UNIVERSITÉ LOUIS PASTEUR STRASBOURG

Thèse présentée pour obtenir le titre de

Docteur de l'Université Louis Pasteur Strasbourg 1

Discipline : Sciences du Vivant
Spécialité : Bioinformatique

par

Anne Friedrich

De la mutation structurale aux phénotypes des pathologies humaines : vers une approche intégrative des mutations et de leurs conséquences.

Soutenue publiquement le 21 décembre 2007 devant le jury :

Directeur de thèse	Olivier POCH, IGBMC, Illkirch
Rapporteur interne	Bruno KIEFFER, IGBMC, Illkirch
Rapporteur externe	Christophe BEROUD, IURC, Montpellier
Rapporteur externe	Marie-Hélène LE DU, CEA Saclay, Gif-sur-Yvette
Examinateur	Thierry TOURSEL, AFM, Evry
Membre invité	Luc MOULINIER, IGBMC, Illkirch

à mes parents,

Remerciements

Avant tout, je tiens à exprimer ma profonde reconnaissance à Bruno Kieffer, Christophe Béroud, Marie-Hélène Le Du et Thierry Toursel pour avoir accepté de juger ce travail de thèse.

Je voudrais également remercier Dino Moras et Jean-Claude Thierry pour m'avoir permis de travailler au sein d'un groupe et d'un département scientifiquement aussi riches que dynamiques.

Un merci tout particulier à mes deux « co-directeurs » de thèse, Olivier et Luc, même si cette appellation ne sera finalement pas officielle, peut être mon seul regret pour cette thèse. Merci à vous deux pour tout ce que vous avez fait pour moi : la façon dont vous avez su être là pour me guider, me soutenir, me motiver parfois, me remettre en place aussi, pour m'avoir fait confiance et m'avoir appris à avoir (de temps en temps) confiance. Merci pour votre complémentarité, votre générosité, cette passion qui vous anime et vos caractères parfois tellement… euh, comment dire… forts ?! qui font de vous ce que vous êtes et est tellement précieux.

Je tiens aussi à remercier les « souteneurs financiers » de cette aventure, somme toute assez nombreux, principalement l'AFM et plus particulièrement Thierry Toursel, pour la confiance accordée et le soutien infaillible depuis le début de l'aventure MS2PH!

Un grand merci à l'équipe de Gilbert Deléage et en particulier à Nicolas G., qui a toujours été disposé et disponible dans les moments où il fallait que tout aille vite, et ils ont été nombreux ! Bon courage à toi pour la dernière (longue) ligne droite de ta thèse. Merci aussi à Raphaël et Nicolas B., tellement efficaces et à notre écoute, ainsi qu'aux nombreux *IBMeurs* que nous avons été amenés à côtoyer.

Bien sûr, merci à toi, le Laboratoire de Bioinformatique et Génomique Intégratives. Comment exprimer tout ce que tu m'as apporté durant les trois dernières années… Ce qu'il y a de particulier dans ce labo, c'est cette ambiance, ce mélange de gens, de genres, de solidarité, de soutien, de discussions animées, de (fous) rires, de confidences parfois… Mais le labo, c'est aussi beaucoup de travail, pas mal d'efficacité et une belle synergie, qui fait qu'on se retrouve à « environ » 25 à ce jour et toujours en belle harmonie.

Comment remercier tout le monde sans oublier personne...

Bon, le labo, c'est avant tout Olivier, « extra-ordinaire » chercheur qui, malgré les innombrables tâches administratives qui lui incombent, nous fait profiter de son don tant qu'il le peut et arrive si souvent à faire émerger l'implicite caché tout au fond de chacun d'entre nous.

Le labo, c'est aussi ses piliers : Raymond, Julie, Odile, Fred et Luc, au labo depuis toujours sans doute… ou presque. Merci d'être là pour nous, alors que vous avez déjà tellement peu de temps pour vous ! Raymond, le Tcl, gscope et la piscine… Julie, les sempiternelles corrections, cette *so british attitude* et le tout avec le sourire *please*… Odile, un calme déroutant, une bonne humeur infaillible et ce don que tu as de transmettre ton savoir. Fred, toujours zen et prêt à rendre service, estampillés ISO9001 dans leurs qualités! Et pis « msieur Lusss » bien sûr, qui m'as fait me lancer dans l'aventure, m'a appris tellement et m'a aussi

offert quelques belles discussions durant nos trajets de retour (et pour ce soir, je propose sandwich merguez !!).

Le labo, c'est encore ce couloir, où nous sommes tellement heureux de n'avoir pas à travailler! Laetitia et Manu, bravo de résister encore et toujours, de nous supporter quand on vient à la rencontre du chef et merci pour votre sympathie et votre disponibilité.

Le labo, c'est bien sûr aussi les « djeuns », qui ne cessent de se démultiplier : Guillaume, Yannick-N, Laurent-P, David et Nicolas, qui ont même réussi à « djeunifier » Radwen... Retranchés dans un bureau où seuls eux veulent bien siéger, jamais à court de blagues graveleuses, toujours quelques bonbecs sous la main, mais aussi et avant tout constamment prêts à rendre service, avec un petit mot sympa pour la bonne humeur en prime. Merci pour votre aide et votre soutien…

Mais le labo c'est aussi la plate-forme : Laurent et mes premiers pas vers les cours, Véro et les si bonnes « papoteries » entre filles et Sophie, dernière blonde arrivée au labo (un peu de renfort, merci !), toujours souriante, à qui il ne manque plus que l'accent de chez nous (t'inquiètes, ca va venir) !

Pis enfin, le labo, c'est ceux qui ont eu la lourde tâche de me supporter durant les derniers mois, les proches au quotidien, ceux avec qui j'ai eu le bonheur de partager mon bureau. Annaïck et sa gentillesse, à qui je tiens à souhaiter *viel Glück für die weiteren*; Hoan, à qui je décerne officiellement le prix du meilleur standardiste vietnamien de l'institut ; Nicolas, toujours une petite anecdote sous le coude pour le sourire du matin ; Wolfgang qui bien qu'un mur nous sépare, a partagé, peut être bien malgré lui parfois, notre quotidien! Et bien sûr mes compagnons d'écriture et de galère : Ravi, toujours serein et serviable, et Yann, à tomber dans la dépression quand j'en sortais et vice versa. Merci pour nos dos à dos des dernières semaines, ces petits mots rassurants, le soutien et tout le reste.

Mais n'oublions pas, le labo, c'est aussi les anciens, allés voir si l'herbe était plus verte ailleurs : les jeunes docteurs, Jean, Aurélie et Fred ; Adeline, ma 1ière voisine de bureau, grâce à qui j'ai pu rapidement acquérir mon statut de pipelette (Véro n'y étant pas tout à fait étrangère non plus !).

Bref, merci le labo ...

Merci à Serge aussi, pour la gestion des serveurs et autres aléas informatiques. Un merci particulier à Laetitia, Anne et notre jeune retraitée, Christiane, pour votre gentillesse et votre efficacité et pour avoir toujours été là pour m'éclairer dans les méandres administratifs de l'institut.

Merci à tous mes proches, famille et amis, grâce à qui j'ai pu conserver ce semblant d'équilibre et qui m'ont permis de garder les yeux ouverts sur la « vraie » vie d'à coté…

Merci Mini et Chrisi, les amies de toujours ; Perrine et Eric dont les visites me sont tellement précieuses ; Anne, Arco, Patty, Célouche et les autres. Merci pour votre amitié, votre soutien, les bons moments et pour être toujours là pour moi.

Merci Jérôme, François et David de vous être occupés de mon homme pendant que je passais des heures solitaires devant mon PC. Attention, là, je le récupère un peu, la tendance va s'inverser!

Un grand merci aussi à ma belle-famille : Cati et Yves, les M&M&M&M, Olive et Aurélie, Manu, Stéphy et leur petite terreur de Lise. Merci pour tout le soutien, les encouragements, les bons moments, New-York et les corrections bien sûr !

Un immense merci à mes parents, et même s'il n'est pas vraiment coutume dans notre famille de s'étaler sur ce genre de considérations, sachez que je vous suis extrêmement reconnaissante pour tout, depuis toujours, et que je vous aime profondément! Et je t'assure maman, « Docteur », c'est pas si mal que ça, t'as pas à rougir devant les copines!

Merci aussi à mon frère Thomas, à Céline et aux filles, Marine et Emmy. Merci d'être là pour moi, toujours. Merci pour le soutien, la confiance et promis, je serai de nouveau là : plus de faux plans et de « débinades » en 2008 !

Enfin, merci, merci mon Cyrillou ! Pour tout, depuis bien longtemps et surtout d'avoir supporté ces derniers mois en ma compagnie. C'est avec regret que je lâcherai mon statut de « princesse » au 1er janvier 2008, mais je tâcherai d'en profiter encore d'ici là !

Dans ces remerciements, je n'oublierai pas non plus les profondeurs abyssales des piscines de la CUS, qui ont su retenir bon nombre de mes angoisses, doutes et autres remises en question durant ces trois années. Merci aussi a Franz Ferdinand, Mika, M, nounours et quelques autres pour avoir systématiquement réussi à me redonner du courage quand l'inspiration venait à me manquer…

Pour finir, je tiens à saluer ces enchaînements d'heureux hasards qui font avancer, évoluer, bouger, pour finalement revenir aux sources ; ces rencontres, scientifiques et humaines ; ceux qui vous font confiance, vous donne confiance et vous font au fur et à mesure trouver votre voie… Merci à tous ceux qui m'ont permis d'en arriver là.

Liste des abréviations

Å	Angström
ADN(c)	Acide DésoxyriboNucléique (complémentaire)
ARN(m)	Acide RiboNucléique (messager)
BIRD	*Biological Integration and Retrieval Data*, système générique de création de systèmes de bases de données
BIRD-QL	*BIRD Query Language*, langage de requête spécifique du système BIRD
CDD	Centre de Données Décrypthon
CDS	*CoDing Sequence*, ou séquence d'ADN codante
CGI	*Common Gateway Interface*
CRIHAN	Centre de Ressources Informatiques de HAute-Normandie
EBI	*European Bioinformatics Institute*
EMBL	*European Molecular Biology Laboratory*
GDB	*human Genome DataBase*
Gmean	*Geometric mean of accuracy*
Go	Gigaoctet
GO	*Gene Ontology*
HapMap	*Haplotype Map*
HGMD	*Human Gene Mutation Database*
HGP	*Human Genome Project*
HTML	*HyperText Markup Language*
HTTP	*HyperText Transfer Protocol*, ou protocole de transfert hypertexte
IBCP	Institut de Biologie et de Chimie des Protéines
IGBMC	Institut de Génétique et de Biologie Moléculaire et Cellulaire
kb	kilo bases, soit 1000 paires de bases
LMS	*Local Maximum Segments*, ou segments de conservation maximale
LOVD	*Leiden Open source Variation Database*
LSDB	*Locus-Specific DataBase*, ou banque de données spécifiques d'un locus
MACS	*Multiple Alignment of Complete Sequences*, ou alignement multiple de séquences complètes
MACSIMS	MACS *Information Management System*
MAGOS	*Multiple AliGnment and mOdelling Server*
MIM	*Mendelian Inheritance in Man*
mm	*mean method*, méthode d'échantillonnage dite des moyennes
MS2PH	de la Mutation Structurale aux Phénotypes des Pathologies Humaines
MS2PH-db	MS2PH-*database*
MSF	*Multiple Sequence Format*

NCBI	*National Center for Biotechnology Information*
NMDB	*National Mutation DataBase*
OMIM	*Online Mendelian Inheritance in Man*
pb	paire de bases
PCR	*Polymerase Chain Reaction*
PDB	*Protein Data Bank*
PHAT	*Predicted Hydrophobic And Transmembrane*
PSIC	*Position-Specific Independent Counts*
RENATER	Réseau National de Télécommunications pour la Technologie, l'Enseignement et la Recherche
sdm	*second derivative method*, méthode d'échantillonnage dite de la dérivée seconde
Se	Sensibilité
sm	*strips method*, méthode d'échantillonnage dite des bandelettes
(ns)SNP	(*non-synonymous*) *Single Nucleotide Polymorphism*, ou mutation ponctuelle isolée
Sp	Spécificité
SQL	*Structured Query Language*
TCP	*Transfer Control Protocol*
TM	TransMembrane
TNSALP	*Tissue-NonSpecific ALkaline Phosphatase*, phosphatase alcaline non tissu-spécifique
To	Téraoctet
UCSC	*University of California, Santa Cruz*
UMD	*Universal Mutation Database*
UniRef	*UniProt Reference clusters*
UTR	*UnTranslated Region*, ou région non traduite
XML	*Extensible Markup Language*

Table des Matières

Avant-Propos 1
Introduction 3
Chapitre 1 - Contexte biologique et bioinformatique 5
1.1. De la découverte de l'ADN… 5
1.2. … à la disponibilité du génome humain 10
Chapitre 2 - Le génome humain et ses variations 15
2.1. Le génome humain 15
2.1.1. Structure des gènes protéiques humains 17
2.1.2. Expression des gènes humains 19
2.1.3. Les protéines : produit de l'expression de gènes 19
2.1.4. Régions intergéniques 20
2.2. Variabilité du génome et évolution 21
2.2.1. Modifications de l'information génétique 23
2.2.2. Origine de l'apparition des mutations 24
2.2.2.1. Mutations spontanées 24
2.2.2.2. Mutations induites 26
2.2.3. Effets des mutations sur le génome 26
2.2.3.1. Substitution 26
2.2.3.2. Insertion et délétion de petite taille 27
2.2.3.3. Autres mutations 28
2.2.4. Conséquences des mutations selon leur localisation dans le génome 29
2.2.4.1. Mutation dans une région intergénique 29
2.2.4.2. Mutation dans une région non transcrite d'un gène 29
2.2.4.3. Mutation dans une région transcrite non codante d'un gène 30
2.2.4.4. Mutation dans une région codante d'un gène 30
2.2.4.4.1. Mutation silencieuse 31
2.2.4.4.2. Mutation exprimée 31
Chapitre 3 - Les maladies génétiques humaines 35
3.1. Qu'est-ce qu'une maladie génétique ? 35
3.1.1. Maladies génétiques et phénotypes 36
3.1.2. Génotype ou les causes moléculaires d'une maladie génétique 37
3.1.2.1. Perte de fonction 37
3.1.2.2. Gain de fonction 38
3.1.2.3. Effet dominant négatif 38
3.1.3. Corrélation génotype/phénotype 38
3.2. Schéma de transmission des maladies monogéniques 39
3.3. Caractérisation du génotype 41
3.3.1. Détermination des gènes responsables de maladies génétiques 42
3.3.1.1. Identification par voie cytogénétique 42
3.3.1.2. Approche positionnelle 42
3.3.1.3. Identification par l'approche du gène candidat 43
3.3.2. Détection expérimentale des mutations 44
3.3.2.1. Identification directe 45
3.3.2.2. Détection indirecte 46
3.4. Sources d'information 47
3.4.1. Banques centrales 49

3.4.2. Banques spécifiques d'un locus 50
3.4.3. Banques de polymorphismes 50
3.4.4. Banques nationales de mutations 51
3.5. Maladies génétiques humaines : quelques chiffres 51
Chapitre 4 - Les mutations et l'analyse de leurs conséquences 55
4.1. Conséquences des mutations sur les protéines 55
4.2. Analyse de protéines 57
4.2.1. Analyse évolutive 57
4.2.1.1. Famille de protéines 58
4.2.1.2. Identification des membres d'une famille protéique 59
4.2.1.3. Analyse des familles de protéines 59
4.2.2. Analyses structurales 61
4.2.2.1. Prédiction de la structure locale des protéines 62
4.2.2.2. Prédiction de la structure globale des protéines 64
Chapitre 5 - Vers une approche intégrative des mutations et de leurs conséquences 67
5.1. Etat de l'art 68
5.2. Projet MS2PH 73
5.2.1. Programme Décrypthon 74
5.2.2. Stratégie déployée dans le cadre de la thèse 74
Matériel & Méthodes 77
Chapitre 6 - Ressources informatiques et bioinformatiques 79
6.1. Équipement et ressources informatiques 79
6.1.1. IGBMC 79
6.1.2. Infrastructures supplémentaires dans le cadre du programme Décrypthon 80
6.1.2.1. Grille de calcul universitaire Décrypthon 81
6.1.2.2. Serveurs de stockage et de fédération des données 82
6.2. Savoir-faire bioinformatique du laboratoire 82
6.2.1. GScope : l'ossature bioinformatique du laboratoire 82
6.2.2. Ecriture de programmes : le Tcl/Tk 83
6.3. Les banques de données biologiques 84
6.3.1. Banques de données généralistes 84
6.3.1.1. UniProt 84
6.3.1.2. UniRef 86
6.3.1.3. PDB 87
6.3.2. Banques de données spécialisées 88
6.3.2.1. GDB 88
6.3.2.2. InterPro 89
6.3.3. Banques de données intégratives 89
6.3.3.1. GeneCards 89
6.3.3.2. Centre de Données Décrypthon 89
6.3.4. Interrogation des banques 90
6.3.4.1. SRS 91
6.3.4.2. Recherche de similarité dans les banques de données 92
6.3.4.3. BIRD-QL 92
6.3.5. PipeAlign : un outil d'analyse de familles protéiques 93
6.3.5.1. Ballast : traitement des résultats des recherches BlastP 94
6.3.5.2. DbClustal : construction de MACS 94
6.3.5.3. RASCAL : parcours et correction des alignements 94
6.3.5.4. LEON : extraction des séquences non homologues 95

6.3.5.5. NorMD : évaluation de la qualité d'un MACS 95
6.3.5.6. Secator et DPC : classification des séquences au sein d'un alignement 95
6.3.6. Edition et mise en forme des alignements multiples 96
6.3.7. MACSIMS : gestion de l'information au sein des alignements multiples 97
6.3.8. OrdAlie : analyse hiérarchisée d'alignements multiples 99
6.4. Analyse structurale des protéines 101
6.4.1. Modélisation de structures tridimensionnelles par homologie 101
6.4.2. Visualisation et mise en forme des structures tridimensionnelles 101
Chapitre 7 - Développements relatifs aux alignements multiples 103
7.1. Méthodes d'échantillonnage des résultats du Blast 103
7.1.1. Méthodes d'échantillonnage basées sur les *E-value* du Blast 104
7.1.1.1. Méthode des bandelettes 104
7.1.1.2. Méthode de la dérivée seconde 105
7.1.1.3. Méthode des moyennes 106
7.1.2. Méthodes basées sur la taxonomie 107
7.1.3. Méthode aléatoire 107
7.2. Calcul de la conservation au sein des alignements multiples 108
7.2.1. Méthode des seuils 108
7.2.2. Score de la distance moyenne 108
7.2.3. Score d'énergie libre 110
7.2.4. Score de la moyenne de la norme des vecteurs 110
Chapitre 8 - MAGOS 113
8.1. Fonctionnement général de MAGOS 113
8.2. Interface de soumission 114
8.3. Protocole implémenté dans MAGOS 115
8.3.1. Sélection de l'empreinte pour la construction du modèle tridimensionnel 115
8.3.2. Construction et annotation de l'alignement multiple 116
8.3.3. Création de l'alignement protéine d'intérêt / empreinte 116
8.3.4. Transferts des fichiers 117
8.3.5. Construction du modèle tridimensionnel 118
8.4. Interface de résultats 118
Chapitre 9 - MS2PH-db 123
9.1. Contenu de la banque 123
9.2. Structure de la banque 124
9.3. Identification des entrées de la banque 124
9.4. Calcul des données relatives à chaque entrée 125
9.4.1. Protocole général 126
9.4.1.1. Sélection des empreintes pour la construction des modèles tridimensionnels 126
9.4.1.2. Construction et annotation des alignements multiples 126
9.4.1.3. Construction des modèles tridimensionnels 127
9.4.2. Implémentation du protocole sur la grille de calcul Décrypthon 127
9.5. Interrogation de MS2PH-db 129
9.6. Interface de résultats 130
9.7. Mise à jour des données 133
Résultats & Discussion 135
Chapitre 10 - Problématique du contenu informationnel des alignements multiples 137
10.1. Alignement multiple de séquences protéiques : un outil central en biologie moderne 138
10.2. Alignement multiple : construction traditionnelle 140
10.3. La construction d'alignements multiples à l'ère post-génomique 141

10.3.1. Réduction *a priori* de l'espace de recherche des séquences 143
10.3.2. Réduction *a posteriori* des séquences détectées 143
10.4. Conclusions et perspectives 144
Chapitre 11 - Alignements multiples dédiés à l'analyse structurale 145
11.1. Echantillonnage des résultats de recherches Blast en vue de la construction d'alignements multiples dédiés à l'analyse structurale 145
11.2. Etude de la conservation des résidus au sein des alignements multiples 167
11.2.1. Détermination automatique de classes de conservation 167
11.2.2. Etude comparative des méthodes intégrées à OrdAlie 169
11.2.2.1. Rappels : paramètres statistiques 169
11.2.2.2. Résultats de l'analyse 170
11.2.2.3. Conclusions et perspectives 171
Chapitre 12 - Caractérisation fonctionnelle et structurale des mutations 173
12.1. Illustration des capacités d'intégration de MAGOS 179
12.1.1. Myotubularine 179
12.1.2. Myopathie myotubulaire liée à l'X 180
12.1.3. Analyse de mutations affectant la myotubularine par MAGOS 180
12.1.3.1. Requête MAGOS 180
12.1.3.2. Données de mutation relatives à la myotubularine 182
12.1.3.3. Analyse des mutations dans l'environnement MAGOS 183
12.1.3.4. Conclusions et perspectives 186
12.2. Statistiques d'utilisation du serveur MAGOS 187
Chapitre 13 - Mise en place des bases de l'infrastructure dédiée à l'analyse globale des mutations 189
13.1. Données génotypiques et phénotypiques 189
13.1.1. Banques de données de mutations 190
13.1.2. Dans un monde idéal… 190
13.1.3. En réalité… 191
13.1.4. Un espoir : vers une mise en place de standards 191
13.1.5. Notre choix 193
13.2. MS2PH-db 194
13.2.1. La myotubularine : une entrée de MS2PH-db 194
13.2.1.1. Carte d'identité de la myotubularine : vue informationnelle 194
13.2.1.2. Interconnexion des vues informationnelles, structurales et évolutives 196
13.2.2. MS2PH-db en quelques chiffres 198
13.3. L'absolue nécessité de l'infrastructure Décrypthon dans le cadre de MS2PH-db 199
13.3.1. Déploiement de la grille de calcul universitaire Décrypthon 199
13.3.2. Implémentation des outils sur la grille de calcul Décrypthon 199
13.3.3. Intérêt de la grille Décrypthon dans le cadre de MS2PH-db 200
13.4. Conclusions et perspectives 201
Chapitre 14 - Vers une typification des mutations 203
14.1. Prédiction des conséquences des mutations 203
14.1.1. Discrimination des mutations neutres/délétères 203
14.1.2. Prédiction d'un phénotype clinique à partir du génotype d'un individu 204
14.2. Présentation des deux « systèmes » étudiés 205
14.2.1. Système 1 : l'hypophosphatasie 206
14.2.1.1. Maladie et phénotypes cliniques 206
14.2.1.2. Phosphatase alcaline non tissu-spécifique 206
14.2.1.3. Relation génotype/phénotype 207

14.2.1.3.1. Transmission sur le mode dominant 207
14.2.1.3.2. Transmission sur le mode récessif 207
14.2.2. Système 2 : la myopathie myotubulaire liée à l'X 208
14.3. Analyses préliminaires relatives aux deux systèmes 208
14.3.1. Données génotypiques et phénotypiques 208
14.3.1.1. Données de mutations relatives à la TNSALP 209
14.3.1.2. Données de mutations relatives à la myotubularine 209
14.3.2. Prédictions des effets délétères/neutres des substitutions 209
14.3.3. Vers une typification automatisée des positions délétères 211
14.3.3.1. Données fonctionnelles 211
14.3.3.2. Données évolutives 213
14.3.3.3. Données structurales 214
14.3.3.4. Caractérisation des substitutions 219
14.3.3.5. Effets cliniques liés à un génotype 220
14.3.4. Vers la mise en place d'un système expert 222
Conclusions & Perspectives 223
Chapitre 15 - Conclusions et perspectives 225
Annexes 229
Références Bibliographiques 267

Table des Figures

Figure 1 Dogme Central de la biologie moléculaire et flux de l'information génétique.. 8
Figure 2 Evolution du nombre de projets disponibles sur le site GOLD. 9
Figure 3 De la disponibilité de la séquence complète du génome humain. 11
Figure 4 Représentation d'une paire de chromosomes homologues. 16
Figure 5 Structure du gène humain : de l'ADN génomique à la protéine. 17
Figure 6 Exemple d'épissage alternatif. 18
Figure 7 Repliement des protéines selon les quatre niveaux de structuration. 20
Figure 8 Echange d'informations génétiques entre chromosomes homologues. 22
Figure 9 Remaniements chromosomiques entraînant une anomalie de structure 24
Figure 10 Réplication semi-conservative du génome.. 25
Figure 11 Mutations ponctuelles au niveau de l'ADN. 26
Figure 12 Glissement de réplication. 28
Figure 13 Code génétique universel. 31
Figure 14 Conséquences des mutations sur la synthèse de la protéine. 33
Figure 15 Arbres généalogiques : schémas de transmission des maladies monogéniques 40
Figure 16 Identification de mutation par clonage positionnel. 43
Figure 17 Séquençage automatique d'un gène : exemple d'électrophérogramme 45
Figure 18 Détection indirecte de mutations par analyse de liaison. 47
Figure 19 Augmentation du nombre d'entrées dans MIM et OMIM. 48
Figure 20 Variété de fonctions des protéines référencées dans OMIM. 52
Figure 21 Répartition des types de mutations à l'origine de maladies génétiques humaines.. 53
Figure 22 Classification des acides aminés d'après leurs propriétés physico-chimiques 56
Figure 23 Relations d'homologie. 58
Figure 24 Vue schématique d'un MACS.. 60
Figure 25 Structure tridimensionnelle d'une protéine en interaction avec son ligand. 61
Figure 26 Schématisation de la grille de calcul Décrypthon. 81
Figure 27 Evolution du nombre d'entrées de la banque Swiss-Prot depuis 1986.. 85
Figure 28 Evolution du nombre d'entrées de la banque TrEMBL depuis 1996 86
Figure 29 Evolution du nombre d'entrées de la banque PDB depuis 1976 87
Figure 30 Page d'accueil du serveur Web SRS de l'IGBMC. 91
Figure 31 Exemple de requête BIRD-QL. 93
Figure 32 Aperçu de la cascade de programmes constituant PipeAlign. 96
Figure 33 Editeur d'alignements multiples Seqlab. 97
Figure 34 Etapes successives de MACSIMS. 98
Figure 35 Alignement multiple annoté par MACSIMS. 99
Figure 36 Représentation des conservations adoptée dans OrdAlie. 100
Figure 37 Méthode des bandelettes 105
Figure 38 Méthode de la dérivée seconde 106
Figure 39 Méthode des moyennes. 106
Figure 40 Exemple de requête BIRD-QL dans le cadre du filtre Eucaryote. 107

Figure 41 Calcul de la distance D_i dans l'espace à R dimensions. 109
Figure 42 « Rosace des acides aminés » 111
Figure 43 Scores de conservation par la méthode de la norme des vecteurs moyens. 111
Figure 44 Organisation du serveur MAGOS 114
Figure 45 Interface de soumission de MAGOS. 115
Figure 46 Schématisation des calculs effectués sur le serveur *Star* 117
Figure 47 Page d'accueil des résultats de MAGOS. 118
Figure 48 Interface de résultat Jmol fournie par MAGOS. 119
Figure 49 Fenêtre externe relative aux numéros d'accession des séquences alignées. 120
Figure 50 Interface de résultat Jmol : alignement par paires utilisé pour la construction du modèle tridimensionnel 121
Figure 51 Sélection des protéines en entrée de MS2PH-db. 125
Figure 52 Schématisation du fonctionnement de nos applications sur la grille. 128
Figure 53 Formulaire d'interrogation « simple » de MS2PH-db 129
Figure 54 Capture d'écran du résultat d'une requête simple 130
Figure 55 Carte d'identité d'une protéine de la banque, la phosphatase alcaline. 131
Figure 56 Interface Jmol d'interconnexion des différentes « vues ». 132
Figure 57 Fonctionnalités supplémentaires intégrées à l'interface Jmol. 133
Figure 58 Rôle central des alignements multiples de séquences protéiques 139
Figure 59 Aperçu des principaux programmes d'alignements multiples disponibles 141
Figure 60 Principales étapes des méthodes de typification des colonnes de conservation. 168
Figure 61 Représentation schématique de la constitution en domaines de la myotubularine. 179
Figure 62 Page d'accueil des résultats obtenus par MAGOS pour la myotubularine. 181
Figure 63 Interconnexion de la séquence de la myotubularine dans l'alignement et son modèle tridimensionnel 182
Figure 64 Localisation des mutations sur le modèle tridimensionnel de la myotubularine. 184
Figure 65 Représentation des mutations situées à proximité de la poche catalytique 185
Figure 66 Visualisation de la localisation de la mutation R421Q 186
Figure 67 Fréquence des mutations observées le long d'un gène 192
Figure 68 Carte d'identité de la myotubularine dans MS2PH-db. 195
Figure 69 Localisation de l'ensemble des mutations affectant la myotubularine 197
Figure 70 Visualisation des résidus situés à moins de 12 Å de l'arginine en position 421 de la séquence de la myotubularine. 198
Figure 71 Prédiction des effets délétères de substitutions par SIFT, PolyPhen et SNAP 210
Figure 72 Prédictions des effets des substitutions faux-sens par SIFT 211
Figure 73 Données fonctionnelles relatives à la TNSALP et à la myotubularine 212
Figure 74 Conservation le long de la TNSALP et de la myotubularine 214
Figure 75 Localisation des mutations sur les modèles structuraux 215
Figure 76 Découpage du modèle tridimensionnel de la TNSALP en domaines structuraux 216
Figure 77 Eléments de structure secondaire le long de la TNSALP et de la myotubularine 217
Figure 78 Acessibilité à la surface des résidus constituant la TNSALP et la myotubularine 218
Figure 79 Proximité des mutations par rapport aux sites « remarquables » 219
Figure 80 Localisation des mutations dominantes dans la TNSALP. 221

Table des Tableaux

Tableau 1 Chronologie non exhaustive des évènements marquants en biologie, informatique et bioinformatique de 1944 à 2001. 7
Tableau 2 Génome humain : quelques statistiques 19
Tableau 3 Statistiques d'OMIM, au 15 août 2007. 51
Tableau 4 Programmes de prédiction de structures secondaires. 63
Tableau 5 Règles de prédiction respectées par PolyPhen. 71
Tableau 6 Performance des méthodes de prédiction des effets délétères de substitution. 72
Tableau 7 Banques de données intégrées au Centre de Données Décrypthon (août 2007). 90
Tableau 8 Sensibilité, spécificité et *Gmean* associés à chacune des 14 méthodes de caractérisation des conservations au sein des alignements multiples.. 171
Tableau 9 Récapitulatif des mutations faux-sens affectant la myotubularine 183
Tableau 10 Temps de calcul pour la construction de deux alignements multiples annotés. 200
Tableau 11 Classes de conservation mises en évidence au sein des alignements multiples des séquences eucaryotes homologues de la TNSALP et de la myotubularine. 213

Avant-Propos

Les travaux présentés dans ce manuscrit de thèse ont été effectués au Laboratoire de Bioinformatique et Génomique Intégratives de l'IGBMC, sous la direction d'Olivier Poch et la supervision de Luc Moulinier, dans le cadre d'un projet intitulé MS2PH, acronyme du titre « de la Mutation Structurale aux Phénotypes des Pathologies Humaines ».

MS2PH est un projet pilote du programme Décrypthon, initié par l'AFM, le CNRS et IBM afin de mettre à la disposition des biologistes une grille de calcul puissante, ainsi que des banques et des outils bioinformatiques susceptibles de répondre aux besoins émergeants liés aux projets scientifiques de grande envergure.

MS2PH est un projet ambitieux ayant pour objectif, à terme, le déploiement et la mise à disposition d'une infrastructure à même de faciliter la compréhension des relations qui existent, au niveau des protéines impliquées dans les maladies génétiques humaines, entre la séquence de la protéine, son évolution, sa structure tridimensionnelle et les pathologies qui y sont associées. Deux aspects de ce projet peuvent ainsi être distingués : le premier est centré sur l'implémentation ou le développement d'outils bioinformatiques en vue d'appréhender de manière efficace ces relations, le second est clairement axé sur les études à haut débit, l'ensemble des protéines impliquées dans des maladies monogéniques humaines étant concerné. Nous avons, tout au long de nos travaux, traité ces deux aspects en parallèle, pour aboutir à la mise en place des bases de l'infrastructure dédiée à l'analyse globale des relations qui lient le phénotype d'un individu à son génotype dans le cadre des maladies monogéniques humaines.

Ce manuscrit est divisé en trois parties :

La première partie consiste en une introduction qui, après avoir exposé le contexte scientifique général de nos études (Chapitre 1), décrit dans ses grandes lignes le génome humain et sa variabilité (Chapitre 2), pour se concentrer par la suite sur les maladies génétiques humaines (Chapitre 3). Elle aborde enfin les moyens offerts par la bioinformatique pour l'étude des protéines (Chapitre 4), avant de présenter les outils de prédiction des effets des mutations sur les protéines disponibles, ainsi que la démarche que nous avons suivie au cours de nos travaux (Chapitre 5).

La deuxième partie, *Matériel & Méthodes*, relate l'environnement bioinformatique dans lequel j'ai évolué au cours de ces trois années (Chapitre 6) et explique les aspects méthodologiques des développements bioinformatiques effectués durant ma thèse (Chapitre 7 : méthodes d'échantillonnage, Chapitre 8 : serveur MAGOS et Chapitre 9 : la banque de données MS2PH-db).

La troisième partie, intitulée *Résultats & Discussion*, débute par l'exposé de la problématique liée à l'accumulation de données biologiques à l'ère post-génomique et de ses conséquences dans les domaines de l'extraction et de l'analyse d'une information pertinente, notamment dans le cadre d'études à haut débit (Chapitre 10). La réflexion que nous avons menée à ce sujet se base plus particulièrement sur les alignements multiples. Elle a donné lieu à une publication, présentée en annexe, à paraître au courant de l'année 2008 dans un livre dédié aux alignements multiples. Le Chapitre suivant rebondit sur cette problématique et présente une étude relative aux méthodes d'échantillonnage et à la réduction du nombre des séquences à aligner tout en maintenant un contenu informationnel suffisant (Chapitre 11). Cette étude, menée dans le cadre de la construction d'alignements multiples dédiés à l'analyse structurale, est présentée dans la Publication n°1 et est complétée, dans la suite du chapitre, par des résultats récents concernant la caractérisation et l'analyse automatiques de la conservation des résidus au sein des alignements multiples. La présentation du serveur Web MAGOS (Publication n°2), qui fournit un espace de travail réunissant les informations de séquence, de structure, de fonction et d'évolution relatives à une protéine d'intérêt, est appuyée par l'illustration des possibilités d'analyse intégrative qu'il offre (Chapitre 12). Enfin, MS2PH-db, la banque de données consacrées aux protéines impliquées dans les maladies monogéniques humaines, qui représente les bases de l'infrastructure dédiée à l'analyse globale des mutations et s'appuie sur la puissance de la plateforme technologique Décrypthon, fait l'objet d'un nouveau chapitre (Chapitre 13). Cette troisième partie s'achève sur un chapitre « exploratoire » (Chapitre 14) qui constitue une transition entre les résultats de nos études et les perspectives de nos travaux et qui présente une analyse préliminaire visant à mettre en application nos développements, afin d'ouvrir la voie vers la mise en place d'un système générique de prédiction des effets d'un génotype sur le phénotype clinique d'un individu.

Introduction

Chapitre 1 - Contexte biologique et bioinformatique

1.1. De la découverte de l'ADN…

La biologie a connu un essor fulgurant durant le 20ième siècle et a été profondément bouleversée au cours de la dernière décennie par le passage à l'ère post-génomique, la généralisation de la biologie à haut débit et l'avènement d'une discipline relativement jeune, la bioinformatique. Cette dernière, qui englobe la collecte, l'organisation et l'analyse de données biologiques au moyen de l'outil informatique, est rapidement devenue indissociable des études relatives à la compréhension du vivant.

Le Tableau 1 présente les principaux évènements marquants en biologie, informatique et bioinformatique, de la découverte de l'ADN comme support de l'information génétique en 1944 au séquençage du génome humain en 2001. Les évènements relatifs à la biologie, l'informatique et la bioinformatique sont représentés respectivement en noir, bleu et vert. En rouge quelques évènements plus spécifiquement liés à la génétique.

