E. Au, J. Hiding, A. W. Bell, J. Fernandez-Rodriguez, S. Lesimple, H. Nagaya, L. Roy, S. J. Gosline, M. Hallett, J. Paiement, R. E. Kearney, T. Nilsson and J. J. Bergeron
 "Quantitative proteomics analysis of the secretory pathway." *Cell*, **2006**, 127 (6), 1265-81.

Goldstein, J. L. and M. S. Brown
 "Regulation of the mevalonate pathway." *Nature*, **1990**, 343 (6257), 425-30.

Gruhler, A., J. V. Olsen, S. Mohammed, P. Mortensen, N. J. Faergeman, M. Mann and O. N. Jensen
 "Quantitative phosphoproteomics applied to the yeast pheromone signaling pathway." *Mol Cell Proteomics*, **2005**, 4 (3), 310-27.

Harborne, J. B.
 "Recent advances in the ecological chemistry of plant terpenoids", **1991**, Oxford, Clarendon Press.

Harker, M., N. Holmberg, J. C. Clayton, C. L. Gibbard, A. D. Wallace, S. Rawlins, S. A. Hellyer, A. Lanot and R. Safford
 "Enhancement of seed phytosterol levels by expression of an N-terminal truncated *Hevea brasiliensis* (rubber tree) 3-hydroxy-3-methylglutaryl-CoA reductase." *Plant Biotechnol J*, **2003**, 1 (2), 113-21.

He, T., K. Alving, B. Feild, J. Norton, E. G. Joseloff, S. D. Patterson and B. Domon
 "Quantitation of phosphopeptides using affinity chromatography and stable isotope labeling." *J Am Soc Mass Spectrom*, **2004**, 15 (3), 363-73.

Heazlewood, J. L., P. Durek, J. Hummel, J. Selbig, W. Weckwerth, D. Walther and W. X. Schulze
 "PhosPhAt: a database of phosphorylation sites in *Arabidopsis thaliana* and a plant-specific phosphorylation site predictor." *Nucleic Acids Res*, **2008**, 36 (Database issue), D1015-21.

Heintz, D., S. Gallien, S. Wischgoll, A. K. Ullmann, C. Schaeffer, A. K. Kretzschmar, A. van Dorsselaer and M. Boll
 "Differential membrane proteome analysis reveals novel proteins involved in the degradation of aromatic compounds in *Geobacter metallireducens*." *Mol Cell Proteomics*, **2009**,

Hem, S., V. Rofidal, N. Sommerer and M. Rossignol
 "Novel subsets of the *Arabidopsis thaliana* plasma membrane phosphoproteome identify phosphorylation sites in secondary active transporters." *Biochem Biophys Res Commun*, **2007**, 363 (2), 375-80.

Hemmerlin, A., J. F. Hoefler, O. Meyer, D. Tritsch, I. A. Kagan, C. Grosdemange-Billiard, M. Rohmer and T. J. Bach
 "Cross-talk between the cytosolic mevalonate and the plastidial methylerythritol phosphate pathways in tobacco bright yellow-2 cells." *J Biol Chem*, **2003**, 278 (29), 26666-76.

Huang, Y., J. M. Triscari, G. C. Tseng, L. Pasa-Tolic, M. S. Lipton, R. D. Smith and V. H. Wysocki
 "Statistical characterization of the charge state and residue dependence of low-energy CID peptide dissociation patterns." *Anal Chem*, **2005**, 77 (18), 5800-13.

Huang, Z. H., T. Shen, J. Wu, D. A. Gage and J. T. Watson

"Protein sequencing by matrix-assisted laser desorption ionization-postsource decay-mass spectrometry analysis of the N-Tris(2,4,6-trimethoxyphenyl)phosphine-acetylated tryptic digests." *Anal Biochem*, **1999**, 268 (2), 305-17.

Hunter, T.

"Protein kinases and phosphatases: the yin and yang of protein phosphorylation and signaling." *Cell*, **1995**, 80 (2), 225-36.

Hunter, T.

"Signaling--2000 and beyond." *Cell*, **2000**, 100 (1), 113-27.

Jalili, P. R., D. Sharma and H. L. Ball

"Enhancement of ionization efficiency and selective enrichment of phosphorylated peptides from complex protein mixtures using a reversible poly-histidine tag." *J Am Soc Mass Spectrom*, **2007**, 18 (6), 1007-17.

Javot, H., V. Lauvergeat, V. Santoni, F. Martin-Laurent, J. Guclu, J. Vinh, J. Heyes, K. I. Franck, A. R. Schaffner, D. Bouchez and C. Maurel

"Role of a single aquaporin isoform in root water uptake." *Plant Cell*, **2003**, 15 (2), 509-22.

Krebs, E. G. and J. A. Beavo

"Phosphorylation-dephosphorylation of enzymes." *Annu Rev Biochem*, **1979**, 48 923-59.

Kweon, H. K. and K. Hakansson

"Selective zirconium dioxide-based enrichment of phosphorylated peptides for mass spectrometric analysis." *Anal Chem*, **2006**, 78 (6), 1743-9.

Larsen, M. R., T. E. Thingholm, O. N. Jensen, P. Roepstorff and T. J. Jorgensen

"Highly selective enrichment of phosphorylated peptides from peptide mixtures using titanium dioxide microcolumns." *Mol Cell Proteomics*, **2005**, 4 (7), 873-86.

Laule, O., A. Furholz, H. S. Chang, T. Zhu, X. Wang, P. B. Heifetz, W. Gruissem and M. Lange

"Crosstalk between cytosolic and plastidial pathways of isoprenoid biosynthesis in *Arabidopsis thaliana*." *Proc Natl Acad Sci U S A*, **2003**, 100 (11), 6866-71.

Li, X., S. A. Gerber, A. D. Rudner, S. A. Beausoleil, W. Haas, J. Villen, J. E. Elias and S. P. Gygi

"Large-scale phosphorylation analysis of alpha-factor-arrested *Saccharomyces cerevisiae*." *J Proteome Res*, **2007**, 6 (3), 1190-7.

Lopez-Otin, C. and C. M. Overall

"Protease degradomics: a new challenge for proteomics." *Nat Rev Mol Cell Biol*, **2002**, 3 (7), 509-19.

Mann, M. and O. N. Jensen

"Proteomic analysis of post-translational modifications." *Nat Biotechnol*, **2003**, 21 (3), 255-61.

Mann, M., S. E. Ong, M. Gronborg, H. Steen, O. N. Jensen and A. Pandey

"Analysis of protein phosphorylation using mass spectrometry: deciphering the phosphoproteome." *Trends Biotechnol*, **2002**, 20 (6), 261-8.

Manning, G., D. B. Whyte, R. Martinez, T. Hunter and S. Sudarsanam

"The protein kinase complement of the human genome." *Science*, **2002**, 298 (5600), 1912-34.

McLachlin, D. T. and B. T. Chait

"Improved beta-elimination-based affinity purification strategy for enrichment of phosphopeptides." *Anal Chem*, **2003**, 75 (24), 6826-36.

McNulty, D. E. and R. S. Annan

"Hydrophilic interaction chromatography reduces the complexity of the phosphoproteome and improves global phosphopeptide isolation and detection." *Mol Cell Proteomics*, **2008**, 7 (5), 971-80.

Mikesh, L. M., B. Ueberheide, A. Chi, J. J. Coon, J. E. Syka, J. Shabanowitz and D. F. Hunt

"The utility of ETD mass spectrometry in proteomic analysis." *Biochim Biophys Acta*, **2006**, 1764 (12), 1811-22.

Molina, H., D. M. Horn, N. Tang, S. Mathivanan and A. Pandey

"Global proteomic profiling of phosphopeptides using electron transfer dissociation tandem mass spectrometry." *Proc Natl Acad Sci U S A*, **2007**, 104 (7), 2199-204.

Nandi, D., P. Tahiliani, A. Kumar and D. Chandu

"The ubiquitin-proteasome system." *J Biosci*, **2006**, 31 (1), 137-55.

Niittyla, T., A. T. Fuglsang, M. G. Palmgren, W. B. Frommer and W. X. Schulze

"Temporal analysis of sucrose-induced phosphorylation changes in plasma membrane proteins of *Arabidopsis*." *Mol Cell Proteomics*, **2007**, 6 (10), 1711-26.

Nuhse, T. S., A. R. Bottrill, A. M. Jones and S. C. Peck

"Quantitative phosphoproteomic analysis of plasma membrane proteins reveals regulatory mechanisms of plant innate immune responses." *Plant J*, **2007**, 51 (5), 931-40.

Nuhse, T. S., A. Stensballe, O. N. Jensen and S. C. Peck

"Phosphoproteomics of the *Arabidopsis* plasma membrane and a new phosphorylation site database." *Plant Cell*, **2004**, 16 (9), 2394-405.

Oda, Y., T. Nagasu and B. T. Chait

"Enrichment analysis of phosphorylated proteins as a tool for probing the phosphoproteome." *Nat Biotechnol*, **2001**, 19 (4), 379-82.

Ohyama, K., M. Suzuki, K. Masuda, S. Yoshida and T. Muranaka
 "Chemical phenotypes of the hmg1 and hmg2 mutants of Arabidopsis demonstrate the in-planta role of HMG-CoA reductase in triterpene biosynthesis." *Chem Pharm Bull (Tokyo)*, **2007**, 55 (10), 1518-21.

Palumbo, A. M., J. J. Tepe and G. E. Reid
 "Mechanistic insights into the multistage gas-phase fragmentation behavior of phosphoserine- and phosphothreonine-containing peptides." *J Proteome Res*, **2008**, 7 (2), 771-9.

Paradela, A. and J. P. Albar
 "Advances in the analysis of protein phosphorylation." *J Proteome Res*, **2008**, 7 (5), 1809-18.

Peng, L., C. H. Hocart, J. W. Redmond and R. E. Williamson
 "Fractionation of carbohydrates in Arabidopsis root cell walls shows that three radial swelling loci are specifically involved in cellulose production." *Planta*, **2000**, 211 (3), 406-14.

Pinkse, M. W., P. M. Uitto, M. J. Hilhorst, B. Ooms and A. J. Heck
 "Selective isolation at the femtomole level of phosphopeptides from proteolytic digests using 2D-NanoLC-ESI-MS/MS and titanium oxide precolumns." *Anal Chem*, **2004**, 76 (14), 3935-43.

Polevoda, B. and F. Sherman
 "N-terminal acetyltransferases and sequence requirements for N-terminal acetylation of eukaryotic proteins." *J Mol Biol*, **2003**, 325 (4), 595-622.

Reiland, S., G. Messerli, K. Baerenfaller, B. Gerrits, A. Endler, J. Grossmann, W. Gruissem and S. Baginsky
 "Large-scale Arabidopsis phosphoproteome profiling reveals novel chloroplast kinase substrates and phosphorylation networks." *Plant Physiol*, **2009**, 150 (2), 889-903.

Rohmer, M.
 "The discovery of a mevalonate-independent pathway for isoprenoid biosynthesis in bacteria, algae and higher plants." *Nat Prod Rep*, **1999**, 16 (5), 565-74.

Roth, K. D., Z. H. Huang, N. Sadagopan and J. T. Watson
 "Charge derivatization of peptides for analysis by mass spectrometry." *Mass Spectrom Rev*, **1998**, 17 (4), 255-74.

Sadagopan, N., M. Malone and J. T. Watson
 "Effect of charge derivatization in the determination of phosphorylation sites in peptides by electrospray ionization collision-activated dissociation tandem mass spectrometry." *J Mass Spectrom*, **1999**, 34 (12), 1279-82.

Sadagopan, N. and J. T. Watson
 "Investigation of the tris(trimethoxyphenyl)phosphonium acetyl charged derivatives of peptides by electrospray ionization mass spectrometry and tandem mass spectrometry." *J Am Soc Mass Spectrom*, **2000**, 11 (2), 107-19.

Sadoul, K., C. Boyault, M. Pabion and S. Khochbin
 "Regulation of protein turnover by acetyltransferases and deacetylases." *Biochimie*, **2008**, 90 (2), 306-12.

Schroeder, M. J., J. Shabanowitz, J. C. Schwartz, D. F. Hunt and J. J. Coon
 "A neutral loss activation method for improved phosphopeptide sequence analysis by quadrupole ion trap mass spectrometry." *Anal Chem*, **2004**, 76 (13), 3590-8.

Shi, S. D., M. E. Hemling, S. A. Carr, D. M. Horn, I. Lindh and F. W. McLafferty
 "Phosphopeptide/phosphoprotein mapping by electron capture dissociation mass spectrometry." *Anal Chem*, **2001**, 73 (1), 19-22.

Steen, H., B. Kuster, M. Fernandez, A. Pandey and M. Mann
 "Tyrosine phosphorylation mapping of the epidermal growth factor receptor signaling pathway." *J Biol Chem*, **2002**, 277 (2), 1031-9.

Steen, H., B. Kuster and M. Mann
 "Quadrupole time-of-flight versus triple-quadrupole mass spectrometry for the determination of phosphopeptides by precursor ion scanning." *J Mass Spectrom*, **2001**, 36 (7), 782-90.

Stensballe, A., O. N. Jensen, J. V. Olsen, K. F. Haselmann and R. A. Zubarev
 "Electron capture dissociation of singly and multiply phosphorylated peptides." *Rapid Commun Mass Spectrom*, **2000**, 14 (19), 1793-800.

Sugiyama, N., H. Nakagami, K. Mochida, A. Daudi, M. Tomita, K. Shirasu and Y. Ishihama
 "Large-scale phosphorylation mapping reveals the extent of tyrosine phosphorylation in Arabidopsis." *Mol Syst Biol*, **2008**, 4 193.

Sutherland, E. W., Jr. and W. D. Wosilait
 "Inactivation and activation of liver phosphorylase." *Nature*, **1955**, 175 (4447), 169-70.

Suzuki, M., Y. Kamide, N. Nagata, H. Seki, K. Ohyama, H. Kato, K. Masuda, S. Sato, T. Kato, S. Tabata, S. Yoshida and T. Muranaka

Partie III : Protéomique Quantitative

Chapitre 1 : La spectrométrie de masse quantitative en analyse protéomique

Chapitre 2 : Méthode de quantification protéomique par comptage de spectres ; application à *Geobacter metallireducens*

Chapitre 3 : Validation et application d'une méthode de quantification protéomique avec marquage au $^{13}\text{C}_9$ -TMPP.

Chapitre 1 : La spectrométrie de masse quantitative en analyse protéomique

Depuis une quinzaine d'années, la spectrométrie de masse a été largement utilisée pour analyser des échantillons biologiques et est devenue un outil indispensable pour la recherche en protéomique. La spectrométrie de masse a été appliquée avec succès pour l'identification et la caractérisation des protéines d'un mélange complexe mais les résultats obtenus n'ont longtemps été que qualitatifs. Cependant, la simple identification d'une protéine exprimée dans un système biologique n'est pas suffisante pour répondre à de nombreuses questions biologiques et une dimension quantitative est de plus en plus requise (par exemple la comparaison des niveaux d'expression protéiques entre deux conditions biologiques).

Les informations quantitatives peuvent être de deux types :

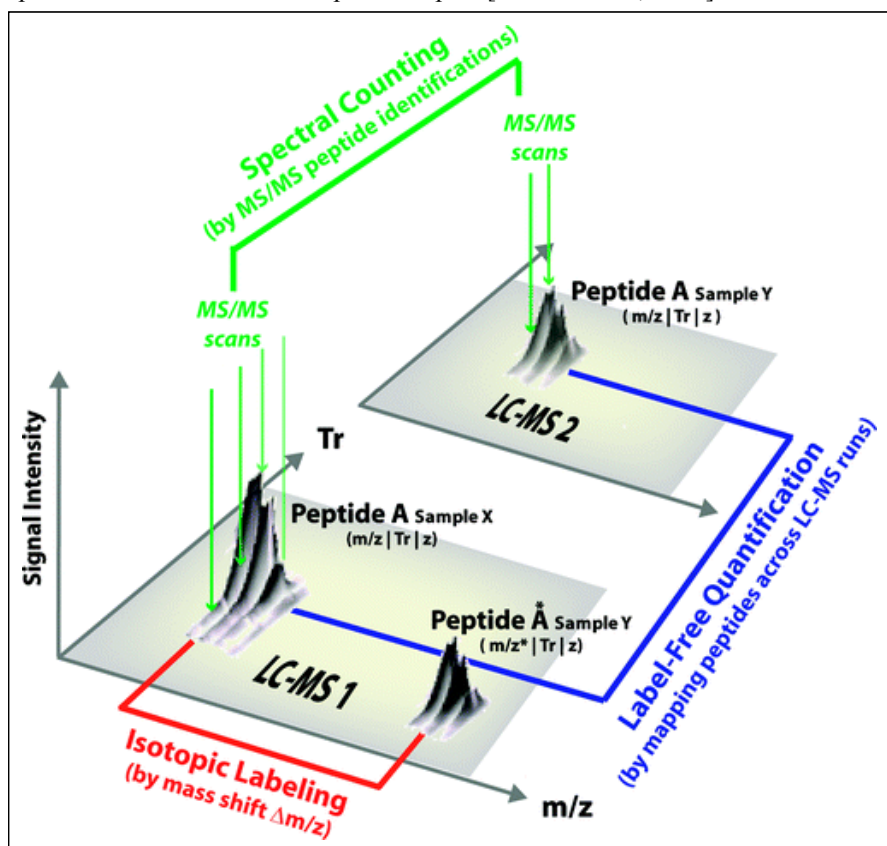
- Le changement relatif de l'abondance d'une protéine entre plusieurs états.
- La quantité absolue de la protéine dans le(s) échantillon(s) analysé(s) (par exemple la quantité en ng/mL d'un biomarqueur ou le nombre de copies/cellule d'une protéine).

Les premières approches de protéomique quantitative ont été largement dédiées à des fins de quantification relative tandis que les méthodes de quantification absolue n'ont été développées que plus récemment. Les principales méthodologies développées dans ces deux champs seront présentées dans la suite de ce chapitre.

1. La quantification relative en analyse protéomique

La quantification des différences dans l'expression de protéines entre plusieurs états physiologiques d'un système biologique est devenue une des tâches les plus importantes de la protéomique. Initialement, l'analyse protéomique comparative et quantitative était réalisée par gel d'électrophorèse bidimensionnel. Dans cette approche, les profils des gels de 2 échantillons protéiques séparés sur 2 gels distincts sont comparés et les protéines induites ou réprimées sont identifiées par une différence d'intensité du spot leur correspondant entre les 2 gels. Toutefois, les gels d'électrophorèse présentent certaines limitations : gamme dynamique restreinte, sous-représentation des protéines membranaires ou de pI extrêmes (comme décrit en partie bibliographique, Chapitre 2 2.1.1.) et faible reproductibilité des gels. Le problème de faible reproductibilité a été largement résolu par la technologie DIGE (« Differential Imaging Gel Electrophoresis ») dans laquelle plusieurs échantillons protéiques marqués avec différentes « étiquettes » chimiques fluorescentes sont séparés et comparés sur le même gel [Unlu et al., 1997]. Malgré les avantages de cette technologie, notamment en terme de séparation des protéines isoformes, la protéomique quantitative par gels a été supplantée par les approches protéomiques de quantification basées sur l'utilisation des données de spectrométrie de masse. Ces approches, schématisées en Figure 1, seront présentées dans la suite de ce chapitre.

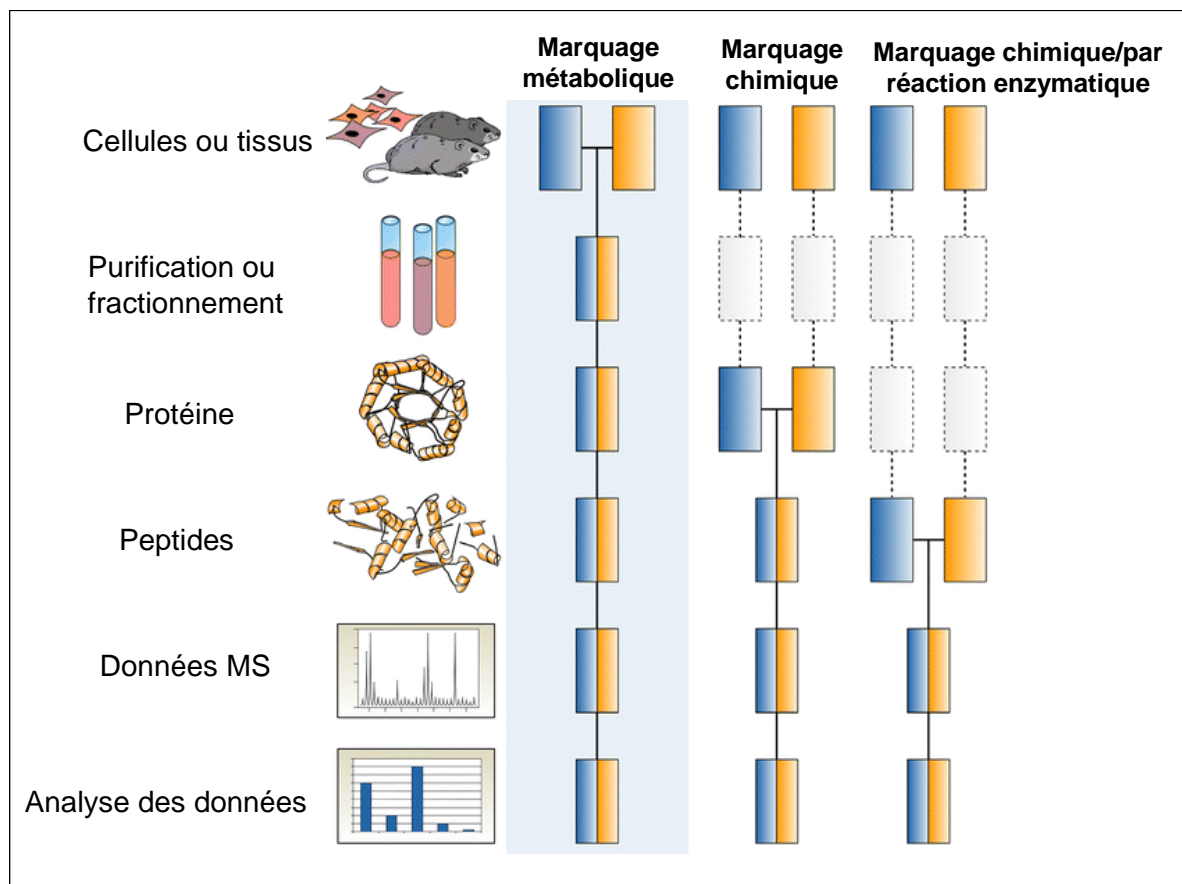
Figure 1 : Résumé des principales approches de quantification par spectrométrie de masse en analyse protéomique. Les détails des différentes approches sont présentés dans la suite du chapitre. D'après [Mueller et al., 2008]



1.1. Les méthodes de marquage par isotopes stables

Une approche majeure pour la quantification relative des protéines est basée sur la dilution isotopique. Cette approche est basée sur le fait qu'un peptide marqué par des isotopes stables (^{13}C à la place de ^{12}C , ^2H à la place de ^1H , ^{18}O à la place de ^{16}O ou ^{15}N à la place de ^{14}N) est chimiquement identique à son équivalent natif et par conséquent, les 2 peptides vont se comporter de manière identique en analyse chromatographique et en spectrométrie de masse. Etant donné que le spectromètre de masse peut mesurer la différence de masse entre les 2 formes du peptide (issues de 2 conditions biologiques distinctes), la quantification relative du peptide (entre les 2 conditions biologiques) est réalisée en comparant les intensités respectives des signaux MS des 2 formes du peptide. Cette approche a été introduite en 1999 par 3 laboratoires [Gygi et al., 1999; Oda et al., 1999; Pasa-Tolic et al., 1999] et a été largement adoptée par la communauté protéomique. Trois approches principales, qui diffèrent par le mode de marquage et l'étape d'introduction du marquage, existent aujourd'hui : i) le marquage métabolique, ii) le marquage chimique, iii) le marquage par réaction enzymatique.

Figure 2 : Mode opératoire des principales approches de quantification relative par marquage aux isotopes stables. Les rectangles bleus et oranges représentent 2 conditions expérimentales. Les lignes horizontales indiquent l'étape lors de laquelle les échantillons sont combinés. L'étape de marquage précède directement la combinaison des 2 échantillons. Les lignes en pointillés indiquent les étapes lors desquelles des variabilités expérimentales peuvent survenir. D'après [Bantscheff et al., 2007]



1.1.1. Marquage métabolique

La technique de marquage métabolique consiste à cultiver les cellules en présence d'un milieu de culture enrichi en ^{15}N pour réaliser le marquage complet des protéines et permettre ainsi la quantification de tous les peptides détectés par MS [Oda et al., 1999; Conrads et al., 2001]. Le marquage métabolique a beaucoup gagné en popularité grâce à l'introduction de l'approche « SILAC » (« Stable Isotope Labeling by Amino acids in Cell culture ») par l'équipe de M. Mann en 2002 [Ong et al., 2002] dans laquelle des acides aminés marqués aux isotopes stables sont introduits dans le milieu de culture. Plusieurs acides aminés marqués ont été utilisés pour cette approche, par exemple la leucine (deutérée) qui permet de marquer jusqu'à 70 % des peptides tryptiques [Foster et al., 2003]. Toutefois, la plupart du temps, l'utilisation conjointe de la $^{13}\text{C}_6$ -arginine et de la $^{13}\text{C}_6$ -lysine dans le milieu de culture est utilisée car elle permet d'obtenir, après digestion tryptique des protéines, des peptides contenant au moins un acide aminé marqué (excepté le peptide C-terminal des protéines). L'identification des peptides est réalisée grâce au spectre de fragmentation d'au moins une des formes du peptide (« légère » ou « lourde ») et la quantification relative est réalisée en comparant les

intensités des massifs isotopiques du peptide sous forme native et du peptide sous forme marquée dans le spectre MS. Dans l'approche SILAC, contrairement au marquage métabolique ^{15}N , le nombre de molécules marquées introduites dans le peptide est défini et ne dépend pas de la formule chimique du peptide, ce qui facilite le traitement des données.

Le principal avantage du marquage métabolique est sa réalisation à la première étape de l'expérience, au niveau des cellules, ce qui permet de s'affranchir des variabilités expérimentales qui pourraient être générées par un traitement en parallèle des échantillons dans les étapes à suivre de l'expérience. Cependant, l'approche SILAC est limitée au matériel biologique généré en culture et n'est donc généralement pas applicable aux tissus, aux fluides biologiques ou aux applications cliniques. Quelques études ont toutefois montré la faisabilité du marquage métabolique *in vivo* sur des organismes multicellulaires (*Caenorhabditis elegans*, *Drosophila melanogaster* [Krijgsveld et al., 2003], sur la plante [Ippel et al., 2004] ou encore sur le rat [Wu et al., 2004]) mais elles restent très ponctuelles. Enfin, le problème de la conversion métabolique des arginines en prolines a récemment été rapporté et peut compliquer le traitement des données [Van Hoof et al., 2007].

1.1.2. Marquage chimique

Une grande variété de réactifs chimiques pour la quantification relative par marquage aux isotopes stables a été rapportée. Dans ces approches, le marquage est réalisé *in vitro* soit au niveau protéique soit au niveau peptidique (Figure 1). Tous les réactifs chimiques développés pour ces approches ciblent des fonctions réactives des protéines ou des peptides (groupements thiol, amine, carboxyle) et les deux protéomes à comparer sont marqués respectivement avec la forme « légère » et « lourde » du réactif (excepté pour les stratégies de marquages isobariques pour lesquelles les différentes formes du réactif ont la même masse). Deux grands types de stratégies par marquage chimique aux isotopes stables existent : les stratégies de marquage isotopique et les stratégies de marquage isobarique.

1.1.2.1. Stratégies de marquage isotopique

Avec les stratégies de marquage chimique isotopique, la quantification relative est réalisée, de manière similaire aux stratégies de marquage métabolique, par comparaison des intensités des massifs isotopiques du peptide sous forme native et du peptide sous forme marquée dans le spectre MS. La stratégie pionnière dans ce domaine, et même dans le domaine de la spectrométrie de masse quantitative en analyse protéomique, fut l'approche introduite en 1999 par l'équipe de R. Aebersold : l'approche « ICAT » (« Isotope-Coded Affinity Tag ») [Gygi et al., 1999]. Le réactif est constitué d'un groupement réactif pour les thiols, d'un groupement biotine pour la purification des peptides marqués (par interaction sur une phase avidine) et d'une région faisant le lien entre les 2 groupements et portant 8 atomes de deuterium à la place des hydrogènes pour la forme lourde du réactif. Après le succès

initial de l'approche ICAT, plusieurs variantes du réactif ont été élaborées pour améliorer le recouvrement des peptides marqués ou leurs propriétés chromatographiques (particulièrement leur co-élution en chromatographie liquide en phase inverse) [Hansen et al., 2003; Li et al., 2003; Oda et al., 2003; Gartner et al., 2007]. Comme la cystéine est un acide aminé plutôt « rare » dans les séquences protéiques, les méthodes visant le marquage et la purification des peptides contenant cet acide aminé permettent de réduire la complexité du mélange de peptides analysés. Cet avantage peut rapidement devenir une limitation car ces approches ne permettent pas de quantifier les protéines qui ne contiennent pas de cystéines. D'autres approches comme l'approche « ICPL » (« Isotope-Coded Protein Label ») [Schmidt et al., 2005] visent à modifier les fonctions amines (des peptides ou des protéines) et permettent ainsi une quantification relative d'un plus grand nombre de peptides (parfois au détriment du nombre de protéines).

1.1.2.2. Stratégies de marquage isobarique

Avec les stratégies de marquage chimique isobarique, la quantification relative est réalisée par comparaison des intensités d'ions fragments spécifiques dans les spectres MS/MS des peptides marqués. Ces stratégies illustrées par les approches « iTRAQ » (« isobaric Tag for Relative and Absolute Quantification ») [Ross et al., 2004] et « TMT » (« Tandem Mass Tags ») [Thompson et al., 2003] utilisent un concept novateur puisque le marquage génère un ion fragment spécifique (appelé ion reporteur) pour la quantification dans les spectres MS/MS. Par conséquent, les spectres MS sont relativement simples puisqu'un peptide marqué par les différentes formes du réactif présente une masse unique avec la co-élution parfaite des différentes formes du peptide marqué. Seul son spectre MS/MS atteste de ce marquage par les différentes formes du réactif. Dans sa dernière version, le marquage iTRAQ permet de réaliser la quantification relative de 8 échantillons simultanément [Choe et al., 2007], ce qui est particulièrement utile pour suivre l'évolution d'un système biologique à de multiples instants ou pour comparer l'influence de nombreuses conditions biologiques différentes sur un protéome donné.

1.1.3. Marquage par réaction enzymatique

Le marquage des peptides aux isotopes stables peut également être réalisé par différentes protéases comme la trypsine, la Lys-N ou la Glu-C [Mirgorodskaya et al., 2000; Yao et al., 2001; Rao et al., 2005]. Dans ces approches, la digestion est réalisée dans de l'eau lourde ($H_2^{18}O$) et des échanges enzymatiques de l'oxygène se produisent sur les groupements carboxyle des peptides générés. La trypsine est utilisée préférentiellement car elle catalyse efficacement l'échange de 2 atomes d'oxygène C-terminaux, ce qui conduit à une augmentation de masse de 4 Da soit le minimum requis pour séparer les enveloppes isotopiques d'un peptide différentiellement marqué. Cette méthode qui permet en théorie de marquer tous les peptides générés par la protéolyse, peut être appliquée à de très faibles

quantités de matériel et est compatible avec pratiquement tous les types de préparation d'échantillon. L'incorporation des isotopes ^{18}O peut également être réalisée post-digestion [Staes et al., 2004]. Dans l'approche de marquage par réaction enzymatique, le marquage est réalisé au niveau du peptide et toutes les étapes précédentes dans la préparation des échantillons sont réalisées en parallèle et peuvent conduire à une variabilité expérimentale entre les 2 échantillons à comparer (Figure 2). En plus, si la trypsine n'est pas complètement inactivée ou éliminée après échange isotopique, elle peut être à l'origine d'un ré-échange des isotopes ^{18}O dans des solvants contenant de l'eau naturelle [Staes et al., 2004]. Ce type de ré-échange peut aussi être observé en l'absence de trypsine et à très faible pH. Enfin, il a également été noté que l'incorporation des isotopes ^{18}O est très lente et incomplète pour les peptides très acides [Gevaert et al., 2008].

La quantification relative utilisant les méthodes de marquage par isotopes stables nécessite l'utilisation de spectromètres de masse suffisamment résolutifs pour distinguer les massifs isotopiques des peptides d'intérêt des signaux interférents co-élués et de masse très proche (correspondant à d'autres peptides ou à d'autres molécules chimiques). La qualité de la quantification relative dépend également de la gamme de linéarité de l'instrument. En effet, pour des signaux MS relativement intenses, le détecteur du spectromètre de masse utilisé peut saturer et par conséquent ne plus présenter une réponse linéaire en fonction de la quantité de l'analyte détectée, ce qui peut entraîner une quantification relative inexacte. Ce problème est le plus souvent rencontré sur des instruments de type Q-TOF utilisant un détecteur avec TDC (« Time-to-Digital Converter »).

1.2. Les méthodes sans marquage (« label-free »)

Actuellement, deux grandes stratégies de quantification relative sans marquage (« label-free ») peuvent être distinguées : i) la stratégie qui consiste à mesurer et comparer l'intensité des signaux MS des peptides appartenant à une protéine donnée entre plusieurs analyses et ii) la stratégie qui consiste à compter et comparer le nombre de spectres MS/MS identifiant les peptides d'une protéine donnée entre plusieurs analyses. Ces deux types de stratégies sont présentés dans la suite du chapitre.

1.2.1. Comparaison des intensités des signaux MS

Dans la stratégie de comparaison des intensités des signaux MS, les chromatogrammes d'ions extraits de chaque peptide dans une expérience LC-MS ou LC-MS/MS sont utilisés comme mesure quantitative de l'abondance du peptide et sont comparés aux signaux respectifs issus d'une ou plusieurs autres expériences pour obtenir des informations de quantification relative [Bondarenko et al., 2002; Chelius et al., 2002; Wang et al., 2003; Wiener et al., 2004; Wang et al., 2006]. Cette stratégie de quantification relative peut être réalisée à partir d'analyses LC-MS ou LC-MS/MS.

Lorsque la stratégie est réalisée à partir d'analyses LC-MS/MS, un bon équilibre entre l'acquisition des spectres MS et l'acquisition des spectres MS/MS doit être trouvé. En effet, si d'un côté un séquençage extensif des peptides par MS/MS est requis pour identifier le plus grand nombre de protéines, une quantification robuste par la mesure de l'intensité des peptides nécessite d'acquérir un nombre de spectres MS suffisant pour obtenir un nombre de points qui permette de définir correctement le pic chromatographique du peptide. Ainsi, même pour les instruments présentant les meilleures vitesses d'acquisition, la quantification la plus juste est obtenue au détriment de la couverture du protéome à l'étude et inversement.

Pour résoudre ce problème, une solution consiste à mener deux expériences séparées, l'une se focalisant sur l'identification du plus grand nombre de peptides possible par MS/MS et l'autre étant réalisée en MS seulement afin d'optimiser l'échantillonnage des signaux des peptides intacts. Les analyses LC-MS/MS peuvent être réalisées dans une première étape de manière extensive pour constituer une large banque d'identification des peptides définis par leur séquence, leur rapport m/z et leur temps de rétention (approche « AMT », « Accurate Mass and retention Time », [Smith et al., 2003; Pasa-Tolic et al., 2004; Fang et al., 2006]). Les analyses LC-MS des échantillons à comparer sont effectuées dans un second temps en réalisant l'identification des différentes espèces par comparaison de leurs coordonnées (rapport m/z, temps de rétention) avec la base AMT et la quantification relative des différentes espèces par comparaison de l'intensité de leur signal MS entre les analyses (hauteur ou aire du pic chromatographique du chromatogramme d'ion extrait). Une autre stratégie consiste à réaliser d'abord les analyses LC-MS des échantillons à comparer afin de réaliser les quantifications relatives des différentes espèces par comparaison de l'intensité de leur signal MS. Dans ce cas, l'identification des espèces quantifiées est réalisée dans un second temps à travers de nouvelles analyses LC-MS/MS avec la possibilité d'établir des listes d'inclusion pour favoriser l'identification des espèces d'intérêt.

La stratégie de quantification relative « label-free » par comparaison des intensités des signaux MS nécessite l'utilisation de spectromètres de masse très résolutifs (par exemple de type Q-TOF ou Orbitrap) et de systèmes chromatographiques très reproductibles pour une détermination des coordonnées (rapport m/z, temps de rétention) les plus précises possibles. Il est également nécessaire de disposer de programmes informatiques très performants pour les traitements post-analytiques, notamment pour les corrections de fluctuations chromatographiques, afin de réaliser une quantification relative fiable.

1.2.2. Comparaison des nombres de spectres (« spectral counting »)

La stratégie de comptage et de comparaison du nombre de spectres (approche « spectral counting », [Washburn et al., 2001; Liu et al., 2004; Gilchrist et al., 2006]) est basée sur l'observation empirique que plus une protéine donnée est présente dans un échantillon, plus le nombre de spectres MS/MS collectés pour les peptides de cette protéine est important dans l'analyse LC-MS/MS de

l'échantillon. L'approche spectral counting transforme donc la fréquence à laquelle un peptide est identifié en une mesure de l'abondance du peptide. Les nombres de spectres des peptides sont ensuite considérés au niveau protéique comme un index de l'abondance des protéines [Gao et al., 2003; Liu et al., 2004; Colinge et al., 2005; Ishihama et al., 2005]. La quantification relative peut ensuite être réalisée en comparant l'index d'abondance de chacune des protéines entre les différents échantillons étudiés. La quantification relative des protéines par l'approche spectral counting est très intuitive et très simple à mettre en œuvre mais doit être considérée avec précaution si le nombre de spectres MS/MS collectés pour les peptides d'une protéine est très faible. Les raisons d'un faible nombre de spectres MS/MS pour une protéine sont multiples et incluent la faible abondance de la protéine, les protéines présentant une séquence courte, les propriétés physicochimiques spécifiques affectant l'observabilité d'un peptide, la complexité de l'échantillon, etc. [Mueller et al., 2008]. [Old et al., 2005] ont montré que bien qu'il était possible de détecter un changement d'abondance d'un facteur 3 pour une protéine avec seulement 4 spectres MS/MS disponibles, ce nombre augmentait exponentiellement pour de plus faibles variations (15 spectres MS/MS pour un changement d'abondance d'un facteur 2). Néanmoins, la corrélation entre la quantité de protéines et le nombre de spectres MS/MS a conduit certains chercheurs à étendre l'utilisation de l'approche spectral counting à l'estimation des niveaux d'expression absolus des protéines. Ces approches seront décrites dans le paragraphe 2.1.2. de ce chapitre.

La stratégie de quantification relative « label-free » par l'approche spectral counting implique la réalisation d'un maximum de spectres MS/MS informatifs dans les analyses pour une quantification de qualité. Les instruments de type trappe ionique présentent l'avantage d'enchaîner les cycles MS et MS/MS à très grande vitesse et sont donc des instruments de choix pour cette stratégie. Le traitement post-analytique des analyses dans cette approche ne nécessite pas l'utilisation de programmes informatiques très perfectionnés.

Globalement, les approches de quantification sans marquage sont à l'heure actuelle les approches de spectrométrie de masse quantitative décrites comme les moins précises car l'ensemble des variations dans les différentes étapes de la préparation des échantillons se répercutent sur les données finales obtenues. Toutefois, ces modes de quantification présentent l'avantage de ne pas avoir de limites en termes de nombre d'expériences comparables et de fournir une plus grande gamme dynamique de quantification que les méthodes utilisant le marquage aux isotopes stables.

L'ensemble des stratégies de quantification relative présentées dans ce paragraphe a été développé pour explorer la dynamique de protéomes « entiers » et fournissent une comparaison relative des abondances des protéines entre quelques échantillons présentant divers états physiologiques. Ces stratégies ont été notamment largement appliquées pour étudier diverses pathologies humaines, en comparant les échantillons provenant de patients malades et les échantillons provenant de patients sains, dans le but de « découvrir » des biomarqueurs et de comprendre la

pathogénèse d'une maladie [Hsich et al., 1996; Ahmed et al., 2004; Chen et al., 2005; Le et al., 2005; Miguet et al., 2009].

2. La quantification absolue en analyse protéomique

A la différence des stratégies de quantification relative, les stratégies de quantification absolue en analyse protéomique visent à déterminer les concentrations des protéines dans des échantillons biologiques. Ces données de quantification absolue, obtenues de manière ciblée sur les protéines d'intérêt, sont très importantes dans le contexte de la protéomique clinique lors de la phase de « validation » de biomarqueurs et permettent de comparer véritablement les données entre laboratoires [Pan et al., 2009]. Ces données de quantification absolue peuvent également être obtenue à plus grande échelle dans le but de réaliser des modélisations mathématiques et des simulations de processus biologiques dans des systèmes biologiques complexes [Malmstrom et al., 2009]. Les stratégies développées au cours des dernières années dans le domaine de la quantification absolue en analyse protéomique seront présentées dans la suite de ce chapitre.

2.1.1. Les stratégies de dilution isotopique

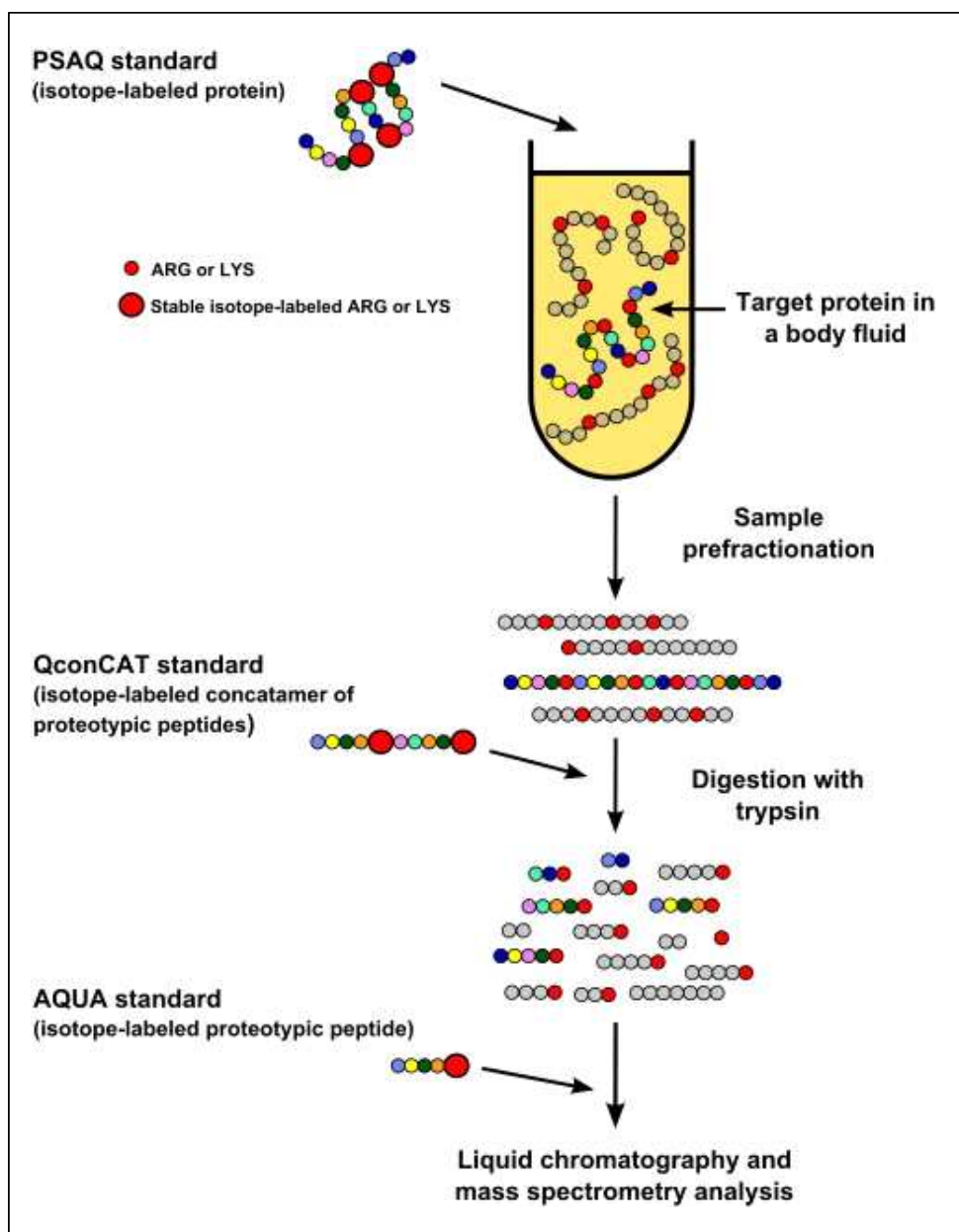
Traditionnellement, le dosage immunoenzymatique sur support solide (« ELISA », « Enzyme-Linked ImmunoSorbent Assay ») est la principale méthode utilisée pour la quantification ciblée d'une protéine. Toutefois, pour certaines protéines d'intérêt, cette approche est limitée par le manque de disponibilité d'anticorps hautement spécifiques et requiert un fort investissement en temps et en argent pour développer un dosage ELISA de qualité. Les stratégies de protéomique quantitative ciblée, et notamment les approches de dilution isotopique apparaissent comme de bonnes alternatives au dosage ELISA. Le concept de dilution isotopique, reconnu depuis des dizaines d'années comme l'approche de référence pour la quantification par spectrométrie de masse des petites molécules [Viswanathan et al., 2007], a récemment été transféré à la quantification absolue des protéines dans les échantillons biologiques. Les stratégies de dilution isotopique reposent sur l'addition de quantités définies de standards marqués aux isotopes stables, qui présentent un comportement chromatographique identique aux composés natifs mais qui peuvent être distingués par leur différence de masse.

Dans le cas le plus simple, la quantification absolue peut être réalisée par l'addition d'une quantité connue d'un peptide standard marqué aux isotopes stables dans un digest protéique et par comparaison des signaux MS du standard et du peptide endogène. Cette méthode introduite par [Gerber et al., 2003] est connue sous l'acronyme « AQUA » (« Absolute QUAntification »). La disponibilité commerciale et l'utilisation facile des peptides AQUA a rendu cette approche très attractive dans le domaine de la validation de biomarqueurs d'intérêt clinique [Pan et al., 2005]. Toutefois, des problèmes de spécificité des peptides standard utilisés peuvent se poser car on peut

trouver plusieurs peptides isobariques présents dans le mélange. En plus, il faut définir soigneusement les quantités de peptide standard ajoutées dans l'échantillon. Cette quantité peut varier pour les différentes protéines d'intérêt dont l'abondance peut être très différente dans un échantillon. Ces deux problèmes ont été largement résolus grâce à la méthode dite MRM (« Multiple Reaction Monitoring ») réalisable par un spectromètre de masse de type triple quadripôle [Lange et al., 2008]. Dans cette méthode, le premier et le troisième quadripôle agissent comme des filtres pour sélectionner spécifiquement des valeurs prédéfinies de m/z correspondant à un ion peptidique et à un ion fragment de ce peptide alors que le second quadripôle sert de cellule de collision. La combinaison des caractéristiques temps de rétention, rapport m/z du peptide et d'un fragment permet d'être spécifique dans l'attribution du signal MS au peptide et d'étendre la gamme de quantification à 5 ordres de grandeur [Wolf-Yadlin et al., 2007; Picotti et al., 2009]. Le choix des peptides standard synthétiques utilisés dans la méthode AQUA est important pour assurer la protéotypicité des peptides choisis (peptides observables expérimentalement et qui sont spécifiques d'une protéine ou d'une protéine isoforme). Ce choix peut être réalisé empiriquement (expériences préliminaires, utilisation des bases de données type PeptideAtlas [Desiere et al., 2006]) ou par prédiction de la protéotypicité des peptides candidats [Kuster et al., 2005; Mallick et al., 2007; Fusaro et al., 2009]. La disponibilité des peptides standard synthétiques choisis en quantités précisément connues est une limitation à l'utilisation large de cette approche.

La stratégie « QconCAT » (« Quantification CONCATemer ») [Pratt et al., 2006] permet de surmonter partiellement cette limitation. Dans cette stratégie, un gène artificiel est utilisé pour l'expression, le marquage et la purification de la protéine artificielle correspondante qui représente un concatémère de peptides tryptiques d'une ou plusieurs protéines connues. Cette approche permet d'étendre la gamme de peptides protéotypiques accessibles (par exemple les peptides hydrophobes). En plus de fournir de multiples peptides standard pour la quantification, cette approche implique l'ajout de la protéine synthétique avant l'étape de digestion dans le processus de préparation des échantillons (Figure 3) et permet donc de limiter un biais potentiel lors de la digestion protéique. Toutefois, les construits artificiels QconCAT peuvent présenter un taux de digestion différents des protéines ciblées [Rivers et al., 2007] et ne sont pas compatibles avec les méthodologies de fractionnement extensives des échantillons. Ces deux limitations peuvent être surmontées avec la stratégie « PSAQ » (« Protein Standard Absolute Quantification ») [Brun et al., 2007] dans laquelle le standard interne de quantification absolue d'une protéine est constitué de l'équivalent intégral de la protéine (protéine recombinante) marquée aux isotopes stables. Comme l'ajout du standard est réalisé à la première étape de la préparation des échantillons (Figure 3), cette méthode est compatible avec les stratégies de préfractionnement et permet de s'affranchir de potentiels biais dans toutes les étapes suivantes de la préparation. L'utilisation de cette méthode pour mesurer précisément la concentration d'un anticorps monoclonal total dans un échantillon, en dépit de nombreuses étapes de préparation, a démontré tout l'intérêt de l'approche PSAQ [Heudi et al., 2008]. Une limitation actuelle de la méthode PSAQ est le coût et la difficulté de production des protéines standard [Brun et al., 2009].

Figure 3 : Les principales stratégies de dilution isotopique pour la quantification absolue ciblée de protéines. Les différentes stratégies diffèrent par la nature du standard et l'étape à laquelle celui-ci est ajouté. D'après [Brun et al., 2009]



2.1.2. Les méthodes alternatives

Les stratégies de dilution isotopique combinées à l'approche MRM sont sensibles, très reproductibles [Addona et al., 2009] et pourraient en théorie être appliquée à la quantification absolue d'un protéome entier. Cependant, ce travail impliquerait la préparation de milliers de peptides standard marqués de concentration connue et serait donc extrêmement coûteux en temps et en argent. Des méthodes ne nécessitant pas l'utilisation de peptides marqués aux isotopes stables ont récemment été introduites pour la quantification absolue de protéines. Ces méthodes qui mesurent l'abondance

absolue des protéines à partir des données collectées dans les expériences d'analyses protéomiques de routine constituent une alternative à l'approche MRM à moindre coût.

Une de ces stratégies utilise les intensités des signaux MS des peptides en considérant que l'abondance d'une protéine peut être correctement estimée à partir de la moyenne des intensités des signaux MS des trois peptides présentant le meilleur facteur de réponse en ionisation électrospray [Silva et al., 2006; Cheng et al., 2009]. L'addition d'une protéine en quantité connue dans l'échantillon permet donc d'estimer les concentrations des autres protéines.

Une autre stratégie de quantification absolue est basée sur l'approche spectral counting (présentée en 1.2.2. de ce chapitre). Cette stratégie part de l'observation que plus une protéine est abondante dans un mélange, plus le nombre de spectres MS/MS collectés pour les peptides de cette protéine est élevé [Liu et al., 2004]. Il est donc possible d'estimer l'abondance relative des différentes protéines présentes dans un mélange donné en comparant le nombre de spectres MS/MS collectés pour l'ensemble des peptides de chacune de ces protéines. Toutefois, les différentes protéines d'un échantillon biologique peuvent être très différentes et ces différences peuvent influencer sur le nombre de spectres MS/MS collectés pour chacune d'entre elles. Par exemple, la digestion enzymatique d'une protéine présentant une séquence « courte » génère moins de peptides que la digestion enzymatique d'une protéine de séquence plus « longue ». A quantité de protéine égale, moins de spectres MS/MS sont donc collectés pour la protéine « courte ». Une estimation plus juste de l'abondance relative des protéines dans un échantillon passe donc par la normalisation du nombre de spectres MS/MS collectés pour les peptides en tenant compte des particularités des différentes protéines. Cette normalisation peut être réalisée pour chaque protéine en divisant le nombre de spectres MS/MS observés pour la protéine par la taille ou la masse de la protéine [Blondeau et al., 2004; Bergeron et al., 2007; Bernay et al., 2009]. Ce type de normalisation permet de corriger les estimations des abondances relatives des protéines de l'échantillon de manière satisfaisante. [Lu et al., 2007] ont proposé dans leur approche « APEX » (« Absolute Protein Expression measurement ») une méthode de normalisation plus poussée qui tient compte des propensions prédites des différents peptides à être identifiés dans les analyses MS/MS. Les résultats obtenus avec ce mode de traitement ont bien corrélé avec ceux obtenus par d'autres méthodes. L'approche spectral counting permet donc une estimation de l'abondance relative des protéines présentes dans un échantillon de manière simple. La connaissance de la quantité d'une ou plusieurs protéines présentes dans les mélanges (ou ajoutées artificiellement) permet d'estimer l'abondance absolue de l'ensemble des protéines du mélange.

Les deux méthodes alternatives de quantification absolue présentées ici conduisent à des erreurs sur l'estimation des concentrations des protéines plus importantes que la stratégie combinant la dilution isotopique et l'analyse MRM. Par contre, elles permettent d'accéder plus facilement à la quantification absolue d'une grande partie du protéome à l'étude. Très récemment, une étude a combiné les différentes approches de quantification absolue pour réaliser la quantification absolue de 50 % du protéome théorique de *Leptospira interrogans*. L'approche combinant la dilution isotopique et l'analyse MRM a permis de quantifier précisément 19 protéines d'abondance différentes qui ont servi de points de calibration pour la quantification de l'ensemble du protéome par les deux autres

approches. Les résultats obtenus ont mis en évidence la puissance de la stratégie générale employée qui a conduit à une erreur de quantification moyenne d'un facteur 2 pour l'approche de [Silva et al., 2006] et d'un facteur 3 pour l'approche spectral counting.