Date	Auteur	Evènement
1944	Avery	Démontre que l'ADN est le support de l'information génétique (Avery *et al.*, 1944)
1953	Watson, Crick	Etablissement de la structure en double hélice de l'ADN (Watson and Crick, 1953)
	IBM 650	Premier ordinateur commercial
	Sanger	Détermination de la séquence des chaînes A et B de l'insuline (Sanger and Thompson, 1953a; Sanger and Thompson, 1953b)
1955	Morton	Mise au point d'une méthode de calcul de la liaison génétique chez l'homme (LOD score) (Morton, 1955)
1956	Tijo, Levan	Preuve que le nombre diploïde de chromosomes chez l'homme est 46 (Tijo and Levan, 1956)
	Anfinsen	La structure tridimensionnelle d'une protéine est fonction de sa séquence (Anfinsen and Redfield, 1956)
	IBM	Commercialisation des premiers disques durs
1957	Ingram	Démonstration que l'anémie falciforme est la conséquence d'un seul changement dans la sous-unité β de l'hémoglobine
1958	Crick	Enonciation du Dogme Central de la biologie moléculaire (Crick, 1958)
1959	Lejeune	La trisomie du chromosome 21 est associée au syndrome de Down (Lejeune *et al.*, 1959)
1962	Nirenberg, Matthaei	Déchiffrage du code génétique (Matthaei *et al.*, 1962)
1965	Monod, Jacob, Wolf	Découverte des mécanismes de la régulation génétique impliqués dans le Dogme Central

	Dayhoff	Premier atlas de séquences et structures de protéines (Dayhoff *et al.*, 1965)
	Moore	Loi de Moore sur l'évolution de la puissance de calcul
1967	Fitch	Construction d'arbres phylogénétiques (Fitch and Margoliash, 1967)
1968	McKusik	Publication de la première édition du livre « Mendelian Inheritance in Man MIM »
1969	ARPANET	Premières interconnexions universitaires
1970	Needlman, Wunsch	Algorithme d'alignement global optimal entre deux séquences de protéines (Needleman and Wunsch, 1970)
1974	Cerf, Kahn	Développement du concept d'Internet et du protocole TCP
	Chou	Algorithme de prédiction des structures secondaires des protéines (Chou and Fasman, 1974)
1977	Sanger	Méthode de séquençage de séquences nucléiques (approche enzymatique) (Sanger *et al.*, 1977)
	Maxam, Gilbert	Méthode de séquençage de séquences nucléiques (approche chimique) (Maxam and Gilbert, 1977)
1978	Sanger	Séquençage du bactériophage phiX174 (5386 pb) (Sanger *et al.*, 1978)
1980	Botstein, White, Slonick, Davis	Conception des bases de la cartographie par le polymorphisme de la longueur des fragments de restriction
	EMBL	Création d'une banque européenne de séquences nucléiques
1981	Smith, Waterman	Algorithme d'alignement local optimal entre deux séquences (Smith and Waterman, 1981)
	IBM	Premier ordinateur sous le nom de *Personal Computer* (PC)
	Anderson	Séquençage du génome mitochondrial humain (Anderson *et al.*, 1981)
1982	GenBank	Création de la banque américaine de séquences nucléiques
1983	Mullis	Invention de la Réaction de Polymérisation en Chaîne (PCR)
1984	Gouy	ACNUC, logiciel d'interrogation de banques de séquences (Gouy *et al.*, 1984)
1985	Lipman, Pearson	FASTA, programme de recherche de similarité dans les banques de données (Lipman and Pearson, 1985)
	Sony, Philips	Création d'un nouveau support numérique, le *Compact Disc* (CD)
1986	Swiss-Prot	Création de la banque de séquences protéiques
	DDBJ	Création de la banque japonaise de séquences nucléiques
	Roderick	Apparition du terme « *genomic* »
1987	Hoffman	Identification du gène de la dystrophie musculaire de Duchenne par la stratégie du clonage positionnel (Hoffman *et al.*, 1987)
	Applied Biosystems	Commercialisation du premier séquenceur automatisé
	McKusick	Première carte génétique du génome humain (McKusick and Ruddle, 1987)
	Puces à ADN	Apparition de la technologie
1988	HUGO	Coordonne le décryptage mondial du génome humain
	Higgins	CLUSTAL : programme d'alignement multiple de séquences (Higgins and Sharp, 1988)
1989	Fields	Système double-hybride permettant la détection des interactions entre deux protéines

	Internet	Internet succède à ARPANET
1990	Altschul	BLAST, programme de recherche de similarité dans les banques de données (Altschul *et al.*, 1990)
	Berners-Lee	Publication du premier document HTML
	HGP	Initiation du « *Human Genome Project* », visant à décrypter l'intégralité du génome humain
		Premier essai de thérapie génique
1991	Adams	Premier séquençage à grande échelle d'ADNc (EST) (Adams *et al.*, 1991)
	Roberts	GRAIL : programme de localisation de gènes
1992		Séquençage du chromosome III de *Saccharomyces cerevisiae*
		13 gènes de maladies génétiques identifiés par clonage positionnel
1993	Cohen	Première carte physique du génome humain (Cohen *et al.*, 1993)
	Boguski	dbEST : banque de données internationale d'EST (Boguski *et al.*, 1993)
	Etzold	SRS : logiciel d'interrogation de banques (Etzold and Argos, 1993)
1995	Fleischmann	Séquençage du premier organisme vivant, *Haemophilus influenza* (Fleischmann *et al.*, 1995)
	DVD Forum	Création d'un support numérique de stockage de haute capacité, le DVD
1996	Walsh	Séquençage du premier génome eucaryote, *Saccharomyces cerevisiae*
	Affymetrix	Commercialisation de la première puce à ADN
1998	W3C	Création du format XML
2000	Adams	Séquençage du génome de *Drosophila melanogaster* (Adams *et al.*, 2000)
	Ashburner	Création de la banque d'annotation Gene Ontology (Ashburner *et al.*, 2000)
2001	Lander	Séquence préliminaire du génome humain par HGP (Lander *et al.*, 2001)
	Venter	Séquence préliminaire du génome humain par Celera Genomics (Venter *et al.*, 2001)

Tableau 1 - Chronologie non exhaustive des évènements marquants en biologie, informatique et bioinformatique de 1944 à 2001.
(traduit et adapté de http://bio.cc/Bioinformatics/history_of_bioinformatics.html, http://villemin.gerard.free.fr/Wwwgvmm/Histoire/Informat.htm et (Pasternak, 1999))

C'est au milieu du 20ième siècle que la biologie et notre vision de la vie ont évolué, grâce à deux évènements majeurs :

- la découverte de l'ADN comme support de l'information génétique (Avery *et al.*, 1944), suivie, quelques années plus tard, par l'élucidation de sa structure en double hélice par Watson et Crick (Watson and Crick, 1953),
- l'énoncé du Dogme Central de la Biologie Moléculaire (Crick, 1958), introduit par Francis Crick en 1958, qui démontre le rôle central des gènes et des protéines au sein de l'organisation de la cellule.

Le Dogme Central est un modèle représentant le flux de l'information génétique entre l'ADN, l'ARN et la protéine. Il est expliqué par trois processus majeurs (Figure 1) :

- la réplication, qui permet la duplication de l'information portée par l'ADN, au cours d'un processus impliquant de nombreuses enzymes,
- la transcription, qui permet le passage de l'ADN à l'ARN messager (ARNm). Au sein des cellules eucaryotes, l'ARNm subit une étape de maturation et migre du noyau vers le cytoplasme,
- la traduction, par l'intermédiaire des ribosomes, qui permet la synthèse des protéines, pièces centrales dans l'édifice des activités cellulaires, en se basant sur l'information portée par l'ARNm.

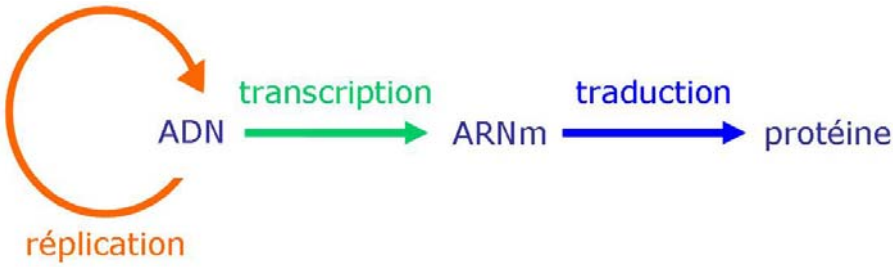


Figure 1 - Dogme Central de la biologie moléculaire et flux de l'information génétique.

Le modèle stipule que le flux de l'information génétique est à sens unique. Il a cependant été démontré que ce dernier est une simplification de la réalité : il est en effet possible, par l'intermédiaire de la transcriptase inverse, de passer de l'ARN à l'ADN (Baltimore, 1970; Temin and Mizutani, 1970) et l'ARN peut également se dupliquer (August *et al.*, 1965).

La découverte de ces concepts révolutionnaires coïncide avec le début d'une évolution rapide dans la recherche et le développement de nouvelles techniques en biologie. Elle va contribuer à l'émergence d'une nouvelle discipline, la bioinformatique qui deviendra rapidement un domaine de recherche à part entière.

Toute une partie du monde de la biologie va ainsi se consacrer à l'étude des gènes tant au niveau de leur séquence, de leur structure, de leur régulation que de leur fonction.

Au milieu des années 60, suite à la disponibilité des premières séquences protéiques, un travail de regroupement et d'archivage de séquences est initié par Margaret Dayhoff qui crée alors la première banque de séquences protéiques, publiée dans l'*Atlas of Protein Sequence and Structure* (Dayhoff, 1965). Il devient rapidement évident qu'une solution informatisée serait plus adaptée à la gestion des données de séquences. C'est dans ce contexte que naît la bioinformatique, se concentrant, dans un premier temps, sur la mise en place de solutions de stockage et donc sur le développement des premières banques de données biologiques.

En 1977, deux méthodes de séquençage des acides nucléiques apparaissent simultanément : une approche enzymatique (Sanger *et al.*, 1977) et une approche chimique (Maxam and Gilbert, 1977). La première méthode prend le pas sur la seconde, jugée trop toxique pour le manipulateur et offre la possibilité à tous les laboratoires de séquencer leurs gènes d'intérêt. Néanmoins, le séquençage de génome complet restera longtemps un challenge et, outre les petits génomes de virus, il faudra attendre le début des années 90 pour le décryptage de génomes complets d'organismes appartenant à l'un des trois domaines du vivant (*Bacteria*, *Archaea* et *Eucarya*) (Woese *et al.*, 1990).

A la fin du 20ième siècle, un profond changement d'échelle s'effectue, déclenché par le passage de l'ère du gène à celui du génome et l'avènement de la génomique. Les projets de séquençage de génomes complets se sont rapidement multipliés grâce aux progrès des techniques de séquençage à haut débit, de sorte que la communauté scientifique dispose actuellement des séquences de 609 génomes complets (GOLD, http://www.genomesonline.org, au 16 juillet 2007) dont 45 archées, 499 bactéries et 65 eucaryotes, sachant qu'un nombre bien plus important de projets de séquençage est en cours (Figure 2).

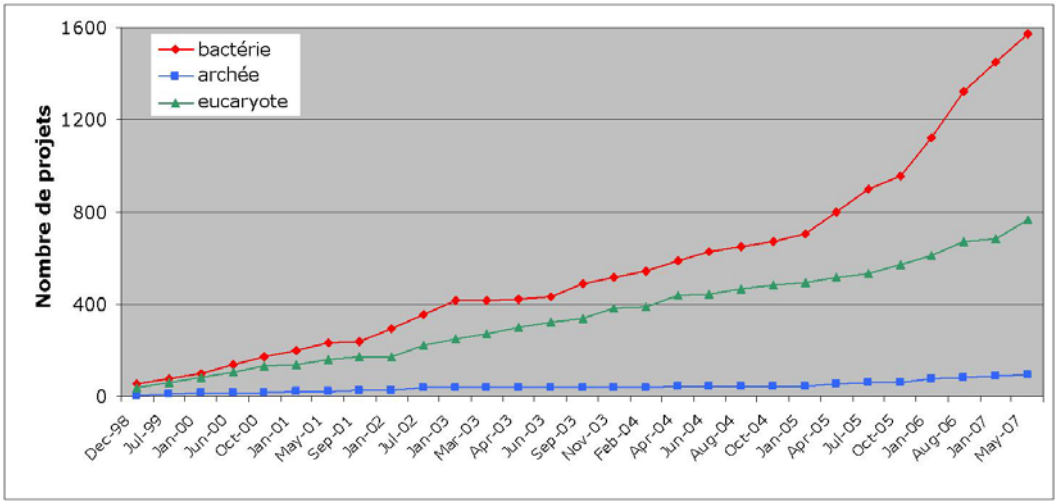


Figure 2 - Evolution du nombre de projets disponibles sur le site GOLD, en fonction du groupe phylogénétique.

Ainsi, la biologie a été profondément bouleversée non seulement par l'entrée dans l'ère post-génomique et la disponibilité des séquences de nombreux génomes complets, mais aussi par la généralisation de nouvelles technologies à haut débit, souvent caractérisées par le suffixe « omique ». Parmi celles-ci, la transcriptomique et la protéomique s'intéressent respectivement à l'ensemble des ARNm et des protéines présents dans une cellule à un moment particulier ; la génomique structurale désigne la détermination de la structure tridimensionnelle des protéines à grande échelle mais parfois également la cartographie et

l'étude de la structure du génome ; l'interactomique se définit quant à elle comme l'étude de l'ensemble des interactions ayant lieu au cœur d'une cellule.

Cette révolution dans les sciences du vivant est également indissociable des avancées réalisées en bioinformatique, dont l'essor a été catalysé par la généralisation de l'utilisation des technologies à haut débit. En effet, l'accumulation massive de données hétérogènes, tant par leur nature que par leur qualité, a nécessité le développement et la mise en place d'outils d'acquisition, d'interrogation et de gestion de l'information performants. L'analyse efficace et rapide de ces données n'est possible que grâce au développement subséquent d'algorithmes dédiés à la bio-analyse, souvent complémentés par des études statistiques, qui permettent d'en extraire une signification biologique pertinente, en proposant notamment des prédictions ou des modélisations biologiques. Notons également que l'essor de la bioinformatique repose en grande partie sur l'outil Internet, qui a permis la diffusion, au sein de la communauté scientifique, des données, outils et programmes au fur et à mesure de leur développement.

1.2. … à la disponibilité du génome humain

Le début du 21ième siècle a été particulièrement marqué par la mise à disposition de la séquence complète du génome humain, après une dizaine d'années d'efforts de la communauté internationale.

L'idée du décryptage intégral de notre patrimoine génétique est née en 1985 à l'*Imperial Cancer Research* de Londres. Le budget considérable alloué pour ce projet, les efforts de nombreux laboratoires internationaux pour sa réalisation et la coordination mondiale de ce décryptage par HUGO (*HUman Genome Organization*), puis HGP (*Human Genome Project*), ont permis la publication d'une version préliminaire du génome humain et de ses 2,9 milliards de nucléotides en 2001 (Lander *et al.*, 2001; Venter *et al.*, 2001). Cette première version, ou *draft*, ne couvrait cependant pas la totalité du génome (3,2 milliards de nucléotides), en raison de l'importance des régions centromériques et télomériques des chromosomes essentiellement constituées de séquences d'ADN fortement biaisées en composition et hautement répétitives. Il a cependant été considéré que cette version préliminaire couvrait les parties les plus importantes du génome, ces séquences répétées ne contenant que peu, ou pas, de gènes (Bork and Copley, 2001).

Depuis 2004, le séquençage et surtout l'assemblage du génome humain peuvent être considérés comme achevés (IHGSC, 2004), bien que ce dernier soit encore régulièrement mis à jour : la version courante (août 2007) de l'assemblage du génome humain date de septembre 2006 (version 36.2). Plusieurs serveurs Web, comme l'UCSC (*University of California, Santa Cruz*, http://genome.ucsc.edu/), le NCBI (*National Center for Biotechnology Information*, http://www.ncbi.nlm.nih.gov/mapview/map_search.cgi?taxid=9606), et

ENSEMBL (http://www.ensembl.org/Homo_sapiens/index.html) offrent à présent la possibilité de naviguer à travers le génome humain et fournissent des outils adaptés à l'analyse de séquences.

Comme l'illustre de façon humoristique la Figure 3, la connaissance de la séquence complète du génome humain n'est pas une fin en soi. Sa disponibilité annonce plutôt de formidables perspectives en termes de recherche scientifique et ouvre la voie à la possibilité d'étudier les gènes humains dans leur contexte génomique, de tenter de définir leur fonction cellulaire ainsi que leurs conditions d'expression afin d'obtenir une meilleure compréhension globale de la cellule, du tissu, voire de l'organisme en tant que « système biologique ». Notons que c'est de la volonté d'intégrer ces différents niveaux d'information qu'une approche récente des sciences de la vie a émergé, à savoir la biologie des systèmes. Celle-ci repose sur la mise en commun des compétences de biologistes, d'informaticiens, de mathématiciens et de physiciens afin de modéliser et de comprendre le fonctionnement de systèmes biologiques dans leur globalité.

La finalisation du projet HGP représente donc le point de départ d'une réelle révolution dans le domaine des sciences du vivant.

Figure 3 - De la disponibilité de la séquence complète du génome humain : une quantité impressionnante de données disponibles à interpréter. (http://cagle.msnbc.com/news/gene/gene14.asp)

Un certain nombre d'initiatives centrées sur le génome humain ont vu le jour parallèlement au projet initial. Le comité HGNC, pour *HUGO Gene Nomenclature Committee*, s'est vu attribuer pour mission l'approbation des noms et symboles assignés aux nouveaux gènes mis en évidence. La GDB (pour *human Genome DataBase*) (Pearson, 1991) se définit comme la banque de données du projet de cartographie du génome humain et a été développée en

parallèle du projet HGP. Chaque locus, gène ou fragment d'ADN anonyme y possède un nom et une localisation plus ou moins précise.

La banque HPRD (Peri *et al.*, 2003), pour *Human Protein Reference Database*, se veut une banque de référence en matière de protéines humaines et a été construite dans l'esprit d'une plateforme initiale pour des études de biologie systémique chez l'homme. Une quantité importante d'informations hétérogènes est disponible pour chaque entrée comme les modifications post-traductionnelles, des données d'interactions, de localisation cellulaire, les maladies associées ainsi que des liens vers d'autres banques. La qualité des informations est garantie par leur vérification manuelle réalisée par des experts.

La disponibilité du génome humain a également ouvert la voie à l'exploitation du protéome humain, auquel les biologistes se sont rapidement intéressés. A l'annonce de la publication de la version préliminaire de la séquence du génome, l'organisation HUPO (*HUman Proteome Organization*) a vu le jour et a notamment permis la mise en place d'un standard pour la représentation des données de protéomique, HUPO-PSI (*Proteomics Standard Initiative*) (Hermjakob, 2006).

De plus, profitant des avancées technologiques et de l'extraordinaire élan d'optimisme induit par la réussite d'un projet qui semblait totalement improbable à ses débuts, de nombreuses initiatives ont été lancées pour rebondir autour des données issues du génome humain, telles que l'établissement de cartes du transcriptome humain ou la détermination de l'interactome humain (Ideker and Valencia, 2006). Ce dernier s'appuie en grande partie sur les résultats du projet de l'ORFeome humain (Lamesch *et al.*, 2007), qui met à disposition un répertoire de clones des phases ouvertes de lectures codant pour des protéines chez l'homme. Citons également le projet « Repliement du protéome humain » (http://www.systemsbiology.org/Technology/Data_Visualization_and_Analysis/Human_Proteome_Folding_Project), basé sur la puissance informatique de millions d'ordinateurs et ayant pour but de modéliser les repliements de toutes les protéines humaines, afin, notamment, de mieux comprendre leurs fonctions.

C'est dans le domaine de la recherche biomédicale que se révèle directement tout le potentiel de l'élucidation du génome humain. En effet, disposer de la séquence de l'ensemble des gènes permet d'envisager la mise en évidence de tous les gènes impliqués dans des maladies génétiques. L'étude de leur fonction et des effets de leurs variations au regard de l'ensemble des données mises à disposition par l'intermédiaire du HGP et des projets dérivés devraient permettre, à terme, d'accélérer le développement de nouvelles solutions thérapeutiques.

De nombreux projets se sont par conséquents concentrés sur les variations du génome humain. Le projet international HapMap (pour *Haplotype Map*) (Consortium, 2003), initié en 2002, a permis la réalisation d'un catalogue des variations génétiques les plus fréquentes chez l'homme, complétant ainsi les données du HGP. La nature des variants y est décrite,

ainsi que leur position dans le génome humain, leur distribution au sein d'une population et entre les populations dans différentes parties du monde (Europe, Asie, Afrique). La carte d'haplotypes du génome humain a notamment permis d'accélérer la découverte de gènes associés à des maladies communes (asthme, cancer, diabète, maladies cardiovasculaires) et de favoriser la création d'outils de diagnostic et le choix de cibles thérapeutiques. Les conséquences directes de la complémentarité des données du HGP et de HapMap peuvent être illustrées par les progrès réalisés dans l'identification de gènes responsables de maladies monogéniques qui représentait, il y a quelques années encore, un travail herculéen, monopolisant une équipe complète de chercheurs durant plusieurs années, sans garantie de succès. Aujourd'hui, grâce à l'évolution des techniques expérimentales, à la disponibilité de cartes génétiques précises et au développement d'outils bioinformatiques et statistiques adaptés, ce travail d'identification est devenu routinier et peut être accompli en quelques semaines par une seule personne (Collins *et al.*, 2003).

Des banques spécifiques regroupant des variations pathogènes ont également vu le jour. Ainsi la HGMD (Stenson *et al.*, 2003), *Human Gene Mutation Database*, par exemple, regroupe les données de mutations associées à l'apparition de maladies génétiques humaines recensées dans la littérature et apporte un solide complément à OMIM (McKusick, 2007), *Online Mendelian Inheritance in Man*, banque de données de référence en matière de gènes impliqués dans des maladies humaines et les phénotypes associés.

L'association HGVS (*Human Genome Variation Society*, http://www.hgvs.org/) s'est quant à elle donnée pour mission de promouvoir non seulement la collecte, la documentation mais aussi la libre distribution d'informations sur les variations génomiques. Elle a également pris en charge la mise en place d'une nomenclature officielle relative à leurs référencements.

Si des standards, voire des ontologies, ont d'ores et déjà été mis en place pour la gestion des données génomiques et de leurs variations ainsi que pour les données protéomiques, des lacunes sont à déplorées en termes d'homogénéisation et de représentation des données phénotypiques, pourtant essentielles pour appréhender les conséquences d'une mutation. La difficulté principale réside dans la multitude de notions auxquelles le terme phénotype peut se référer, incluant les caractéristiques morphologiques, biochimiques, physiologiques ou comportementales d'un individu. Un projet centré sur le phénome humain, le *Human Phenome Project* (Freimer and Sabatti, 2003), a pour but de clarifier, en termes pratiques, la constitution d'un phénotype et de délimiter les composants phénotypiques du phénome.

Ainsi, les avancées réalisées conjointement en génétique, génomique, protéomique et bioinformatique constituent une nouvelle dimension au cœur de la recherche biologique et biomédicale.

Les biologistes disposent à présent d'un répertoire d'outils d'analyse remarquablement performants, qui s'appuient non seulement sur les données recueillies sur l'homme mais aussi sur l'ensemble des autres organismes vivants. En l'espace de 50 ans à peine, la recherche en biologie est donc passée de l'étude d'un seul gène à l'étude de génomes et de protéomes d'organismes complets et l'analyse des organismes à un niveau de détail moléculaire impressionnant est dorénavant réalisable.

Les efforts particuliers réalisés en termes d'étude et de collecte d'informations relatives aux variations du génome humain offrent la possibilité de mieux appréhender leurs conséquences potentielles et d'expliquer quels mécanismes provoqueront ou favoriseront l'apparition d'une maladie. Grâce à la somme d'informations disponibles au travers des diverses banques de données et au développement consubstantiel de nouveaux outils d'analyse *in silico*, il devient envisageable de mieux comprendre la relation ancestrale qui lie le génotype de l'homme à son phénotype en analysant, de manière globale, le chemin allant de la mutation et son réseau de conséquences aux informations phénotypiques associées.

Pour mieux comprendre l'ensemble des paramètres qui entrent en jeu dans cette relation et en saisir au mieux la complexité, la suite de cette introduction sera divisée en quatre parties. Nous nous concentrerons dans le prochain chapitre sur la composition et la structure du génome humain ainsi que sur les variations qui l'affectent au cours du temps. Nous consacrerons le Chapitre 3 aux maladies génétiques humaines avant de présenter dans le Chapitre 4 les conséquences potentielles des mutations sur les protéines et des méthodes permettant leur analyse. Finalement, nous nous focaliserons dans le Chapitre 5 sur la prédiction des effets délétères des mutations avant d'introduire les travaux réalisés dans le cadre de cette thèse.

Chapitre 2 - Le génome humain et ses variations

Ce qui, de tout temps, est apparu comme particulièrement incroyable aux yeux de l'homme, est, d'une part, la formidable ressemblance des êtres humains entre eux, au niveau de leur apparence au sens large et, d'autre part, leur incroyable diversité, faisant de chaque être humain un être unique. C'est au travers de l'intérêt des hommes pour les études de la transmission et de la variabilité de caractères physiques qu'est née la génétique, science qui répond à des préoccupations très anciennes concernant la transmission « familiale » de caractères.

La génétique moderne remonte aux travaux de Mendel dans les années 1860, qui fut le premier à établir les lois de l'hérédité (Mendel, 1866). Ce dernier comprit, par des observations réalisées sur les pois (*Pisum sativum*), qu'un caractère héréditaire pouvait exister sous différentes versions, les unes dominantes, les autres récessives. Il énonça les lois de la transmission de certains traits héréditaires. Comme nous l'avons évoqué dans le Chapitre 1, les avancées réalisées depuis en termes de compréhension du vivant et de son évolution sont absolument incroyables et nous savons notamment que chaque individu est porteur du matériel génétique qui le caractérise, par l'intermédiaire de son génome.

Dans ce chapitre, nous nous intéresserons dans un premier temps à la constitution globale du génome humain, pour nous concentrer dans un deuxième sur sa variabilité.

2.1. Le génome humain

Le génome humain est constitué de plus de trois milliards de paires de bases nucléotidiques, qui s'assemblent en 22 paires de chromosomes autosomes et une paire de chromosomes sexuels (XX chez la femme et XY chez l'homme). Les deux chromosomes d'une paire sont dits homologues, à l'exception de la paire XY. Les bases nucléotidiques sont au nombre de quatre : deux bases puriques, l'adénine et la guanine (respectivement A et G) et deux bases pyrimidiques, la thymine et la cytosine (respectivement T et C) qui ont la particularité de s'assembler par paires (A:T et G:C) (Figure 4).

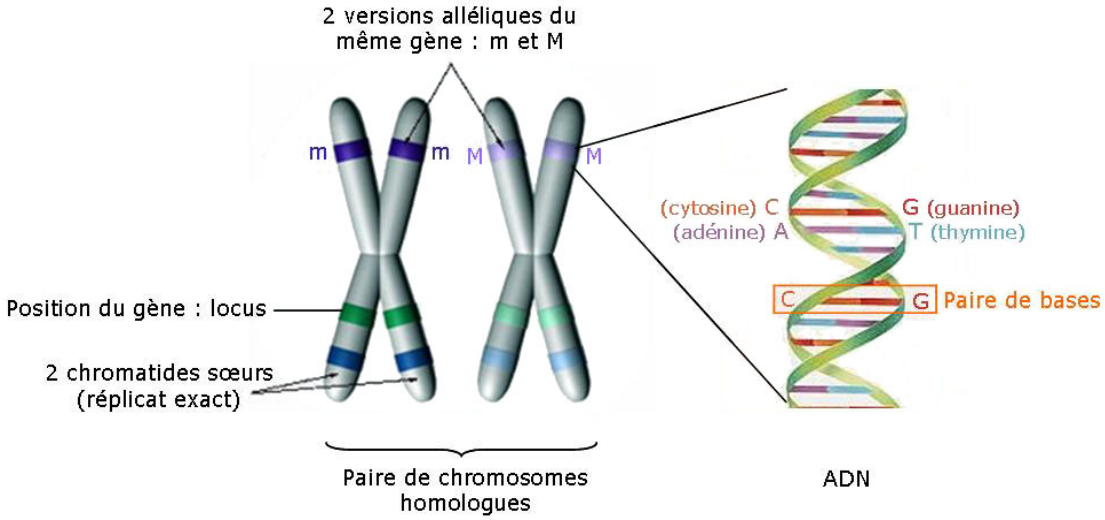


Figure 4 - Représentation d'une paire de chromosomes homologues.
(adaptée de http://www.stanford.edu/group/hopes/basics/dna/f_b11homolgs.jpg et http://www.labolife.info/fr/Labolife.php?page=0)

Dans les mois qui précédaient la publication de la séquence du génome humain, une estimation du nombre de gènes entre 80 000 et 100 000 était encore acceptée (Pennisi, 2000). Une fois la séquence complète du génome disponible, la première surprise fut de découvrir à quel point ce nombre était surestimé. En effet, celui-ci a été revu à la baisse et a été envisagé aux alentours de 30 000 lors de l'analyse de la première version du génome, pour être aujourd'hui plutôt évalué aux alentours de 20 000 à 25 000 (IHGSC, 2004). Les estimations varient encore d'un scientifique à l'autre (Pennisi, 2007), du fait des différents programmes de prédiction utilisés mais aussi de par la définition même de la notion de gène. En effet, plus les connaissances scientifiques s'élargissent, plus il est difficile de donner une définition communément acceptée d'un gène (Gerstein *et al.*, 2007; Pearson, 2006). La définition classique, et simpliste, d'un gène s'est longtemps référée à une séquence d'ADN qui code pour une protéine. C'était sans compter sur le rôle prépondérant, longtemps sous-estimé, des différentes formes d'ARN et la découverte de l'importance et de la diversité des ARN non codants (ARN de transfert, ARN ribosomaux, ARN interférents, micro ARN, etc. (revue dans (Mattick and Makunin, 2006))) qui a bouleversé les principes établis sur les processus fondamentaux de la cellule.

La fraction codante du génome humain ne représente qu'une part très réduite de ce dernier, estimée à 1,2% du génome euchromatique (ensemble des régions peu condensées de la chromatine, actives pour la transcription) (IHGSC, 2004).

Nous décrirons, dans les sections suivantes, la structure traditionnelle des gènes humains, en nous focalisant sur les gènes protéiques, la régulation de leur expression, le produit de leur expression et nous aborderons enfin les rôles que peuvent assurer les régions intergéniques au sein du génome.

2.1.1. Structure des gènes protéiques humains

La structure interne d'un gène protéique humain est relativement complexe. Elle est qualifiée de discontinue, du fait que la séquence transcrite soit constituée d'exons et d'introns (Figure 5). Les introns font partie intégrante des segments d'ADN transcrits par l'ARN polymérase, mais les transcrits primaires subissent ensuite une étape de maturation. Celle-ci comprend à la fois la mise en place d'éléments de protection contre la dégradation de l'ARNm (Ross, 1995; Wang *et al.*, 2002) par l'intermédiaire d'une coiffe en 5' et d'une queue polyA en 3' et l'épissage, qui permet l'élimination des introns grâce à la présence de sites donneurs et accepteurs à leurs extrémités. On peut noter que les exons contiennent également des régions non codantes aux extrémités 5' et 3' de l'ARNm, les régions 5' UTR et 3' UTR (pour *UnTranslated Region*). Seule la région centrale, encore appelée région codante ou CDS (pour *CoDing Sequence*), sera traduite.

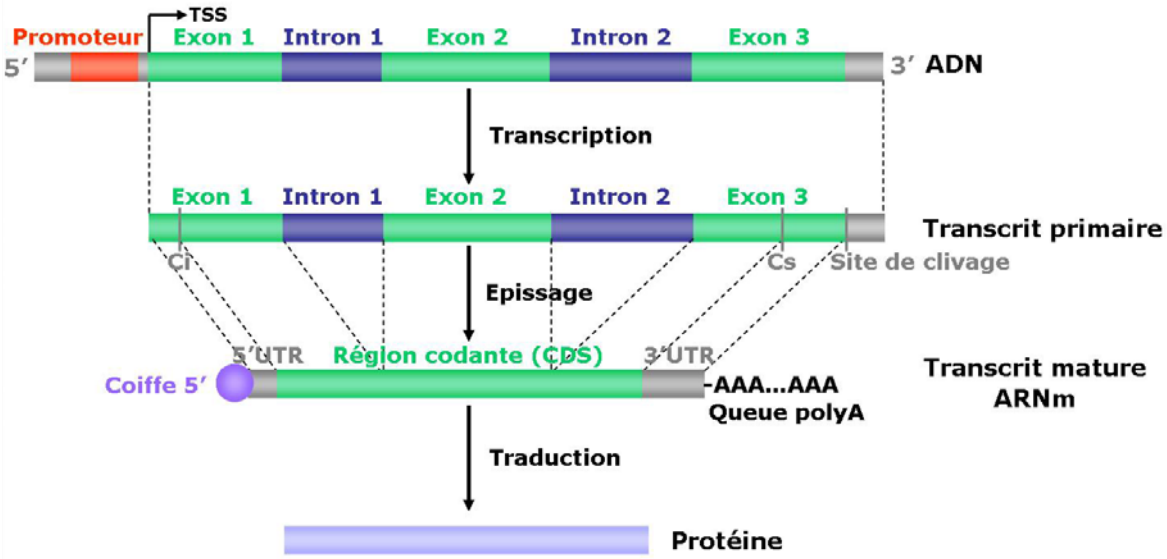


Figure 5 - Structure du gène humain : de l'ADN génomique à la protéine.

Un degré supérieur de complexité est ajouté aux gènes de nombreux eucaryotes par l'épissage alternatif qui consiste à manipuler, pour un gène donné, son répertoire d'exons afin d'obtenir des ARNm différents. Ce processus permet d'augmenter le nombre de protéines codées par un même gène et donc, potentiellement, le nombre de fonctions d'un gène (Figure 6). Il peut avoir lieu soit directement dans le CDS et entraîner la production de protéines différentes, soit dans les régions 5' et 3' UTR, avec comme conséquences possibles une variation de la stabilité, de la régulation, de la localisation, etc. du transcrit mature et, au final, une modification potentielle du niveau de production de la protéine. L'épissage alternatif peut aboutir, à partir d'un seul gène, à une variété impressionnante de protéines (Graveley, 2001) et ce phénomène est loin d'être une exception. Il semble même, au contraire,

être une règle générale puisque, selon diverses estimations, de 50% à 80% des gènes humains possèdent un ou plusieurs variants d'épissage (Johnson *et al.*, 2003; Kampa *et al.*, 2004; Modrek and Lee, 2002).

L'épissage alternatif contribue ainsi à l'augmentation de la diversité protéomique du génome humain et la fréquence élevée de ce phénomène peut apporter une explication à la surestimation initiale du nombre de gènes portés par le génome humain (*voir* section 2.1, p. 15).

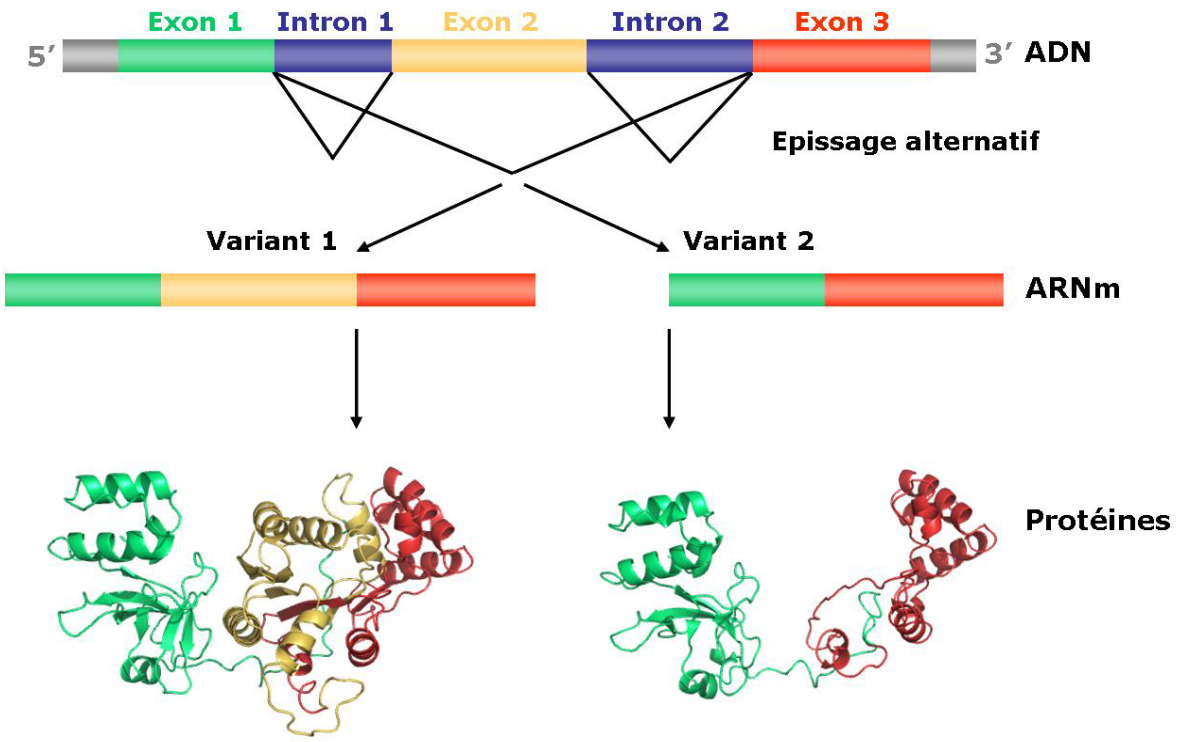


Figure 6 - Exemple d'épissage alternatif. A partir d'un même gène, deux transcrits matures de taille et de composition différentes sont obtenus suite à l'épissage alternatif. Ces ARNm seront ensuite traduits en deux protéines différentes, qui auront potentiellement des fonctions et/ou des partenaires d'interaction différents.

Le promoteur du gène est situé en son extrémité 5', en amont du site d'initiation de la transcription et est impliqué dans la régulation de l'expression du gène (*voir* section 2.1.2, p. 19).

La structure des gènes est très variable tant par leur taille, leur organisation, que par le nombre et la longueur des introns et des exons qui les constituent (Tableau 2). Pour illustration, la taille du plus petit gène codant pour une protéine chez l'homme est de 500 nucléotides. Celui-ci ne contient pas d'intron et code pour une histone. En opposition, le gène humain le plus grand est le gène codant pour la dystrophine : ce dernier est constitué de 2,4 millions de nucléotides et 79 exons.

	Génome humain : 21 667 gènes codants identifiés			
	Total	Minimum	Maximum	Moyenne
Exons	-	2 pb	22 753 pb	293 pb
Introns	-	2 pb	1 033 509 pb	5 675 pb
Gènes à 1 exon	1 869	-	-	-
Transcrits à 1 exon	4 094	-	-	-
Exons par gène	-	0	316	8
Exons par transcrit	-	1	312	9

Tableau 2 - Génome humain : quelques statistiques. Le nombre de gènes identifiés a été extrait du site web d'Ensembl : http://www.ensembl.org/Homo_sapiens/index.html, les statistiques sont issues du NCBI (assemblage 36, version 2) : http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=36&ver=2.

2.1.2. Expression des gènes humains

Les gènes impliqués dans les mêmes processus biologiques se trouvent en des endroits disparates du génome, souvent dans des chromosomes différents. Ceci implique la mise en jeu de mécanismes très élaborés, pour l'obtention d'une expression coordonnée de ces gènes, qui s'effectue principalement au niveau de la transcription des gènes.

Très schématiquement, les régions en amont de la séquence transcrite du gène sont responsables de la régulation de l'expression de ce gène. Celle-ci variera en fonction du type cellulaire, des conditions environnementales, de l'état de la cellule par rapport au cycle cellulaire ou au stade de développement. Les différentes séquences responsables de la régulation de l'expression d'un gène répondent à des signaux intracellulaires spécifiques, de façon à ce qu'idéalement le gène ne soit transcrit qu'au moment et à l'endroit appropriés. Ces séquences de régulation de l'expression sont appelées régions promotrices ou régulatrices (*enhancer* ou *silencer*). Les facteurs de transcription permettent l'activation ou la répression de l'expression d'un gène en se fixant spécifiquement sur de courtes séquences d'ADN, les sites de liaison aux facteurs de transcription, situées au sein de ces régions régulatrices. C'est la présence de ces sites, mais aussi l'ordre de leur agencement et la distance les séparant qui contribuent à la spécificité d'expression d'un gène.

2.1.3. Les protéines : produit de l'expression de gènes

Si les gènes représentent l'unité informationnelle génétique, les protéines représentent les unités fonctionnelles majeures. Ces dernières peuvent être classées selon leur fonction biologique et incluent : les enzymes, responsables de la catalyse de milliers de réactions chimiques au cœur des cellules ; les protéines de structure comme la tubuline, la kératine ou le collagène ; les protéines de transport à l'exemple de l'hémoglobine ; les protéines de régulation telles que les facteurs de transcription ; les molécules de signalement comme les

hormones et leurs récepteurs ; les protéines du système immunitaire ; les protéines assurant un rôle mécanique telles que l'actine et la myosine.

Une séquence protéique est définie par la succession d'acides aminés qui la constitue, mais, schématiquement, c'est sous une forme tridimensionnelle précise qu'elle est active. En effet, ce repliement complexe positionne à proximité l'un de l'autre, dans un arrangement d'une extrême précision, les atomes de quelques acides aminés souvent éloignés dans la séquence, afin de former les sites efficients de la protéine : site catalytique, site de liaison d'antigène, site de reconnaissance d'un récepteur, site d'interaction entre deux protéines, etc. Le repliement de la protéine peut être considéré comme un processus hiérarchique, durant lequel la séquence s'organise en quatre niveaux successifs (Figure 7).

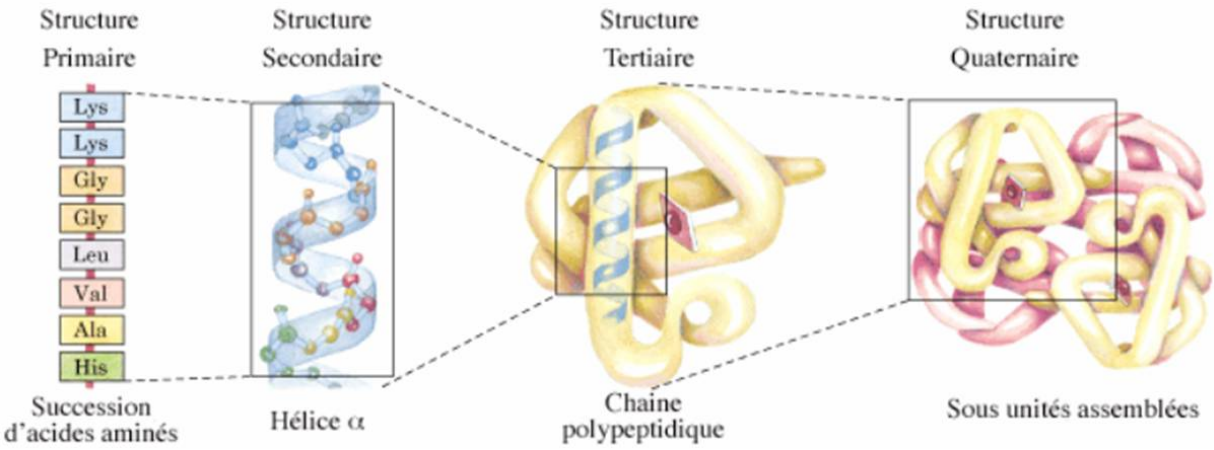


Figure 7 - Repliement des protéines selon les quatre niveaux de structuration.

La séquence des acides aminés liés dans la chaîne polypeptidique constitue la structure primaire de la protéine. La structure secondaire est un premier niveau de repliement, adopté par des portions de la protéine, résultant d'interactions entre des acides aminés voisins sur la chaîne. Deux motifs de repliement caractéristiques peuvent ainsi se former : les hélices α et les feuillets β, réunis par des boucles ou des demi-tours. La structure tertiaire correspond au repliement de la protéine dans l'espace, à l'agencement des éléments de structures secondaires entre eux et à l'organisation spatiale des chaînes latérales. Enfin, certaines protéines sont constituées de plusieurs chaînes polypeptidiques, ou sous-unités, qui s'arrangent spatialement en une structure quaternaire.

2.1.4. Régions intergéniques

La proportion de régions intergéniques du génome humain est actuellement estimée à 97% (IHGSC, 2004). Une grande partie de ces régions intergéniques est constituée d'ADN satellite des régions centromériques et télomériques, mais aussi d'autres séquences répétées et de pseudogènes.

On distingue deux types majeurs de séquences répétées : les séquences répétées en tandem et les séquences répétées dispersées.

Les séquences répétées en tandem comprennent les minisatellites, constitués de répétitions de motifs, dont la taille peut varier de 14 à 100 nucléotides et les microsatellites dont la taille du motif varie de 1 à 13 nucléotides. Ces derniers sont fréquemment polymorphes et distribués uniformément le long du génome, c'est pourquoi ils sont souvent utilisés comme marqueurs génétiques et servent en particulier à l'établissement de cartes génétiques.

Les séquences répétées dispersées sont de plus grande taille. Les deux familles les plus connues sont les SINE (*Small Interspersed Elements*), dont la taille varie entre 130 et 500 paires de bases (pb), et les LINE (*Long Interspersed Elements*), dont la taille est de quelques kilobases (kb). Ces éléments sont répartis sur l'ensemble du génome. Les séquences ALU, qui constituent une des familles les plus abondantes de SINE, sont, par exemple, connues pour favoriser les recombinaisons anormales dans le génome humain et seraient ainsi à l'origine de nombreuses pathologies (Deininger and Batzer, 1999).

L'abondance des séquences répétées pose la question de leur rôle dans l'expression et l'évolution des génomes. Plusieurs études (revue dans (Makalowski, 2000)) ont montré que ces éléments n'étaient pas aussi neutres qu'on pourrait le penser et qu'en plus de représenter des points chauds de recombinaison, ils constituent une source importante de motifs intervenant aussi bien dans la régulation de la transcription que dans les signaux de polyadénylation.

Les pseudogènes sont quant à eux des copies endommagées et souvent, mais pas toujours, inactives de gènes existants. Ils peuvent s'assimiler aux fossiles de notre génome (Mighell *et al.*, 2000; Vanin, 1985).

Ainsi, notre vision des régions intergéniques, longtemps considérées comme étant principalement constituées de « *junk DNA* », a grandement évolué au cours des 20 dernières années et même si les évidences du rôle primordial qu'elles sont amenées à jouer s'accumulent, il reste encore aujourd'hui beaucoup de chemin à parcourir avant de mesurer l'ampleur réelle de leur importance, nos connaissances étant encore relativement limitées (Makalowski, 2003).

2.2. Variabilité du génome et évolution

Chaque être humain est unique, c'est pourquoi nous ne devrions pas parler du génome humain mais plutôt des génomes humains. En effet, à l'exception des « vrais » jumeaux, chaque individu dispose de son génome propre et l'on considère que le génome de deux individus non apparentés varie en moyenne de 0,1%, toutes ethnies confondues.

La diversité génétique se reflète dans l'existence de variations alléliques, chaque gène pouvant exister sous plusieurs formes. Une variation allélique est traditionnellement décrite comme un polymorphisme si l'on retrouve plus d'un allèle à un certain locus dans la population avec une fréquence supérieure à 1%.

Les principaux mécanismes conduisant à la variabilité des génomes sont le brassage génétique, assuré par la reproduction sexuée, ainsi que les modifications accidentelles de l'ADN. Le brassage inter et intrachromosomique peut notamment être illustré par les chiasmas, ou crossing-over (Figure 8), qui sont des recombinaisons réciproques impliquant un échange, de longueur égale, de séquences homologues entre deux chromosomes d'une même paire. Les modifications accidentelles de l'ADN peuvent être représentées par les mutations et les réarrangements chromosomiques. La majorité de ses évènements surviennent au cours des divisions cellulaires.

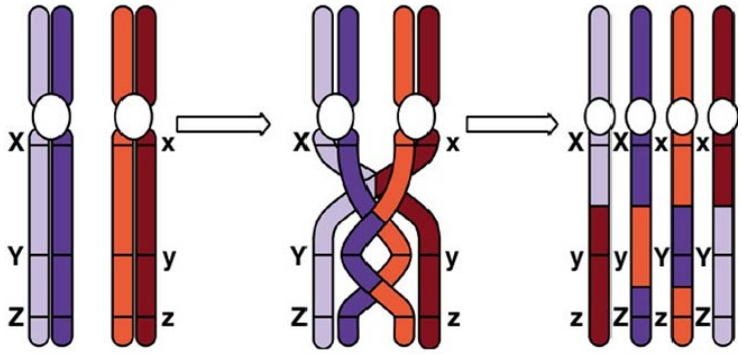


Figure 8 - Echange d'informations génétiques entre chromosomes homologues, par formation de crossing-over au moment de la méiose. Ce processus augmente de manière significative la diversité du contenu informationnel des gamètes, en créant des chromatides uniques.
(adaptée de http://www.uic.edu/classes/bios/bios100/lecturesf04am/lect16.htm)

Ces variations génomiques, qui peuvent apparaître spontanément ou être induites par des facteurs extérieurs, représentent le moteur de l'évolution, mais elles peuvent aussi être associées à l'apparition de maladies génétiques. On peut distinguer les modifications dites germinales, qui affectent les gamètes et sont donc potentiellement transmissibles à la descendance, des modifications somatiques, qui affectent les autres cellules d'un individu et ne sont pas transmissibles d'une génération à l'autre. Ces dernières n'interviennent pas dans le processus de l'évolution.

Les modifications les moins favorables à la survie de l'individu, que l'on considère comme délétères, sont éliminées par le jeu de la sélection naturelle au cours de l'évolution. Les modifications avantageuses, quant à elles, ont tendance à s'accumuler. La plupart des variations sont cependant considérées comme neutres, ne conférant ni avantage ni

inconvénient à l'individu les portant : elles pourront se fixer ou réapparaître par le jeu du hasard.

Nous ferons, dans les prochaines sections, un tour d'horizon des différents types de modifications accidentelles auxquelles est sujet le génome humain, pour nous concentrer ensuite sur ses variations les plus courantes, les mutations. Nous nous intéresserons particulièrement à l'origine de leur apparition ainsi qu'à leurs conséquences, d'une part sur le génome et d'autre part au niveau du produit de l'expression des gènes.

2.2.1. Modifications de l'information génétique

Le génome humain peut être considéré comme une entité dynamique, qui change au cours du temps, à la suite des effets cumulés de lésions locales de la séquence, les mutations, mais aussi du fait de réarrangements plus larges, pouvant engendrer différents types de remaniements chromosomiques.

Les réarrangements chromosomiques entraînant une anomalie du nombre de chromosomes sont généralement liés à une mauvaise ségrégation des chromosomes au moment de la première division méiotique à l'origine de la formation des gamètes. En effet, au cours de cette première division cellulaire, une cellule diploïde, contenant deux copies de chaque chromosome, donne naissance à deux gamètes haploïdes, comprenant chacun un chromosome de chaque paire. Lorsque les deux chromosomes d'une même paire migrent vers la même cellule fille, un des gamètes formés aura deux copies d'un chromosome et son complémentaire aucune : on parle d'aneuploïdie du fait que les gamètes ne contiennent pas le nombre attendu de chromosomes. En cas de fécondation, le zygote formé sera lui aussi aneuploïde, vis-à-vis de quoi la nature est très intolérante. La monosomie n'est pas compatible avec la vie, mis à part quand celle-ci concerne les chromosomes sexuels. Certaines trisomies sont quant à elles viables, mais entraînent l'apparition de symptômes sévères, la plus connue étant la trisomie du chromosome 21, associée au syndrome de Down (Lejeune *et al.*, 1959).

Une aneuploïdie peut également apparaître au niveau d'une cellule somatique, suite à l'absence de disjonction d'un couple de chromosomes au moment de la mitose. Ces cellules aneuploïdes perdent généralement leur capacité à se diviser et le défaut n'est donc pas propagé. Il arrive cependant qu'il soit à l'origine de cycles anormaux de division cellulaire conduisant à l'apparition de cancers.

Les réarrangements chromosomiques entraînant une anomalie de structure des chromosomes peuvent concerner un ou plusieurs chromosomes et avoir lieu avant une division cellulaire. Les principaux types de remaniements (Figure 9) nécessitent une double cassure du brin d'ADN. Ils seront dits équilibrés s'ils n'entraînent pas la perte de matériel génétique et déséquilibrés dans le cas contraire. De manière générale, les remaniements

équilibrés n'ont pas de conséquence sur le sujet porteur, alors que les remaniements déséquilibrés se traduisent par des manifestations cliniques d'autant plus sévères que la perte ou le gain de matériel est plus important.

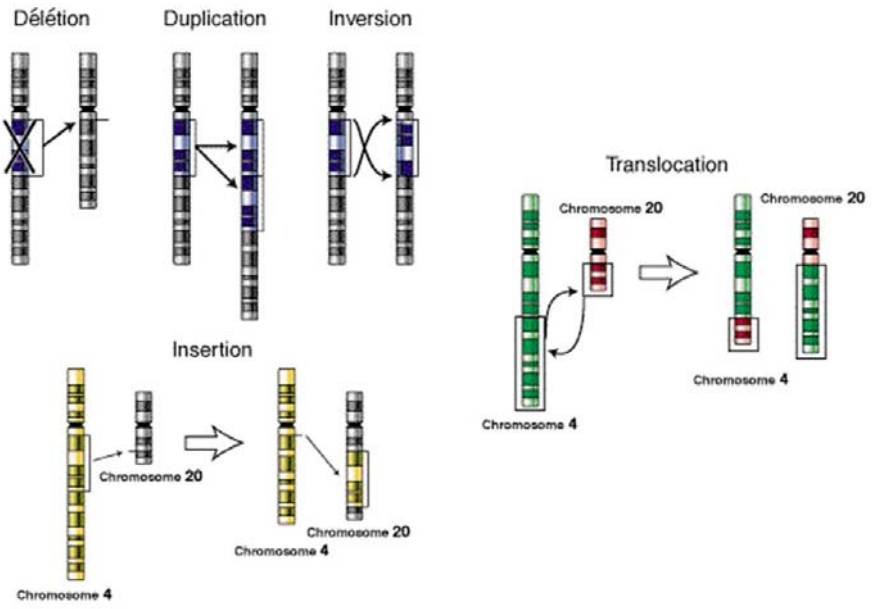


Figure 9 - Remaniements chromosomiques entraînant une anomalie de structure.
(adaptée de http://en.wikipedia.org/wiki/Image:Types-of-mutation.png)

2.2.2. Origine de l'apparition des mutations

Les mutations n'entraînant pas de remaniement chromosomique majeur peuvent être spontanées, si elles apparaissent naturellement au cours de la réplication du génome, ou induites par des agents externes, encore appelés agents mutagènes.

2.2.2.1. Mutations spontanées

Avant chaque division cellulaire, le génome humain et ses trois milliards de nucléotides doivent être copiés de manière précise dans leur intégralité, ce qui représente un véritable défi. La réplication du génome est principalement assurée par l'ADN polymérase, un complexe enzymatique qui se trouve être d'une remarquable efficacité. Au cours de la réplication, chaque brin parental d'une molécule d'ADN sert de matrice pour la production d'un nouveau brin, lui-même déterminé par la complémentarité des bases. Chaque nouvelle molécule d'ADN double brin est constituée d'un brin parental et d'un brin néoformé (Figure 10) : on parle alors de réplication semi-conservative.

Bien que d'une incroyable fidélité, l'ADN polymérase introduit des erreurs lors de la copie du matériel génétique. Une grande partie de ces erreurs est corrigée immédiatement par des mécanismes complexes et efficaces de réparation de l'ADN (Kunkel, 2004; Umar and Kunkel,

1996), notamment grâce à l'activité exonucléolytique associée à la polymérase, encore appelée édition, ainsi qu'à un système post-réplicatif de réparation des mésappariements qui corrige les erreurs de réplication ayant échappé à l'édition (de Wind and Hays, 2001).

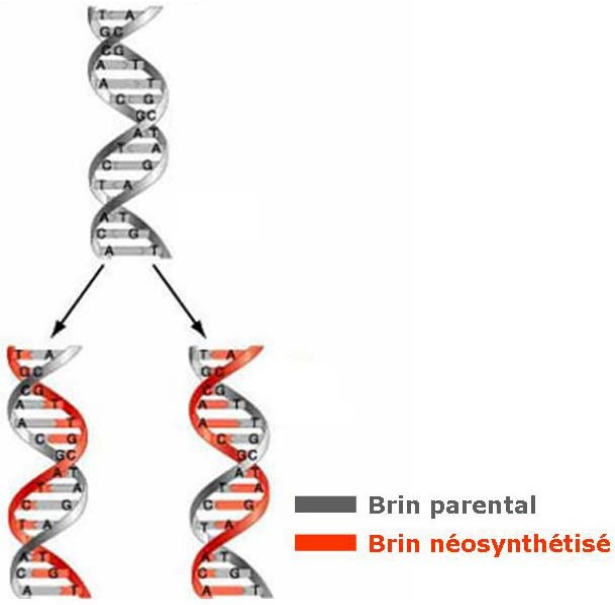


Figure 10 - Réplication semi-conservative du génome. L'ADN polymérase se sert de chacun des brins de l'ADN parental comme matrice pour la synthèse de deux molécules d'ADN double brin, constituée chacune d'un brin parental associé à un brin néosynthétisé.
(adaptée de http://fig.cox.miami.edu/~cmallery/150/gene/mol_gen.htm)

Un défaut dans un gène du système de réparation de l'ADN peut avoir de graves conséquences sur l'individu et peut, par exemple, provoquer l'apparition d'une hypersensibilité de la peau aux rayons ultraviolets, appelée *xeroderma pigmentosum* (Bootsma and Hoeijmakers, 1991), conduisant à des cancers cutanés.

La fréquence d'apparition des mutations spontanées chez l'homme est estimée à environ un nucléotide tous les 10^9 nucléotides répliqués (Meyers, 2005), reflétant l'efficacité du système dans sa globalité.

Il faut noter que la fidélité de la réplication de l'ADN est également fonction de la nature de la séquence d'ADN à répliquer : en effet, les mutations spontanées présentent un biais important dans leur apparition, tant au niveau de la distribution des sites qu'à celui du type d'altération causée au sein de la séquence génomique (Drake, 1991; Maki, 2002). Les microsatellites par exemple, très utilisés en cartographie génétique, constituent une fraction particulière du génome, avec un taux de mutation plus élevé que le reste du génome. De même, le système immunitaire humain est un exemple surprenant de point chaud de mutations bénéfiques au cœur des cellules somatiques, on parle même d'hypermutabilité dans ce cas, qui semble être la conséquence du comportement inhabituel du système de

réparation des mésappariements de l'ADN (Casali *et al.*, 2006). Le taux de mutation observé est de six à sept ordres de magnitude plus élevée et un vaste répertoire d'anticorps émerge à partir de ce processus. Dans ce cas précis, la cellule utilise de manière favorable les mutations générées.

2.2.2.2. Mutations induites

Dans certaines circonstances, le taux de mutations peut être considérablement augmenté par l'intervention d'agents environnementaux, appellés agents mutagènes. Ces derniers peuvent être des facteurs physiques (rayons ultraviolets, radiations ionisantes, etc.) ou chimiques (agents désaminants, alkylants, etc.). Ils sont le plus souvent des agents cancérigènes et nocifs du fait de leur présence non contrôlée, et parfois non contrôlable, dans l'environnement. Ils agissent généralement sur un brin de l'ADN parental, produisant un changement de structure qui affecte la complémentarité du nucléotide altéré.

2.2.3. Effets des mutations sur le génome

Quelque soit l'origine de l'apparition d'une mutation, elle est susceptible de provoquer à elle seule une cascade de modifications pouvant entraîner des changements importants à plusieurs niveaux, de la séquence génomique aux traits physiques de l'individu.

Nous allons à présent détailler les effets directs d'une mutation sur la séquence du génome les plus connus.

2.2.3.1. Substitution

Une substitution caractérise une mutation ponctuelle, qui se traduit par le changement d'un nucléotide par un autre. Parmi les substitutions, on peut distinguer les transitions (modification d'une base purique A ou G par une base purique ou d'une base pyrimidique C ou T par une base pyrimidique), des transversions (modification d'une base purique par une base pyrimidique ou l'inverse) (Figure 11).

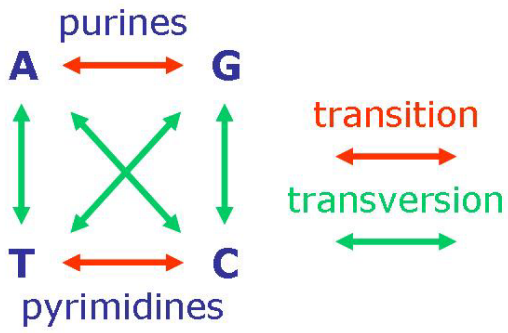


Figure 11 - Mutations ponctuelles au niveau de l'ADN : les possibilités de substitutions des quatre bases nucléotidiques.

Dans la nature, le taux de transitions observé est environ deux fois supérieur à celui des transversions (Wakeley, 1994). Si les phénomènes de substitutions devaient se produire de manière aléatoire, du fait qu'il y a deux fois plus de possibilités théoriques de transversion que de transition, nous devrions rencontrer des proportions inverses. Ceci s'explique, d'une part, par le fait qu'au cours de la réplication, l'ADN polymérase incorpore plus fréquemment une base erronée de même nature chimique et, d'autre part, par les propriétés des mécanismes de correction, qui reconnaissent plus facilement une transversion qu'une transition, la première engendrant une anomalie de la double hélice plus prononcée.

De plus, le dinucléotide CpG est ce que l'on appelle un « point chaud » de mutation (Cooper and Krawczak, 1989). En effet, la méthylation de l'ADN intervient préférentiellement en position 5 des cytosines précédant les guanines. Or, les cytosines et les cytosines méthylées subissent, avec une certaine fréquence, des événements de désamination qui les transforment respectivement en uraciles et en thymines. À la différence du mésappariement U:G, très efficacement réparé par un complexe enzymatique spécifique, la réparation du mésappariement T:G est beaucoup plus problématique, puisque la thymine est un composant normal de l'ADN. Par conséquent, en raison du caractère palindromique du dinucléotide CpG et de sa méthylation, la réparation des mésappariements T:G peut conduire à des transitions CpG vers TpG ou CpA, qui représentent environ 35% de l'ensemble des mutations ponctuelles.

Les substitutions d'un seul nucléotide sont couramment appelées SNP, pour *Single Nucleotide Polymorphism*. Ces « polymorphismes » peuvent en théorie être bi, tri ou tétra allélique, mais on retrouvera en pratique préférentiellement deux nucléotides différents à une même position d'un gène. Notons que le terme polymorphisme est dans ce cas souvent utilisé par abus de langage, car employé sans tenir compte de la fréquence de la modification au sein de la population.

Les SNP représentent la source de variation la plus courante au sein du génome humain avec une représentativité d'environ 90% (Collins *et al.*, 1998). Il a été estimé que 93% des gènes humains contiennent au moins un SNP et que 98% des gènes sont à proximité (+/- 5 kb) d'un SNP (Chakravarti, 2001).

2.2.3.2. Insertion et délétion de petite taille

L'ajout d'un ou plusieurs nucléotides est appelé insertion ; la perte d'un ou plusieurs nucléotides est appelée délétion.

Les insertions/délétions, bien que susceptibles de se produire tout le long du génome, sont particulièrement fréquentes au niveau des séquences répétées. En effet, au moment de la duplication du génome, celles-ci peuvent être à l'origine d'un « glissement de réplication », au cours duquel le brin matrice, ou sa copie, peuvent dévier de leurs positions relatives en

formant des boucles d'ADN simple brin par appariement de bases complémentaires de sorte qu'une partie de la matrice sera omise ou copiée deux fois selon le cas (Figure 12). Ce phénomène explique, par exemple, la grande variabilité des séquences des microsatellites par la création de nouveaux variants qui s'ajoutent à la collection des allèles déjà présents au sein de la population.

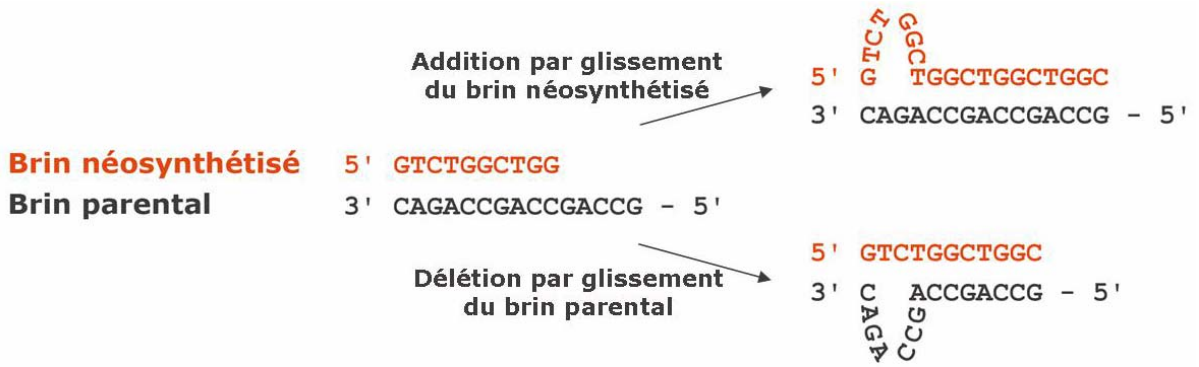


Figure 12 - Glissement de réplication. Lors de la réplication, un glissement affectant le brin parental peut engendrer une délétion sur le brin néoformé, alors qu'un glissement de ce dernier peut être à l'origine d'une insertion.

Il arrive également qu'un ou plusieurs nucléotides consécutifs soient remplacés par un ou plusieurs autres nucléotides (le même nombre ou un nombre différent). On parle dans ce dernier cas d'indel, pour désigner l'insertion et la délétion simultanées de nucléotides, dont plusieurs points chauds ont récemment été mis en évidence (Chuzhanova *et al.*, 2003). Bien que la plupart des indels ait été décrite comme étant associée à l'apparition de maladies génétiques chez l'homme, ils peuvent occasionnellement représenter des variants polymorphiques au sein de la population (Fernie and Hobart, 1997).

2.2.3.3. Autres mutations

Les insertions/délétions peuvent avoir une taille plus importante, sans qu'on parle directement de remaniement chromosomique. Il arrive en effet qu'une délétion de grande taille entraîne la perte d'un exon ou d'un gène entier. Ces mutations restent cependant relativement peu fréquentes, excepté pour les gènes liés au chromosome X, où ce genre de remaniement est plus fréquent et peut affecter de 5 à 95% de la région codante d'un gène. Dans le cas de la myopathie de Duchenne, plus de 70% des mutations répertoriées rapportent la perte d'un ou plusieurs exons (Aartsma-Rus *et al.*, 2006) dans le gène de la dystrophine.

2.2.4. Conséquences des mutations selon leur localisation dans le génome

Comme nous l'avons déjà souligné, une mutation dans le génome peut n'avoir aucune conséquence et être ce qu'on appelle une mutation neutre, dont l'existence ne sera même pas soupçonnée du fait de l'absence de modification de trait associée. Une faible proportion des mutations survenant dans le génome a cependant des répercussions au niveau du produit de l'expression des gènes, qui peuvent se traduire par des répercussions physiques (modification d'un trait) voire cliniques (apparition de symptômes pathologiques). Nous parlerons de mutations délétères dans ce dernier cas et nous verrons dans les prochaines sections que leurs répercussions sur le produit de l'expression des gènes dépendront de leurs localisations dans le génome et dans le gène.

2.2.4.1. Mutation dans une région intergénique

Lorsque l'unité transcriptionnelle et le promoteur d'un gène sont intacts, son expression peut être altérée par ce qu'on nomme un « effet de position », modifiant son environnement chromatinien habituel. Un réarrangement peut séparer le gène d'éléments régulateurs distants ou, au contraire, placer à proximité d'un gène l'*enhancer* ou le *silencer* d'un autre gène. L'implication d'anomalies moléculaires à l'origine d'effets de position en pathologie humaine est probablement sous-estimée à ce jour, faute de critère simple permettant de déterminer leur caractère causal.

De plus, comme nous l'avons déjà précisé dans la section 2.1.4, p. 20, il est important de rappeler que les connaissances actuelles sur l'ensemble des fonctionnalités représentées au sein des régions intergéniques sont relativement limitées. Nous ne sommes par conséquent actuellement pas à même de mesurer toute l'importance et les répercussions des mutations s'y produisant.

2.2.4.2. Mutation dans une région non transcrite d'un gène

Une mutation localisée dans le promoteur d'un gène peut, par l'intermédiaire de plusieurs mécanismes, affecter l'expression de ce gène. La mutation peut altérer un site consensus de liaison d'un ou plusieurs facteurs de transcription et perturber l'expression génique normale. La mutation du promoteur peut également altérer sa structure, entraînant la modification négative ou positive de l'expression du gène par l'intermédiaire de structures secondaires.

Une mutation peut de la même façon affecter une région régulatrice de type *enhancer* ou *silencer*, souvent tissu-spécifique et modifier l'expression du gène de manière locale ou généralisée. Des mutations affectant une région de contrôle de la régulation des locus de la β-globine, appelée LCR (pour *Locus Control Regions*), ont par exemple été rapportées. Elles laissent les séquences du gène et de son promoteur intactes, mais entraînent une abolition

complète de l'expression de ce gène et sont à l'origine de l'apparition d'une maladie appelée la β-thalassémie (Olivieri, 1999).

2.2.4.3. Mutation dans une région transcrite non codante d'un gène

Les mutations affectant une partie non codante d'un gène auront un effet délétère si elles affectent une région essentielle à l'expression correcte de ce gène, en particulier celles nécessaires au bon déroulement de l'épissage. En effet, une mutation, en modifiant les séquences consensus des sites donneurs et accepteurs d'épissage ou le site de branchement intronique, peut être à l'origine de la suppression d'un exon ou de la rétention d'un intron. Des mutations hors de ces sites particuliers peuvent également dévoiler des sites cryptiques d'épissage, conduisant à un épissage erroné. Ces mutations ne se font pas toujours en phase : elles peuvent entraîner bien plus que la perte/gain d'un exon/intron et être à l'origine d'un décalage du cadre de lecture, qui engendrera la synthèse d'une protéine présentant une partie C-terminale totalement erronée.

Précisons que les mutations entraînant l'abolition de sites d'épissage ne sont pas exclusivement situées au sein des introns, mais peuvent également affecter des positions essentielles au sein des exons. Récemment, la description de nouvelles séquences consensus exoniques, les ESE (*exonic splicing enhancers*) et les ESS (*exonic splicing silencers*), favorisant ou réprimant l'épissage de l'exon qui les contient, a permis de comprendre pourquoi des modifications de la séquence codante d'un gène peuvent être responsables de l'épissage complet de l'exon où elles sont positionnées (Cartegni *et al.*, 2002).

La proportion des mutations entraînant des défauts d'épissage a été estimée à environ 10% des mutations délétères totales affectant des gènes humains (Krawczak *et al.*, 2007), mais ce chiffre est sans doute sous-estimé, car seules les mutations affectant les sites donneurs et accepteurs d'épissage sont prises en compte.

De même, une mutation du signal de polyadénylation peut également perturber la maturation de l'extrémité 3' de l'ARN et être responsable d'une diminution de la stabilité du transcrit.

2.2.4.4. Mutation dans une région codante d'un gène

Au moment de la traduction, la suite des acides aminés qui vont constituer la protéine est déduite de l'enchaînement des triplets de nucléotides, encore appelés codons, de l'ARNm par la relation unique qui existe entre un codon donné et son acide aminé correspondant, selon la règle édictée par le code génétique universel (Figure 13). 64 combinaisons des quatre nucléotides A, C, U et G en triplets sont possibles et codent pour 20 acides aminés naturels. On parle de la dégénérescence du code génétique du fait que plusieurs codons correspondent au même acide aminé. C'est souvent le troisième nucléotide d'un codon qui

peut être interchangé sans modification de l'acide aminé correspondant : ce phénomène est appelé le « *wobble* ». De plus, trois triplets ne codent pour aucun acide aminé et représentent ce qu'on appelle les « codons stop », qui indiquent la fin de la protéine.

1ère position	2ème position												3ème position
	U			C			A			G			
U	UUU	Phe (F)	Phénylalanine	UCU	Ser (S)	Sérine	UAU	Tyr (Y)	Tyrosine	UGU	Cys (C)	Cystéine	U
	UUC	Phe (F)		UCC	Ser (S)		UAC	Tyr (Y)		UGC	Cys (C)		C
	UUA	Leu (L)	Leucine	UCA	Ser (S)		UAA	Stop		UGA	Stop		A
	UUG	Leu (L)		UCG	Ser (S)		UAG			UGG	Trp (W)	Tryptophane	G
C	CUU	Leu (L)		CCU	Pro (P)	Proline	CAU	His (H)	Histidine	CAU	Arg (R)	Arginine	U
	CUC	Leu (L)		CCC	Pro (P)		CAC	His (H)		CGC	Arg (R)		C
	CUA	Leu (L)		CCA	Pro (P)		CAA	Gln (Q)	Glutamine	CGA	Arg (R)		A
	CUG	Leu (L)		CCG	Pro (P)		CAG	Gln (Q)		CGG	Arg (R)		G
A	AUU	Ile (I)	Isoleucine	ACU	Thr (T)	Thréonine	AAU	Asn (N)	Asparagine	AGU	Ser (S)	Sérine	U
	AUC	Ile (I)		ACC	Thr (T)		AAC	Asn (N)		AGC	Ser (S)		C
	AUA	Ile (I)		ACA	Thr (T)		AAA	Lys (K)	Lysine	AGA	Arg (R)	Arginine	A
	AUG	Met (M)	Méthionine	ACG	Thr (T)		AAG	Lys (K)		AGG	Arg (R)		G
G	GUU	Val (V)	Valine	GCU	Ala (A)	Alanine	GAU	Asp (D)	Acide aspartique	GGU	Gly (G)	Glycine	U
	GUC	Val (V)		GCC	Ala (A)		GAC	Asp (D)		GGC	Gly (G)		C
	GUA	Val (V)		GCA	Ala (A)		GAA	Glu (E)	Acide glutamique	GGA	Gly (G)		A
	GUG	Val (V)		GCG	Ala (A)		GAG	Glu (E)		GGG	Gly (G)		G

Figure 13 - Code génétique universel. On remarque un « flottement » au niveau du 3ème nucléotide du codon. Le code universel accepte quelques exceptions, notamment au niveau des mitochondries.

Les conséquences sur le produit de l'expression d'un gène d'une mutation au niveau d'une partie codante de ce gène peuvent être de plusieurs natures.

2.2.4.4.1. Mutation silencieuse

De par la dégénérescence du code génétique, la substitution d'un acide nucléique par un autre peut être sans conséquence au niveau de la séquence en acides aminés de la protéine correspondante : en effet, le nouveau triplet peut tout à fait coder pour la même protéine. On parle dans ce cas de mutation silencieuse (Figure 14).

2.2.4.4.2. Mutation exprimée

Les mutations non silencieuses qui affectent la partie codante d'un gène auront un effet direct sur la séquence de la protéine codée par ce gène, qui pourra avoir ou non des répercussions sur sa fonctionnalité.

Si la mutation observée au niveau de l'ADN est la substitution d'un nucléotide par un autre, on parlera de mutation ponctuelle au niveau de la protéine. On peut classer les mutations ponctuelles en trois catégories (Figure 14) :

- mutations non-sens : le triplet après mutation code pour un des trois codons stop. Il est courant que l'apparition d'un codon stop prématuré soit à l'origine d'un ARNm instable, souvent rapidement dégradé *in vivo*, par une forme de surveillance, la voie de dégénérescence de l'ARNm à médiation non-sens (Culbertson, 1999; Hentze and Kulozik, 1999). Lorsque l'ARNm est suffisamment stable pour être traduit, ces mutations sont à l'origine de la synthèse de protéines tronquées, dont la taille est inférieure à celle de la protéine sauvage. Les conséquences d'une mutation non-sens sont directement liées à la quantité et à la nature de l'information « perdue » ou « gardée »,
- mutations faux-sens : le triplet muté code pour un des 19 autres acides aminés naturels. La protéine mutée aura dans ce cas la même longueur que la protéine sauvage. Certaines de ces mutations sont tout à fait tolérées par la protéine qui continuera à assurer ses fonctions, alors que d'autres peuvent avoir un effet délétère. Les conséquences des mutations faux-sens dépendent principalement de leurs positions au sein de la protéine et des propriétés physico-chimiques des acides aminés impliqués dans la substitution. Nous reviendrons en détails sur les conséquences possibles d'une mutation ponctuelle au niveau de la structure et/ou de la fonction de la protéine dans le Chapitre 4,
- mutation d'un codon stop : la mutation modifie un codon stop en un acide aminé et allonge la taille de la protéine. Les effets sur la structure et la fonction de la protéine dépendront généralement de l'emplacement du prochain codon stop.