Chapitre 2 : Méthode de quantification protéomique par comptage de spectres ; application à *Geobacter metallireducens*

Cette étude a été réalisée en collaboration avec l'équipe du Professeur Matthias Boll de l'institut de biochimie de l'Université de Leipzig en Allemagne.

1. Contexte de l'étude

1.1. Métabolisme anaérobie des composés aromatiques

Les composés aromatiques forment une grande classe de composés naturels ou générés par l'homme et qui ne peuvent être minéralisés en CO₂ que par des microorganismes. Cette minéralisation peut être réalisée par des microorganismes anaérobies ou aérobies. Les bactéries aérobies métabolisant les composés aromatiques utilisent le dioxygène comme cosubstrat pour la plupart des processus clé [Fuchs, 2008]. Par exemple, les oxygénases catalysent les hydroxylations exergoniques du cycle aromatique ainsi que sa coupure ultérieure. Les bactéries anaérobies utilisent des stratégies différentes pour l'attaque du cycle aromatique. Jusqu'à présent, le métabolisme anaérobie des composés aromatiques a essentiellement été étudié chez les bactéries anaérobies facultatives [Boll et al., 2002; Gibson et al., 2002; Diaz, 2004; Fuchs, 2008]. Les réactions impliquées dans le catabolisme des composés aromatiques chez les bactéries anaérobies facultatives, comme *Thauera aromatica*, sont présentées en Figure 1. Par contre, le métabolisme des composés aromatiques dans les bactéries anaérobies obligatoires reste très peu connu. La Figure 1 illustre le rôle clé de l'enzyme ATP-dépendant benzoyl-CoA reductase dans la voie de dégradation des composés aromatiques chez les bactéries anaérobies facultatives en participant à la réduction du noyau benzénique. Or, les protéines constitutives de cette enzyme ne sont pas codées dans le génome des bactéries anaérobies obligatoires [Wischgoll et al., 2005]. Ces dernières doivent donc utiliser des enzymes et des mécanismes différents pour dégrader les composés aromatiques.

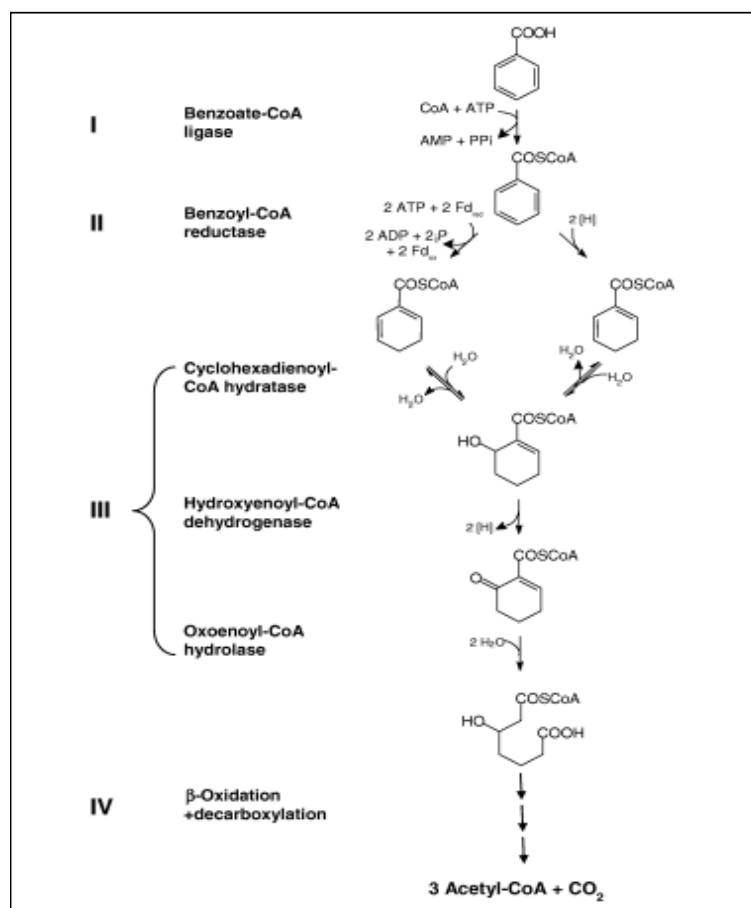


Figure 1 : Réactions impliquées dans le métabolisme du composé aromatique modèle benzoate chez la bactérie anaérobie facultative *T. aromatica*.

Etape 1 : Activation du benzoate en benzoyl-CoA catalysée par la benzoate-CoA ligase.

Etape 2 : Réduction du noyau benzénique par l'enzyme clé benzoyl-CoA reductase. Le transfert des électrons au noyau benzénique est couplé à l'hydrolyse stoechiométrique de l'ATP.

Etape 3 : β-oxydation modifiée conduisant à l'ouverture du cycle.

Etape 4 : Série de β-oxydation conduisant à la formation de 3 acetyl-CoA et de CO₂. L'acetyl-CoA est ensuite complètement minéralisé en CO₂ dans le cycle de Krebs

D'après [Wischgoll et al., 2005]

1.2. *Geobacter metallireducens*

La bactérie *Geobacter metallireducens*, l'organisme modèle de cette étude, est une bactérie anaérobie obligatoire dont le génome est séquencé. Cet organisme appartient à la famille des *Geobacteraceae* (δ -proteobactéries). *G. metallireducens* a été isolé et décrit par [Lovley et al., 1988; Lovley et al., 1993]. En plus de leur importance dans la dégradation des composés aromatiques, *G. metallireducens* et les autres *Geobacteraceae* sont célèbres pour leur capacité à transférer des électrons aux oxydes métalliques et même à des électrodes. Dans ce dernier cas, de l'électricité peut être produite [Lovley, 2003; Lovley et al., 2004].

Précédemment, une étude différentielle par gel 2D des protéomes solubles de *G. metallireducens* cultivé en condition anaérobie avec ou sans source de carbone aromatique avait fourni un premier ensemble de résultats participant à la compréhension des particularités de la dégradation des composés aromatiques chez les bactéries anaérobies obligatoires [Wischgoll et al., 2005]. Cette étude avait notamment permis de mettre en évidence quelques protéines différenciellement exprimées qui pourraient appartenir à un complexe supposé réaliser la réduction du noyau benzénique à la place des enzymes ATP-dépendent benzoyl-CoA reductases des bactéries anaérobies facultatives. En combinant l'ensemble des données protéomiques obtenues avec des résultats de qRT-PCR, deux clusters de gènes impliqués dans le métabolisme des composés aromatiques chez *G. metallireducens* et dont l'expression était induite lors de la culture avec une source de carbone aromatique avaient

également pu être mis en évidence (clusters dénommés IA/B et II). Toutefois, ces résultats n'avaient fourni qu'une vision restreinte des processus et des constituants impliqués dans la dégradation des composés aromatiques probablement parce que certaines catégories de protéines (membranaires, pI extrêmes) ne sont pas compatibles avec le gel 2D [Santoni et al., 2000; Braun et al., 2007].

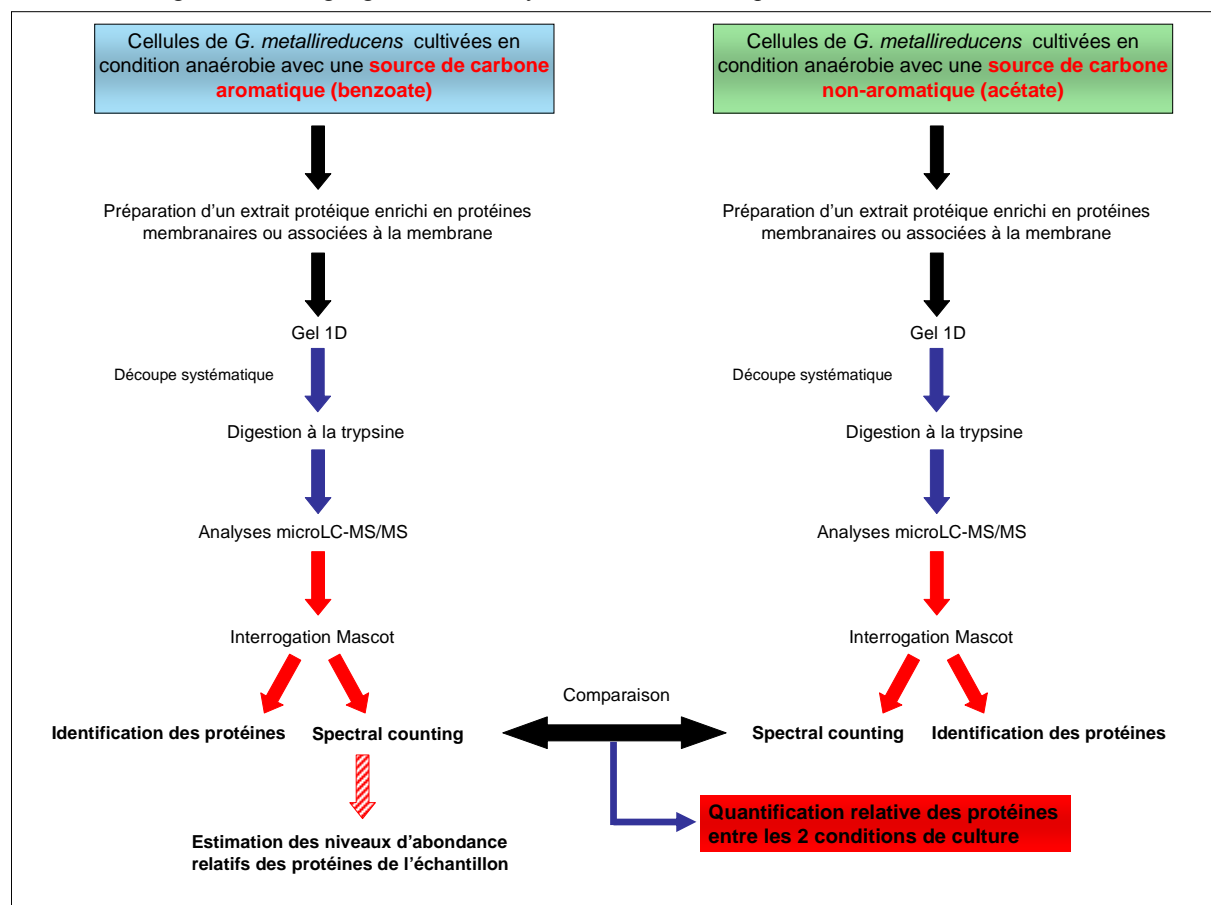
1.3. Objectif de l'étude

Dans ce contexte, afin d'améliorer notre compréhension du catabolisme des composés aromatiques chez les bactéries anaérobies obligatoires, nous avons réalisé une analyse différentielle des protéomes de *G. metallireducens* cultivés en condition anaérobie avec ou sans source de carbone aromatique à partir d'un mode de préparation permettant l'analyse des protéines membranaires ou liées à la membrane. Pour cela, nous avons réalisé la mise en place au laboratoire d'une approche de quantification sans marquage, de type « spectral counting », à grande échelle.

2. Stratégie d'analyse protéomique

La stratégie générale d'analyse protéomique est détaillée en Figure 2

Figure 2 : Stratégie générale d'analyse différentielle des protéomes de *G. metallireducens*



2.1. Préparation des échantillons

Les extraits protéiques ont été préparés à partir des cellules de *G. metallireducens* cultivées en condition anaérobie avec ou sans source de carbone aromatique. Les extraits protéiques, préparés de manière à conduire à un enrichissement en protéines membranaires ou associées à la membrane (préparation d'une fraction protéique de type microsomale, vue dans la le Chapitre 2 de la partie II des résultats), ont été séparés sur gel 1D. Les différentes pistes du gel ont été découpées systématiquement en 96 bandes qui ont ensuite été digérées à la trypsin. Des duplicats biologiques ont été réalisés sur ces préparations. Les protocoles détaillés des préparations d'échantillons sont décrits dans la publication des résultats.

2.2. Analyses par spectrométrie de masse

De manière similaire à l'étude réalisée en Partie I des résultats, Chapitre 2. 2., la taille de la piste de gel et la quantité de protéines déposée ont été adaptées à la sensibilité du système instrumental utilisé pour les analyses LC-MS/MS : le système microHPLC équipé d'une colonne chromatographique de 300 μm de diamètre interne couplé à une trappe ionique HCT Ultra (Bruker Daltonics). Le paramétrage des méthodes chromatographiques et de spectrométrie de masse a été soigneusement réglé pour les besoins de l'étude. Ainsi, la complexité des mélanges peptidiques issus de la digestion de chaque bande de gel 1D nous a conduit à réaliser les séparations chromatographiques avec un gradient de modification de la composition de la phase mobile de pente relativement douce (10-70 % acétonitrile sur 120 minutes). Le paramétrage de l'acquisition du spectromètre de masse a également été réglé pour permettre une quantification par approche spectral counting de qualité tout en conservant une couverture du protéome analysé satisfaisante. Ainsi, la sélection des ions parents à fragmenter a été réalisée sur un mode de « semi-exclusion », c'est-à-dire que chaque ion parent sélectionné a été fragmenté puis exclu de la sélection sur un temps assez court ($\sim 1/3$ de la durée du pic chromatographique correspondant). Ce paramétrage offre la possibilité d'obtenir plusieurs spectres du même ion sans que la totalité des spectres MS/MS ne soit réalisée sur une population trop restreinte d'ions. De plus, le nombre de spectres MS/MS moyennés pour donner le spectre MS/MS final de chaque ion sélectionné a été réglé de manière à permettre l'attribution d'une séquence peptidique pour un nombre maximal de spectres MS/MS avec le couplage microLC-MS/MS utilisé dans cette étude. Il a donc fallu trouver un compromis entre la qualité des spectres MS/MS utilisés pour les identifications (le rapport signal/bruit d'un spectre MS/MS final augmente avec le nombre de spectres MS/MS moyennés pour le constituer) et le temps d'analyse qui leur est consacré (le temps consacré à la constitution d'un spectre MS/MS final augmente avec le nombre de spectres MS/MS moyennés et conduit à une diminution du nombre de spectres MS/MS finaux réalisés dans toute l'analyse). Avec le couplage instrumental utilisé dans notre étude, ce nombre a été optimisé à 6. Enfin, le traitement post-analytique (génération des « peak lists ») a été réalisé sans moyennner les

spectres MS/MS finaux qui ont pu être réalisés sur le même précurseur pour ne pas induire de biais dans la quantification.

L'ensemble des données MS/MS a été soumis à une interrogation Mascot dans une version target-decoy de la banque protéique de *G. metallireducens* téléchargée sur le site du NCBI. Les seuils de score des identifications peptidiques ont été établis pour maintenir un FDR maximal de 1 % sur l'identification des protéines. Les protocoles analytiques sont décrits dans la publication des résultats.

2.3. Quantification des protéines identifiées

La quantification relative des protéines entre les deux conditions de culture a été réalisée par une approche spectral counting (vu en Chapitre 1 1.2.2. de cette partie des résultats). La quantification relative des protéines de *G. metallireducens* cultivé avec une source de carbone aromatique a également été réalisée par une approche spectral counting (vu en Chapitre 1 2.1.2. de cette partie des résultats). Le détail de l'ensemble des traitements est décrit dans la publication des résultats.

L'analyse comparative (benzoate/acetate) de la régulation transcriptionnelle de certains gènes correspondant à des protéines différentiellement exprimées a été réalisée par qRT-PCR et comparée à l'analyse protéomique différentielle par spectral counting.

L'activité des enzymes qui voient leur expression augmenter dans la culture avec une source de carbone aromatique a également été déterminée lorsque cela était possible.

3. Résultats publiés

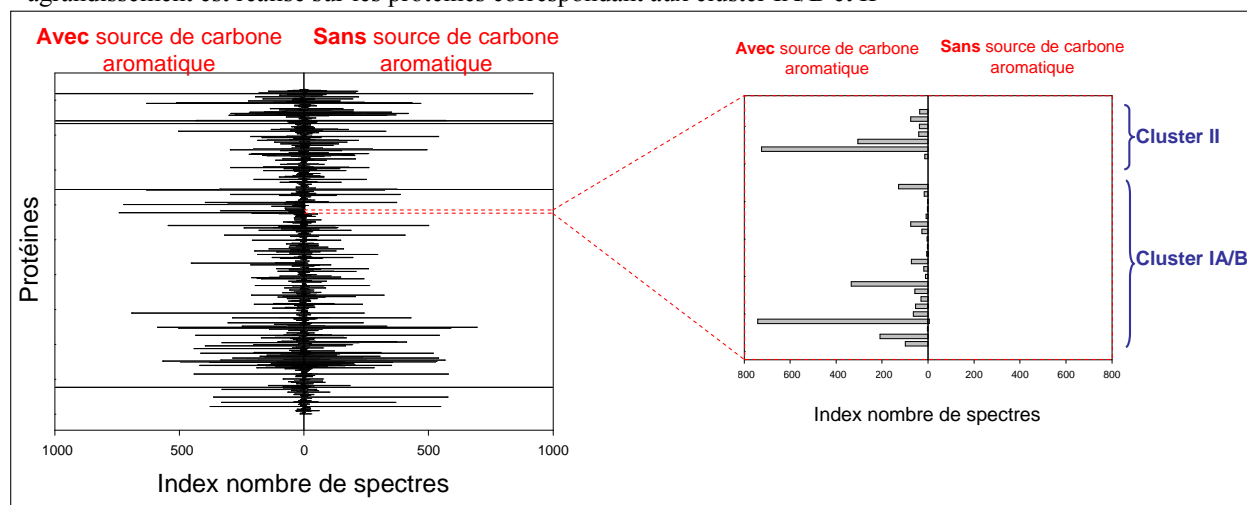
Les résultats obtenus dans cette étude ont fait l'objet d'une publication acceptée dans le journal *Molecular and Cellular Proteomics* en juin 2009.

La stratégie de préparation des échantillons a permis un enrichissement efficace en protéines membranaires ou associées à la membrane puisque sur les 931 protéines identifiées dans l'étude, environ 1/3 contiennent au moins un domaine transmembranaire prédit et/ou sont décrites comme membranaires ou liées à la membrane (et cela malgré l'annotation de localisation cellulaire des protéines très incomplète chez *G. metallireducens*). L'analyse protéomique différentielle par spectral counting a permis de mettre en évidence 130 protéines différentiellement exprimées entre les deux conditions de culture. Parmi ces protéines, nous avons notamment pu démontrer la surexpression, lors de la culture avec une source de carbone aromatique, de la majorité des produits des 2 clusters de gènes (IA/B et II) précédemment décrits comme impliqués dans le métabolisme des composés aromatiques chez *G. metallireducens* (Figure 3). L'intégration des données obtenues à partir i) de l'analyse protéomique et du traitement spectral counting, ii) de l'analyse quantitative sur l'expression des gènes et iii) des mesures d'activités enzymatiques a fourni de nouveaux éléments permettant d'améliorer la compréhension des processus clé de la dégradation des composés aromatiques chez les bactéries anaérobies obligatoires.

Les résultats obtenus ont notamment permis de formuler plusieurs hypothèses sur les particularités de l'étape clé du métabolisme des composés aromatiques chez les bactéries anaérobies obligatoires, la réduction du cycle aromatique, qui sont les suivantes :

- La réduction du cycle aromatique par transfert d'électrons ne semble pas être couplée à l'hydrolyse stoechiométrique de l'ATP comme chez les bactéries anaérobies facultatives.
- Un ensemble de 8 protéines (gi|78194549-41, codées par un cluster de 8 gènes BamBCDEFGHI) surexprimées dans les cellules cultivées en présence de benzoate, dont des homologues n'ont jusque là été identifiées que chez des bactéries anaérobies obligatoires, est proposé comme impliqué dans la réduction du cycle aromatique.
- Le complexe benzoyl-CoA reductase, pour l'instant non caractérisé, semble être associé à la membrane probablement par interaction de certains de ses constituants avec d'autres protéines impliquées dans le mécanisme de transfert d'électrons au cycle aromatique.
- Le constituant supposé du complexe contenant le site actif semble contenir un cofacteur tungstène plutôt que molybdène.

Figure 3 : Comparaison des abondances des protéines identifiées entre les 2 conditions de culture par l'approche spectral counting. L'ensemble des protéines identifiées dans l'étude est distribué sur l'ordonnée du graphique. Un index du nombre de spectres normalisé pour chaque protéine est indiqué sur l'abscisse du graphique. Un agrandissement est réalisé sur les protéines correspondant aux cluster IA/B et II



Differential Membrane Proteome Analysis Reveals Novel Proteins Involved in the Degradation of Aromatic Compounds in *Geobacter metallireducens**[§]

Dimitri Heintz^{‡§}, Sébastien Gallien^{§¶}, Simon Wischgoll^{||}, Anja Kerstin Ullmann^{**}, Christine Schaeffer[¶], Antje Karen Kretzschmar^{**}, Alain van Dorsseleer[¶], and Matthias Boll^{||‡‡}

Aromatic compounds comprise a large class of natural and man-made compounds, many of which are of considerable concern for the environment and human health. In aromatic compound-degrading anaerobic bacteria the central intermediate of aromatic catabolism, benzoyl coenzyme A, is attacked by dearomatizing benzoyl-CoA reductases (BCRs). An ATP-dependent BCR has been characterized in facultative anaerobes. In contrast, a previous analysis of the soluble proteome from the obligately anaerobic model organism *Geobacter metallireducens* identified genes putatively coding for a completely different dearomatizing BCR. The corresponding BamBCDEFGHI complex is predicted to comprise soluble molybdenum or tungsten, selenocysteine, and FeS cluster-containing components. To elucidate key processes involved in the degradation of aromatic compounds in obligately anaerobic bacteria, differential membrane protein abundance levels from *G. metallireducens* grown on benzoate and acetate were determined by the MS-based spectral counting approach. A total of 931 proteins were identified by combining one-dimensional sodium dodecyl sulfate-polyacrylamide gel electrophoresis with liquid chromatography-tandem mass spectrometry. Several membrane-associated proteins involved in the degradation of aromatic compounds were newly identified including proteins with similarities to modules of NiFe/heme b-containing and energy-converting hydrogenases, cytochrome *bd* oxidases, dissimilatory nitrate reductases, and a tungstate ATP-binding cassette transporter system. The transcriptional regulation of differentially expressed genes

was analyzed by quantitative reverse transcription-PCR; in addition benzoate-induced *in vitro* activities of hydrogenase and nitrate reductase were determined. The results obtained provide novel insights into the poorly understood degradation of aromatic compounds in obligately anaerobic bacteria. *Molecular & Cellular Proteomics* 8:2159–2169, 2009.

Aromatic compounds comprise the second most abundant class of natural compounds that can be fully degraded to CO₂ by aerobic and anaerobic microorganisms. In aerobic bacteria and fungi the key reactions involved in the degradation pathways of low molecular weight aromatic growth substrates are catalyzed by oxygenases that have been extensively studied in the past 50 years. In contrast, anaerobic microorganisms use a completely different enzyme inventory for the activation of chemically inert side chains or the dearomatization and cleavage of the aromatic ring. In the past 10–15 years, initial insights into the function of novel enzyme reactions involved in the catabolism of aromatic growth substrates in facultative anaerobes have been obtained (for recent reviews, see Refs. 1–4). In contrast, much less is known about the catabolism of aromatic compounds in obligate anaerobes such as Fe(III)-reducing, sulfate-reducing, or fermenting bacteria.

Benzoyl coenzyme A is a key intermediate in the anaerobic degradation of aromatic compounds in both facultative and obligate anaerobes. It serves as substrate for dearomatizing benzoyl-CoA reductases (BCRs)¹ that dearomatize the aromatic ring by two-electron reduction yielding cyclohexa-1,5-diene-1-carboxyl-CoA (Fig. 1 and Refs. 5–7). A BCR enzyme has so far only been isolated and studied in the denitrifying, facultatively anaerobic *Thauera aromatica* (8). The extremely oxygen-sensitive enzyme has an $\alpha\beta\gamma\delta$ composition and harbors three [4Fe-4S]^{1+/2+} clusters (9). It couples the mecha-

From the [‡]Institut de Biologie Moléculaire des Plantes, CNRS-UPR2357, Université Louis-Pasteur, 67083 Strasbourg, France, ^{||}Institute of Biochemistry, University of Leipzig, 04103 Leipzig, Germany, ^{**}RNomics group, Fraunhofer Institut für Zelltherapie und Immunologie, 04103 Leipzig, Germany, and [¶]Laboratoire de spectrométrie de masse BioOrganique, Institut Pluridisciplinaire Hubert Curien-Département Sciences Analytiques et Interactions Ioniques et Biomoléculaires, Université Louis-Pasteur, CNRS UMR7178, 25 rue Becquerel, 67087 Strasbourg, France

Received, February 5, 2009, and in revised form, May 1, 2009

Published, MCP Papers in Press, June 3, 2009, DOI 10.1074/mcp.M900061-MCP200

¹ The abbreviations used are: BCR, benzoyl-CoA reductase; ABC, ATP-binding cassette; 2D, two-dimensional; 1D, one-dimensional; ETF, electron-transferring flavoprotein; qPCR, quantitative PCR; Bam, benzoic acid metabolism.

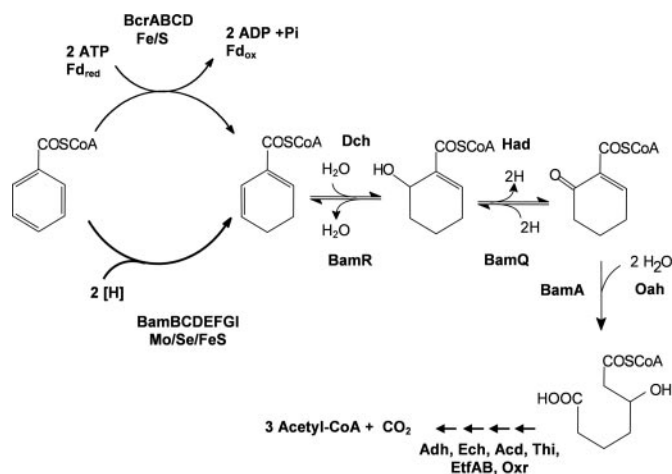


FIG. 1. Gene products involved in the benzoyl-CoA degradation pathway in facultative and obligate anaerobes. Abbreviations for enzymes from facultative anaerobes are: *Bcr*, benzoyl-CoA reductase (ATP-dependent); *Dch*, dienoyl-CoA hydratase; *Had*, 6-hydroxycyclohexenyl-CoA dehydrogenase; *Oah*, 2-oxocyclohexenyl-CoA hydroxylase. The abbreviation for an enzyme from obligate anaerobes is: *Bam*, benzoic acid metabolism. Additional abbreviations are: *Adh*, alcohol dehydrogenase (3-OH-acyl-CoA dehydrogenase); *Ech*, enoyl-CoA hydratase; *Acd*, acyl-CoA dehydrogenase; *Thi*, thiolase; *EtfAB*, electron-transferring flavoprotein subunits A and B; *Oxr*, oxidoreductase homologous to membrane-bound heterodisulfide reductases.

nistically difficult reduction of the aromatic ring to a stoichiometric ATP hydrolysis (10). The coupling of electron transfer to an exergonic reaction is suggested to be essential to overcome the high redox barrier for electron transfer to the aromatic ring. Initial evidence for a mechanism comprising single electron transfer and protonation steps according to the Birch reduction in chemical synthesis has been obtained (11, 12). The further conversion of the dienoyl-CoA product proceeds by enzymes of the so-called benzoyl-CoA degradation pathway yielding the CoA ester of an aliphatic dicarboxylic acid (usually 3-hydroxypimelyl-CoA), which is then further converted to three acetyl-CoA and CO₂ by β -oxidation and a decarboxylating glutaryl-CoA dehydrogenase (Fig. 1). No homologues of ATP-dependent BCR enzymes are present in genomes of aromatic compound-degrading, obligately anaerobic *Geobacter* species or *Syntrophus aciditrophicus* (13, 14).

A previous analysis of the benzoate-induced soluble proteome of the Fe(III)-respiring *Geobacter metallireducens* separated by 2D gel electrophoresis revealed two benzoate-induced gene clusters (IA/B and II with one IA and IB separated by a transposon element (14)). They comprise genes coding for enzymes of the benzoyl-CoA degradation pathway (*bam* (benzoic acid metabolism) genes, clusters IA and II) and β -oxidation reactions (cluster IB). Heterologous expression and characterization of BamY (14), BamR (15), and BamA (16) from several strict anaerobes suggested that the benzoyl-CoA degradation pathways are identical in both facultative and obligate anaerobes (Fig. 1). However, a putative protein complex consisting of eight soluble protein components, Bam-

BCDEFGHI (BamB-I), was suggested to replace the ATP-dependent BCR from facultative anaerobes. This assumption was supported by the finding that the genes coding for similar BamB-I complexes are only present in the genomes of obligately anaerobic bacteria that use aromatic growth substrates (13, 14). The individual proteins of the BamB-I complex show similarities to soluble components of NADH:quinone oxidoreductases (BamGHI), electron-transferring components of hydrogenases (BamC and selenocysteine-containing BamF), soluble heterodisulfide reductases (BamDE), and to molybdenum or tungsten-containing aldehyde:ferredoxin oxidoreductase-like proteins (BamB). In accordance, growth of *G. metallireducens* on benzoate depended on molybdenum/tungsten and selenium, which confirmed the essential role of the BamB-I complex in anaerobic aromatic metabolism of *G. metallireducens* (14).² BamB was considered as the active site-containing component, whereas the BamC-I components were suggested to be involved in electron transfer from an unknown donor to BamB. No ATP-binding motif was found in the BamB-I complex, suggesting that electron transfer to benzoyl-CoA may be independent of ATP hydrolysis. Thus, the presence of additional, so far unknown membrane protein components putatively involved in energy conversion processes was hypothesized (14).

To obtain a more complete picture of the processes and protein components involved in the aromatic metabolism of obligately anaerobic bacteria, the membrane proteome of the model organism *G. metallireducens* grown on benzoate and acetate was analyzed by combining one-dimensional (1D) SDS-PAGE sample fractionation with an LC-MS/MS-based protein identification. The differential protein expression levels and the relative protein abundance levels were determined by the spectral counting approach that is based on the empirical observation that the MS/MS sampling rate of a particular peptide is directly related to the abundance of this peptide in the sample (17–22). The transcriptional regulation of selected genes with annotated function was additionally analyzed by RT-quantitative PCR (qPCR) analysis; where possible, *in vitro* activities of benzoate-induced enzymes were determined.

MATERIALS AND METHODS

Cultivation of *G. metallireducens*—*G. metallireducens* (German Collection of Microorganisms and Cell Cultures (DSMZ) number 7210) was cultured anaerobically with acetate (30 mM) or benzoate (5 mM) as carbon source and Fe(III)-citrate as electron acceptor as described previously (23). The growth was monitored by cell counting using a Neubauer chamber. For all experiments, cells were always harvested in the exponential growth phase at $1.5\text{--}2.5 \times 10^8$ cells ml⁻¹ by centrifugation (20,000 $\times g$) after which they were stored in liquid nitrogen.

Sample Preparation and 1D SDS-PAGE Separation—Frozen cells of *G. metallireducens* grown on either acetate or benzoate as carbon

² D. Heintz, S. Gallien, S. Wischgoll, A. K. Ullmann, C. Schaeffer, A. K. Kretzschmar, A. van Dorsseleer, and M. Boll, unpublished results.

source were suspended in buffer containing 50 mM NH_4HCO_3 , pH 7.8, and 3 mM DTT (1 g of cells in 2 ml of buffer). Cell lysates were obtained by passage through a French pressure cell at 137 megapascals. After centrifugation at $100,000 \times g$ (1 h at 4 °C) the pellet was washed and centrifuged twice using the same buffer. The pellet was suspended in buffer containing 9 M urea, 50 mM NH_4HCO_3 , pH 7.8, 3 mM DTT, and PMSF. After ultracentrifugation the supernatant was discarded, and the remaining pellet was suspended with 1% SDS, 50 mM Tris/HCl, pH 8, and 3 mM DTT and centrifuged again. The supernatant was used for proteome analysis. The protein content was determined using a modified Bradford protocol as described previously (24). Five hundred micrograms of protein resuspended in a solution of 1% SDS, 50 mM Tris-HCl, pH 8.3, and 3 mM DTT were diluted with an equal volume of 1D SDS-PAGE sample buffer (62.5 mM Tris/HCl, pH 8, 2% SDS, 30% glycerol, and 0.01% bromophenol blue (25)) and boiled for 5 min. The samples were cooled and vigorously shaken for 1 h at room temperature. Prior to electrophoresis the samples were centrifuged at $19,000 \times g$ for 5 min at 20 °C. The supernatants were loaded onto a 1D SDS-PAGE gel using a 4% acrylamide stacking gel and a 15% acrylamide resolving gel. After electrophoretic separation the gels were fixed in 50% ethanol and 3% phosphoric acid for 2 h, washed three times with bidistilled water (10 min), and stained overnight with colloidal Coomassie Brilliant Blue (0.08% Coomassie Brilliant Blue G-250, 1.6% *ortho*-phosphoric acid, 8% ammonium sulfate, and 20% methanol (26)).

In-gel Digestion and Mass Spectrometry Analysis—After gel image analysis, the overall separating gel band of each lane containing the different protein bands was excised in 96 slices ($\sim 1 \times 10$ mm) using a custom-built grid. In-gel digestion was performed with an automated protein digestion system (MassPREP Station, Waters). The gel slices were washed twice with 50 μl of 25 mM NH_4HCO_3 and 50 μl of acetonitrile. Cysteine residues were reduced by 50 μl of 10 mM DTT at 57 °C and alkylated by 50 μl of 55 mM iodoacetamide. After dehydration with acetonitrile, the proteins were cleaved in-gel with 10 μl of 12.5 ng μl^{-1} modified porcine trypsin (Promega) in 25 mM NH_4HCO_3 at 37 °C for 16 h. Tryptic peptides were extracted with 60% acetonitrile in 0.5% formic acid followed by a second extraction with 100% acetonitrile. The peptide extracts obtained were analyzed by micro-LC-MS/MS using an Agilent 1100 Series capillary LC system (Agilent Technologies) coupled to a High Capacity Trap Ultra ion trap (Bruker Daltonics). Peptide mixtures were loaded on a Zorbax 300SB-C₁₈ trap column (300 $\mu\text{m} \times 5$ mm; Agilent Technologies) using 0.1% trifluoroacetic acid for 3 min at 50 $\mu\text{l} \text{ min}^{-1}$. After washing, flow was reversed through the trap column, and the peptides were eluted with a gradient of 10–70% acetonitrile in 0.05% trifluoroacetic acid delivered over 120 min at a flow rate of 4 $\mu\text{l} \text{ min}^{-1}$ through a reverse phase capillary column (Zorbax 300SB-C₁₈, 300- μm inner diameter \times 15 cm; Agilent Technologies). The High Capacity Trap Ultra ion trap was externally calibrated with standard compounds. The general mass spectrometric parameters were as follows: capillary voltage, -4000 V; drying gas, 6 liters min^{-1} ; drying temperature, 300 °C. The system was operated by automatic switching between MS and MS/MS modes. MS scanning was performed in the standard-enhanced resolution mode at a scan rate of 8100 $m/z \text{ s}^{-1}$ with an ion charge control of 100,000 in a maximal fill time of 200 ms; five scans were averaged to obtain an MS spectrum. The three most abundant peptides and preferentially doubly charged ions were selected on each MS spectrum for further isolation and fragmentation. The MS/MS scanning was performed in the ultrascan resolution mode at a scan rate of 26,000 $m/z \text{ s}^{-1}$ with an ion charge control of 300,000, and a total of six scans were averaged to obtain the MS/MS spectrum. The complete system was fully controlled by ChemStation Revision B.01.03 (Agilent Technologies) and EsquireControl 6.1 Build 78 (Bruker Dalton-

ics) software. Mass data collected during LC-MS/MS analyses were processed using the software tool DataAnalysis 3.4 Build 169 and converted into *.mgf files.

Protein Identification—The MS/MS data were analyzed using the Mascot 2.2.0 algorithm (Matrix Science, London, UK) for a search against an in-house generated protein database composed of protein sequences of *G. metallireducens* downloaded from NCBI (on December 20, 2007) concatenated with reversed copies of all sequences (2×3532 entries). Spectra were searched with a mass tolerance of 0.5 Da for MS and MS/MS data, allowing a maximum of one missed cleavage with trypsin and with carbamidomethylation of cysteines and oxidation of methionines specified as variable modifications. Protein identifications were validated when at least two peptides with high quality MS/MS spectra (less than five points below the Mascot threshold score of identity at 95% confidence level) were detected. In the case of one-peptide hits, the score of the unique peptide had to be greater (minimal “difference score” of 20) than the 95% significance Mascot threshold. The estimated false discovery rate by searching the target-decoy database (27) was found to lie below 0.8%.

Cellular component gene ontology annotation of *G. metallireducens* was downloaded from the Gene Ontology Annotation Database. *In silico* transmembrane domain prediction was performed using the Phobius Web server (28).

Differential Protein Expression and Relative Protein Abundance—To estimate the differential expression of each protein *i*, the differential absolute protein expression measurement scheme (21, 22, 29) was used. From all high quality MS/MS spectra (less than five points below the Mascot threshold score of identity at 95% confidence level) associated with each validated protein, we calculated the fraction f_i of interpreted spectra for a given protein *i* in the experiment as $f_i = n_i/N$ where n_i is the number of spectra from protein *i* and *N* is the total number of observed MS/MS spectra in the experiment. On the basis of these measures of f_i , the test statistic for differential expression of a protein was calculated as

$$Z_i = \frac{f_{i,1} - f_{i,2}}{\sqrt{f_{i,0}(1 - f_{i,0})/N_1 + f_{i,0}(1 - f_{i,0})/N_2}} \quad (\text{Eq. 1})$$

where the numerator represents the difference in sampled proportions of protein *i* in two proteomics experiments, $f_{i,1} = n_{i,1}/N_1$ and $f_{i,2} = n_{i,2}/N_2$, and the denominator represents the standard error of the difference under the null hypothesis in which the two sampled proportions were drawn from the same underlying distribution with the overall proportion $f_{i,0} = (n_{i,1} + n_{i,2})/(N_1 + N_2)$. The abundance ratio (F_i) of a protein *i* between the two experiments (two growth conditions) was calculated as the ratio of the fraction of interpreted spectra $F_i = f_{i,1}/f_{i,2}$.

For proteins with no associated MS/MS spectra in one of the two cell extracts used, the abundance ratio was arbitrarily set as greater than 150 (no MS/MS spectra obtained in cells grown with acetate) or less than $1/150$ (no MS/MS spectra obtained from cells grown with benzoate). Proteins with $|Z| > 2.58$ (99% significance level) and log ratio ($\log_2 F$) greater than 1 were considered significantly induced during growth on benzoate. Proteins with $|Z| > 2.58$ (99% significance level) and log ratio ($\log_2 F$) less than -1 were considered as significantly repressed during growth on benzoate. In this analysis, the results from two biological replicates for extracts from cells grown on benzoate and acetate were pooled and the Z-score and log ratio for each validated protein were calculated.

For the comparison of protein abundances protein weights (*W*) were taken into consideration (18); the relative protein abundance for each of the *k* proteins was calculated as follows.

$$Fr_i = \frac{n_{i,1}/W_i}{\sum_{i=1}^k n_{i,1}/W_i} \quad (\text{Eq. 2})$$

The relative protein abundance of succinate dehydrogenase (NCBI database accession no. gi78194855) was set to 100 and used as reference for the calculation of each relative protein abundance.

Quantitative PCR Analysis—One microgram of total RNA from *G. metallireducens* grown on acetate or benzoate was reverse transcribed using random hexamer primers and the High Capacity® Reverse Transcription kit (Applied Biosystems, Foster City, CA). The cDNA was diluted 1:12.5 and served as the template for qPCR analysis using the TaqMan® 9700 System (Applied Biosystems) with Fast SYBR® Green Master Mix (Applied Biosystems) according to the manufacturer's protocol. Melting curve and agarose gel analyses were used to confirm the specificity of the amplification reactions. All mRNA quantification data were normalized to mRNA levels of a housekeeping enzyme (succinate dehydrogenase). As a positive control induction of *bamF* (coding for a putative methyl viologen-reducing hydrogenase, δ subunit) in bacteria grown on benzoate was verified (induction, ~33-fold). Primer sequences for quantitative real time expression analysis (qRT-PCR) of mRNA levels of different enzymes are listed in supplemental Data 1. Genes were considered significantly induced or repressed during growth on benzoate if the log ratio (\log_2) of the mRNA abundance between the two growth conditions was greater than 1 or less than -1.

Determination of Enzyme Activities—Hydrogenase and nitrate reductase activities were analyzed at 30 °C under anaerobic conditions using oxidized benzyl viologen as artificial electron acceptor (hydrogenase) or the reduced form as electron acceptor (nitrate reductase). Substoichiometric reduction of benzyl viologen was anaerobically carried out by additions of dithionite from a 50 mM stock solution to $A_{600} = 1.4$ – 1.6 in the nitrate reductase and to $A_{600} = 0.05$ in the hydrogenase assay. The test buffer contained 1 mM benzyl viologen, 50 mM Tris/HCl, pH 7.8, and 100–150 μg of solubilized membrane protein fraction obtained after ultracentrifugation of crude extracts from *G. metallireducens* grown on acetate or benzoate. The reaction was carried out in a gas-tight sealed cuvette (400 μl) under a nitrogen atmosphere. For the hydrogenase assay, the reaction was started by injection of 100 μl of H_2 via a gas-tight syringe followed by gentle shaking of the cuvette. For the nitrate reductase activity test, the reaction was started by addition of 0.5–2.5 mM sodium nitrate. As a control, the membrane-bound succinate dehydrogenase activity was determined spectrophotometrically using ferricenium as electron acceptor. Cytochrome *bd* oxidase activity was tested by oxygen determination using an optode (Fibox3, PreSens Precision Sensing GmbH, Regensburg, Germany) and an oxygen sensor spot in a 1.5-ml gas-tight glass vial at room temperature. Menaquinol (0.5–1.5 mM), NADH (0.5–1 mM), acetate (10–20 mM), and benzoate (1–2 mM) were tested as electron donors for cell extracts/whole cells.

RESULTS AND DISCUSSION

Differential Analysis of the Membrane Proteome of *G. metallireducens*

In a recent study we analyzed and compared the soluble proteomes of *G. metallireducens* cells grown on benzoate and acetate by 2D gel electrophoresis coupled to mass spectrometric analysis. By combining the data obtained with results from reverse transcription-PCR, 44 benzoate-induced genes were found to be organized in the gene clusters IA/B and II (14). They comprise genes for all predicted enzymes involved

in the conversion of benzoyl-CoA to acetyl-CoA and CO_2 ; an eight gene-containing cluster (*bamB*–I) was proposed to code for a novel BCR complex.

To reveal benzoate-induced membrane-bound/associated proteins, the spectral counting approach was performed with tryptic digests from membrane preparations obtained from cells grown on benzoate and acetate. After thorough removal of soluble or loosely membrane-attached proteins by washing/extraction steps, the membrane fractions were separated by 1D SDS-PAGE, digested in-gel, and analyzed by LC-MS/MS. Although the urea treatment step is generally accepted for membrane protein preparations, one has to consider that the chaotropic urea agent may break interactions between some non-integral membrane protein subunits, which may result in a loss of some proteins. MS/MS spectra analysis identified 931 proteins in the genome of *G. metallireducens* with a very low false positive rate (below 0.8% as estimated by searching a target-decoy database (27)). Among these, 804 proteins were identified with at least two peptides, and 127 proteins were identified with one peptide; for the latter the spectra associated with the best score identifications are shown in supplemental Data 2. To date, 325 of these 931 proteins have a cellular component gene ontology annotation (indicated in supplemental Data 7). 53% of them (173 proteins) are annotated as membrane-bound/associated proteins, which represents 40% of all annotated membrane-bound/associated proteins from *G. metallireducens*. In addition, *in silico* transmembrane domain prediction allowed the identification of 250 proteins with at least one transmembrane domain (supplemental Data 7). About one-third of all proteins identified in this study (303 proteins) contain at least one predicted transmembrane domain and/or are predicted as membrane-bound/associated despite the very incomplete cellular component annotation of *G. metallireducens*.

The differential absolute protein expression measurement method (21, 22, 29) was used to identify proteins with a significant differential expression by performing a test statistic (*Z*-test). To estimate abundance changes during growth on benzoate and acetate, an abundance ratio was calculated for each protein. To avoid overconsideration of small changes in protein abundances and of changes in the abundances of very minor proteins, we imposed $|Z|$ to be greater than 2.58 (corresponding to 99% confidence) and the \log_2 abundance ratio to be greater than 1 to consider a protein as benzoate-induced. To estimate potential biological and analytical variabilities, preparations of the membrane fractions and analysis of the tryptic peptides were carried out with extracts from two different cell batches both harvested in the exponential growth phase. The lists of all data obtained are available in supplemental Data 2–6; the spectral counting data are shown in supplemental Data 7. Using the framework described above, spectral counting analysis revealed 130 proteins with an altered expression levels in cells grown on benzoate and acetate. From the 100 benzoate-induced proteins, 23 were

TABLE I

Abundance of membrane protein components involved in central energy metabolism in *G. metallireducens* cells grown on benzoate and acetate

Protein annotation	Molecular mass	NCBI database accession no.	log ₂ ratio of protein abundance (benzoate/acetate)	Relative protein abundance (benzoate)
	<i>kDa</i>	<i>gi</i>		
Proton-translocating NADH:quinone oxidoreductase, chain N	52.1	78195787	-0.7	8.4
Proton-translocating NADH:quinone oxidoreductase, chain M	57.2	78195788	-0.1	37.4
NADH:plastoquinone oxidoreductase, chain 5	73.3	78195789	-0.4	26.7
NADH:ubiquinone oxidoreductase, chain 4L	11.1	78195790	-0.6	42.3
NADH:ubiquinone/plastoquinone oxidoreductase, chain 6	18.2	78195791	-0.1	74.8
NADH:quinone oxidoreductase, chain I	15.1	78195792	-0.1	51.4
Respiratory chain NADH dehydrogenase, subunit 1	37.5	78195793	-0.5	55.5
NADH dehydrogenase I, G subunit, putative	89.0	78195794	-0.2	17.5
NADH dehydrogenase I, F subunit	64.3	78195795	-0.3	21.0
NADH dehydrogenase (ubiquinone), 24-kDa subunit	19.2	78195796	-0.1	9.3
NADH dehydrogenase I, D subunit	43.7	78195797	-0.7	35.0
NADH (or F ₄₂₀ H ₂) dehydrogenase, subunit C	18.5	78195798	-0.3	60.6
NADH dehydrogenase (ubiquinone), 20-kDa subunit	18.3	78195799	0.0	34.3
NADH:ubiquinone/plastoquinone oxidoreductase, chain 3	13.3	78195800	-0.6	38.5
Fumarate reductase, iron-sulfur protein	27.2	78194853	1.0	164.4
Succinate dehydrogenase or fumarate reductase, flavoprotein	70.7	78194854	0.5	330.6
Succinate dehydrogenase, cytochrome <i>b</i> ₅₅₈ subunit	24.0	78194855	-0.1	100.0
H ⁺ -Transporting two-sector ATPase, A subunit	25.5	78195804	0.6	4.5
ATP synthase F ₀ , C subunit	9.4	78195805	-1.0	73.9
H ⁺ -transporting two-sector ATPase, δ/ε subunit	15.2	78195850	-0.1	17.7
ATP synthase F ₁ , β subunit	51.0	78195851	0.4	87.7
H ⁺ -Transporting two-sector ATPase, γ subunit	31.9	78195852	0.1	38.7
ATP synthase F ₁ , α subunit	54.6	78195853	0.2	66.5
H ⁺ -Transporting two-sector ATPase, δ subunit	19.4	78195854	-1.0	45.0
H ⁺ -Transporting two-sector ATPase, B/B' subunit	22.6	78195855	-1.0	22.1
H ⁺ -Transporting two-sector ATPase, B/B' subunit	15.6	78195856	-0.9	31.5
Inorganic H ⁺ -pyrophosphatase	70.1	78195685	0.0	126.2

identified as membrane-bound/associated proteins encoded by genes of the *bam* clusters. Further 18 newly identified proteins with a minimum of 4.5-fold change during growth on benzoate were selected for further investigations.

Detection of Membrane Proteins Involved in Overall Energy Metabolism

To demonstrate the efficiency and comparability of the membrane proteome analysis in cells grown on benzoate and acetate, the observed abundance ratio of components from membrane complexes involved in general energy metabolism are exemplarily discussed in the following. A non-differential abundance of the housekeeping proteins in cells grown on benzoate and acetate was expected.

All predicted components of NADH:menaquinone oxidoreductase, succinate dehydrogenase, and H⁺-F₀F₁-ATP synthase were identified in cells grown on benzoate and acetate at almost equal abundances (Table I). None of the 27 subunits of the three complexes passed the established filtering thresholds to consider a protein as differentially expressed during growth on benzoate and acetate. These results indicate that both cell types were in comparable physiological states (exponential growth phase) and that differences in abundance of other membrane proteins cannot be

assigned to differences in cultivation or sample preparation and demonstrate the reliability of the spectral counting approach for differential membrane proteome analysis.

The genome of *G. metallireducens* contains three copies of a gene cluster annotated as components of (mena)quinol: cytochrome *c* oxidoreductases (usually termed cytochrome *bc*₁ complexes consisting of four subunits). However, neither of the corresponding components was identified in cells grown on benzoate or acetate. This finding indicates that no such respiratory complex is involved in electron transfer from menaquinol to the terminal acceptor Fe(III) and that a so far unknown menaquinol-oxidizing membrane component should exist. Thus, the predicted cytochrome *bc*₁ complexes are considered to exhibit other functions. Recently a special role of such a complex (gi78194581–gi78194584) in *p*-cresol degradation was suggested (30, 31). The complex was proposed to mediate electron transfer from the *p*-cresol methylhydroxylase reaction to the menaquinone pool. Accordingly the genes of this complex were only induced during growth on *p*-cresol but not during growth on benzoate or acetate.

Interestingly in cells grown on both acetate and benzoate the gene putatively coding for a proton-translocating pyrophosphatase was identified in equal amounts (Table I). This finding is in line with the fact that during growth on both

TABLE II

Differential identification of BamB-I components in the membrane and soluble protein fraction

Semiquantitative identification in the soluble fraction was carried in a previous analysis after separation on 2D gels (14).

NCBI database accession no.	Gene product	Amino acid similarities to	Relative abundance	
			Membrane fraction ^a	Soluble fraction
<i>gi</i>				
78194549	BamB	Aldehyde ferredoxin oxidoreductases	1.6	+++
78194548	BamC	FeS subunit of hydrogenases	<1	+++
78194547	BamD	Soluble subunit of heterodisulfide reductases	<1	++
78194546	BamE	Soluble subunit of heterodisulfide reductases	<1	+++
78194545	BamF	SeCys-containing subunit of hydrogenases	19	-
78194543	BamG	NADH:ubiquinone oxidoreductases (NuoG)	12	-
78194542	BamH	NADH:ubiquinone oxidoreductases (NuoF)	<1	-
78194541	BamI	NADH:ubiquinone oxidoreductases (NuoE)	1	+++

^a Arbitrary units; the abundance of the structural gene of succinate dehydrogenase (gi78194855) was set to 100.

benzoate and acetate AMP + PP_i-forming carboxylic acid-CoA ligases are involved in initial ATP-dependent benzoate/acetate activations. Obviously obligate anaerobes commonly couple the exergonic hydrolysis of PP_i to H⁺/Na⁺ ion translocation across the cytoplasmic membrane as it has been demonstrated for the fermenting *Syntrophus gentianae* (32).

Identification of Proteins Encoded by the Benzoate-induced Gene Clusters IA/B and II

A previous study revealed the benzoate-induced gene clusters IA/B and II (14). Analysis of the soluble proteomes only identified 13 of the 44 predicted benzoate-induced gene products, although RT-PCR analysis indicated that the 44 genes were all induced during growth on benzoate. The reason for this discrepancy could be due to (i) the presence of membrane/membrane-associated proteins that were not suitable for 2D gel electrophoretic analysis, (ii) very low expression level of these proteins, or (iii) posttranscriptional regulatory processes. In contrast, the membrane proteome analysis of this work identified 23 of the 44 putative gene products as benzoate-induced.

In Table II, the differential identification of individual components of the putative benzoyl-CoA dearomatizing BamB-I complex in the soluble and membrane proteome is shown. To compare abundances of the proteins in cells grown on benzoate, we additionally calculated the relative protein abundance of each protein by normalizing the spectral count by the protein molecular weight (see "Material and Methods"). BamBCDE and BamI had been identified in the soluble fraction in the previous study (14).² These components were also identified in the membrane fraction of benzoate-grown cells albeit at low (BamB) or very low amounts (BamC-E and BamI; Table II). Although topology prediction annotated the selenocysteine-containing BamF and the BamG as soluble proteins, they were only identified at relatively high abundance in the membrane fraction, indicating that at least these components of the putative BamB-I complex were attached to the membrane. Obviously most of the BamB-E and BamHI components were removed during the multiple washing steps of

sample preparation. In summary, the results obtained suggest that the predicted BamB-I complex is at least partly associated to the membrane with BamFG apparently showing a higher affinity to the membrane than the other components. As expected components of the BamB-I complex were absent in the membrane fraction of cells grown on acetate (no spectrum observed).