Si la mutation observée au niveau de l'ADN est une insertion, une délétion ou un indel, il faut une fois de plus distinguer plusieurs cas :

- si la modification concerne un nombre de nucléotides multiple de trois, l'effet sur la protéine sera le gain ou la perte d'un ou plusieurs acides aminés dans sa séquence et ses conséquences dépendront principalement de l'emplacement des résidus dans la structure. Schématiquement, si une telle mutation a lieu au cœur d'un élément de structure secondaire de la protéine, celle-ci risque d'être fortement déstabilisée par la modification.
- si la modification concerne un nombre de nucléotides non multiple de trois, on assistera à un changement du cadre de lecture de la protéine, plus couramment appelé *frameshift*. La séquence en acides aminés de la protéine est totalement modifiée après l'endroit où s'effectue la mutation et l'apparition d'un codon stop peut se faire à n'importe quel moment : la protéine peut être rapidement tronquée ou tout à fait rallongée. Ces mutations ont souvent des conséquences très lourdes sur la protéine et ses fonctions, à l'exception de celles affectant l'extrémité C-terminale de la protéine qui entraînent l'apparition rapide d'un codon stop.

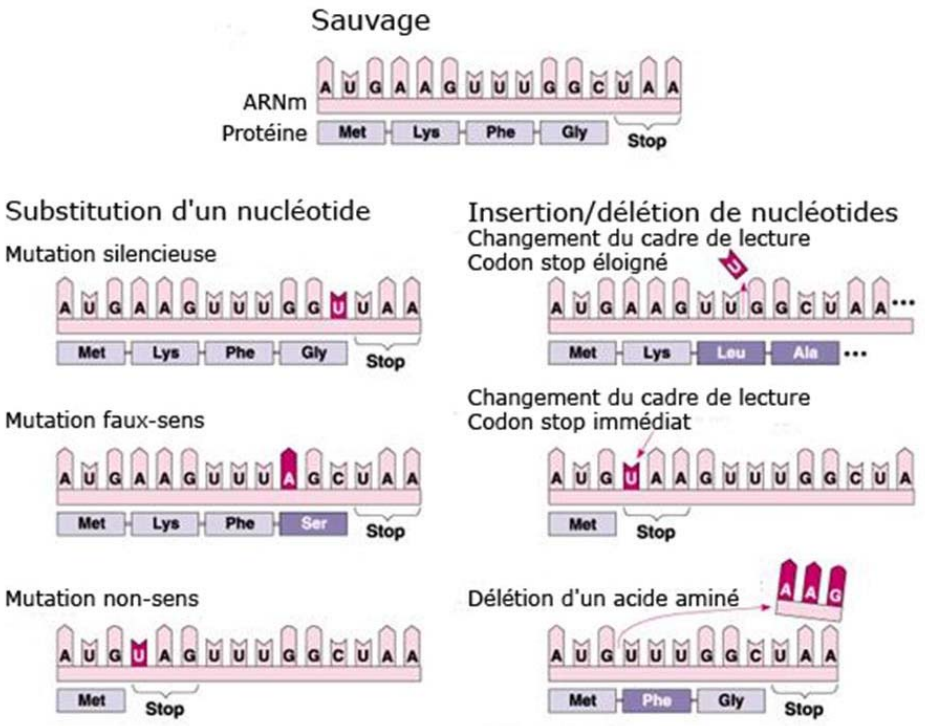


Figure 14 - Conséquences des mutations sur la synthèse de la protéine.
(adaptée de http://fajerpc.magnet.fsu.edu/Education/2010/Lectures/26_DNA_Transcription.htm)

Précisons finalement que, principalement dans le cas des substitutions exprimées ou silencieuses, d'autres effets que ceux cités peuvent être observés, à une fréquence beaucoup plus rare. En effet, une substitution dans une région codante peut être à l'origine de l'apparition d'un site d'épissage auparavant masqué, susceptible de modifier entièrement le produit de l'expression du gène. La substitution peut également modifier la dynamique du temps de demi-vie de l'ARNm, en affectant sa structure secondaire, ou être à l'origine de problèmes traductionnels si le nouveau triplet formé représente ce qu'on appelle un codon rare.

Chapitre 3 - Les maladies génétiques humaines

Comme nous venons de le voir, les mutations sont à la base des fondements de l'évolution, mais elles peuvent également avoir des répercussions pathogènes soit en étant la cause directe de l'apparition d'une maladie, soit en augmentant la susceptibilité d'un individu à une maladie.

Le taux de mutation relativement faible habituellement observé chez l'homme peut donc être considéré comme un équilibre entre l'apparition occasionnelle de « modifications avantageuses », qui auront tendance à s'accumuler au cours de l'évolution, et l'émergence de maladies dans une petite proportion des membres de l'espèce humaine.

Nous allons dans ce chapitre nous intéresser plus précisément à ce qu'est une maladie génétique, à ses modes de transmissions et à la manière dont les variations du génome pourront agir sur le phénotype de l'individu.

3.1. Qu'est-ce qu'une maladie génétique ?

Une maladie génétique peut être définie, dans un premier temps, comme une pathologie, engendrée par une « anomalie » portée par le génome de l'individu, qui sera transmissible à sa descendance.

Schématiquement, on peut caractériser une maladie génétique par :

- son génotype, qu'on peut définir comme étant la cause moléculaire de la maladie,
- son phénotype ou les manifestations « observables » du génotype,
- la relation qui lie le génotype au phénotype.

Bien que ce soit le génotype qui dicte le phénotype, une maladie génétique est avant tout la manifestation d'un ensemble de symptômes. Le phénotype caractérise donc la maladie, dans le sens où il est le premier décelé, mais trouve son origine dans le génotype. Les relations qui lient le génotype au phénotype sont généralement très complexes. De manière surprenante, les manifestations cliniques peuvent être extrêmement différentes selon la nature de la mutation présente dans un même gène. A l'opposé, des pathologies qui apparaissent très spécifiques par leur combinaison unique de symptômes, présentent parfois une hétérogénéité génétique très importante avec, dans certains cas, plusieurs gènes dont les mutations entraînent des maladies cliniquement non distinguables. Un diagnostic préliminaire peut ainsi être établi à partir de l'analyse du phénotype, mais ce dernier doit

être confirmé par la détermination précise du génotype associé à l'apparition des symptômes.

La modification du phénotype d'un individu due à un changement dans son empreinte génétique n'est cependant pas automatiquement associée à une maladie génétique. En effet, un certain nombre de mutations ne vont engendrer que des changements mineurs et font partie intégrante du jeu de l'évolution.

D'un point de vue « philosophique », il peut donc être intéressant de se demander à partir de quand on peut considérer que la modification d'un phénotype est associée à une maladie et non plus à un continuum de la diversité humaine. Il n'y a bien sûr pas de réponse stricte à cette question, relative à la sensibilité de chacun et très bien illustrée par la formule « *One man's trait is another man's disease* » (Hawley and Mori, 2004). La personne la plus à même de juger si un trait s'apparente à une maladie est sans doute la personne chez qui ce trait s'exprime, mais c'est cependant l'environnement dans lequel elle évolue et le regard des autres qui rendra ce trait « handicapant » voire pathologique. Quoi qu'il en soit, lorsqu'une modification de phénotype entraîne des conséquences purement cosmétiques, on ne devrait pas parler de maladie. Nous considérerons le terme de « maladie génétique » comme s'appliquant exclusivement à des traits qui aboutissent à des problèmes médicaux.

De façon schématique, quatre grands groupes de maladies génétiques sont actuellement distinguables : (i) les maladies héréditaires à transmission mendélienne ; (ii) les maladies mitochondriales, dont l'hérédité est dite maternelle du fait que les gènes mitochondriaux sont exclusivement apportés par l'ovocyte au moment de la formation du zygote ; (iii) les maladies par aberration chromosomique ; (iv) les maladies multifactorielles, dont la répartition chez les apparentés ne suit pas les lois de Mendel. Les maladies multifactorielles, encore appelées maladies polygéniques ou à hérédité complexe, sont causées par un ensemble de facteurs génétiques et environnementaux. Notons que les allèles impliqués ne sont, dans la très grande majorité des cas, pas délétères, mais confèrent ce qu'on appelle une susceptibilité accrue à la maladie.

Nous allons principalement nous concentrer dans ce chapitre, et tout au long du manuscrit, sur les maladies héréditaires à transmission mendélienne, qui sont des maladies monogéniques, induites par des défauts dans un seul gène.

3.1.1. Maladies génétiques et phénotypes

Nous avons défini le phénotype comme étant les manifestations observables du génotype. Des précisions de deux ordres sont à apporter à cette définition. D'une part, ces « manifestations observables » sont à considérer à plusieurs niveaux et couvrent des caractéristiques allant du microscopique au macroscopique. En effet, les conséquences d'un génotype particulier aux niveaux moléculaire, cellulaire et tissulaire ainsi que ses influences

physiques et/ou comportementales observables directement sur l'individu font partie intégrante de la description d'un phénotype. Nous avons donc affaire à un ensemble d'effets reliés, décrivant divers degrés d'impacts d'un gène sur un organisme.

D'autre part, il existe une relation complexe entre le génotype, l'environnement et les manifestations phénotypiques, pour laquelle il est extrêmement difficile de mesurer l'influence de chacun. Les facteurs environnementaux peuvent être des facteurs génotypiques, représentés par l'ensemble des relations existant entre les deux allèles de chacun des gènes d'un individu, des facteurs épigénétiques, c'est-à-dire des modifications transmissibles et réversibles de l'expression des gènes ne s'accompagnant pas de changements des séquences nucléotidiques (Jirtle and Skinner, 2007), ou des facteurs extérieurs, comme l'exposition environnementale ou le mode de vie de l'individu. Ces facteurs environnementaux peuvent notamment être à l'origine d'une variabilité interindividuelle des manifestations cliniques pour une même mutation.

Enfin, les phénotypes peuvent être classés en plusieurs catégories en fonction de leur schéma d'émergence. On peut distinguer : les phénotypes congénitaux, qui sont exprimés dès la naissance, comme un déficit de coagulation du sang lié à l'hémophilie ; les phénotypes développementaux, citons comme exemple les premiers symptômes de la Chorée de Huntington qui n'apparaissent qu'après l'âge de 30 ans ; et les phénotypes inductibles, qui ne s'exprimeront chez l'individu que si ce dernier est exposé à un agent particulier, comme l'intolérance au lactose, ou hypolactasie, qui se révèle lors de l'absorption de lait, généralement avant l'âge adulte et est due à un défaut de régulation dans l'activité de la lactase.

3.1.2. Génotype ou les causes moléculaires d'une maladie génétique

Les mutations délétères, auxquelles des phénotypes cliniques sont associés, ont dans un premier temps un effet au niveau moléculaire. Sur le plan fondamental, dans le cadre des maladies génétiques, les modifications affectant directement le produit de l'expression d'un gène peuvent être classées en trois catégories que nous allons détailler dans cette section.

3.1.2.1. Perte de fonction

Une mutation peut être à l'origine de la perte de fonction du produit de l'expression d'un gène. Dans cette situation, il faut distinguer les allèles amorphes ou nuls, responsables de la synthèse d'une protéine totalement inactive ou de la perte totale d'expression d'une protéine, et les allèles hypomorphes responsables de la synthèse d'une protéine partiellement inactive ou de l'expression partielle d'une protéine.

Les mutations par perte de fonction sont, pour la majorité, retrouvées dans des maladies récessives, qui nécessitent pour se manifester une atteinte des deux allèles.

Elles peuvent cependant aussi être retrouvées dans les maladies dites dominantes par haplo-insuffisance, pour lesquelles la perte de 50% de l'activité du gène est suffisante pour entraîner un phénotype clinique. Citons comme exemple l'hypercholestérolémie familiale, dont la forme hétérozygote est néanmoins beaucoup moins sévère que la forme homozygote (Hobbs *et al.*, 1992).

3.1.2.2. Gain de fonction

Dans cette situation, plus rare, il faut distinguer les allèles hypermorphes des allèles néomorphes.

On qualifie d'hypermorphe un allèle qui correspond à une surexpression du gène sauvage ou qui code une forme hyperactive de son produit.

Un allèle néomorphe entraînera soit la production d'une protéine assurant une fonction différente de celle de la protéine sauvage, soit celle d'une protéine assurant la même fonction, exprimée au cours du développement au mauvais endroit et/ou au mauvais moment. Par conséquent, ces mutations touchent de préférence des régions régulatrices du génome.

De manière générale, les allèles « gain de fonction » sont dominants par rapport à l'allèle sauvage.

3.1.2.3. Effet dominant négatif

Cette situation correspond à la frontière entre les mutations entraînant une perte de fonction et celles entraînant un gain de fonction. On parle d'effet dominant négatif lorsque la protéine codée par un allèle muté, dénommé allèle antimorphe, perd non seulement sa fonction, mais interfère aussi avec la fonction de l'allèle normal chez les hétérozygotes.

De telles mutations caractérisent les gènes codant les protéines de structure ou capables de former des homodimères ou des hétérodimères : ces mutations entraînent des modifications conformationnelles qui affectent la fonction de la protéine ou du complexe sain.

3.1.3. Corrélation génotype/phénotype

Un sujet d'étude primordial en génétique humaine est l'analyse des relations qui existent entre le génotype d'un individu et son phénotype. Cette étape est essentielle pour mieux comprendre les fonctions remplies par le gène d'intérêt, en déduire le mécanisme de la maladie et être capable, à terme, d'élaborer des solutions thérapeutiques.

Les phénotypes moléculaires, qu'on peut considérer comme étant les effets directs de la mutation sur le produit ou l'expression des gènes, sont généralement plus faciles à appréhender dans le contexte actuel d'explosion des connaissances relatives à l'expression, la

structure et la fonction des gènes et de leurs produits. Les conséquences morphologiques ou comportementales sur l'individu porteur de la mutation sont beaucoup plus difficiles à lier à un gène et à une mutation particulière. Outre les difficultés liées à la compréhension globale du « système » dans lequel la protéine s'inscrit, des problèmes relatifs à l'expression des phénotypes peuvent être rencontrés, comme :

- la pénétrance incomplète, illustrée par l'existence d'individus porteurs d'un allèle délétère mais chez lesquels le phénotype usuellement observé ne s'exprime pas. La pénétrance est la mesure quantitative de la fréquence de la manifestation d'un phénotype lorsqu'un allèle particulier est présent dans une population,
- la variabilité phénotypique, que nous avons déjà évoquée dans la section 3.1.1 et qui peut entraîner une hétérogénéité clinique entre deux individus porteurs de la même mutation.

Enfin, pour mener à bien des études de corrélation entre le génotype d'un individu et son phénotype, il est bien sûr indispensable de disposer d'informations précises sur les mutations et les phénotypes qui y sont associés, mais il est également essentiel de disposer d'informations sur le polymorphisme, pour mesurer et comprendre pourquoi ces derniers n'influencent pas le bon fonctionnement du gène.

3.2. Schéma de transmission des maladies monogéniques

Une maladie est considérée comme génétique si elle est transmissible d'une génération à l'autre. Il faut toujours garder à l'esprit, dans le cadre des études des maladies génétiques humaines, que l'homme est diploïde. L'homme possède donc deux allèles de chaque gène, identiques (homozygotie) ou différents (hétérozygotie). Une maladie génétique sera transmise sur le mode dominant si les conséquences de la mutation de l'une des deux copies du gène ne sont pas compensées par la copie normale, sinon on parlera de transmission en mode récessif. La transmission d'allèles, même délétères, peut se faire de manière transparente si elle est récessive.

Précisons que ce n'est pas le gène, mais la mutation qui caractérise le mode de transmission. La dominance/récessivité n'est pas une propriété intrinsèque d'un allèle particulier mais décrit plutôt sa relation avec l'allèle lui correspondant sur le chromosome homologue. Ceci explique pourquoi des mutations dans un même gène peuvent être transmises selon un schéma différent. Dans le cas de l'hypophosphatasie, maladie caractérisée par une déminéralisation des os et due à une mutation délétère affectant la phosphatase alcaline non tissu-spécifique, les formes sévères de la maladie sont transmises sur le mode récessif tandis que les formes modérées peuvent être transmises sur le mode dominant ou récessif (Mornet and Simon-Bouy, 2004).

Les études familiales, illustrées notamment par la construction d'arbres généalogiques (Figure 15), représentent une voie de choix pour l'étude de la transmission des caractères d'une génération à l'autre.

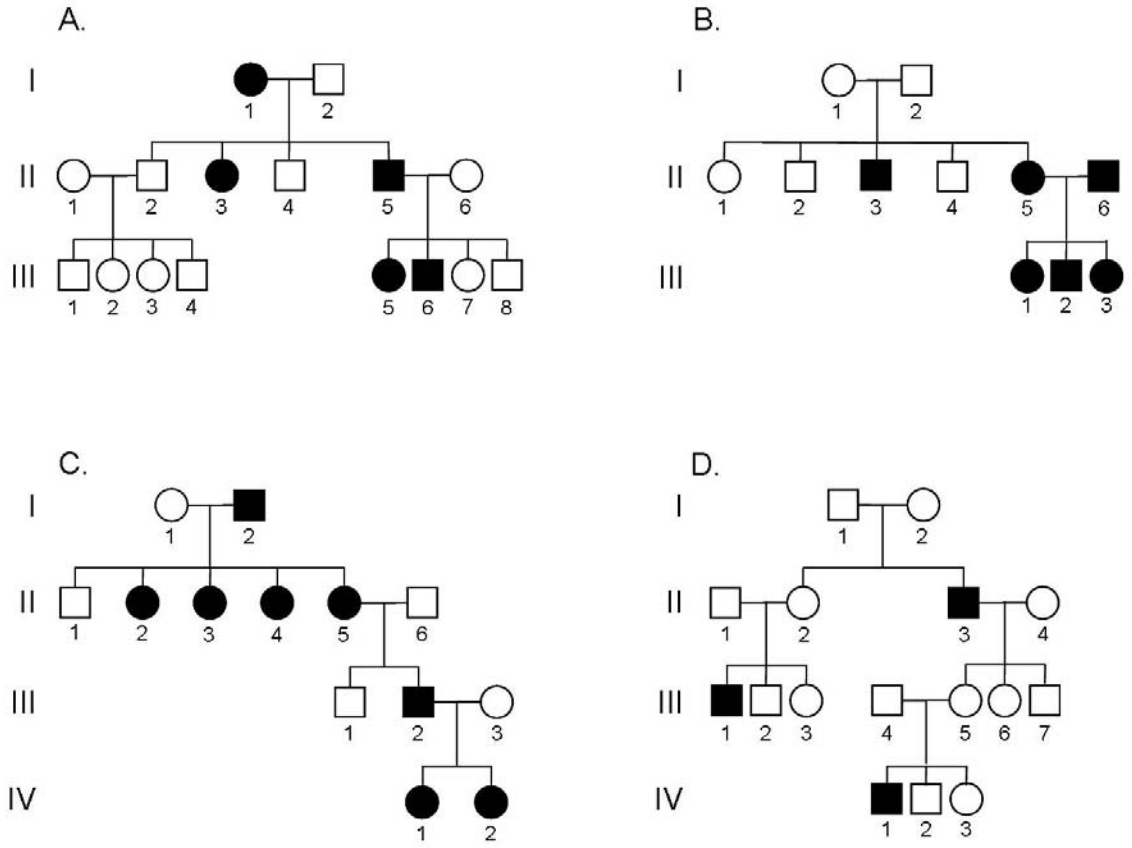


Figure 15 - Arbres généalogiques : schémas de transmission des maladies monogéniques. Les ronds représentent les femmes et les carrés les hommes. Les ronds et carrés noirs représentent les individus malades. Les chiffres romains désignent les générations et les chiffres arabes précisent les individus de chaque génération. (A) arbre caractéristique d'une transmission autosomique dominante ; (B) arbre caractéristique d'une transmission autosomique récessive ; (C) arbre caractéristique d'une transmission liée à l'X et dominante ; (D) arbre caractéristique d'une transmission liée à l'X et récessive.

On distingue généralement trois schémas principaux de transmission des maladies monogéniques :

- la transmission autosomique dominante au cours de laquelle l'apparition d'une maladie est liée à la mutation d'un gène porté par un chromosome autosome. Dans ce cas, la présence d'un seul allèle muté chez un individu suffit pour provoquer l'apparition de la maladie. On retrouvera classiquement, au sein des arbres généalogiques, autant d'hommes que de femmes atteints au cours des générations successives, au moins un parent atteint pour chaque individu atteint, mais aussi deux parents atteints ayant des enfants sains (Figure 15 A),

- la transmission autosomique récessive, cas dans lequel la mutation se situe dans un gène porté par un autosome, mais où les deux allèles de ce gène doivent être mutés pour que la maladie apparaisse. Les individus ne possédant qu'un seul allèle muté, qui ne déclarent pas la maladie, sont considérés comme des porteurs sains. On trouvera dans un arbre généalogique caractéristique d'une maladie autosomique récessive des enfants atteints nés de deux parents non-atteints, les garçons et les filles étant touchés à une fréquence équivalente. Tous les descendants de deux parents atteints seront eux-mêmes atteints (Figure 15 B),
- la transmission liée au sexe, qui présente un schéma particulier selon le chromosome sexuel porteur d'une part et le sexe de l'individu d'autre part :
    - un caractère lié au chromosome Y présente une hérédité dite holandrique et se transmet uniquement de père en fils. Le chromosome Y porte principalement les gènes de la différenciation sexuelle : la plupart des mutations au niveau de ces gènes entraînent une stérilité des individus la portant, voire une inversion sexuelle avec stérilité. Les maladies liées à l'Y sont par conséquent généralement dues à l'apparition d'une néomutation,
    - un caractère lié au chromosome X se manifestera selon un schéma particulier : chez les hommes, toute mutation d'un gène porté par ce chromosome déterminera l'apparition de la maladie, en raison de l'absence d'un second chromosome X susceptible de compenser le défaut du gène muté. On déterminera le caractère dominant ou récessif d'un gène lié à l'X grâce aux phénotypes des femmes qui possèdent deux chromosomes X. Dans le cas d'un état lié à l'X dominant, toutes les filles d'un homme atteint seront touchées, alors qu'une femme atteinte pourra donner naissance aussi bien à des filles qu'à des garçons atteints. Globalement, autant de femmes que d'hommes manifesteront le trait (Figure 15 C). Dans le cas d'une maladie liée à l'X et récessive, tous les fils d'une mère atteinte seront atteints, les pères atteints ne transmettront jamais le caractère à leur fils, des parents non atteints peuvent donner naissance à des fils atteints et globalement, la maladie sera plus fréquente chez les garçons que chez les filles (Figure 15 D).

3.3. Caractérisation du génotype

Caractériser de manière précise une maladie génétique et comprendre ses fondements nécessitent d'une part l'identification du gène responsable de la maladie et d'autre part l'identification des mutations à l'origine de la maladie. Nous aborderons les méthodes disponibles pour réaliser ces identifications dans les prochaines sections.

3.3.1. Détermination des gènes responsables de maladies génétiques

Appréhender dans leur globalité les relations entre un phénotype et son génotype dépend des capacités à caractériser le ou les gènes responsables de ce phénotype ainsi que leurs rôles et interrelations.

Du fait de l'influence de nombreux facteurs génétiques et environnementaux sur la plupart des maladies génétiques humaines et les phénotypes qui leur sont associés, l'identification et la caractérisation des gènes spécifiques de certaines variations influençant une maladie sont particulièrement fastidieuses.

La mise en place du programme Génome Humain et d'autres programmes de séquençage de génomes a permis de disposer d'un grand nombre de ressources permettant d'améliorer considérablement les moyens de recherche des gènes responsables des maladies génétiques humaines, notamment grâce à la disponibilité de séquences, de cartes génétiques, de données d'expression, etc.

Les premiers gènes identifiés en tant que locus morbides ont souvent pu l'être grâce au phénotype lui-même. Ainsi, l'implication de l'hémoglobine dans l'émergence de l'anémie falciforme a, par exemple, été mise en évidence grâce à la disponibilité de données biochimiques précises, qui ont permis la purification du produit du gène muté.

Lorsqu'un biologiste ne dispose pas, ou de peu, d'informations relatives à la fonction du produit du gène morbide, plusieurs techniques sont à sa disposition, la stratégie déployée reposant sur les informations génétiques et moléculaires disponibles.

3.3.1.1. Identification par voie cytogénétique

Des observations cytogénétiques sur des individus malades peuvent révéler des remaniements chromosomiques, comme des délétions ou des translocations, qui permettent de guider les recherches vers une région précise du génome.

Cette méthode a permis de mettre en évidence le gène de la dystrophine comme étant responsable de la myopathie de Duchenne. En effet, la translocation d'une région du chromosome X sur le chromosome 21 d'une fille myopathe a désigné cette région comme porteuse du gène muté et l'analyse des gènes présents dans cet intervalle a permis l'identification du gène morbide (Monaco *et al.*, 1986).

3.3.1.2. Approche positionnelle

L'approche positionnelle est la méthode de choix lorsque l'on commence une étude sans *a priori* sur le produit du gène responsable d'une maladie génétique. En effet, cette méthode est basée sur des analyses de liaison génétique, qui sous-tendent que les gènes physiquement proches sur un chromosome sont généralement hérités conjointement : ils apparaissent par

conséquent comme étant génétiquement liés. L'approche positionnelle repose donc sur l'identification des marqueurs génétiques qui coségrègent avec la maladie lors de la méiose.

Cette approche nécessite le recensement de familles au sein desquelles la maladie d'intérêt est exprimée. Pour chaque individu, un échantillon sanguin est prélevé, qui servira au génotypage d'un grand nombre de marqueurs répartis à intervalles réguliers le long des chromosomes. Le but est d'identifier les marqueurs dont les allèles sont hérités spécifiquement avec le phénotype associé à la maladie en réalisant un travail de focalisation progressive qui, grâce aux cartes du génome, permet d'identifier et de réduire progressivement un intervalle de liaison contenant le gène muté. L'établissement des cartes génétiques précises suite au séquençage du génome humain a considérablement amélioré la pratique de cette stratégie. L'inventaire de l'ensemble des gènes présents sur cet intervalle a ainsi été dressé, et parmi ceux-ci on recherche celui présentant des mutations qui ségrègent spécifiquement chez les individus malades.

3.3.1.3. Identification par l'approche du gène candidat

L'identification du gène morbide par l'approche du gène candidat n'est sans doute pas la méthode la plus efficace, puisqu'elle repose sur un « pari », néanmoins raisonné, sur le type de protéine qui pourrait être responsable de l'émergence du phénotype.

Après avoir pris les mesures des symptômes d'une maladie génétique ou lorsque le biologiste dispose d'informations sur ses bases biochimiques, un gène candidat sera en effet choisi dans le répertoire disponible des gènes plausibles.

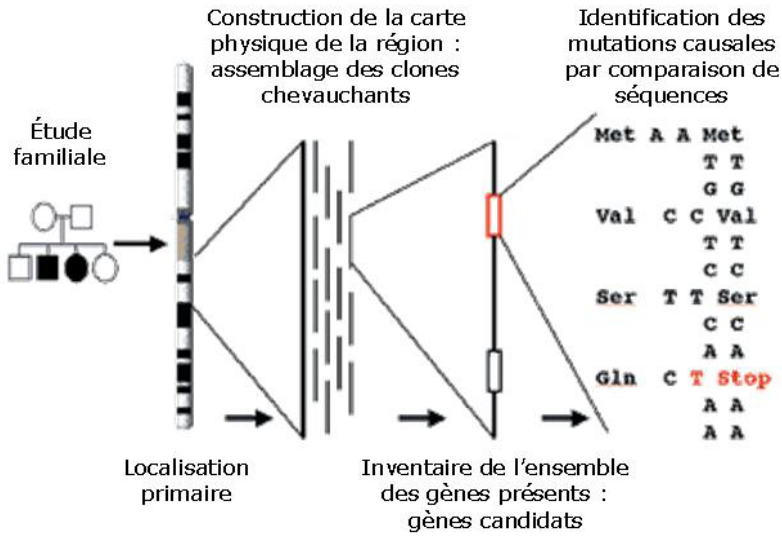


Figure 16 - Identification de mutation par clonage positionnel, complétée par une approche du gène candidat.
(adaptée de http://images.the-scientist.com/content/figures/images/yr2003/feb10/chart2.gif)

Cette approche peut également compléter l'approche positionnelle, on parle alors d'approche positionnelle du gène candidat (Figure 16). Ue fois la partie du chromosome contenant le gène morbide identifiée, les cartes génétiques et transcriptionnelles les plus récentes sont examinées pour prendre connaissance des régions codantes déjà cartographiées dans le même intervalle que le gène morbide et un ou plusieurs gènes candidats sont alors déterminés.

L'efficacité de cette méthode a été augmentée par l'accumulation, au cours des dernières années, de ressources bioinformatiques multiples à même de croiser un certain nombre de données, éventuellement relatives à plusieurs organismes. Les données d'expression représentent de précieuses informations pour la détermination de gènes candidats, permettant de voir si l'expression d'un gène est cohérente avec le phénotype de la maladie observée. Les données relatives à la fonction des produits des gènes, que l'on peut retrouver dans certaines banques de données ou prédire à partir d'outils bioinformatiques, peuvent également renseigner le scientifique sur la potentialité d'un gène à devenir un « bon candidat ».

3.3.2. Détection expérimentale des mutations

La démonstration ultime qu'un gène est responsable de l'émergence d'un phénotype malade passe par l'identification de mutations dans ce gène, qui pourront être reliées à l'apparition de la maladie. De par l'existence d'une variabilité naturelle entre les séquences de plusieurs individus, l'identification de plusieurs mutations significatives dans le même gène qui ségrègent entre des familles indépendantes, mais présentant des caractéristiques cliniques similaires, offre la preuve la plus convaincante de la relation de cause à effet entre le gène identifié et la maladie et ouvre des perspectives de diagnostic.

La détection de mutations est couramment utilisée dans le cadre de diagnostics cliniques ou d'études de prévention et il est bien sûr essentiel de disposer de protocoles fiables, afin de réduire au maximum les erreurs de diagnostics. Le recours au diagnostic clinique est généralement effectué en cas d'antécédents familiaux et dans l'intérêt du patient, mais de nombreuses questions éthiques s'ouvrent autour de certains tests proposés (Ensenauer *et al.*, 2005).

Depuis l'avènement de la PCR, qui permet l'amplification spécifique d'un fragment d'ADN, de nombreuses stratégies ont été développées pour l'identification des mutations pathogènes. Le choix d'une technique ou d'une stratégie dépendra du fait que le gène d'intérêt ait été identifié, qu'il soit cloné ou non, mais aussi de sa taille, de la nature et de la fréquence des mutations.

Les analyses génotypiques peuvent être classées en deux catégories : la détection directe, qui permet la mise en évidence de la lésion génétique elle-même, et la détection indirecte, qui permet la mise en évidence d'un gène muté, en se basant sur des marqueurs polymorphes.

Nous passerons rapidement en revue les méthodes les plus couramment employées dans les deux paragraphes suivants.

3.3.2.1. Identification directe

La mise en évidence d'une lésion génomique est de difficulté variable selon la nature du gène et le type de lésion. Dans tous les cas, pour explorer directement un gène, il faut disposer d'une sonde clonée spécifique de ce gène et en connaître la (ou les) séquence saine.

L'ensemble des mutations affectant un gène d'intérêt peut être identifié par séquençage direct de l'ADN, mais cette technique n'est pas la plus rapide. En effet, elle est peu adaptée au séquençage de nombreux échantillons car relativement lourde. De plus, lorsque les mutations sont présentes à l'état hétérozygote, les mutations ponctuelles peuvent être difficiles à détecter et nécessitent de travailler avec une séquence d'ADN de très haute qualité pour la mise en évidence d'un double signal à une position précise (Figure 17).

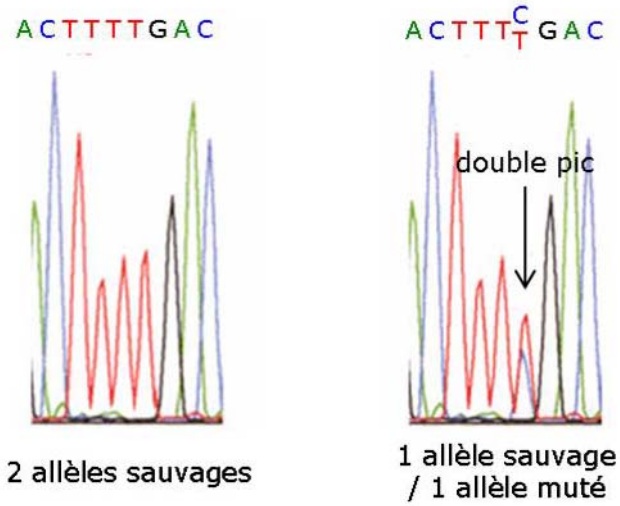


Figure 17 - Séquençage automatique d'un gène : exemple d'électrophérogramme. Chaque nucléotide est représenté par un pic dont la couleur révèle la nature. A gauche, profil obtenu dans le cas d'un individu homozygote, qui porte deux allèles sauvages. A droite, profil obtenu dans le cas d'un individu hétérozygote, porteur d'un allèle sauvage et d'un allèle muté. L'allèle muté possède un C en position 6 alors que l'allèle sauvage possède un T : on remarque deux pics superposés à cette position. (adaptée de http://www.uvm.edu/~cgep/Education/Sequence.html)

D'autres techniques courantes permettent de déterminer si un segment d'ADN porte ou non un changement de nucléotides, sans avoir d'*a priori* sur la mutation en question.

L'emplacement de la mutation dans le gène est repéré afin de limiter l'étape ultime de séquençage qui identifie définitivement la mutation.

La méthode dite du « polymorphisme de la conformation de l'ADN simple brin » (Orita *et al.*, 1989) est basée sur le principe que l'ADN double brin dénaturé adopte une conformation tridimensionnelle spécifique. La différence de migration sur gel entre l'ADN sauvage et l'ADN muté est typique de la présence d'une mutation. Cette méthode est très sensible et permet de détecter toutes sortes de mutations, même ponctuelles, mais elle est limitée à des fragments d'environ 200 pb, c'est pourquoi les exons du gène d'intérêt sont testés de manière indépendante.

La méthode de l'analyse des hétéroduplex (Keen *et al.*, 1991) est, quant à elle, basée sur le fait que les deux brins d'ADN peuvent être séparés par dénaturation, puis réassociés pour reformer un duplex. L'ADN testé sous forme simple brin est ici associé à l'ADN sauvage double brin. Si les deux molécules sont identiques, elles formeront un homoduplex. Elles formeront un hétéroduplex dans le cas contraire, c'est-à-dire si l'ADN testé contient une mutation. Une mutation ponctuelle suffit à modifier la mobilité électrophorétique entre les deux entités, permettant la mise en évidence aisée de la présence d'une mutation dans l'ADN testé. La méthode du polymorphisme de conformation et celle de l'analyse des hétéroduplex peuvent être combinées pour une meilleure efficacité dans la détection des mutations (Kozlowski and Krzyzosiak, 2001).

D'autres méthodes, que nous ne détaillerons pas ici, peuvent également être utilisées, comme l'électrophorèse en gradient de dénaturation, la coupure chimique des mésappariements ou encore le test de troncation de la protéine (*voir* (Pasternak, 1999) pour plus de détails).

Enfin, les développements récents dans l'automatisation et la miniaturisation des technologies ont permis le développement de nouveaux standards et une évolution dans la philosophie des diagnostics moléculaires à l'ère post-génomique. La technologie des puces à ADN est devenue synonyme de détection de mutations à grande échelle et assure fiabilité et rapidité dans leur identification, en permettant de tester des centaines de variations ponctuelles ou plus importantes, en une seule expérience (Craig and Stephan, 2005). Les puces à ADN, communément appelées *DNA microarray*, sont dans ce cas constituées d'oligonucléotides courts (les sondes), suffisamment spécifiques pour discriminer des séquences différant d'un seul nucléotide, immobilisés sur un support solide selon une disposition ordonnée. L'analyse repose sur l'hybridation entre les sondes de la puce et l'ADN testé marqué par un fluorochrome.

3.3.2.2. Détection indirecte

Le diagnostic indirect par analyse de liaison avec des marqueurs polymorphes est envisagé lorsque la mutation délétère n'est pas connue et dans la mesure où des données concernant

la localisation chromosomique précise du gène en cause sont disponibles. La mise en œuvre de cette méthode suppose l'existence d'une certitude diagnostique absolue et l'absence d'hétérogénéité génique de la maladie.

Le principe du diagnostic indirect consiste à distinguer le chromosome porteur du gène pathologique de son homologue normal, non plus par la détection de la lésion génomique elle-même, mais par l'étude de marqueurs localisés dans le voisinage de celle-ci, à une distance génétique faible et connue, voire à l'intérieur du gène en question. Cette méthode nécessite de disposer de l'ADN de plusieurs individus affectés et apparentés au premier degré (parent/enfant/fratrie). En se basant sur des marqueurs polymorphiques d'ADN, la phase entre les allèles marqueurs et la maladie est établie (Figure 18). Cette information peut être utilisée pour traquer la partie du chromosome portant le gène d'intérêt dans d'autres membres de la famille.

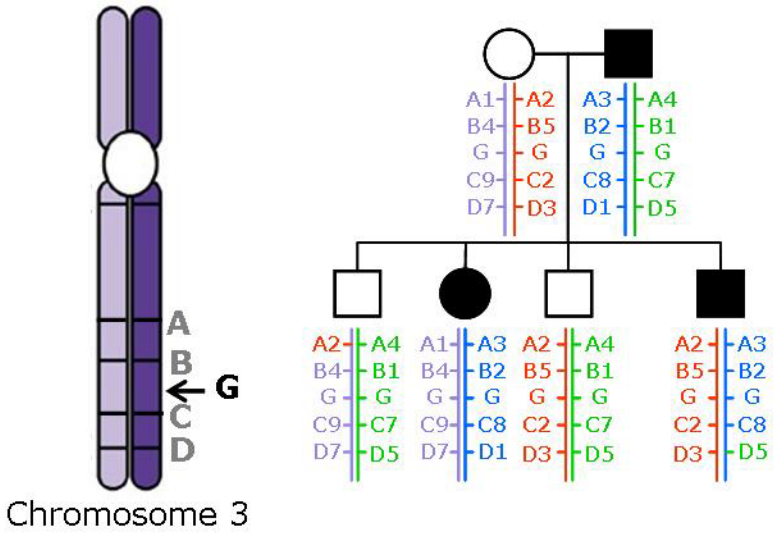


Figure 18 - Détection indirecte de mutations par analyse de liaison : cas d'une maladie autosomique dominante dont le locus morbide G se situe sur le chromosome 3. Les allèles de quatre loci marqueurs (A, B, C et D) sont représentés. Cette étude permet de définir quels allèles des marqueurs sont associés avec l'allèle responsable de la maladie (la phase). On remarque ici que les allèles A3, B2 et C8 sont systématiquement associés à l'allèle morbide. Dans l'arbre généalogique, les ronds représentent les femmes, les carrés les hommes ; les symboles noirs représentent les individus malades.