The benzoate-induced cluster IB contains a number of genes coding for enzymes involved in β -oxidation reactions of aromatic catabolism, nine of which were identified as benzoate-induced in the membrane fraction (supplemental Data 2, 5, and 6). It has been reported that acyl-CoA dehydrogenases and their *in vivo* electron acceptors, electron-transferring flavoproteins (ETFs), are often found to be attached to the membrane by interaction with a membrane-bound ETF:quinone oxidoreductase (33). In accordance, two subunits of ETFs (gi78194527-8) and glutaryl-CoA dehydrogenase (BamM; gi78194537) were identified as benzoate-induced in both the soluble and the membrane proteome (14). As *G. metallireducens* contains no genes coding for homologues of ETF:quinone oxidoreductases, the primary electron-accepting membrane protein of reduced ETF was unclear. It has been proposed that *oxr* genes (gi78194527 and gi78194532), located adjacent to the genes coding for two subunits of ETF (*etfAB*), code for an alternative ETF:quinone oxidoreductase complex. Indeed both *Oxr* copies were identified as benzoate-induced membrane proteins. They show similarities to heme b/FeS cluster-containing, membrane-bound heterodisulfide reductases from *Methanosarcinales* species. Topology prediction revealed the presence of five to six transmembrane helices, which explains why they were not identified after 2D gel electrophoretic separation in the previous study. The role of benzoate-induced proteins, annotated as two subunits of succinyl-CoA synthetases, identified in both the soluble and membrane fractions is unknown so far.

Identification of Novel Benzoate-induced Proteins

Some of the newly identified proteins showed similarities to components of membrane-bound hydrogenases, cytochrome

bd oxidase, nitrate reductase, transporter-related proteins, and a number of proteins with unclear/unknown functions (Table III and supplemental Data 2, 5, and 6). To verify the induction of the genes coding for benzoate-induced membrane proteins, the transcriptional regulation of selected genes was analyzed. For this purpose total RNA was isolated from cells grown on benzoate and acetate in the exponential growth phase and converted to cDNA by reverse transcriptase reactions, respectively. Quantitative analysis of gene expression was performed by RT-qPCR using appropriate oligonucleotide primer pairs. Primers for amplifying cDNA from the structural genes of succinate dehydrogenase and inorganic H⁺-pyrophosphatase served as positive controls, which are expressed in cells grown on benzoate and acetate cells at equal amounts (Table III). Primers amplifying cDNA from two benzoate genes, *bamB* and *bamF* (Table II and Ref. 14), served as positive control for benzoate-induced genes (Table III). An intergenic, non-coding region served as negative control as described previously (14).

Overall the proteomics and transcription analysis data are in good agreement. Of 29 benzoate-induced genes of the clusters IA/B that were previously identified by RT-PCR (14), 23 of the corresponding gene products (~80%) were also identified in this work at higher abundance in cells grown on benzoate (supplemental Data 7). In addition, a further 19 benzoate-induced genes were newly identified by RT-qPCR analysis in the study (Table III). For 11 of these proteins, proteomics and RT-qPCR data were in full agreement. No opposite significant change between mRNA and protein levels was observed. However, a few genes showed a significant change only at the mRNA or only at the protein level. These cases are presented in Table III and are discussed in the following. In the case of NiFe/heme b hydrogenase and nitrate reductase results from *in vitro* enzyme activity determinations are presented (Table IV); again succinate dehydrogenase activity measurements served as a control.

NiFe/Heme b-containing Hydrogenase—All three subunits of a putative NiFe/heme B-containing hydrogenase complex (gi78195776–78) were identified at a clearly higher abundance in cells grown on benzoate (Table III). Using oligonucleotide primers deduced from the three subunits of hydrogenase, a clear induction of gene transcription was observed with cDNA from cells grown on benzoate when compared with cDNA from acetate-grown cells (Table III). Using a spectrophotometric assay monitoring the H₂ and cell extract-dependent reduction of benzyl viologen, hydrogenase activities were determined mainly in the membrane protein fraction of extracts from cells grown on benzoate (Table IV). The activity was 15 times higher than in extract from cells grown on acetate; this is in good accordance with data obtained from proteome and RT-qPCR analysis. The genome of *G. metallireducens* contains two further gene clusters coding for putative soluble hydrogenases (e.g. gi78193586 and gi78195766 annotated as the large NiFe-containing subunit of the putative

hydrogenase complexes). However, using oligonucleotide primers for amplifying DNA fragments of the genes putatively coding for the large subunits, expression was negligible and non-differential in cells grown on benzoate and acetate (not shown). Thus, only one of the three gene clusters in the genome of *G. metallireducens* that putatively code for hydrogenases was found to be induced during growth on benzoate.

Membrane-bound three subunit-containing NiFe/heme b hydrogenases usually catalyze the oxidation of dihydrogen to protons, and reducing equivalents are transferred to quinones (menaquinone in the case of *G. metallireducens*). The induction of such a hydrogenase during growth on benzoate was not expected. A possible function as a low potential electron donor for benzoyl-CoA reduction is rather unlikely as the presence of the heme b-containing transmembrane subunit rather suggests electron transfer to menaquinone. The question arises whether the enzyme uses external or endogenously produced H₂ as substrate *in vivo*. Notably the gas phase of the culture medium did not contain dihydrogen.

Energy-converting Hydrogenases—Next to the 14 clustered genes annotated as subunits of complex I of the respiratory chain (Table I), an additional protein assigned to a putative NADH:quinone oxidoreductase subunit was identified as more abundant in the membrane fraction in cells grown on benzoate (gi78192847). The gene belongs to a cluster comprising seven open reading frames of which six are annotated as putative subunits of NADH:quinone oxidoreductases/energy-converting hydrogenases and one is annotated as a transcriptional regulator (Table III). Two additional products from this cluster were identified with an abundance ratio above the threshold (gi78192845 and gi78192848). However, the number of MS/MS spectra was too low to consider them as significantly differentially expressed using the highly stringent criteria established in this study. Topology prediction revealed that gi78192848–51 contain 42 highly hydrophobic transmembrane domains, which may explain the low recovery in the MS analysis. Highly hydrophobic proteins have been reported to escape analysis due to the lack of tryptic digestion sites close to the membrane domains (34). Interestingly an additional highly similar gene cluster is present in the genome (gi78195048–53) of which three proteins were identified (Table III). To further test the induction of both gene clusters during growth on benzoate, the induction of selected genes of each cluster was tested by RT-qPCR analysis. The results obtained indicate that transcription of genes from cluster gi78195048–53 was clearly induced during growth on benzoate, whereas expression of cluster gi78192845–51 was moderately induced (Table III). The similarities between the six genes of the two newly identified benzoate-induced gene clusters were highest with putative proteins from other *Geobacter* species and related Deltaproteobacteria (50–65% amino acid sequence identity). When compared with enzymes from Enterobacteriaceae or methanogens, similarities were clearly higher to components of energy converting-hydroge-

TABLE III

Analysis of membrane-bound/associated proteins that are more abundant in cells grown on benzoate and transcriptional analysis of selected genes induced during growth on benzoate

Only proteins with annotated functions are shown; proteins are grouped according to the clustering of the corresponding genes. Proteins with bold data met the spectral counting criteria and are considered as induced during growth on benzoate. Gene induction was determined by RT-qPCR. Proteins marked by a check in the right column are considered as induced during growth on benzoate. The assignment is based on either protein abundance (additionally marked by checks) or by mRNA abundance (log ratio >1.1). Dashes indicate protein abundances below the criteria for unambiguous protein identification. TMH, number of transmembrane helices; ND, not determined.

Protein annotation in genome	Molecular weight	gi	TMH	Protein abundance			mRNA abundance		
				Benzoate ^a	log ratio (benzoate/acetate)	Z (benzoate/acetate)	Induced in benzoate growth condition	log ratio (benzoate/acetate)	Induced in benzoate growth condition
Succinate dehydrogenase subunit C (control 1)	24,012	78194855	5	100.00	-0.1	1.25		0.0	
Inorganic H ⁺ -pyrophosphatase (control 2)	70,100	78195685	16	126.20	0.0	0.80		0.1	
BamB (positive control 1)	73,802	78194549	0	1.65	>7.2	4.14	✓	> 5.3	✓
BamF (positive control 2)	23,900	78194545	0	18.83	>7.2	8.71	✓	> 5.3	✓
Nitrate reductase, α subunit	134,326	78192805	0	17.40	2.7	14.40	✓	2.9	✓
Nitrate reductase, β subunit	54,118	78192806	0	10.77	3.2	7.70	✓	1.7	✓
Nitrate reductase, δ subunit	22,822	78192807	0	—	—	—	ND	—	✓
Nitrate reductase, γ subunit	26,501	78192808	5	3.63	2.2	2.62	✓	4.0	✓
Complex I/energy-converting hydrogenase	27,614	78192845	0	2.78	1.9	2.13	✓	1.1	✓
Transcriptional regulator, IclR family	29,438	78192846	0	7.18	-0.2	0.54			
Complex I/energy-converting hydrogenase	54,267	78192847	0	1.77	>7.2	3.68	✓	1.6	✓
Complex I/energy-converting hydrogenase	51,165	78192848	11	0.63	>7.2	2.12			✓
Complex I/energy-converting hydrogenase	22,810	78192849	7	—	—	—	ND	1.2	✓
Complex I/energy-converting hydrogenase	32,014	78192850	8	—	—	—	ND	2.3	✓
Complex I/energy-converting hydrogenase	70,567	78192851	16	—	—	—	ND	1.0	✓
Complex I/energy-converting hydrogenase	26,303	78195048	0	2.19	0.2	0.30		1.6	✓
Complex I/energy-converting hydrogenase	55,236	78195049	0	2.55	0.2	0.48		2.1	✓
Complex I/energy-converting hydrogenase	51,177	78195050	13	—	—	—	ND		✓
Complex I/energy-converting hydrogenase	23,043	78195051	7	0.28	-1.1	0.67		1.7	✓
Complex I/energy-converting hydrogenase	32,263	78195052	7	—	—	—	ND		✓
Complex I/energy-converting hydrogenase	70,158	78195053	14	—	—	—	ND	2.5	✓
Na ⁺ /solute symporter	61,848	78193214	13	50.63	5.2	20.24	✓	5.2	✓
Protein of unknown function DUF485	11,918	78193215	2	13.43	>7.2	4.75	✓		✓
Tungstate ABC transporter-related protein	26,615	78193512	0	3.61	3.8	3.30	✓	1.6	✓
Binding protein inner membrane component	23,847	78193513	5	1.34	1.2	1.00	✓	2.0	✓
Extracellular tungstate-binding protein	29,161	78193514	0	9.88	2.8	5.10	✓		✓
Protein of unknown function DUF214	91,594	78193984	11	35.02	5.2	20.50	✓	2.2	✓
ABC transporter-related protein	24,352	78193985	0	33.39	6.8	10.58	✓	1.3	✓
AMP-dependent synthetase and ligase	91,691	78194077	0	3.35	3.9	5.95	✓		✓
Cytochrome <i>bd</i> ubiquinol oxidase, subunit II	37,435	78194391	8	5.82	>7.2	5.54	✓		✓
Cytochrome <i>bd</i> ubiquinol oxidase, subunit I	49,956	78194392	9	14.23	3.8	9.04	✓	-0.3	✓
Transcriptional regulator, BadM/Rrf2 family	14,604	78194393	0	6.58	>7.2	3.68	✓	-0.5	✓
Ruberythrin	19,131	78195684	0	6.02	3.0	3.35	✓	-1.0	✓
NiFe hydrogenase, cytochrome <i>b</i> subunit	25,101	78195776	4	1.79	>7.2	2.51	✓	3.4	✓
NiFe hydrogenase, large subunit	62,264	78195777	0	8.12	>7.2	8.45	✓	3.2	✓
NiFe hydrogenase, small subunit	40,167	78195778	0	3.83	>7.2	4.66	✓	4.2	✓

^a Arbitrary units; the abundance of the structural gene of succinate dehydrogenase (gj78194855) was set to 100.

TABLE IV
Specific activities of hydrogenase and nitrate reductase in cell extracts of *G. metallireducens*

Activity was determined spectrophotometrically using reduced benzyl viologen as electron donor (nitrate reductase) or the oxidized form as electron acceptor (hydrogenase). For the succinate dehydrogenase control, ferricenium was used as electron acceptor. One milliunit is referred to as conversion of $1 \text{ nmol min}^{-1} \text{ H}_2/\text{nitrate/succinate}$, respectively. Mean values \pm S.D. from three independent determinations are presented.

Enzyme	Extracts				-Fold increase benzoate/acetate (membrane fraction)
	Benzoate-grown cells		Acetate-grown cells		
	Soluble	Membrane	Soluble	Membrane	
	<i>milliunits mg⁻¹</i>				
Hydrogenase	47 \pm 5	157 \pm 12	10 \pm 4	10 \pm 2	16
Nitrate reductase	146 \pm 8	450 \pm 50	48 \pm 12	125 \pm 20	4
Succinate dehydrogenase (control)	155 \pm 40	525 \pm 70	270 \pm 50	774 \pm 200	0.7

nases than to components of related complex I of the respiratory chain.

Energy-converting hydrogenases usually couple the exergonic reduction of protons to H_2 (e.g. with formate as electron donor) to the transport of H^+/Na^+ across the cytoplasmic membrane (35). They usually consist of two soluble components and two or more integral membrane protein components and are homologous to the proton-translocating/energy-conserving modules of NADH:quinone oxidoreductases. Examples are hydrogenases 3 and 4 from *Escherichia coli* or the Ech hydrogenases from methanogens. Notably in *G. metallireducens* the benzoate-induced gene clusters putatively coding for energy-converting complexes cannot be assigned to "real" hydrogenases as the deduced large subunit (e.g. gi78192847) does not contain the N- and C-terminal conserved CXXC motif, which is involved in NiFe cofactor binding (35). In *G. metallireducens* such an energy-conserving complex would represent an attractive candidate for a benzoate-induced membrane complex that is involved in a membrane potential-driven electron transport from an unknown donor to BCR.

Cytochrome *bd* Oxidase and Rubrerythrin—Two membrane proteins annotated as subunits of cytochrome *bd* oxidases were identified at clearly higher abundance in cells grown on benzoate when compared with cells grown on acetate (Table III). In contrast, RT-qPCR analysis revealed that the levels of the corresponding mRNAs were almost equal in cells grown on benzoate and acetate (Table III). This discrepancy suggests the presence of additional posttranscriptional regulatory elements. Attempts to verify the induction of cytochrome *bd* oxidase by *in vitro* activity measurements were not feasible because of very high background reactions caused by the presence of high amounts of Fe(II), which was derived either from the growth medium or from the numerous cytochromes in the dark reddish cell extracts.

Cytochrome *bd* oxidases usually catalyze the reduction of dioxygen by (mena)quinol with a very high affinity to dioxygen (36, 37). When part of a respiratory chain, such enzymes serve as terminal oxidases and are usually expressed at low oxygen concentrations in some aerobic organisms. In obligate anaerobes they rather serve as powerful oxygen-scavenging en-

zymes for protection from oxidative stress. It is conceivable that the obligately anaerobic *G. metallireducens* has an increased demand for an oxygen detoxification system during growth on benzoate for protection of the assumed extremely oxygen-sensitive BCR complex. The induction of an oxygen-scavenging enzyme during anaerobic growth on an aromatic substrate has recently been demonstrated in the facultatively anaerobic *T. aromatica*: here a benzoate-induced dienoyl-CoA oxidase with a high affinity to dioxygen as electron acceptor was identified (38). The requirement for increased oxygen protection when an extremely oxygen-sensitive enzyme complex has to be synthesized has been studied extensively in the case of nitrogenase (39).

Next to the cytochrome *bd* oxidase, rubrerythrin, another protein supposed to be involved in oxygen protection (40), was identified at clearly higher abundance in cells grown on benzoate (log ratio >3 ; Table III). But as in the case of cytochrome *bd* oxidase, transcription of the gene was similarly induced in cells grown on benzoate and acetate. Thus, it appears that the regulatory pattern of both putative oxygen-scavenging proteins follows similar principles probably involving posttranscriptional regulation processes.

Dissimilatory Nitrate Reductase—Surprisingly the products of three genes putatively coding for a membrane-bound, four subunit-containing dissimilatory nitrate reductase complex were identified at clearly higher abundance in cells grown on benzoate than in cells grown on acetate (Table III). In accordance, a clear induction of three genes during growth on benzoate was verified by RT-qPCR analysis. Moreover membrane extracts from cells grown on benzoate exhibited a 4-fold higher nitrate reductase activity than similarly prepared extracts from cells grown on acetate. Together these results indicate an up-regulation on the transcriptional level, although cells were always grown with Fe(III)-citrate as terminal electron acceptor. On the first view the induction of dissimilatory nitrate reductase during growth on benzoate appears curious. A possible explanation is based on the assumption that both nitrate reductase and BamB contain a molybdo-/tungstopterin cofactor. Thus, growth on an aromatic growth substrate is expected to induce the genes coding for enzymes involved in molybdenum/tungsten uptake and molybdo-/tungstopterin

cofactor synthesis. A co-induction of several molybdo-/tungstopterin cofactor-containing enzymes when molybdo-/tungstopterin cofactor synthesis is induced is conceivable. *E.g.* the *modA* gene codes for a molybdenum-sensing transcriptional regulator of genes involved in molybdenum uptake, synthesis of the molybdenum/tungsten cofactor, and molybdenum/tungsten cofactor-containing polypeptides (41).

Tungstate Uptake System—The products of the benzoate-induced genes gi78193512 and gi78193514 belong to a cluster of genes coding for a putative ABC transporter system with the former representing the gene coding for the nucleotide binding component and the latter gene coding for the periplasmic binding protein (Table III). The third permease component of a typical ABC transporter system (gi78193513) was identified with a too low number of MS/MS spectra to pass the Z-test (despite an abundance ratio above threshold) most probably because of its extreme hydrophobicity (more than 60% of the amino acids are predicted to be involved in transmembrane helices formation). RT-qPCR analysis confirmed the induction of the corresponding genes (Table III). The putative periplasmic binding protein component showed exceptionally high similarities to tungstate-binding proteins of ABC transporters referred to as TupA (42) from several bacteria (up to 87% amino acid sequence identity to other *Geobacter* species and up to 56% identity to annotated TupA proteins from *Ralstonia* species). In particular the unique typical features of TupA proteins that distinguish them from molybdate-binding proteins were identified (*e.g.* the conserved TTTS motif near the N terminus and the SRGD_XSGT motif (43)). The genes coding for the putative tungstate uptake system are part of a cluster containing several further genes coding for annotated enzymes involved in molybdo-/tungstopterin cofactor biosynthesis (gi78193508–19). Notably both the molybdo- and tungstopterin cofactors are usually synthesized by the same set of enzymes, whereas molybdate and tungstate uptake systems are highly specific (42).

The identification of a putative tungstate uptake system is remarkable as BamB, the proposed active site-containing component of benzoyl-CoA reductase from *G. metallireducens*, is similar to aldehyde:ferredoxin oxidoreductases that also often contain a tungstopterin cofactor (44, 45). Consequently BamB is rather supposed to be a tungsten rather than a molybdenum cofactor-containing protein.

Further Newly Identified Membrane Proteins That Were More Abundant in Cells Grown on Benzoate—A number of further benzoate-induced proteins were identified in the membrane fraction (supplemental data). Although many of them code for unknown proteins, some of them show similarities to other proteins (Table III and supplemental data). *E.g.* next to the putative ABC transporter for tungstate uptake two additional benzoate-induced transporter related proteins, annotated as components of a Na⁺/proline transporter (gi78193214–5) and an additional ABC transporter (gi78193984–5), were identified. RT-qPCR analysis con-

firmed a high induction of the corresponding genes (Table III). Furthermore a benzoate-induced membrane-associated, AMP-dependent ligase was found at higher abundance in cells grown on benzoate that was also slightly induced on the transcriptional level (Table III). The specific function of these and other benzoate-induced unknown proteins remains to be elucidated.

Conclusions: Membrane Proteins Involved in Aromatic Catabolism of G. metallireducens

The integration of the data obtained from (i) membrane proteome analysis by the spectral counting approach/1D SDS-PAGE fractionation, (ii) quantitative gene expression determinations, and (iii) enzyme activity measurements enabled a number of novel insights into key processes of aromatic degradation in obligately anaerobic bacteria. The so far non-characterized dearomatizing benzoyl-CoA reductase complex is proposed to be membrane-associated possibly by interaction with components similar to those of energy-converting hydrogenases. This suggestion is first based on the presence of the BamFG components of the putative dearomatizing BamB–I complex in the membrane fraction and the inability to identify them in the soluble proteome. Second two benzoate-induced gene clusters putatively coding for proteins with high similarities to components of energy-converting hydrogenases were identified. This finding implies that electron transfer to the aromatic ring is not coupled to a stoichiometric ATP hydrolysis but may rather be driven by a membrane potential involving the energy-conserving hydrogenase modules. The identification of benzoate-induced genes coding for an ABC transporter that contains all characteristic and unique features of tungstate uptake systems indicates that the putative active site-containing component of benzoyl-CoA reductase, BamB, contains rather a tungsten- than molybdenum-containing cofactor. The induction of a dissimilatory nitrate reductase by benzoate can be explained by common regulatory circuits of molybdo-/tungstopterin-containing proteins, whereas the role of benzoate-induced NiFe/heme b-containing uptake hydrogenase remains to be studied. Growth on benzoate also increased the level of oxygen-scavenging proteins most probably by posttranscriptional regulation. Thus, anaerobic aromatic degradation appears to be associated with a higher demand for protection from oxidative stress, a phenomenon that is probably common for all anaerobic bacteria using aromatic growth substrates.

* This work was funded by the Deutsche Forschungsgemeinschaft Grant BO 1565/6-1.

§ The on-line version of this article (available at <http://www.mcponline.org>) contains supplemental material.

§ Both authors contributed equally to this work.

‡‡ To whom correspondence should be addressed: Inst. of Biochemistry, Brüderstrasse 34, D-04103 Leipzig, Germany. Fax: 49-341-9736919; E-mail: boll@uni-leipzig.de.

REFERENCES

- Boll, M., Fuchs, G., and Heider, J. (2002) Anaerobic oxidation of aromatic compounds and hydrocarbons. *Curr. Opin. Chem. Biol.* **6**, 604–611
- Diaz, E. (2004) Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility. *Int. Microbiol.* **7**, 173–180
- Fuchs, G. (2008) Anaerobic metabolism of aromatic compounds. *Ann. N.Y. Acad. Sci.* **1125**, 82–99
- Gibson, J., and Harwood, C. S. (2002) Metabolic diversity in aromatic compound utilization by anaerobic microbes. *Annu. Rev. Microbiol.* **56**, 345–369
- Boll, M. (2005) Dearomatizing benzene ring reductases. *J. Mol. Microbiol. Biotechnol.* **10**, 132–142
- Boll, M. (2005) Key enzymes in the anaerobic aromatic metabolism catalyzing Birch-like reductions. *Biochim. Biophys. Acta* **1707**, 34–50
- Boll, M., and Fuchs, G. (2005) Unusual reactions involved in anaerobic metabolism of phenolic compounds. *Biol. Chem.* **386**, 989–997
- Boll, M., and Fuchs, G. (1995) Benzoyl-coenzyme A reductase (dearomatizing), a key enzyme of anaerobic aromatic metabolism. ATP dependence of the reaction, purification and some properties of the enzyme from *Thauera aromatica* strain K172. *Eur. J. Biochem.* **234**, 921–933
- Boll, M., Fuchs, G., Meier, C., Trautwein, A., and Lowe, D. J. (2000) EPR and Mossbauer studies of benzoyl-CoA reductase. *J. Biol. Chem.* **275**, 31857–31868
- Unciuleac, M., and Boll, M. (2001) Mechanism of ATP-driven electron transfer catalyzed by the benzene ring-reducing enzyme benzoyl-CoA reductase. *Proc. Natl. Acad. Sci.* **98**, 13619–13624
- Möbitz, H., and Boll, M. (2002) A Birch-like mechanism in enzymatic benzoyl-CoA reduction: a kinetic study of substrate analogues combined with an *ab initio* model. *Biochemistry* **41**, 1752–1758
- Thiele, B., Rieder, O., Golding, B. T., Müller, M., and Boll, M. (2008) Mechanism of enzymatic Birch reduction: stereochemical course and exchange reactions of benzoyl-CoA reductase. *J. Am. Chem. Soc.* **130**, 14050–14051
- McInerney, M. J., Rohlin, L., Mouttaki, H., Kim, U., Krupp, R. S., Rios-Hernandez, L., Sieber, J., Struchtemeyer, C. G., Bhattacharyya, A., Campbell, J. W., and Gunsalus, R. P. (2007) The genome of *Syntrophus aciditrophicus*: life at the thermodynamic limit of microbial growth. *Proc. Natl. Acad. Sci. U.S.A.* **104**, 7600–7605
- Wischgoll, S., Heintz, D., Peters, F., Erxleben, A., Sarnighausen, E., Reski, R., Van Dorsselaer, A., and Boll, M. (2005) Gene clusters involved in anaerobic benzoate degradation of *Geobacter metallireducens*. *Mol. Microbiol.* **58**, 1238–1252
- Peters, F., Shinoda, Y., McInerney, M. J., and Boll, M. (2007) Cyclohexa-1,5-diene-1-carbonyl-coenzyme A (CoA) hydratases of *Geobacter metallireducens* and *Syntrophus aciditrophicus*: evidence for a common benzoyl-CoA degradation pathway in facultative and strict anaerobes. *J. Bacteriol.* **189**, 1055–1060
- Kuntze, K., Shinoda, Y., Moutakki, H., McInerney, M. J., Vogt, C., Richnow, H. H., and Boll, M. (2008) 6-Oxocyclohex-1-ene-1-carbonyl-coenzyme A hydrolases from obligately anaerobic bacteria: characterization and identification of its gene as a functional marker for aromatic compounds degrading anaerobes. *Environ. Microbiol.* **10**, 1547–1556
- Andersen, J. S., Wilkinson, C. J., Mayor, T., Mortensen, P., Nigg, E. A., and Mann, M. (2003) Proteomic characterization of the human centrosome by protein correlation profiling. *Nature* **426**, 570–574
- Blondeau, F., Ritter, B., Allaire, P. D., Wasiak, S., Girard, M., Hussain, N. K., Angers, A., Legendre-Guillemain, V., Roy, L., Boismenu, D., Kearney, R. E., Bell, A. W., Bergeron, J. J., and McPherson, P. S. (2004) Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling. *Proc. Natl. Acad. Sci. U.S.A.* **101**, 3833–3838
- Gao, B. B., Clermont, A., Rook, S., Fonda, S. J., Srinivasan, V. J., Wojtkowski, M., Fujimoto, J. G., Avery, R. L., Arrigg, P. G., Bursell, S. E., Aiello, L. P., and Feener, E. P. (2007) Extracellular carbonic anhydrase mediates hemorrhagic retinal and cerebral vascular permeability through prekallikrein activation. *Nat. Med.* **13**, 181–188
- Gilchrist, A., Au, C. E., Hiding, J., Bell, A. W., Fernandez-Rodriguez, J., Lesimple, S., Nagaya, H., Roy, L., Gosline, S. J., Hallett, M., Paiement, J., Kearney, R. E., Nilsson, T., and Bergeron, J. J. (2006) Quantitative proteomics analysis of the secretory pathway. *Cell* **127**, 1265–1281
- Lu, P., Rangan, A., Chan, S. Y., Appling, D. R., Hoffman, D. W., and Marcotte, E. M. (2007) Global metabolic changes following loss of a feedback loop reveal dynamic steady states of the yeast metabolome. *Metab. Eng.* **9**, 8–20
- Wang, R., and Marcotte, E. M. (2008) The proteomic response of *Mycobacterium smegmatis* to anti-tuberculosis drugs suggests targeted pathways. *J. Proteome Res.* **7**, 855–865
- Lovley, D. R., and Phillips, E. J. (1988) Novel mode of microbial energy metabolism: organic carbon oxidation coupled to dissimilatory reduction of iron or manganese. *Appl. Environ. Microbiol.* **54**, 1472–1480
- Ramagli, L. S., and Rodriguez, L. V. (1985) Quantitation of microgram amounts of protein in two-dimensional polyacrylamide gel electrophoresis sample buffer. *Electrophoresis* **559**–563
- Laemmli, U. K. (1970) Cleavage of structural proteins during the assembly of the head of bacteriophage T4. *Nature* **227**, 680–685
- Neuhoff, V., Stamm, R., Pardowitz, I., Arold, N., Ehrhardt, W., and Taube, D. (1990) Essential problems in quantification of proteins following colloidal staining with Coomassie brilliant blue dyes in polyacrylamide gels, and their solution. *Electrophoresis* **11**, 101–117
- Elias, J. E., and Gygi, S. P. (2007) Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nat. Methods* **4**, 207–214
- Käll, L., Krogh, A., and Sonnhammer, E. L. (2007) Advantages of combined transmembrane topology and signal peptide prediction—the Phobius web server. *Nucleic Acids Res.* **35**, W429–W432
- Lu, P., Vogel, C., Wang, R., Yao, X., and Marcotte, E. M. (2007) Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation. *Nat. Biotechnol.* **25**, 117–124
- Johannes, J., Bluschke, A., Jehmlich, N., von Bergen, M., and Boll, M. (2008) Purification and characterization of active-site components of the putative p-cresol methylhydroxylase membrane complex from *Geobacter metallireducens*. *J. Bacteriol.* **190**, 6493–6500
- Peters, F., Heintz, D., Johannes, J., van Dorsselaer, A., and Boll, M. (2007) Genes, enzymes, and regulation of para-cresol metabolism in *Geobacter metallireducens*. *J. Bacteriol.* **189**, 4729–4738
- Schöcke, L., and Schink, B. (1998) Membrane-bound proton-translocating pyrophosphatase of *Syntrophus gentianae*, a syntrophically benzoate-degrading fermenting bacterium. *Eur. J. Biochem.* **256**, 589–594
- Parker, A., and Engel, P. C. (2000) Preliminary evidence for the existence of specific functional assemblies between enzymes of the beta-oxidation pathway and the respiratory chain. *Biochem. J.* **345**, 429–435
- Jansson, M., Wårell, K., Levander, F., and James, P. (2008) Membrane protein identification: N-terminal labeling of nontryptic membrane protein peptides facilitates database searching. *J. Proteome Res.* **7**, 659–665
- Hedderich, R. (2004) Energy-converting [NiFe] hydrogenases from archaea and extremophiles: ancestors of complex I. *J. Bioenerg. Biomembr.* **36**, 65–75
- Borisov, V. B. (1996) Cytochrome bd: structure and properties. *Biokhimiya* **61**, 786–799
- Jünemann, S. (1997) Cytochrome bd terminal oxidase. *Biochim. Biophys. Acta* **1321**, 107–127
- Thiele, B., Rieder, O., Jehmlich, N., von Bergen, M., Müller, M., and Boll, M. (2008) Aromatizing cyclohexa-1,5-diene-1-carbonyl-coenzyme A oxidase. Characterization and its role in anaerobic aromatic metabolism. *J. Biol. Chem.* **283**, 20713–20721
- Fay, P. (1992) Oxygen relations of nitrogen fixation in cyanobacteria. *Microbiol. Rev.* **56**, 340–373
- Lehmann, Y., Meile, L., and Teuber, M. (1996) Rubrerythrin from *Clostridium perfringens*: cloning of the gene, purification of the protein, and characterization of its superoxide dismutase function. *J. Bacteriol.* **178**, 7152–7158
- Grunden, A. M., and Shanmugam, K. T. (1997) Molybdate transport and regulation in bacteria. *Arch. Microbiol.* **168**, 345–354
- Andreesen, J. R., and Makdessi, K. (2008) Tungsten, the surprisingly positively acting heavy metal element for prokaryotes. *Ann. N.Y. Acad. Sci.* **1125**, 215–229
- Makdessi, K., Andreesen, J. R., and Pich, A. (2001) Tungstate Uptake by a highly specific ABC transporter in *Eubacterium acidaminophilum*. *J. Biol. Chem.* **276**, 24557–24564
- Roy, R., Menon, A. L., and Adams, M. W. (2001) Aldehyde oxidoreductases from *Pyrococcus furiosus*. *Methods Enzymol.* **331**, 132–144
- Roy, R., and Adams, M. W. (2002) Characterization of a fourth tungsten-containing enzyme from the hyperthermophilic archaeon *Pyrococcus furiosus*. *J. Bacteriol.* **184**, 6952–6956

4. Etude complémentaire

4.1. Objectif

Les hypothèses formulées sur les particularités de l'étape de réduction du cycle aromatique chez la bactérie anaérobie obligatoire *G. metallireducens* nous ont incité à réaliser des expériences complémentaires. Ces expériences ont visé à isoler et à caractériser l'enzyme responsable de cette réaction chez *G. metallireducens*, la benzoyl-CoA reductase (BCR) dont les caractéristiques générales semblent être complètement différentes des BCRs connues jusque là chez les bactéries anaérobies facultatives.

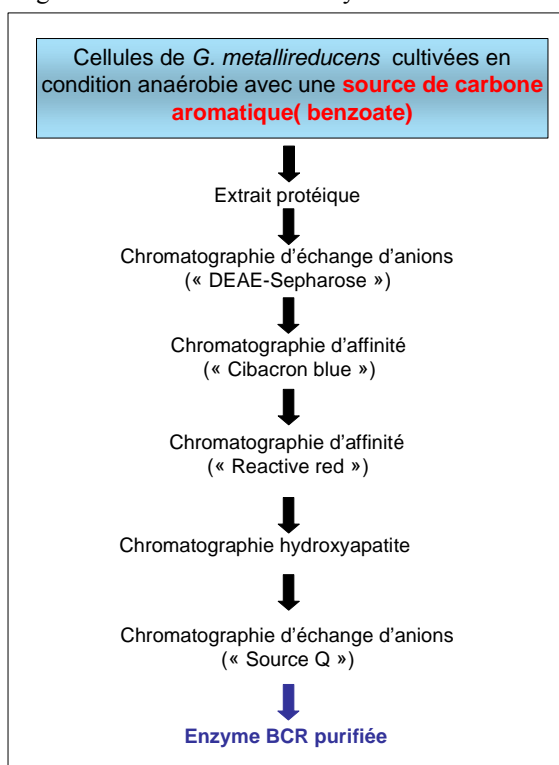
4.2. Mise en œuvre

Jusque là, la seule enzyme BCR ayant été isolée et étudiée provient de la bactérie anaérobie facultative *T. aromatica* [Boll et al., 1995]. Le test d'activité utilisé dans cette étude n'étant pas applicable pour le nouveau type d'enzyme, un nouveau test d'activité spectrophotométrique a dû être élaboré. Celui-ci utilise la réaction inverse, c'est-à-dire l'oxydation du cyclohexa-1,5-dienoyl-1-carboxyl-CoA avec un accepteur artificiel d'électrons (2,6-dichlorophenolindolphenol, DCIP).

A partir de cellules de *G. metallireducens* cultivées en condition anaérobie en présence de benzoate, un extrait protéique soluble a été préparé. L'enzyme BCR a été purifiée par 5 étapes successives de fractionnement chromatographique (Figure 4) en suivant la fraction protéique d'intérêt grâce au test d'activité spectrophotométrique développé.

L'enzyme purifiée a ensuite pu être caractérisée. Il fut procédé à l'analyse de ses propriétés moléculaires, de ses constituants protéiques, de ses cofacteurs, de sa stabilité, de sa spécificité pour des substrats et de ses propriétés cinétiques par différentes techniques (gel 1D, gel natif, analyses nanoLC-MS/MS, chromatographie d'exclusion stérique, spectroscopie UV/vis et EPR). Les analyses nanoLC-MS/MS des digests peptidiques des constituants protéiques de l'enzyme séparés sur gel 1D ont été réalisées sur un système nanoHPLC-Chip (Agilent Series 1100) couplé à une trappe ionique HCT Ultra (Bruker Daltonics). Etant donné la très faible complexité des mélanges analysés, une puce microfluidique comprenant une colonne analytique « courte » (43 mm) a été utilisée. Le détail des expériences est décrit dans la publication des résultats.

Figure 4 : Purification de l'enzyme BCR

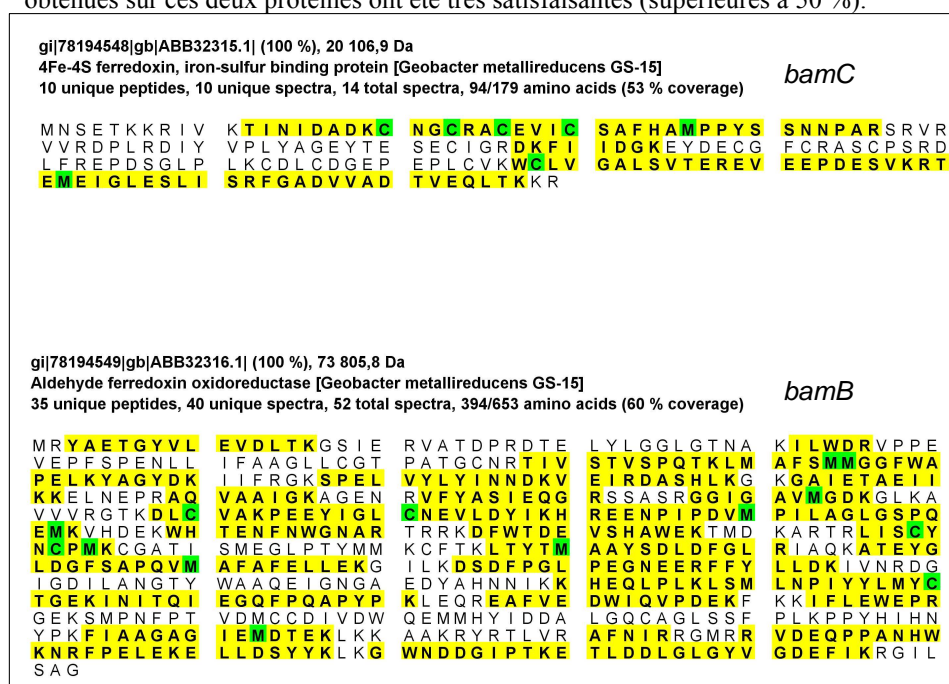


4.3. Résultats publiés

Les résultats obtenus dans cette étude ont fait l'objet d'une publication acceptée dans le journal *Proceedings of the National Academy of Sciences of the United States of America* en août 2009.

L'étude a permis d'identifier et de caractériser finement l'enzyme BCR de *G. metallireducens* qui constitue le prototype d'une toute nouvelle classe d'enzymes BCR. Cette enzyme, très sensible à l'oxygène et présentant une masse moléculaire d'environ 185 kDa, est constituée de sous unités de 73 et 20 kDa (respectivement la protéine gi|78194549 codée par le gène *bamB* et la protéine gi|78194548 codée par le gène *bamC*, Figure 5) suggérant une composition $\alpha_2\beta_2$. L'étude a également révélé le contenu en cofacteur de chaque unité $\alpha\beta$ (0.9 W, 15 Fe et 12.5 S ; un cluster $[3\text{Fe-4S}]^{0/1+}$ et trois clusters $[4\text{Fe-4S}]^{+1/+2}$).

Figure 5 : Sous-unités de l'enzyme BCR identifiées par LC-MS/MS. Les peptides surlignés en jaune ont été identifiés par LC-MS/MS. Les couvertures de séquence obtenues sur ces deux protéines ont été très satisfaisantes (supérieures à 50 %).



5. Conclusion et perspectives

L'ensemble des travaux réalisés sur l'étude de la dégradation des composés aromatiques chez la bactérie anaérobie obligatoire *G. metallireducens* illustre l'intérêt de la protéomique (particulièrement la protéomique quantitative) à grande échelle pour l'étude différentielle de protéomes. Celle-ci a permis de confirmer des hypothèses formulées précédemment sur les protéines impliquées dans le catabolisme étudié mais aussi d'établir de nouvelles observations et hypothèses sur des mécanismes associés à ce processus grâce à certains résultats qui n'étaient pas attendus. Certaines hypothèses ont ensuite pu être vérifiées grâce à l'étude complémentaire réalisée spécifiquement sur l'enzyme clé benzoyl-CoA reductase.

Ces travaux illustrent également l'intérêt des études intégratives qui combinent les aspects complémentaires de la protéomique et de la transcriptomique et qui permettent une vision plus globale des processus étudiés.

Les perspectives de ce travail seraient d'arriver à confirmer les hypothèses formulées sur le mécanisme et les composés impliqués dans le transfert d'électrons lors de l'étape clé de réduction des cycles aromatiques et de donner une signification biologique à l'expression différentielle observée de certaines protéines restant pour l'instant ininterprétée.

Identification and characterization of the tungsten-containing class of benzoyl-coenzyme A reductases

Johannes W. Kung^a, Claudia Löffler^a, Katerina Dörner^b, Dimitri Heintz^c, Sébastien Gallien^d, Alain Van Dorsselaer^d, Thorsten Friedrich^b, and Matthias Boll^{a,1}

^aInstitute of Biochemistry, University of Leipzig, 04103 Leipzig, Germany; ^bInstitute of Biochemistry, University of Freiburg, 79104 Freiburg, Germany; ^cInstitute de Biologie Moléculaire des Plantes, Centre National de la Recherche Scientifique (CNRS)-Unité Propre de Recherche 2357, Université de Strasbourg, 67083 Strasbourg, France; and ^dLaboratoire de Spectrométrie de Masse BioOrganique, IPHC-DSA, Institut Pluridisciplinaire Hubert Curien Département des Sciences Analytiques, Université de Strasbourg, CNRS, 67087 Strasbourg, France

Edited by Caroline S. Harwood, University of Washington, Seattle, WA, and approved July 30, 2009 (received for review May 8, 2009)

Aromatic compounds are widely distributed in nature and can only be biomineralized by microorganisms. In anaerobic bacteria, benzoyl-CoA (BCoA) is a central intermediate of aromatic degradation, and serves as substrate for dearomatizing BCoA reductases (BCRs). In facultative anaerobes, the mechanistically difficult reduction of BCoA to cyclohexa-1,5-dienoyl-1-carboxyl-CoA (dienoyl-CoA) is driven by a stoichiometric ATP hydrolysis, catalyzed by a soluble, three [4Fe-4S] cluster-containing BCR. In this work, an *in vitro* assay for BCR from the obligately anaerobic *Geobacter metallireducens* was established. It followed the reverse reaction, the formation of BCoA from dienoyl-CoA in the presence of various electron acceptors. The benzoate-induced activity was highly specific for dienoyl-CoA ($K_m = 24 \pm 4 \mu\text{M}$). The corresponding oxygen-sensitive enzyme was purified by several chromatographic steps with a 115-fold enrichment and a yield of 18%. The 185-kDa enzyme comprised 73- and 20-kDa subunits, suggesting an $\alpha_2\beta_2$ -composition. MS analysis revealed the subunits as products of the benzoate-induced *bamBC* genes. The $\alpha\beta$ unit contained 0.9 W, 15 Fe, and 12.5 acid-labile sulfur. Results from EPR spectroscopy suggest the presence of one [3Fe-4S]^{0/+1} and three [4Fe-4S]^{+1/+2} clusters per $\alpha\beta$ unit; oxidized BamBC exhibited an EPR signal typical for a W(V) species. The FeS clusters and the W-cofactor could only be fully reduced by dienoyl-CoA. BamBC represents the prototype of a previously undescribed class of dearomatizing BCRs that differ completely from the ATP-dependent enzymes from facultative anaerobes.

anaerobic aromatic degradation | AOR | geobacter

Both naturally occurring and man-made aromatic compounds are ubiquitous and many are highly persistent and harmful to man and the environment. Only microorganisms are capable of fully degrading aromatic compounds to CO₂. Under aerobic conditions, bacteria use oxygenases to insert oxygen atoms into the ring to relieve ring resonance. In the absence of oxygen, this biochemical strategy is not an option, and anaerobic bacteria use completely different enzymes that relieve ring resonance by a reductive mechanism (1–4).

In anaerobic bacteria benzoyl-CoA (BCoA) is a central intermediate in the degradation pathways of many aromatic compounds. BCoA serves as the substrate for BCoA reductases (BCR), which dearomatize the aromatic ring by reduction yielding cyclohexa-1,5-diene-1-carboxyl-CoA (dienoyl-CoA; Fig. 1) (5–8). A BCR enzyme has so far only been isolated and studied in the denitrifying, facultative anaerobe *Thauera aromatica* (9). This extremely oxygen-sensitive enzyme has an $\alpha\beta\gamma\delta$ -composition and harbors three [4Fe-4S]^{+1/+2} clusters (10). It catalyzes electron transfer from reduced ferredoxin to the substrate with concomitant ATP hydrolysis (11), a reaction that was previously considered unique to nitrogenases (12). The coupling of electron transfer to an exergonic reaction is essential

to overcome the high redox barrier for aromatic ring reduction. Initial evidence for a mechanism via radical intermediates according to a Birch-type reduction has been obtained (13, 14).

Homologs of ATP-dependent BCRs from *T. aromatica* are present in the genomes of all facultative anaerobes described so far, which degrade low molecular aromatic growth substrates (5). In contrast, homologs of these genes are absent in the genomes of *Geobacter* species [Fe(III)-respiring] or *Syntrophus aciditrophicus* (fermenting), which are known to degrade benzoate (15–17). Also, the assay for ATP-dependent BCRs from facultative anaerobes failed with extracts from *Geobacter metallireducens* (16). The lack of an assay has so far prevented the study of an ATP independent BCR.

Recently, eight clustered benzoate-induced genes were identified, which are proposed to code for an uncharacterized BCR (Fig. 1B) (16). Homologs of this putative BamBCDEFGHI (BamB-I, encoded by gi78194549–41) complex (*bam*, benzoic acid metabolism) have so far only been identified in obligately anaerobic bacteria that use low molecular aromatic growth substrates. The individual proteins of the BamB-I complex show similarities to soluble components of NADH:quinone oxidoreductases (BamGHI), electron transfer components of hydrogenases (BamC and SeCys containing BamF), soluble heterodisulfide reductases (BamDE), and to Mo- or W-containing aldehyde:ferredoxin oxidoreductases (AORs, BamB). Accordingly, growth of two obligate anaerobes on benzoate depended on Mo or W and Se (16, 18). BamB was hypothesized to function as the active-site-containing component, whereas the remaining components were suggested to be involved in electron transfer from an unknown donor to BamB. Membrane proteome analysis of *G. metallireducens* revealed that at least the BamF and BamG components are membrane-associated (19). No ATP-binding motif was found in the BamB-I complex, suggesting that electron transfer to BCoA may be independent of ATP hydrolysis.

Here, we describe the isolation and characterization of the prototype of a new class of BCR enzymes. The catalytic components of this enzyme contain W and FeS clusters. The general features of the enzymes differ completely from that of known ATP-dependent BCRs even though the substrate and its product are identical.

Author contributions: M.B. designed research; J.W.K., C.L., K.D., D.H., and S.G. performed research; T.F. contributed new reagents/analytic tools; J.W.K., D.H., S.G., A.V.D., T.F., and M.B. analyzed data; and J.W.K. and M.B. wrote the paper.

The authors declare no conflict of interest.

This article is a PNAS Direct Submission.

¹To whom correspondence should be addressed. E-mail: boll@uni-leipzig.de.

This article contains supporting information online at www.pnas.org/cgi/content/full/0905073106/DCSupplemental.

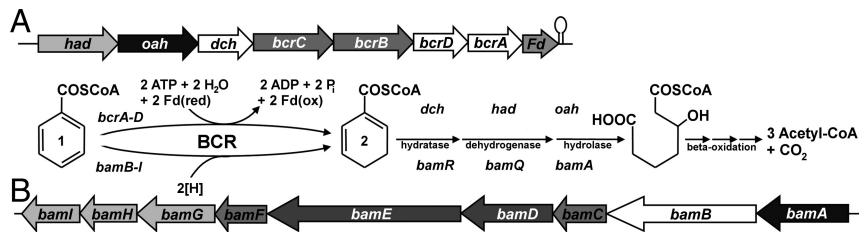


Fig. 1. Reaction of BCRs and organization of genes involved in BCoA degradation in *T. aromatica* and *G. metallireducens*. The BCoA degradation pathways proceed via identical intermediates in all anaerobic bacteria. All enzymes involved are highly similar in both organisms with the exception of BCR involved in the reduction of BCoA (1) to dienoyl-CoA (2). (A) Cluster of genes coding for BCR (*bcrA-D*) and other enzymes (*had*, *oah*, and *dch*) of the BCoA degradation pathway in *T. aromatica*. Note that BcrA-D couples ring reduction to a stoichiometric ATP hydrolysis (upper BCR reaction). (B) Cluster of genes coding for putative BCR (*bamB-I*) and another enzyme (*bamA*, a homolog for *oah*) involved in the BCoA degradation pathway in *G. metallireducens*. For simpler presentation, the *bamRQ* genes are not shown. The reaction catalyzed by BamB-I is suggested to be ATP independent (lower BCR reaction).

Results

Enzyme Assay for BCR from *G. metallireducens*. To date, in vitro enzyme activities have only been determined for ATP-dependent BCRs from facultative anaerobes using low-potential artificial electron donors (9). Although indirect evidence was provided that BCoA reduction should yield dienoyl-CoA in all anaerobes (8), the assay failed with extracts derived from obligate anaerobes grown on an aromatic growth substrate (16). To circumvent the dependence on an electron-activating system for the forward reaction, we determined the reverse reaction, the electron acceptor-dependent oxidation of dienoyl-CoA. This reaction is driven by aromatization and should be highly exergonic in the presence of appropriate electron acceptors. The dienoyl-CoA used throughout this study was enzymatically synthesized from BCoA using purified BCR from *T. aromatica* (20).

Soluble extracts from *G. metallireducens* cells grown on benzoate catalyzed the time-, protein-, and dienoyl-CoA-dependent reduction of 2,6-dichlorophenol indolphenol (DCPIP) in a pseudofirst order reaction as determined in a spectrophotometric assay. HPLC analysis of samples taken at different time points confirmed the formation of BCoA from dienoyl-CoA (a representative HPLC assay with the purified enzyme is shown in Fig. S1). One DCPIP was reduced per BCoA formed. The dienoyl-CoA oxidizing activity was oxygen-sensitive (for details, see below). For this reason, all steps were carried out under strictly anoxic conditions; DCPIP could not be replaced by oxygen as electron acceptor. Because the dienoyl-CoA aromatizing activity was sensitive to dilution ($<0.1 \text{ mg mL}^{-1}$), the enzyme assay routinely contained 5 mg mL^{-1} BSA. The activity was present in extracts of cells grown on benzoate ($400\text{--}600 \text{ nmol min}^{-1} \text{ mg}^{-1}$), but was virtually absent in extracts from cells grown on acetate ($<1 \text{ nmol min}^{-1} \text{ mg}^{-1}$). Addition of MgATP or MgADP (5 mM each) had no effect on the activity. After ultracentrifugation, 45% of the dienoyl-CoA oxidizing activity was found in the membrane fraction. Addition of 500 mM KCl to the extract buffer released 95% of the aromatizing activity into the soluble protein fraction. After ultracentrifugation and a desalting step, the soluble protein fraction was used for purification.

Purification of Dienoyl-CoA:Acceptor Oxidoreductase. With the spectrophotometric assay established, the isolation of the dienoyl-CoA aromatizing enzyme was carried out. Purification was accomplished in five chromatographic steps, including DEAE-Sepharose anion exchange chromatography (elution at 225 mM KCl at pH 7.8), Cibacron Blue and Reactive Red affinity chromatography (flow through, respectively), hydroxyapatite chromatography (elution at 80 mM sodium phosphate at pH 7.8), and Source Q anion exchange chromatography (elution at 210 mM KCl at pH 7.8). The order of columns used and the integration of the two affinity chromatography steps were essential for complete removal of contaminants. A typical purification protocol is shown in Table 1. A 115-fold enrichment with a yield of 18% was achieved. From 20 g cells (wet mass), $\approx 6 \text{ mg}$ protein with a specific activity of $68 \mu\text{mol min}^{-1} \text{ mg}^{-1}$ were obtained.

Molecular Properties, Genes, and Cofactors of the Dienoyl-CoA Aromatizing Enzyme. SDS/PAGE analysis of the enzyme fractions obtained during the purification procedure revealed the enrichment of protein bands migrating at 65 and 20 kDa in an almost 1:1 ratio after normalization for the molecular masses (Fig. 2A). The excised bands were analyzed by nanoliquid chromatography-coupled tandem MS (nanoLC-MS/MS) after enzymatic digestion with trypsin. By using MASCOT MS/MS data searches against *G. metallireducens* protein database, the resulting peptides were identified to belong to the *bamC* gene product (gi78194548, 52% sequence coverage, 20-kDa subunit) and to the *bamB* gene product (gi78194549, 60% sequence coverage, 65-kDa subunit; for MS data, see Table S1). Therefore, the dienoyl-CoA oxidizing enzyme is henceforth termed BamBC. The masses as deduced from the amino acid sequences of BamB and BamC were 73 and 21 kDa, respectively. The discrepancy between the molecular masses of BamB as deduced from the amino acid sequence (73 kDa) and as determined by SDS PAGE (65 kDa) is unclear. The peptides identified by MS did not indicate a loss of a corresponding 8-kDa peptide at the C or N terminus. BamBC had a native molecular mass of $185 \pm 10 \text{ kDa}$ as determined by gel filtration, suggesting an $\alpha_2\beta_2$ composition.

Table 1. Protocol for purification of BamBC

Purification step	Volume, mL	Total activity, U	Protein, mg	Specific activity, U mg^{-1}	Yield, %	Enrichment, fold
100,000 $\times g$ supernatant	79	1,778	2,955	0.6	100	1.0
Dialysis	80	2,102	2,880	0.7	118	1.2
DEAE	310	1,273	322	3.9	71	6.6
CibacronBlue/ReactiveRed	320	615	102	6.0	35	10
Hydroxyapatite	50	452	22	20	26	33
SourceQ	24	321	6.0	68	18	115

One U refers to the oxidation of $1 \mu\text{mol min}^{-1}$ dienoyl-CoA. Activities were determined with DCPIP as electron acceptor.

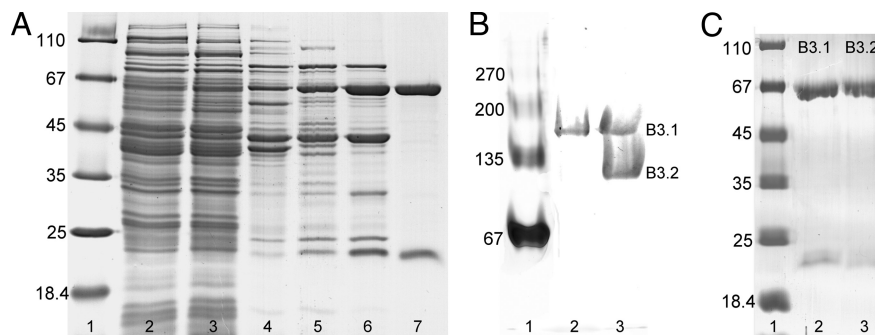


Fig. 2. Purification of BamBC and activity staining. **A** SDS-polyacrylamide gel of dienoyl-CoA:acceptor oxidoreductase activity containing fractions obtained during the purification of BamBC from *G. metallireducens*, 7 μg of protein per lane. Lane 1, molecular mass standard; lane 2, cell extract; lane 3, dialysed extract; lane 4, DEAE Sepharose; lane 5, Cibacron Blue and Reactive Red agarose; lane 6, hydroxyapatite; lane 7, SourceQ Sepharose. **(B)** Native gel electrophoresis and activity staining. Lane 1, BSA standard (Coomassie-stained); lane 2, soluble extract of cells grown benzoate; lane 3, purified BamBC. Lanes 2 and 3 were activity-stained. **(C)** SDS polyacrylamide gel analysis of bands activity-stained bands (B3.1 and B3.2). Lane 1, molecular mass standard; lane 2, excised band at 180 kDa in **B**; lane 3, excised band at 100 kDa in **B**.

Purified BamBC was subjected to native gel electrophoresis. Subsequent activity staining with dienoyl-CoA as electron donor, DCPIP as primary and a staining redox dye as secondary electron acceptor revealed two protein bands of ≈ 180 and 100 kDa (Fig. 2*B*). Both bands were excised and analyzed by denaturing gel electrophoresis. In both cases two bands migrating at the corresponding molecular masses of BamB and BamC were obtained; the assignment of the proteins bands to BamBC was confirmed by MS analysis. Thus, the two bands obtained after native gel electrophoresis apparently represent the $\alpha_2\beta_2$ and $\alpha\beta$ forms of BamBC (Fig. 2*C*). In extracts from cells grown on benzoate, only one band ≈ 180 kDa was obtained on native gels after activity staining (Fig. 2*B*). Obviously, in cell extracts, BamBC was more resistant to dissociation during native gel electrophoresis than in the as-isolated form.

The amount of iron and acid labile sulfur per mol $\alpha\beta$ -unit was 15.2 ± 0.6 mol Fe and 12.7 ± 1.4 mol acid labile sulfur as determined colorimetrically (mean values \pm SDs). Inductively coupled plasma (ICP) MS analysis of metals and Se revealed the presence of 0.9 W, 1.2 Zn, and 2.1 Ca per $\alpha\beta$ unit; the amount of Mn, Co, Ni, Cu, V, Se, and Mo was <0.05 mol per mol $\alpha\beta$ unit.

Factors Affecting BamBC Activity. BamBC as isolated in the absence of a reducing agent had a half life in air of ≈ 3 h. Sensitivity was greatly enhanced to a half-life <30 s when the enzyme was incubated with dithionite or dienoyl-CoA before oxygen exposure. Under anoxic conditions in the oxidized state, it was stable at -20°C for weeks in the presence of 5% PEG₄₀₀₀. Virtually no loss of activity was observed when BamBC was incubated with 5 mM sodium cyanide for up to 2 h at 4°C .

UV/vis and EPR Spectroscopy. The dark brownish enzyme as isolated was considered to be in the oxidized state. It exhibited a typical UV/vis spectrum of iron-sulfur proteins with a shoulder between 400 and 450 nm (Fig. 3*A*). Virtually complete reduction of BamBC (1–5 μM) was accomplished by excess of dienoyl-CoA (50 μM). In contrast, addition of sodium dithionite (50 μM) bleached the spectrum of oxidized BamBC only by 10% when compared with dienoyl-CoA. The difference spectrum (oxidized minus dienoyl-CoA reduced form) exhibited a maximum at 409 nm (Fig. 3*A*, *Inset*); the molecular extinction coefficients as determined are presented in Table 2.

Oxidized BamBC as isolated showed two distinct EPR signals (Fig. 3*B* and *C*). They are indicative for $S = 1/2$ signals of a $[3\text{Fe-4S}]^{+1}$ cluster (optimally developed at 10 K) and a W(V) species (40 K). At higher temperatures, the W(V) signal changed from an axial ($g_x = 1.893$, $g_y = g_z = 1.986$) to a rhombic signal

($g_z = 2.013$). On addition of excess dienoyl-CoA, both signals disappeared. In parallel complex and very broad, fast relaxing signals optimally developed at 10 K and 20 mW spanning >70 mT. The signal is interpreted as a mixture of interacting and noninteracting low-spin $[4\text{Fe-4S}]^{+}$ clusters. The features between 330 and 370 mT showed nearly identical power saturation and temperature dependencies, which did not allow the extraction of subspectra from single species. On addition of dithionite, the $[3\text{Fe-4S}]^{+1}$ signal disappeared, whereas the W(V) signal remained, although at lower intensity. In agreement with UV/vis spectroscopy, the $[4\text{Fe-4S}]$ clusters were hardly reducible to the paramagnetic $+1$ -state by dithionite. In accordance to metal analysis, one BamBC is suggested to contain a $\text{W}^{+4/+5/+6}$ -cofactor, one $[3\text{Fe-4S}]^{+1/0}$, and three $[4\text{Fe-4S}]^{+1/+2}$ clusters.

Substrate Preference and Kinetic Properties. BamBC was highly specific for 1,5-dienoyl-CoA. None of the following dienoyl-CoA/monoenoil-CoA analogs was converted as tested by HPLC analysis: 1,3-dienoyl-CoA, 1,4-dienoyl-CoA, 1-enoil-CoA, 2-enoil-CoA, and 3-enoil-CoA (0.2 mM each). Benzaldehyde, crotonaldehyde, and acetaldehyde were not oxidized (2 mM each), and they had no inhibitory effect on dienoyl-CoA oxidation. BamBC used the electron acceptors benzyl viologen (350%

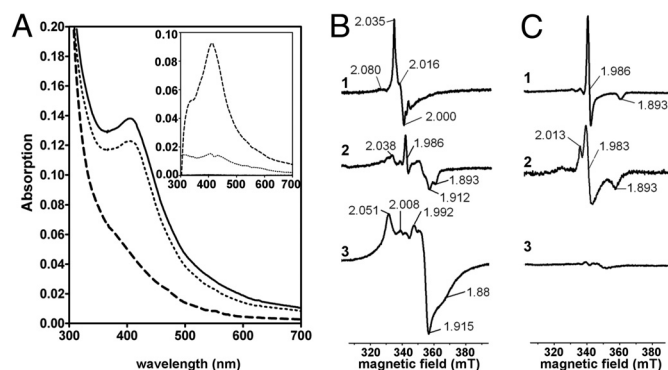


Fig. 3. UV/vis and EPR spectroscopy of BamBC. **(A)** UV/vis spectra of BamBC (*Inset*, difference spectra): solid line, as isolated (1 μM); dotted line, reduced with dithionite (50 μM); and dashed line, reduced with 1,5-dienoyl-CoA (50 μM). **(B)** Selected EPR spectra of FeS clusters at optimal conditions: B1, as isolated at 10 K, 20 mW after subtraction of spectrum C1; B2, dithionite reduced at 20 K, 5 mW; and B3 dienoyl-CoA reduced at 10 K, 20 mW. **(C)** EPR spectra of W(V): C1, as isolated at 40 K, 5 mW; C2, as isolated at 80 K, 5 mW; and C3, dienoyl-CoA reduced at 40 K, 5 mW. The numbers refer to g values. All EPR spectra were recorded at 0.6 mT modulation amplitude.

Table 2. Properties of BamBC

Property	Value
Reaction catalyzed	Dienoyl-CoA + acceptor _{ox} → Benzoyl-CoA + acceptor _{red}
Native molecular mass	185 ± 10 kDa
Subunits	BamB (73 kDa), BamC (21 kDa)
Molecular composition	$\alpha_2\beta_2$
Cofactor content (per $\alpha\beta$)	W: 0.9 Ca: 2.1 Zn: 1.2 Fe 15.2 ± 0.6 Acid labile S: 12.7 ± 1.4
UV/vis absorption maxima	oxidized: $\epsilon_{280} = 290,000 \text{ M}^{-1} \text{ cm}^{-1}$ oxidized: $\epsilon_{417} = 82,500 \text{ M}^{-1} \text{ cm}^{-1}$ oxidized–reduced: $\epsilon_{409} = 44,900 \text{ M}^{-1} \text{ cm}^{-1}$
K_m (dienoyl-CoA)	24 ± 4 μM
Catalytic no.	52 s ⁻¹ (per $\alpha\beta$)
Half-life on air	30 s (reduced), 3 h (oxidized)

activity) and methyl viologen (360% activity) at higher rates than DCPIP (100% activity, 0.4 mM each). In all cases, a stoichiometry of two electrons transferred per dienoyl-CoA oxidized was observed. BamBC did not catalyze the reduction of BCoA using Ti(III)-citrate (5 mM), sodium dithionite (1 mM), or reduced methyl viologen (0.5 mM) as electron donors. BamBC was active in a broad pH range from 5 to 9 with an optimum at pH 6.8. The initial rates of the reaction followed Michaelis–Menten kinetics with an apparent K_m for dienoyl-CoA of 24 ± 4 μM (mean value ± SD). The presence of BCoA had virtually no inhibitory effect on the initial rate at concentrations up to 0.5 mM by using 0.1 mM dienoyl-CoA.

Discussion

Properties of BamBC and Assignment to a Previously Undescribed Class of BCoA Reductases. In this work, the prototype of a previously undescribed class of tungsten containing BCR enzymes was identified and characterized (properties are summarized in Table 2). Although the forward reaction could not be followed with BamBC in the absence of the electron activation components, the following properties support the contention that BamBC indeed represent the active site components of BCR from *G. metallireducens*. (i) The reaction was highly specific for dienoyl-CoA, whereas none of the other dienoyl-CoA/monoenoil-CoA isomers tested was converted. (ii) Only dienoyl-CoA, but not the artificial reductant dithionite, completely reduced the enzyme as determined by UV/vis and EPR spectroscopy. (iii) Both transcription of the *bamBC* genes (16), as well as the dienoyl-CoA aromatizing activity were highly induced by an aromatic growth substrate. (iv) Highly similar homologs of the *bamBC* genes are present only in strict anaerobes that degrade aromatic compounds (Fig. 4). (v) The identification of W as cofactor fits well with the recently identified benzoate-induced tungstate transporter (19).

EPR spectroscopy and sequence comparisons with similar enzymes indicate that BamB represents the tungstopterin and [4Fe-4S] cluster-containing active site component, whereas BamC represents a two [4Fe-4S] and a [3Fe-4S] cluster-containing electron transfer subunit (16). The BamB subunit shows high amino acid sequence similarities (45–90%) identities to homologues from other *Geobacter* and *Syntrophus* species, and up to 30% identity to AOR-like enzymes from organisms that do not use aromatic growth substrates (16). Phylogenetic analysis suggests that BamB enzymes represent a separate class within the AOR family (Fig. 4). AORs usually catalyze the oxidation of aldehydes to the corresponding carboxylic acids, and the electrons are transferred to ferredoxin (21, 22). Because BamBC did not oxidize any of the tested aldehydes, the biological functions of AORs and BamBC appear to differ fundamentally.

In the case of BamBC, an unusual role of the tungsten cofactor is proposed as tungsto- or molybdo-pterin enzymes usually catalyze oxygen atom transfer reactions (23). The W(V) low spin EPR signal only disappeared on dienoyl-CoA addition supporting its catalytic relevance. Like other AOR enzymes, BamBC contained Ca, which is involved in pterin cofactor binding (24). The presence of Zn and an additional Ca in BamBC is rather unusual; both may function as Lewis acids for the stabilization of negatively charged radical/nonradical intermediates.

BamBC used the low potential electron acceptor methyl viologen without reaching thermodynamic equilibrium, suggesting that the E° of the BCoA/dienoyl-CoA redox couple is far < -500 mV (E° methyl viologen_{ox/red} = -448 mV) (25). The low potential can be rationalized by the largely exergonic aromatization reaction, and explains why dienoyl-CoA, but not dithionite, was capable of fully reducing BamBC.

The BamBC components of *G. metallireducens* BCR share no similarities to the known ATP-dependent BCRs from facultative

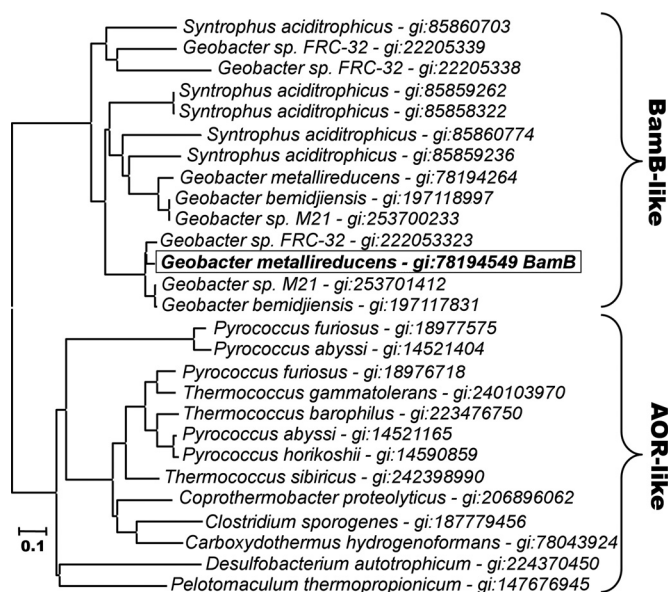


Fig. 4. Phylogenetic analysis of BamB and other AOR-like enzymes. The amino acid sequences used in the BamB cluster derive from the genomes of *Geobacter* and *Syntrophus* species that are known to degrade aromatic compounds. Note that many paralogs may be present in one organism. The sequences of AORs depicted are those with the highest similarities to BamB genes. For phylogenetic analysis (maximum likelihood method) the Mega4 software package was used (<http://www.megasoftware.net/>).

anaerobes. Obviously, the capability of dearomatizing BCoA has evolved twice in nature in two completely different ways.

The BamD-I components are hypothesized to constitute a complex electron transfer system that acts in conjunctions with BamBC. Obviously, they are not tightly attached to BamBC and are not required for dienoyl-CoA oxidation. BamD-I are predicted to contain numerous additional FeS clusters, two flavins and a SeCys (16).

Possible Scenarios for the Forward Reaction of BCR from *G. metallireducens*. BCRs are suggested to dearomatize the aromatic ring in a Birch-like reduction reaction via single electron transfer steps. The assumed extremely low redox potential for the first electron transfer requires an activation reaction. There are two plausible scenarios for ATP-independent electron activations in BCR from *G. metallireducens*.

A previous membrane proteome analysis revealed benzoate-induced proteins with similarities to components of energy-converting hydrogenases (19). The proteins lack the Ni-containing active site domain, which excludes a role in a hydrogenase reaction. Instead, they could represent attractive components involved in a membrane potential-driven electron transfer to BamBC via the BamDEFGHI modules. Alternatively, an electron bifurcation may drive the unfavorable electron transfer to the aromatic ring. This process was originally described for an electron transferring flavoprotein containing two FAD cofactors. It coupled the exergonic electron transfer from NADH to crotonyl-CoA to the endergonic one from NADH to ferredoxin (26, 27). The flavin cofactors were considered essential for this bifurcation of electrons. Notably, amino acid sequence comparisons suggested that BamE and BamH both contain flavin-binding motifs (16). However, by using various combinations of electron donors/acceptors, no evidence for such a bifurcation could be obtained (see *Materials and Methods*). It has to be considered that, even under very mild extract preparation conditions, BamBC activity was always distributed in both the soluble and membrane protein fraction. Also, BamD-I did not copurify. In summary, cell rupture obviously destroyed the correct assembly of the electron activation machinery. A screening for more stable BamB-I variants in other obligate anaerobes is required to overcome these problems in the future.

Materials and Methods

Cultivation of Cells and Preparation of Extracts. *G. metallireducens* (DSMZ 7210) was anaerobically cultivated in a 200-L-fermenter at 30 °C in a mineral salt medium (28) with benzoate (5 mM) or acetate (30 mM) as sole carbon source and nitrate (15 mM) as electron acceptor. For the preparation of crude extracts, frozen cells were anaerobically suspended at 4 °C in 20 mM triethanolamine/HCl, 5 mM MgCl₂, pH 7.8 (referred to as buffer A, 2 mL g⁻¹ cell, wet mass), containing 500 mM KCl, 0.02 mg DNase I, and 0.02 mg dithioerythritol. Cell disruption was accomplished by alternate freezing in liquid nitrogen and thawing (three times). The 100,000 × g pellet (1 h at 4 °C) was washed in buffer A containing 150 mM KCl (1 mL g⁻¹ wet cells) and centrifuged again. Combined supernatants were used for further studies.

Synthesis of CoA Esters. BCoA was synthesized from benzoic acid anhydride and CoA (29). Dienoyl-CoA was enzymatically synthesized from BCoA by using enriched BCR from *T. aromatica*; the product was purified by preparative HPLC as described (7). Other CoA-esters were synthesized as described earlier (20).

Enzyme Assays. All assays were carried out in anaerobically sealed cuvettes at 30 °C under a N₂ atmosphere (100%). Dienoyl-CoA:acceptor oxidoreductase activity was determined based on the absorbance change of DCPIP during reduction ($\Delta\epsilon_{700} = 5,900 \text{ M}^{-1} \text{ cm}^{-1}$ at pH 7.3, self-determined). The assay mixture (400 μL) contained 150 mM Mops/KOH, 15 mM MgCl₂, 150 mM NaCl, 5 mg mL⁻¹ BSA pH 7.3 (referred to as buffer B), plus 0.4 mM DCPIP, 0.2 mM dienoyl-CoA, and 1–10 μg of protein. Alternatively, benzyl viologen ($\epsilon_{578} = 12,000 \text{ M}^{-1} \text{ cm}^{-1}$) or methyl viologen ($\epsilon_{730} = 2,400 \text{ M}^{-1} \text{ cm}^{-1}$) (30) were used as electron acceptors in the spectrophotometric assay. In the discontinuous

assay (400 μL), 25 μL samples were taken at different time points and analyzed by HPLC as described (20). This assay was used for product analysis.

To test the forward reaction in cell extract preparations, Ti(III)-citrate (5 mM), dithionite (5 mM), and reduced methyl viologen (1 mM) were used as electron donors. As ferredoxin reducing system 2-oxoglutarate (5 mM) plus CoA (0.5 mM) were used. For potential electron bifurcation reactions NAD(P)⁺ (1 mM), and menadione (0.2 mM) were combined with all of the donors listed above.

Purification of BamBC. All steps were performed at 4 °C in an anaerobic glove box (N₂:H₂, 95:5, by vol.). Extracts of 20 g cells (wet mass) were dialysed overnight against 2.5-L buffer A containing 150 mM KCl. The 3,500 × g supernatant (80 mL) was applied to a DEAE-Sepharose column (Fast Flow, volume 100 mL, diameter 5.1 cm; GE Healthcare), which had been equilibrated with buffer A. The column was washed at a flow rate of 6 mL min⁻¹ with 3 bed vol of buffer A, followed by 5 bed vol of 120 mM NaCl in buffer A. Activity was eluted in a step gradient at 225 mM NaCl in buffer A within 310 mL. The fractions containing activity were applied onto Cibacron Blue Agarose 3GA Type 3000-CL (volume, 50 mL; diameter, 5.1 cm; Sigma-Aldrich) and Reactive Red Agarose 120 Type 3000-CL (volume, 30 mL; diameter, 2.6 cm; Sigma-Aldrich) columns connected directly to each other at a flow rate of 4 mL min⁻¹; both columns had been equilibrated with 150 mM NaCl in buffer A. Activity was obtained in the flow-through (320 mL). The pooled fraction were applied to a hydroxyapatite column (MacroPrep, ceramic hydroxyapatite 40 μm , 30 mL, 2.6 diameter; BioRad) at a flow rate of 4 mL min⁻¹ equilibrated with buffer A. The column was washed with 3 bed volumes of 20 mM potassium phosphate pH 7.8; activity eluted in a step gradient at 80 mM potassium phosphate, pH 7.8 (50 mL). For the last purification step, a Source 15Q column (HiLoad, volume 8 mL, 1.6 cm diameter; GE Healthcare) was equilibrated with buffer A containing 150 mM NaCl (flow rate 2 mL min⁻¹). The fractions applied eluted in a linear 150–300 mM NaCl gradient (50 bed volumes) at \approx 210 mM NaCl. The dienoyl-CoA:acceptor oxidoreductase activity containing fractions were concentrated in microconcentrators (30-kDa exclusion limit, Vivaspin 6; Sartorius) by centrifugation. The enzyme solution was slowly diluted under gentle stirring with a 20% (wt/vol) PEG₄₀₀₀ stock solution in buffer B to a final concentration of 5% (wt/vol) and kept frozen in anaerobic glass vials at –20 °C. In this form, the enzyme was stable for several weeks.

Determination of Molecular Mass. Native mass of BamBC was determined by analytical gel filtration via Superdex 200 (10/300 GL column, GE Healthcare; 25 mL column volume; 0.5 mL min⁻¹ flow rate) using 20 mM triethanolamine-HCl buffer, pH 7.8, 4 mM MgCl₂, 150 mM KCl. The column was calibrated with apoferritin (443 kDa), catalase (245 kDa), BSA (67 kDa), and carboanhydrase (29 kDa).

Determination of Metal Cofactors and Acid-Labile Sulfur. The content of iron and acid labile sulfur was determined colorimetrically by the methods of Lovenberg (31) and Beinert (32), respectively. W, Mo, V, Zn, Ca, Cu, Ni, Co, Mn, and Se were determined by ICP-MS analysis at the Helmholtz Centre for Environmental Research, Leipzig, Germany (Wennrich, Division for Analytics and Ecotoxicology). To exclude metal contamination, the last purification step (Source 15Q) was carried out by using chemicals of the highest purity available (RotiPuran Water, Roth; NaCl and NaOH TraceSelect grade, Sigma-Aldrich, respectively). The amount of metals and selenium in a protein-free buffer sample control was negligible.

Determination of Kinetic Properties and Oxygen Sensitivity. For K_m determination, the dienoyl-CoA concentration was varied from 3 to 200 μM by fitting the initial rates to Michaelis-Menten curves using the Prism software package (GraphPad). For determination of the pH dependence of the reaction, buffer B was supplemented by sodium acetate and Tris/HCl (0.15 M each). For oxygen inactivation assays, the enzyme was gently stirred in air at 4 °C, an anaerobically incubated enzyme served as control. The spectrophotometric assay was started by enzyme addition after different incubation times. To test the effect of cyanide (5 mM), the enzyme was anaerobically incubated at 4 °C for different time intervals before the assays were started (2 h). The substrate preference was tested with dienoyl-CoA/monoenoyl-CoA analogs at 0.2 mM concentrations; aldehydes as indicated were tested at 2 mM concentrations.

MS Analysis of BamBC. Gel bands of interest were digested *in gel* with trypsin and peptides were extracted as described previously (33). The resulting peptide were analyzed by nanoLC-MS/MS on an Agilent 1100 Series HPLC-Chip/MS system (Agilent Technologies) coupled to an HCT Ultra ion trap (Bruker Daltonics) for the SDS/PAGE analysis of the enzyme fraction, or on a nanoACQUITY Ultra-Performance-LC (UPLC, Waters) coupled to SYNAPT hybrid quad-

rupole orthogonal acceleration time-of-flight tandem mass spectrometer (Waters) for the SDS/PAGE analysis of the bands obtained after native gel analysis. The MS/MS data were analyzed using the MASCOT 2.2.0. algorithm (Matrix Science) to search against an in-house generated target-decoy protein database from *G. metallireducens*.

UV/vis Spectroscopy. UV/vis spectra of purified BamBC (1–5 μ M in buffer B without BSA) were recorded with a UV-1650PC Shimadzu spectrophotometer in a gas-tight quartz cuvette under anaerobic conditions. Dithionite (0.05 mM) or dienoyl-CoA (0.05 mM) were added from anaerobic stock solutions in buffer without BSA.

EPR Spectroscopy. EPR measurements were conducted with a Bruker EMX 1/6 spectrometer operating at X-band. The sample temperature was controlled with an Oxford instrument ESR-9 helium flow cryostat. The magnetic field was calibrated using a strong or a weak pitch standard. Samples were taken from the BamBC as isolated (300 μ M in buffer A plus 5% PEG₄₀₀₀, wt/vol), and after

1 min incubation with 4 mM dithionite or by 4.6 mM of dienoyl-CoA. EPR conditions were: microwave frequency, 9.44 GHz; modulation amplitude, 0.6 mT; time constant, 0.164 s; scan rate, 17.9 mT min⁻¹.

Further Determinations. SDS/PAGE (12.5%) was carried out according to Laemmli. Proteins were visualized using SimplyBlue SafeStain (Invitrogen). Protein was routinely determined by the method of Bradford using BSA as standard. Native gel electrophoresis and activity staining were carried out as described (20). Briefly, it followed the dienoyl-CoA-dependent reduction of 3-(4',5'-dimethylthiazol-2-yl)-2,4-diphenyltetrazolium bromide in the presence of DCPIP. Excised bands were incubated at 95 °C in 20 μ L SDS sample buffer for 10 min and analyzed by SDS/PAGE (12.5% wt/vol).

ACKNOWLEDGMENTS. We thank Nasser Gad'on and Georg Fuchs (University of Freiburg) for help with the cultivation, Rainer Wennrich (University of Leipzig) for ICP-MS analysis, and Gary Sawers (University of Halle, Halle, Germany) for careful proofreading the manuscript. This work was funded by the Deutsche Forschungsgemeinschaft Grants BO 1565/5-2 and BO 1565/10-1.

- Boll M, Fuchs G, Heider J (2002) Anaerobic oxidation of aromatic compounds and hydrocarbons. *Curr Opin Chem Biol* 6:604–611.
- Carmona M, et al. (2009) Anaerobic catabolism of aromatic compounds: A genetic and genomic view. *Microbiol Mol Biol Rev* 73:71–133.
- Fuchs G (2008) Anaerobic metabolism of aromatic compounds. *Ann NY Acad Sci* 1125:82–99.
- Gibson J, Harwood CS (2002) Metabolic diversity in aromatic compound utilization by anaerobic microbes. *Annu Rev Microbiol* 56:345–369.
- Boll M (2005) Dearomatizing benzene ring reductases. *J Mol Microbiol Biotechnol* 10:132–142.
- Boll M (2005) Key enzymes in the anaerobic aromatic metabolism catalysing Birch-like reductions. *Biochim Biophys Acta* 1707:34–50.
- Boll M, et al. (2000) Nonaromatic products from anoxic conversion of benzoyl-CoA with benzoyl-CoA reductase and cyclohexa-1,5-diene-1-carbonyl-CoA hydratase. *J Biol Chem* 275:21889–21895.
- Peters F, Shinoda Y, McInerney MJ, Boll M (2007) Cyclohexa-1,5-diene-1-carbonyl-coenzyme A (CoA) hydratases of *Geobacter metallireducens* and *Syntrophus aciditrophicus*: Evidence for a common benzoyl-CoA degradation pathway in facultative and strict anaerobes. *J Bacteriol* 189:1055–1060.
- Boll M, Fuchs G (1995) Benzoyl-coenzyme A reductase (dearomatizing), a key enzyme of anaerobic aromatic metabolism. ATP dependence of the reaction, purification and some properties of the enzyme from *Thauera aromatica* strain K172. *Eur J Biochem* 234:921–933.
- Boll M, Fuchs G, Meier C, Trautwein A, Lowe DJ (2000) EPR and Mossbauer studies of benzoyl-CoA reductase. *J Biol Chem* 275:31857–31868.
- Boll M, Albracht SS, Fuchs G (1997) Benzoyl-CoA reductase (dearomatizing), a key enzyme of anaerobic aromatic metabolism. A study of adenosinetriphosphatase activity, ATP stoichiometry of the reaction and EPR properties of the enzyme. *Eur J Biochem* 244:840–851.
- Buckel W, Hetzel M, Kim J (2004) ATP-driven electron transfer in enzymatic radical reactions. *Curr Opin Chem Biol* 8:462–467.
- Thiele B, Rieder O, Golding BT, Müller M, Boll M (2008) Mechanism of enzymatic Birch reduction: Stereochemical course and exchange reactions of benzoyl-CoA reductase. *J Am Chem Soc* 130:14050–14051.
- Mobitz H, Boll M (2002) A Birch-like mechanism in enzymatic benzoyl-CoA reduction: A kinetic study of substrate analogues combined with an ab initio model. *Biochemistry* 41:1752–1758.
- McInerney MJ, et al. (2007) The genome of *Syntrophus aciditrophicus*: Life at the thermodynamic limit of microbial growth. *Proc Natl Acad Sci* 104:7600–7605.
- Wischgoll S, et al. (2005) Gene clusters involved in anaerobic benzoate degradation of *Geobacter metallireducens*. *Mol Microbiol* 58:1238–1252.
- Butler JE, et al. (2007) Genomic and microarray analysis of aromatics degradation in *Geobacter metallireducens* and comparison to a *Geobacter* isolate from a contaminated field site. *BMC Genomics* 8:180.
- Peters F, Rother M, Boll M (2004) Selenocysteine-containing proteins in anaerobic benzoate metabolism of *Desulfococcus multivorans*. *J Bacteriol* 186:2156–2163.
- Heintz D, et al. (2009) Membrane proteome analysis reveals novel benzoate induced proteins in *G. metallireducens*. *Mol Cell Prot* 8:2159–2169.
- Thiele B, et al. (2008) Aromatizing cyclohexa-1,5-diene-1-carbonyl-coenzyme A oxidase. Characterization and its role in anaerobic aromatic metabolism. *J Biol Chem* 283:20713–20721.
- Chan MK, Mukund S, Kletzin A, Adams MW, Rees DC (1995) Structure of a hyperthermophilic tungstopterin enzyme, aldehyde ferredoxin oxidoreductase. *Science* 267:1463–1469.
- Roy R, Menon AL, Adams MW (2001) Aldehyde oxidoreductases from *Pyrococcus furiosus*. *Methods Enzymol* 331:132–144.
- Hille R (2005) Molybdenum-containing hydroxylases. *Arch Biochem Biophys* 433:107–116.
- Hu Y, Faham S, Roy R, Adams MW, Rees DC (1999) Formaldehyde ferredoxin oxidoreductase from *Pyrococcus furiosus*: The 1.85 Å resolution crystal structure and its mechanistic implications. *J Mol Biol* 286:899–914.
- Wardman G (1989) Reduction potentials of one-electron couples NADH-dye reductase and a non-haem iron protein involving free radicals in aqueous solution. *J Phys Chem Ref Data* 18:1637–1755.
- Herrmann G, Jayamani E, Mai G, Buckel W (2008) Energy conservation via electron-transferring flavoprotein in anaerobic bacteria. *J Bacteriol* 190:784–791.
- Li F, et al. (2008) Coupled ferredoxin and crotonyl coenzyme A (CoA) reduction with NADH catalyzed by the butyryl-CoA dehydrogenase/Etf complex from *Clostridium kluyveri*. *J Bacteriol* 190:843–850.
- Lovley DR, et al. (1993) *Geobacter metallireducens* gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals. *Arch Microbiol* 159:336–344.
- Schachter D, Taggart JV (1953) Benzoyl coenzyme A and hippurate synthesis. *J Biol Chem* 203:925–934.
- White H, Strobl G, Feicht R, Simon H (1989) Carboxylic acid reductase: A new tungsten enzyme catalyses the reduction of non-activated carboxylic acids to aldehydes. *Eur J Biochem* 184:89–96.
- Lovenberg W, Buchanan BB, Rabinowitz JC (1963) Studies on the chemical nature of clostridial ferredoxin. *J Biol Chem* 238:3899–3913.
- Beinert H (1983) Semi-micro methods for analysis of labile sulfide and of labile sulfide plus sulfane sulfur in unusually stable iron-sulfur proteins. *Anal Biochem* 131:373–378.
- Gallien S, et al. (2009) Ortho-proteogenomics: Multiple proteomes investigation through orthology and a new MS-based protocol. *Genome Res* 19:128–135.

Supplemental methods

In gel digestion and mass spectrometry analysis.

Gel bands of interest were cut and dried under vacuum. In-gel digestion was performed with an automated protein digestion system, MassPREP Station (Micromass, Manchester, UK). The gel slices were washed three times in a mixture containing 25 mM NH_4HCO_3 :ACN [1:1, v/v]. The cysteine residues were reduced by 50 μl of 10 mM dithiothreitol at 57°C and alkylated by 50 μl of 55 mM iodacetamide. After dehydration with acetonitrile, the proteins were cleaved in the gel with 40 μl of 12.5 ng/ μl of modified porcine trypsin (Promega, Madison, WI, USA) in 25 mM NH_4HCO_3 at room temperature for 14 hours. The resulting tryptic peptides were extracted with 60% acetonitrile in 0.5% formic acid, followed by a second extraction with 100% (v/v) acetonitrile.

The resulting peptide extracts were analyzed by nanoLC-MS/MS on an Agilent 1100 Series HPLC-Chip/MS system (Agilent Technologies, Palo Alto, USA) coupled to an HCT Ultra ion trap (Bruker Daltonics, Bremen, Germany) for the SDS-PAGE analysis of the enzyme fraction. Chromatographic separations were conducted on a chip containing a Zorbax 300SB-C18 (75 μm inner diameter \times 43 mm) column and a Zorbax 300SB-C18 (40 nL) enrichment column (Agilent Technologies). Peptide mixtures were loaded on the Zorbax 300SB-C18 (40 nL) enrichment column using 0.1% formic acid at 3.75 $\mu\text{L min}^{-1}$. After washing, the peptides were eluted with a gradient 4-40% acetonitrile in 0.1% formic acid delivered over 7 min at a flow rate of 300 nL min^{-1} through the Zorbax 300SB-C18 (75 μm inner diameter \times 43 mm) analytical column. HCT Ultra ion trap was externally calibrated with standard compounds. The general mass spectrometric parameters were as follows: capillary voltage, -1750V; dry gas, 3 liters/min; dry temperature, 300 °C. The system was operated with automatic switching between MS and MS/MS modes using. The MS scanning was performed in the standard-enhanced resolution mode at a scan rate of 8,100 m/z per second with an aimed ion charge control of 100,000 in a maximal fill time of 200 ms and a total of 4 scans were averaged to obtain MS spectrum. The three most abundant peptides and preferentially doubly charged ions were selected on each MS spectrum for further isolation and fragmentation. The MS/MS scanning was performed in the ultrascan resolution mode at a scan rate of 26,000 m/z per second with an aimed ion charge control of 300,000 and a total of 6 scans were averaged to obtain MS/MS spectrum. The complete system was fully controlled by ChemStation Rev. B.01.03 (Agilent Technologies) and EsquireControl 6.1 Build 78 (Bruker Daltonics) softwares. Mass data collected during LC-MS/MS analyses were processed using the software tool DataAnalysis 3.4 Build 169 and converted into *.mgf files

The MS/MS data were analyzed using the MASCOT 2.2.0. algorithm (Matrix Science, London, UK) for search against an in-house generated protein database composed of protein sequences of *G. metallireducens* downloaded from <http://www.ncbi.nlm.nih.gov/sites/entrez> (on December 20, 2007) concatenated with reversed copies of all sequences (2 \times 3,532 entries). For ion trap analyses, spectra were searched with a mass tolerance of 0.5 Da for MS and MS/MS data, allowing a maximum of one missed cleavage with trypsin and with carbamidomethylation of cysteines and oxidation of methionines specified as variable modifications. Protein identifications were validated when at least three peptides with high quality MS/MS spectra (less than 12 points below Mascot's threshold score of identity at 95% confidence level) were detected. No protein was identified in reversed sequences highlighting the high reliability of the identifications.

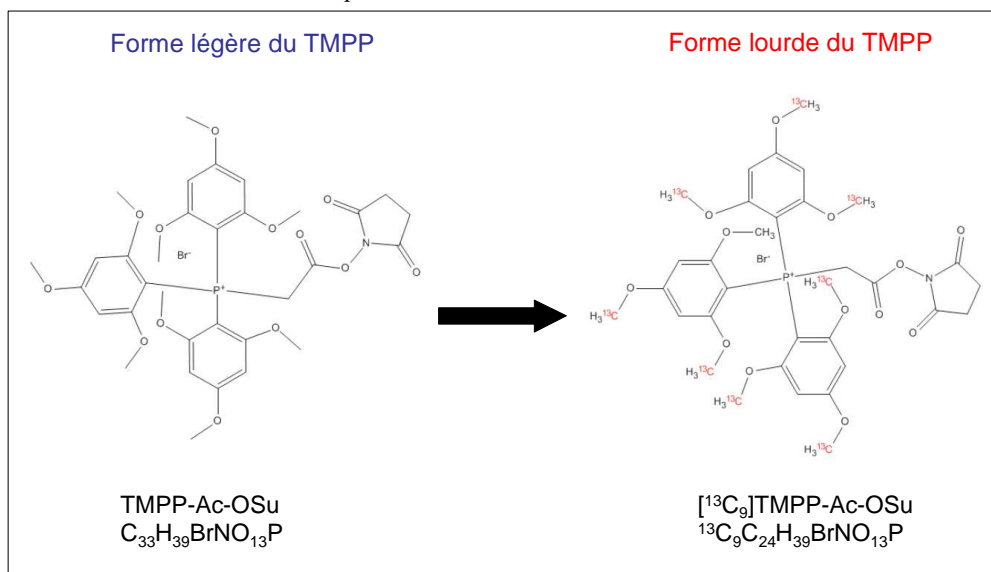
Chapitre 3 : Validation et application d'une méthode de quantification protéomique avec marquage au $^{13}\text{C}_9$ -TMPP.

L'approche N-TOP développée pour améliorer la caractérisation des extrémités N-terminales des protéines s'est avérée efficace pour valider/corriger un grand nombre d'erreurs d'annotation des codons d'initiation des protéines chez *M. smegmatis* (Partie I des résultats). Pour profiter de ses avantages dans le contexte des études protéomiques différentielles, nous avons décidé de conférer une dimension quantitative à cette méthode en développant une stratégie de marquage isotopique différentiel.

1. Développement et validation de la méthode de quantification (« quantitative N-terminal Oriented Proteomics », « qN-TOP »)

La stratégie de marquage isotopique différentiel développée ici repose sur l'utilisation d'une forme lourde du réactif TMPP. Pour cela, nous avons opté pour substituer 9 atomes de Carbone « naturel » du réactif par 9 atomes de Carbone 13 (Figure 1). Ce choix permet d'obtenir une différence de masse suffisante pour séparer par spectrométrie de masse les massifs isotopiques des deux formes d'un peptide différentiellement marqué et est compatible avec les contraintes de la synthèse chimique du réactif. Cette synthèse a été réalisée par la société « Alsachim » (Strasbourg, France), spécialisée dans la synthèse à façon de molécules marquées aux isotopes stables. Le TMPP « naturel » joue le rôle de « marqueur léger » et la forme modifiée avec les carbones ^{13}C de « marqueur lourd ». Cette nouvelle stratégie qN-TOP (« quantitative N-Terminal Oriented Proteomics ») utilisant les deux formes du réactif doit permettre de réaliser la quantification relative des protéines marquées des deux échantillons à comparer.

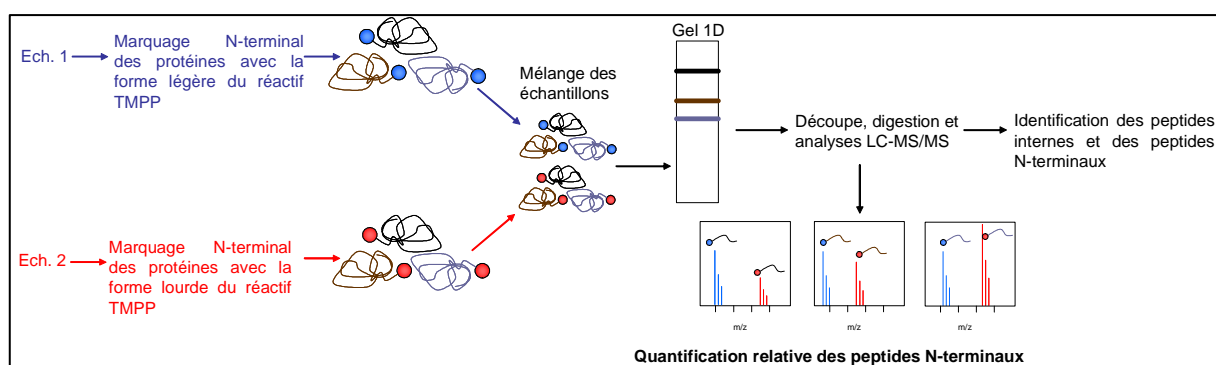
Figure 1 : Elaboration de la forme lourde du TMPP par substitution de 9 atomes de Carbone « naturel » du réactif par 9 atomes de Carbone 13



1.1. Principe

La nouvelle approche qN-TOP est basée sur le marquage isotopique différentiel des extrémités N-terminales des protéines de deux extraits à comparer par le TMPP lourd et le TMPP léger (Figure 2). Les extraits protéiques marqués différemment sont ensuite mélangés puis séparés par gel 1D. L'étape gel 1D reste indispensable dans notre stratégie pour éliminer les excès de réactifs et de ses composés de dégradation. Après découpe des bandes et digestion trypsique, l'ensemble des peptides de digestion est analysé par LC-MS/MS. Comme pour la stratégie N-TOP, les analyses permettent d'identifier les peptides internes de la protéine élués dans la première partie du gradient chromatographique et d'identifier les peptides N-terminaux marqués au TMPP dans la seconde partie du gradient chromatographique. De plus, les peptides N-terminaux issus des 2 échantillons peuvent être quantifiés de manière relative par comparaison des intensités des signaux MS correspondant au peptide N-terminal marqué avec la forme légère du réactif et au peptide N-terminal marqué avec la forme lourde.

Figure 2 : Principe de l'approche qN-TOP

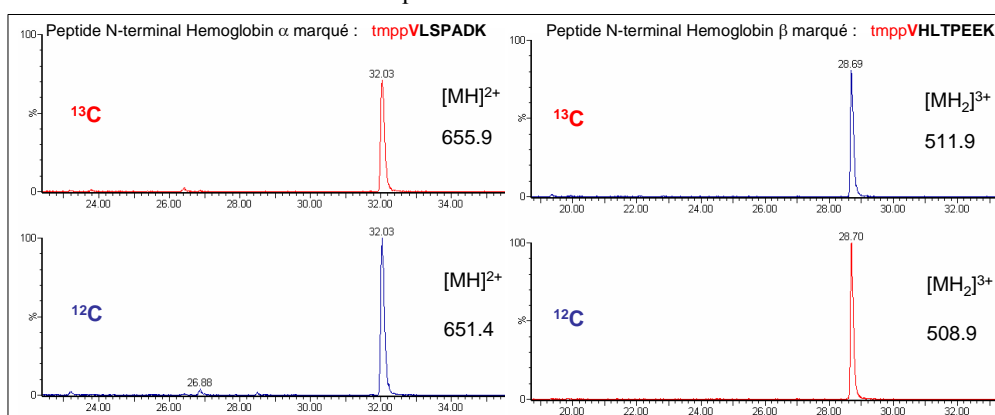


1.2. Validation de la méthode de quantification qN-TOP sur des protéines modèles

Nous avons procédé à la validation de la méthode de quantification qN-TOP sur des protéines modèles. Pour cela, nous avons dans un premier temps préparé un mélange de 2 protéines modèles : les chaînes alpha et beta d'hémoglobine purifiées. Le mélange a ensuite été divisé en deux fractions et chacune des fractions a été soumise indépendamment au marquage N-terminal avec le TMPP léger ou lourd. Ces deux fractions ont été mélangées dans différents ratios compris entre 0.1 et 10 (ratio théorique $^{12}\text{C}/^{13}\text{C}$ des peptides N-terminaux marqués au TMPP) et les différentes préparations protéiques ont été séparées par gel 1D, digérées à la trypsine puis analysées par nanoLC-MS/MS sur un système nanoAcquity UPLC couplé à un Q-TOF Synapt (Waters). Les préparations et les analyses ont été répétées plusieurs fois.

Nous avons pu vérifier que les peptides N-terminaux différenciellement marqués étaient parfaitement co-élus dans nos analyses (Figure 3).

Figure 3 : Comparaison des chromatogrammes d'extraits d'ions des peptides N-terminaux différenciellement marqués



1.2.1. Comparaison de plusieurs modes d'analyse et de traitement pour la quantification relative des peptides N-terminaux

1.2.1.1. Mode d'acquisition automatique standard

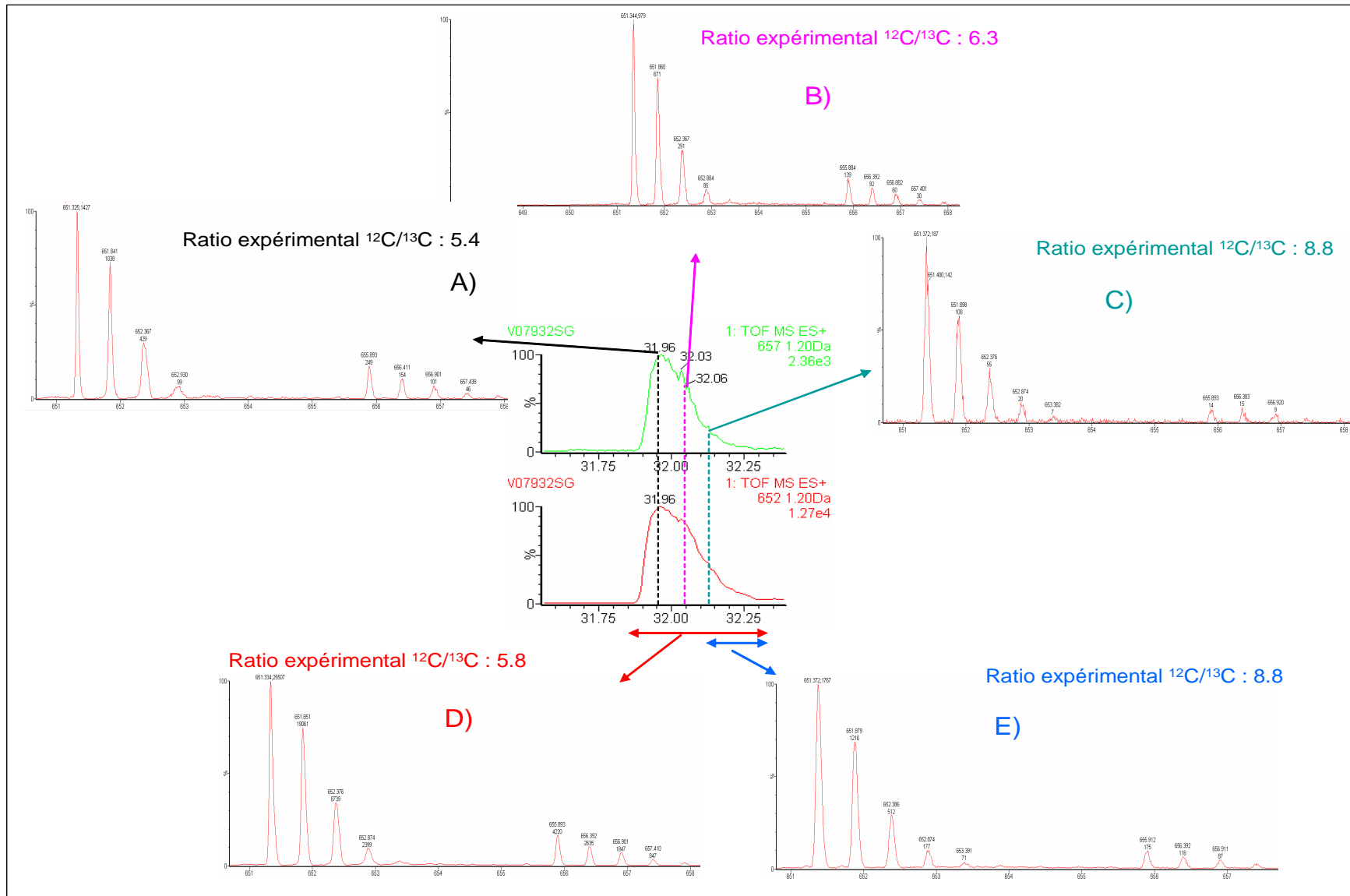
Dans un premier temps, la quantification relative du peptide N-terminal de la chaîne α d'hémoglobine différenciellement marqué a été réalisée en comparant les chromatogrammes d'ions extraits (englobant l'ensemble du massif isotopique) du peptide différenciellement marqué analysé en mode d'acquisition automatique nanoLC-MS/MS (mode d'analyse standard). Ce traitement a permis d'observer une forte limitation de la gamme de linéarité de l'instrument. Ainsi, par exemple, pour le mélange préparé dans un ratio théorique $^{12}\text{C}/^{13}\text{C}$ de 10, un ratio expérimental de 5.4 a été obtenu au

sommet du pic chromatographique (intensité du pic de base : 1427 coups/s) du peptide N-terminal de l'hémoglobine chaîne α (Figure 4 A). Lorsque la comparaison des sommes des intensités des pics des massifs isotopiques du peptide différenciellement marqué est réalisée à différents instants après le sommet du pic chromatographique, le ratio expérimental $^{12}\text{C}/^{13}\text{C}$ obtenu augmente. Ainsi, un ratio de 6.3 (Figure 4 B) a pu être mesuré environ 6 secondes après le sommet du pic (intensité du pic de base : 979 coups/s) et un ratio de 8.8 (Figure 4 C) a été mesuré environ 12 secondes après le sommet (intensité du pic de base : 187 coups/s). Etant donné que le peptide marqué avec la forme lourde du TMPP est parfaitement co-élué avec le peptide marqué avec la forme légère du TMPP, cette augmentation du ratio expérimental à mesure que l'intensité diminue confirme ce problème de linéarité de la mesure.

Si la comparaison des sommes des intensités des pics des massifs isotopiques du peptide différenciellement marqué est réalisée sur toute la durée du pic chromatographique, le ratio $^{12}\text{C}/^{13}\text{C}$ expérimental obtenu est également bien inférieur au ratio théorique : 5.8 au lieu de 10 (Figure 4 D). Ce mode de traitement n'est donc pas adapté pour la réalisation d'une quantification relative fiable. Par contre, si cette comparaison est réalisée sur la fin du pic chromatographique (spectres MS pour lesquelles l'intensité des signaux considérés est inférieure à 250 c/s), on obtient un ratio expérimental plus juste : 8.8 (Figure 4 E).

Validation et application d'une méthode de quantification protéomique avec marquage au $^{13}\text{C}_9$ -TMPP

Figure 4 : Comparaison des spectres MS du peptide N-terminal de la chaîne α d'hémoglobine différenciellement marqué à différents instant de son élution. Les chromatogrammes d'ions extraits du peptide marqué au TMPP lourd et léger sont indiqués au centre. A) Spectre MS obtenu au sommet du pic chromatographique du peptide. B) Spectre MS obtenu ~6s après le sommet. C) Spectre MS obtenu ~12s après le sommet. D) Spectre MS résultant de la somme des spectres obtenus sur toute la durée du pic chromatographique. E) Spectre MS résultant de la somme des spectres obtenus sur la dernière partie du pic chromatographique (spectres MS pour lesquelles l'intensité des signaux considérés est inférieure à 250 c/s)



1.2.1.2. Mode d'acquisition pDRE

Le spectromètre de masse utilisé pour ces analyses (Synapt Waters) dispose d'un mode d'analyse particulier, le mode « pDRE » (« programmable Dynamic Range enhancement) dans lequel sont alternées :

- Des mesures MS pour lesquelles la totalité des ions présents dans l'interface du spectromètre sont transmis.
- Des mesures MS pour lesquelles seule une fraction des ions présents est transmise (5 % dans nos analyses).

La comparaison des mesures MS dans les 2 conditions permet de corriger les intensités des différents ions saturés dans les spectres MS. Nous avons donc dans un deuxième temps ré-analysé les différents échantillons avec le mode pDRE pour essayer d'améliorer la gamme de linéarité de la mesure de quantification. Cette fois, pour le mélange préparé dans un ratio théorique $^{12}\text{C}/^{13}\text{C}$ de 10, on obtient au sommet du pic chromatographique (intensité du pic de base : 1047 coups/s) du peptide N-terminal de la chaîne α d'hémoglobine un ratio expérimental de 10.7. Le mode pDRE semble donc être adapté pour améliorer la gamme de linéarité de l'instrument.