3.4. Sources d'information

Dans le cadre des variations génétiques humaines et des phénotypes qui leurs sont associés, nous disposons actuellement d'un grand nombre de banques de données diverses et variées, qui contiennent différents niveaux d'informations.

MIM (McKusick, 1998), pour *Mendelian Inheritance in Man*, et OMIM (Hamosh *et al.*, 2005), pour *Online Mendelian Inheritance in Man*, sa version en ligne, sont les références en matière de maladie génétique humaine. OMIM est une banque de données, accessible sur le site du NCBI (http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM), et n'est autre que la version électronique de l'encyclopédie créée par Victor McKusick et ses collaborateurs en 1960 et maintenue depuis à l'université John Hopkins de Baltimore. La banque connaît une croissance exponentielle, notamment depuis les avancées techniques réalisées dans la mise en évidence des gènes (Figure 19) dans les années 90. OMIM représente un catalogue des gènes humains et des maladies associées et la force de cette banque réside dans la qualité et la diversité de son contenu. L'accès à des informations tant cliniques, génétiques que moléculaires y est possible. Chaque maladie génétique est répertoriée, décrite et annotée. Chaque entrée d'OMIM, en plus de la position du gène dans la cartographie génétique humaine, donne une description du syndrome, des informations sur son mode de transmission, le nom et la séquence du gène, un certain nombre de mutations ainsi que de nombreuses références vers des articles spécifiques.

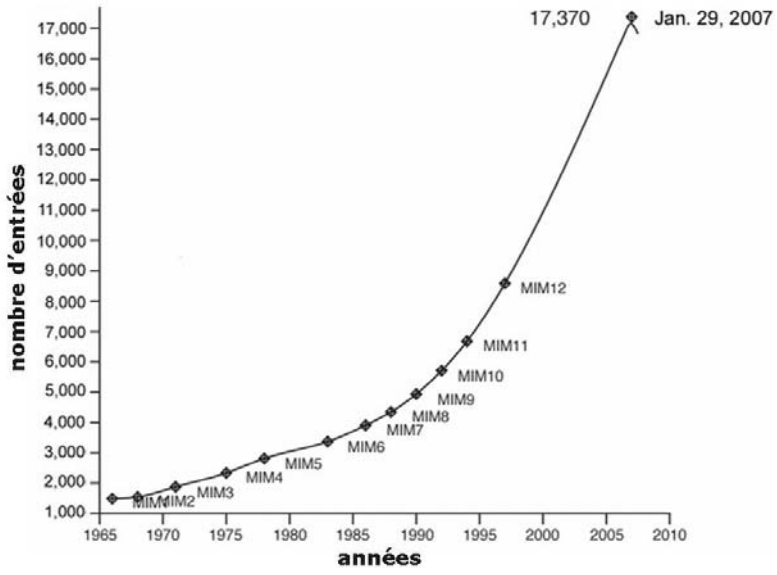


Figure 19 - Augmentation du nombre d'entrées dans MIM et OMIM. (adaptée de (McKusick, 2007))

D'autres banques remarquables ont été développées plus récemment, adaptées non seulement au grand public mais aussi à un public plus spécialisé de médecins et/ou chercheurs. GeneClinics (Tarczy-Hornoch *et al.*, 2000) (http://www.geneclinics.org/) est à la fois une ressource électronique génétique qui décrit des maladies héréditaires, un service d'information sur le rôle des tests génétiques de diagnostic et un service de conseil pour des patients ayant des maladies génétiques. GeneClinics est également destiné à promouvoir l'utilisation de services génétiques pour une prise de décision sur des tests génétiques. Les données sont saisies par des experts de chacune des maladies, garantissant l'accès à des

données fiables de grande qualité. Citons encore Orphanet (Ayme *et al.*, 1998) (http://www.orpha.net/), un serveur d'informations sur les maladies rares dont l'objectif est de concourir à l'amélioration de leur prise en charge et de leur traitement, qu'elles soient d'origine génétique, auto-immune ou infectieuse.

Les banques de données spécialisées dans la centralisation des variations génétiques peuvent être classées en quatre catégories, que nous allons décrire succinctement dans les prochaines sections. Nous reprendrons une description un peu plus précise de ces banques, en nous concentrant sur leurs avantages et leurs inconvénients respectifs, dans le Chapitre 13 - *Mise en place des bases de l'infrastructure dédiée à l'analyse globale des mutations*, p. 189.

3.4.1. Banques centrales

Les banques dites « centrales » regroupent en leur sein une multitude de données de mutation associée à un grand nombre de gènes impliqués dans des maladies génétiques. Chaque donnée y est présentée avec des détails limités. Généralement, ces banques référencent exclusivement des mutations délétères, les effets phénotypiques associés n'étant que peu décrits et souvent limités au nom de la maladie.

La banque centrale de référence est la HGMD (Stenson *et al.*, 2003), pour *Human Genome Mutation Database*, maintenue à l'université de Cardiff en partenariat avec BioBase, une compagnie privée. Une version publique de la banque est disponible (http://www.hgmd.cf.ac.uk/ac/index.php), mais seules les données datant d'il y a plus de deux ans y sont accessibles librement. La HGMD regroupe des données issues de la littérature, chaque entrée étant vérifiée précautionneusement et validée contre la séquence de référence du gène, permettant de disposer de données homogènes par rapport à la description et à la numérotation des mutations. Chaque mutation n'est référencée qu'une fois dans la base.

Swiss-Prot (Boeckmann *et al.*, 2003), banque généraliste de séquences protéiques, peut également être considérée comme une banque centrale de mutations. En effet, sans se targuer de centraliser de manière exhaustive les mutations connues affectant une protéine, Swiss-Prot référence un grand nombre de variants pour de nombreuses protéines humaines.

Citons enfin la PMD (Kawabata *et al.*, 1999), pour *Protein Mutant Database*, qui centralise des données, non exclusivement humaines, sur les mutations naturelles et artificielles extraites de la littérature et qui sont un peu plus ardues à exploiter à cause de leur référencement par article et non par protéine.

3.4.2. Banques spécifiques d'un locus

Les banques spécifiques d'un locus, couramment appelées LSDB, pour *Locus-Specific DataBase*, contiennent, comme leur nom l'indique, des informations relatives à un gène ou à une famille de gènes spécifiques impliqués dans le même type de maladie. Ces banques sont directement maintenues par des cliniciens ou des scientifiques spécialistes du gène ou de la maladie d'intérêt et répertorient des données publiées ou saisies par les expérimentateurs avant publication. Les informations stockées dans ces banques sont de très grande qualité et beaucoup plus détaillées que dans les banques centrales.

Un bon exemple de LSDB est HbVar (Giardine *et al.*, 2007), banque de données relationnelles relatives aux variants de l'hémoglobine et à la thalassémie. Elle recense pour chaque variant non seulement des données précises sur la mutation, la pathologie associée, l'hématologie, la mobilité électrophorétique, mais aussi des données structurales et fonctionnelles ainsi que des informations sur l'occurrence ethnique, etc.

Près de 700 LSDB sont actuellement dénombrées et disponibles à travers le web. Ces banques sont référencées sur le site de la *Human Genome Variation Society* (http://www.hgvs.org/). Le nombre de LSBD disponibles ne cesse de croître, mais toutes ne sont pas maintenues de manière régulière. De plus, les formats de ces banques sont très variables, mais des efforts sont déployés pour leur homogénéisation et des outils génériques sont mis à la disposition des cliniciens, notamment par l'intermédiaire des systèmes UMD, pour *Universal Mutation Database* (Beroud *et al.*, 2000) et LOVD, pour *Leiden Open (source) Variation Database* (Fokkema *et al.*, 2005).

3.4.3. Banques de polymorphismes

De nombreuses informations sur les variations de séquences chez l'homme sont également disponibles dans les banques de polymorphismes, comme dbSNP (Sherry *et al.*, 1999) et HGVbase (Fredman *et al.*, 2002).

Le nom de polymorphisme est ici employé par abus de langage dans le sens où, de manière générale, la fréquence de la variation dans la population n'est pas prise en compte. De plus, ces banques ne tentent pas clairement de lier un génotype à un phénotype et, bien que les variations neutres soient surreprésentées au sein de ces banques, certaines variations délétères y sont également référencées.

Ces banques représentent une source d'informations importante pour compléter l'étude d'un gène ou d'une région d'intérêt, mais qu'il faut néanmoins manipuler avec précaution.

3.4.4. Banques nationales de mutations

Le spectre des mutations observées pour chaque gène ou maladie diffère généralement selon les groupes de populations, mais aussi entre les groupes ethniques s'ils se situent dans des zones géographiques différentes. C'est pourquoi un nouveau type de banques de mutations fait son apparition : les banques nationales de mutations, ou NMDB pour *National Mutation DataBase*. Ces dernières représentent un outil de choix dans l'optimisation des services nationaux de diagnostics. De plus, le fait de disposer d'informations relatives aux populations permet de calculer des paramètres couramment utilisés dans les analyses statistiques des mutations, comme la fréquence ou la pénétrance d'une mutation, la pénétrance reflétant la probabilité qu'une personne portant une mutation développe les caractéristiques causées par cette mutation.

Actuellement, à peine une dizaine de NMDB ont été développées, elles sont référencées en partie sur le site http://www.goldenhelix.org/.

3.5. Maladies génétiques humaines : quelques chiffres

Au 15 août 2007, OMIM contenait 2 072 gènes dont la séquence est connue et auxquels au moins un variant allélique est associé (requête réalisée sur le site web d'OMIM, http://www.ncbi.nlm.nih.gov/sites/entrez?db=OMIM&TabCmd=Limits, en limitant la recherche aux variants alléliques connus). Plusieurs phénotypes distincts pouvant être associés à un même gène, selon la mutation qui l'affecte, le nombre de phénotypes, dont les bases moléculaires sont connues, est légèrement supérieur (2 147) dans le tableau des statistiques d'OMIM (Tableau 3).

Classification des entrées	Autosome	Lié à l'X	Lié à l'Y	Mitochondrial	Total
* Gène avec séquence connue	11 106	507	48	37	11 698
+ Gène avec séquence connue et phénotype	351	31	0	0	382
# Description phénotypique, bases moléculaires connues	1 940	179	2	26	2 147
% Phénotype ou locus mendélien, bases moléculaires inconnues	1 464	133	4	0	1 601
Autres, principalement phénotypes avec bases mendéliennes soupçonnées	1 981	143	2	0	2 126
Total	16 842	993	56	63	17 954

Tableau 3 - Statistiques d'OMIM, au 15 août 2007.

Les gènes impliqués dans les maladies génétiques humaines peuvent être classés selon les fonctions assurées par le produit de leur expression. La Figure 20 illustre la diversité de ces fonctions, symbolisées par leur classement *Gene Ontology*. La distribution de ces fonctions est

très proche de celle obtenue au niveau de l'annotation du protéome humain (Pruess *et al.*, 2003).

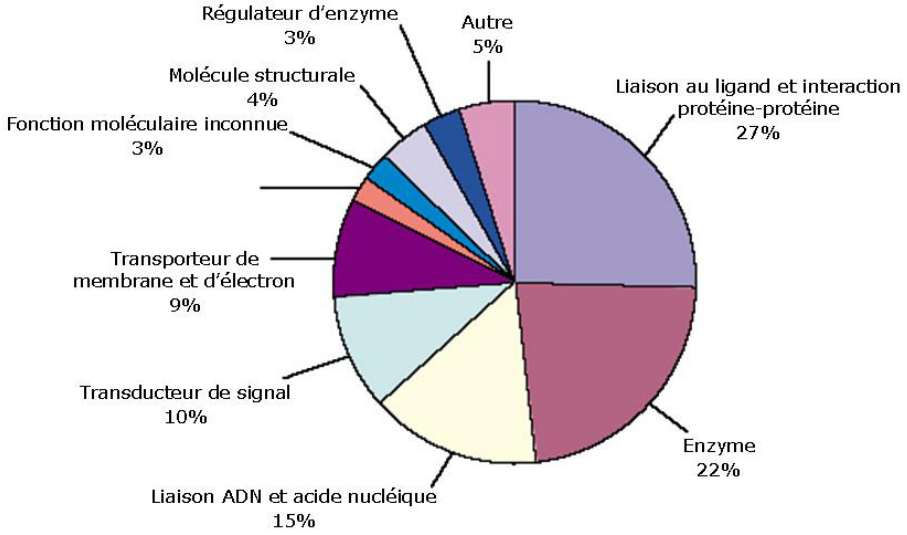


Figure 20 - Variété de fonctions (définie selon *Gene Ontology* GO) des protéines référencées dans OMIM (protéines issues de Swiss-Prot, pour lesquelles des mutations sont renseignées). Une protéine peut avoir plusieurs termes GO associés. (adaptée de (Steward *et al.*, 2003))

L'étude des positions, au sein des gènes, des mutations associées à des maladies génétiques humaines a révélé qu'elles ne sont pas réparties de manière uniforme le long de ces séquences d'ADN (Antonarakis *et al.*, 2001), mais qu'elles sont au contraire fortement influencées par l'environnement local de la séquence (Krawczak *et al.*, 1998). Le dinucléotide CpG est, par exemple, un point chaud de mutation et une étude comparative des mutations délétères et de l'usage des codons dans les gènes humains a mis en évidence une hypermutabilité des triplets de nucléotides codant pour l'arginine et la glycine. Les mutations engendrant des décalages du cadre de lecture prennent, quant à elles, préférentiellement position dans des endroits du gène où deux nucléotides identiques se succèdent. De tels constats permettent d'envisager qu'à plus ou moins long termes, les différents types de mutation, leur localisation et leur fréquence relative pourront être prédits à partir de la séquence du gène d'intérêt.

De plus, les mutations délétères sont, de manière générale, observées à des positions hautement conservées au sein de la séquence, indiquant une importance fonctionnelle ou structurale de ces positions (Ferrer-Costa *et al.*, 2002). La cartographie des substitutions faux-sens délétères à la surface des protéines a permis de mettre en évidence qu'elles prennent préférentiellement place dans des poches ou des « vides » structuraux et qu'elles entraînent une altération de la fonction de la protéine (Stitziel *et al.*, 2003).

Le spectre des mutations issues de la HGMD est représenté dans la Figure 21. La majorité des mutations répertoriées au sein de cette banque, la plus complète en termes de maladies génétiques humaines, sont des substitutions d'un seul nucléotide (67,8%), les mutations faux-sens représentant à elles seules près de la moitié des mutations totales de la banque. De plus, Antonarakis et ses collègues (Antonarakis *et al.*, 2000) ont montré que les transitions sont nettement plus fréquentes que les transversions (62,5% contre 37,5% environ), ce qui reste en accord avec les observations génériques des taux de variations dans la nature (*voir* section 2.2.3.1, p. 26).

Le second type de mutations le plus courant est représenté par les petites insertions et petites délétions de moins de 20 nucléotides, sachant que ces dernières impliquent la plupart du temps un seul nucléotide. Environ 87% des petites délétions entraînent une modification du cadre de lecture du fait que le nombre de nucléotides insérés ou perdus ne soit pas un multiple de trois.

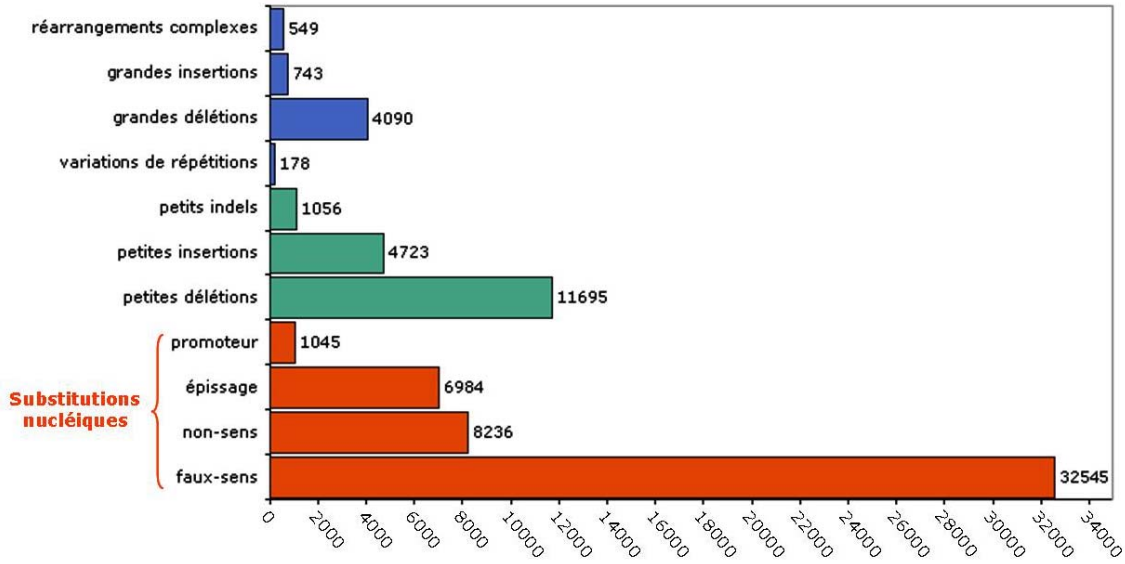


Figure 21 - Répartition des types de mutations à l'origine de maladies génétiques humaines. Données extraites de la banque de données HGMD (http://www.hgmd.cf.ac.uk/) au 31 juillet 2007.

Enfin, une étude (Jimenez-Sanchez *et al.*, 2001) a montré l'existence d'une corrélation entre la fonction de la protéine impliquée dans une maladie et certaines caractéristiques de cette dernière, comme l'âge de déclaration de la maladie ou son schéma de transmission. Les maladies associées à des gènes codant pour des protéines de toutes les catégories fonctionnelles peuvent apparaître à n'importe quelle étape de la vie, mais les gènes codant des facteurs de transcription sont surreprésentés au niveau des maladies se déclarant *in utero* : ces protéines jouent un rôle prépondérant au niveau de l'orchestration du développement néonatal. Ces maladies sont généralement transmises sur le mode dominant. Beaucoup de maladies se déclarant avant l'âge d'un an sont dues à un défaut dans une

enzyme et se transmettent sur le mode récessif. L'enfant est généralement sain à la naissance du fait qu'il a accès au système homéostatique métabolique de sa mère à travers le placenta durant la grossesse et les premiers symptômes n'apparaissent par conséquent qu'une fois l'enfant né.

Chapitre 4 - Les mutations et l'analyse de leurs conséquences

Une des difficultés principales, dans le cadre de l'étude des maladies génétiques humaines, consiste à distinguer, au niveau d'un gène d'intérêt, une mutation délétère à l'origine de modifications phénotypiques, de variations neutres, qui seront sans conséquence.

Pour être à même de surmonter ces difficultés, il est essentiel de comprendre comment une mutation est susceptible d'affecter le produit de l'expression du gène morbide.

Nous nous concentrerons dorénavant sur les mutations affectant directement les régions codantes d'un gène. Nous verrons dans ce chapitre les effets moléculaires possibles de telles mutations en décrivant les « points sensibles » des protéines. Nous dévoilerons ensuite comment il est possible, grâce aux avancées scientifiques permettant une meilleure compréhension de l'entité protéine, mais aussi grâce à la quantité de données disponibles au sein des banques et aux nombreux développements bioinformatiques, de caractériser une protéine par l'intermédiaire d'analyses évolutives et structurales. Nous nous focaliserons principalement sur les concepts de ces analyses, que nous exemplifierons par certaines méthodes et par quelques outils bioinformatiques développés pour assurer et faciliter ces investigations.

4.1. Conséquences des mutations sur les protéines

Les mutations non-sens aboutissant à la formation de l'un des trois codons stop, les mutations responsables d'un décalage du cadre de lecture (y compris les anomalies de l'épissage) et les mutations du codon d'initiation de la traduction entraînent généralement l'absence de formation d'une quelconque protéine, ou la formation d'une protéine tronquée dont l'activité sera nulle ou très réduite.

En revanche, les conséquences des mutations faux-sens ou des délétions, des insertions ou des indels en phase (y compris les anomalies de l'épissage) sont moins flagrantes. Ces mutations sont responsables d'un changement de la séquence protéique et sont susceptibles d'affecter par exemple la stabilité de la protéine, son adressage intracellulaire, sa maturation, son assemblage dans une structure multimérique, les sites essentiels à son activité enzymatique ou à l'interaction avec des ligands, etc.

Le changement de la séquence de la protéine peut également être sans conséquence sur sa fonction. Ainsi, lorsqu'un changement nucléotidique, de type faux-sens ou délétion, insertion ou indel en phase, survient au sein de la séquence codante d'un gène, un faisceau

d'arguments est nécessaire pour déterminer les conséquences de cette anomalie dans la pathologie.

Ces conséquences vont dépendre, d'une part, du contexte fonctionnel et structural de la position du/des résidus modifiés et, d'autre part, de la nature du changement induit. En effet, de nombreuses positions au sein de la protéine ne requièrent pas la présence d'un acide aminé spécifique, mais plutôt le respect de propriétés physico-chimiques particulières (hydrophobicité, volume, charge…) du résidu. De ce fait, certaines substitutions sont tout à fait tolérées au sein des protéines (Bowie *et al.*, 1990) : on peut parler, à une position donnée, de compatibilité des résidus sauvages et mutés.

Les résidus enfouis au cœur de la protéine sont extrêmement importants pour son repliement et sa stabilité par exemple, et seuls les résidus hydrophobes ou neutres sont généralement tolérés. Les substitutions qui permettent la conservation de ces propriétés d'hydrophobicité sont dans ce cas souvent bien supportées.

L'enjeu de la discrimination des mutations délétères par rapport aux mutations neutres peut être schématisé par la capacité à prédire les substitutions ou modifications qui seront tolérées par la protéine. Le diagramme de Venn ayant trait aux propriétés des acides aminés (Figure 22) donne une idée de la compatibilité des résidus en fonction des caractéristiques physico-chimiques à maintenir.

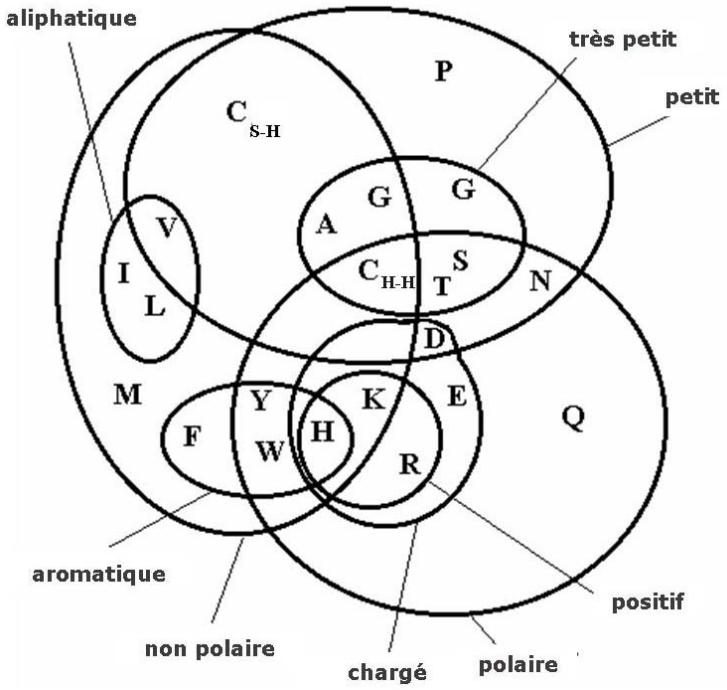


Figure 22 - Classification des acides aminés d'après leurs propriétés physico-chimiques (Taylor, 1986) : diagramme de Venn.

A priori, une mutation aura un effet délétère si elle altère de manière directe la fonction de la protéine en affectant un résidu essentiel ou si elle déstabilise la structure de la protéine, l'empêchant ainsi d'assurer son rôle fonctionnel.

Les résidus fonctionnels critiques que l'on peut relever sur une protéine sont les suivants :

- les résidus du site actif, impliqués directement dans la catalyse,
- les résidus impliqués dans une liaison particulière (au ligand, au calcium, à un ion métallique, etc.) ou dans l'interaction avec d'autres protéines,
- les résidus modifiés post-traductionnellement, etc.

Les caractéristiques structurales critiques que l'on peut noter par rapport à la position relative de la variation dans la protéine sont les suivantes :

- les résidus du cœur hydrophobe de la protéine, garant de la stabilité globale de la protéine,
- les résidus impliqués dans un pont disulfure,
- les résidus impliqués dans une structure secondaire,
- les résidus à la surface et leur polarité, les résidus formant une surface d'interaction,
- les résidus structuralement proches de résidus fonctionnels et impliqués dans leur stabilisation.

Au cours de l'évolution, les résidus fonctionnellement et structuralement importants subissent une forte pression de sélection. C'est pourquoi deux types principaux d'études peuvent être envisagés : les analyses de séquences basées sur la conservation des résidus au cours de l'évolution et les analyses structurales, les deux pouvant être complémentaires.

4.2. Analyse de protéines

Caractériser une protéine dans son contexte évolutif ou dans son contexte structural est devenu courant en biologie moderne, grâce à l'accumulation de données au sein des banques d'une part et aux développements bioinformatiques qui ont été réalisés autour de ces problématiques d'autre part.

4.2.1. Analyse évolutive

L'étude d'une séquence protéique dans son contexte évolutif constitue une étape cruciale dans sa caractérisation et est basée sur l'analyse de sa famille. Après avoir défini la notion de famille de protéines, nous verrons comment retrouver les membres d'une famille à partir d'une protéine d'intérêt et les bases de ces analyses.

4.2.1.1. Famille de protéines

Les séquences d'une même famille protéique, dites homologues, partagent un ancêtre commun. Celui-ci est représenté par sa séquence, à partir de laquelle des mutations se sont accumulées au cours de l'évolution pour aboutir à des protéines distinctes. L'homologie entre deux séquences peut laisser supposer, mais sans le prouver, que les protéines codées par ces gènes assurent une fonction similaire ou proche. Or, l'évolution d'une famille de gènes peut comporter de nombreux évènements. Fitch a introduit dès 1970 (Fitch, 1970) différents termes pour classer les relations d'homologie qui peuvent exister entre deux séquences (Figure 23) :

- les gènes orthologues sont des gènes homologues qui ont évolué à partir d'un gène ancestral suite à un évènement de spéciation,
- les gènes paralogues sont des gènes homologues qui ont évolué à partir d'un gène ancestral suite à un évènement de duplication,
- les gènes xénologues sont des gènes homologues présents au sein d'un même organisme, l'un ayant été hérité « verticalement » et le second par transfert horizontal, c'est-à-dire lors de l'acquisition du matériel génétique d'une autre espèce.

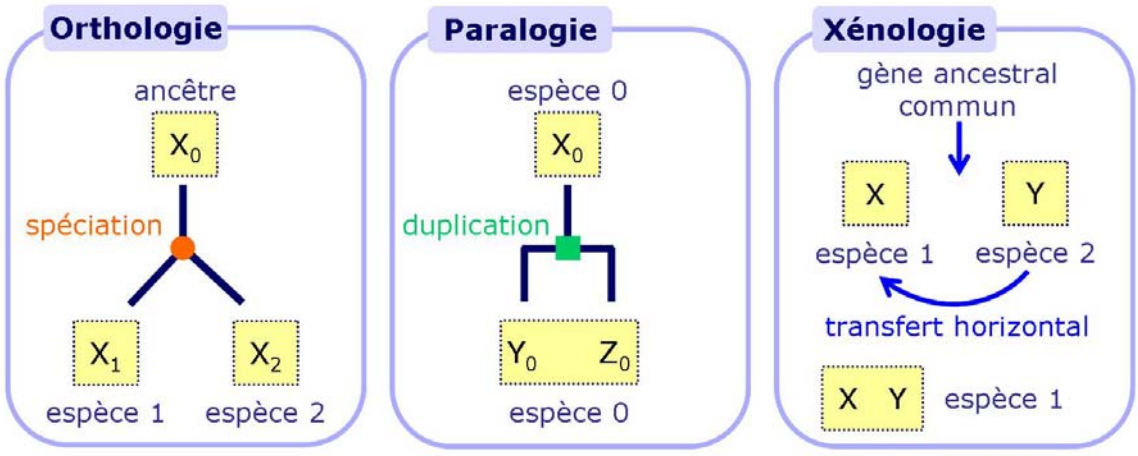


Figure 23 - Relations d'homologie.

Les familles de gènes paralogues ont souvent des fonctions proches, mais pas nécessairement identiques, tandis que les gènes orthologues ont généralement conservé la même fonction. La distinction entre gènes paralogues et orthologues est donc importante dans le cadre d'études évolutives.

4.2.1.2. Identification des membres d'une famille protéique

Typiquement, la mise en évidence des membres d'une famille protéique se fait par l'intermédiaire de recherches de similarité dans les banques de données de séquences. Un certain nombre de logiciels bioinformatiques ont été développés pour réaliser aisément ces recherches. Les programmes les plus fréquemment utilisés sont BlastP (Altschul *et al.*, 1997), Fasta (Pearson and Lipman, 1988) et Psi-Blast (Altschul *et al.*, 1997), ce dernier permettant de détecter des similarités entre des protéines plus éloignées.

Ces programmes fournissent un alignement local des régions les mieux conservées entre la protéine d'intérêt et chacune des protéines détectées au sein de la banque. Pour chaque alignement, un score et une estimation statistique de sa signification biologique, la valeur d'« expect » (ou *E-value*), sont fournis. Les séquences associées à une *E-value* inférieure ou égale à 0,001 sont couramment considérées comme présentant une similarité significative et retenues comme séquences homologues pour la suite des analyses. Ce seuil de 0,001 n'a évidemment pas de valeur absolue mais constitue un compromis, la plupart du temps satisfaisant, entre le nombre de faux positifs et le nombre de faux négatifs. Seules des analyses complémentaires permettront de définir si les séquences sélectionnées dérivent effectivement d'un ancêtre commun.

Un certain nombre de banques de données présentant des informations relatives aux familles de protéines, à leurs domaines fonctionnels, à leurs signatures et organisations dans différents organismes sont également disponibles. InterPro (Mulder *et al.*, 2007), Pfam (Bateman *et al.*, 2004) et Prosite (Hulo *et al.*, 2006) constituent, par exemple, des répertoires de familles de protéines et d'informations complémentaires permettant de disposer d'un très bon aperçu de la composition d'une famille. La banque COG (*Cluster of Orthologous Genes*) (Tatusov *et al.*, 2003) regroupe, quant à elle, des familles d'orthologues provenant de différents génomes complets.

4.2.1.3. Analyse des familles de protéines

La « voie royale » pour l'analyse des familles de protéines est l'alignement multiple de séquences, en particulier quand celles-ci sont alignées sur toute leur longueur pour obtenir un alignement multiple de séquences complètes, ou MACS, acronyme de *Multiple Alignment of Complete Sequences*. Il consiste à aligner plusieurs séquences homologues dans leur intégralité (alignement global) et permet ainsi de replacer une protéine dans le contexte de sa famille. Un grand nombre de questions de la biologie moderne peuvent être adressées par l'intermédiaire du MACS (Lecompte *et al.*, 2001b), qui offre l'avantage de fournir à la fois une vision synthétique des relations évolutives au sein de la famille et de l'organisation en domaines des différents membres (Figure 24). Il représente également une aide précieuse

pour la prédiction et la modélisation de structures secondaires et tertiaires de nouvelles séquences et constitue le point de départ indispensable d'une analyse phylogénétique.

Enfin, il est particulièrement adapté pour l'étude de la conservation des résidus entre les séquences d'une même famille dans sa globalité et permet aussi de mettre en évidence des résidus conservés de manière différentielle entre sous-familles, qui indiquent généralement des résidus de spécificité et révèlent parfois l'existence de groupes de protéines paralogues dont la fonction peut avoir évolué.

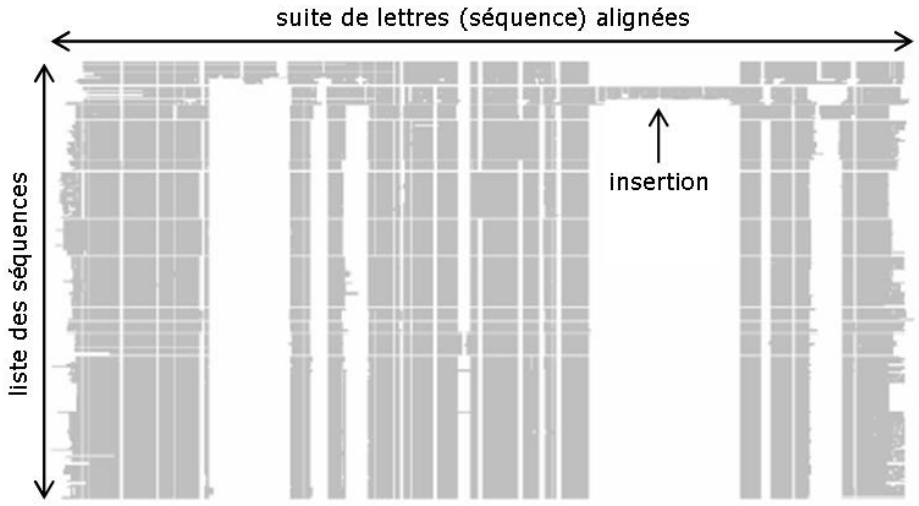


Figure 24 - Vue schématique d'un MACS. L'alignement des séquences sur toute leur longueur peut, selon le cas, nécessiter la création d'insertions dans un groupe de séquences, correspondant à des lacunes, ou *gap* dans d'autres.

De nombreux programmes d'alignement multiple de séquences ont été développés au cours des 20 dernières années (revue dans (Edgar and Batzoglou, 2006)), le plus connu et couramment utilisé étant sans doute ClustalW (Thompson *et al.*, 1994). Les programmes les plus récents combinent plusieurs approches algorithmiques et permettent la construction automatique d'alignements multiples de qualité contenant plusieurs centaines de séquences en quelques minutes.

Nous reviendrons plus en détails sur les avantages des alignements multiples et leur construction dans le Chapitre 10 - *Problématique du contenu informationnel des alignements multiples*, p. 137.

Les ressources présentant des informations relatives aux familles de protéines que nous avons citées au préalable (InterPro, Pfam et Prosite) fournissent, en plus des séquences de la famille, l'alignement de domaines spécifiques des séquences protéiques de la famille. Elles ne présentent cependant que des alignements locaux des régions les mieux conservées et caractéristiques de la famille.

4.2.2. Analyses structurales

Comme nous l'avons déjà évoqué, une protéine assurera son rôle biologique sous une forme tridimensionnelle précise. Au cœur de la cellule, l'environnement des protéines est majoritairement aqueux. Schématiquement, dans le cas d'une protéine globulaire, la chaîne polypeptidique va se replier de manière spontanée, afin d'enfouir les chaînes latérales de ses résidus hydrophobes au cœur de la protéine et de présenter à la surface les chaînes polaires. En première approximation, ces protéines possèdent donc un cœur hydrophobe garant, en partie, de la stabilité de la molécule et une surface hydrophile, l'ensemble adoptant une structure compacte.

Mises à part les interactions hydrophobes, d'autres interactions interviennent dans la stabilisation des structures tridimensionnelles, dont les principales sont :

- les ponts disulfures qui représentent un lien covalent pouvant se former entre deux cystéines structuralement proches,
- les liaisons hydrogène, qui stabilisent également les structures secondaires et qui peuvent impliquer les squelettes polypeptidiques (chaînes principales) et les chaînes latérales,
- les ponts salins qui se forment entre deux acides aminés ionisés.

Lors du repliement de la protéine, sa surface s'organise de façon à présenter les motifs qui peuvent participer de certaines fonctions, à l'exemple des domaines de liaison spécifiques à des molécules. Typiquement, ces domaines de liaison seront localisés dans des cavités, crevasses ou autres types de surfaces, contenant un ensemble de chaînes latérales d'acides aminés disposées de manière à pouvoir créer des liaisons non covalentes avec d'autres molécules (Figure 25).

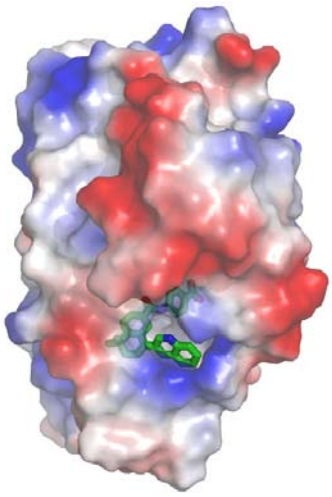


Figure 25 - Structure tridimensionnelle d'une protéine en interaction avec son ligand (PDB : 1DKF).

Ainsi, tous les facteurs impliqués dans la stabilité et la spécificité de surface d'une protéine entrent en jeu pour que celle-ci puisse assurer son rôle fonctionnel de manière précise.

L'ensemble de ces facteurs n'est cependant connu que pour un nombre restreint de protéines. En effet, si, à l'heure de la généralisation des techniques expérimentales à haut débit, l'obtention de la séquence d'une protéine est devenue relativement rapide, la détermination expérimentale de sa structure reste toujours difficile, longue et coûteuse. Pour illustrer ce constat, on peut noter que la principale banque internationale de structures tridimensionnelles, la PDB (Kouranov *et al.*, 2006), contient 45 213 entrées au 14 août 2007, alors que la banque de séquences protéiques UniProt (Consortium1, 2007) contient 4 672 908 entrées au 24 juillet 2007.