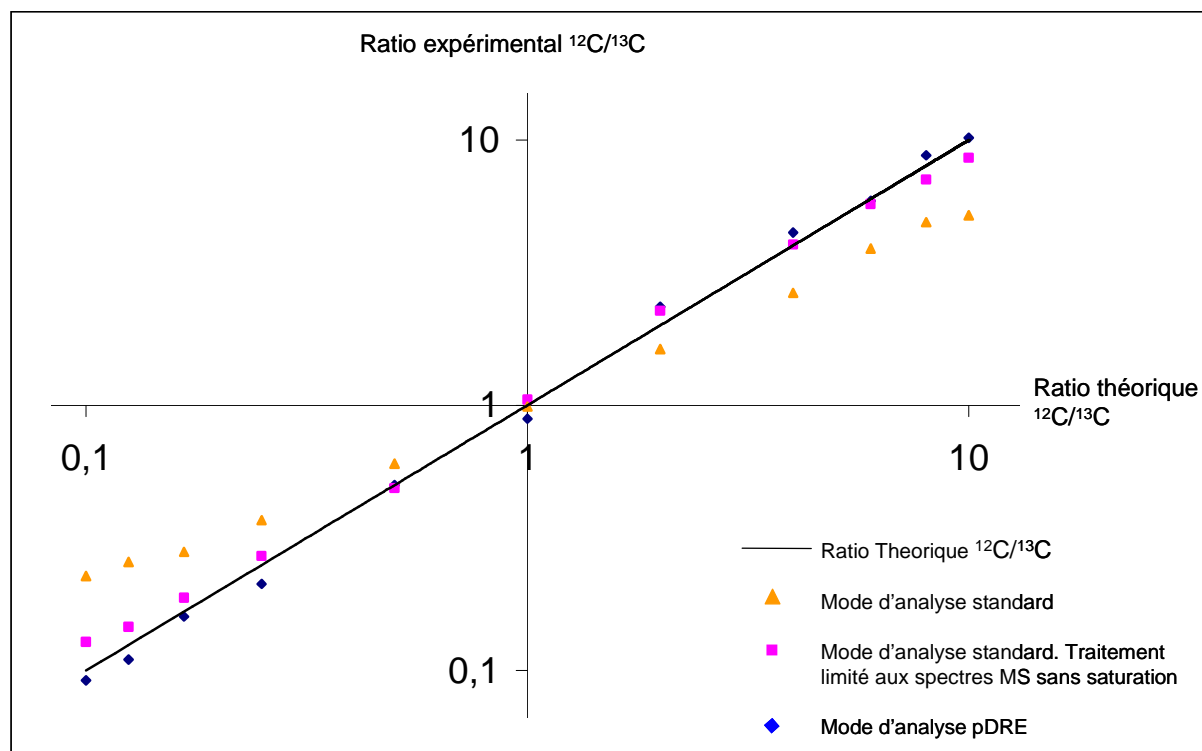
Nous avons donc procédé à la quantification relative du peptide N-terminal différentiellement marqué de la chaîne α d'hémoglobine dans tous les mélanges (ratios compris entre 0.1 et 10) en comparant les 3 types de procédures suivants :

- Comparaison en **mode d'analyse standard** des sommes des intensités des pics des massifs isotopiques du peptide différentiellement marqué sur **l'ensemble de son pic chromatographique**.
- Comparaison en **mode d'analyse standard** des sommes des intensités des pics des massifs isotopiques du peptide différentiellement marqué sur l'ensemble des **spectres MS** pour lesquels l'intensité des pics considérés était **inférieure au seuil de saturation** (250 coups/s).
- Comparaison en **mode d'analyse pDRE** des sommes des intensités des pics des massifs isotopiques du peptide différentiellement marqué sur **l'ensemble de son pic chromatographique**.

La comparaison des 3 types de procédure est illustrée en Figure 5. Il apparaît que le mode d'analyse standard traité sans précaution particulière conduit à une mauvaise linéarité ($R < 0.992$) et à une erreur importante sur les quantifications (jusqu'à plus de 50 %) alors qu'avec un traitement limité aux spectres ne présentant pas de saturation, on obtient une meilleure linéarité ($R \sim 0.997$) et une plus faible erreur (erreur < 30 %). Enfin, comme prévu, les analyses réalisées en mode pDRE présentent la meilleure linéarité ($R \sim 0.998$) et la meilleure justesse (erreur < 20 %). Toutefois, ces bénéfices sont obtenus au détriment de la vitesse d'acquisition et de traitement des mesures. En effet, les durées des cycles d'acquisition MS et MS/MS sont presque doublées avec ce mode d'analyse. Celui-ci n'est donc pas adapté pour l'analyse protéomique haut-débit pour laquelle un grand nombre de spectres MS/MS doit pouvoir être généré. Le mode d'analyse standard est donc requis pour l'application de la méthode

de qN-TOP sur un échantillon biologique. Par conséquent, la possibilité de réaliser de manière satisfaisante les expériences de quantification en mode d'analyse standard avec un traitement limité aux spectres ne présentant pas de saturation sera investigué dans la suite du travail.

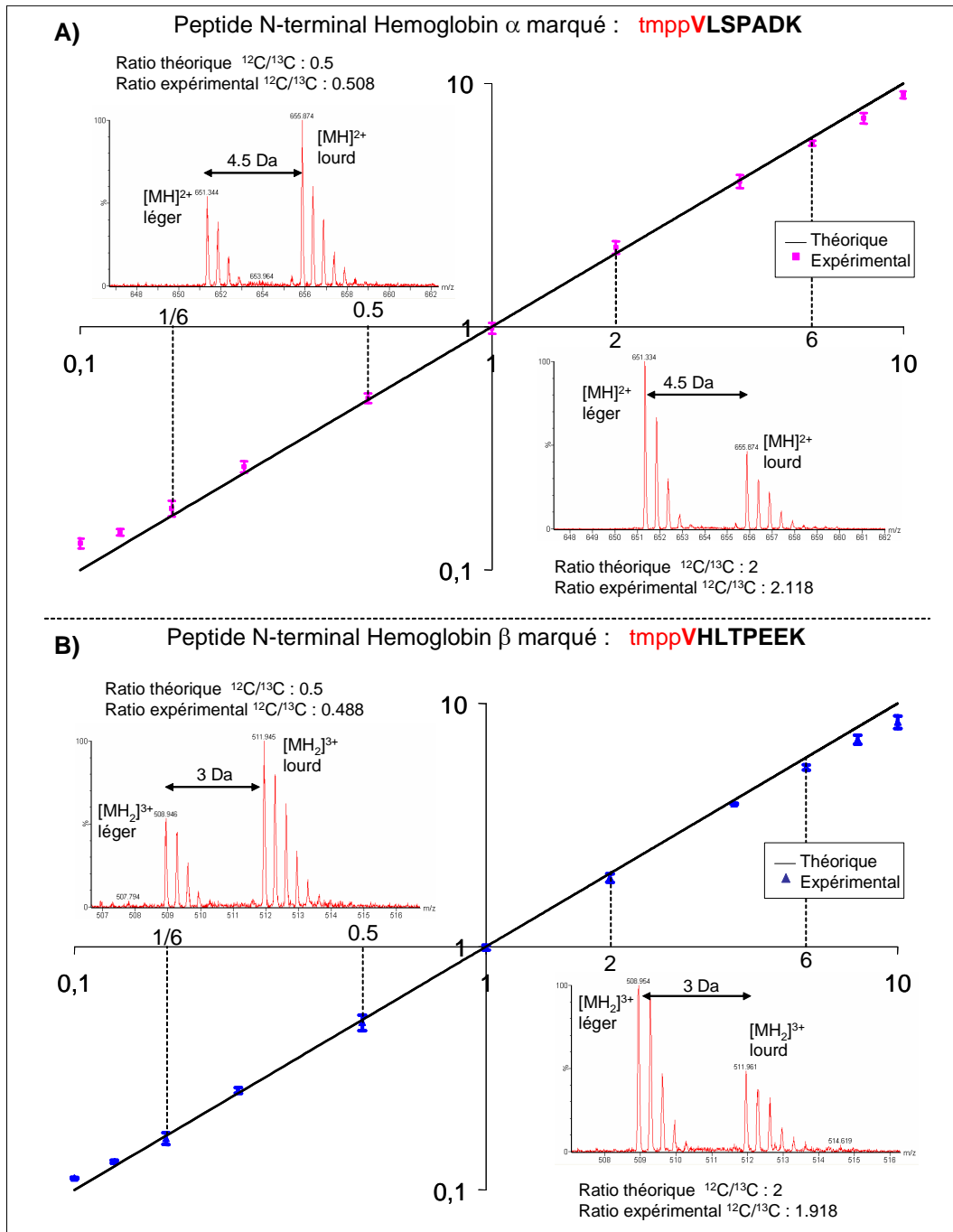
Figure 5 : Comparaison des 3 types de procédure (mode d'analyse et traitement) utilisées pour la quantification relative du peptide N-terminal de la chaîne α d'hémoglobine différemment marqué au TMPP.



1.2.2. Linéarité, justesse et reproductibilité de la méthode de quantification qN-TOP

En procédant à l'analyse en mode standard et au traitement limité aux spectres MS sans saturation pour les différents réplicats, nous avons observé une linéarité et une justesse satisfaisante pour la quantification des 2 peptides N-terminaux pour des ratios compris entre 1/6 et 6 (coefficients de corrélation >0.999 , erreur $<10\%$). Pour les ratios en dehors de cette gamme, on observe une erreur plus importante ($<30\%$) (Figure 4). La reproductibilité est satisfaisante puisque les écarts-types relatifs sont tous inférieurs à 10% .

Figure 4 : Evaluation de la justesse, la linéarité et la reproductibilité des ratios d'abondance des peptides N-terminaux des chaînes α (A) et β (B) d'hémoglobine différemment marqués



Comme les résultats obtenus sur le mélange des 2 protéines modèles étaient corrects, nous avons complexifié l'échantillon en ajoutant deux protéines standard supplémentaires (fsmR et myoglobine) dans notre mélange de départ et en reproduisant l'expérience pour des ratios compris entre 1.1 et 18.2. Les peptides N-terminaux des 2 nouvelles protéines sont également parfaitement co-élués (Figure 5). On retrouve aussi pour les 4 peptides N-terminaux une erreur de quantification

inférieure à 10 % pour les ratios inférieurs à 6 et un peu plus importante pour les ratios supérieurs à 6 (erreur < 30 %) (Tableau 1).

Figure 5 : Comparaison des chromatogrammes d'extraits d'ions des peptides N-terminaux différenciellement marqués de fsmsR et myoglobine

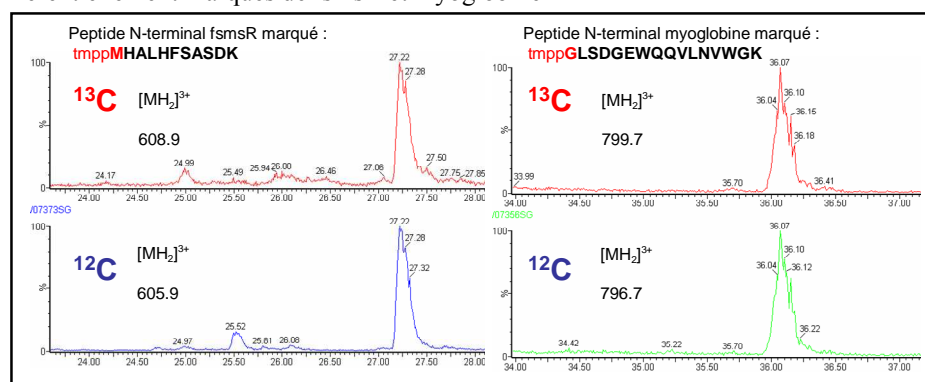


Tableau 1 : Quantification relative des peptides N-terminaux des 4 protéines modèles du mélange

Ratio théorique peptides N-ter Hémoglobine a / Hémoglobine b / myoglobine / fsmsR	Ratio expérimental peptides N-ter Hémoglobine a / Hémoglobine b / myoglobine / fsmsR	Erreur (%)
1.10 / 1.10 / 1.10 / 1.10	1.16 / 1.05 / 1.02 / 1.12	5.5 / 4.5 / 7.3 / 1.8
2.30 / 2.30 / 2.30 / 2.30	2.21 / 2.18 / 2.23 / 2.22	3.9 / 5.2 / 3.0 / 3.5
8.90 / 8.90 / 8.90 / 8.90	7.91 / 7.83 / 6.53 / 8.42	11.1 / 12.0 / 26.6 / 5.4
18.20 / 18.20 / 18.20 / 18.20	14.15 / 12.93 / 12.85 / 15.63	22.3 / 29.0 / 29.4 / 14.1

Les résultats de quantification obtenus sur l'application de la stratégie qN-TOP à des mélanges de protéines modèles se sont avérés corrects mais seulement dans une gamme de linéarité restreinte (ratios compris entre 1/6 et 6). Le problème de faible gamme de linéarité obtenu sur l'analyse des protéines modèles conjugué au mode de traitement manuel des analyses, qui exclue toute automatisation, va encore s'amplifier pour le traitement quantitatif de « vrais » échantillons biologiques. En effet, on peut s'attendre à observer un bruit de fond plus élevé pour l'analyse des échantillons biologiques qui va compliquer le traitement quantitatif des peptides N-terminaux différenciellement marqués. Ainsi, pour la quantification, il sera nécessaire de ne considérer que les signaux MS des peptides différenciellement marqués qui :

- ne présentent pas de saturation (intensité < 250 c/s)
- qui présentent un rapport signal/bruit suffisant pour les distinguer du bruit de fond

Or, ces deux conditions sont difficiles à remplir simultanément lors de l'analyse d'échantillons biologiques. La non-satisfaction de ces conditions conduira le plus souvent à la sous-estimation des réelles différences entre les échantillons.

Les limitations de linéarité observées ici, et qui sont également à l'origine des problèmes de traitement des signaux pour la quantification, ne sont pas liées à la stratégie qN-TOP mais sont intrinsèquement liées à l'instrument utilisé : le Q-TOF Synapt de Waters. En effet, nous avons vu avec

l'utilisation du mode pDRE que la stratégie qN-TOP pouvait conduire à des quantifications satisfaisantes sur une plus grande gamme de linéarité et avec un traitement du signal facilité. Les problèmes de gamme de linéarité restreinte sont le plus souvent rencontrés sur les instruments de type Q-TOF utilisant un détecteur avec TDC (« Time-to-Digital Converter »), ce qui est le cas du spectromètre de masse Synapt de Waters utilisé dans cette étude.

2. Application de la stratégie qN-TOP à l'étude différentielle de processus protéolytiques

Malgré les limitations instrumentales mises en évidence dans le paragraphe précédent et pénalisant la qualité de la quantification différentielle pouvant être réalisée avec la stratégie qN-TOP, nous avons tout de même appliqué la stratégie à une étude différentielle de processus protéolytiques. Cette application est réalisée dans le but d'évaluer l'efficacité de la stratégie pour identifier les sites protéolytiques d'un protéome complexe et de vérifier que la stratégie ne présente pas de biais sur les identifications en fonction du type de marquage (lourd ou léger). Cette application devra être réalisée par la suite sur un instrument présentant de meilleures performances essentiellement en terme de gamme dynamique du détecteur pour pouvoir tirer pleinement parti du potentiel de la stratégie qN-TOP d'un point de vue « quantification ».

La stratégie qN-TOP, qui permet à la fois de faciliter la caractérisation des extrémités N-terminales des protéines et de quantifier les peptides N-terminaux, est utilisée dans cette étude car elle est particulièrement bien adaptée à l'étude des processus protéolytiques (processomes et dégradomes) qui génèrent des nouvelles extrémités N-terminales pour les protéines précurseurs subissant ces processus.

2.1. Contexte de l'étude

2.1.1. Les processus protéolytiques

La protéolyse est une des modifications post-traductionnelles des protéines les plus importantes. Elle consiste en l'hydrolyse irréversible des liaisons peptidiques et isopeptidiques et affecte toutes les protéines à certains moments de leur « cycle de vie ». La protéolyse peut conduire à la dégradation complète de la protéine ou seulement à des coupures spécifiques de la séquence. Dans ce dernier cas, cette protéolyse fortement contrôlée qu'on appelle processus protéolytique, permet aux protéases de moduler précisément les fonctions des protéines. Par des processus protéolytiques précis, les protéases participent à tous les processus biologiques comme la réplication de l'ADN, la prolifération, la différenciation et la migration des cellules, le développement neuronal, l'hémostase ou

encore l'immunité, la cicatrisation et l'apoptose [Barrett, 2004]. Les protéases sont également impliquées dans beaucoup de pathologies parmi lesquelles le cancer [Doucet et al., 2008].

Comme pour toutes les modifications post-traductionnelles, si la protéolyse n'est pas considérée dans les analyses protéomiques, une quantité considérable d'information est perdue et l'annotation fonctionnelle des composants du protéome est erronée. Il est donc nécessaire dans l'analyse protéomique d'identifier les modifications dans le N- et C-terminome, résultant souvent de la protéolyse, et de développer des approches efficaces pour l'analyse de ces sous-protéomes.

2.1.2. L'étude des processus protéolytiques

Les méthodes décrites pour l'enrichissement des peptides N-terminaux (Partie I des résultats, Chapitre 2. 1.1.2.) des protéines précurseurs sont de façon générale également bien adaptées pour l'enrichissement des « nouveaux » peptides N-terminaux générés par des processus protéolytiques. Toutefois, généralement, l'étude des processus protéolytiques, qu'ils soient générés *in vivo* en réponse à une condition biologique donnée ou *in vitro* par l'ajout d'une protéase dont on essaie d'identifier les substrats [Van Damme et al., 2005; Vande Walle et al., 2007; Lamkanfi et al., 2008; Van Damme et al., 2009], nécessite une dimension quantitative. L'ajout de cette dimension est généralement réalisé par l'utilisation de différentes formes isotopiques du réactif utilisé ou grâce à une stratégie de marquage par réaction enzymatique (digestion enzymatique dans un milieu H_2^{16}O ou H_2^{18}O) [Van Damme et al., 2005; Vande Walle et al., 2007].

2.1.3. Le modèle du jeûne prolongé : adaptations métaboliques

La masse et la composition corporelle d'un animal peuvent varier considérablement au cours de son cycle annuel en fonction de contraintes écophysiologiques comme la reproduction, l'hibernation et la mue. Le stockage des réserves énergétiques et leur mobilisation ultérieure sont donc des étapes essentielles à la survie de l'animal dans son milieu naturel, puisqu'elles permettent de subvenir aux besoins de l'organisme dans des situations où les apports alimentaires sont espacés dans le temps. Le jeûne prolongé d'un animal se caractérise par une mobilisation séquentielle des substrats énergétiques, hydrates de carbone (phase 1) puis lipides (phase 2) et enfin protéines (phase 3). Une meilleure connaissance des processus protéolytiques dans ces situations permettrait de mieux comprendre les mécanismes de gestion des réserves énergétiques en réponse à la privation de nourriture à long terme.

2.1.4. Objectif de l'étude

Nous avons appliqué la stratégie qN-TOP à l'étude du processome hépatique de rats soumis à un jeûne expérimental plus ou moins prolongé. Nous nous sommes intéressés au protéome hépatique dans cette étude car le foie étant un carrefour métabolique, le protéome hépatique est sans doute le plus susceptible de subir les « événements protéolytiques » générés en réponse au jeûne expérimental. Nous avons donc comparé les « processomes » hépatiques de rats nourris, en phase 2 et en phase 3 du jeûne prolongé. De plus, nous avons réalisé une étude différentielle par comptage des spectres (« Spectral Counting ») pour comparer les niveaux d'expression des protéines entre les 3 groupes. La mise en relation des quantifications différentielles des formes protéiques processées avec les niveaux d'expression des protéines « totales » (forme(s) processée(s) + forme non-processée) pourrait également permettre de déterminer d'éventuelles inductions de processus protéolytiques en réponse à la situation biologique donnée.

2.2. Stratégie d'analyse

2.2.1. Préparation des échantillons

Neuf rats ont été logés individuellement dans une chambre à température contrôlée (25 +/- 1°C) avec des photopériodes constantes (12 H lumière / 12 H obscurité) et avec accès libre à de l'eau et à un régime standard. Une fois atteinte la masse approximative de 270 g, les rats ont été divisés en 3 groupes au hasard. 3 rats ont été immédiatement sacrifiés comme animaux contrôle. Ces contrôles étaient dans un état post-absorptif avec un estomac encore plein comme il le fut noté durant l'échantillonnage des tissus. Ces rats contrôles sont par la suite référencés comme « animaux nourris ». Les autres rats ont été soumis au jeûne expérimental jusqu'en phase 2 ou 3. L'identification des phases 2 ou 3 du jeûne a été réalisée par pesée journalière [Cherel et al., 1991; Koubi et al., 1991] et l'état métabolique des rats en phase 2 ou 3 a été contrôlé par analyse des métabolites du plasma. 3 rats ont été sacrifiés en phase 2 du jeûne expérimental et 3 rats en phase 3. Le foie a été rapidement retiré des rats fraîchement tués et placé dans l'azote liquide avant stockage à -80°C.

Environ 300 mg de tissu hépatique issu de chaque prélèvement (3 rats nourris, 3 rats en phase 2 du jeûne, 3 rats en phase 3 du jeûne) ont ensuite été broyés dans l'azote liquide et les protéines ont été extraites dans 3.5 mL de la solution tampon de marquage au TMPP (Chapitre 2 de la partie I des résultats) qui contient des inhibiteurs de protéases. Les concentrations protéiques de chaque extrait ont été estimées par dosage Bradford (~15 µg/µL de protéine dans chaque extrait). Pour chaque préparation protéique, un prélèvement volumétrique correspondant à 80 µg de protéines a été réalisé et dédié à l'étude différentielle des protéomes hépatiques. Un autre prélèvement de 80 µg de protéines a été réalisé et dédié à l'étude différentielle des processomes hépatiques.

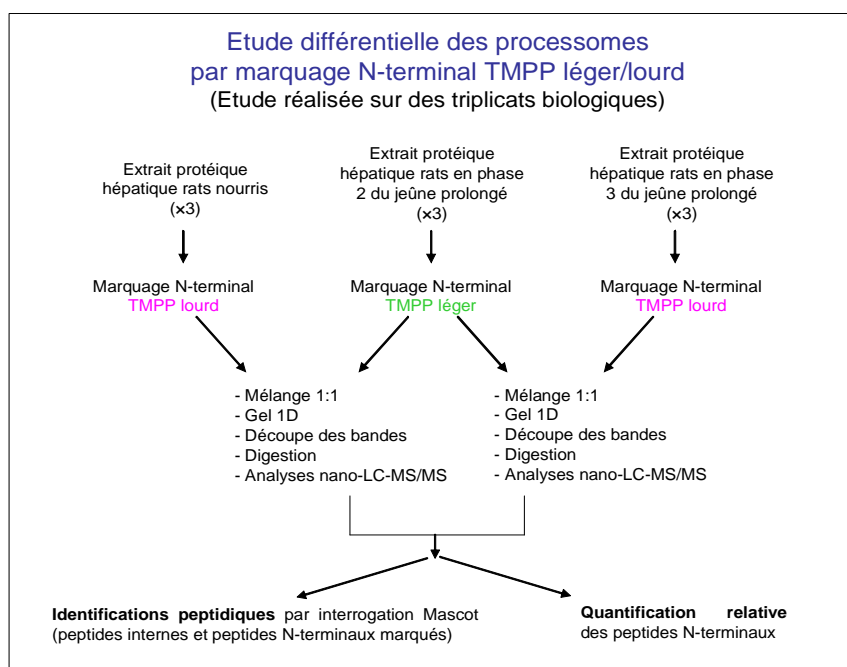
2.2.2. Etude différentielle des processomes hépatiques chez le rat dénutri

Les extraits protéiques hépatiques prélevés de chaque préparation (3 rats nourris, 3 rats en phase 2 du jeûne, 3 rats en phase 3 du jeûne) ont été soumis, après réduction/alkylation des cystéines, au protocole de préparation qN-TOP comme indiqué dans la Figure 6. Après marquage au TMPP léger ou lourd et mélange des extraits protéiques différenciellement marqués des différents groupes, les 3 échantillons ont ensuite été séparés par gel 1D. Chaque piste de gel 1D a été découpée systématiquement en 35 bandes. Après réduction, alkylation et digestion trypsique des protéines présentes dans les bandes, les peptides de digestion ont été extraits puis analysés par nanoLC-MS/MS sur un système nanoAcquity UPLC couplé à un Q-TOF Synapt (Waters). La complexité des mélanges peptidiques et le caractère relativement hydrophobe des peptides N-terminaux marqués au TMPP en RP-HPLC nous ont conduit à réaliser les séparations chromatographiques avec un gradient de modification de la composition de la phase mobile de pente douce (10-50 % acétonitrile sur 80 minutes). La fenêtre d'isolation des peptides à fragmenter a été paramétrée pour s'assurer de ne pas isoler un peptide marqué à la fois avec le réactif TMPP lourd et avec le réactif TMPP léger (fenêtre réduite à 2 m/z à la place des 3 m/z utilisés habituellement). Le détail des paramétrages des analyses nanoLC-MS/MS est décrit dans la partie expérimentale générale.

L'identification des peptides internes a été réalisée par interrogation Mascot des données MS/MS dans une version target-decoy de la banque protéique de rat téléchargée sur le site du NCBI avec une tolérance de 30 ppm sur la masse des précurseurs et 0.1 Da sur la masse des ions fragments, avec la carbamidomethylation obligatoire des cystéines, en autorisant une coupure manquée avec la digestion trypsique et l'oxydation des méthionines. Les seuils de score des identifications peptidiques ont été établis pour maintenir un FDR maximal de 1 % sur l'identification des protéines.

L'identification des peptides N-terminaux marqués au TMPP a été réalisée par interrogation Mascot des données MS/MS dans une version target-decoy de la banque protéique de rat téléchargée sur le site du NCBI avec une tolérance de 30 ppm sur la masse des précurseurs et 0.1 Da sur la masse des ions fragments, avec la carbamidomethylation obligatoire des cystéines, en autorisant une coupure manquée avec la digestion semi-trypsique, l'oxydation des méthionines, l'acétylation N-terminale des protéines, le marquage TMPP (léger ou lourd) en N-terminal des peptides, sur les lysines et sur les tyrosines. Les seuils de score des identifications peptidiques ont été établis pour maintenir un FDR maximal de 1 % sur l'identification des peptides marqués au TMPP.

Figure 6 : Stratégie d'analyse différentielle des processomes hépatiques par marquage N-terminal au TMPP léger/lourd



2.2.3. Etude différentielle des protéomes hépatiques chez le rat dénutri

Après réduction/alkylation des cystéines, les extraits protéiques hépatiques prélevés de chaque préparation (3 rats nourris, 3 rats en phase 2 du jeûne, 3 rats en phase 3 du jeûne) ont été séparés par gel 1D. Chacune des 9 pistes de gel 1D a été découpée systématiquement en 35 bandes. Après réduction, alkylation et digestion trypsique des protéines dans les bandes, les peptides de digestion ont été extraits puis analysés par nanoLC-MS/MS sur un système nanoHPLC-Chip (Agilent Series 1100) couplé à une trappe ionique HCT Ultra (Bruker Daltonics).

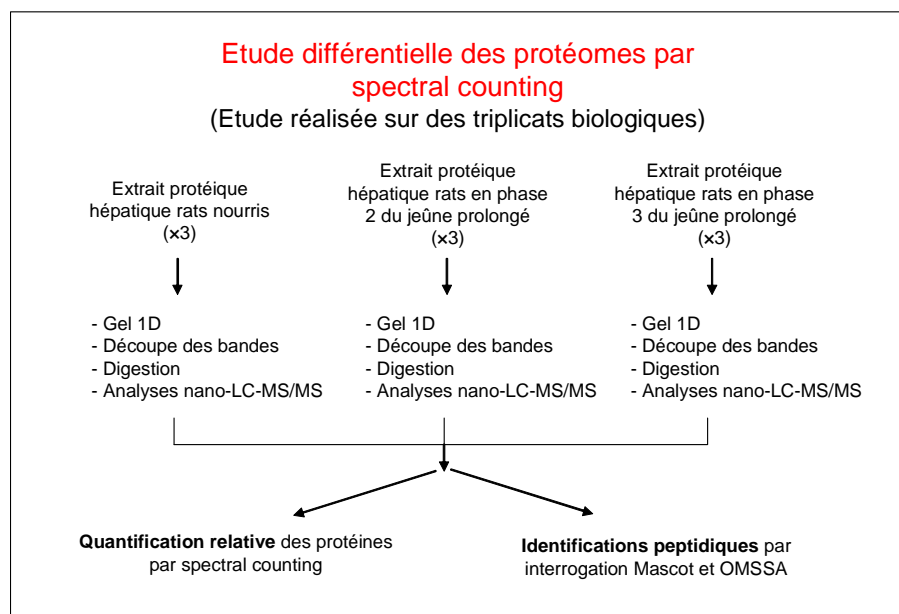
Les analyses chromatographiques ont été réalisées dans des conditions originales. En effet, le modificateur utilisé dans la phase mobile pour réaliser le gradient d'éluion des peptides n'a pas été l'acétonitrile mais le méthanol. Ce changement a été réalisé en réponse à la pénurie d'acétonitrile qui a touché les fournisseurs au cours de l'année 2008-2009. Pour s'assurer que ce changement ne pénalisait pas significativement les performances analytiques de notre système, nous avons comparé les analyses nanoLC-MS/MS réalisées les mêmes mélanges peptidiques plus ou moins complexes en utilisant soit le méthanol soit l'acétonitrile comme modificateur. L'utilisation du méthanol a nécessité la réoptimisation des gradients chromatographiques. Par contre, l'augmentation de la pression dans le système résultant de l'utilisation du méthanol n'a pas été critique (~50 % d'augmentation à la pression maximale relevée au cours de l'analyse) et l'efficacité d'ionisation électrospray des peptides s'est trouvée améliorée. Ces observations n'ont pas été documentées dans cette thèse mais une publication reprenant ces résultats ainsi que ceux obtenus sur un couplage nanoHPLC-Chip-tripe quadripôle en utilisant également le méthanol comme modificateur de la phase mobile chromatographique est en cours de rédaction. Le paramétrage du spectromètre de masse a également été optimisé pour les besoins de

l'étude. En effet, de la même manière que pour l'étude appliquée à *G. metallireducens* (Partie III des résultats, Chapitre 2), l'acquisition du spectromètre de masse a été paramétrée pour permettre une quantification par approche spectral counting de qualité tout en conservant une couverture du protéome analysé satisfaisante. Ainsi, la sélection des ions parents à fragmenter a été réalisée sur un mode de « semi-exclusion ». Le nombre de spectres MS et MS/MS moyennés pour donner un spectre final a été respectivement réduit à 3 et 2, soient les valeurs permettant l'attribution d'une séquence peptidique pour un nombre maximal de spectres MS/MS avec le couplage nanoLC-MS/MS utilisé dans l'étude. Enfin, le traitement post-analytique (génération des « peak lists ») a été réalisé sans moyenner les spectres MS/MS finaux réalisés sur le même précurseur pour ne pas induire de biais dans la quantification. Le traitement des données brutes a également été réalisé sans procéder à la détermination des états de charge des précurseurs pour ne pas induire de biais dans la comparaison des moteurs de recherche utilisés par la suite pour les identifications (car si Mascot utilise cette information dans le processus d'identification, ce n'est pas le cas d'OMSSA). Le détail des paramétrages des analyses nanoLC-MS/MS est décrit dans la partie expérimentale générale.

Les données MS/MS ont été soumises à une interrogation Mascot et à une interrogation OMSSA dans une version target-decoy de la banque protéique de rat téléchargée sur le site du NCBI avec une tolérance de 0.5 Da sur la masse des précurseurs et 0.5 Da sur la masse des ions fragments, avec la carbamidomethylation obligatoire des cystéines, en autorisant l'oxydation des méthionines et une coupure manquée avec la digestion trypsique. Les seuils de score des identifications peptidiques réalisées dans les recherches Mascot et OMSSA ont été établis pour maintenir un FDR maximal de 1 % sur l'identification des protéines.

La quantification relative des protéines entre les 3 groupes a été réalisée par une approche spectral counting (Figure 7).

Figure 7 : Stratégie d'analyse différentielle des protéomes hépatiques par spectral counting



2.3. Résultats préliminaires

2.3.1. Etude différentielle des processomes hépatiques chez le rat dénutri.

L'étude différentielle des processomes hépatiques chez le rat dénutri a permis d'appliquer la stratégie qN-TOP sur des échantillons biologiques complexes en étant bien conscient des limites instrumentales pour la quantification. En effet, nous avons vu précédemment que la faible gamme de linéarité du spectromètre de masse utilisé pénalisait la quantification relative des peptides N-terminaux différentiellement marqués et obligeait au traitement manuel des données. Les résultats préliminaires obtenus permettent tout de même d'évaluer la pertinence de l'approche qN-TOP pour de telles études.

2.3.1.1. Identification des protéines

Nous avons établi les scores seuils des identifications peptidiques pour obtenir un FDR d'environ 1 % sur l'identification des protéines. Pour les identifications de protéines avec au moins 2 peptides, chacune des identifications peptidiques doit être supérieure au score seuil d'identification Mascot « Score identité-10 ». Pour les identifications de protéines avec un seul peptide, les identifications peptidiques doivent être supérieures au score seuil d'identification Mascot « Score identité+17 ». Ces critères permettent de « valider » l'identification de 1757 protéines uniques (FDR~1.5 %).

2.3.1.2. Identification des processus protéolytiques

2.3.1.2.1. Les peptides marqués en N-terminal au TMPP

Dans cette étude, la procédure d'identification des peptides N-terminaux marqués au TMPP est différente de celle qui a été utilisée pour la validation/correction des codons d'initiation des protéines chez *M. smegmatis* (Partie I des résultats, Chapitre 2).

En effet, dans l'étude portant sur *M. smegmatis*, seuls les peptides N-terminaux qui permettaient de valider/corriger les codons d'initiation prédits des protéines étaient conservés dans la procédure de « validation » des identifications. Cette sélection était réalisée en écartant les potentiels peptides N-terminaux marqués au TMPP dont le premier acide aminé n'était pas codé par un codon d'initiation, ATG, GTG ou TTG, ou par le premier codon suivant un codon d'initiation ou pour lesquels d'autres peptides étaient identifiés en amont dans la même séquence protéique. Ce mode de sélection a permis d'améliorer les performances d'identification des peptides N-terminaux en limitant l'espace de recherche aux seuls peptides d'intérêt dans notre étude et d'éliminer la quasi-totalité des

peptides qui auraient pu présenter une réaction parasite mineure sur une lysine ou une tyrosine (Partie I des résultats, Chapitre 2. 3.4.2.4.).

En revanche, dans cette étude de processomes, l'utilisation de ce mode de sélection n'est pas pertinente car d'une part on ne peut préjuger des sites de coupure protéolytique et d'autre part une protéine peut être présente et détectée dans l'échantillon sous forme processée et non-processée. Par conséquent, ici, l'ensemble des peptides identifiés comme marqués en N-terminal au TMPP a été conservé pour identifier ces potentiels sites de coupure protéolytique. Pour s'assurer que les identifications des peptides potentiellement marqués en N-terminal au TMPP ne résultent pas d'un marquage aspécifique sur une lysine ou une tyrosine, les interrogations Mascot ont été effectuées en permettant la modification de ces 2 acides aminés au TMPP.

Les interrogations Mascot ont permis d'identifier 251 séquences peptidiques uniques marquées en N-terminal au TMPP sans ambigüité (FDR~2 %). Parmi ces séquences 64 correspondent au peptide N-terminal d'une protéine précurseur (premier acide aminé codé par le codon d'initiation ou le codon suivant) et 187 correspondent au peptide N-terminal d'une protéine processée. Seulement 64 séquences correspondant au peptide N-terminal d'une protéine précurseur ont été identifiées dans l'étude car une grande majorité des protéines sont acétylées en N-terminal chez les mammifères (~76 % des protéines complètement acétylées et 8 % partiellement acétylées chez l'homme [Arnesen et al., 2009]) et ne peuvent donc pas être modifiées par le TMPP.

Parmi les 187 séquences qui correspondent aux peptides N-terminaux de formes protéiques processées, 57 % correspondent à des processus protéolytiques dans les 50 premiers résidus de la protéine et semblent donc résulter de la coupure d'un peptide de transit ou signal (le plus souvent observée dans les 50 premiers résidus des protéines). Les 43 % restant correspondent à des processus protéolytiques après les 50 premiers résidus et semblent donc résulter de processus protéolytiques additionnels ou de la dégradation des protéines.

Sur les 251 séquences peptidiques uniques marquées en N-terminale au TMPP, 159 (64 %) ont été identifiées à la fois sous forme marquée au TMPP léger et sous forme marquée au TMPP lourd, 51 (20 %) uniquement sous forme marquée au TMPP léger et 41 (16 %) uniquement sous forme marquée au TMPP lourd. Cette distribution des identifications des séquences peptidiques marquées en N-terminal au TMPP indique qu'il ne semble pas y avoir de biais sur les identifications en fonction du type de marquage.

2.3.1.2.2. Les peptides marqués aspécifiquement au TMPP

Nous avons également relevé l'ensemble des spectres qui correspondent à un marquage TMPP aspécifique pour évaluer de manière semi-quantitative l'étendue de ces marquages sur l'analyse extensive d'un échantillon biologique complexe. Nous avons vu précédemment (Partie I des résultats, Chapitre 2. 1.2.) que la régiosélectivité du marquage TMPP de 95 % sur les amines libres N-terminales (pKa ~7.8) vis à vis des amines libres des lysines (pKa ~11) à un pH de 8.2 [Huang et al., 1999] induisait une réaction parasite mineure sur les lysines. Ici, nous avons relevé 1990 spectres

correspondant à des peptides présentant un marquage TMPP en N-terminal. Un traitement *in silico* de la banque protéique de rat révèle la possibilité d'obtenir approximativement 22 peptides marqués sur une lysine par peptide N-terminal. Si la réactivité du TMPP était identique en N-terminal et sur les lysines, on observerait 43780 (1990×22) spectres correspondant à des peptides présentant un marquage TMPP sur une lysine. Or, ici, seulement 1609 spectres correspondant à des peptides présentant un marquage TMPP sur une lysine ont été répertoriés et indiquent une spécificité de marquage d'environ 97 % (voisine des 95 % rapportée précédemment). La spécificité de la réaction en N-terminal est sans doute encore plus importante car, bien qu'il soit possible que certaines protéines soient présentes dans l'échantillon sous une forme clivée et une forme intacte, la grande majorité des protéines de l'échantillon sont acétylées en N-terminal et ne peuvent donc pas être modifiées par le TMPP. On a pu observer que les marquages aspécifiques n'étaient observés que sur les peptides très abondants car plus de 90 % des spectres sont issus de peptides entièrement tryptiques (qui représentent la grande majorité du signal des peptides dans une digestion tryptique) et plus d'1/3 de ces spectres sont issus de peptides appartenant aux 5 protéines majoritaires de l'échantillon. En ce qui concerne le marquage au TMPP sur les tyrosines, on observe encore 5 fois moins de spectres que pour le marquage sur les lysines. La réaction parasite de marquage partiel des tyrosines peut donc être évitée de manière relativement efficace par l'ajout d'hydroxylamine.

S'il est important de distinguer les peptides marqués au TMPP en N-terminal des peptides marqués au TMPP aspécifiquement qui pourraient conduire à des interprétations biologiques erronées, leur présence ne pénalise que de façon limitée l'identification des peptides réellement marqués en N-terminal et des peptides internes non modifiés. En effet, comme pour les peptides marqués en N-terminal, les peptides marqués aspécifiquement sont élués chromatographiquement après l'ensemble des peptides internes et ils ne représentent environ qu'1 % du signal issu de l'ensemble des peptides internes (plus de 150 000 spectres correspondant aux peptides internes non modifiés ont été relevés).

2.3.1.2.3. Perspectives

Plusieurs méthodes sont envisagées pour améliorer l'identification des peptides marqués en N-terminal au TMPP au niveau du marquage chimique et au niveau du traitement des données MS/MS.

Une première méthode consiste à modifier le mode opératoire en insérant une étape de « blocage » des amines libres des chaînes latérales des lysines des protéines avant de réaliser le marquage N-terminal au TMPP pour augmenter encore le caractère spécifique de cette dernière réaction. Ce « blocage » des lysines pourrait être réalisé par guanidination, une réaction décrite comme convertissant spécifiquement les lysines en homoarginines [Keough et al., 2000; Beardsley et al., 2002; Yamaguchi et al., 2008]. Par conséquent, les lysines ne sont plus du tout disponibles pour le marquage TMPP et il devrait être possible de s'affranchir complètement des quelques réactions

aspécifiques décrites précédemment. Toutefois, cette réaction de guanidination des lysines présente plusieurs inconvénients :

- Des réactions aspécifiques de guanidination sur les amines libre N-terminales ont également été rapportées [Beardsley et al., 2002] et pourraient pénaliser notre étude.
- L'ajout de cette étape de guanidination supplémentaire pourrait également interférer avec la réaction de marquage au TMPP (tampon, excès de réactif).
- La conversion des lysines en homoarginines empêche la trypsine d'hydrolyser les liaisons peptidiques en C-terminal des lysines modifiées. Par conséquent, lors de la digestion trypsique, seules les liaisons peptidiques en C-terminal des arginines pourront être hydrolysées. Les peptides générés, et notamment les peptides N-terminaux, présenteront donc une masse moléculaire plus élevée. Sachant que le marquage TMPP conduit aussi à une augmentation de masse significative des peptides N-terminaux, cette augmentation de masse induite par la guanidination pourrait également gêner l'identification des peptides N-terminaux par spectrométrie de masse.

Cette étape de « blocage » des amines libres des chaînes latérales des lysines des protéines avant marquage TMPP devra être testée et comparée à notre mode opératoire « habituel » pour évaluer son intérêt.

Les autres méthodes consistent à améliorer le processus d'identification des peptides marqués au TMPP lors des recherches des données MS/MS dans les banques de données. Une recherche séquentielle sur le même principe que celle décrite dans la Partie I des résultats, Chapitre 2. 3., pourrait être une alternative efficace. Dans ce cas, une sous-banque protéique générée à partir de l'ensemble des protéines identifiées avec au moins un peptide interne pourrait être utilisée pour les identifications des peptides marqués en N-terminal au TMPP et réduirait considérablement l'espace de recherche. Il sera également intéressant d'utiliser un deuxième moteur de recherche, par exemple OMSSA pour réaliser les identifications des peptides marqués au TMPP.

2.3.1.3. Quantification relative des peptides N-terminaux

La quantification relative des peptides N-terminaux différenciellement marqués n'a pas été réalisée à ce stade de l'étude en raison des problèmes de linéarité de quantification observés sur le spectromètre de masse utilisé. Ces problèmes auraient conduit à réaliser un traitement manuel, assez laborieux et susceptible d'aplanir les différences d'abondance des peptides N-terminaux entre les échantillons des différents groupes. Pour aller plus loin dans cette quantification, 2 possibilités s'offrent à nous : i) Nous pouvons procéder à de nouvelles analyses en suivant la même méthodologie mais en utilisant un spectromètre de masse présentant une plus large gamme de linéarité (par exemple pour ce qui concerne les instruments du laboratoire, soit une trappe ionique soit un autre Q-TOF équipé d'un détecteur avec ADC [« Analog-to-Digital Converter »] beaucoup plus performant en terme de gamme de linéarité ; ii) Nous pouvons tirer partie des identifications des processus

protéolytiques déjà réalisées dans l'étude pour réaliser une quantification ciblée des peptides N-terminaux détectés grâce à une approche MRM (« Multiple Reaction Monitoring ») appliquée sur le spectromètre de masse de type triple quadripôle récemment acquis au laboratoire (Agilent 6410 triple Quad LC/MS). A ce stade, les 2 possibilités sont en cours d'étude.

2.3.2. Etude différentielle des protéomes hépatiques chez le rat dénutri

2.3.2.1. Identification des protéines

Pour l'étude différentielle des protéomes hépatiques chez le rat dénutri, nous avons eu recours à l'utilisation combinée de 2 moteurs de recherche : Mascot et OMSSA. L'utilisation combinée de plusieurs moteurs de recherche pour bénéficier de leurs résultats complémentaires est une approche en plein développement. Plusieurs études ont rapporté que l'utilisation de plusieurs moteurs de recherches conduisait à l'identification d'un plus grand nombre de peptides et par conséquent à des identifications de protéines plus fiables et à l'identification d'un plus grand nombre de protéines [Resing et al., 2004; Elias et al., 2005; Kapp et al., 2005; Price et al., 2007]. Même si pour pouvoir pleinement bénéficier des avantages de cette approche, il est nécessaire de pouvoir combiner les scores des identifications par les différents moteurs pour obtenir un score probabiliste final, ici, nous avons procédé différemment. Ainsi, dans un premier temps, nous avons établi les scores seuils des identifications peptidiques pour chacun des moteurs indépendamment pour obtenir un FDR inférieur à 1 % sur l'identification des protéines. Dans un deuxième temps nous avons filtré l'identification des protéines par l'utilisation combinée des scores seuils fixés pour les 2 moteurs.

Scores seuils Mascot sur les identifications peptidiques :

- Au moins 2 peptides participent à l'identification des protéines : score identité-11.
- 1 seul peptide participe à l'identification des protéines : score identité+4.

Scores seuils OMSSA sur les identifications peptidiques :

- Au moins 2 peptides participent à l'identification des protéines : $-\log(\text{E-value})=0.05$.
- 1 seul peptide participe à l'identification des protéines : $-\log(\text{E-value})=2.7$.

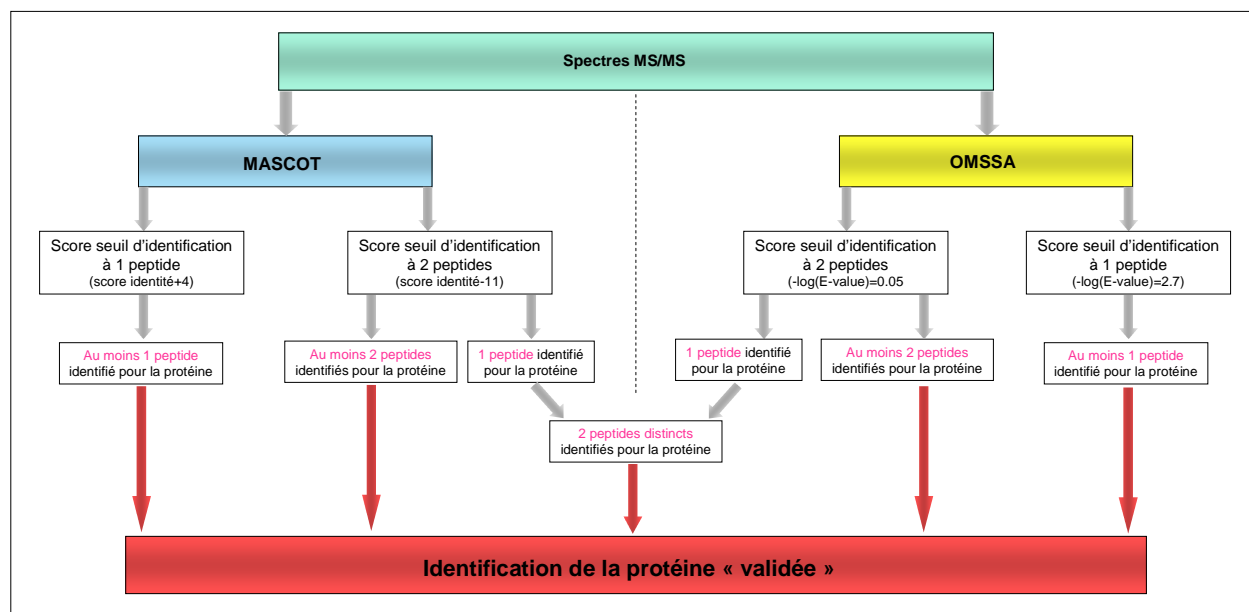
Combinaison des 2 moteurs de recherche :

- Quand au moins 2 peptides participent à l'identification de la protéine, leur score d'identification doit être supérieur au score seuil Mascot et/ou au score seuil OMSSA d'identification à 2 peptides. C'est-à-dire que l'identification de la protéine est « validée » si au moins 2 peptides ont un score d'identification passant le seuil de score Mascot ou passant le seuil de score OMSSA ou passant les 2 seuils. L'identification de la protéine est également validée si au moins un peptide a un score d'identification passant le seuil de score Mascot et un autre peptide a un score d'identification passant le seuil de score OMSSA.

- Quand 1 peptide participe à l'identification de la protéine, son score d'identification doit être supérieur au score seuil Mascot et/ou au score seuil OMSSA d'identification à 1 peptide pour que l'identification de la protéine soit « validée ».

Le mode opératoire pour la « validation » de l'identification d'une protéine par l'utilisation combinée des 2 moteurs de recherche est présenté dans la Figure 8.

Figure 8 : Mode opératoire pour la validation de l'identification d'une protéine par l'utilisation combinée des 2 moteurs de recherche.



Nous avons réalisé en Figure 9 la comparaison des résultats obtenus avec l'un ou l'autre des moteurs de recherche ou une combinaison des 2 moteurs, en termes de :

- ✓ Nombre de spectres moyen par piste de gel 1D passant les seuils.
- ✓ Nombre de peptides moyen par piste de gel 1D passant les seuils.
- ✓ Nombre de protéines identifiées sur toute l'étude en appliquant les différents seuils sur les identifications peptidiques.

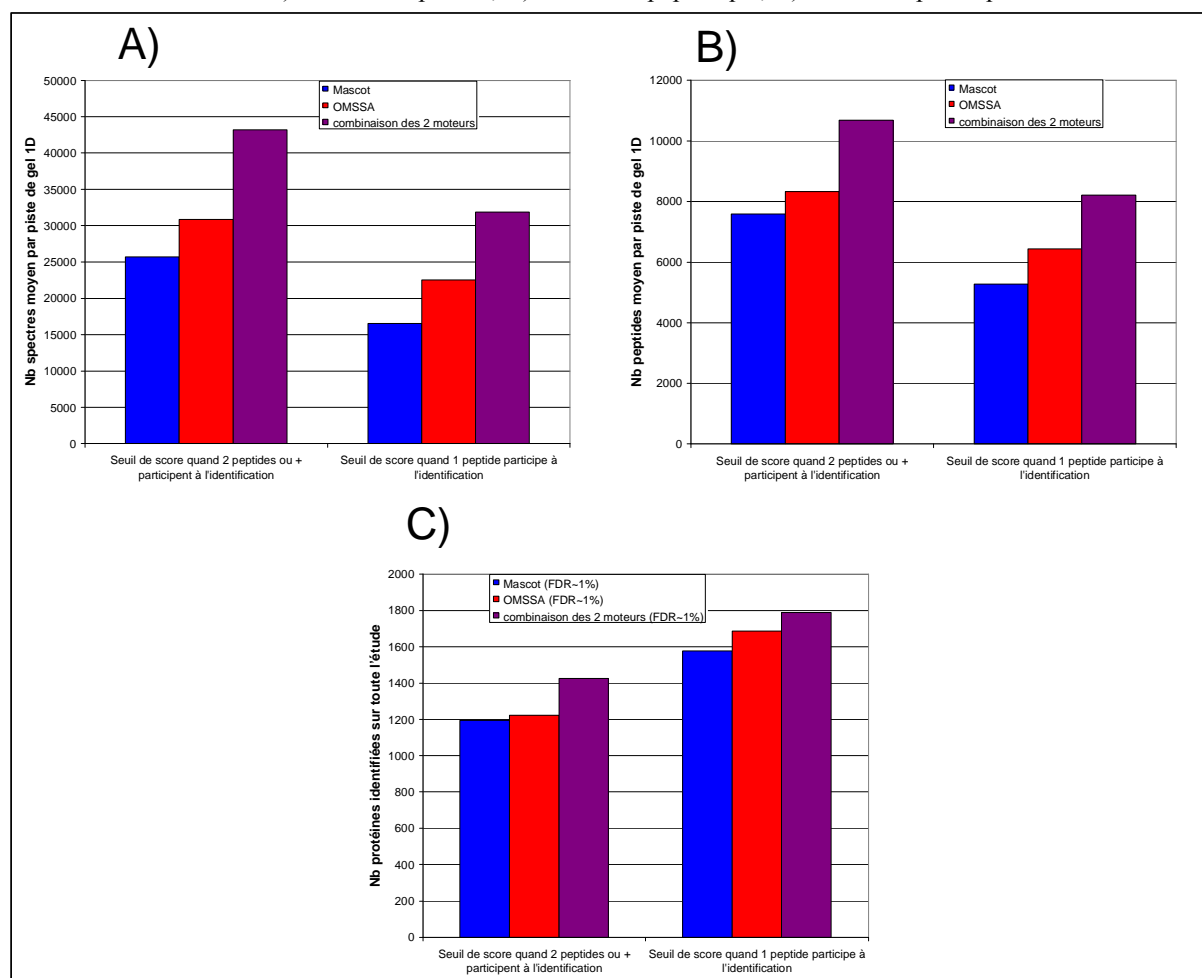
Nous avons pu observer dans cette comparaison qu'OMSSA permettait d'identifier un nombre très légèrement supérieur de protéines par rapport à Mascot (1222 contre 1194 protéines avec le seuil à 2 peptides, 1686 contre 1576 protéines avec le seuil à 1 peptide). L'utilisation combinée des 2 moteurs permet d'augmenter encore légèrement le nombre de protéines identifiées (1425 protéines avec le seuil à 2 peptides et 1789 protéines avec le seuil à 1 peptide).

Les différences entre les 2 moteurs sont plus larges au niveau de l'identification peptidique (~10-20 % de peptides identifiés en plus avec OMSSA) et surtout au niveau de l'identification spectrale (~20-30 % de spectres identifiés en plus avec OMSSA). De même, l'utilisation combinée des 2 moteurs de recherche améliore encore très significativement les performances d'identification puisque cette stratégie permet d'identifier près de 50 % de peptides en plus et presque le double de spectres par rapport à Mascot. L'utilisation combinée des 2 moteurs de recherche permet donc non seulement

d'augmenter le nombre de protéines identifiées mais aussi d'améliorer la couverture des séquences protéiques identifiées grâce à davantage de peptides et enfin de fournir un nombre bien plus important de spectres qui permettent d'améliorer la fiabilité des quantifications par spectral counting.

Finalement, en combinant les résultats sur les identifications des protéines à 1 peptide et 2 peptides ou plus obtenus avec l'utilisation conjointe des 2 moteurs de recherche, 1806 protéines uniques ont pu être identifiées (FDR : 1.1 %).

Figure 9 : Comparaison des résultats d'identification obtenus avec Mascot, OMSSA ou la combinaison des 2 moteurs de recherche A) au niveau spectral, B) au niveau peptidique, C) au niveau protéique.



2.3.2.2. Quantification relative des protéines

La quantification relative des protéines par spectral counting a été limitée aux protéines identifiées par au moins deux peptides dans l'étude (1425 protéines). Nous avons réalisé cette sélection afin de nous assurer qu'un nombre de spectres suffisamment élevé est considéré pour le traitement quantitatif de chaque protéine car la fiabilité de la quantification dépend de ce nombre de spectres.

Tous les spectres attribués à cet ensemble de protéines et passant le seuil de score d'identification à 2 peptides ont été considérés pour la quantification par spectral counting (total d'environ 260 000 spectres considérés dans toute l'étude). Le nombre de spectres attribués à chaque protéine dans une expérience donnée a été normalisé en le divisant par le nombre total de spectres attribués à toutes les protéines dans la même expérience. Contrairement à l'étude précédente (Partie III des résultats, Chapitre 2), comme les analyses ont été réalisées sur des triplicats biologiques, nous avons pu utiliser comme test statistique pour l'estimation de la significativité des différences d'expression protéique, un test d'analyse de la variance (« ANOVA », « ANalysis Of Variance) qui a l'avantage de prendre en compte les variabilités intra-groupes. Le traitement ANOVA de l'ensemble des données normalisées a permis la mise en évidence de 328 protéines différemment exprimées entre les groupes au risque $p < 0.05$. Des expériences complémentaires sur l'analyse du niveau d'expression des gènes par puces à ADN (puces Affymetrix) ont également été réalisées sur les mêmes prélèvements et leur mise en relation avec les données de protéomique quantitative sont en cours. De plus, des analyses western blot sur des protéines d'intérêt détectées comme différemment exprimées sont également en cours et permettront, en fonction de leurs cohérences avec les résultats de protéomique quantitative, d'établir des faisceaux d'évidence qui pourront augmenter la confiance dans les interprétations biologiques des résultats.

2.4. Conclusion et perspectives

Les résultats préliminaires obtenus sur l'étude différentielle des protéomes hépatiques chez le rat dénutri ont permis d'identifier plus de 1800 protéines uniques. L'optimisation des méthodes nanoLC-MS/MS et la combinaison de plusieurs moteurs de recherche se sont donc avérées très efficaces pour améliorer non seulement l'identification des protéines mais aussi le traitement quantitatif par l'approche spectral counting puisque qu'un total d'environ 260000 spectres a pu être considéré (soit environ le double du nombre de spectres qui auraient été considérés en utilisant le moteur de recherche Mascot seulement).

Cette première étude différentielle par l'approche qN-TOP a montré que cette approche présentait un potentiel intéressant pour l'identification des sites protéolytiques d'un protéome complexe. En effet, 187 séquences peptidiques uniques marquées en N-terminal au TMPP (lourd et/ou léger) et correspondant au peptide N-terminale d'une protéine processée ont pu être identifiées et cela sans biais apparent sur les identifications en fonction du type de marquage. Cette première application de l'approche qN-TOP témoigne donc de la faisabilité d'un projet d'étude des processus protéolytiques générés *in vivo* en réponse à une condition biologique donnée avec cette approche. Toutefois, l'aspect quantitatif de notre étude n'a pour l'instant pu être exploité à cause des limitations instrumentales auxquelles nous avons été confrontés. Ces limitations seront prises en compte dans la suite du projet pour nous permettre de tirer pleinement parti du potentiel de notre nouvelle stratégie qN-TOP.

Conclusion

Cette troisième partie des résultats a été consacrée au développement et à l'application de méthodologies de protéomique quantitative.

Nous avons mis en place au laboratoire une stratégie de quantification de type spectral counting qui a été appliquée à deux études protéomiques différentielles. Cette stratégie a nécessité un paramétrage affiné des instruments de chromatographie et de spectrométrie de masse pour obtenir une quantification fiable. Nous avons également pu constater que l'utilisation combinée de plusieurs moteurs de recherche pour l'attribution des spectres constituait un avantage certain pour ce type d'approche. Dans nos travaux, nous nous sommes également efforcés de compléter nos résultats d'analyses protéomiques quantitatives avec d'autres résultats obtenus :

- Au niveau transcriptionnel de manière large (puce à ADN pour l'étude différentielle des protéomes hépatiques chez le rat dénutri) ou plus ciblée (analyses PCR quantitatives de l'expression de quelques gènes d'intérêt pour l'étude appliquée à *G. metallireducens*).
- Au niveau protéique par d'autres techniques : Western blot, mesures d'activités.

L'utilisation de méthodes d'analyses différentes permet de confirmer mutuellement les résultats pour une confiance accrue dans l'interprétation biologique de ceux-ci. Ici, les techniques complémentaires utilisées font appel à d'autres techniques que la spectrométrie de masse. Toutefois, une technique biochimique comme le western blot pourrait sans doute être substituée avantageusement par la stratégie de quantification ciblée par approche MRM qui est en cours de mise en place au laboratoire.

Bien que la protéomique quantitative par les approches sans marquage semble devenir la technique de quantification protéomique privilégiée dans les laboratoires, les approches avec marquage peuvent être très efficaces, notamment pour la quantification ciblée d'une sous-population de peptides. La stratégie qN-TOP que nous avons développée permet par exemple la quantification relative ciblée de peptides N-terminaux à grande échelle qui n'aurait pas été possible avec les méthodes sans marquage.

Bibliographie

- Addona, T. A., S. E. Abbatiello, B. Schilling, S. J. Skates, D. R. Mani, D. M. Bunk, C. H. Spiegelman, L. J. Zimmerman, A. J. Ham, H. Keshishian, S. C. Hall, S. Allen, R. K. Blackman, C. H. Borchers, C. Buck, H. L. Cardasis, M. P. Cusack, N. G. Dodder, B. W. Gibson, J. M. Held, T. Hiltke, A. Jackson, E. B. Johansen, C. R. Kinsinger, J. Li, M. Mesri, T. A. Neubert, R. K. Niles, T. C. Pulsipher, D. Ransohoff, H. Rodriguez, P. A. Rudnick, D. Smith, D. L. Tabb, T. J. Tegeler, A. M. Variyath, L. J. Vega-Montoto, A. Wahlander, S. Waldemarson, M. Wang, J. R. Whiteaker, L. Zhao, N. L. Anderson, S. J. Fisher, D. C. Liebler, A. G. Paulovich, F. E. Regnier, P. Tempst and S. A. Carr**
"Multi-site assessment of the precision and reproducibility of multiple reaction monitoring-based measurements of proteins in plasma." *Nat Biotechnol*, **2009**, 27 (7), 633-41.
- Ahmed, N., G. Barker, K. T. Oliva, P. Hoffmann, C. Riley, S. Reeve, A. I. Smith, B. E. Kemp, M. A. Quinn and G. E. Rice**
"Proteomic-based identification of haptoglobin-1 precursor as a novel circulating biomarker of ovarian cancer." *Br J Cancer*, **2004**, 91 (1), 129-40.
- Arnesen, T., P. Van Damme, B. Polevoda, K. Helsens, R. Evjenth, N. Colaert, J. E. Varhaug, J. Vandekerckhove, J. R. Lillehaug, F. Sherman and K. Gevaert**
"Proteomics analyses reveal the evolutionary conservation and divergence of N-terminal acetyltransferases from yeast and humans." *Proc Natl Acad Sci U S A*, **2009**, 106 (20), 8157-62.
- Bantscheff, M., M. Schirle, G. Sweetman, J. Rick and B. Kuster**
"Quantitative mass spectrometry in proteomics: a critical review." *Anal Bioanal Chem*, **2007**, 389 (4), 1017-31.
- Barrett, A. J.**
"Bioinformatics of proteases in the MEROPS database." *Curr Opin Drug Discov Devel*, **2004**, 7 (3), 334-41.
- Beardsley, R. L. and J. P. Reilly**
"Optimization of guanidination procedures for MALDI mass mapping." *Anal Chem*, **2002**, 74 (8), 1884-90.
- Bergeron, J. J. and M. Hallett**
"Peptides you can count on." *Nat Biotechnol*, **2007**, 25 (1), 61-2.
- Bernay, B., M. C. Gaillard, V. Guryca, A. Emadali, L. Kuhn, A. Bertrand, I. Detraz, C. Carcenac, M. Savasta, E. Brouillet, J. Garin and J. M. Elalouf**
"Discovering new bioactive neuropeptides in the striatum secretome using in vivo microdialysis and versatile proteomics." *Mol Cell Proteomics*, **2009**, 8 (5), 946-58.
- Blondeau, F., B. Ritter, P. D. Allaire, S. Wasiak, M. Girard, N. K. Hussain, A. Angers, V. Legendre-Guillemain, L. Roy, D. Boismenu, R. E. Kearney, A. W. Bell, J. J. Bergeron and P. S. McPherson**
"Tandem MS analysis of brain clathrin-coated vesicles reveals their critical involvement in synaptic vesicle recycling." *Proc Natl Acad Sci U S A*, **2004**, 101 (11), 3833-8.
- Boll, M. and G. Fuchs**
"Benzoyl-coenzyme A reductase (dearomatizing), a key enzyme of anaerobic aromatic metabolism. ATP dependence of the reaction, purification and some properties of the enzyme from *Thaueria aromatica* strain K172." *Eur J Biochem*, **1995**, 234 (3), 921-33.
- Boll, M., G. Fuchs and J. Heider**
"Anaerobic oxidation of aromatic compounds and hydrocarbons." *Curr Opin Chem Biol*, **2002**, 6 (5), 604-11.
- Bondarenko, P. V., D. Chelius and T. A. Shaler**
"Identification and relative quantitation of protein mixtures by enzymatic digestion followed by capillary reversed-phase liquid chromatography-tandem mass spectrometry." *Anal Chem*, **2002**, 74 (18), 4741-9.
- Braun, R. J., N. Kinkl, M. Beer and M. Ueffing**
"Two-dimensional electrophoresis of membrane proteins." *Anal Bioanal Chem*, **2007**, 389 (4), 1033-45.
- Brun, V., A. Dupuis, A. Adrait, M. Marcellin, D. Thomas, M. Court, F. Vandenesch and J. Garin**
"Isotope-labeled protein standards: toward absolute quantitative proteomics." *Mol Cell Proteomics*, **2007**, 6 (12), 2139-49.
- Brun, V., C. Masselon, J. Garin and A. Dupuis**
"Isotope dilution strategies for absolute quantitative proteomics." *J Proteomics*, **2009**, 72 (5), 740-9.
- Chelius, D. and P. V. Bondarenko**
"Quantitative profiling of proteins in complex mixtures using liquid chromatography and mass spectrometry." *J Proteome Res*, **2002**, 1 (4), 317-23.
- Chen, Y., S. W. Kwon, S. C. Kim and Y. Zhao**
"Integrated approach for manual evaluation of peptides identified by searching protein sequence databases with tandem mass spectra." *J Proteome Res*, **2005**, 4 (3), 998-1005.
- Cheng, F. Y., K. Blackburn, Y. M. Lin, M. B. Goshe and J. D. Williamson**

"Absolute protein quantification by LC/MS(E) for global analysis of salicylic acid-induced plant protein secretion responses." *J Proteome Res*, **2009**, 8 (1), 82-93.

Cherel, Y. and Y. Le Maho

"Refeeding after the late increase in nitrogen excretion during prolonged fasting in the rat." *Physiol Behav*, **1991**, 50 (2), 345-9.

Choe, L., M. D'Ascenzo, N. R. Relkin, D. Pappin, P. Ross, B. Williamson, S. Guertin, P. Pribil and K. H. Lee

"8-plex quantitation of changes in cerebrospinal fluid protein expression in subjects undergoing intravenous immunoglobulin treatment for Alzheimer's disease." *Proteomics*, **2007**, 7 (20), 3651-60.

Colinge, J., D. Chiappe, S. Lagache, M. Moniatte and L. Bougueleret

"Differential proteomics via probabilistic peptide identification scores." *Anal Chem*, **2005**, 77 (2), 596-606.

Conrads, T. P., K. Alving, T. D. Veenstra, M. E. Belov, G. A. Anderson, D. J. Anderson, M. S. Lipton, L. Pasa-Tolic, H. R. Udseth, W. B. Chrisler, B. D. Thrall and R. D. Smith

"Quantitative analysis of bacterial and mammalian proteomes using a combination of cysteine affinity tags and ¹⁵N-metabolic labeling." *Anal Chem*, **2001**, 73 (9), 2132-9.

Desiere, F., E. W. Deutsch, N. L. King, A. I. Nesvizhskii, P. Mallick, J. Eng, S. Chen, J. Eddes, S. N. Loevenich and R. Aebersold

"The PeptideAtlas project." *Nucleic Acids Res*, **2006**, 34 (Database issue), D655-8.

Diaz, E.

"Bacterial degradation of aromatic pollutants: a paradigm of metabolic versatility." *Int Microbiol*, **2004**, 7 (3), 173-80.

Doucet, A., G. S. Butler, D. Rodriguez, A. Prudova and C. M. Overall

"Metadegradomics: toward in vivo quantitative degradomics of proteolytic post-translational modifications of the cancer proteome." *Mol Cell Proteomics*, **2008**, 7 (10), 1925-51.

Elias, J. E., W. Haas, B. K. Faherty and S. P. Gygi

"Comparative evaluation of mass spectrometry platforms used in large-scale proteomics investigations." *Nat Methods*, **2005**, 2 (9), 667-75.

Fang, R., D. A. Elias, M. E. Monroe, Y. Shen, M. McIntosh, P. Wang, C. D. Goddard, S. J. Callister, R. J. Moore, Y. A. Gorby, J. N. Adkins, J. K. Fredrickson, M. S. Lipton and R. D. Smith

"Differential label-free quantitative proteomic analysis of *Shewanella oneidensis* cultured under aerobic and suboxic conditions by accurate mass and time tag approach." *Mol Cell Proteomics*, **2006**, 5 (4), 714-25.

Foster, L. J., C. L. De Hoog and M. Mann

"Unbiased quantitative proteomics of lipid rafts reveals high specificity for signaling factors." *Proc Natl Acad Sci U S A*, **2003**, 100 (10), 5813-8.

Fuchs, G.

"Anaerobic metabolism of aromatic compounds." *Ann N Y Acad Sci*, **2008**, 1125 82-99.

Fusaro, V. A., D. R. Mani, J. P. Mesirov and S. A. Carr

"Prediction of high-responding peptides for targeted protein assays by mass spectrometry." *Nat Biotechnol*, **2009**, 27 (2), 190-8.

Gao, J., G. J. Opiteck, M. S. Friedrichs, A. R. Dongre and S. A. Hefta

"Changes in the protein expression of yeast as a function of carbon source." *J Proteome Res*, **2003**, 2 (6), 643-9.

Gartner, C. A., J. E. Elias, C. E. Bakalarski and S. P. Gygi

"Catch-and-release reagents for broadscale quantitative proteomics analyses." *J Proteome Res*, **2007**, 6 (4), 1482-91.

Gerber, S. A., J. Rush, O. Stemman, M. W. Kirschner and S. P. Gygi

"Absolute quantification of proteins and phosphoproteins from cell lysates by tandem MS." *Proc Natl Acad Sci U S A*, **2003**, 100 (12), 6940-5.

Gevaert, K., F. Impens, B. Ghesquiere, P. Van Damme, A. Lambrechts and J. Vandekerckhove

"Stable isotopic labeling in proteomics." *Proteomics*, **2008**, 8 (23-24), 4873-85.

Gibson, J. and S. H. C

"Metabolic diversity in aromatic compound utilization by anaerobic microbes." *Annu Rev Microbiol*, **2002**, 56 345-69.

Gilchrist, A., C. E. Au, J. Hiding, A. W. Bell, J. Fernandez-Rodriguez, S. Lesimple, H. Nagaya, L. Roy, S. J. Gosline, M. Hallett, J. Paiement, R. E. Kearney, T. Nilsson and J. J. Bergeron

"Quantitative proteomics analysis of the secretory pathway." *Cell*, **2006**, 127 (6), 1265-81.

Gygi, S. P., B. Rist, S. A. Gerber, F. Turecek, M. H. Gelb and R. Aebersold

"Quantitative analysis of complex protein mixtures using isotope-coded affinity tags." *Nat Biotechnol*, **1999**, 17 (10), 994-9.

Hansen, K. C., G. Schmitt-Ulms, R. J. Chalkley, J. Hirsch, M. A. Baldwin and A. L. Burlingame

- "Mass spectrometric analysis of protein mixtures at low levels using cleavable ¹³C-isotope-coded affinity tag and multidimensional chromatography." *Mol Cell Proteomics*, **2003**, 2 (5), 299-314.
- Heudi, O., S. Barteau, D. Zimmer, J. Schmidt, K. Bill, N. Lehmann, C. Bauer and O. Kretz**
 "Towards absolute quantification of therapeutic monoclonal antibody in serum by LC-MS/MS using isotope-labeled antibody standard and protein cleavage isotope dilution mass spectrometry." *Anal Chem*, **2008**, 80 (11), 4200-7.
- Hsich, G., K. Kenney, C. J. Gibbs, K. H. Lee and M. G. Harrington**
 "The 14-3-3 brain protein in cerebrospinal fluid as a marker for transmissible spongiform encephalopathies." *N Engl J Med*, **1996**, 335 (13), 924-30.
- Huang, Z. H., T. Shen, J. Wu, D. A. Gage and J. T. Watson**
 "Protein sequencing by matrix-assisted laser desorption ionization-postsource decay-mass spectrometry analysis of the N-Tris(2,4,6-trimethoxyphenyl)phosphine-acetylated tryptic digests." *Anal Biochem*, **1999**, 268 (2), 305-17.
- Ippel, J. H., L. Pouvreau, T. Kroef, H. Gruppen, G. Versteeg, P. van den Putten, P. C. Struik and C. P. van Mierlo**
 "In vivo uniform (¹⁵N)-isotope labelling of plants: using the greenhouse for structural proteomics." *Proteomics*, **2004**, 4 (1), 226-34.
- Ishihama, Y., Y. Oda, T. Tabata, T. Sato, T. Nagasu, J. Rappsilber and M. Mann**
 "Exponentially modified protein abundance index (emPAI) for estimation of absolute protein amount in proteomics by the number of sequenced peptides per protein." *Mol Cell Proteomics*, **2005**, 4 (9), 1265-72.
- Kapp, E. A., F. Schutz, L. M. Connolly, J. A. Chakel, J. E. Meza, C. A. Miller, D. Fenyo, J. K. Eng, J. N. Adkins, G. S. Omenn and R. J. Simpson**
 "An evaluation, comparison, and accurate benchmarking of several publicly available MS/MS search algorithms: sensitivity and specificity analysis." *Proteomics*, **2005**, 5 (13), 3475-90.
- Keough, T., M. P. Lacey and R. S. Youngquist**
 "Derivatization procedures to facilitate de novo sequencing of lysine-terminated tryptic peptides using postsource decay matrix-assisted laser desorption/ionization mass spectrometry." *Rapid Commun Mass Spectrom*, **2000**, 14 (24), 2348-56.
- Koubi, H. E., J. P. Robin, G. Dewasmes, Y. Le Maho, J. Frutoso and Y. Minaire**
 "Fasting-induced rise in locomotor activity in rats coincides with increased protein utilization." *Physiol Behav*, **1991**, 50 (2), 337-43.
- Krijgsveld, J., R. F. Ketting, T. Mahmoudi, J. Johansen, M. Artal-Sanz, C. P. Verrijzer, R. H. Plasterk and A. J. Heck**
 "Metabolic labeling of *C. elegans* and *D. melanogaster* for quantitative proteomics." *Nat Biotechnol*, **2003**, 21 (8), 927-31.
- Kuster, B., M. Schirle, P. Mallick and R. Aebersold**
 "Scoring proteomes with proteotypic peptide probes." *Nat Rev Mol Cell Biol*, **2005**, 6 (7), 577-83.
- Lamkanfi, M., T. D. Kanneganti, P. Van Damme, T. Vanden Berghe, I. Vanoverberghe, J. Vandekerckhove, P. Vandenabeele, K. Gevaert and G. Nunez**
 "Targeted peptidocentric proteomics reveals caspase-7 as a substrate of the caspase-1 inflammasomes." *Mol Cell Proteomics*, **2008**, 7 (12), 2350-63.
- Lange, V., P. Picotti, B. Domon and R. Aebersold**
 "Selected reaction monitoring for quantitative proteomics: a tutorial." *Mol Syst Biol*, **2008**, 4 222.
- Le, L., K. Chi, S. Tyldesley, S. Flibotte, D. L. Diamond, M. A. Kuzyk and M. D. Sadar**
 "Identification of serum amyloid A as a biomarker to distinguish prostate cancer patients with bone lesions." *Clin Chem*, **2005**, 51 (4), 695-707.
- Li, J., H. Steen and S. P. Gygi**
 "Protein profiling with cleavable isotope-coded affinity tag (cICAT) reagents: the yeast salinity stress response." *Mol Cell Proteomics*, **2003**, 2 (11), 1198-204.
- Liu, H., R. G. Sadygov and J. R. Yates, 3rd**
 "A model for random sampling and estimation of relative protein abundance in shotgun proteomics." *Anal Chem*, **2004**, 76 (14), 4193-201.
- Lovley, D. R.**
 "Cleaning up with genomics: applying molecular biology to bioremediation." *Nat Rev Microbiol*, **2003**, 1 (1), 35-44.
- Lovley, D. R., S. J. Giovannoni, D. C. White, J. E. Champine, E. J. Phillips, Y. A. Gorby and S. Goodwin**
 "Geobacter metallireducens gen. nov. sp. nov., a microorganism capable of coupling the complete oxidation of organic compounds to the reduction of iron and other metals." *Arch Microbiol*, **1993**, 159 (4), 336-44.
- Lovley, D. R., D. E. Holmes and K. P. Nevin**
 "Dissimilatory Fe(III) and Mn(IV) reduction." *Adv Microb Physiol*, **2004**, 49 219-86.

Lovley, D. R. and E. J. Phillips

"Novel Mode of Microbial Energy Metabolism: Organic Carbon Oxidation Coupled to Dissimilatory Reduction of Iron or Manganese." *Appl Environ Microbiol*, **1988**, 54 (6), 1472-1480.

Lu, P., C. Vogel, R. Wang, X. Yao and E. M. Marcotte

"Absolute protein expression profiling estimates the relative contributions of transcriptional and translational regulation." *Nat Biotechnol*, **2007**, 25 (1), 117-24.

Mallick, P., M. Schirle, S. S. Chen, M. R. Flory, H. Lee, D. Martin, J. Ranish, B. Raught, R. Schmitt, T. Werner, B. Kuster and R. Aebersold

"Computational prediction of proteotypic peptides for quantitative proteomics." *Nat Biotechnol*, **2007**, 25 (1), 125-31.

Malmstrom, J., M. Beck, A. Schmidt, V. Lange, E. W. Deutsch and R. Aebersold

"Proteome-wide cellular protein concentrations of the human pathogen *Leptospira interrogans*." *Nature*, **2009**, 460 (7256), 762-5.

Miguet, L., G. Bechade, L. Fornecker, E. Zink, C. Felden, C. Gervais, R. Herbrecht, A. van Dorsselaer, L. Mauvieux and S. Sanglier-Cianferani

"Proteomic analysis of malignant B-cell derived microparticles reveals CD148 as a potentially useful antigenic biomarker for mantle cell lymphoma diagnosis." *J Proteome Res*, **2009**, 8 (7), 3346-54.

Mirgorodskaya, O. A., Y. P. Kozmin, M. I. Titov, R. Korner, C. P. Sonksen and P. Roepstorff

"Quantitation of peptides and proteins by matrix-assisted laser desorption/ionization mass spectrometry using (18)O-labeled internal standards." *Rapid Commun Mass Spectrom*, **2000**, 14 (14), 1226-32.

Mueller, L. N., M. Y. Brusniak, D. R. Mani and R. Aebersold

"An assessment of software solutions for the analysis of mass spectrometry based quantitative proteomics data." *J Proteome Res*, **2008**, 7 (1), 51-61.

Oda, Y., K. Huang, F. R. Cross, D. Cowburn and B. T. Chait

"Accurate quantitation of protein expression and site-specific phosphorylation." *Proc Natl Acad Sci U S A*, **1999**, 96 (12), 6591-6.

Oda, Y., T. Owa, T. Sato, B. Boucher, S. Daniels, H. Yamanaka, Y. Shinohara, A. Yokoi, J. Kuromitsu and T. Nagasu

"Quantitative chemical proteomics for identifying candidate drug targets." *Anal Chem*, **2003**, 75 (9), 2159-65.

Old, W. M., K. Meyer-Arendt, L. Aveline-Wolf, K. G. Pierce, A. Mendoza, J. R. Sevinisky, K. A. Resing and N. G. Ahn

"Comparison of label-free methods for quantifying human proteins by shotgun proteomics." *Mol Cell Proteomics*, **2005**, 4 (10), 1487-502.

Ong, S. E., B. Blagoev, I. Kratchmarova, D. B. Kristensen, H. Steen, A. Pandey and M. Mann

"Stable isotope labeling by amino acids in cell culture, SILAC, as a simple and accurate approach to expression proteomics." *Mol Cell Proteomics*, **2002**, 1 (5), 376-86.

Pan, S., R. Aebersold, R. Chen, J. Rush, D. R. Goodlett, M. W. McIntosh, J. Zhang and T. A. Brentnall

"Mass spectrometry based targeted protein quantification: methods and applications." *J Proteome Res*, **2009**, 8 (2), 787-97.

Pan, S., H. Zhang, J. Rush, J. Eng, N. Zhang, D. Patterson, M. J. Comb and R. Aebersold

"High throughput proteome screening for biomarker detection." *Mol Cell Proteomics*, **2005**, 4 (2), 182-90.

Pasa-Tolic, L., P. K. Jensen, G. A. Anderson, M. Lipton, K. K. Peden, S. Martinovic, N. Tolic, J. E. Bruce and R. D. Smith

"High throughput proteome-wide precision measurements of protein expression using mass spectrometry." *Journal of the American Chemical Society* **1999**, 121 (34), 7949-7950.

Pasa-Tolic, L., C. Masselon, R. C. Barry, Y. Shen and R. D. Smith

"Proteomic analyses using an accurate mass and time tag strategy." *Biotechniques*, **2004**, 37 (4), 621-4, 626-33, 636 passim.

Picotti, P., B. Bodenmiller, L. N. Mueller, B. Domon and R. Aebersold

"Full dynamic range proteome analysis of *S. cerevisiae* by targeted proteomics." *Cell*, **2009**, 138 (4), 795-806.

Pratt, J. M., D. M. Simpson, M. K. Doherty, J. Rivers, S. J. Gaskell and R. J. Beynon

"Multiplexed absolute quantification for proteomics using concatenated signature peptides encoded by QconCAT genes." *Nat Protoc*, **2006**, 1 (2), 1029-43.

Price, T. S., M. B. Lucitt, W. Wu, D. J. Austin, A. Pizarro, A. K. Yocum, I. A. Blair, G. A. FitzGerald and T. Grosser

"EBP, a program for protein identification using multiple tandem mass spectrometry datasets." *Mol Cell Proteomics*, **2007**, 6 (3), 527-36.

Rao, K. C., V. Palamalai, J. R. Dunlevy and M. Miyagi

"Peptidyl-Lys metalloendopeptidase-catalyzed 18O labeling for comparative proteomics: application to cytokine/lipopolysaccharide-treated human retinal pigment epithelium cell line." *Mol Cell Proteomics*, **2005**, 4 (10), 1550-7.

Resing, K. A., K. Meyer-Arendt, A. M. Mendoza, L. D. Aveline-Wolf, K. R. Jonscher, K. G. Pierce, W. M. Old, H. T. Cheung, S. Russell, J. L. Wattawa, G. R. Goehle, R. D. Knight and N. G. Ahn
 "Improving reproducibility and sensitivity in identifying human proteins by shotgun proteomics." *Anal Chem*, **2004**, 76 (13), 3556-68.

Rivers, J., D. M. Simpson, D. H. Robertson, S. J. Gaskell and R. J. Beynon
 "Absolute multiplexed quantitative analysis of protein expression during muscle development using QconCAT." *Mol Cell Proteomics*, **2007**, 6 (8), 1416-27.

Ross, P. L., Y. N. Huang, J. N. Marchese, B. Williamson, K. Parker, S. Hattan, N. Khainovski, S. Pillai, S. Dey, S. Daniels, S. Purkayastha, P. Juhasz, S. Martin, M. Bartlet-Jones, F. He, A. Jacobson and D. J. Pappin
 "Multiplexed protein quantitation in *Saccharomyces cerevisiae* using amine-reactive isobaric tagging reagents." *Mol Cell Proteomics*, **2004**, 3 (12), 1154-69.

Santoni, V., M. Molloy and T. Rabilloud
 "Membrane proteins and proteomics: un amour impossible?" *Electrophoresis*, **2000**, 21 (6), 1054-70.

Schmidt, A., J. Kellermann and F. Lottspeich
 "A novel strategy for quantitative proteomics using isotope-coded protein labels." *Proteomics*, **2005**, 5 (1), 4-15.

Silva, J. C., M. V. Gorenstein, G. Z. Li, J. P. Vissers and S. J. Geromanos
 "Absolute quantification of proteins by LCMSE: a virtue of parallel MS acquisition." *Mol Cell Proteomics*, **2006**, 5 (1), 144-56.

Smith, R. D., G. A. Anderson, M. S. Lipton, C. Masselon, L. Pasa-Tolic, H. Udseth, M. Belov, Y. Shen and T. D. Veenstra
 "High-performance separations and mass spectrometric methods for high-throughput proteomics using accurate mass tags." *Adv Protein Chem*, **2003**, 65 85-131.

Staes, A., H. Demol, J. Van Damme, L. Martens, J. Vandekerckhove and K. Gevaert
 "Global differential non-gel proteomics by quantitative and stable labeling of tryptic peptides with oxygen-18." *J Proteome Res*, **2004**, 3 (4), 786-91.

Thompson, A., J. Schafer, K. Kuhn, S. Kienle, J. Schwarz, G. Schmidt, T. Neumann, R. Johnstone, A. K. Mohammed and C. Hamon
 "Tandem mass tags: a novel quantification strategy for comparative analysis of complex protein mixtures by MS/MS." *Anal Chem*, **2003**, 75 (8), 1895-904.

Unlu, M., M. E. Morgan and J. S. Minden
 "Difference gel electrophoresis: a single gel method for detecting changes in protein extracts." *Electrophoresis*, **1997**, 18 (11), 2071-7.

Van Damme, P., L. Martens, J. Van Damme, K. Hugelier, A. Staes, J. Vandekerckhove and K. Gevaert
 "Caspase-specific and nonspecific in vivo protein processing during Fas-induced apoptosis." *Nat Methods*, **2005**, 2 (10), 771-7.

Van Damme, P., S. Maurer-Stroh, K. Plasman, J. Van Durme, N. Colaert, E. Timmerman, P. J. De Bock, M. Goethals, F. Rousseau, J. Schymkowitz, J. Vandekerckhove and K. Gevaert
 "Analysis of protein processing by N-terminal proteomics reveals novel species-specific substrate determinants of granzyme B orthologs." *Mol Cell Proteomics*, **2009**, 8 (2), 258-72.

Van Hoof, D., M. W. Pinkse, D. W. Oostwaard, C. L. Mummery, A. J. Heck and J. Krijgsveld
 "An experimental correction for arginine-to-proline conversion artifacts in SILAC-based quantitative proteomics." *Nat Methods*, **2007**, 4 (9), 677-8.

Vande Walle, L., P. Van Damme, M. Lamkanfi, X. Saelens, J. Vandekerckhove, K. Gevaert and P. Vandenabeele
 "Proteome-wide Identification of HtrA2/Omi Substrates." *J Proteome Res*, **2007**, 6 (3), 1006-15.

Viswanathan, C. T., S. Bansal, B. Booth, A. J. DeStefano, M. J. Rose, J. Sailstad, V. P. Shah, J. P. Skelly, P. G. Swann and R. Weiner
 "Quantitative bioanalytical methods validation and implementation: best practices for chromatographic and ligand binding assays." *Pharm Res*, **2007**, 24 (10), 1962-73.

Wang, G., W. W. Wu, W. Zeng, C. L. Chou and R. F. Shen
 "Label-free protein quantification using LC-coupled ion trap or FT mass spectrometry: Reproducibility, linearity, and application with complex proteomes." *J Proteome Res*, **2006**, 5 (5), 1214-23.

Wang, W., H. Zhou, H. Lin, S. Roy, T. A. Shaler, L. R. Hill, S. Norton, P. Kumar, M. Anderle and C. H. Becker
 "Quantification of proteins and metabolites by mass spectrometry without isotopic labeling or spiked standards." *Anal Chem*, **2003**, 75 (18), 4818-26.

Washburn, M. P., D. Wolters and J. R. Yates, 3rd

"Large-scale analysis of the yeast proteome by multidimensional protein identification technology." *Nat Biotechnol*, **2001**, 19 (3), 242-7.

Wiener, M. C., J. R. Sachs, E. G. Deyanova and N. A. Yates

"Differential mass spectrometry: a label-free LC-MS method for finding significant differences in complex peptide and protein mixtures." *Anal Chem*, **2004**, 76 (20), 6085-96.

Wischgoll, S., D. Heintz, F. Peters, A. Erxleben, E. Sarnighausen, R. Reski, A. Van Dorsselaer and M. Boll

"Gene clusters involved in anaerobic benzoate degradation of *Geobacter metallireducens*." *Mol Microbiol*, **2005**, 58 (5), 1238-52.

Wolf-Yadlin, A., S. Hautaniemi, D. A. Lauffenburger and F. M. White

"Multiple reaction monitoring for robust quantitative proteomic analysis of cellular signaling networks." *Proc Natl Acad Sci U S A*, **2007**, 104 (14), 5860-5.

Wu, C. C., M. J. MacCoss, K. E. Howell, D. E. Matthews and J. R. Yates, 3rd

"Metabolic labeling of mammalian organisms with stable isotopes for quantitative proteomic analysis." *Anal Chem*, **2004**, 76 (17), 4951-9.

Yamaguchi, M., D. Nakayama, K. Shima, H. Kuyama, E. Ando, T. A. Okamura, N. Ueyama, T. Nakazawa, S. Norioka, O. Nishimura and S. Tsunasawa

"Selective isolation of N-terminal peptides from proteins and their de novo sequencing by matrix-assisted laser desorption/ionization time-of-flight mass spectrometry without regard to unblocking or blocking of N-terminal amino acids." *Rapid Commun Mass Spectrom*, **2008**, 22 (20), 3313-9.

Yao, X., A. Freas, J. Ramirez, P. A. Demirev and C. Fenselau

"Proteolytic ¹⁸O labeling for comparative proteomics: model studies with two serotypes of adenovirus." *Anal Chem*, **2001**, 73 (13), 2836-42.

CONCLUSION GENERALE

CONCLUSION GENERALE

L'objectif principal de ce travail de thèse était de **développer de nouvelles approches d'aide à l'annotation génomique (protéogénomique)** basées sur les **méthodologies de la protéomique** pour **améliorer la qualité des banques de séquences protéiques** disponibles pour la communauté scientifique.

Les travaux réalisés pour atteindre cet objectif sont présentés dans la **partie I des résultats** dans laquelle nous avons notamment développé :

- *Une nouvelle stratégie pour améliorer la détermination des peptides N-terminaux des protéines (stratégie N-TOP) afin de faciliter l'identification des codons d'initiation des protéines qui représente sans doute un des plus grands défis de l'annotation génomique.*
- *Des stratégies optimisées de recherche de données MS/MS dans les banques de séquences (génomiques ou protéiques) pour faciliter la mise en évidence d'erreurs dans les banques protéiques générées par prédiction in silico.*

Ces stratégies ont pu être appliquées avec succès dans le cadre de différentes études en permettant d'**améliorer la qualité des banques protéiques**

- ✓ **Par la correction d'erreurs de séquençage des génomes.** Une stratégie optimisée de recherche des données MS/MS dans le génome complet de *M. smegmatis* combinée aux techniques de bioinformatique et de re-séquençage génomique ciblé a permis la détection et la correction des ICDSs dans le génome de cet organisme (**Chapitre 1**).
- ✓ **Par la correction d'un grand nombre d'erreurs issues de l'annotation génomique in silico des codons d'initiation des protéines et par l'identification de séquences protéiques non prédites.** La stratégie N-TOP combinée avec une stratégie optimisée de recherche des données MS/MS et avec les techniques de génomique comparative a permis de valider / corriger plus de 600 codons d'initiations de protéines annotés dans le génome de *M. smegmatis* et, par propagation, dans le génome des autres mycobactéries. Plusieurs séquences protéiques non prédites ont également pu être mises en évidence (**Chapitre 2**).
- ✓ **Par la détermination exacte de la séquence protéique** d'une enzyme issue d'un organisme dont le **génomme n'est pas séquencé**. La stratégie N-TOP combinée avec une stratégie de séquençage *de novo* des peptides analysés par nanoLC-MS/MS et d'analyse RT-PCR a rendu possible l'annotation exacte de la séquence de l'enzyme cholest-4-en-3-one- Δ^1 -dehydrogenase

impliquée dans une étape-clé du métabolisme anoxique du cholestérol chez *Sterolibacterium denitrificans* dont le génome n'est pas séquencé (**Chapitre 3**).

- ✓ **Par la détermination de séquences protéiques non prédites, codées** dans des **segments de génomes non assemblés** ou dans des **segments de génomes non séquencés** dans le contexte de l'étude de communautés bactériennes par **méta-protéo-génomique**. La stratégie de recherche séquentielle optimisée des données de nanoLC-MS/MS, dans une série de banques de séquences protéiques conçues en adéquation avec l'étude, a conduit à l'identification extensive de protéines exprimées dans une communauté bactérienne présente sur un site pollué à l'arsenic. La stratégie de recherche séquentielle a permis de limiter l'impact sur les identifications des différentes erreurs pouvant être générées au cours du processus de constitution des banques protéiques (**Chapitre 4**).

La stratégie N-TOP a également ouvert d'autres perspectives dans deux grands champs très importants de la protéomique : la **caractérisation des phosphorylations** des protéines et la **protéomique quantitative**.

La **partie II des résultats** est consacrée au **développement d'une stratégie d'analyse phosphoprotéomique bénéficiant des avantages du marquage au TMPP**. Etant donné l'importance des phosphorylations des protéines en biologie, l'analyse des phosphopeptides est primordiale mais reste difficile de part leurs caractéristiques intrinsèques (faible efficacité d'ionisation, caractère hydrophile, faible efficacité de fragmentation) et leur faible abondance (**Chapitre 1**). La stratégie originale développée, reposant sur l'analyse simultanée des phosphopeptides marqués au TMPP et des phosphopeptides natifs, s'est avérée très efficace pour révéler certaines **catégories de phosphopeptides ignorées par les techniques d'analyses habituelles** sans la nécessité de réaliser des analyses LC-MS/MS supplémentaires. Cette stratégie a pu être appliquée à l'analyse comparative des phosphoprotéomes d'*Arabidopsis thaliana* sauvage et mutant *hmgr1-1* (**Chapitre 2**).

La **partie III des résultats** est consacrée au développement et à l'application de méthodologies de **protéomique quantitative**. L'analyse protéomique par spectrométrie de masse a longtemps fourni des résultats uniquement qualitatifs. Néanmoins, nombreux sont les biologistes à s'intéresser au caractère quantitatif de ces résultats et au cours des dernières années, différentes stratégies utilisant le marquage aux isotopes stables ou non ont été développées pour répondre à ces besoins (**Chapitre 1**). L'**étude comparative du protéome de *Geobacter metallireducens*** cultivé en condition anaérobie avec ou sans source de carbone aromatique a nécessité la mise en place d'une **approche de quantification sans marquage, par comptage des spectres (« spectral counting »)**, à grande échelle (**Chapitre 2**). Cette approche s'est avérée efficace non seulement pour confirmer des hypothèses

biologiques mais aussi pour donner de **nouvelles pistes de réflexion**. Une approche de même type a été appliquée dans le cadre d'un deuxième projet portant sur **l'étude comparative de protéomes hépatiques** chez le rat dénutri. (**Chapitre 3**). Comme dans cette dernière étude nous nous sommes également intéressés à **l'analyse différentielle des processomes hépatiques** chez le rat dénutri, il a également fallu développer une **approche de quantification relative utilisant le marquage par isotopes stables**. Etant donné que la stratégie N-TOP est bien adaptée pour la détection de processus protéolytiques, nous lui avons conféré **une dimension quantitative** par l'élaboration d'une forme lourde du réactif TMPP. Cette nouvelle **stratégie qN-TOP** a fourni des résultats préliminaires relativement prometteurs.

Enfin, il faut souligner que lors de chaque étude protéomique, il est possible d'utiliser :

- ✓ Diverses techniques séparatives des protéines et des peptides.
- ✓ Différents type de spectromètres de masse.
- ✓ Différents paramétrage pour les séparations chromatographique et l'analyse par spectrométrie de masse.
- ✓ Différentes stratégies d'identification (voire de quantification).
- ✓ Différents types de banques de données.

Au cours de cette thèse, nous nous sommes efforcés de réaliser des choix stratégiques à tous ces niveaux et ceci le plus finement possible en fonction de la question biologique posée et des caractéristiques de l'échantillon de départ.

Il apparaît donc que l'analyse protéomique est loin d'être une méthodologie totalement aboutie et standardisée. Elle demande, de la part de ceux qui veulent l'utiliser à un haut niveau, une grande maîtrise des différents points mentionnés ci-dessus.

Finalement, notre travail montre qu'il existe encore de nombreuses possibilités pour augmenter le champ d'application de l'analyse protéomique. Il sera possible d'améliorer sa sensibilité et la qualité des identifications des protéines (degré de confiance, modifications post-traductionnelles, taux de recouvrement) à condition que les efforts pour les développements méthodologiques soient poursuivis dans les laboratoires spécialisés.

PARTIE EXPERIMENTALE

PARTIE EXPERIMENTALE

1. Partie I des résultats

1.1. Chapitre 1

Le détail des expériences et des protocoles analytiques est décrit dans la publication des résultats.

1.2. Chapitre 2

Le détail des expériences et des protocoles analytiques de l'application de la stratégie N-TOP sur les protéines modèles et de la première application de la stratégie N-TOP sur *M. smegmatis* est décrit dans la publication des résultats.

Le détail des expériences de nanoLC-MS/MS de la deuxième application de la stratégie N-TOP sur *M. smegmatis* est décrit ci-dessous :

Les extraits peptidiques ont été analysés par nanoLC-MS/MS sur un système chromatographique nanoACQUITY Ultra-Performance-LC (UPLC, Waters, Milford, MA) couplé à un spectromètre de masse « SYNAPT hybrid quadrupole orthogonal acceleration time-of-flight tandem mass spectrometer » (Waters, Milford, MA). Les mélanges peptidiques ont été chargés sur une précolonne de concentration Symmetry C18 (180 μm de diamètre interne \times 20 mm de longueur, particules de 5 μm ; Waters) en utilisant une phase mobile aqueuse avec 0.1 % d'acide formique à 5 $\mu\text{L min}^{-1}$. Après dessalage, les peptides ont été élués avec un gradient 1-70 % acétonitrile (+0.1 % d'acide formique) délivré sur 120 minutes à un débit de 400 nL min^{-1} au travers de la colonne analytique BEH130 C18 (75 μm de diamètre interne \times 200 mm de longueur, particules de 1.7 μm ; Waters). Les paramètres généraux du spectromètre de masse ont été les suivants : tension capillaire, 3500 V ; tension de cône, 35 V. Pour les expériences de spectrométrie de masse en tandem, le système a opéré avec un basculement automatique entre les modes MS et MS/MS. Les 3 ions les plus intenses, préférentiellement doublement et triplement chargés, ont été sélectionnés sur chaque spectre MS pour être isolés et fragmentés en mode CID avec 2 énergies réglées en utilisant le « collision energy profile ». L'ensemble du système a été complètement contrôlé par le programme MassLynx 4.1 (SCN 566, Waters, Milford, MA). Les données brutes collectées pendant les analyses nanoLC-MS/MS ont été traitées et converties avec le programme ProteinLynx Browser 2.3 (23, Waters, Milford, MA) en fichier « peak list » .pkl.

1.3. Chapitre 3

Le détail des expériences et des protocoles analytiques est décrit dans la publication des résultats.

1.4. Chapitre 4

Le détail des expériences et des protocoles analytiques est décrit dans la publication en cours de soumission jointe au chapitre.

2. Partie II des résultats

2.1. Chapitre 2

2.1.1. Préparation des fractions protéiques microsomales

Chaque fraction microsomale a été préparée à partir de 20 g de feuilles provenant de 8 plantes différentes (pour minimiser les variabilités biologiques d'une plante à l'autre). Trois fractions microsomales ont été préparées par condition biologique (*Arabidopsis thaliana* sauvage et mutant *hmgr1-1*) et donc 24 plantes différentes ont été utilisées en tout par condition biologique. Pour chaque préparation, les 20 g de feuilles ont été broyées avec un homogénéiseur de type « ultra-thurax » pendant 2 min à 4°C dans le tampon de broyage (250 mM sucrose, 100 mM HEPES/KOH, pH 7.5, 15 mM EGTA, 5 % glycérol, 0.5 % polyvinylpyrrolidone K25, 3 mM dithiothréitol, 1 mM phenylmethylsulfonyl fluoride, 50 mM sodium pyrophosphate, 25mM sodium fluoride, 1 mM sodium molybdate). La solution a ensuite été filtrée à travers une gaze (0.8mm de diamètre) retenant les débris cellulaires et centrifugée à 3000 g pendant 10 min à 4°C pour éliminer les débris cellulaires qui se retrouvent dans le culot. Les fractions microsomales ont finalement été récupérées grâce à une deuxième étape de centrifugation à 100000 g pendant 1 heure à 4°C. Une étape finale de lavage des microsomes avec du tampon de broyage a permis d'éliminer une grande partie des protéines solubles contaminant la préparation. Le tube qui a servi à la préparation a ensuite été plongé dans l'azote liquide pour décoller le culot et le transférer dans un tube (2 mL) qui sera utilisé pour la suite de la préparation.

2.1.2. Solubilisation et digestion des fractions protéiques microsomales

Chaque fraction protéique microsomale a été reprise dans 100 µL d'une solution : 9M Urée, 2 % CHAPS, 5mM DTT. La digestion a été réalisée par ajout de 20 µg de trypsine solubilisée dans 1.3 mL d'une solution à 50 mM de bicarbonate d'ammonium. Après 1 heure d'incubation à 37°C, on ajoute 10 µg de trypsine solubilisée dans 50µL d'une solution à 50 mM de bicarbonate d'ammonium. Après 16 heures d'incubation à 37°C, chaque mélange a été acidifié avec de l'acide acétique (5 % final). La préparation a ensuite été centrifugée à 15000 g pendant 15 minutes et le culot obtenu a été éliminé. Le surnageant obtenu est conservé pour être analysé (40 nmol de peptides estimés dans 700 µL de solution).

2.1.3. Dessalage des peptides

Les mélanges peptidiques issus de chaque fraction protéique microsomale ont été chargés sur des micro-colonnes « faites maison » préparées avec 100 μL de phase POROS oligo R3 et équilibrées avec une solution à 0.1 M d'acide acétique. Après lavage avec 500 μL d'une solution à 0.1 M d'acide acétique, les peptides ont été élués avec 500 μL d'une solution 80 % ACN/ 0.1M acide acétique. Les éluats ont été partagés en deux fractions égales qui ont été chacune évaporée à sec.

2.1.4. Marquage TMPP

Pour chaque mélange peptidique issu d'une fraction protéique microsomale, la moitié de l'échantillon dessalé et évaporé à sec a été repris dans le tampon de marquage TMPP (tampon tris-HCl 50mM pH 8.3). Une solution à 0.1 M de TMPP dans ACN : eau (2 : 8, v/v) a été ajoutée à une stoechiométrie estimée 200:1 et la réaction a été réalisée pendant 1 H.

2.1.5. Enrichissement des phosphopeptides par IMAC fer

Pour chaque mélange peptidique issu d'une fraction protéique microsomale, la partie de l'échantillon soumise au marquage TMPP a été acidifiée à l'aide d'une solution à 0.1 M d'acide acétique (200 μL final) et la partie de l'échantillon non modifiée a été reprise également dans 200 μL d'une solution à 0.1 M d'acide acétique. Les 2 parties de chaque échantillon ont finalement été rassemblées. Les 400 μL de solution provenant de chacune des fractions protéiques microsomales ont été chargées sur des micro-colonnes « faites maison » préparées avec 100 μL de matrice IMAC NTA Fe^{3+} équilibrées avec une solution à 0.1 M d'acide acétique. Après lavage avec 500 μL d'une solution à 0.1 M d'acide acétique puis 500 μL d'une solution 0.1 M NaCl, 25 % ACN, 1 % acide acétique, les peptides ont finalement été élués avec 450 μL d'une solution 50mM NaH_2PO_4 pH 9.4 et acidifiés avec 5 % HCOOH final. Après dessalage sur colonne POROS oligo R3 (comme décrit précédemment), les éluats IMAC des 3 échantillons issus d'*Arabidopsis thaliana* sauvage ont été rassemblés d'une part et les éluats IMAC des 3 échantillons issus du mutant *hmgr1-1* ont été rassemblés d'autre part pour minimiser les variabilités techniques (environ 1.5 mL final). Les échantillons finaux ont été analysés directement par nanoLC-MS/MS ou ont subi un enrichissement par chromatographie d'affinité TiO_2 ou un fractionnement RP-HPLC à pH basique avant analyses nanoLC-MS/MS. Les fractions non retenues de l'enrichissement IMAC ont été rassemblées par condition biologique, leur volume a été réduit pour être soumis à un enrichissement par chromatographie d'affinité TiO_2 ou par une combinaison de la précipitation au phosphate de calcium avec la chromatographie d'affinité TiO_2 .

2.1.6. Enrichissement des phosphopeptides par TiO_2

L'enrichissement des phosphopeptides par chromatographie d'affinité TiO_2 a été réalisé sur des pointes de pipette « UptiTip coated Titanium » (Interchim, Montluçon, France). Les mélanges peptidiques ont été enrichis selon le protocole préconisé par le fabricant.

2.1.7. Précipitation des phosphopeptides au phosphate de calcium

A partir de 50 μL de solution peptidique, 2 μL d'une solution Na_2HPO_4 à 0.5 M ont été ajoutés suivi de 0.5 μL d'une solution de $\text{NH}_3\cdot\text{H}_2\text{O}$ à 2 M. Le pH de la solution doit être voisin de 10. Si ce n'est pas le cas, la solution de $\text{NH}_3\cdot\text{H}_2\text{O}$ à 2 M est ajoutée de nouveau jusqu'à obtenir le pH souhaité. Ensuite, 2 μL d'une solution de CaCl_2 à 2 M ont été ajoutés et le mélange a été vortexé. Un précipité est apparu. Le mélange a ensuite été centrifugé à 20000 g pendant 10 minutes. Le surnageant a été éliminé. 60 μL d'une solution de CaCl_2 à 80 mM ont été ajoutés pour suspendre et laver le précipité et le mélange a de nouveau été vortexé puis centrifugé à 20000 G pendant 10 minutes. Le surnageant a été éliminé et le précipité a été redissout dans 20 μL d'une solution à 5 % HCOOH .

2.1.8. Analyses nanoLC-MS/MS

Les extraits peptidiques enrichis en phosphopeptides ont été analysés par nanoLC-MS/MS sur un système chromatographique Agilent 1100 Series HPLC-Chip/MS system (Agilent Technologies, Palo Alto, USA) couplé à un spectromètre de masse HCT Ultra ion trap (Bruker Daltonics, Bremen, Germany). Les séparations chromatographiques ont été réalisées sur une puce microfluidique contenant une colonne Zorbax 300SB-C18 (75 μm de diamètre interne \times 150 mm de longueur, particules de 3.5 μm) et une précolonne d'enrichissement Zorbax 300SB-C18 (40 nL) (Agilent Technologies). Les mélanges peptidiques ont été chargés sur la précolonne d'enrichissement Zorbax 300SB-C18 en utilisant une phase mobile aqueuse avec 0.1 % d'acide formique à 3.75 $\mu\text{L min}^{-1}$. Après dessalage, les peptides ont été élués avec un gradient 8-70 % acétonitrile (+0.1 % d'acide formique) délivré sur 47 minutes à un débit de 300 nL min^{-1} au travers de la colonne analytique Zorbax 300SB-C18. Le spectromètre de masse HCT Ultra ion trap a été calibré avec des composés de référence. Les paramètres généraux du spectromètre de masse ont été les suivants : tension capillaire, -1750V; gaz de séchage, 3 litres/min; température de séchage, 300°C. Le système a opéré avec un basculement automatique entre les modes MS et MS/MS. L'analyse MS a été réalisée dans le mode de résolution "standard-enhanced" à une vitesse de balayage de 8100 m/z par seconde avec un contrôle de la charge d'ion à 100000 dans un temps de remplissage maximale de 200 ms et un total de 4 spectres MS ont été moyennés pour obtenir les spectres MS finaux. Les 3 ions les plus intenses, préférentiellement doublement ou triplement chargés ont été sélectionnés dans chaque spectre MS pour être isolés et fragmentés. L'analyse MS/MS a été réalisée dans le mode de résolution "ultrascan" à une vitesse de balayage de 26000 m/z par seconde avec un contrôle de la charge d'ion à 300000 et un total de 2

spectres MS/MS ont été moyennés pour obtenir les spectres MS/MS finaux. ». L'ensemble du système a été complètement contrôlé par les programmes ChemStation Rev. B.01.03 (Agilent Technologies) et EsquireControl 6.1 Build 78 (Bruker Daltonics). Les données brutes collectées pendant les analyses nanoLC-MS/MS ont été traitées et converties avec le programme DataAnalysis 3.4 Build 169 en fichiers « peak list » .mgf.

2.1.9. Fractionnement RP-HPLC à pH basique

Le fractionnement a été réalisé sur un système chromatographique microHPLC (Agilent Series 1100) équipé d'un micro-collecteur de fractions. Une partie des extraits peptidiques enrichis en phosphopeptides par chromatographie d'affinité IMAC (environ 1/5) ont été injectés sur une colonne Zorbax 300Extend-C18 (300 µm de diamètre interne × 150 mm de longueur, particules de 3.5 µm) après réduction du volume, élimination de l'acétonitrile et basification avec l'ajout de triéthylamine. Les échantillons ont été chargés à 4 µL min⁻¹ avec (A) une solution de triéthylamine à 72 mM triéthylamine titrée à pH 10 avec de l'acide acétique. Après 15 minutes de conditions isocratiques à 100 % A, l'élution des peptides a été réalisée avec (B) une solution de triéthylamine à 72 mM et d'acide acétique à 52 mM dans l'acétonitrile. Nous avons utilisé un gradient 0-55 % B en 55 minutes. Des fractions de 8µL ont été collectées toutes les 2 minutes sur une période de 48 minutes entre 20 et 68 minutes d'analyse. L'acétonitrile a été éliminé par évaporation à sec des fractions collectées. Les fractions ont finalement été reprises en ajoutant 4 µL d'une solution à 0.1 % HCOOH.