Afin de pallier le manque de données expérimentales disponibles, un certain nombre de développements bioinformatiques ont permis de mettre en place des outils de prédiction *in silico* de la structure d'une protéine. Il faut distinguer les méthodes qui visent à prédire localement la structure des protéines des méthodes visant à fournir un modèle global. Ces deux approches sont néanmoins liées puisque les prédictions locales fournissent bien souvent des éléments permettant de complémenter ou de réaliser la prédiction globale.

4.2.2.1. Prédiction de la structure locale des protéines

La séquence en acides aminés d'une protéine encode sa structure tridimensionnelle et il est courant de prédire un certain nombre de caractéristiques structurales à partir de cet enchaînement de résidus.

Les principaux éléments structuraux pouvant être prédits à partir de la structure primaire d'une protéine sont :

- les éléments de structures secondaires, dont la prédiction constitue un vaste domaine d'étude depuis les années 70 (revue dans (Rost, 2001)). Les premières méthodes publiées utilisaient le fait que, dans les structures tridimensionnelles, certains acides aminés soient retrouvés préférentiellement dans certaines classes de structures secondaires ; au courant des années 80, l'environnement local des résidus dans la séquence a été pris en compte. Les méthodes actuelles se sont diversifiées et utilisent couramment l'information de séquences homologues (Heringa, 2000). Elles permettent d'obtenir un taux de bonne prédiction supérieur à 75%. Des approches différentes étant utilisées selon les programmes, il est courant d'obtenir des résultats contradictoires, c'est pourquoi il est conseillé de comparer plusieurs méthodes afin d'obtenir un consensus. Le Tableau 4 présente quelques méthodes (non exhaustives) de prédiction de structures secondaires,
- les régions *coiled-coils*, répétitions d'heptapeptides, qui sont des hélices oligomérisées enroulées les unes autour des autres pour former des super-hélices. Plusieurs

programmes de prédiction des coiled-coils existent, citons COILS (Lupas *et al.*, 1991), PAIRCOIL (Berger *et al.*, 1995) et MULTICOIL (Wolf *et al.*, 1997), par exemple,

- les régions transmembranaires, ou plus précisément les hélices α transmembranaires. Tous les programmes de prédiction reposent sur les profils d'hydrophobicité de la chaîne polypeptidique impliquant 16 à 25 résidus et supposent que leur structuration sera en hélice α (revue dans (Chen *et al.*, 2002)),
- l'accessibilité au solvant des résidus, calculée par des programmes, comme PHDacc (Rost, 1996) ou PROFacc (Rost and Liu, 2003), permettant de prédire si un résidu se situe à la surface de la protéine ou non. PHDacc est basé sur un système de réseaux de neurones et calcule la probabilité d'accessibilité au solvant,
- les régions désordonnées, qui n'adoptent pas de structure tridimensionnelle bien définie et qui peuvent être impliquées dans de nombreux processus biochimiques importants (reconnaissance moléculaire, transduction de signal) (Wright and Dyson, 1999). Différents paramètres, comme la composition en acides aminés, la complexité de la séquence, l'hydrophobicité ou la charge nette, sont pris en compte par les programmes de prédiction disponibles, comme DISOPRED2 (Ward *et al.*, 2004), GlobPlot2 (Linding *et al.*, 2003) et FoldIndex (Prilusky *et al.*, 2005). Une stratégie couramment employée pour améliorer l'efficacité de prédiction des programmes disponibles est d'utiliser plusieurs méthodes performantes et d'en combiner les résultats.

Programmes	Commentaires
GOR (Garnier *et al.*, 1978)	Prédiction de trois états (hélice, feuillet et coude). Repose sur une analyse statistique de la fréquence des résidus observés dans une fenêtre de 17 acides aminés.
NNSSP (Salamov and Solovyev, 1997)	Prédiction de la structure secondaire par homologie avec les séquences de structure connue. Le choix de la matrice d'évaluation est critique, un réseau neuronal combine les résultats de six matrices différentes.
PHD (Rost, 1996)	Combine les informations des alignements multiples avec trois réseaux neuronaux
Predator (Frishman and Argos, 1995)	Méthode particulièrement efficace pour la prédiction des feuillets. Utilise des alignements locaux qui sont ensuite pondérés.
PSIpred (Jones, 1999)	Utilise une matrice de positions générée par Psi-Blast et l'utilise pour son réseau neuronal comme données. Demande l'existence d'homologues pour une prédiction fiable.

Tableau 4 - Programmes de prédiction de structures secondaires.

4.2.2.2. Prédiction de la structure globale des protéines

Les méthodes de prédiction de structure locale représentent souvent une première étape vers la prédiction des structures globales des protéines. Les stratégies de prédiction employées ne sont pas les mêmes suivant les informations disponibles sur la séquence d'intérêt.

La modélisation par homologie est la méthode de choix pour la prédiction de la structure tridimensionnelle d'une protéine d'intérêt. Son principe repose sur le fait que des protéines homologues, qui partagent des séquences fortement similaires, ont des structures tridimensionnelles proches. Il est donc possible d'utiliser la structure connue d'une protéine homologue, qui servira d'empreinte pour modéliser la structure d'une protéine d'intérêt.

La modélisation par homologie se base sur un alignement de qualité entre les deux séquences. Les coordonnées atomiques de l'empreinte sont affectées à la séquence d'intérêt et les chaînes latérales sont positionnées de façon à minimiser l'énergie du système. Le seuil de similarité nécessaire pour réaliser cette modélisation est variable et dépend de la longueur de l'alignement (Sander and Schneider, 1991). On considère néanmoins que les deux séquences doivent partager environ 30% d'identité pour que la modélisation soit de bonne qualité. Il faut savoir, qu'en théorie, plus le pourcentage d'identité entre les deux séquences est élevé, meilleur sera le modèle.

MODELLER (Sali *et al.*, 1995), Geno3D (Combet *et al.*, 2002), SWISS-MODEL (Schwede *et al.*, 2003), 3D-JIGSAW (Bates *et al.*, 2001) sont des exemples d'outils performants de modélisation par homologie.

Lorsqu'il n'existe pas d'empreinte satisfaisante de la protéine d'intérêt, la méthode dite du *threading* (ou enfilage) peut être utilisée pour tenter de prédire la structure. On estime qu'il y a actuellement une chance sur deux pour que le repliement d'une séquence pour laquelle on ne dispose pas d'empreinte satisfaisante soit déjà connu. Cette méthode consiste à utiliser les motifs structuraux connus pour y « enfiler » la séquence. On identifie par *threading* les repliements les mieux adaptés à une séquence en utilisant des profils dérivés de structures tridimensionnelles connues (Bryant and Lawrence, 1993). Au final, après le choix d'un support structural et l'alignement de la séquence cible sur le support, le modèle tridimensionnel est obtenu à l'aide des outils de modélisation par homologie. La difficulté supplémentaire réside dans la mise en évidence de la compatibilité de la séquence avec une structure existante.

Enfin, dans les cas où la modélisation par homologie et le *threading* ne fournissent pas de réponse, soit parce que la séquence à prédire adopte une structure qui n'a jamais été observée, soit parce que la structure compatible n'est pas détectée, l'utilisation de méthode de prédiction *ab initio* est envisageable (Bonneau and Baker, 2001). Celle-ci consiste à rechercher la structure de la séquence d'intérêt en se basant uniquement sur sa séquence : un modèle physique simplifié de la chaîne protéique est construit et une recherche exhaustive

de l'espace conformationnel est réalisée pour obtenir une structure de moindre énergie. Ces constructions sont très gourmandes en ressources informatiques, c'est pourquoi la méthode de prédiction *ab initio* n'est applicable que sur des protéines de petite taille.

Chapitre 5 - Vers une approche intégrative des mutations et de leurs conséquences

Comme nous l'avons vu tout au long de cette introduction, l'ère scientifique dans laquelle nous évoluons ouvre de larges perspectives en termes d'avancées dans la compréhension du vivant dans sa globalité. Une époustouflante quantité de données biologiques hétérogènes ne cesse de s'accumuler au sein de diverses banques de données grâce, notamment, à la mise au point de techniques expérimentales de pointe, mais aussi grâce aux développements bioinformatiques qui ne cessent de s'améliorer. Il est à présent possible de stocker, de centraliser, de gérer et d'analyser cette somme toujours grandissante de données et il devient envisageable de commencer à répondre à des questions biologiques de plus en plus pointues, et ce, en ne se focalisant plus sur l'étude d'un gène ou d'une protéine particulière, mais en étendant les analyses sur des ensembles de gènes ou protéines, voire sur des génomes ou des protéomes entiers.

En termes de recherche biomédicale, dont le but principal est d'obtenir une meilleure compréhension des maladies génétiques humaines afin de pouvoir mettre en œuvre le développement de diagnostics ou de solutions thérapeutiques efficaces, un enjeu primordial est la capacité à comprendre et prédire les effets des variations génétiques sur le phénotype d'un individu les portant. La bioinformatique semble à même de fournir un large choix d'outils de qualité dans cette optique et ses avantages sont multiples : les moyens informatiques disponibles permettent de gérer une vaste quantité de données, hétérogènes de surcroît, ce qui est devenu absolument primordial, d'une part de par le nombre de gènes connus pour être impliqués dans une maladie génétique et, d'autre part, de par la quantité de mutations ou polymorphismes mis en évidence et qu'il faut gérer. De plus, la mise en place de solutions bioinformatiques est, de manière générale, très peu onéreuse et beaucoup plus rapide que le déploiement de techniques et protocoles purement expérimentaux. Il faut toutefois garder à l'esprit que les études bioinformatiques ont comme ambition de faciliter le travail des biologistes, en proposant des hypothèses de travail sur lesquelles il est nécessaire de porter un regard critique. Seule l'expertise biologique et les vérifications expérimentales permettront de valider de manière définitive les hypothèses.

Une difficulté, dans le cadre de l'étude des maladies génétiques humaines, consiste à distinguer, au niveau d'un gène d'intérêt, les mutations délétères qui seront à l'origine de modifications phénotypiques, des variations neutres qui seront sans conséquence.

Disposer de programmes fiables et performants permettrait de traiter le volume de données générées par les programmes de séquençage de SNP à haut débit afin de faciliter, voire d'automatiser, la sélection de SNP dans le cadre d'études génétiques pour fixer des priorités entre les gènes candidats. De tels programmes participeraient également à la compréhension des bases moléculaires des maladies génétiques humaines, étape fondamentale dans le processus de développement de nouvelles solutions thérapeutiques.

Enfin, dans l'idéal, si de nouveaux développements permettaient de disposer de méthodes de prédictions extrêmement fiables, le diagnostic clinique et le conseil génétique pourraient se baser dessus.

Nous présenterons, dans une première section, un certain nombre de méthodes visant à discriminer les mutations délétères des mutations neutres disponibles, avant de présenter le projet MS2PH dans le cadre duquel cette thèse a été réalisée.

5.1. Etat de l'art

Des méthodes d'analyse de protéines dédiées à la prédiction des effets des mutations ont d'ores et déjà été mises à profit dans le développement de systèmes bioinformatiques (Mooney, 2005).

Parmi ces outils de prédiction, rares sont ceux qui ne sont pas spécifiques des SNP non-synonymes, que nous signifierons dorénavant par leur appellation courante, nsSNP. Une approche originale a cependant été rapportée (Fairbrother *et al.*, 2004) pour l'analyse des variations (synonymes ou non) qui perturbent les sites activateurs d'épissage exoniques (ESE pour *Exonic Splicing Enhancer*). Celle-ci est basée sur l'étude des alignements des SNP prédits comme affectant des sites ESE humains. Les auteurs ont montré, par comparaison avec le génome du chimpanzé, que ces SNP subissent une pression de sélection.

Précisons que les mutations affectant des régions non codantes du génome n'ont pas reçu la même attention que celles affectant les régions codantes, non pas parce que ces mutations sont considérées comme moins importantes, mais plutôt à cause de la difficulté de collecter des informations sur les effets de telles mutations. En effet, comprendre comment des variations affectent l'expression des gènes est actuellement considéré comme un challenge majeur en génétique humaine (Hudson, 2003) : aux vues des connaissances actuelles, il n'est pas systématiquement possible de différencier les variations affectant directement la régulation d'un gène des effets environnementaux. Les quelques études qui se sont concentrées sur ce problème ont couplé des approches bioinformatiques à des données expérimentales relatives au niveau d'expression de gènes par l'intermédiaire de puces à ADN (Pastinen *et al.*, 2004; Wittkopp *et al.*, 2004). A l'heure actuelle, estimer si un variant affectant une position prédite comme faisant partie d'un site de régulation d'un gène

modifiera, ou non, son niveau d'expression ne peut se faire qu'en vérifiant la conservation de cette position chez des organismes modèles.

Plusieurs groupes de recherche se sont attelés au développement d'outils de prédiction des effets des nsSNP sur la fonction d'une protéine, avec plus ou moins de succès. Rappelons que les nsSNP comptent pour environ la moitié des mutations délétères chez l'homme (Stenson *et al.*, 2003) et que ces mutations restent les mieux renseignées en termes phénotypiques.

L'efficacité d'une méthode de prédiction repose sur le choix des facteurs prédictifs ainsi que sur la méthode de prise de décision.

Les méthodes actuelles de prédiction peuvent être schématiquement divisées en deux catégories : celles exclusivement basées sur des données de séquences et celles qui combinent des données de séquences et de structure et qui dépendent par conséquent généralement de la disponibilité de données structurales.

Les facteurs prédictifs les plus couramment utilisés sont la conservation des résidus au cours de l'évolution (Bromberg and Rost, 2007; Chasman and Adams, 2001; Ng and Henikoff, 2003; Saunders and Baker, 2002), la compatibilité des résidus sauvages et mutés (Bromberg and Rost, 2007; Chasman and Adams, 2001; Ng and Henikoff, 2003; Ramensky *et al.*, 2002; Saunders and Baker, 2002), la proximité d'un site fonctionnel (actif, liaison au ligand) (Chasman and Adams, 2001; Ramensky *et al.*, 2002; Saunders and Baker, 2002), le changement de charge ou d'hydrophobicité engendré par la substitution d'un résidu enfoui (Chasman and Adams, 2001; Ramensky *et al.*, 2002; Saunders and Baker, 2002), l'impact d'un changement dans le cœur hydrophobe sur la solubilité de la protéine, la destruction d'un pont disulfure ou l'insertion d'une proline dans une hélice (Bromberg and Rost, 2007; Chasman and Adams, 2001; Ramensky *et al.*, 2002; Saunders and Baker, 2002) ainsi que la rigidité moléculaire (Chasman and Adams, 2001; Saunders and Baker, 2002).

Les approches considérées pour la prise de décision sont très diverses et nous ne les détaillerons pas. Elles incluent l'utilisation de fonctions de score (Ng and Henikoff, 2003), de règles de décision (Ramensky *et al.*, 2002), de machine à vecteurs de support (Vapnik, 1998) ou SVM pour *Support Vector Machine* (Bao and Cui, 2005) et de réseaux de neurones (Bromberg and Rost, 2007).

Précisons qu'une difficulté majeure lors du développement de méthodes prédictives réside dans la définition des jeux de données initiales et, en particulier, dans le choix du jeu d'allèles neutres utilisé comme élément de comparaison (Care *et al.*, 2007). En effet, la plupart des approches ont été développées en se basant sur des jeux de mutations neutres qui contiennent par définition un certain nombre de faux-positifs : les banques de SNP, qui ne contiennent pas de lien au phénotype ; des jeux de données basés sur des résidus acceptés dans des espèces proches, alors que certaines mutations délétères chez l'homme sont tout à

fait tolérées dans des organismes proches (Kondrashov *et al.*, 2002) ; des données issues d'expériences de mutagénèse, pour lesquelles les effets délétères nécessitant des conditions environnementales précises peuvent ne pas être décelés.

A l'heure actuelle, les trois méthodes qui fournissent les résultats les plus satisfaisants semblent être SIFT (Ng and Henikoff, 2003), PolyPhen (Ramensky *et al.*, 2002) et SNAP (Bromberg and Rost, 2007).

SIFT (*Sorting Intolerant From Tolerant*) est exclusivement basé sur des données de conservation entre séquences homologues et la construction d'une matrice de substitution spécifique des positions dans la séquence (PSSM pour *Position-Specific Substitution Matrice*). SIFT calcule des probabilités normalisées pour toutes les substitutions possibles à chaque position et les probabilités inférieures à un seuil prédéfini seront considérées comme délétères. SIFT autorise la soumission d'une séquence d'intérêt et utilise les informations de l'alignement multiple, lui aussi soumis ou construit automatiquement, pour prédire si une substitution sera tolérée ou non à chaque position de la séquence. Le principal avantage d'une méthode basée exclusivement sur des données de séquences est qu'elle ne nécessite pas l'accès à des données structurales et peut par conséquent être appliquée à un nombre plus important de protéines. L'inconvénient majeur réside dans la réduction des capacités de prédiction lorsque le nombre d'homologues est limité.

PolyPhen (*Polymorphism Phenotyping*) utilise des données de séquences ainsi que de structure, quand la structure expérimentale de la protéine d'intérêt ou d'un homologue partageant plus de 50% d'identité est disponible. La position de la substitution est caractérisée par :

- sa conservation au cours de l'évolution, un score de substitution est calculé en utilisant le logiciel PSIC (*Position-Specific Independent Counts*) (Sunyaev *et al.*, 1999),
- ses annotations fonctionnelles extraites de Swiss-Prot,
- sa cartographie sur la structure tridimensionnelle d'une protéine proche disponible.

Les substitutions sont classées en trois niveaux : bénigne, potentiellement délétère et probablement délétère. Les règles de décision appliquées sont présentées dans le Tableau 5.

	Règles			Prédiction	Base de la prédiction
	score PSIC	propriété fonctionnelle	Propriété de la substitution		
1	indifférent	annoté comme site fonctionnel	indifférent	probablement délétère	annotation de séquence
2	indifférent	annoté comme site de formation de liaison	indifférent	probablement délétère	annotation de séquence
3	indifférent	dans région TM annotée	Score PHAT de la substitution < 0	potentiellement délétère	annotation de séquence
4	indifférent	dans région TM prédite		potentiellement délétère	prédiction de séquence
5	≤ 0,5	indifférent	indifférent	bénin	alignement multiple
6	> 1,0	à moins de 3 Å d'un ligand	indifférent	probablement délétère	structure
7		à moins de 3 Å de résidus fonctionnels	indifférent	probablement délétère	structure
8	dans l'intervalle (0,5 - 1,5)	accessibilité ≤ 15%	changement propension surface accessible ≥0,75	potentiellement délétère	structure
9			changement volume chaîne latérale ≥60	potentiellement délétère	structure
10			changement volume chaîne latérale ≤-60	potentiellement délétère	structure
11		accessibilité ≤ 5%	changement propension surface accessible ≥1,0	probablement délétère	structure
12			changement volume chaîne latérale ≥80	probablement délétère	structure
13			changement volume chaîne latérale ≤-80	probablement délétère	structure
14	dans l'intervalle (1,5 - 2,0)		changement propension surface accessible ≥1,0	probablement délétère	structure
15			changement volume chaîne latérale ≥80	probablement délétère	structure
16			changement volume chaîne latérale ≤-80	probablement délétère	structure
17		indifférent	indifférent	potentiellement délétère	structure
18	> 2,0	indifférent	indifférent	probablement délétère	alignement multiple

Tableau 5 - Règles de prédiction respectées par PolyPhen. TM : transmembrane ; PHAT : matrice de substitution spécifique des régions transmembranaires.
(adapté de http://genetics.bwh.harvard.edu/pph/pph_help.html)

SNAP (*Screening for Non-Acceptable Polymorphism*) est un outil de prédiction qui se base exclusivement sur des données de séquences. Il incorpore, d'une part, des données relatives à la conservation des résidus au sein de la famille et il réalise, d'autre part, des prédictions de structures locales (structures secondaires et accessibilité au solvant) à partir de la séquence

protéique. Dans une version dérivée, SNAP^{annotated}, le programme incorpore également des annotations de Swiss-Prot, ainsi que les prédictions de SIFT.

Le Tableau 6 fournit une idée des performances de chacune des méthodes présentées. Les résultats obtenus sont très prometteurs et attestent de la puissance de chacune de ces méthodes. En effet, le pourcentage de bonnes prédictions se situe entre 74 et 79% selon la méthode considérée. On remarque que SIFT et PolyPhen offrent une sensibilité moins élevée que SNAP (63,4 et 66,9% contre 80,2%), mais leur spécificité est supérieure. Les résultats des quatre méthodes testées restent ainsi comparables dans leur globalité. Nous reviendrons succinctement sur les performances de ces méthodes dans le cadre du Chapitre 14 - *Vers une typification des mutations*, p. 203.

programme	non-traité	délétère bien prédit / total prédit	délétère bien prédit / total délétère (sensibilité)	neutre bien prédit / total prédit	Neutre bien prédit / total neutre (spécificité)	Proportion de SNP bien prédits
SIFT	2374	79,8 ± 0,6	63,4 ± 1,2	70,1 ± 2,7	84,3 ± 1,2	74,0 ± 1,4
PolyPhen	1647	79,1 ± 0,7	66,9 ± 1,4	71,8 ± 2,7	82,7 ± 1,1	74,9 ± 1,3
SNAP	0	76,7 ± 0,7	80,2 ± 0,9	79,8 ± 2,7	76,2 ± 2,2	78,2 ± 1,3

Tableau 6 - Performance des méthodes de prédiction des effets délétères de substitution sur un jeu de nsSNP. Le jeu de données initiales est constitué de 39 987 nsSNP délétères et 40 830 nsSNP neutres. (adapté de (Bromberg and Rost, 2007))

En principe, les données structurales devraient apporter une aide à la prédiction des effets des mutations. Cependant, en pratique, traduire une connaissance structurale sur le phénotype est loin d'être trivial et les méthodes basées exclusivement sur la séquence restent très puissantes. Ceci démontre une nouvelle fois la robustesse de l'étude de la conservation des résidus dans les protéines homologues et reflète l'importance structurale et fonctionnelle des contraintes évolutives subies par la protéine. Tous s'accordent pour dire que l'information évolutive est le facteur prédictif le plus puissant, mais que les informations structurales sont indispensables lorsque peu de séquences homologues sont disponibles.

Les méthodes fondées sur des données structurales expérimentales, qui demandent de disposer de la structure de la protéine ou d'une structure proche, sont aujourd'hui limitées par le fait que peu de structures soient disponibles, comparativement aux séquences. Il est cependant légitime de penser, aux vues du développement de projets de génomique structurale (Berman and Westbrook, 2004) ou de projets de modélisation de structure tridimensionnelles de génomes complets (Peitsch *et al.*, 2000), que le nombre de structures résolues expérimentalement va augmenter de manière exponentielle. L'applicabilité de ces méthodes n'en sera qu'accrue. De plus, il est possible de disposer d'informations structurales

grâce à la seule séquence protéique, comme l'accessibilité au solvant ou les éléments de structures secondaires.

Les méthodes de prédiction de l'effet délétère des mutations d'ores et déjà disponibles permettent d'obtenir des résultats performants. Si des expérimentalistes se basent sur certaines de ces méthodes pour faciliter leurs recherches (Cohen *et al.*, 2004; Letourneau *et al.*, 2005), ils ne sont pour le moment *a priori* qu'assez peu nombreux. Certains d'entre eux déplorent le manque de contrôle qu'il est possible d'appliquer sur la « boîte noire » que représente l'outil de prédiction. Clarifier les bases de la prise de décision, voire les rendre ajustables par l'utilisateur, et expliciter la prédiction par rapport à ces dernières pourraient aider les chercheurs dans leurs démarches et leur permettre d'associer leur propre jugement à la prise de décision.

Ces méthodes de prédiction sont, de plus, assez peu adaptées pour être à même d'aider directement au diagnostic clinique. En effet, elles ne prennent pas en compte la diploïdie de l'individu porteur de la/des mutations (homo/hétérozygotie) ni la relation des allèles entre eux (dominance/récessivité). Les effets de la mutation sur la protéine ne peuvent par conséquent pas être directement extrapolés aux effets cliniques sur l'individu.

A notre connaissance, seul le serveur GenoPred (www.genopred.uvsq.fr), développé spécifiquement autour de l'hypophosphatasie, maladie liée à un déficit en phosphatase alcaline non tissu-spécifique (Mornet and Simon-Bouy, 2004), aborde ces problématiques. Sans constituer un outil de diagnostic clinique, ce système, qui s'appuie sur des modèles statistiques robustes et se base sur des analyses de conservation de séquences aussi bien que sur des données fonctionnelles et structurales, permet la prédiction d'un phénotype clinique précis à partir de la description des deux allèles d'un patient.

5.2. Projet MS2PH

C'est dans ce contexte scientifique général que s'inscrit le projet MS2PH, acronyme de « de la Mutation Structurale aux Phénotypes des Pathologies Humaines », dans le cadre duquel les travaux présentés dans cette thèse ont été réalisés.

Comme nous venons de l'aborder dans cette introduction, la fonction de la plupart des gènes est déterminée par la structure tridimensionnelle du produit de leur expression. En se concentrant sur les gènes impliqués dans des maladies génétiques et en étudiant les informations relatives à leur structure, nous devrions théoriquement être capables d'expliquer pourquoi différentes mutations dans le même gène sont à l'origine de phénotypes distincts.

MS2PH a pour but, à terme, le développement d'outils bioinformatiques et la mise à disposition de sources d'informations permettant de faciliter la compréhension des relations qui existent, au niveau des protéines impliquées dans les maladies génétiques humaines,

entre la séquence de la protéine, son évolution, sa structure tridimensionnelle et les pathologies qui y sont associées.

L'ampleur de ce projet nécessite donc le traitement d'un grand nombre de données hétérogènes ainsi que des ressources informatiques puissantes, pour pouvoir intégrer et analyser la grande quantité de données. C'est pourquoi MS2PH fait partie des projets pilotes du programme Décrypthon.

5.2.1. Programme Décrypthon

Le programme Décrypthon (www.decrypthon.fr) a été initié par l'AFM (Association Française contre les Myopathies), le CNRS et la société informatique IBM (*International Business Machine*), pour répondre aux besoins émergeants de projets scientifiques de grande envergure qui nécessitent des traitements informatiques à haut débit et l'intégration de données multiples. Le programme Décrypthon a pour ambition de mettre à la disposition des biologistes une plateforme technologique puissante adaptée à leurs besoins. Cette plateforme est notamment constituée d'une grille de calcul permettant d'accélérer les analyses génomiques et protéomiques à haut débit et d'un espace de gestion et de stockage de données hétérogènes, interconnecté à la grille de calcul, permettant de réaliser aisément des analyses intégratives puissantes.

De par son rôle de projet pilote, les besoins informatiques, tant en termes de puissance de calcul que de gestion et de stockage des données inhérentes à MS2PH, ont été pris en compte lors du déploiement de l'infrastructure Décrypthon, que nous décrirons plus en détails dans le Chapitre 6 - *Ressources informatiques et bioinformatiques*, p. 80.

5.2.2. Stratégie déployée dans le cadre de la thèse

MS2PH s'appuie sur une approche globale de la mutation affectant la séquence codante d'un gène et ses conséquences structurales et phénotypiques.

La démarche suivie, dans le cadre des études présentées, peut être découpée en quatre étapes :

(i) la caractérisation des résidus d'une séquence protéique dans un contexte évolutif et le développement de stratégies basées sur l'analyse d'alignements multiples de séquences complètes,

(ii) la caractérisation des résidus d'une séquence protéique, dans un contexte fonctionnel et structural, et le développement d'un environnement de travail permettant une approche intégrative des données de séquences, de fonction, de structure et d'évolution,

(iii) l'association de données de mutation et de données phénotypiques aux informations de séquence/fonction/structure/évolution relatives à des protéines d'intérêt et le développement d'une banque de données dédiée aux protéines impliquées dans les maladies génétiques humaines permettant une approche intégrative de l'ensemble de ces données,

(iv) l'analyse préliminaire des mutations affectant deux protéines d'intérêt, la phosphatase alcaline non tissu-spécifique et la myotubularine, pour lesquelles les relations génotype/phénotype sont particulièrement bien documentées. La typification de leurs mutations a pour objectif la mise en place des premières règles d'association entre protéines, mutations, impacts structuraux et phénotypes, afin d'ouvrir la voie vers la mise en place d'un système générique de prédiction des effets d'un génotype sue le phénotype clinique d'un individu.

Matériel & Méthodes

Chapitre 6 - Ressources informatiques et bioinformatiques

A l'ère post-génomique, et dans un contexte où les études à haut débit se généralisent, nous faisons face à d'énormes quantités d'informations à traiter, qui évoluent et s'enrichissent quotidiennement. Dans ces conditions, l'analyse bioinformatique doit faire appel à des banques de données régulièrement mises à jour, à l'automatisation de certains traitements ainsi qu'au stockage organisé des résultats obtenus.

Au cours de nos travaux, nous avons non seulement pu bénéficier de l'infrastructure existante au Laboratoire de Bioinformatique et Génomique Intégratives et des services fournis par la plateforme de bioinformatique de Strasbourg (http://bips.u-strasbg.fr/), récemment labellisée ISO 9001, mais aussi de l'infrastructure offerte par le programme Décrypthon.

Dans ce chapitre, les équipements informatiques dont nous avons pu profiter seront décrits avant d'introduire succinctement le savoir-faire bioinformatique du laboratoire. Nous présenterons ensuite les principales banques de données sur lesquelles nos études se sont appuyées, ainsi que les moyens mis à notre disposition pour les interroger. Nous terminerons ce chapitre en nous intéressant plus précisément aux outils utilisés dans le cadre des études relatives aux alignements multiples de séquences protéiques d'une part et aux structures tridimensionnelles de protéines d'autre part.

6.1. Équipement et ressources informatiques

L'équipement informatique et son bon fonctionnement tiennent une part importante dans le déroulement de travaux de recherche et plus particulièrement lors d'études bioinformatiques à haut débit. Les paragraphes suivants décrivent l'environnement informatique dans lequel l'ensemble des travaux présentés dans ce manuscrit a été réalisé, le nom usuel des machines étant indiqué en italique.

6.1.1. IGBMC

Nous disposons à l'IGBMC de trois serveurs dédiés à la bioinformatique et aux calculs intensifs.

Titus est le support du serveur Web ainsi que des services interactifs fournis par la plateforme de bioinformatique de Strasbourg. C'est un serveur SUN quadri-processeurs,

d'un Gigaoctet (Go) de mémoire vive. Son architecture est de type 64 bits et il dispose d'une installation Solaris 9.

Les deux autres serveurs sont des serveurs de calcul. *Beaufort* est un cluster de six machines Compaq ES40 quadri-processeurs. Cinq machines disposent de 4 Go de mémoire vive et la sixième de 16 Go. Le serveur *Star* est quant à lui composé d'un cluster de six serveurs SUN Enterprise V40z et de deux serveurs SUN M2200. Deux serveurs sont dédiés au stockage des données et à l'accès disque et utilisent le système d'exploitation Solaris 10. Les six autres serveurs utilisent la version Fedora Linux 7 et sont dédiés aux calculs. Le système total est constitué de huit quadri-processeurs Opteron disposant de 4, 16 ou 32 Go de mémoire vive selon les machines.

Le serveur de disque est basé sur un équipement Sun V480 (Solaris 9), qui dispose de 8 Téraoctets (To) d'espace disque, partagé avec les autres serveurs *via* NFS (*Network File System*).

6.1.2. Infrastructures supplémentaires dans le cadre du programme Décrypthon

L'augmentation constante du volume de données à traiter et les développements algorithmiques toujours plus complexes nécessitant des temps de calcul souvent longs impliquent que de nombreux projets scientifiques demandent, pour être menés à bien, de disposer d'une grande puissance de calcul et d'espace de stockage considérable.

Les technologies dites de « grille », encore appelées *Grid Computing*, peuvent idéalement répondre à ces besoins émergeants en biologie. En effet, elles permettent de partager, de manière sécurisée, un ensemble de ressources informatiques hétérogènes et délocalisées, pouvant être représenté par des ordinateurs personnels ou des supercalculateurs. Ces ressources sont mises en réseau et partagées grâce à des solutions logicielles dédiées. Elles peuvent ainsi générer, à un instant donné, un système virtuel doté d'une très grande puissance de calcul et d'une capacité de stockage en rapport.

L'objectif du programme Décrypthon est d'aider à accélérer la recherche en sciences du vivant et, plus particulièrement, en protéomique et génomique, en offrant aux scientifiques un accès à la technologie de grille. L'infrastructure mise en place s'appuie d'une part sur une grille dite « universitaire » constituée de supercalculateurs, et d'autre part sur une grille dite « d'internautes », the *World Community Grid*. Cette dernière est basée sur l'utilisation des ordinateurs personnels d'internautes souhaitant participer au programme Décrypthon et mettant à disposition des scientifiques des « cycles de calcul » non utilisés par leur machine.

Notre implication dans le programme Décrypthon à travers le projet MS2PH, nous a permis de bénéficier de la puissance de la grille de calcul universitaire Décrypthon, ainsi que des serveurs de stockage, qui seront décrits au cours des deux prochaines sections.

6.1.2.1. Grille de calcul universitaire Décrypthon

La grille de calcul universitaire du programme Décrypthon est constituée de supercalculateurs fournis par IBM aux centres universitaires et organismes contributeurs de Bordeaux I, Lille I, Paris VI, Orsay, l'ENS (Ecole Normale Supérieure) de Lyon et du CRIHAN (Centre de Ressources Informatiques de HAute-Normandie) à Rouen. Ces serveurs sont dotés des technologies IBM Power 4+ et Power 5 de dernière génération.

La grille est pilotée par un serveur central situé à l'université d'Orsay. L'ensemble de ces ressources est connecté par le réseau à haut débit RENATER 2 (Réseau National de Télécommunications pour la Technologie, l'Enseignement et la Recherche) qui relie les établissements français d'enseignement supérieur et de recherche (Figure 26).

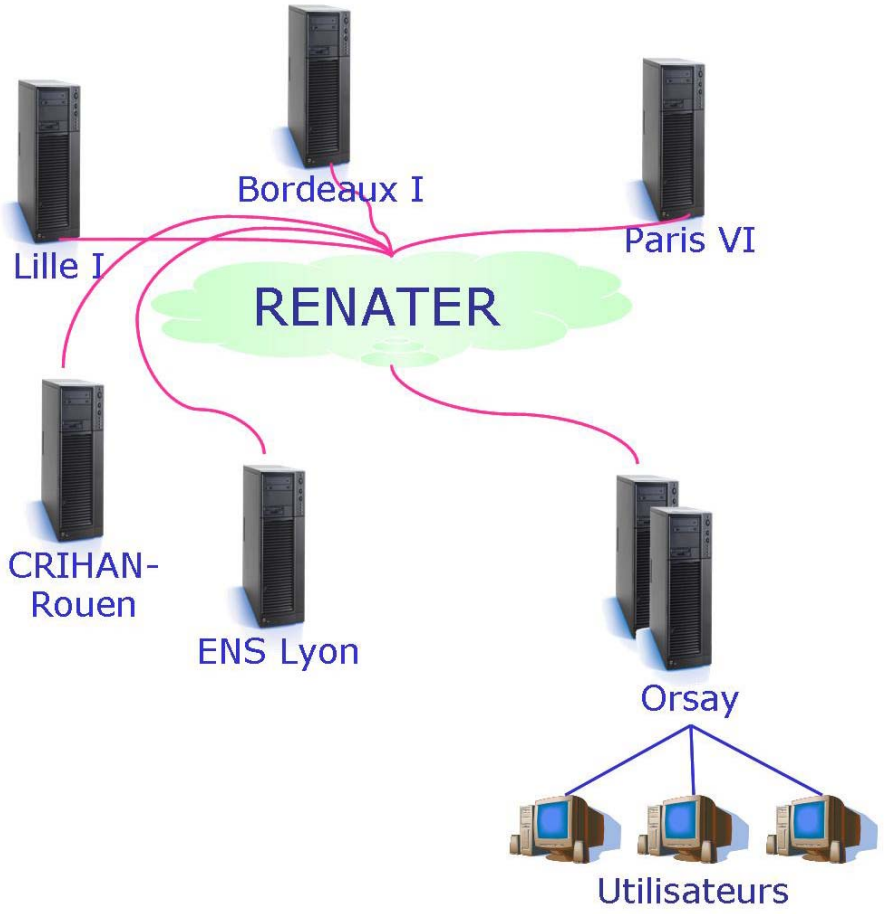


Figure 26 - Schématisation de la grille de calcul Décrypthon.

La puissance de calcul initiale de la grille universitaire est de 500 Gigaflops, répartis sur 112 nœuds de calcul, auxquels peuvent s'ajouter, si nécessaire, une partie de la puissance des machines déjà présentes dans les universités et organismes partenaires.

Des calculs fortement couplés entre eux peuvent y être traités en parallèle. Le fait que cette grille soit constituée d'un réseau à haut débit, disponible 24 heures sur 24, et dotée de machines puissantes (plusieurs Gigaoctets de mémoire vive par poste) permet le partage et l'échange, à une fréquence élevée, de données communes de taille importante.

La suite logicielle Grid MP™ de la société United Devices a, dans un premier temps, été utilisée pour piloter l'intégralité des ressources de la grille. Grid MP™ permet d'optimiser la planification, l'exécution et le suivi des programmes informatiques de chaque projet scientifique utilisant la puissance de la grille. Un logiciel de gestion de grille appelé DIET, pour *Distributed Interactive Engineering Toolbox*, modulaire et adaptable est voué à le remplacer dans le cadre d'une collaboration avec l'ENS de Lyon.

6.1.2.2. Serveurs de stockage et de fédération des données

De nombreux programmes déployés sur la grille de calcul demandent l'accès à diverses banques de données, dont les tailles interdisent généralement un transfert entre chaque nœud d'une grille de calcul. C'est en partie pour pallier ce problème que la plateforme Décrypthon s'est dotée d'une baie de stockage de 20 To, installée au CRIHAN, qui supporte le Centre de Données Décrypthon (*voir* section 6.3.3.2, p. 89). Les nœuds de calcul de la grille peuvent communiquer directement avec ce dernier en passant par le réseau RENATER 2. Cette baie de stockage est également destinée à l'archivage des résultats obtenus sur la grille de calcul.