2.1.10. Identification des phosphopeptides

L'ensemble des données MS/MS ont été soumises à une interrogation Mascot dans une version target-decoy de la banque protéique d'*Arabidopsis thaliana* téléchargée sur le site de Uniprot (<http://www.uniprot.org/>) en novembre 2007, avec une tolérance de 0.5 Da sur la masse des précurseurs et des fragments, en autorisant une coupure manquée par la trypsine et avec l'oxydation des méthionines, la phosphorylation des sérines, thréonines et tyrosine ainsi que la modification N-terminale des peptides au TMPP spécifiées comme modifications variables. L'attribution des sites de phosphorylation a été validée quand une différence de score minimale de 5 a pu être observée sur les scores de corrélation Mascot entre l'attribution du site la plus sûre et la deuxième attribution sur la même séquence peptidique.

2.1.11. Quantification relative des phosphopeptides entre les extraits protéiques d'*A. thaliana* sauvage et mutant *hmgr1-1* par approche spectral counting

Cette quantification relative et les tests statistiques liés ont été réalisés selon le principe décrit dans [Heintz et al., 2009] à la différence qu'ici la quantification a été réalisée au niveau du phosphopeptide et pas de la protéine.

3. Partie III des résultats

3.1. Chapitre 2

Le détail des expériences et des protocoles analytiques est décrit dans la publication des résultats.

3.2. Chapitre 3

3.2.1. Expériences nanoLC-MS/MS dans l'étude différentielle des protéomes hépatiques chez le rat dénutri

Les extraits peptidiques ont été analysés par nanoLC-MS/MS sur un système chromatographique Agilent 1100 Series HPLC-Chip/MS system (Agilent Technologies, Palo Alto, USA) couplé à un spectromètre de masse HCT Ultra ion trap (Bruker Daltonics, Bremen, Germany). Les séparations chromatographiques ont été réalisées sur une puce microfluidique contenant une colonne Zorbax 300SB-C18 (75 μm de diamètre interne \times 150 mm de longueur, particules de 3.5 μm) et une précolonne d'enrichissement Zorbax 300SB-C18 (40 nL) (Agilent Technologies). Les mélanges peptidiques ont été chargés sur la précolonne d'enrichissement Zorbax 300SB-C18 en utilisant une phase mobile aqueuse avec 0.1 % d'acide formique à 3.75 $\mu\text{L min}^{-1}$. Après dessalage, les peptides ont été élués avec un gradient 10-85 % méthanol (+0.1 % d'acide formique) délivré sur 26 minutes à un débit de 300 nL min^{-1} au travers de la colonne analytique Zorbax 300SB-C18. Le spectromètre de masse HCT Ultra ion trap a été calibré avec des composés de référence. Les paramètres généraux du spectromètre de masse ont été les suivants : tension capillaire, -1750V; gaz de séchage, 3 litres/min; température de séchage, 300 °C. Le système a opéré avec un basculement automatique entre les modes MS et MS/MS. L'analyse MS a été réalisée dans le mode de résolution "standard-enhanced" à une vitesse de balayage de 8100 m/z par seconde avec un contrôle de la charge d'ion à 100000 dans un temps de remplissage maximale de 200 ms et un total de 3 spectres MS ont été moyennés pour obtenir les spectres MS finaux. Les 3 ions les plus intenses, préférentiellement doublement ou triplement chargés ont été sélectionnés dans chaque spectre MS pour être isolés et fragmentés. L'analyse MS/MS a été réalisée dans le mode de résolution "ultrascan" à une vitesse de balayage de 26000 m/z par seconde avec un contrôle de la charge d'ion à 300000 et un total de 2 spectres MS/MS ont été moyennés pour obtenir les spectres MS/MS finaux. ». L'ensemble du système a été complètement contrôlé par les programmes ChemStation Rev. B.01.03 (Agilent Technologies) et EsquireControl 6.1 Build 78 (Bruker Daltonics). Les données brutes collectées pendant les analyses nanoLC-MS/MS ont été traitées et converties avec le programme DataAnalysis 3.4 Build 169 en fichiers « peak list » .mgf.

3.2.2. Expériences nanoLC-MS/MS dans l'étude différentielle des processomes hépatiques chez le rat dénutri

Les extraits peptidiques ont été analysés par nanoLC-MS/MS sur un système chromatographique nanoACQUITY Ultra-Performance-LC (UPLC, Waters, Milford, MA) couplé à un spectromètre de masse « SYNAPT hybrid quadrupole orthogonal acceleration time-of-flight tandem mass spectrometer » (Waters, Milford, MA). Les mélanges peptidiques ont été chargés sur une précolonne de concentration Symmetry C18 (180 μm de diamètre interne \times 20 mm de longueur, particules de 5 μm ; Waters) en utilisant une phase mobile aqueuse avec 0.1 % d'acide formique à 5 $\mu\text{L min}^{-1}$. Après dessalage, les peptides ont été élués avec un gradient 10-50 % acétonitrile (+0.1 % d'acide formique) délivré sur 80 minutes à un débit de 400 nL min^{-1} au travers de la colonne analytique BEH130 C18 (75 μm de diamètre interne \times 200 mm de longueur, particules de 1.7 μm ; Waters). Les paramètres généraux du spectromètre de masse ont été les suivants : tension capillaire, 3500 V ; tension de cône, 35 V. Pour les expériences de spectrométrie de MS en tandem, le système a opéré avec un basculement automatique entre les modes MS et MS/MS. Les 3 ions les plus intenses, préférentiellement doublement et triplement chargés, ont été sélectionnés sur les spectres MS pour être isolés et fragmentés en mode CID avec 2 énergies réglées en utilisant le « collision energy profile ». L'ensemble du système a été complètement contrôlé par le programme MassLynx 4.1 (SCN 566, Waters, Milford, MA). Les données brutes collectées pendant les analyses nanoLC-MS/MS ont été traitées et converties avec le programme ProteinLynx Browser 2.3 (23, Waters, Milford, MA) en fichier « peak list » .pkl.

Heintz, D., S. Gallien, S. Wischgoll, A. K. Ullmann, C. Schaeffer, A. K. Kretzschmar, A. van Dorsselaer and M. Boll

"Differential membrane proteome analysis reveals novel proteins involved in the degradation of aromatic compounds in *Geobacter metallireducens*." *Mol Cell Proteomics*, **2009**,

TRAVAUX SUPPLEMENTAIRES

TRAVAUX SUPPLEMENTAIRES

Plusieurs autres études protéomiques ont pu être réalisées au cours de ce travail de thèse. Dans cette partie « travaux supplémentaires » sont joints une publication déjà parue et les « abstracts » de quatre autres publications qui ont été soumises.

Trois de ces cinq publications traitent d'études protéomiques différentielles par gel 2D. Ces études ont été réalisées sur les protéomes apoplastiques de feuilles de plantes alimentaires sensibles ou résistantes à la toxicité du manganèse.

Les deux autres publications traitent de la caractérisation de deux communautés microbiennes sur deux sites plus ou moins contaminés à l'arsenic. Ces caractérisations ont été réalisées à différents points de vue (chimique, ARN 16S, phylogénie, diversité des gènes de résistance à l'arsenic) et notamment du point de vue protéomique.

Contrairement à l'étude méta-protéo-génomique présentée dans le Chapitre 4 de la Partie I des résultats, ici, les analyses métagénomiques ont été réalisées :

- ✓ Sur les extraits protéiques issus de tous les micro-organismes présents dans les sédiments des ruisseaux (sans enrichissement spécifique en micro-organismes bactériens).
- ✓ Sans connaissance préalable des séquences génomiques des micro-organismes majeurs des sites.

L'étude de communautés bactériennes dans leur milieu naturel est une discipline en plein développement à l'heure actuelle. Jusque-là, ces études étaient réalisées le plus souvent par analyse métagénomique mais nous pensons que l'analyse métagénomique a également un rôle majeur à jouer dans ce type d'étude. C'est pour cette raison que cette thématique prend une importance croissante au laboratoire à l'heure actuelle.

RESEARCH PAPER

Characterization of leaf apoplastic peroxidases and metabolites in *Vigna unguiculata* in response to toxic manganese supply and silicon

Hendrik Führs¹, Stefanie Götze¹, André Specht¹, Alexander Erban², Sébastien Gallien³, Dimitri Heintz⁴, Alain Van Dorsselaer³, Joachim Kopka², Hans-Peter Braun⁵ and Walter J. Horst^{1,*}

¹ Institute of Plant Nutrition, Faculty of Natural Sciences, Leibniz University Hannover, Herrenhäuser Str. 2, D-30419 Hannover, Germany

² Max Planck Institute of Molecular Plant Physiology, Am Mühlenberg 1, D-14476 Potsdam-Golm, Germany

³ Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC-DSA, ULP, CNRS, UMR7178 ; 25 rue Becquerel, F-67087 Strasbourg, France

⁴ Institut de Biologie Moléculaire des Plantes (IBMP) CNRS-UPR2357, ULP, F-67083 Strasbourg, France

⁵ Institute of Plant Genetics, Faculty of Natural Sciences, Leibniz University Hannover, Herrenhäuser Str. 2, D-30419 Hannover, Germany

Received 7 November 2008; Accepted 26 January 2009

Abstract

Previous work suggested that the apoplastic phenol composition and its interaction with apoplastic class III peroxidases (PODs) are decisive in the development or avoidance of manganese (Mn) toxicity in cowpea (*Vigna unguiculata* L.). This study characterizes apoplastic PODs with particular emphasis on the activities of specific isoenzymes and their modulation by phenols in the Mn-sensitive cowpea cultivar TVu 91 as affected by Mn and silicon (Si) supply. Si reduced Mn-induced toxicity symptoms without affecting the Mn uptake. Blue Native-PAGE combined with Nano-LC-MS/MS allowed identification of a range of POD isoenzymes in the apoplastic washing fluid (AWF). In Si-treated plants Mn-mediated induction of POD activity was delayed. Four POD isoenzymes eluted from the BN gels catalysed both H₂O₂-consuming and H₂O₂-producing activity with pH optima at 6.5 and 5.5, respectively. Four phenols enhanced NADH-peroxidase activity of these isoenzymes in the presence of Mn²⁺ (*p*-coumaric = vanillic >> benzoic > ferulic acid). *p*-Coumaric acid-enhanced NADH-peroxidase activity was inhibited by ferulic acid (50%) and five other phenols (50–90%). An independent component analysis (ICA) of the total and apoplastic GC-MS-based metabolome profile showed that Mn, Si supply, and the AWF fraction (AWF_{H₂O}, AWF_{NaCl}) significantly changed the metabolite composition. Extracting non-polar metabolites from the AWF allowed the identification of phenols. Predominantly NADH-peroxidase activity-inhibiting ferulic acid appeared to be down-regulated in Mn-sensitive (+Mn, –Si) and up-regulated in Mn-tolerant (+Si) leaf tissue. The results presented here support the previously hypothesized role of apoplastic NADH-peroxidase and its activity-modulating phenols in Mn toxicity and Si-enhanced Mn tolerance.

Key words: BN-PAGE, cowpea, leaf apoplast, metabolome, manganese toxicity, phenolics, proteome.

Introduction

Manganese (Mn) in plants is an essential micronutrient (Marschner, 1995). However, at supra-optimum supply Mn readily becomes toxic to plants. Mn toxicity in crops is a widely distributed plant disorder mainly on acidic and

insufficiently drained soils with low redox potentials thus leading to high amounts of plant-available Mn (Horst, 1988).

In cowpea, Mn-resistant cultivars do not differ in Mn accumulation from Mn-sensitive cultivars (Horst, 1980;

* To whom correspondence should be addressed: E-mail: horst@pflern.uni-hannover.de

© 2009 The Author(s).

This is an Open Access article distributed under the terms of the Creative Commons Attribution Non-Commercial License (<http://creativecommons.org/licenses/by-nc/2.0/uk/>) which permits unrestricted non-commercial use, distribution, and reproduction in any medium, provided the original work is properly cited.

Führs *et al.*, 2008). Therefore, in this species Mn resistance is regarded as Mn tolerance (Horst, 1983). Typical Mn stress-induced toxicity symptoms in cowpea develop primarily on older leaves as distinct brown spots located in the leaf apoplast of the epidermis starting at the leaf base, then spreading to the tip, followed by chlorosis, and, finally, leaf shedding (Horst and Marschner, 1978*b*; Horst, 1982).

The brown spots consist of oxidized Mn and oxidized phenolic compounds (Wissemeier and Horst, 1992). Hence, the oxidation of Mn²⁺ and phenols mediated by apoplastic PODs was proposed to be a key reaction leading to Mn toxicity (Fecht-Christoffers *et al.*, 2006). Class III apoplastic PODs (EC 1.11.17) belong to multigenic families (Passardi *et al.*, 2004) with various functions in plant growth (for more information see Passardi *et al.*, 2005). PODs are polyfunctional enzymes that undergo two reaction cycles: the peroxidase–oxidase cycle (with NADH as substrate also called NADH-peroxidases) resulting in H₂O₂ production (Halliwell, 1978) and the peroxidase cycle (with guaiacol as phenol substrate also called guaiacol-peroxidase) leading to H₂O₂ consumption (Fecht-Christoffers *et al.*, 2003*a*, *b*). H₂O₂-producing POD activity was intensively studied with respect to numerous exogenous factors like ambient pH (Bolwell *et al.*, 1995, 2001), phenol composition (Halliwell, 1978; Fecht-Christoffers *et al.*, 2006), and Mn²⁺ concentration *in vivo* (Yamazaki and Piette, 1963; Halliwell, 1978).

Fecht-Christoffers *et al.* (2006, 2007) investigated H₂O₂-producing activity of apoplastic peroxidases of cowpea *in vitro* and found that not only Mn²⁺ but also phenols are required to induce NADH-peroxidase activity. Increasing Mn concentrations in the leaf tissue and the AWF affected the total apoplastic phenol concentration and composition. Crosswise combining of AWF metabolites with AWF proteins from cultivars differing in Mn tolerance revealed a significant effect on NADH-peroxidase activity. They concluded that the apoplastic phenol composition and its interaction with PODs are decisive in the development or avoidance of Mn toxicity.

Silicon is a beneficial element for most plants (Epstein, 1999), and alleviates heavy metal toxicities, for example, aluminium and Mn toxicity. The alleviative effect of Si on Mn toxicity was described for common bean and cowpea (Horst and Marschner, 1978*a*; Iwasaki *et al.*, 2002*a*, *b*), cucumber (Rogalla and Römheld, 2002; Shi *et al.*, 2005), and pumpkin (Iwasaki and Matsumura, 1999). For cowpea, Horst and Marschner (1978*a*) found that leaf Mn was more evenly distributed in Si-treated cowpea plants. Horst *et al.* (1999) demonstrated a reduction in apoplastic Mn concentrations due to Si supply and concluded that Si changes apoplastic Mn-binding properties, even though this could only partly explain Si-mediated alleviation of Mn toxicity (Iwasaki *et al.*, 2002*b*). It was found that toxicity symptoms and guaiacol-peroxidase activities were more closely related to apoplastic Si concentrations than to apoplastic Mn concentrations, indicating a more direct involvement of Si nutrition in detoxification of apoplastic Mn.

The work presented here specifically addressed the hypothesis that the activities of specific apoplastic perox-

idases and their modulation by metabolites are decisive for Mn toxicity and Si-induced enhanced Mn tolerance in the Mn-sensitive cowpea cultivar TVu 91.

Materials and methods

Plant material

Cowpea [*Vigna unguiculata* (L.) Walp., cv. TVu 91] was grown hydroponically in a growth chamber under controlled environmental conditions at 30/27 °C day/night temperatures, 75±5% relative humidity, and a photon flux density of 150 µmol m⁻¹ s⁻¹ photosynthetic active radiation (*PAR*) at mid-plant height during a 16 h photoperiod. After germination in 1 mM CaSO₄ for 7 d, seedlings were transferred to a constantly aerated nutrient solution with four plants in one 5.0 l pot. The composition of the nutrient solution was (µM): Ca(NO₃)₂ 1000, KH₂PO₄ 100, K₂SO₄ 375, MgSO₄ 325, FeEDDHA 20, NaCl 10, H₃BO₃ 8, MnSO₄ 0.2, CuSO₄ 0.2, ZnSO₄ 0.2, Na₂MoO₄ 0.05. Silicon-treated plants (+Si) received Si in form of Aerosil (Horst and Marschner, 1978*a*; chemically clean silicic acid, solubility in water: 0.6–0.75 mg l⁻¹ or 20–26.5 µM). After preculture for 14 d, the Mn concentration in the nutrient solution was increased from 0.2 µM (–Mn) to 50 µM (+Mn) for 4 d or 6 d. The nutrient solution was changed two to three times per week to avoid nutrient deficiencies.

Extraction of water-soluble and ionically bound apoplastic proteins and metabolites

Apoplastic washing fluid (AWF) was extracted by a vacuum infiltration/centrifugation technique according to Fecht-Christoffers *et al.* (2003*a*, *b*). Leaves were infiltrated with chilled dH₂O by reducing the pressure to –35 hPa followed by a slow relaxation. AWF_{H₂O} was recovered by centrifugation at 1324 *g* for 5 min at 4 °C. Afterwards, the same leaves were infiltrated with chilled 0.5 M NaCl solution and AWF_{NaCl} was recovered as described above. Malate dehydrogenase (MDH) activity in both AWF fractions showed a cytoplasmic contamination of less than 1% (data not shown). Until further analysis the AWF was stored at –80 °C.

Quantification of toxicity symptoms

For the quantification of Mn toxicity symptoms, the density of brown spots was counted on a 1.54 cm² area at the base and tip on the upper side of the second oldest middle trifoliate leaf and calculated on 1 cm² base.

Manganese analysis

Manganese in the bulk-leaf tissue was determined in the second oldest middle trifoliate leaf after dry ashing at 480 °C for 8 h, dissolving the ash in 6 M HCl with 1.5% (w/v) hydroxylammonium chloride, and then diluting (1:10 v/v) with double demineralized water. Apoplastic Mn concentrations were measured in 1:10 dilutions of the AWF. Both

measurements were carried out by optical inductively-coupled plasma-emission spectroscopy (Spectro Analytical Instruments GmbH, Kleve, Germany).

Silicon analysis

Monomeric Si concentration in the AWF was determined according to Iwasaki *et al.* (2002a, b). AWF and a standard solution (0–100 µg Si ml⁻¹ AWF) were mixed with 250 µl of staining solution (1:1 mix of 0.08 M H₂SO₄ and 20 g l⁻¹ (NH₄)₆Mo₇O₂₄·4H₂O). After 30 min of incubation 250 µl of freshly prepared ascorbic acid (0.1 g 25 ml⁻¹) and 250 µl tartaric acid (0.85 g 25 ml⁻¹) were added. Samples were measured at λ=811 nm in a Microplate-Reader (µQuant, BioTek Instruments, Germany).

Determination of the protein concentration in the AWF and AWF concentrates

The protein concentration in the AWF for the calculation of specific enzyme activities was determined according to Bradford (1976). The protein concentration of AWF concentrates was measured for 1D BN-PAGE using the 2-D Quant Kit[®] (GE Healthcare, USA) according to the manufacturer's instructions.

Determination of specific peroxidase activities in the AWF

For the measurement of H₂O₂-consuming guaiacol-peroxidase activities in the AWF, the oxidation of the substrate guaiacol was determined spectrophotometrically at λ=470 nm (UVIKON 943, BioTek Instruments GmbH, Neufahrn, Germany). Samples were mixed with guaiacol solution (20 mM guaiacol in 10 mM Na₂HPO₄ buffer, pH 6) and 0.03% (v/v) H₂O₂. For calculation of enzyme activities the molar extinction coefficient 26.6 l (mmol cm)⁻¹ was used.

For the measurement of the H₂O₂-producing NADH-peroxidase activity in the AWF, samples were mixed with MnCl₂ (16 mM), *p*-coumaric acid (1.6 mM) and NADH (0.22 mM). The NADH oxidation-dependent decline in absorption at λ=340 nm was determined. For calculation of enzyme activities the molar extinction coefficient 1.13 l (mmol cm)⁻¹ was used.

1D BN-PAGE of apoplastic proteins and POD activity staining

For protein separation by electrophoresis under native conditions, the proteins of the AWF were concentrated at 4 °C by using centrifugal concentrators with a molecular mass cut-off at 5 kDa (Vivaspin 6, Vivascience, Hannover, Germany). Running conditions were used according to the manufacturer's instructions.

Proteins were separated via BN-PAGE according to Jansch *et al.* (1996). Protein samples were combined with Coomassie Blue solution [5% (w/v) Serve Blue G and 750 mM aminocaproic acid] and 10% (v/v) glycerol (100%). Samples were loaded onto a native acrylamide gel with

a 4% (w/v) stacking gel and a 12% to 20% (w/v) gradient separation gel. Electrophoresis was carried out at 100 V and 6–8 mA for 45 min followed by 13 h at 15 mA (max. 500 V).

NADH-peroxidase activity in the gel was determined by NBT staining to detect O₂⁻ radicals or by DAB staining (data not shown) to detect H₂O₂. The staining solution finally consisted of 16 mM MnCl₂, 1.6 mM *p*-coumaric acid, 0.22 mM NADH, and 2.5 mg ml⁻¹ NBT in order to detect O₂⁻ radicals, that are proposed to be produced during the NADH-peroxidase activity of PODs (Halliwell, 1978) because a direct detection of H₂O₂ by DAB staining was difficult due to the high gel background caused by Coomassie. Gels were stained for 30 min at room temperature. The gels were afterwards soaked in 20 mM guaiacol (in 10 mM Na₂HPO₄) and 0.03% (v/v) H₂O₂ for 3 min to detect guaiacol-peroxidase activity.

For preparative BN-PAGE guaiacol-peroxidase staining was carried out only for a few seconds in order to reduce enzyme damage by product-enzyme interaction.

Electroelution of specific POD isoenzymes for further physiological characterization

Four POD isoenzymes (P1, P3, P5, and P6 in Fig. 3C) were chosen for electroelution from BN gels that was carried out according to Wehrhahn and Braun (2002). POD isoenzymes were cut from the gel and incubated for 30 min in cathodic buffer [50 mM Tricine, 15 mM BIS-TRIS, 0.1 % (w/v) Coomassie 250 G, pH 7 adjusted at 4 °C] and transferred into the chambers of an electroeluter (CBS SCIENTIFIC, Del Mar, USA). The gel pieces containing the POD isoenzymes were filled into the electroeluter containing elution buffer (25 mM Tricine, 7.5 mM BIS-TRIS, pH 7.0 adjusted at 4 °C). Electroelution was carried out for 5 h and 4 °C at 350 V and 6–10 mA, using dialysis membranes (Medicell, Kleinfeld) with a MWCO of 12–14 kDa under constant buffer circulation (Econopump, Bio-Rad Laboratories, CA, USA). Until further characterization, eluates were stored at -80 °C.

Determination of the pH optimum of the guaiacol-peroxidase and NADH-peroxidase activity of POD isoenzymes

For guaiacol-peroxidase measurements, 6 µl eluate was mixed with guaiacol (20 mM) in 0.1 M succinate buffer with the pH values 5, 5.5, 6, 6.5, and 7. The reaction was started by adding 0.3% (v/v) H₂O₂. The increase in absorption was measured at λ=470 nm using a Microplate Reader. For calculation of enzyme activities the molar extinction coefficient 26.6 l (mmol cm)⁻¹ was used.

NADH-peroxidase activity measurements were made by combining MnCl₂, *p*-coumaric acid, and NADH in final concentrations of 16 mM, 1.6 mM, and 0.66 mM, respectively, with 7.5 µl protein eluate in 0.1 M succinate buffer (as described above). The decline in absorption was determined using a Microplate Reader at λ=340 nm. For

calculation of enzyme activities the molar extinction coefficient $1.13 \text{ l (mmol cm)}^{-1}$ was used.

Determination of cofactor specificity for NADH-peroxidase activity of POD isoenzymes

The same experimental set-up as for the determination of the pH optimum was followed using succinate buffer (pH 5.5). *p*-Coumaric acid was substituted by benzoic acid, caffeic acid, chlorogenic acid, ferulic acid, gallic acid, protocatechuic acid, syringic acid, vanillic acid, and *p*-hydroxybenzoic acid in four different concentrations (1.66 mM, 0.166 mM, 0.0166 mM, and 0.00166 mM) in the measuring solution. In order to simplify this report, benzoic acid as an aromatic carboxylic acid is termed as a phenolic acid, too. For each phenol concentration specific extinction coefficients were determined and used for enzyme activity calculation (see Supplementary Table S1 at *JXB* online).

Determination of changes in NADH-peroxidase activity of POD isoenzymes as affected by combining different phenols with p-coumaric acid

To detect the effects of different phenols on *p*-coumaric acid-stimulated NADH-peroxidase activity of different isoenzymes separated by BN-PAGE 0.166 mM *p*-coumaric acid was combined with benzoic acid, caffeic acid, chlorogenic acid, ferulic acid, gallic acid, protocatechuic acid, syringic acid, vanillic acid, and *p*-hydroxybenzoic acid each at a concentration of 0.0166 mM. All other factors were kept as described for the measurement of cofactor specificity. Activity was expressed as a percentage of *p*-coumaric acid induced NADH-peroxidase activity. For each phenol concentration, specific extinction coefficients were determined and used for enzyme activity calculation (see Supplementary Table S1 at *JXB* online).

Mass spectrometric protein analysis and data interpretation

Marked BN-PAGE bands stained for guaiacol-peroxidase activity were cut and dried under vacuum. In-gel digestion was performed with an automated protein digestion system, MassPREP Station (Micromass, Manchester, UK). The gel slices were washed three times in a mixture containing 25 mM NH_4HCO_3 :acetonitrile (1:1, v/v). The cysteine residues were reduced by 50 μl of 10 mM dithiothreitol at 57 °C and alkylated by 50 μl of 55 mM iodacetamide. After dehydration with acetonitrile, the proteins were cleaved in the gel with 40 μl of 12.5 ng μl^{-1} of modified porcine trypsin (Promega, Madison, WI, USA) in 25 mM NH_4HCO_3 at room temperature for 14 h. The resulting tryptic peptides were extracted with 60% acetonitrile in 0.5% formic acid, followed by a second extraction with 100% (v/v) acetonitrile.

Nano-LC-MS/MS analysis of the resulting tryptic peptides was performed using using an Agilent 1100 series HPLC-Chip/MS system (Agilent Technologies, Palo Alto, USA) coupled to an HCT Ultra ion trap (Bruker Daltonics,

Bremen, Germany). Chromatographic separations were conducted on a chip containing a Zorbax 300SB-C18 (75 μm inner diameter \times 150 mm) column and a Zorbax 300SB-C18 (40 nl) enrichment column (Agilent Technologies).

HCT Ultra ion trap was externally calibrated with standard compounds. The general mass spectrometric parameters were as follows: capillary voltage, -1750 V ; dry gas, 3.0 l min^{-1} ; dry temperature, $300 \text{ }^\circ\text{C}$. The system was operated with automatic switching between MS and MS/MS modes. The MS scanning was performed in the standard-enhanced resolution mode at a scan rate of 8100 m/z s^{-1} with an aimed ion charge control of 100 000 in a maximal fill time of 200 ms and a total of four scans were averaged to obtain a MS spectrum. The three most abundant peptides and preferentially doubly charged ions were selected on each MS spectrum for further isolation and fragmentation. The MS/MS scanning was performed in the ultrascan resolution mode at a scan rate of $26\,000 \text{ m/z s}^{-1}$ with an aimed ion charge control of 300 000 and a total of six scans were averaged to obtain an MS/MS spectrum. The complete system was fully controlled by ChemStation Rev. B.01.03 (Agilent Technologies) and EsquireControl 6.1 Build 78 (Bruker Daltonics) softwares. Mass data collected during LC-MS/MS analyses were processed using the software tool DataAnalysis 3.4 Build 169 and converted into .mgf files. The MS/MS data were analysed using the MASCOT 2.2.0. algorithm (Matrix Science, London, UK) to search against an in-house generated protein database composed of protein sequences of Viridiplantae downloaded from <http://www.ncbi.nlm.nih.gov/sites/entrez> (on 6 March 2008) concatenated with reversed copies of all sequences ($2 \times 478\,588$ entries). Spectra were searched with a mass tolerance of 0.5 Da for MS and MS/MS data, allowing a maximum of 1 missed cleavage by trypsin and with carbamidomethylation of cysteines, oxidation of methionines, and N-terminal acetylation of proteins specified as variable modifications. Protein identifications were validated when at least two peptides with high quality MS/MS spectra (Mascot ion score greater than 31) were detected. In the case of one-peptide hits, the score of the unique peptide must be greater (minimal 'difference score' of 6) than the 95% significance Mascot threshold (Mascot ion score >51). For the estimation of the false positive rate in protein identification, a target-decoy database search was performed (Elias and Gygi, 2007).

GC-MS-based metabolite profiling

For GC-MS analysis, polar metabolite fractions were extracted from $60 \text{ mg} \pm 10 \%$ (FW) frozen plant material, ground to a fine powder, with methanol/chloroform. The fraction of polar metabolites was prepared by liquid partitioning into water/methanol (polar fraction) and chloroform (non-polar fraction) as described earlier (Roessner *et al.*, 2000; Wagner *et al.*, 2003). Metabolite samples were derivatized by methoxyamination, using a 20 mg ml^{-1} solution of methoxyamine hydrochloride in pyridine, and subsequent trimethylsilylation, with *N*-methyl-*N*-

(trimethylsilyl)-trifluoroacetamide (Fiehn *et al.*, 2000; Roessner *et al.*, 2000). A C₁₂, C₁₅, C₁₉, C₂₂, C₂₈, C₃₂, and C₃₆ n-alkane mixture was used for the determination of retention time indices (Wagner *et al.*, 2003). Ribitol and deuterated alanine were added for internal standardization. Samples were analysed using GC-TOF-MS (ChromaTOF software, Pegasus driver 1.61; LECO, <http://www.leco.com>). Four sample types (\pm Mn and \pm Si), each with five replicates, comprised an experimental data set of 20 chromatograms. The chromatograms and mass spectra were evaluated using the TagFinder software (Luedemann *et al.*, 2008).

Sample preparation for the metabolite profiling of the AWF was adapted to the respective volumes and metabolite concentrations. In this case 200 μ l of AWF_{H₂O} and AWF_{NaCl} were extracted to obtain a polar metabolite fraction, without further addition of water. The volume of methanol/chloroform was reduced to 50% as were the reagents for methoxyamination and silylation. Four sample types (two Mn treatments, and two Si treatments), each with four to five replications, in total 35 chromatograms, were analysed as described above.

In parallel free phenols (in the following termed non-polar apoplastic fraction) were extracted from AWF_{H₂O} and AWF_{NaCl}. First AWF was alkalinized with 0.5 N NaOH (ratio 1:1) overnight. Afterwards samples were acidified by adding 5 N HCl (ratio 0.1125:1). Phenols were then extracted by shaking with diethylether (ratio 1:1). Samples were then dried under nitrogen atmosphere and prepared for GC-MS analysis as described for AWF. Four sample types (two Mn treatments and two Si treatments), each with five to six replications, resulted in 48 chromatograms, which were processed as described.

GC-MS metabolite profiles were processed after conversion into NetCDF file format using the TagFinder (Luedemann *et al.*, 2008) and NIST05 software (<http://www.nist.gov/srd/mslist.htm>). The mass spectral and retention index (RI) collection of the Golm metabolome database (Kopka *et al.*, 2005; Schauer *et al.*, 2005) was used for manually supervised metabolite identification. Yet non-identified metabolic components were disregarded for the present study. Peak height representing a mass specific arbitrary detector response was used for screening the relative changes of metabolite pools. The initial mass specific responses were normalized by leaf fresh weight and ribitol recovery. AWF metabolite profiles were normalized to ribitol recovery and AWF total volume of partitioned polar (water/methanol) and non-polar (chloroform) AWF fractions.

Statistical analysis of GC-MS profiles

Prior to statistical data assessment, response ratios were calculated based on the mean response of each metabolic feature from all samples of an experimental data set. Response ratios were subsequently log₁₀-transformed. Independent component analysis (ICA) and missing value substitution was as described earlier (Scholz *et al.*, 2005). ICA was carried out using the first five principal compo-

nents obtained from a set of manually identified metabolites represented by at least three specific mass fragments each. Basic calculations of relative changes in abundance of specific metabolites due to Mn and Si treatment were made with the Microsoft Excel 2000 software program and respective embedded algorithms. For pairwise comparisons thresholds of 2-fold change in pool size and $P < 0.05$ (t test,) were applied or levels of significance indicated, namely ***, **, and * representing $P < 0.001$, 0.01, and 0.05, respectively. Logarithmic transformation of response ratios approximated the required Gaussian normal distribution of metabolite profiling data (Schaarschmidt *et al.*, 2007).

Statistical analysis of Mn and Si concentrations and apoplastic enzyme activities

Statistical analysis, if not mentioned otherwise, was carried out using SAS Release v8.0 (SAS Institute, Cary, NC). Results from analysis of variance are given according to their level of significance as ***, **, and * for $P < 0.001$, 0.01, and 0.05, respectively. Pairwise comparisons were by using Student's t test.

Results

Exposing the plants to 50 μ M Mn supply rapidly increased the Mn tissue concentration in the second oldest trifoliate leaf over the 4 d treatment period (Fig. 1A). This led to typical Mn toxicity symptoms (brown spots) after 2 d increasing up to 70 spots cm⁻² after 4 d of Mn treatment (Fig. 1B). Silicon supply did not affect leaf Mn accumulation (Fig. 1A). However, in contrast to plants cultivated without Si, Si-treated plants developed only slight Mn toxicity symptoms (2–5 spots cm⁻²) after 4 d of Mn treatment (Fig. 1B).

Since our previous work indicated a particular role of the apoplast in the expression of Mn toxicity and Mn tolerance in cowpea, our studies were focused on the AWF in particular. In this study, the leaves were submitted to a fractionated AWF extraction procedure yielding a free water-soluble fraction (AWF_{H₂O}) and an ionically bound NaCl-extractable (AWF_{NaCl}) fraction. The Mn concentration in the AWF_{H₂O} increased rapidly after 1 d of toxic Mn supply and then it tended to decrease again (Fig. 2A). Silicon application consistently enhanced the monomeric Si concentration in the AWF_{H₂O} (Fig. 2B) compared with non Si-treated plants, without consistently affecting the apoplastic Mn concentration (Fig. 2A). In the AWF_{NaCl}, the Mn concentration of the second trifoliate leaf steeply increased after 1 d Mn treatment and remained stable at a higher level than in the AWF_{H₂O} (Fig. 2C). In Si-treated plants, the Mn concentrations were slightly higher. Silicon treatment enhanced the monomeric Si concentration (Fig. 2D), but with Mn treatment duration this difference disappeared.

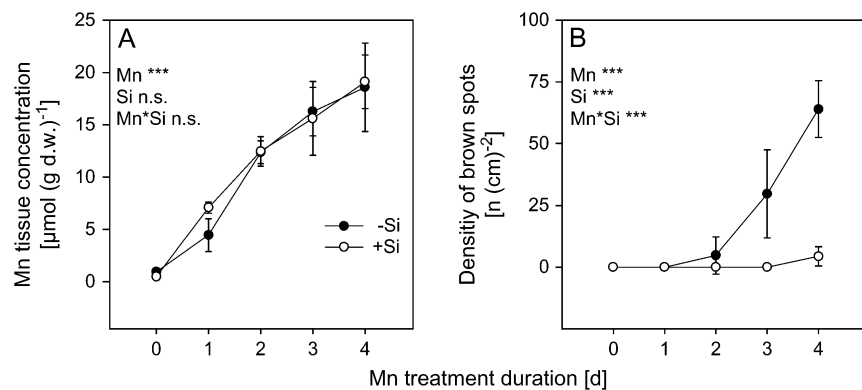


Fig. 1. Effect of Mn treatment duration and Si supply on (A) the Mn tissue concentration and (B) the density of brown spots of the second oldest trifoliate leaves of the Mn-sensitive cowpea cultivar TVu 91. After 2 weeks of preculture at $0.2 \mu\text{M}$ Mn the Mn supply was increased to $50 \mu\text{M}$ for 4 d. Silicon was supplied throughout plant culture. Results of the analysis of variance are given according to their level of significance as ***, ** or * for $P < 0.001$, 0.01 , 0.05 , respectively. Values are means \pm SD with $n=16$.

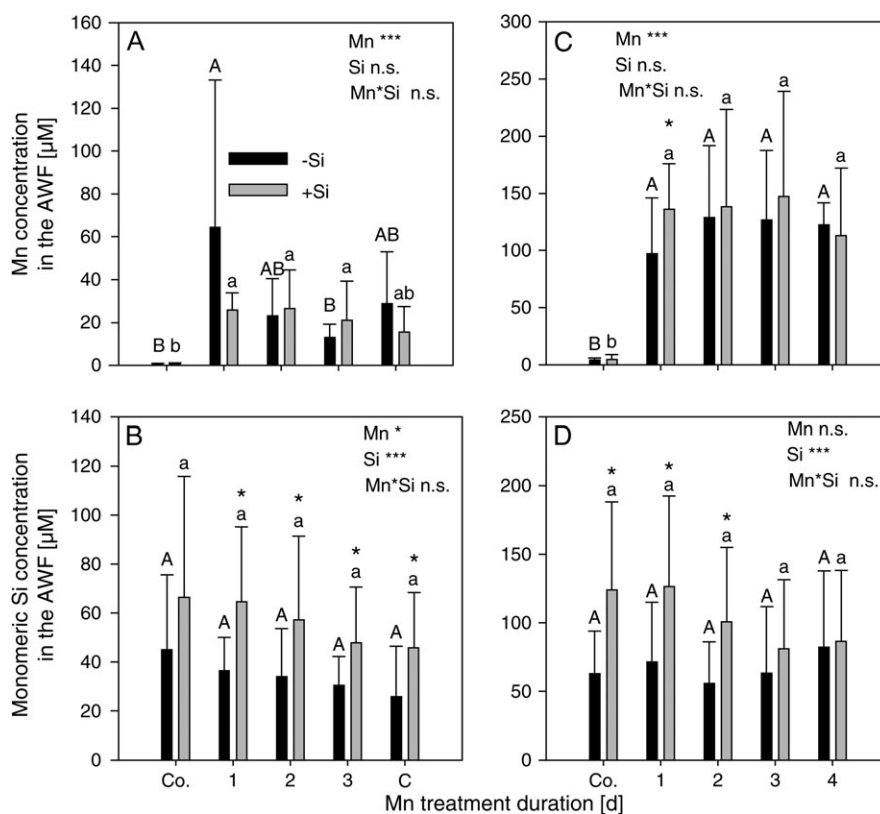


Fig. 2. Effect of Mn treatment duration and Si supply on the Mn concentration (A, C) and the monomeric Si concentration (B, D) in the water-soluble apoplasmic fraction (A, B), and in the ionically bound apoplasmic fraction (C, D) of the second oldest trifoliate leaves of the Mn-sensitive cowpea cultivar TVu 91. After 2 weeks of preculture at $0.2 \mu\text{M}$ Mn, the Mn supply was increased to $50 \mu\text{M}$ for 4 d. Silicon was supplied throughout plant culture. Results of the analysis of variance are given according to their level of significance as ***, ** or * for $P < 0.001$, 0.01 , or 0.05 , respectively. Upper case and lower case letters indicate significant differences between Mn treatment duration of -Si and +Si-treated plants, respectively, at $P < 0.05$. An asterisk on top of the columns indicates significant differences between the Si treatments for at least $P < 0.05$ according to Tukey. Values are means \pm SD with $n=16$.

In order to demonstrate the capability of the POD isoenzymes to catalyse both H_2O_2 -producing and -consuming POD activities, AWF_{NaCl} was separated by BN-PAGE and PODs in-gel stained first for NADH-peroxidase followed by staining for guaiacol-peroxidase activity (Fig. 3A, B). Despite the quite low NADH-peroxidase activity stain-

ing intensity the gels revealed that each isoenzyme showed both activities. Staining with guaiacol visualized major isoenzymes more clearly: one isoenzyme smaller than P1 and four isoenzymes greater than P1, all with low activity levels (Fig. 3B). After 6 d of Mn treatment, three additional guaiacol-peroxidase bands appeared greater than the P6

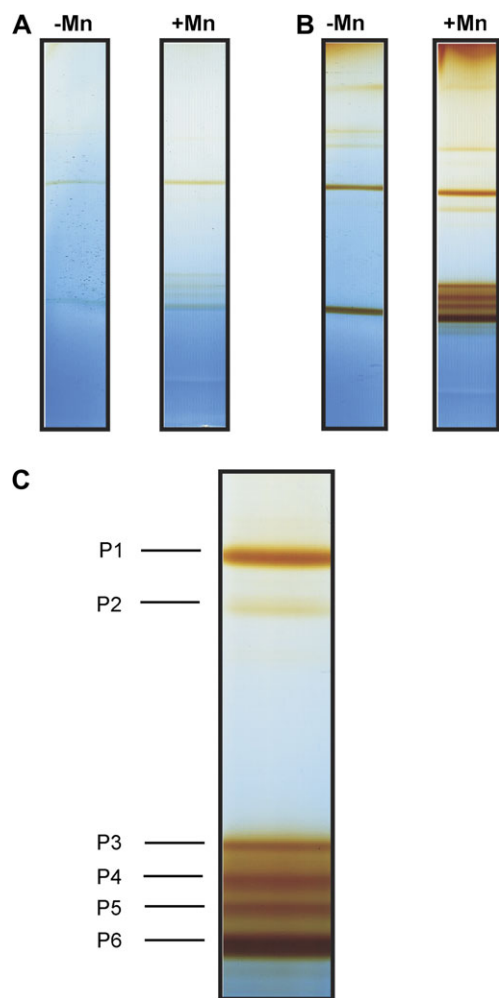


Fig. 3. AWF_{NaCl}-proteins of the second oldest trifoliolate leaf of the Mn-sensitive cultivar TVu 91 stained for (A) NADH-peroxidase and (B) guaiacol-peroxidase activity after separation by BN-PAGE. After preculture with 0.2 μM Mn (-Mn) for 14 d, plants received 50 μM (+Mn) Mn for 6 d. Fifty μl of concentrated AWF_{NaCl} containing ionically bound proteins (-Mn 60 μg , +Mn 112 μg) were loaded onto the gels. Proteins were NBT-stained for NADH-peroxidase (A) at pH 5.0 with 16 mM MnCl₂, 1.66 mM *p*-coumaric acid, 0.625 mg ml⁻¹ NBT, and 0.22 mM NADH. For guaiacol-peroxidase, proteins were stained (B) in 18 mM guaiacol (in 9 mM Na₂HPO₄) and 0.03% H₂O₂ at pH 6.0. Close up (C) shows marked isoenzymes (P1, P3, P5, P6) that were chosen for elution and further characterization of pH optima and substrate specificity.

isoenzyme and one with a MW smaller than P1. One isoenzyme with a MW greater than P1 disappeared owing to elevated Mn supply. An extensive study of in-gel activity-stained BN gels loaded rigorously with the same protein quantities comparing Mn treatments with and without Si supply and differentiating between AWF_{H₂O} and AWF_{NaCl} proteins revealed that all isoenzymes were qualitatively present in both Mn treatments, but elevated Mn supply led to an increased abundance of isoenzymes P3 and P5, especially in the water-soluble fractions (see Supplementary Figs S1 and S2 at *JXB* online). In Mn-control plants Si-treatment did not affect the POD isoenzyme pattern. Silicon

delayed but not suppressed the Mn-mediated increase in the number of POD isoenzymes in the AWF_{H₂O} (see Supplementary Fig. S1 at *JXB* online).

Figure 3C shows a close-up of those POD isoenzymes (clearly appearing after 4 d of Mn treatment), which were chosen for further characterization after elution of the proteins from the gels: P1, P3, P5, and P6, whereas P2 and P4 were only sequenced. The eluted isoenzymes P3, P5, and P6, showed both NADH-peroxidase and guaiacol-peroxidase activities (Fig. 4A, B). The specific activity was highest for P6 followed by P5. The POD isoenzyme P1 had very little guaiacol-peroxidase activity. The pH optimum for all isoenzymes showing activity was consistently 6.5 for guaiacol-peroxidase activity (Fig. 4A) and pH 5.5 for NADH-peroxidase activity (Fig. 4B).

All marked POD activity-stained protein bands (Fig. 3) were cut; proteins were digested and analysed by liquid chromatography-coupled mass spectrometry (LC-MS/MS). MS/MS searches did not always lead to a positive identification in cowpea (*Vigna unguiculata*) since its genome has not yet been sequenced, but can lead to the identification of peptides in related sequences of green plants (Viridiplantae) downloaded from <http://www.ncbi.nlm.nih.gov/sites/entrez>. Forty-four unique proteins were identified in the green plants database. To estimate the false positive rate of identification, a target-decoy database was performed (Elias and Gygi, 2007), and no additional protein was identified in reversed sequences, suggesting that our dataset contained very few or no false-positive identifications. A list of all resulting peptides, as well as their identities, is given as supplementary data (see Supplementary Table S2 at *JXB* online). Among these peptides, 11 peptides belonging to class III peroxidases could be identified (Fig. 5). At least three overlapping peptides provide evidence for at least three distinct gene products. Three peptides with amino acid substitutions were exclusively found in POD isoenzyme P1 when extracted with NaCl from Mn-treated plants (Figs 3, 5; see Supplementary Table S2 at *JXB* online).

Since apoplastic NADH-peroxidase proved to react most sensitively to toxic Mn supply and this enzyme has been attributed a key role in the expression of Mn toxicity (Fecht-Christoffers *et al.*, 2006, 2007), the NADH-peroxidase activity of the isoenzymes was further characterized for interaction with different commercially available phenols (Fig. 6) at the optimum pH identified above with *p*-coumaric acid and Mn as a cofactors. Among the 10 phenols tested, *p*-coumaric acid and vanillic acid proved to be the most effective cofactors for all isoenzymes particularly at the highest concentration level. Benzoic acid showed only little activity at the higher concentrations even though the response pattern was similar, whereas ferulic acid activated NADH-peroxidase activity only at a lower concentration. All other phenols did not induce NADH-peroxidase activity. As shown above (Fig. 4A, B) the isoenzyme P6 showed by far the highest activity.

The potential inhibitory effect of phenols on NADH-peroxidase activity was studied by adding eight phenols to the reaction mixture and monitoring their effect on

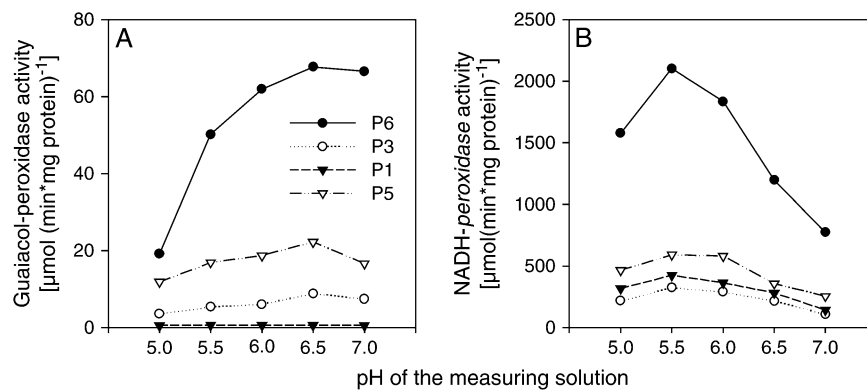


Fig. 4. Determination of the pH optimum of (A) the guaiacol-peroxidase activity and of (B) the NADH-peroxidase activity of four POD isoenzymes of the Mn-sensitive cowpea cultivar TVu 91. POD isoenzymes were eluted from BN gels that separated a mixture of AWF_{H₂O} and AWF_{NaCl} extracted from the second oldest trifoliate leaf of Mn-treated (4 d) and \pm Si-treated (as described in the Materials and methods) plants. Measurements were done in succinate buffer with pH values between 5.0 and 7.0 using 0.5 steps between the pH values. Measuring solution (0.1 M succinate buffer) for the determination of NADH-peroxidase activity consisted of 16 mM MnCl₂, 1.66 mM *p*-coumaric acid, and 0.22 mM NADH, measuring solution for guaiacol-peroxidase activity consisted of 18 mM guaiacol (in 90 mM succinate buffer) and 0.03% H₂O₂.

p-coumaric acid-stimulated enzyme activity (Fig. 7). Benzoic acid and vanillic acid did not reduce the *p*-coumaric acid-stimulated NADH-peroxidase activity and even enhanced it. All other phenols inhibited NADH-peroxidase activity by about 50% (ferulic and syringic acid) and by >90% for the other phenols. This was true for all isoenzymes.

Since metabolites were shown to affect apoplastic PODs strongly (see above and Fecht-Christoffers *et al.*, 2006), the bulk-leaf metabolome was studied in a broad range approach using GC-MS and independent component analyses (ICA) (Scholz *et al.*, 2004). Applying ICA, sample clusters were investigated according to the major variances due to the treatment-induced qualitative and quantitative changes of metabolite pools. This variance criterion was augmented by subsequent pairwise or multiple probability-based statistical significance testing.

In our factorial experimental designs both Mn and Si treatment proved to be among the most important independent components (Fig. 8A) of our data sets resulting from the bulk-leaf tissue. The analysis revealed that Mn (IC01) and Si (IC04) treatments induced significant changes in the metabolome. Silicon treatment clearly induced significant conditional differences among the Mn control treatment but only slight differences in Mn-treated plants. The Mn effect was mainly caused by changes in the concentrations of amino acids (serine, threonine, asparagine, aspartic acid), phenylalcohols (coniferylalcohol), organic acids (gluconic acid), and sugar alcohols (sorbitol) as revealed by ICA loadings. The Si effect was mainly due to differences in sugars (galactose) and organic acids (gluconic acid).

In view of the particular role of the activity of apoplastic peroxidases in Mn toxicity additionally the AWF_{H₂O} and the AWF_{NaCl} were subjected to a metabolomic analysis. The ICA showed clear differences between the AWF fractions (IC01, Fig. 8B). Also, manganese treatment induced separate clustering in both AWF fractions (IC02). In this

approach Si did not affect the sample clustering according to treatment-mediated metabolite differences. As revealed by ICA loadings, metabolites mainly responsible for the differential clustering of AWF_{H₂O} and AWF_{NaCl} were GABA, organic acids (malic acid, ribonic acid, gluconic acid), amino acids (threonine), and sugars (xylose, erythrose, fucose) among many currently unidentified metabolites. The clustering according to the Mn treatment was mainly caused by organic acids (maleic acid, malic acid, nicotinic acid, itaconic acid), amino acids (threonine, alanine), sugars (xylose, fructose, tagatose), and phenols (3-hydroxybenzoic acid).

Further fractionation of the leaf apoplastic metabolome by an extraction method specifically yielding non-polar metabolites revealed a clustering of samples according to the infiltration solution, confirming the strong experimental impact of the AWF fraction on the result (Fig. 8C). Loadings derived from ICA showed that among other currently unknown metabolites, mainly organic acids (fumaric acid, malic acid, succinic acid, citric acid, 3-oxoglutaric acid) and phenylpropanoids (*cis*- and *trans*-cinnamic acid, *p*-hydroxybenzoic acid) were responsible for this clustering.

Quantification of relative changes between treatments yielded five different phenols in this non-polar extract (Table 1) among them ferulic acid, *p*-hydroxybenzoic acid, and *p*-coumaric acid which had shown considerable inhibiting or enhancing effects, respectively, on *in vitro* NADH-peroxidase activity. Ferulic acid and *p*-coumaric acid were analytically separated into respective *cis*- and *trans*-isomers, whereas in the *in vitro* NADH-peroxidase activity-enhancing/inhibiting tests (Figs 6, 7) commercially available isomer mixtures were used. Both ferulic acid isomers showed a significant 2–4-fold reduction in abundance in Mn-treated plants compared with control plants in the AWF_{H₂O} fraction. A comparison of \pm Si treatments revealed a significantly increased abundance of benzoic acid and of ferulic

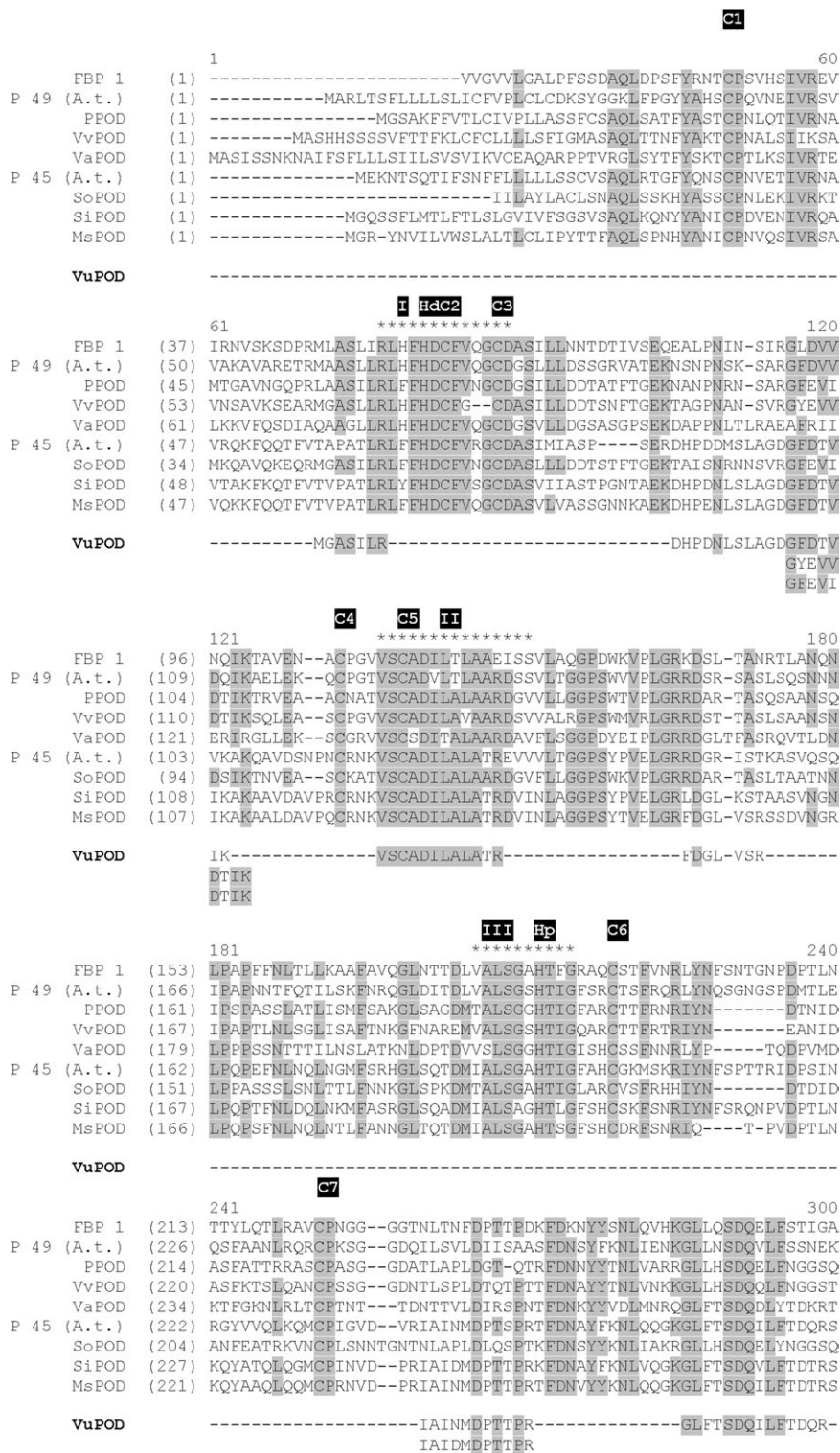


Fig. 5. Alignment of determined and deduced amino acid (aa) sequences of peroxidases of various plant species and all 11 nano LC-MS/MS-identified peroxidase peptide sequences from cowpea. Amino acid positions conserved in at least 50% of the sequences are underlaid in grey. Asterisks (*) indicate the conserved distal haem-binding domain (I), the central conserved domain of unknown function (II), and the proximal haem-binding domain. The eight cysteines (C1–C8) and the distal (Hd) and proximal (Hp) histidines are indicated, too. Abbreviations: FBP1, French Bean Peroxidase 1 (acc no. AF149277); P49 (A.t.), POD isoenzyme 49 from *Arabidopsis thaliana* (acc. no. O23237); PPOD from *Populus* ssp. (acc. no. AAX53172); VvPOD from *Vitis vinifera* (acc. no. CAO48839); VaPOD from *Vigna angularis* (acc. no. BAA01950); P45 (A.t.) POD isoenzyme 45 from *Arabidopsis thaliana* (acc. no. Q96522); SoPOD from *Spinacia oleracea* (acc. no. CAA71493); SiPOD from *Sesamum indicum* (acc. no. ABB89209), MsPOD from *Medicago sativa* (acc. no. CAC38106); VuPOD, POD peptide sequences of *Vigna unguiculata* (this study).



Fig. 5. (continued)

acid isomers (more than 3-fold) in Si-treated plants only in the AWF_{NaCl} fraction. In Si-treated plants, high Mn supply led to increased concentrations of benzoic acid in the AWF_{H_2O} fraction and to decreased abundance of ferulic acid compared with plants grown at low Mn supply. A comparison of +Mn/+Si with +Mn/-Si (Mn toxicity-showing) plants showed significantly decreased *p*-hydroxybenzoic acid concentrations. A major, however not significant, increase in abundance of *cis*-ferulic acid is indicated in the +Mn/+Si plants not showing Mn toxicity symptoms. NADH-*peroxidase* activity enhancing *p*-coumaric acid showed no changes in abundance in each of the comparisons.

A three-factorial ANOVA showed benzoic acid, *p*-hydroxybenzoic acid, and ferulic acid to be significantly affected by Mn (Table 2). Silicon treatment significantly affected *p*-hydroxybenzoic acid and *cis*-ferulic acid. Highly significant differences between the apoplastic fractions were found for all identified phenylpropanoids except ferulic acid and benzoic acid. Also, the infiltration solution had a clear impact on *p*-hydroxybenzoic acid, *p*-coumaric acid, and *trans*-sinapic acid. None of the two or three way interactions were significant (not presented).

Discussion

Effect of Mn and Si on apoplastic Mn fractions

Manganese is readily taken up by plants independent of the Si supply, but the expression of toxicity symptoms was suppressed by Si treatment (Fig. 1A, B) which is in line with results previously published for cowpea (Horst *et al.*, 1999; Iwasaki *et al.*, 2002a, b). This Si-enhanced Mn tolerance has been explained entirely in cucumber (Rogalla and Römheld, 2002) or partly in cowpea (Iwasaki *et al.*, 2002a, b) by a reduction of the free Mn in the apoplast through enhanced strong binding of Mn by the cell walls in Si-treated plants. However, in the present study neither the AWF_{H_2O} (Fig. 2A)

nor the 5-fold higher AWF_{NaCl} (Fig. 2C) Mn concentrations differed clearly owing to Si treatment. This might be explained by different growing conditions of the plants and Mn extraction procedures. Nevertheless, this clearly shows that, in cowpea, the expression of Mn toxicity cannot be explained just on the basis of the free and exchangeable Mn concentration in the leaf apoplast, in agreement with the conclusion drawn by Iwasaki *et al.* (2002a, b). They postulated a particular role of the monomeric Si in enhancing Mn tolerance. Indeed, also in our study the monomeric Si concentration was consistently higher in Si-treated plants in the AWF_{H_2O} (Fig. 2B) and initially also in the AWF_{NaCl} (Fig. 2D) fraction. The decreasing concentration of monomeric Si with increasing Mn treatment duration in the latter fraction possibly due to polymerization and/or strong binding in the cell walls (incrustation) may explain why Si treatment did not prevent but only delayed the formation of brown spots (Fig. 1B) with extended Mn treatment duration.

Manganese and Si-induced changes of peroxidase activities

All isoenzymes were shown to perform both reaction cycles (Figs 3, 4). Mn treatment led to an increased abundance of POD isoenzymes (Fig. 3; see Supplementary Fig. S1 at JXB online; Fecht-Christoffers *et al.*, 2003b) thus explaining enhanced apoplastic POD activities (Fecht-Christoffers *et al.*, 2006). Silicon treatment only delayed but not suppressed the Mn-mediated increased abundance of POD isoenzymes (see Supplementary Fig. S1 at JXB online), which is in line with the delayed but not prevented development of Mn toxicity symptoms (Fig. 1B). Using higher protein loadings BN-PAGE separation of AWF_{H_2O} and AWF_{NaCl} protein did not reveal qualitative but only quantitative differences in POD isoenzyme patterning between the infiltration solutions, indicating that all detected isoforms are principally water-soluble (see

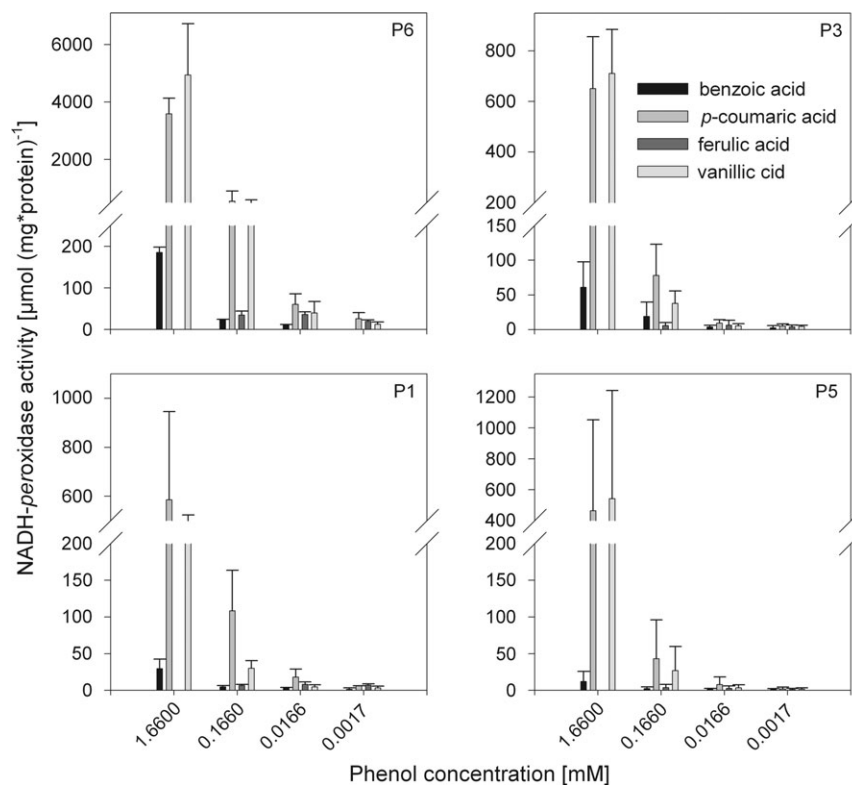


Fig. 6. Effect of different phenols on NADH-peroxidase activity of four POD isoenzymes. POD isoenzymes were eluted from BN-gels (see Fig. 4 and Materials and methods). Measuring solution (0.1 M succinate buffer, pH 5.5) consisted of 16 mM MnCl_2 , 0.22 mM NADH, and phenols (benzoic acid, *p*-coumaric acid, ferulic acid, caffeic acid, chlorogenic acid, gallic acid, protocatechuic acid, syringic acid, and vanillic acid) in different concentrations (1.6 mM, 0.166 mM, 0.016 mM, and 0.0016 mM). Only the four displayed phenols induced NADH-peroxidase activity. For the calculation of enzyme activities, extinction coefficients were adapted (see Supplementary Table S1 at *JXB* online). Results are from two independent experiments including plant growth and protein separation.

Supplementary Fig. S2 at *JXB* online), even though a low protein loading could lead to the opposite conclusion (Fig. 3; see Supplementary Fig. S1 at *JXB* online). The results confirm a particular role of PODs in the $\text{AWF}_{\text{H}_2\text{O}}$ in the modulation of Mn toxicity (Fecht-Christoffers *et al.*, 2006, 2007).

Characterization of the identified peroxidases

The sequencing of the POD activity-showing 1D-BN protein bands P1 to P6 revealed that each band was composed of more than one protein (see Supplementary Table S2 at *JXB* online) confirming BN/SDS-PAGE results previously published by Fecht-Christoffers *et al.* (2003b). All bands led to the identification of at least one peptide with high sequence homology to peroxidases in the NCBI green plants database. In total, 11 different peptides have been identified belonging to the class III secretory peroxidase family including sequences for the conserved so-called 'domain II' (Hiraga *et al.*, 2001)/'domain D' (Delannoy *et al.*, 2003) (Fig. 5, see Supplementary Table S2 at *JXB* online). Three overlapping peptide sequences provide evidence for the presence of at least three distinct genes encoding for class III secretory peroxidases (Fig. 5). Three peptides with amino acid substitutions (including the overlapping peptide

sequences: Fig. 5) were exclusively found in AWF_{NaCl} -extracted isoenzyme P1 from Mn-treated plants (Figs 3, 5; see Supplementary Table S2 at *JXB* online) indicating specific apoplastic binding properties.

As MS analyses did not result in complete POD sequences, one can only speculate about the total number of distinct class III secretory peroxidases in *Vigna unguiculata*. Based on *in gel* activity stainings, peroxidases of a wide range of MW were detected (Fig. 3; see Supplementary Figs S1 and S2 at *JXB* online). There are several possibilities leading to such great differences in the MW of the isoenzymes. (i) Class III peroxidases belong to a large multigenic family even though they are distinct proteins (Passardi *et al.*, 2004) with MWs ranging 28 kDa up to 60 kDa (Hiraga *et al.*, 2001). (ii) A protein oligomer showing peroxidase activity is conceivable, such as a peroxidase dimer. (iii) Depending on the degree of *N*-glycosylation, the native MW may vary thus leading to changes in the MW in the order of P3–P6 (Fig. 3). (iv) Other apoplastic proteins than class III peroxidases might also have peroxidative activity, i.e. oxidoreductase and/or auxin-binding (germin-like) proteins, even though the sequencing results did not identify proteins that could perform a peroxidative reaction (see Supplementary Table S2 at *JXB* online).

The role of pH in controlling apoplastic POD isoenzyme activities

A pH optimum seems to be necessary for POD self-protection (Olsen *et al.*, 2003). In addition, the pH could be an important regulatory factor for the relative performance of either the peroxidative or the peroxidative–oxidative reaction cycle of the enzyme. If an apoplastic pH of about 5.0–6.0 as shown for *Vicia faba* (Mühling and Läubli, 2000) is assumed, both POD cycles are expected to have high activities within this range (Fig. 5), indicating that the apoplastic pH is not decisive in regulating the relative contribution of each reaction cycle in response to toxic Mn supply. The determined pH optimum for both POD activities is precisely in the range of the recommended pH of the measuring solutions for POD activity determination *in vitro* in studies investigating lignin formation (Kärkönen *et al.*, 2002). However, in studies on the hypersensitive stress response to leaf pathogens, NADH-peroxidase-mediated H₂O₂ production proved to be related to an alkalization of the apoplast (Bolwell *et al.*, 1995, 1998, 2001; Pignocchi and

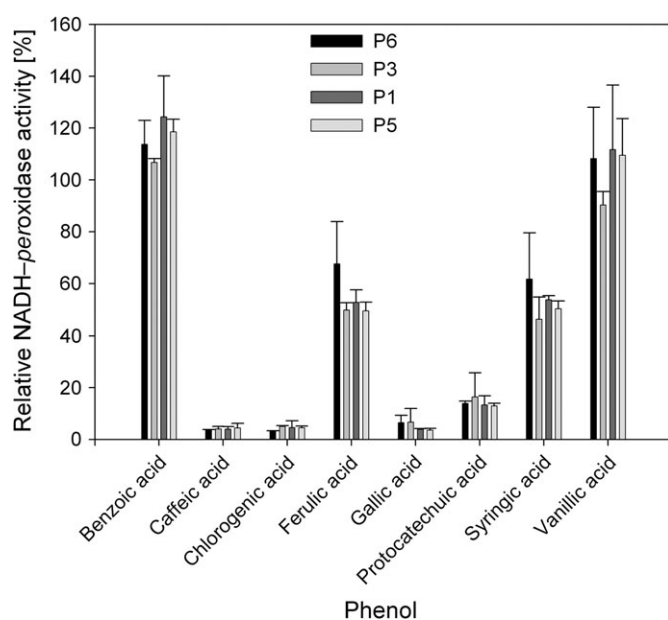


Fig. 7. Effect of combining different phenols with *p*-coumaric acid as the control phenol on the induction capability for NADH-peroxidase activity of four POD isoenzymes. POD isoenzymes were eluted from BN gels (see Fig. 3 and Materials and methods). Measuring solution (0.1 M succinate buffer, pH 5.5) consisted of 0.166 mM *p*-coumaric acid, 16 mM MnCl₂, 0.22 mM NADH, and 0.0166 mM of one of the following phenols to examine interactions between phenols: benzoic acid, ferulic acid, caffeic acid, chlorogenic acid, gallic acid, protocatechuic acid, syringic acid, or vanillic acid. Activities are expressed as relative values in relation to activities when *p*-coumaric acid was applied alone (in the same concentration). For the calculation of enzyme activities, extinction coefficients were adapted (see Supplementary Table S1 at *JXB* online). Results are from two independent experiments including plant growth and protein separation.

Foyer, 2003) suggesting differences between biotic and abiotic stress responses.

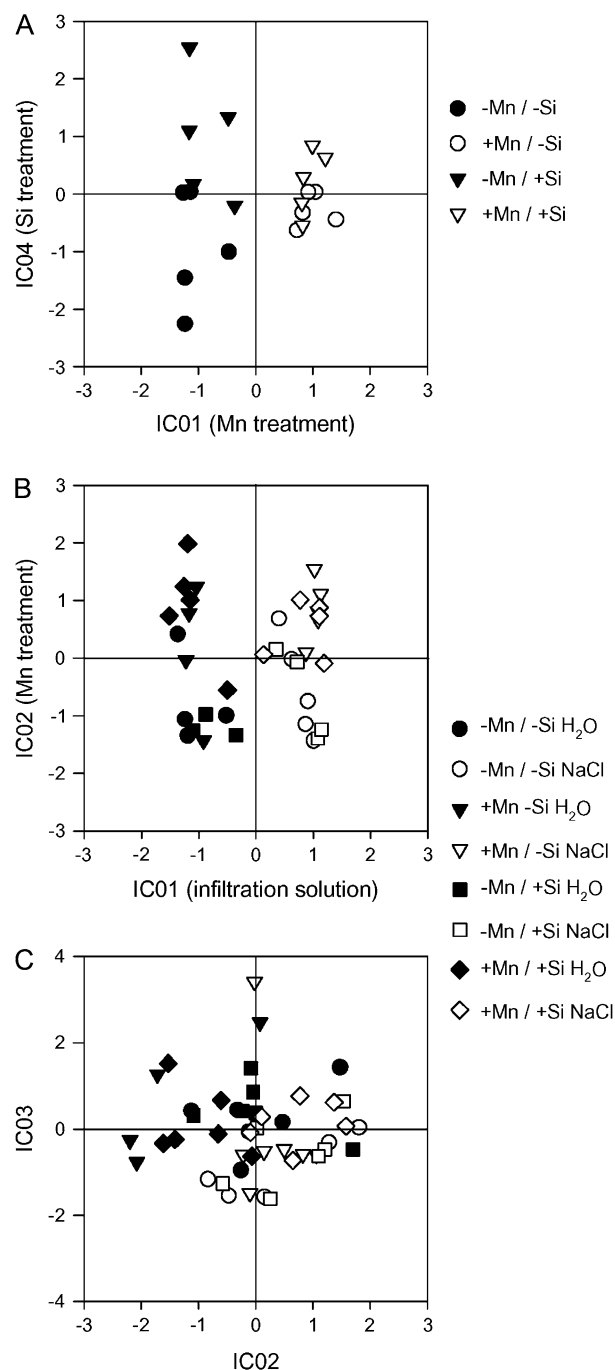


Fig. 8. ICA plot of the GC-MS-accessible (A) bulk-leaf metabolome, (B) the polar AWF metabolites, and (C) the non-polar metabolites extracted from the AWF. The second oldest trifoliolate leaf of the Mn-sensitive cultivar TVu 91 was tested for Mn and Si effects. After 14 d of preculture with or without Si, plants received 50 μM Mn (+Mn) for 3 d or 0.2 μM Mn (–Mn) continuously. Bulk-leaf, AWF- and non-polar apoplastic metabolites were extracted (*n*=5 and 6, respectively) as described in the Materials and methods. ICA was conducted using MetaGeneAlyse at <http://metagenealyse.mpimp-golm.mpg.de>.

Table 1. Identified phenols (GC-MS) in the non-polar fraction of the leaf AWF recovered after infiltration with H₂O or NaCl

Displayed are the relative pool-size changes of each phenol calculated on the basis of response ratios. The effects of these phenols on the NADH-*peroxidase* activity (see Figs 6 and 7) of apoplastic peroxidase isoenzymes are also shown. After 14 d of preculture, \pm Si-treated plants of the Mn-sensitive cowpea cv. TVu91 received 50 μ M Mn for 3 d or 0.2 μ M Mn continuously. Statistical testing of changes in metabolite abundance were calculated using log₁₀-transformed response ratios. An asterisk denotes significant differences at least at $P < 0.05$ ($n=6$), respectively (t test).

Detected metabolites	+Mn/-Mn		+Si/-Si		+Mn+Si/-Mn+Si		+Mn+Si/+Mn-Si		Effect of phenol on peroxidase activity ^a
	AWF _{H₂O}	AWF _{NaCl}	AWF _{H₂O}	AWF _{NaCl}	AWF	AWF _{NaCl}	AWF _{H₂O}	AWF _{NaCl}	
Benzoic acid	1.41 ^c	1.35	0.91	1.32*	1.49*	1.14	0.97	1.12	weak induction/no inhibition
<i>p</i> -Hydroxybenzoic acid ^b	1.47	1.61	0.87	0.65	1.03	1.23	0.61*	0.50*	No induction/50% inhibition
<i>cis-p</i> -Coumaric acid	1.00	0.81	0.95	1.13	1.06	0.62	1.00	0.86	Strong induction
<i>cis</i> -Ferulic acid	0.24*	0.30*	1.03	3.77*	0.70	0.35*	2.96	4.37	Weak induction/50% inhibition
<i>trans-p</i> -Coumaric acid	0.88	0.82	0.84	0.95	0.97	0.59	0.92	0.68	Strong induction
<i>trans</i> -Ferulic acid	0.44*	2.31	1.34	3.61*	0.50	0.37*	1.52	0.57	Weak induction/50% inhibition
<i>trans</i> -Sinapic acid	2.39	0.81	n.d. ⁺	0.72	n.d. ⁺⁺	0.90	n.d. ⁺⁺	0.80	Not examined

^a from Figs. 6 and 7.

^b After identification of *p*-hydroxybenzoic acid, this phenol was additionally tested with respect to NADH-*peroxidase* activity. In addition to the 50% inhibitory effect it showed no induction capability for NADH-*peroxidase* activity for each isoenzyme tested.

^c Numbers are calculated ratios of the response ratios (not log₁₀-transformed) within the individual comparison. ANOVA did not reveal a significant Mn \times Si interaction.

⁺, ⁺⁺ were not detected (n.d.) in +Si and +Mn+Si treatments, respectively.

Table 2. Identified phenols (GC-MS) in the non-polar fraction of the leaf AWF recovered after infiltration with H₂O or NaCl (Inf.).

Displayed are the *p*-values derived from analysis of variance based on log₁₀-transformed response ratios ($n=6$). For the effects of these phenols on the NADH-*peroxidase* activity of apoplastic peroxidase isoenzymes see Figs 6 and 7 as well as Table 1. After 14 d of preculture, \pm Si-treated plants of the Mn-sensitive cowpea cultivar TVu 91 received 50 μ M Mn for 3 d or 0.2 μ M Mn continuously.

Metabolite	Mn	Si	Inf.
Benzoic acid	0.0032	0.2786	0.2615
4-Hydroxybenzoic acid ^a	0.0093	<0.0001	<0.0001
<i>cis</i> -4-Hydroxycinnamic acid	0.4236	0.5636	<0.0001
<i>cis</i> -Ferulic acid	<0.0001	0.0012	0.4433
<i>trans</i> -4-Hydroxycinnamic acid	0.1269	0.1057	<0.0001
<i>trans</i> -Ferulic acid	0.0129	0.2470	0.3870
<i>trans</i> -Sinapic acid	0.4671	0.1685	0.0039

^a After identification of *p*-hydroxybenzoic acid, this phenol was additionally tested with respect to NADH-*peroxidase* activity. In addition to the 50% inhibitory effect, it showed no induction capability for NADH-*peroxidase* activity for each isoenzyme tested.

The role of metabolites in controlling apoplastic POD isoenzyme activities: metabolite profiling

In a broad range metabolomic approach, it has been shown that Mn toxicity induced changes in the bulk-leaf metabolome according to ICA (Fig. 8A; IC01) consistent with our recent results showing that Mn toxicity also affects symplastic reactions using a combined proteomic/transcriptomic and physiological approach (Führs *et al.*, 2008). The involvement of the symplast in Mn toxicity is in line with studies using other plant species showing Mn toxicity-induced reduced CO₂ assimilation capacity (González and Lynch, 1997, 1999; González *et al.*, 1998, common bean;

Nable *et al.*, 1988; Houtz *et al.*, 1988, tobacco) accompanied by reduced chlorophyll contents (Gonzalez and Lynch, 1999; Gonzalez *et al.*, 1998, common bean; Moroni *et al.*, 1991, wheat), and high Mn-accumulation rates in chloroplasts (Lidon *et al.*, 2004, rice). Our metabolomic approach also showed that Si supply led to a particular clustering of the total leaf metabolome as revealed by ICA (Fig. 8A; IC04). This is in agreement with the work of Maksimović *et al.* (2007) on Si/Mn-toxicity interaction in cucumber who concluded that Si supply modulates the phenol metabolism.

A closer investigation of the apoplastic metabolome using AWF_{NaCl} and AWF_{H₂O} revealed that the infiltration solution (IC01; Fig. 8B) was the most important factor explaining differences between the extracted metabolome fractions. Manganese (IC02) but not Si treatment affected both AWF metabolome fractions. The ICA loadings identified organic acids, amino acids, and sugars to be responsible for Mn and infiltration solution-related clusterings (Fig. 8B), whereas phenolic compounds were unexpectedly low since Fecht-Christoffers *et al.* (2006, 2007) reported a Mn-induced change in the apoplastic water-soluble phenol composition (and at later toxicity stages even in phenol concentration) using HPLC separation of leaf AWF_{H₂O} in cowpea (see discussion below). However, GC-MS based metabolite profiling typically covers mostly primary metabolites explaining the relative low abundance of phenolic compounds.

To overcome this problem, an additional special AWF-extraction procedure was applied yielding non-polar metabolites. This resulted in clustering only according to the infiltration solution (Fig. 8C; IC02, see discussion below). ICA loadings revealed, in addition to organic acids, that phenylpropanoids were mainly responsible for the clustering. Among other detected aromatic compounds, ferulic acid was identified as a clearly Mn and Si-affected phenol (Tables 1, 2; see discussion below).

Overall, the broad-range metabolite profiling in the bulk-leaf extract (Fig. 8A, ICA01) and the AWF (Fig. 8B; ICA02) revealed a clear difference related to the Mn treatment. The Si effect was less clearly expressed. A preliminary metabolite-specific evaluation of the metabolites indicates alterations of metabolic pathways mainly related to organic acids, amino acids, and sugars/sugar alcohols. A detailed evaluation and discussion of the qualitative changes in polar apoplastic metabolites is beyond the scope of this paper and will be subject of a subsequent paper.

The role of phenols in controlling apoplastic NADH-peroxidase activity

Analysing the AWF_{H₂O} using HPLC, Fecht-Christoffers *et al.* (2006) separated water-soluble phenols in the apoplast. A Mn treatment not only increased the peak size but also led to at least two additional peaks, which supported their conclusion that the presence of phenols in the apoplast is decisive for the expression of Mn toxicity/Mn tolerance in cowpea leaf tissue. However, they failed to identify the phenols. Our gas chromatography–mass spectrometry approach allowed us to identify five phenols. However, the method does not allow absolute concentrations to be determined but only relative treatment-related concentration changes. Also, it was not possible to identify most phenols directly in the AWF. Therefore, the aqueous AWF was extracted with diethylether which led to a concentration of the phenols but at the same time only yielded non-polar metabolites. Thus, the applied technique did not allow us to identify and quantify all the phenols present in the apoplast which is a major focus of ongoing research. Nevertheless, among the phenols identified (Tables 1, 2), four were found which had been tested for their effect on NADH-peroxidase activity *in vitro*. Only *p*-coumaric acid had a strong activity-enhancing effect. Ferulic acid and *p*-hydroxybenzoic acid had only a weak or lacking stimulating effect, but a strong inhibiting effect when combined with *p*-coumaric acid. Benzoic acid only weakly enhanced and did not inhibit NADH-peroxidase activity (Figs 6, 7; Table 1).

The three-factorial analysis of variance of the treatment-induced changes in the abundance of the phenols (Table 2) revealed that Mn treatment significantly affected the concentrations of benzoic, *p*-hydroxybenzoic and, most clearly, ferulic acid, whereas Si treatment affected *p*-hydroxybenzoic and again most clearly *cis*-ferulic acid. Looking at the comparison of means of the treatment-specific relative pool-size changes of the individual phenols (Table 1), it appears that the change in the concentration in the apoplast of ferulic acid particularly plays a key role in the expression of Mn toxicity symptoms: a reduction of the concentration leading to a reduced inhibition of NADH-peroxidase activity is characteristic for leaves showing Mn toxicity symptoms (+Mn/–Si), while Mn-tolerant leaf tissue (–Mn/+Si; +Mn/+Si) is characterized by an enhanced accumulation. The constitutive effect of Si on an enhanced abundance of ferulic acid seems to be strong enough to counteract the Mn-induced

reducing effect (compare +Mn +Si/–Mn +Si, Table 1). Also, it appears that Si affects the phenol concentration more in the AWF_{NaCl} (as indicated by the high infiltration solution effect on the phenols in Table 2) than in the AWF_{H₂O} corroborating results demonstrating Si-mediated changes of apoplastic Mn-binding properties (Iwasaki *et al.*, 2002a; Rogalla and Römheld, 2002). However, ferulic acid and benzoic acid, in particular, were not affected by the infiltration solution, indicating specific apoplastic binding properties in the apoplast for each phenol regardless of Si nutrition (Table 1). The Si-induced significantly higher abundance of benzoic acid might be of minor importance, given the rather weak NADH-peroxidase activity-enhancing effect (Fig. 8). However, the lowered concentration of NADH-peroxidase activity-inhibiting *p*-hydroxybenzoic acid in the presence of Si at high Mn supply is not in line with the above expressed line of thinking. Thus it appears a more detailed and quantitative investigation of the phenols present in the leaf apoplast is necessary to understand Mn toxicity and Mn tolerance fully.

In conclusion, the results presented here confirm the hypothesized role of apoplastic NADH-peroxidase and its activity-modulating phenols in Mn toxicity and Si-enhanced Mn tolerance. Isoenzyme BN gel-profiling of POD enzymes and their characterization after elution from the gels, and metabolite profiling of the bulk-leaf and the AWF appear to be powerful tools in enhancing the physiological and molecular understanding of Mn toxicity and Mn tolerance.

Supplementary data

Supplementary data can be found at *JXB* online.

Supplementary Fig. S1. 1D BN-PAGE resolution of AWF_{H₂O} and AWF_{NaCl} proteins (16 µg) after 0 and 4 d of Mn treatment of ±Si-treated plants of the Mn-sensitive cowpea cultivar TVu 91.

Supplementary Fig. S2. 1D BN-PAGE resolution of AWF_{H₂O} and AWF_{NaCl} proteins (180 µg) after 0 d and 4 d of Mn treatment of the Mn-sensitive cowpea cultivar TVu 91.

Supplementary Table S1. Extinction coefficients for the calculation of NADH-peroxidase activities of different POD isoenzymes supplied with different phenols in changing concentrations as shown in Figs 4, 6, and 7.

Supplementary Table S2. Peptide sequences of apoplastic leaf proteins sequenced with LC-MS/MS.

Acknowledgements

This work was supported by the Deutsche Forschungsgemeinschaft (grants HO 931-17, HO 931-18/1).

References

Bolwell GP, Butt VS, Davies DR, Zimmerlin A. 1995. The origin of the oxidative burst in plants. *Free Radical Research* **23**, 517–532.

- Bolwell GP, Davies DR, Gerrish C, Auh C-K, Murphy TM.** 1998. Comparative biochemistry of the Oxidative Burst produced by rose and French Bean cells reveals two distinct mechanisms. *Plant Physiology* **116**, 1379–1385.
- Bolwell GP, Page A, Piślewska M, Wojtaszek P.** 2001. Pathogenic infection and the oxidative defences in plant apoplast. *Protoplasma* **217**, 20–32.
- Bradford MM.** 1976. A rapid and sensitive method for the quantitation of microgram quantities of protein utilizing the principle of protein-dye binding. *Analytical Biochemistry* **72**, 248–254.
- Delannoy E, Jalloul A, Assigbetsé K, Marmey P, Geiger JP, Lherminier J, Daniel JF, Martinez C, Nicole M.** 2003. Activity of class III peroxidases in the defense of cotton to bacterial blight. *Molecular Plant-Microbe Interactions* **16**, 1030–1038.
- Elias JE, Gygi SP.** 2007. Target-decoy search strategy for increased confidence in large-scale protein identifications by mass spectrometry. *Nature Methods* **4**, 207–214.
- Epstein E.** 1999. Silicon. *Annual Review of Plant Physiology and Plant Molecular Biology* **50**, 641–664.
- Fecht-Christoffers MM, Braun H-P, Lemaitre-Guillier C, VanDorsselear A, Horst WJ.** 2003b. Effect of Mn toxicity on the proteome of the leaf apoplast in cowpea. *Plant Physiology* **133**, 1935–1946.
- Fecht-Christoffers MM, Führs H, Braun H-P, Horst WJ.** 2006. The role of hydrogen peroxide-producing and hydrogen peroxide-consuming peroxidases in the leaf apoplast of cowpea in manganese tolerance. *Plant Physiology* **140**, 1451–1463.
- Fecht-Christoffers MM, Maier P, Horst WJ.** 2003a. Apoplastic peroxidase and ascorbate are involved in manganese toxicity and tolerance of *Vigna unguiculata*. *Physiologia Plantarum* **117**, 237–244.
- Fecht-Christoffers MM, Maier P, Iwasaki K, Braun H-P, Horst WJ.** 2007. The role of the leaf apoplast in manganese toxicity and tolerance in cowpea (*Vigna unguiculata* L. Walp). In: Sattelmacher B, Horst WJ, eds. *The apoplast of higher plants: compartment of storage, transport, and reactions*. Dordrecht, The Netherlands: Springer, 307–322.
- Fiehn O, Kopka J, Trethewey RN, Willmitzer L.** 2000. Identification of uncommon plant metabolites based on calculation of elemental compositions using gas chromatography and quadrupole mass spectrometry. *Analytical Chemistry* **72**, 3573–3580.
- Führs H, Hartwig M, Molina LEB, Heintz D, Van Dorsselear A, Braun H-P, Horst WJ.** 2008. Early manganese-toxicity response in *Vigna unguiculata* L.: a proteomic and transcriptomic study. *Proteomics* **8**, 149–159.
- González A, Lynch JP.** 1997. Effects of manganese toxicity on leaf CO₂ assimilation of contrasting common bean genotypes. *Physiologia Plantarum* **101**, 872–880.
- González A, Lynch JP.** 1999. Subcellular and tissue Mn compartmentation in bean leaves under Mn toxicity stress. *Australian Journal of Plant Physiology* **26**, 811–822.
- González A, Steffen KL, Lynch JP.** 1998. Light and excess manganese: Implications for oxidative stress in common bean. *Plant Physiology* **118**, 493–504.
- Halliwell B.** 1978. Lignin synthesis: the generation of hydrogen peroxide and superoxide by horseradish peroxidase and its stimulation by manganese (II) and phenols. *Planta* **140**, 81–88.
- Hiraga S, Sasaki K, Ito H, Ohashi Y, Matsui H.** 2001. A large family of class III plant peroxidases. *Plant and Cell Physiology* **42**, 462–468.
- Horst WJ.** 1980. Genotypische Unterschiede in der Mangan-Toleranz von Cowpea (*Vigna unguiculata*). *Angewandte Botanik* **54**, 377–392.
- Horst WJ.** 1982. Quick screening of cowpea genotypes for manganese tolerance during vegetative and reproductive growth. *Zeitschrift für Pflanzenernährung und Bodenkunde* **145**, 423–425.
- Horst WJ.** 1983. Factors responsible for genotypic manganese tolerance in cowpea (*Vigna unguiculata*). *Plant and Soil* **72**, 213–218.
- Horst WJ.** 1988. The physiology of Mn toxicity. In: Webb MJ, Nable RO, Graham RD, Hannam RJ, eds. *Manganese in soil and plants*. Dordrecht/Boston/London: Kluwer Academic Publishers, 175–188.
- Horst WJ, Fecht M, Naumann A, Wissemeyer AH, Maier P.** 1999. Physiology of manganese toxicity and tolerance in *Vigna unguiculata* (L.) Walp. *Journal of Plant Nutrition and Soil Science* **162**, 263–274.
- Horst WJ, Marschner H.** 1978a. Effect of silicon on manganese tolerance of bean plants (*Phaseolus vulgaris* L.). *Plant and Soil* **50**, 287–303.
- Horst WJ, Marschner H.** 1978b. Symptome von Manganüberschuss bei Bohnen (*Phaseolus vulgaris*). *Zeitschrift für Pflanzenernährung und Bodenkunde* **141**, 129–142.
- Houtz RL, Nable RO, Cheniae GM.** 1988. Evidence for effects on the *in vivo* activity of ribulose-bisphosphate carboxylase/oxygenase during development of Mn toxicity in tobacco. *Plant Physiology* **86**, 1143–1149.
- Iwasaki K, Maier P, Fecht M, Horst WJ.** 2002a. Effects of silicon supply on apoplastic manganese concentrations in leaves and their relation to manganese tolerance in cowpea (*Vigna unguiculata* (L.) Walp.). *Plant and Soil* **238**, 281–288.
- Iwasaki K, Maier P, Fecht M, Horst WJ.** 2002b. Leaf apoplastic silicon enhances manganese tolerance of cowpea (*Vigna unguiculata*). *Journal of Plant Physiology* **159**, 167–173.
- Iwasaki K, Matsumura A.** 1999. Effect of silicon on alleviation of manganese toxicity in pumpkin (*Cucurbita moschata* Duch cv. Shintosa). *Soil Science and Plant Nutrition* **45**, 909–920.
- Jänsch L, Kruff V, Schmitz UK, Braun HP.** 1996. New insights into the composition, molecular mass and stoichiometry of the protein complexes of plant mitochondria. *The Plant Journal* **9**, 357–368.
- Kärkönen A, Koutaniemi S, Mustonen M, Syrjänen K, Brunow G, Kilpeläinen I, Teeri TH, Simola LK.** 2002. Lignification related enzymes in *Picea abies* suspension cultures. *Physiologia Plantarum* **114**, 343–353.
- Kopka J, Schauer N, Krueger S, et al.** 2005. GMD@CSB.DB: the Golm metabolome database. *Bioinformatics* **21**, 1635–1638.
- Lidon FC, Barreiro MG, Ramalho JC.** 2004. Manganese accumulation in rice: implications for photosynthetic functioning. *Journal of Plant Physiology* **161**, 1235–1244.
- Luedemann A, Strassburg K, Erban A, Kopka J.** 2008. TagFinder for the quantitative analysis of gas chromatography-mass spectrometry

(GC-MS) based metabolite profiling experiments. *Bioinformatics* **24**, 732–737.

Maksimović JD, Bogdanović J, Maksimović V, Nikolic M. 2007. Silicon modulates the metabolism and utilization of phenolic compounds in cucumber (*Cucumis sativus* L.) grown at excess manganese. *Journal of Plant Nutrition and Soil Science* **170**, 739–744.

Marschner H. 1995. *Mineral nutrition in higher plants*, 2nd edn. London, UK: Academic Press.

Moroni JS, Briggs KG, Taylor GJ. 1991. Chlorophyll content and leaf elongation rate in wheat seedlings as a measure of manganese tolerance. *Plant and Soil* **136**, 1–9.

Mühling KH, Läuchli A. 2000. Light-induced pH and K⁺ changes in the apoplast of intact leaves. *Planta* **212**, 9–15.

Nable RO, Houtz RL, Cheniae GM. 1988. Early inhibition of photosynthesis during development of Mn toxicity in tobacco. *Plant Physiology* **86**, 1136–1142.

Olsen LF, Hauser MJB, Kummer U. 2003. Mechanism of protection of peroxidase activity by oscillatory dynamics. *European Journal of Biochemistry* **270**, 2796–2804.

Passardi F, Cosio C, Penel C, Dunand C. 2005. Peroxidases have more functions than a Swiss army knife. *Plant Cell Reports* **24**, 255–265.

Passardi F, Longet D, Penel C, Dunand C. 2004. The class III peroxidase multigenic family in rice and its evolution in land plants. *Phytochemistry* **65**, 1879–1893.

Pignocchi C, Foyer CH. 2003. Apoplastic ascorbate metabolism and its role in the regulation of cell signalling. *Current Opinion in Plant Biology* **133**, 443–447.

Roessner U, Wagner C, Kopka J, Trethewey RN, Willmitzer L. 2000. Simultaneous analysis of metabolites in potato tuber by gas chromatography–mass spectrometry. *The Plant Journal* **23**, 131–142.

Rogalla H, Römheld V. 2002. Role of leaf apoplast in silicon-mediated manganese tolerance of *Cucumis sativus* L. *Plant, Cell and Environment* **25**, 549–555.

Schaarschmidt S, Kopka J, Ludwig-Müller J, Hause B. 2007. Regulation of arbuscular mycorrhization by apoplastic invertases: enhanced invertase activity in the leaf apoplast affects the symbiotic interaction. *The Plant Journal* **51**, 390–405.

Schauer N, Steinhäuser D, Strelkov S, et al. 2005. GC-MS libraries for the rapid identification of metabolites in complex biological samples. *FEBS Letters* **579**, 1332–1337.

Scholz M, Gatzek S, Sterling A, Fiehn O, Selbig J. 2004. Metabolite fingerprinting: detecting biological features by independent component analysis. *Bioinformatics* **20**, 2447–2454.

Scholz M, Kaplan F, Guy CL, Kopka J, Selbig J. 2005. Non-linear PCA: a missing data approach. *Bioinformatics* **21**, 3887–3895.

Shi Q, Bao Z, Zhu Z, He Y, Qian Q, Yu J. 2005. Silicon-mediated alleviation of Mn toxicity in *Cucumis sativus* in relation to activities of superoxide dismutase and ascorbate peroxidase. *Phytochemistry* **66**, 1551–1559.

Wagner C, Sefkow M, Kopka J. 2003. Construction and application of a mass spectral and retention time index database generated from plant GC/EI-TOF-MS metabolite profiles. *Phytochemistry* **62**, 887–900.

Wehrhahn W, Braun HP. 2002. Biochemical dissection of the mitochondrial proteome from *Arabidopsis thaliana* by three-dimensional gel electrophoresis. *Electrophoresis* **23**, 640–646.

Wissemeyer AH, Horst WJ. 1992. Effect of light intensity on manganese toxicity symptoms and callose formation in cowpea (*Vigna unguiculata* (L.) Walp.). *Plant and Soil* **143**, 299–309.

Yamazaki I, Piette LH. 1963. The mechanism of aerobic oxidase reaction catalysed by peroxidase. *Biochimica et Biophysica Acta* **77**, 47–64.

Proteomic characterization of the leaf apoplast of *Vigna unguiculata* L. in response to short-term toxic manganese supply

Hendrik Führs^{a*}, Mareike Vorholt^a, Sébastien Gallien^b, Dimitri Heintz^c, Alain Van Dorsselaer^b, Hans-Peter Braun^d & Walter J. Horst^a

^a Institute for Plant Nutrition, Faculty of Natural Sciences, Leibniz University Hannover, Herrenhäuser Str. 2, 30419 Hannover, Germany

^b Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC-DSA,

Université de Strasbourg, CNRS, UMR7178 ; 25 rue Becquerel, 67 087 Strasbourg, France

^c Institut de Biologie Moléculaire des Plantes (IBMP), 28 rue Goethe, CNRS-UPR2357, Université de Strasbourg, 67083 Strasbourg, France

^d Institute for Plant Genetics, Faculty of Natural Sciences, Leibniz University Hannover, Herrenhäuser Str. 2, 30419 Hannover, Germany

Abstract : Previous studies investigating the leaf apoplastic water-soluble proteome after longer term Mn treatment in cowpea (*Vigna unguiculata* L.) suggested that the leaf apoplast is the decisive leaf compartment for the development of Mn toxicity. To study short-term effects of excess Mn supply, apoplastic proteins were extracted from leaves of the Mn-sensitive cowpea cultivar TVu 91 and the apoplastic proteome characterized by means of IEF/SDS-PAGE and liquid chromatography-tandem mass spectrometry. Two Apoplastic-Washing-Fluid fractions (AWFH₂O and AWFNaCl) and a cell-wall fraction released from isolated cell walls were analysed. The purity of the generated fractions was tested by determining the activity of malate dehydrogenase as marker-enzyme for cytoplasmic contamination. The cell-wall isolation-procedure proved to be inappropriate for the investigation of strongly bound cell-wall proteins owing to a symplastic contamination of up to 4% and the identification of mainly typical symplastic proteins. The AWF extraction procedures yielded low contaminations (<0.5 %) and only few peptides assigned to typical symplastic proteins. One day of excess Mn allowed the identification of two AWFH₂O and three AWFNaCl extracted spots of changed abundance. The identification of a Mn-induced basic peroxidase (POD) isoenzyme in the AWFNaCl fraction in addition to acidic POD isoenzymes in the AWFH₂O further supports the proposed decisive role of H₂O₂-producing and consuming PODs for the development of Mn toxicity. Identification of further proteins significantly affected by Mn treatment (polygalacturonase-inhibiting proteins and α -galactosidases) suggests Mn excess-induced modification of cell-wall development and functions, whereas detection of others (acetylcholinesterase, GDSL-lipase 1, aspartyl protease) indicates changes in broad-sense signal transduction processes.

Comparative proteomic and physiological characterization of Mn sensitivity and Mn tolerance in barley (*Hordeum vulgare* L.) and rice (*Oryza sativa* L.)

Hendrik Führs^{1†}, Christof Behrens^{4†}, Sébastien Gallien², Dimitri Heintz³, Alain Van Dorsselaer², Hans-Peter Braun⁴ & Walter J. Horst^{1*}

¹ Institute for Plant Nutrition, Faculty of Natural Sciences, Leibniz University Hannover, Herrenhäuser Str. 2, 30419 Hannover, Germany

² Laboratoire de Spectrométrie de Masse Bio-Organique, IPHC-DSA, Université de Strasbourg, CNRS, UMR7178 ; 25 rue Becquerel, 67 087 Strasbourg, France

³ Institut de Biologie Moléculaire des Plantes (IBMP), 28 rue Goethe, CNRS-UPR2357, Université de Strasbourg, 67083 Strasbourg, France

⁴ Institute for Plant Genetics, Faculty of Natural Sciences, Leibniz University Hannover, Herrenhäuser Str. 2, 30419 Hannover, Germany

Running title: Mn sensitivity and tolerance in barley and rice

* Corresponding author:

Email: horst@pflern.uni-hannover.de

†H.F. and C.B. contributed equally to this work.

Abstract

Background and Aims: Research on manganese (Mn) toxicity and tolerance indicates that Mn toxicity develops apoplastically through increased peroxidase activities mediated by phenolics and Mn, and Mn tolerance could be conferred by sequestration of Mn in inert cell compartments. This comparative study focusses on Mn-sensitive barley (*Hordeum vulgare*) and Mn-tolerant rice (*Oryza sativa*) as model organisms to unravel mechanisms of Mn toxicity and/or tolerance in monocots. **Methods:** Bulk-leaf Mn concentrations as well as peroxidase activities and protein concentrations were analysed in apoplastic washing fluid (AWF) in both species. In rice, Mn distribution between leaf compartments and the leaf proteome using 2D IEF/SDS-PAGE and 2D BN/SDS-PAGE were studied.

Key Results: The Mn-sensitivity of barley was confirmed since the formation of brown spots on older leaves was induced by low bulk-leaf and AWF Mn concentrations and strongly enhanced H₂O₂- producing and consuming peroxidase activities. In contrast, by a factor 50 higher Mn concentrations did not produce Mn toxicity symptoms on older leaves in rice. Peroxidase activities lower by a factor of about 100 in the rice-leaf AWF compared to barley support the view of a central role of these peroxidases in the apoplastic expression of Mn toxicity. The high Mn

tolerance of old rice leaves could be related to a high Mn binding capacity of the cell walls. The proteomic studies suggest that the lower Mn tolerance of young rice leaves could be related to Mn excess-induced displacement of Mg and Fe from essential metabolic functions.

Conclusions: The results provide evidence that Mn toxicity in barley involves apoplastic lesions mediated by peroxidases. The high Mn tolerance of old leaves of rice involves a high Mn-binding capacity of the cell walls whereas Mn toxicity in less Mn-tolerant young leaves is related to Mn-induced Mg and Fe deficiencies.

Key words: apoplast, compartmentation, *Hordeum vulgare* cv. Baroness, Mn sensitivity, Mn tolerance, *Oryza sativa* var. *japonica* cv. Guara, proteome, photosynthesis

Adaptative response and persistent metabolic activity of *Euglena* sp. in acid mine drainage microbial community

Florence Goulhen-Chollet^{1†}, David Halter^{1†}, Odile Bruneel², Sébastien Gallien³, Bertrand Chaumande^{1,§}, Corinne Casiot², Guillaume Morin⁴, Gordon E. Brown Jr.^{5,6}, Amélie Bardil², Denis Le Paslier^{7,8}, Christine Schaeffer³, Alain Van Dorsselaer³, Françoise Elbaz-Poulichet², Florence Arsène-Ploetze¹, Philippe N. Bertin^{1*}

† These authors contributed equally to this work

¹ UMR7156 Université de Strasbourg/CNRS, Génétique Moléculaire, Génomique Microbiologie, Département Micro-organismes, Génomes, Environnement, 28 rue Goethe, 67083 Strasbourg Cedex, France.

² Laboratoire Hydrosociences Montpellier, UMR 5569 (CNRS - IRD - Universités Montpellier I et II), Université Montpellier II, CC MSE, Place Eugène Bataillon, 34095 Montpellier Cedex 05, France.

³ Laboratoire de Spectrométrie de Masse Bio-organique, Institut Pluridisciplinaire Hubert Curien, UMR7178 CNRS-Université de Strasbourg, 25 rue Becquerel, 67087 Strasbourg, France.

⁴ IMPMC, Institut de Minéralogie et de Physique des Milieux Condensés (IMPMC), UMR 7590 – CNRS – Universités Paris 6 & 7 – IPGP, 140 rue de Lourmel, 75015 Paris, France.

⁵ Surface & Aqueous Geochemistry Group, Department of Geological and Environmental Sciences, Stanford University, Stanford, California 94305-2115, USA.

⁶ Stanford Synchrotron Radiation Laboratory, SLAC, 2575 Sand Hill Road, MS 69, Menlo Park, California 94025, USA.

⁷ CNRS UMR8030, Génomique Métabolique, 2 rue Gaston Crémieux, 91057 Evry Cedex, France.

⁸ Commissariat à l'Energie Atomique (CEA), Direction des Sciences du Vivant, Institut de Génomique, Genoscope, Laboratoire de Génomique Comparative, 2 rue Gaston Crémieux, 91057 Evry Cedex, France.

§ Present address: Institut Pluridisciplinaire Hubert Curien, Département Sciences Analytiques, ECPM, 25 rue Becquerel, 67087 Strasbourg Cedex 2, France.

* **Corresponding author.**

Mailing address: UMR7156 Université de Strasbourg/CNRS, Génétique Moléculaire, Génomique Microbiologie, Département Micro-organismes, Génomes, Environnement, 28 rue Goethe, 67083

Publication soumise à *Applied and Environmental Microbiology*,

Strasbourg Cedex, France. Phone: +33 3 90 24 20 08. Fax: +33 3 90 24 20 28. E-mail: philippe.bertin@unistra.fr 38

Running title. *Euglena* sp. adaptation and activity in Carnoulès AMD

Keywords. metaproteomics / environmental genomics / *Euglena* / microbial community

Journal section. Environmental Microbiology

Abstract

The Carnoulès mine drainage (France) is characterized by acid waters containing a high concentration of arsenic and iron. Major characteristics of the Reigous creek community were analyzed by combining chemistry, mineralogy, 16S rRNA gene libraries, and metaproteomics. Seventy-two proteins were identified, including several involved in energy or carbohydrate metabolism, and in CO₂ fixation and were shown to originate from bacteria, eukaryotic micro-organisms or plants. These results demonstrated that micro-organisms such as α -, β -, γ -Proteobacteria, Actinobacteria and Firmicutes are metabolically active. However, one of the most dominant groups found in the Carnoulès community in both water and sediments was *Euglena* sp., an acidophilic photosynthetic protist. Therefore, a second sediment sampling was performed leading to the identification of one hundred and seventy eight proteins that specifically originated from *Euglena*. Our observations support a role for this protist in oxygen production at water/sediment interface. In addition, our results suggest that several micro-organisms, including *Euglena* sp., expressed several adaptive responses to harsh conditions and may play a role in the transformation of the different substances present in the

Characterization of an arsenic contaminated environment, Sainte-Marie-aux-mines, France, revealed a large prokaryotic diversity with arsenic-specific adaptation capacities.

Audrey Heinrich-Salmeron^{1*}, Audrey Cordi^{2*}, Sébastien Gallien³, David Halter¹, Christophe Pagnout², Florence Goulhen-Chollet¹, Elham Abbaszadehfard², Alain Van Dorsselaer³, Christine Schaeffer³, Philippe Bertin¹, Pascale Bauda², Florence Arsène-Ploetze^{1,&}.

¹ Laboratoire de Génétique Moléculaire, Génomique, Microbiologie, Département Microorganismes, Génomes, Environnement, UMR7156 Université de Strasbourg/CNRS, 28 rue Goethe, 67083 Strasbourg Cedex, France.

² Laboratoire des Interactions Ecotoxicologie Biodiversité Ecosystèmes (LIEBE), UMR7146, CNRS, Université Paul Verlaine, Campus Bridoux, rue du Général Delestraint, 57070 Metz, France.

³ Laboratoire de Spectrométrie de Masse Bio-organique, Institut Pluridisciplinaire Hubert Curien, UMR7178 Université de Strasbourg/CNRS, 25 rue Becquerel, 67087 Strasbourg, France.

† * Equal contributors

† & Corresponding author : Florence.ploetze@gem.u-strabg.fr

ABSTRACT

The Sainte Marie mine site (France) is characterized by a mildly arsenic contamination and a neutral pH. An integrated study coupling chemistry, phylogeny of cultured and uncultured prokaryotes, metaproteomics and *aox* gene diversity analyses was performed to explore arsenic-dependent adaptation capacities of the microbial community. A large diversity was observed but only some of these prokaryotes were arsenic resistant and most of them belonged to Gamma-proteobacteria. Such bacteria are particularly adapted to persist in this environment as illustrated by our metaproteomic analyses. In addition, 77 *aoxB* genes that we classified in 4 distinct groups were characterized from both cultured and uncultured strains. In depth analyses of these genes allowed the definition of specific signature sequences for Alpha-proteobacteria. Altogether, this study revealed that despite a moderate arsenic contamination, some of the prokaryotes found in this environment have developed arsenic-specific adaptation processes.