Enfin, le serveur de l'ENS de Lyon est doté du logiciel *Websphere Federation Server*, destiné à la gestion du Centre de Données Décrypthon ainsi qu'à la fédération d'autres banques de données, publiques ou privées. Ce système permet en effet de se connecter, à distance, à des banques personnelles, dont les données ne seront pas rapatriées. Le gestionnaire de ces banques décidera des champs auxquels il permet l'accès par l'intermédiaire de la grille, assurant un respect des règles de confidentialité, notamment dans le cadre de la manipulation de données de patients.

Ces serveurs vont ainsi contribuer à l'augmentation du volume de données qui pourront être traitées simultanément sur la grille.

6.2. Savoir-faire bioinformatique du laboratoire

6.2.1. GScope : l'ossature bioinformatique du laboratoire

GScope (Ripp *et al.*, manuscrit en préparation) est une plateforme logicielle de manipulation, de visualisation et d'analyse des données massives, initialement mise en place au laboratoire pour répondre aux besoins en ressources bioinformatiques nécessaires lors de l'annotation du génome de *Pyrococcus abyssi* (Lecompte *et al.*, 2001a). GScope est depuis devenu une

plateforme génomique à part entière, dédiée à l'étude de données massives, qu'il s'agisse de génomes complets ou simplement d'ensembles de gènes ou de protéines.

Le gène et son produit étant au cœur de nombreux projets biologiques, GScope permet d'effectuer des traitements bioinformatiques universels, basés sur des programmes que l'on peut considérer comme « classiques » et applicables à tous types de projets (comme les traitements liés aux séquences), ainsi que sur des outils développés au sein du laboratoire. L'architecture de GScope, programmée en Tcl/Tk, permet l'utilisation de nouveaux modules, dédiés à des activités d'analyses spécifiques, en intégrant des programmes externes ainsi que les nouveaux développements réalisés par les membres du laboratoire.

Nous avons ainsi pu bénéficier de l'expérience du laboratoire dans le développement d'outils intégrés à GScope et contribuer à son expansion : l'ensemble des développements bioinformatiques que nous avons été amenés à réaliser au cours de nos travaux ont vu le jour sous l'environnement GScope, en particulier, les protocoles déployés pour être implémentés sur la grille de calcul ont été élaborés et minutieusement testés et validés au sein de GScope.

Un serveur Web communicant avec la plateforme GScope par l'intermédiaire d'une interface CGI (*Common Gateway Interface*) a également été mis en place. Il permet de partager les résultats des analyses avec la communauté scientifique ou des collaborateurs externes.

6.2.2. Ecriture de programmes : le Tcl/Tk

Afin de rester cohérent avec les développements réalisés au sein de la plateforme GScope, le langage de programmation utilisé au sein du laboratoire est principalement le Tcl/Tk (http://www.tcl.tk).

Le Tcl/Tk a l'avantage d'être un langage de script interprété, qui ne nécessite par conséquent aucune compilation. Il permet de plus le déploiement rapide de nouvelles applications, qui seront parfaitement portables sur n'importe qu'elle plateforme informatique (Windows, MacOS et tout système Unix ou Linux).

Le Tcl/Tk est composé de deux parties : le Tcl (pour *Tool Command Language*), un langage de commande, et le Tk (pour *ToolKit*), extension graphique de haut niveau. Le Tcl/Tk est tout à fait adapté à la lecture et à la manipulation de fichiers, toutes les variables étant considérées comme des chaînes de caractères. De plus, le Tk en fait un langage de programmation qui permet de créer aisément des interfaces graphiques très complètes et indispensables pour le biologiste.

L'utilisation de ce langage s'est révélée cruciale pour l'implémentation de nos programmes sur la grille de calcul Décrypthon : sa syntaxe riche et simple ainsi que sa souplesse, associée à la modularité de nos développements, nous ont permis de réaliser des transferts aisés et rapides en termes d'adaptation aux besoins de la grille.

6.3. Les banques de données biologiques

Une condition indispensable au bon déroulement d'analyses bioinformatiques est de pouvoir disposer d'un accès aisé à des banques de données biologiques régulièrement mises à jour.

Nous décrirons en détail dans cette section les principales banques de données utilisées au cours de ce travail.

6.3.1. Banques de données généralistes

Au cours de notre travail, nous avons exploité de façon routinière plusieurs banques que l'on peut qualifier de généralistes, installées localement sur les serveurs de l'IGBMC et mises à jour automatiquement par la plateforme bioinformatique de Strasbourg.

6.3.1.1. UniProt

UniProt (*United Protein Databases*) (Wu *et al.*, 2006) constitue la réunification de deux banques de séquences protéiques majeures, Swiss-Prot et TrEMBL, en une seule et unique ressource.

La banque Swiss-Prot (Boeckmann *et al.*, 2003), créée en juillet 1986 par Amos Bairoch à Genève, est maintenue et distribuée conjointement par le SIB (*Swiss Insitute of Bioinformatics*) et l'EBI (*European Bioinformatics Institute*). Cette banque de données de séquences protéiques se veut la référence en matière de banques de protéines annotées et représente une véritable exception dans le monde des banques de séquences généralistes : elle privilégie en effet la qualité et la richesse des annotations à l'exhaustivité de sa collection de séquences et présente une redondance minimale, les différentes versions d'une même entrée étant fusionnées.

Chaque protéine entrée dans Swiss-Prot est directement prise en charge par les experts en annotation de cette banque, permettant de maintenir son contenu cohérent et homogène. Les séquences présentes dans cette banque proviennent de quatre origines : la traduction des gènes annotés dans la banque EMBL, banque européenne de référence de séquences nucléotidiques ; certaines protéines issues d'autres banques protéiques ; la consultation de publications scientifiques ; quelques soumissions directes par les auteurs.

Une des caractéristiques principales de Swiss-Prot est la qualité extrêmement élevée de ses annotations. Les différentes catégories d'informations figurant dans la banque comprennent notamment, pour chaque entrée, la ou les fonctions des protéines, les modifications post-traductionnelles connues, les sites et domaines structuraux ou fonctionnels identifiés, des références bibliographiques, les structures secondaires et quaternaires, les similarités avec d'autres protéines, les positions conflictuelles ainsi que des mutations connues associées avec d'éventuelles pathologies, etc. Un des atouts supplémentaires de cette banque repose sur l'introduction d'un nombre très important de références croisées avec d'autres banques

de données (nucléiques, de structure, de motifs, etc.). Cette spécificité se révèle particulièrement utile avec le développement des systèmes d'interrogation de banques. Le flot grandissant de nouvelles séquences issues de projets de séquençage de génomes pénalise cependant la banque Swiss-Prot, dont la richesse des annotations ne peut suivre une telle explosion de données. Ces éléments se reflètent parfaitement dans la croissance quasi linéaire du nombre d'entrées dans cette banque (Figure 27).

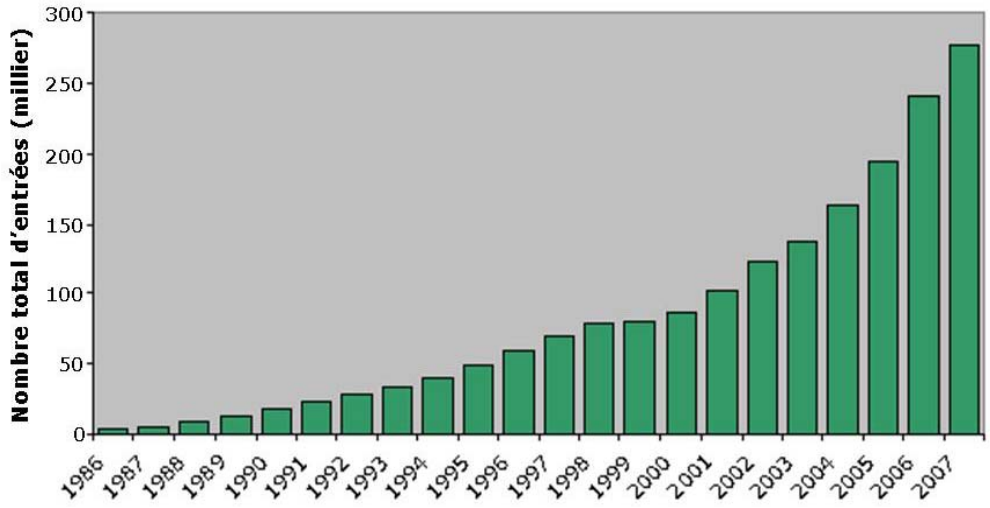


Figure 27 - Evolution du nombre d'entrées de la banque Swiss-Prot depuis 1986. Les données sont extraites du site http://www.expasy.org/sprot/relnotes/#SPstat.

Introduite en 1996, la banque TrEMBL (Boeckmann *et al.*, 2003) pour *Translation from EMBL*, est décrite comme un supplément de Swiss-Prot. En effet, TrEMBL, distribuée par l'EBI, contient la traduction de toutes les parties codantes annotées figurant à l'EMBL, banque européenne de référence de séquences nucléotidiques, à l'exception des protéines figurant dans Swiss-Prot. TrEMBL constitue ainsi un dépôt de séquences non validées et faiblement annotées, qui vont par la suite être examinées par les annotateurs de Swiss-Prot pour y être intégrées. Sa croissance est fulgurante (Figure 28) et correspond aux flux importants de données générés par les projets de séquençage des génomes complets.

TrEMBL se caractérise ainsi par une redondance importante, des séquences souvent incomplètes, annotées automatiquement. Cette banque de données peut désormais être définie comme un complément de Swiss-Prot.

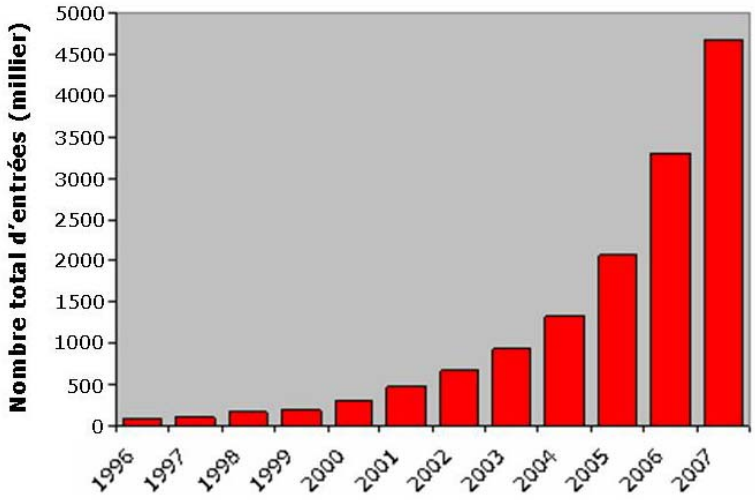


Figure 28 - Evolution du nombre d'entrées de la banque TrEMBL depuis 1996. Les données sont extraites du site http://www.ebi.ac.uk/trembl/Documents/old_trembl_rel_notes.html.

6.3.1.2. UniRef

Les banques UniRef, pour *UniProt Reference clusters* (Suzek *et al.*, 2007), représentent des banques de données non redondantes issues de la banque UniProt. Les séquences y sont regroupées sur la base de leur pourcentage d'identité.

UniRef100 regroupe, dans un seul enregistrement, les séquences identiques et les fragments qu'elles contiennent au sein d'un même organisme.

UniRef90 et UniRef50 regroupent les séquences de tous les organismes sur la base d'un seuil de similarité de 90% et 50% respectivement. Une séquence est choisie comme représentant pour chaque entrée et des références croisées permettent d'avoir accès aux autres séquences du groupe. Ces banques constituent ainsi un ensemble de séquences non redondantes, où chaque séquence est définie comme un cluster de séquences partageant au moins 90% ou 50% d'identité.

L'utilisation des banques UniRef permet d'accélérer les recherches de similarité, mais souvent aussi d'augmenter leur significativité. Ces banques répondent en effet à un besoin de « simplification » de l'information, qui fait écho à la quantité de données actuellement disponibles et à la redondance trouvée dans les banques. Cette redondance peut être illustrée non seulement par la présence de séquences strictement identiques, mais aussi par une « redondance fonctionnelle » entre les séquences proches, qui fait que l'information pertinente pour le biologiste doit parfois être récupérée au sein de séquences plus éloignées.

6.3.1.3. PDB

La PDB (*Protein Data Bank*) (Kouranov *et al.*, 2006) est la principale banque internationale de structures tridimensionnelles. Cette banque a été fondée en 1971 par le BNL (*Brookhaven National Laboratory*) et contenait sept structures. Elle est depuis 1998 sous la tutelle du RCSB (*Research Collaboratory for Structural Bioinformatics*).

Les structures de protéines constituent l'essentiel des entrées de la PDB avec 45 213 entrées au 14 août 2007, mais on y trouve également des structures de molécules d'ARN ou d'ADN (1 770) et de complexes protéines-acides nucléiques (1 825). Ces structures sont déterminées expérimentalement par cristallographie aux rayons X, par Résonance Magnétique Nucléaire (RMN) ou encore par microscopie électronique. On peut noter que plus de 90% des structures déposées dans la PDB sont résolues par la technique de cristallographie aux rayons X.

Les entrées de la banque comprennent des informations sur les structures primaires et secondaires des molécules considérées, les coordonnées atomiques, souvent des détails des expériences (conditions de cristallisation, empilements cristallins, statistiques d'affinement, etc.) ainsi que des références bibliographiques.

Bien que le nombre de structures de macromolécules biologiques connues à l'heure actuelle soit très inférieur à celui des séquences, de par les nombreuses difficultés que soulèvent leurs déterminations, celui-ci croît actuellement à une vitesse comparable à celle observée pour les séquences protéiques il y a quelques années (Figure 29), grâce aux projets massifs de génomique structurale (Berman and Westbrook, 2004).

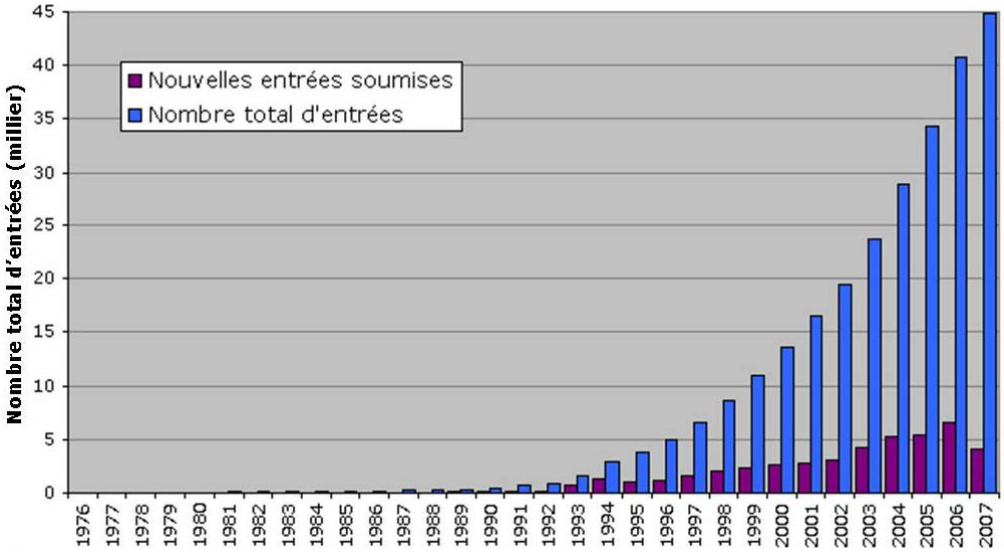


Figure 29 - Evolution du nombre d'entrées de la banque PDB depuis 1976.
(adaptée de http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100)

On peut noter une forte redondance en séquence dans la PDB car plusieurs structures tridimensionnelles peuvent correspondre strictement ou quasi-strictement à la même séquence, selon les mutants étudiés, les partenaires ou ligands introduits dans l'expérience, les conditions d'obtention ou la finesse de résolution de la structure.

Des banques comme CATH (Orengo *et al.*, 1997), SCOP (Murzin *et al.*, 1995) et FSSP (Holm and Sander, 1994) classent les structures issues de la PDB en fonction des domaines structuraux qui les composent. Ces banques représentent en quelques sortes le pendant d'UniRef, par rapport aux banques de séquences, dans cette logique de « simplification » de l'information au sein des banques biologiques.

6.3.2. Banques de données spécialisées

Certaines recherches plus spécifiques réalisées au cours de nos études ont nécessité l'utilisation de banques spécialisées. Nous décrirons les principales banques utilisées dans cette section.

6.3.2.1. GDB

La GDB (Pearson, 1991), pour *human Genome DataBase* , a été créée en 1990 à l'Université John Hopkins de Baltimore. Cette banque de données a été maintenue de 1999 à 2002 par le *Bioinformatics Supercomputing Centre* de Toronto et est à présent gérée par le RTI (*Research Triangle Institute*) aux Etats-Unis. Elle est accessible par le site Web http://www.gdb.org/.

Cette banque se définit comme la banque de données de cartographie du génome humain et regroupe des données relatives à la structure, la diversité et l'évolution de ce dernier. Elle est issue d'une banque de données destinée à la préparation de cartes consensus lors d'ateliers réunissant les spécialistes de la cartographie des gènes (*Human Gene Mapping workshops*) qui procédaient périodiquement à la remise à jour des cartes de chromosomes.

La GDB collecte les données publiées dans la littérature ou directement soumises avant publication, notamment en provenance de grands centres de cartographie. Les données comprennent en particulier la description d'objets cartographiques caractérisés par toute une série d'attributs. Chaque locus, gène ou fragment d'ADN anonyme possède un nom et une localisation plus ou moins précise, qui peut être relative à d'autres loci de la base. Il n'est cependant pas rare d'observer des contradictions entre les ordres proposés.

Des références croisées avec les bases de données de séquences sont également disponibles pour la plupart des entrées de cette banque.

6.3.2.2. InterPro

InterPro (Mulder *et al.*, 2007) est une banque intégrée, constituée d'informations relatives aux familles de protéines, aux domaines protéiques ainsi qu'aux sites fonctionnels des protéines. Elle est distribuée depuis 1999 par l'EBI.

InterPro représente une collaboration entre un certain nombre de banques de données de signatures de protéines, incluant les banques de domaines Pfam (Bateman *et al.*, 2004), Prodom (Bru *et al.*, 2005), Smart (Letunic *et al.*, 2006) et les banques de motifs Prints (Attwood, 2002) et Prosite (Hulo *et al.*, 2006).

Les signatures des protéines sont intégrées manuellement au sein des entrées d'InterPro et sont vérifiées afin de disposer d'informations biologiques et fonctionnelles fiables. InterPro fournit également des liens vers d'autres banques spécialisées, comme Gene Ontology (Ashburner *et al.*, 2000).

6.3.3. Banques de données intégratives

Enfin, dans le cadre de certains développements spécifiques, ainsi que lors de l'utilisation de la grille de calcul Décrypthon, nous avons été amenés à effectuer des analyses basées sur des banques de données intégratives, qui ont la particularité de regrouper en leur sein des données issues de plusieurs banques de données.

6.3.3.1. GeneCards

GeneCards (Rebhan *et al.*, 1997) est une banque de données développée et maintenue depuis 1997 au *Weizmann Institute of Science*. Elle est publiquement accessible à l'adresse suivante : www.genecards.org.

La mise en place de cette banque est née de l'ambition de regrouper des informations hétérogènes, relatives aux gènes humains, dispersées au travers d'une variété de banques de données spécialisées, afin d'offrir pour chacun une vue d'ensemble cohérente. GeneCards fournit un grand nombre de références croisées (ENSEMBL, UniProt, InterPro, GO, OMIM, et HGMD pour ne citer qu'eux) et permet ainsi à l'utilisateur d'accéder aisément à un ensemble de données intégrées.

6.3.3.2. Centre de Données Décrypthon

Le Centre de Données Décrypthon (CDD) est un système d'intégration de données qui regroupe des informations biologiques et biomédicales hétérogènes, issues de diverses banques de données publiques (Tableau 7). Le CDD fait partie intégrante de l'infrastructure mise en place dans le cadre du programme Décrypthon. Il est hébergé par le serveur de stockage situé au CRIHAN (*voir* section 6.1.2.2, p. 82) et a en particulier été développé pour

être exploité dans le cadre de la grille de calcul universitaire Décrypthon. En effet, le CDD permet la mise à disposition directe de l'ensemble des données nécessaires au bon fonctionnement des applications portées sur cette grille, sans avoir besoin de les transférer sur chaque nœud de calcul. Le volume total de données actuellement stockées au CDD est d'environ 1 To.

Le CDD repose sur le système de gestion de base de données DB2 d'IBM. Il a comme caractéristiques principales de pouvoir contrôler des bases de données de divers types et de gérer un très grand volume de données hétérogènes, tout en respectant des contraintes de sécurité et de confidentialité si besoin est.

Banques de données intégrées	Taille (ordre de grandeur, en Go)
GenBank	800
RefSeq	50
UniProt	100
PDB	50
OMIM	1
Genome UCSC	10
Taxonomy	1
GO	10

Tableau 7 - Banques de données intégrées au Centre de Données Décrypthon (août 2007).

Le CDD tire profit du système BIRD (Nguyen *et al.*, manuscrit soumis), pour *Biological Integration and Retrieval Data*, développé par Hoan Nguyen au sein du laboratoire. BIRD est un système générique, écrit en Java, qui permet la création de systèmes de bases de données à façon, à partir de schémas relationnels préexistants, en intégrant des liens croisés entre les données. Un modèle de données configurable est d'ores et déjà disponible pour plusieurs banques de données, notamment celles incorporées au CDD. BIRD permet ainsi une gestion automatique et souple du CDD.

La mise à jour du CDD repose sur l'efficacité et la rapidité d'intégration de BIRD et se fait en collaboration avec la plateforme de bioinformatique de Strasbourg, qui fournit les versions à jour et validées des diverses banques intégrées au CDD.

6.3.4. Interrogation des banques

La plupart des banques de données installées localement sont disponibles sous plusieurs formats, afin d'être interrogeables de différentes manières : soit par des recherches textuelles simples permettant l'interrogation des champs indexés pour chaque entrée des différentes banques, soit par des requêtes croisées exploitant les liens entre ces banques, soit par des recherches de similarité, par l'intermédiaire de l'outil Blast.

Un langage de requête propre a été développé au sein du laboratoire pour faciliter l'extraction de données du Centre de Données Décrypthon.

Nous décrirons dans cette section les principales méthodes d'interrogation des banques que nous avons utilisées au cours de nos travaux.

6.3.4.1. SRS

La plupart des banques de données installées localement à l'IGBMC (banques de séquences nucléiques ou protéiques, banques de structure, banques de motifs, etc.) peuvent être interrogées par l'intermédiaire du système SRS (Etzold and Argos, 1993), pour *Sequence Retrieval System*.

Grâce à ce système, toute collection structurée peut être indexée par l'intermédiaire d'un langage de programmation propre et des liens croisés directs ou indirects peuvent être réalisés entre les banques si des champs communs existent.

SRS permet ainsi de réaliser des requêtes portant sur les champs indexés pour chaque entrée des différentes banques ainsi que des requêtes croisées exploitant les liens existants entre elles. Ces requêtes peuvent être faites de manière interactive (http://bips.u-strasbg.fr/srs) (Figure 30) ou sous forme de lignes de commandes sous UNIX.

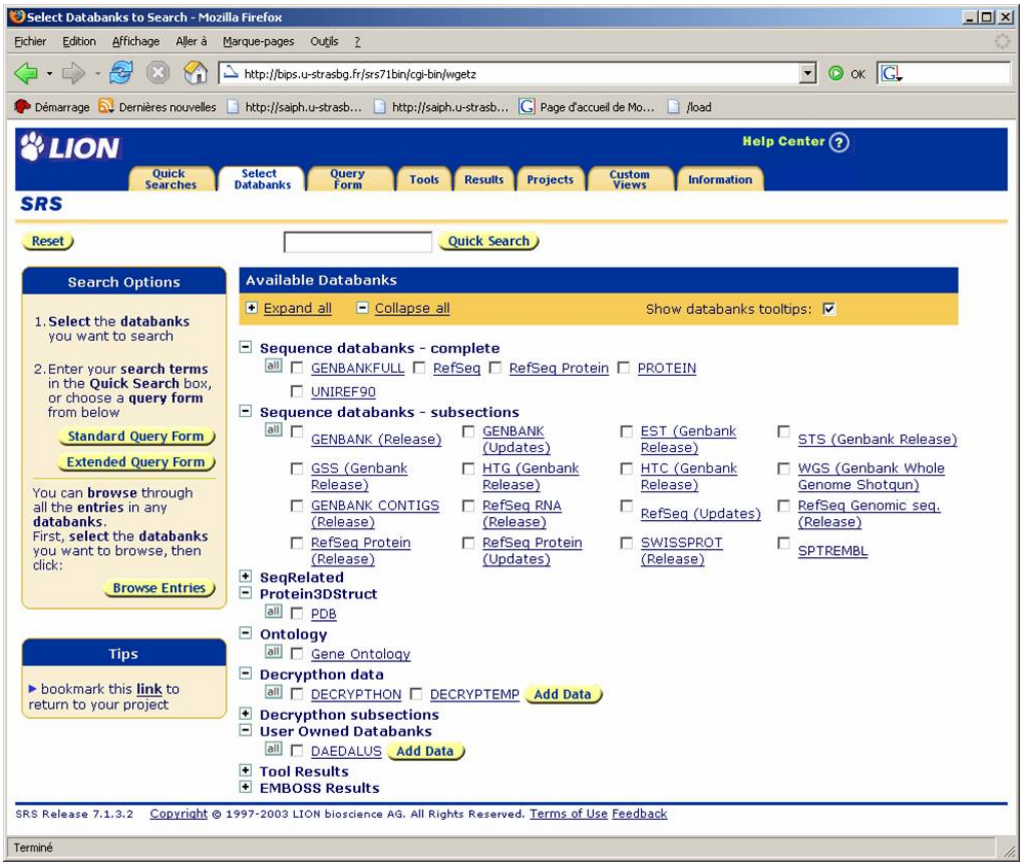


Figure 30 - Page d'accueil du serveur Web SRS de l'IGBMC.

6.3.4.2. Recherche de similarité dans les banques de données

L'interrogation des banques de données peut également se faire directement par l'intermédiaire de recherches de similarité, notamment pour la mise en évidence des séquences appartenant à une même famille protéique. La recherche de similarité fait appel aux alignements par paires et repose sur la comparaison entre une séquence d'intérêt et les séquences d'une banque de données. Au cours de ce processus, les séquences sont comparées afin d'obtenir le plus de correspondances possibles entre les lettres qui les composent.

Nous avons, au cours de nos travaux, effectué les recherches de similarité dans les banques de séquences avec la série des programmes Blast (Altschul *et al.*, 1997) en utilisant les paramètres proposés par défaut. Nous avons en particulier utilisé le programme BlastP, dédié à la recherche de similarité dans les banques de séquences protéiques, à partir d'une séquence protéique. BlastP fournit un alignement local des régions les mieux conservées entre la protéine d'intérêt et chacune des protéines détectées au sein de la banque. Pour chaque alignement, un score et une estimation statistique de sa signification biologique, la valeur d'« expect » (ou *E-value*) sont présentés. Cette dernière prend en compte plusieurs facteurs, comme la longueur des séquences, le score de similarité entre les deux séquences ainsi que la taille et la composition de la banque.

Nous considérons, de manière générale, que les séquences auxquelles une *E-value* inférieure ou égale à 0,001 est associée présentent une similarité significative avec la séquence d'intérêt.

6.3.4.3. BIRD-QL

La richesse du modèle de données de BIRD (*voir* section 6.3.3.2, p. 89) a naturellement conduit au développement d'un langage de requête propre, appelé BIRD-QL, qui permet l'interrogation du CDD.

BIRD-QL peut être considéré comme un dialecte, fait d'assemblage de SQL (*Structured Query Language*) et de fonctions propres à BIRD, qui permet aux bioinformaticiens de formuler des requêtes sophistiquées de manière aisée (Figure 31). Les liens croisés mis en place entre les données peuvent être exploités sans avoir nécessairement besoin de maîtriser l'architecture exacte des tables constituant les banques générées sous BIRD.

BIRD-QL permet ainsi d'exprimer aisément des requêtes élaborées et représente un outil extrêmement performant d'extraction des connaissances, qui participe à l'optimisation et au contrôle de la performance du système.

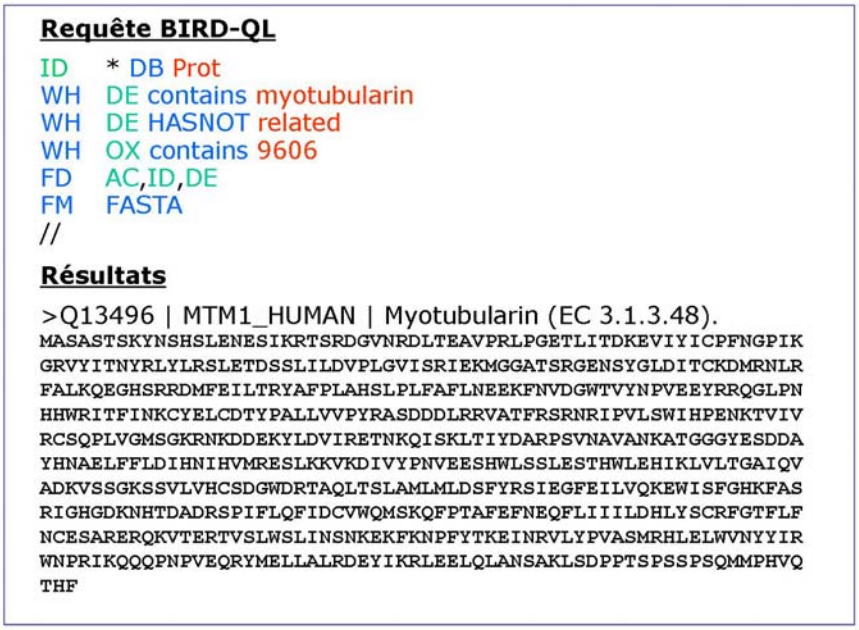


Figure 31 - Exemple de requête BIRD-QL. En bleu, le « dialecte » propre à BIRD-QL, en vert les champs de la banque interrogée et en rouge les conditions spécifiées par l'utilisateur. La requête présentée interroge les banques de séquences protéiques (Prot) et demande de retourner, au format FASTA (dont la ligne de description contient le numéro d'accession de la séquence AC, l'identifiant ID et la définition DE), les protéines humaines (dont l'identifiant de taxonomie est 9606) dont la définition DE contient le mot « myotubularin », mais pas le mot « related ».

Une grande partie des études que nous avons menées est basée sur l'analyse et l'exploitation d'alignements multiples de séquences complètes, que nous appelons couramment MACS, pour *Multiple Alignment of Complete Sequences*.

Nous décrirons dans cette section la méthodologie et les outils adoptés pour la construction d'alignements multiples de bonne qualité et leur analyse.

6.3.5. PipeAlign : un outil d'analyse de familles protéiques

Nous avons construit nos MACS en utilisant des versions plus ou moins modifiées de PipeAlign (Plewniak *et al.*, 2003), un outil d'analyse de famille de protéines qui a été développé au sein du laboratoire. Nous présenterons ici la version d'origine du programme, qui permet la construction automatique d'un MACS hiérarchisé à partir d'une séquence protéique unique ou d'un jeu de séquences en entrée.

PipeAlign est constitué d'une suite de programmes d'analyse de séquences, qui peuvent également être utilisés indépendamment les uns des autres. La première étape est la recherche de similarité dans les banques de données protéiques à l'aide du logiciel BlastP.

Nous présenterons succinctement dans les prochaines sections chacun des autres programmes impliqués dans PipeAlign (Figure 32).

6.3.5.1. Ballast : traitement des résultats des recherches BlastP

Ballast (Plewniak *et al.*, 2000) construit un profil de conservation à partir des séquences détectées par BlastP. La contribution de chaque séquence dans le profil est proportionnelle à sa significativité, c'est-à-dire à son *E-value*. Le profil est ensuite lissé et des pics y sont détectés en utilisant la dérivée seconde du profil lissé. Ces pics définissent les segments de conservation maximale, encore appelés LMS pour *Local Maximum Segments*, qui correspondent aux segments de séquences les mieux conservés entre la séquence initiale et les séquences détectées par BlastP. Les positions des LMS dans chaque séquence sont identifiées et conservées dans un fichier.

6.3.5.2. DbClustal : construction de MACS

DbClustal (Thompson *et al.*, 2000) est un programme d'alignement multiple de séquences complètes qui conjugue les avantages des algorithmes d'alignement global et d'alignement local.

En effet, ClustalW (Thompson *et al.*, 1994), programme d'alignement multiple basé sur l'algorithme d'alignement global développé par Needleman et Wunsch (Needleman and Wunsch, 1970), a longtemps été utilisé au laboratoire pour la construction des MACS. Les inconvénients d'une méthode basée uniquement sur l'alignement global ont cependant été mis en évidence (Thompson *et al.*, 1999), notamment dans le cas d'alignements de séquences contenant des insertions ou des extensions N-terminales ou C-terminales.

DbClustal a été développé pour pallier ces insuffisances. Ce programme reste basé sur l'algorithme de ClustalW mais intègre aussi les informations de conservation locale mises en évidence par Ballast, en se servant des LMS comme points d'ancrage pour la construction de l'alignement multiple global.

6.3.5.3. RASCAL : parcours et correction des alignements

DbClustal étant basé sur un algorithme qui utilise des approximations, il est possible que des erreurs soient introduites au sein de l'alignement multiple. Le programme RASCAL (Thompson *et al.*, 2003), pour *RApid Scanning and Correction of ALignment errors* a été développé pour détecter ces erreurs et les corriger. L'alignement multiple obtenu en sortie du programme DbClustal est divisé horizontalement et verticalement pour former un « quadrillage » au sein duquel les régions bien alignées, et donc fiables, peuvent être différenciées. Les erreurs potentielles d'alignement sont détectées en comparant les profils des régions fiables. RASCAL réaligne chaque région mal alignée en utilisant un algorithme proche de celui implémenté dans ClustalW. La correction de l'alignement est restreinte aux régions les moins fiables, permettant ainsi une stratégie de ré-affinement plus performante.

6.3.5.4. LEON : extraction des séquences non homologues

Un alignement multiple n'ayant de sens que si les séquences protéiques alignées sont homologues, l'étape suivante de PipeAlign a été mise en place pour détecter, au sein du MACS, les séquences n'appartenant pas à la famille d'intérêt. Le programme LEON (Thompson *et al.*, 2004), pour *multiple aLignment Evaluation Of Neighbours*, se base sur les régions fiables, encore appelées *core blocks*, déterminées par RASCAL. LEON profite de la nature transitive des relations d'homologie : l'information des séquences intermédiaires est utilisée pour mettre en évidence les régions conservées des séquences les plus divergentes. Les blocs de conservation de chaque sous-famille du MACS sont ensuite reliés, afin de former des régions contiguës de conservation considérées comme homologues à la séquence initiale. La composition en acides aminés des séquences de l'alignement est également prise en compte, par l'incorporation d'un certain nombre d'algorithmes de détection de segments dont la composition est biaisée. Finalement, les séquences qui ne contiennent aucune région homologue sont retirées du MACS.

En sortie de LEON, on dispose donc d'un MACS de bonne qualité qui ne contient que des séquences partageant au moins une région homologue à la séquence initiale.

6.3.5.5. NorMD : évaluation de la qualité d'un MACS

NorMD (Thompson *et al.*, 2001), pour *Normalized Mean Distance*, est une fonction objective qui peut être utilisée pour évaluer la qualité d'un MACS. Cette fonction combine les avantages des techniques basées sur les scores de colonnes avec la sensibilité des méthodes introduisant des scores de similarité de résidus.

Le score assigné à l'alignement est normalisé par rapport au nombre de séquences que contient cet alignement, leur pourcentage d'identité, leur longueur, etc. Ceci permet de comparer les scores NorMD d'alignements indépendants. Le score assigné par NorMD sera, de manière générale, compris entre 0 et 1. Plus le score est proche de 1, plus la qualité de l'alignement peut être considérée comme satisfaisante. Un seuil de 0,3 a été déterminé, en dessous duquel nous considérons que la qualité d'un alignement n'est pas satisfaisante.

6.3.5.6. Secator et DPC : classification des séquences au sein d'un alignement

La classification des séquences au sein d'un alignement est la dernière étape intégrée à PipeAlign. Les programmes Secator (Wicker *et al.*, 2001) et DPC (Wicker *et al.*, 2002), pour *Density of Points Clustering*, permettent la classification, ou hiérarchisation, des séquences d'un alignement multiple dans des sous-groupes potentiellement fonctionnels, le nombre de sous-groupes créés étant déterminé de façon automatique par ces programmes.

La Figure 32 récapitule les différentes étapes réalisées par PipeAlign. La sortie finale de la suite de programme PipeAlign est un MACS validé de haute qualité, dans lequel les séquences sont classées en sous-familles potentiellement fonctionnelles.

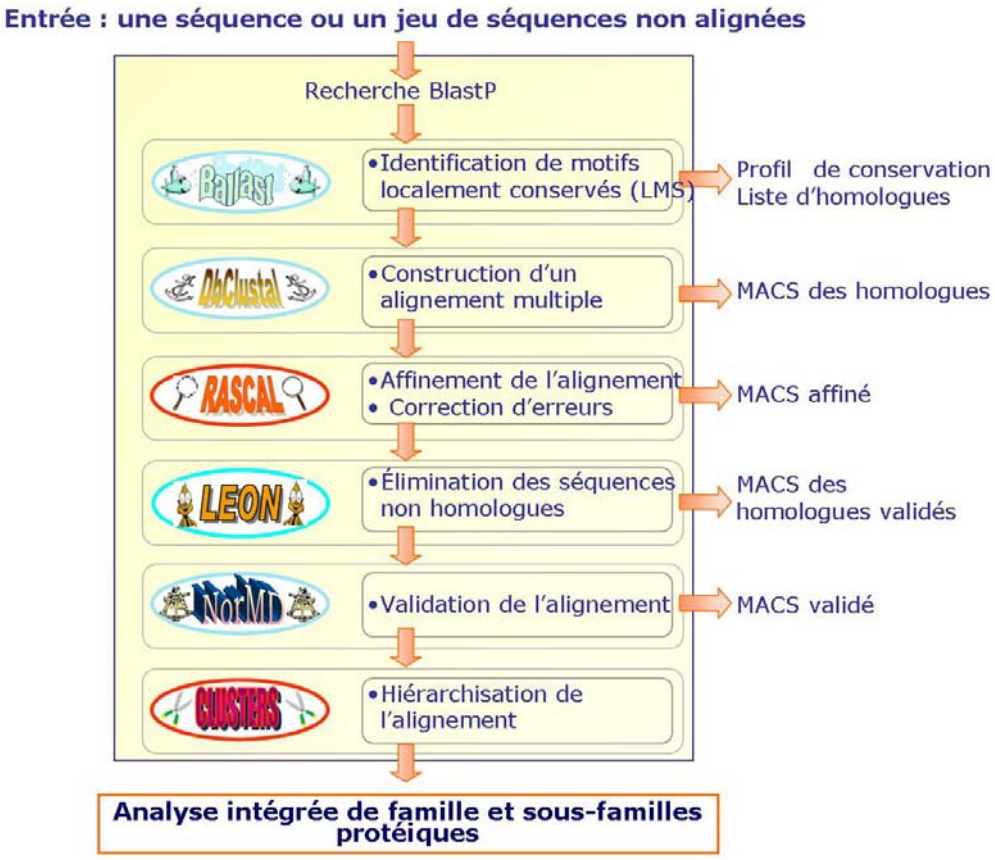


Figure 32 - Aperçu de la cascade de programmes constituant PipeAlign.
(adaptée de (Plewniak *et al.*, 2003))

PipeAlign est mis à disposition de la communauté scientifique par l'intermédiaire du site Web de la plateforme de bioinformatique de Strasbourg (http://bips.u-strasbg.fr/PipeAlign/). Nous disposons de la version locale de chacun des programmes qui constitue la cascade de PipeAlign, nous permettant de concevoir des « PipeAlign » à façon, ajustés au mieux à nos besoins, en fonction des études menées.

6.3.6. Edition et mise en forme des alignements multiples

Quelque soit la méthode automatique utilisée pour la construction d'un alignement multiple de séquences, il est nécessaire de l'affiner et de le valider manuellement en se basant sur l'expertise biologique.

Un certain nombre d'éditeurs d'alignements multiples sont disponibles actuellement pour réaliser ce travail. La correction manuelle des alignements multiples, en se basant

notamment sur la conservation des structures secondaires, a été réalisée dans l'interface graphique Seqlab du package GCG (Wisconsin Package Version 10.2, *Genetics Computer Group* Madison, Wisc.), installé sur *Titus*, qui permet de visualiser, de manipuler et d'éditer les alignements multiples de manière conviviale (Figure 33).

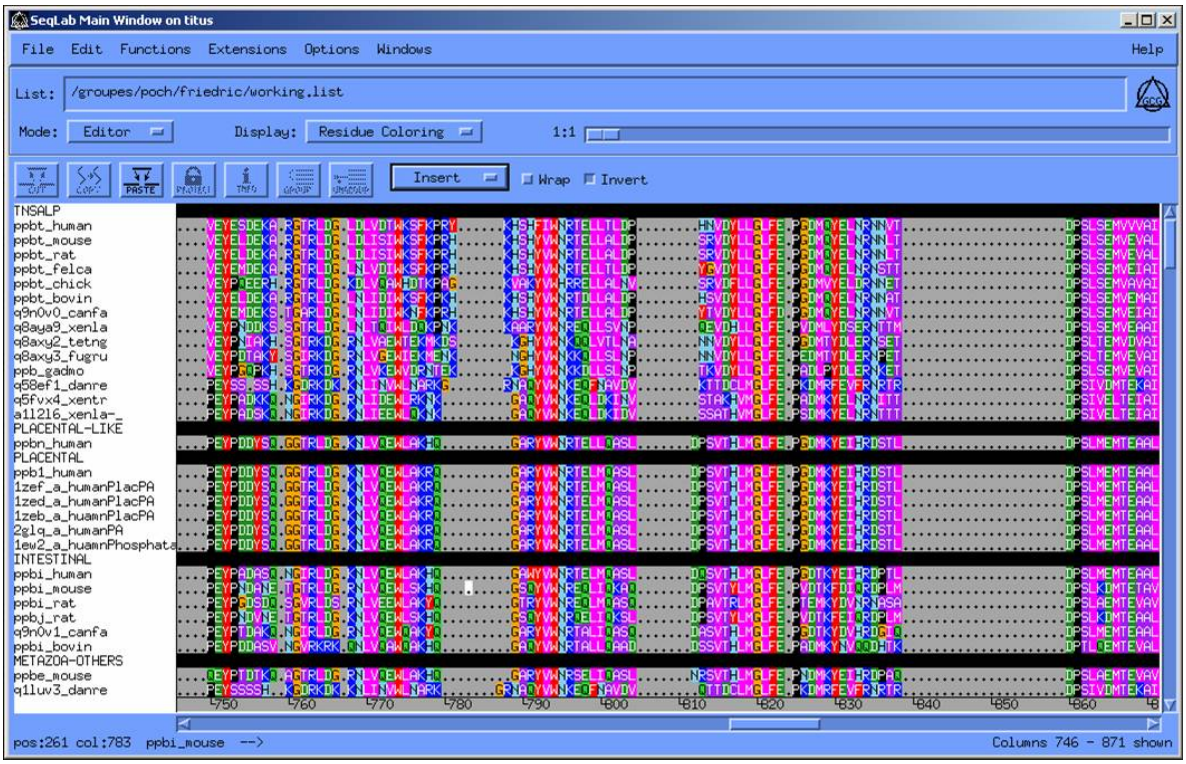


Figure 33 - Editeur d'alignements multiples Seqlab.

6.3.7. MACSIMS : gestion de l'information au sein des alignements multiples

MACSIMS (Thompson *et al.*, 2006), acronyme de *Multiple Alignment of Complete Sequences Information Management System*, est un système de gestion de l'information basé sur l'ontologie des alignements multiples, MAO (Thompson *et al.*, 2005), pour *Multiple Alignment Ontology*, répertorié sur OBO (*Open Biological Ontologies*, http://obofoundry.org/), le site de référence des ontologies biologiques.

MACSIMS permet une annotation automatique des MACS par l'intégration et l'organisation de différents types de données dans le cadre de l'alignement multiple. MACSIMS combine des méthodes exploitant l'analyse des bases de connaissances et la prédiction de séquences *ab initio*. Une étape de validation croisée des informations recueillies, s'appuyant sur les données de conservation au sein de l'alignement multiple, lui-même divisé en sous familles, et la pertinence des informations générées, permettent la mise en évidence des données avérées. L'information validée des séquences connues est alors propagée aux séquences

inconnues, les caractérisant ainsi par des annotations fiables et détaillées. Un résumé des étapes intégrées à MACSIMS est visualisable sur la Figure 34.

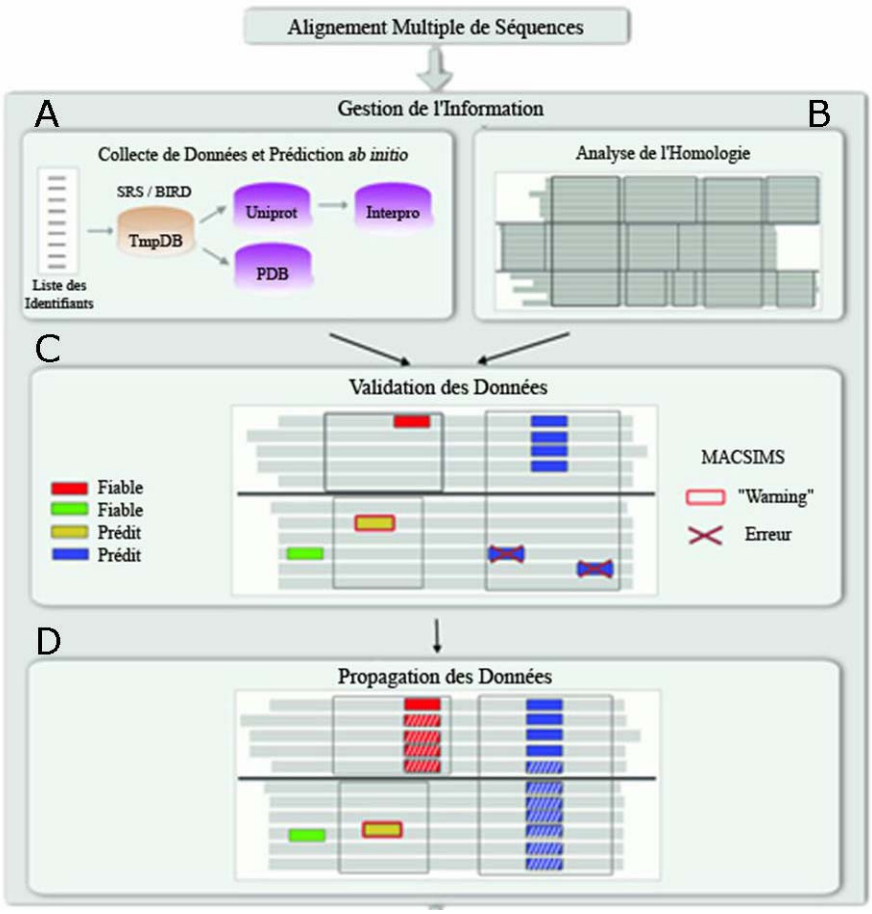


Figure 34 - Etapes successives de MACSIMS. Pour chaque séquence de l'alignement, des informations sont collectées au sein de banques de données publiques et certaines caractéristiques prédites *ab initio* (A). Les blocs de conservation de l'alignement sont déterminés en parallèle (B). Les données recueillies fiables sont validées, les autres sont éliminées (C). Les données fiables sont ensuite transférées des séquences annotées aux séquences inconnues (D). (adaptée de (Thompson *et al.*, 2006))

Les informations relatives aux séquences de l'alignement sont collectées au sein de diverses banques de données, comme UniProt, PDB et InterPro, et sont de plusieurs natures comme des données taxonomiques, des résidus des sites actifs ou de liaison, des résidus modifiés, des variants associés à des maladies génétiques, des domaines Pfam-A et Prosite et des annotations de l'ontologie des gènes GO. Les segments transmembranaires et de faible complexité ainsi que les régions dites « *coiled-coil* » sont prédits à partir de la séquence primaire des protéines de l'alignement.

Les informations collectées ou générées par MACSIMS sont disponibles dans le format XML, un format structuré permettant une exploitation informatique automatique à haut débit. L'éditeur d'alignements multiples Jalview (Clamp *et al.*, 2004) a été adapté pour permettre la

visualisation, de manière simple et conviviale, des alignements annotés et est accessible *via* une interface Web (http://bips.u-strasbg.fr/MACSIMS/) (Figure 35).

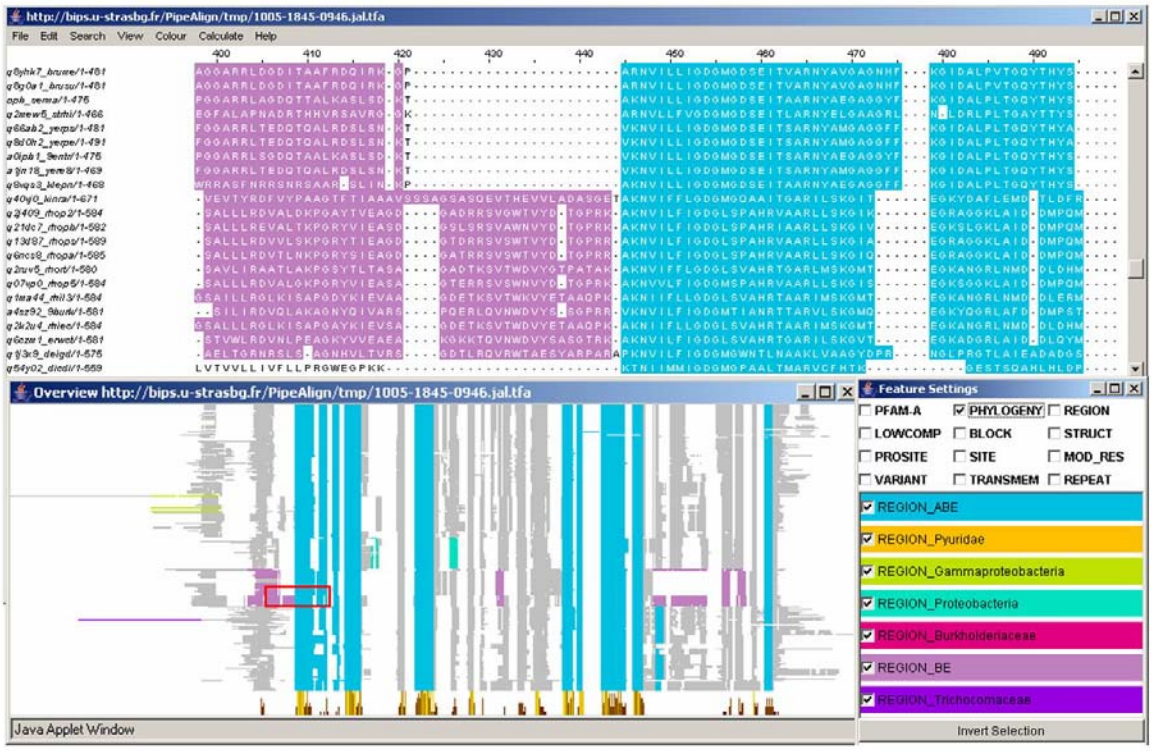


Figure 35 - Alignement multiple annoté par MACSIMS. Trois fenêtres de visualisation Jalview sont présentées : l'alignement multiple annoté est présenté dans la fenêtre du haut. Un aperçu de l'alignement complet est disponible (en bas à gauche) : la partie de l'alignement visualisée dans la fenêtre du haut est délimitée par le rectangle rose. La fenêtre de sélection des informations (en bas à droite) permet de choisir les données à visualiser sur l'alignement.

MACSIMS facilite ainsi la collecte automatique d'informations et l'extraction de connaissances et fournit un outil interactif d'interrogation et de visualisation des résultats.

6.3.8. OrdAlie : analyse hiérarchisée d'alignements multiples

OrdAlie, pour *Ordered Alignment information explorer* (Moulinier *et al.*, manuscrit en préparation) est un outil d'analyse et d'édition des alignements multiples développé au laboratoire par Luc Moulinier. OrdAlie est à même de gérer aussi bien les formats traditionnels des alignements multiples, comme MSF et FASTA (*voir* Annexe 1, p. 231 et Annexe 2, p. 232), que les alignements annotés par MACSIMS, au format XML.

OrdAlie permet une analyse interactive des alignements multiples et l'exploration des séquences, structures et fonctions des protéines, ainsi que des relations évolutives entre ces dernières. Les séquences de l'alignement peuvent être automatiquement ou manuellement classées en sous-familles, permettant une analyse hiérarchique de la conservation des résidus

au niveau de la famille entière ou des sous-familles. Plusieurs méthodes de calcul des conservations sont disponibles (*voir* section 7.2, p. 108), celle basée sur des seuils de conservation le long des colonnes de l'alignement étant utilisée par défaut. La représentation des conservations fournie par OrdAlie (Figure 36) permet d'évaluer aisément la pertinence des groupes constitués et met en évidence à la fois les régions de forte conservation caractéristiques de la famille et les résidus discriminants, c'est-à-dire conservés à l'intérieur d'un sous-groupe et absents dans les autres groupes. Ces oppositions sont révélatrices des spécificités de chaque groupe à l'intérieur de la famille.

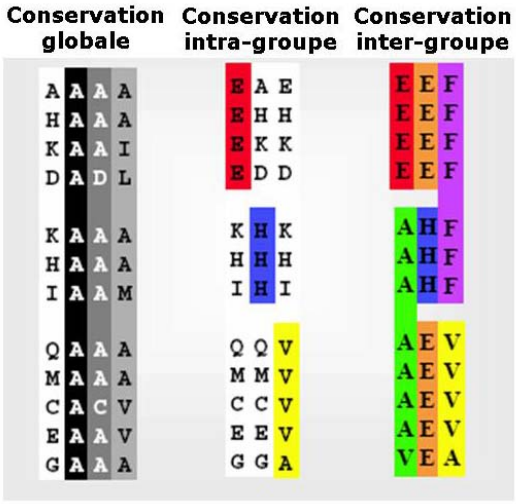


Figure 36 - Représentation des conservations adoptée dans OrdAlie. La conservation globale concerne tous les membres de la famille (noir : conservation stricte, gris foncé : au moins 80% de résidus identiques, gris clair : au moins 60% de résidus partageant les mêmes propriétés physico-chimiques) ; la conservation intra-groupe : plus de 80% d'identité dans un groupe (rouge pour le groupe 1, bleu pour le groupe 2, jaune pour le groupe 3) ; conservation inter-groupe : plus de 80% d'identité entre deux groupes (violet pour les groupes 1 et 2, vert pour les groupes 2 et 3, orange pour les groupes 1 et 3).

Les résidus conservés peuvent également être visualisés dans le contexte de leur environnement structural, par l'intermédiaire de RASMOL (Sayle and Milner-White, 1995), lorsque des structures sont disponibles au sein de l'alignement.

Enfin, un module de construction et de visualisation d'arbres phylogénétiques est intégré à OrdAlie.

6.4. Analyse structurale des protéines

Nous avons également basé une partie de nos études sur l'analyse structurale des protéines.

Nous présenterons dans cette section les logiciels employés pour la construction de modèles tridimensionnels par homologie ainsi que les outils utilisés pour l'analyse visuelle des modèles et structures.

6.4.1. Modélisation de structures tridimensionnelles par homologie

Les modèles tridimensionnels ont principalement été construits en utilisant le logiciel Geno3D (Combet *et al.*, 2002), hébergé par le serveur de l'IBCP de Lyon. Le principal avantage de Geno3D réside dans sa capacité à générer un modèle par homologie de bonne qualité en partant de séquences partageant un faible taux d'identité (entre 20 et 30% d'identité). Geno3D extrait les contraintes géométriques (distances et angles dièdres) des atomes correspondants entre la protéine d'intérêt et l'empreinte, qui sont ensuite fournies au logiciel CNS (Brunger *et al.*, 1998). Une étape de dynamique moléculaire et de minimisation d'énergie est ensuite réalisée, avec et sans contraintes géométriques. Par défaut, trois modèles sont générés et validés en réalisant un alignement structural avec l'empreinte. Le modèle présentant l'énergie minimale est finalement retenu.

Nous avons également été amenés à construire des modèles par homologie en utilisant Modeller (Sali *et al.*, 1995), version 9v1, installé en local sur le serveur *Star*. Modeller a l'avantage de pouvoir utiliser plusieurs séquences empreintes lors de la construction du modèle par homologie.

6.4.2. Visualisation et mise en forme des structures tridimensionnelles

Nous avons utilisé le logiciel PyMOL (http://pymol.sourceforge.net) et l'application Jmol (http://jmol.sourceforge.net/) pour visualiser nos structures et modèles en trois dimensions et les analyser. Jmol présente l'avantage d'être disponible sous forme d'applet et peut ainsi être intégré au sein de pages Web, sans que l'utilisateur n'ait à réaliser d'installation préalable.

Des fonctions permettent de visualiser et d'explorer les différentes parties ou propriétés d'une molécule (comme les charges, les résidus hydrophobes et polaires). Ces deux outils sont capables d'importer tous les formats de modèles moléculaires disponibles et peuvent être contrôlés au moyen de scripts. La qualité et la diversité des représentations disponibles ainsi que les possibilités offertes à un utilisateur averti rendent ces outils indispensables à l'analyse structurale.

Chapitre 7 - Développements relatifs aux alignements multiples

Une partie des études que nous avons menées au cours de cette thèse est basée sur les alignements multiples, l'une des thématiques centrales du laboratoire. Le Chapitre 10 - *Problématique du contenu informationnel des alignements multiples*, p. 137 sera dédié aux alignements multiples et présentera plus en détails leurs applications, leur construction ainsi que leurs analyses.

Le présent chapitre sera dédié aux méthodes développées, implémentées ou simplement testées au cours de nos travaux et relatives, d'une part, à l'échantillonnage des résultats de recherches de similarité pour la construction d'alignements à façon et, d'autre part, aux méthodes de calcul de la conservation des résidus au sein des alignements.

7.1. Méthodes d'échantillonnage des résultats du Blast

L'ère post-génomique est caractérisée par une quantité considérable et sans cesse croissante de données disponibles au sein des diverses banques biologiques. Ceci a de lourdes conséquences sur la manière de mener à bien les analyses de séquences. Afin de faciliter l'accès à différents niveaux d'informations, des stratégies d'organisation et de « simplification » des données voient le jour petit à petit.

La construction et l'exploitation d'alignements multiples font partis des champs d'application dans lesquels de telles stratégies se doivent d'être mises en place. En effet, comme l'illustre par la succession de programmes constituant PipeAlign (*voir* section 6.3.5, p. 93), la construction d'un alignement multiple commence par une recherche de similarité dans les banques de séquences. A l'heure actuelle, ces dernières mettent couramment en évidence plusieurs centaines, voire plusieurs milliers de séquences à partir d'une simple séquence d'intérêt. Extraire une information pertinente d'une telle quantité de séquences se révèle difficile.

Un challenge actuel réside par conséquent dans la sélection intelligente et efficace des séquences avant de réaliser les alignements multiples. Ces stratégies reposent sur une simplification de l'information, mais doivent se faire en fonction de la question biologique posée, pour ne pas altérer le message essentiel aux besoins de l'analyse.

Ces idées seront développées dans le cadre du Chapitre 10 - *Problématique du contenu informationnel des alignements multiples*, p. 137 et du Chapitre 11 - *Alignements multiples dédiés à l'analyse structurale*, p. 145.

C'est dans ce contexte que nous avons été amenés à développer et/ou implémenter des méthodes de sélection des séquences mises en évidence par des recherches Blast. Celles-ci nous ont permis de développer des versions ajustées de PipeAlign, en intégrant une étape de sélection des séquences avant la construction de l'alignement.

Nous décrirons dans ce chapitre chacune des méthodes de sélection déployées.

7.1.1. Méthodes d'échantillonnage basées sur les *E-value* du Blast

Au cours de nos travaux, nous avons étudié la possibilité de réduire le nombre de séquences à aligner tout en conservant le potentiel informationnel, en particulier structural, des alignements multiples associés aux jeux de séquences sélectionnés. Nous avons pour cela implémenté trois méthodes d'échantillonnage des résultats de recherches de similarité Blast, basées sur les *E-value* associées aux séquences détectées.

Ces méthodes reposent sur l'hypothèse qu'aux niveaux structural et fonctionnel, les séquences les plus proches entre elles ne sont pas celles qui apportent le plus d'informations pertinentes. Dans ce contexte, la diversité est généralement plus informative (Przybylski and Rost, 2002).

Les trois méthodes présentées dans cette section choisissent ainsi des séquences représentatives de la dispersion des *E-value*. En effet, l'*E-value* calculée par le programme Blast étant notamment représentative de la longueur des séquences alignées, de leur score de similarité et de la composition de la banque, les séquences qui présentent des *E-value* très proches lors d'une recherche de similarité sont potentiellement proches entre elles.

7.1.1.1. Méthode des bandelettes

Dans le cadre de la méthode des bandelettes, encore appelée *strips method* ou *sm*, le nombre maximal de séquences que l'on souhaite sélectionner doit être fixé au préalable. Nous désignerons ce nombre par x.

Le graphique logarithmique des *E-value* du Blast est divisé en x bandelettes de même largeur. Si E_1 est l'*E-value* la plus basse mise en évidence lors de la recherche Blast et E_n l'*E-value* la plus élevée ($E_n \leq 0{,}001$), la largeur des bandelettes est calculée de la manière suivante :

$$\text{Largeur} = \frac{1}{x}\left(\log(E_1) - \log(E_n)\right)$$

La séquence associée à l'*E-value* la plus basse est sélectionnée (Figure 37) au sein de chaque bandelette non vide.

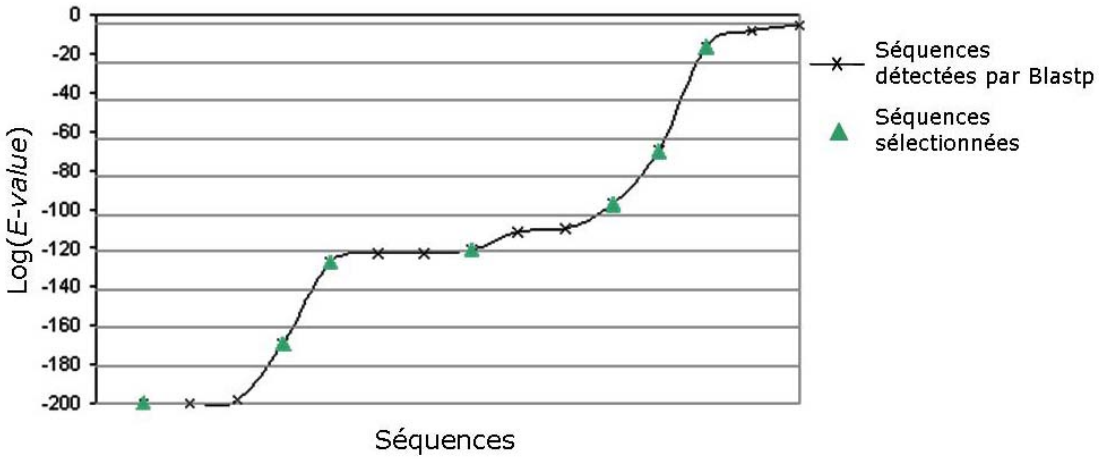


Figure 37 - Méthode des bandelettes. Le graphique logarithmique des *E-value* est dans cet exemple divisé en 10 bandelettes de même largeur. Les triangles verts schématisent les séquences sélectionnées.

7.1.1.2. Méthode de la dérivée seconde

La méthode de la dérivée seconde, encore appelée *second derivative method* ou *sdm*, permet de sélectionner les séquences situées à des points d'inflexion de la courbe logarithmique des *E-value*.

Pour cela, la fonction de variation V de la courbe des *E-value* est calculée. Nous aurons : $V_{(i+1)} = E_{i+1} - E_i$.

Le sens de variation de V est déterminé en étudiant le signe de sa propre dérivée V', c'est-à-dire la dérivée seconde de la courbe logarithmique des *E-value*. Lorsque cette dernière passe d'une valeur positive à une valeur négative, la séquence associée à la valeur négative est sélectionnée (Figure 38).

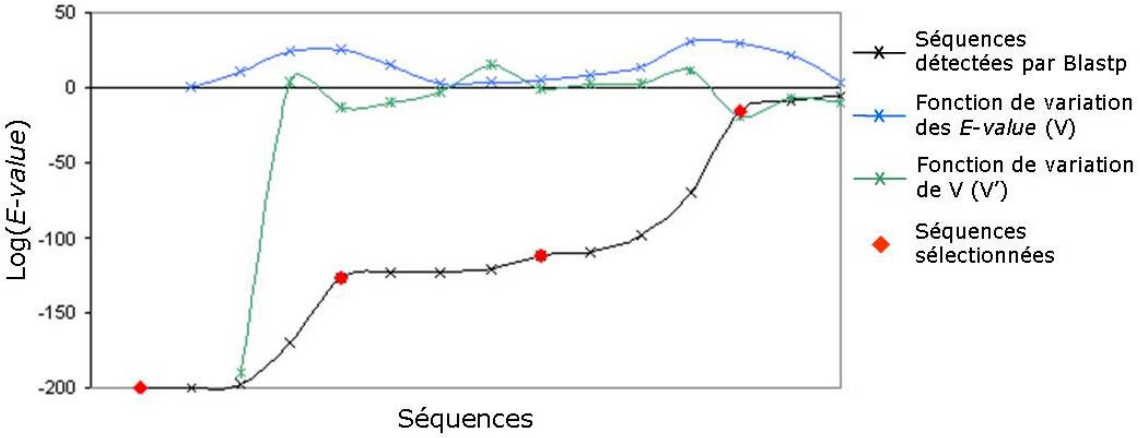


Figure 38 - Méthode de la dérivée seconde. Les séquences situées aux points d'inflexion de la courbe de variation des *E-value* sont sélectionnées (losange rouge).

7.1.1.3. Méthode des moyennes

Dans le cadre de la méthode dite des moyennes, encore appelée *mean method* ou *mm*, un seuil est défini comme étant la différence moyenne entre les logarithmes des *E-value* successives rencontrées au sein du Blast (Errami *et al.*, 2003).

Si on considère n comme étant le nombre de séquences retrouvées avec une *E-value* inférieure ou égale à 0,001, E_n l'*E-value* associée à la $n^{ième}$ séquence et E_1 l'*E-value* la plus basse mise en évidence lors de la recherche Blast, le seuil peut être calculé par la formule suivante :

$$\text{Seuil} = \frac{\log(E_n) - \log(E_1)}{n-1}$$

Une séquence i sera sélectionnée par cette méthode lorsque : $\log(E_i) - \log(E_{i+1}) \geq$ Seuil (Figure 39).

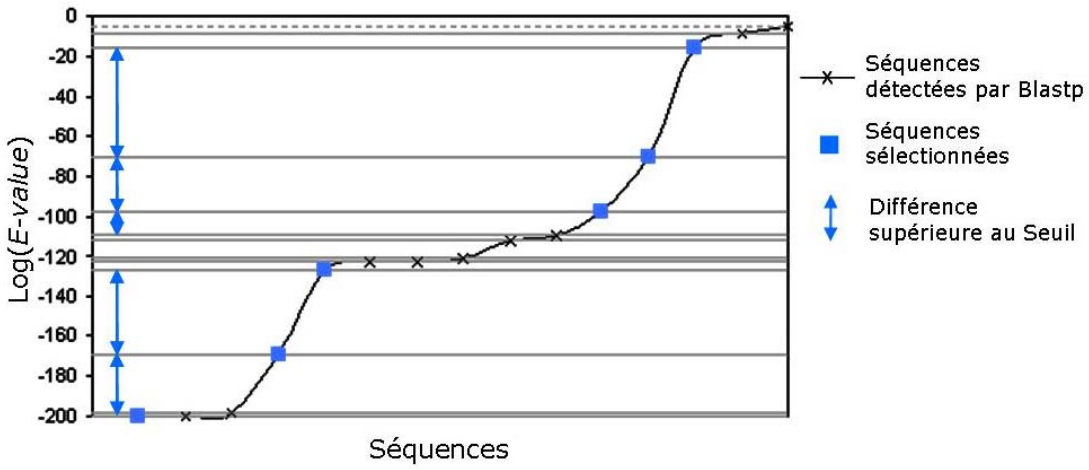


Figure 39 - Méthode des moyennes. Les flèches bleues illustrent les différences des log(*E-value*) successives supérieures au Seuil pré-calculé et les carrés bleus les séquences sélectionnées.

7.1.2. Méthodes basées sur la taxonomie

Nous avons également implémenté deux méthodes de sélection des séquences basées sur la taxonomie des organismes dont elles proviennent : les filtres dits Eucaryote et Métazoaire. Pour que ces filtres soient applicables, la recherche de similarité doit être menée dans la banque de séquences UniProt.

Ces méthodes tirent profit des développements réalisés dans le cadre de BIRD et sont basées sur des requêtes BIRD-QL (*voir* section 6.3.4.3, p. 92). La Figure 40 présente un exemple de requête réalisée dans le cadre du filtre Eucaryote.

Grâce aux numéros d'accession des séquences mises en évidence avec une *E-value* inférieure ou égale à 0,001 au cours des recherches Blast, le champ contenant les informations de taxonomie d'UniProt est interrogé afin de discriminer les séquences issues d'organismes eucaryotes ou métazoaires des autres.

```
ID      P12345 DB Prot
WH      OC contains Eukaryota
FD      ID,AC
//
...
//
ID      P23456 DB Prot
WH      OC contains Eukaryota
FD      ID,AC
//
```

Figure 40 - Exemple de requête BIRD-QL dans le cadre du filtre Eucaryote.

Nous nous sommes pour le moment limités à la mise en place des filtres Eucaryote et Métazoaire, mais il est évident que la simplicité du système de requête offert par BIRD-QL rend tout à fait envisageable le développement à façon d'autres filtres, basés sur n'importe quel champ des banques du Centre de Données Décrypthon indexé par BIRD.

7.1.3. Méthode aléatoire

Une méthode de sélection des séquences, reposant sur un système aléatoire, a également été implémentée afin de servir de témoin dans le cadre de nos analyses. Cette méthode prend simplement en paramètre le taux de réduction qu'elle doit appliquer.

$$\text{Taux de réduction} = \frac{NbSéquencesSélectionnées}{NbSéquencesDétectées}$$

7.2. Calcul de la conservation au sein des alignements multiples

L'étude de la conservation des résidus au sein des alignements multiples est une étape primordiale lors de l'analyse d'une famille protéique.

Différentes méthodes de calcul de la conservation des résidus que nous détaillerons dans les quatre prochaines sections sont implémentées dans OrdAlie (*voir* section 6.3.8, p. 99).

7.2.1. Méthode des seuils

La méthode dite des seuils est basée sur l'utilisation d'un seuil de conservation prédéfini. Par défaut, le seuil est positionné à 80%, mais ce dernier est modifiable par l'utilisateur.

Soit un seuil de conservation x% ; une colonne est considérée comme conservée si x% des résidus sont identiques dans la colonne, en tenant compte des insertions (ou *gap*).

7.2.2. Score de la distance moyenne

Le score de la distance moyenne, ou *mean distance*, a été développé par Thompson et ses collaborateurs (Thompson *et al.*, 1997) dans le cadre de la construction de profils de conservation au sein de ClustalX.

Le score de « qualité » d'une colonne dépend de la variabilité des acides aminés en son sein. Il est basé sur l'utilisation d'une matrice de substitution, qui présente une probabilité de substitution pour chaque paire d'acides aminés. La matrice BLOSUM 62 (Henikoff and Henikoff, 1992) (*voir* Annexe 3, p. 233) est utilisée dans OrdAlie.

Considérons un alignement de longueur M qui contient N séquences. Celui-ci peut être schématisé de la manière suivante :

```
A11 A12 A13 ..........  A1M
A21 A22 A23 ..........  A2M
.
.
AN1 AN2 AN3 ..........  ANM
```

Soit C une matrice de substitution de taille R, dans laquelle C(i,j) est le score de substitution assigné à la paire d'acides aminés (i,j).

Pour calculer un score de conservation pour la $j^{ième}$ colonne de l'alignement, un espace des séquences à R dimensions est défini.

Pour la $j^{ième}$ position de l'alignement, chaque séquence consiste en un résidu auquel un point S de l'espace est assigné. S a R dimensions et pour la séquence i, la $r^{ième}$ dimension est définie de la manière suivante :

$$S_r = C\left(r, A_{ij}\right)$$

Une valeur consensus (X) est ensuite calculée pour la j[ième] position de l'alignement. Cette valeur a également R dimensions et est définie de la manière suivante :

$$X_r = \frac{\sum_{i=1}^{R} \left(F_{ij} \times C(i,r) \right)}{N}$$

Où F_{ij} est le nombre de résidus i à la position j de l'alignement.

La distance D_i entre chaque séquence i et le consensus X peut ensuite être calculée dans l'espace à R dimensions (Figure 41).

$$D_i = \sqrt{\sum_{r=1}^{R} \left(X_r - S_r \right)^2}$$

Le score de qualité MD pour la position j de l'alignement correspond à la moyenne des distances D_i.

$$MD_j = \frac{\sum_{i=1}^{M} D_i}{N}$$

Ce score est normalisé en le multipliant par le pourcentage de séquences qui ont un résidu et non un *gap* à cette position.

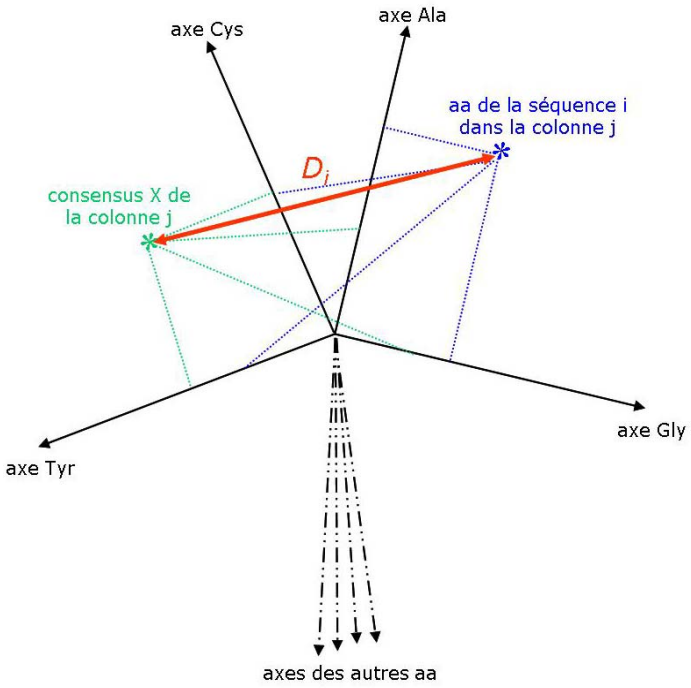


Figure 41 - Calcul de la distance D_i entre l'acide aminé (aa) de la séquence i et le consensus X à la position j de l'alignement, dans l'espace à R dimensions.

7.2.3. Score d'énergie libre

Lockless et Ranganathan (Lockless and Ranganathan, 1999) ont introduit une mesure de la conservation à une position de l'alignement en mesurant la déviation de la fréquence d'un acide aminé dans une colonne par rapport à la fréquence de cet acide aminé dans l'alignement complet, tout en tenant compte de la fréquence « naturelle » de l'acide aminé en question.

Considérons un alignement de N séquences. La probabilité de retrouver un acide aminé x n fois dans une colonne de cet alignement est définie de la manière suivante :

$$P(x) = \frac{N!}{n_x!(N-n_x)!} p_x^{n_x} (1-p_x)^{N-n_x}$$

Où n_x est le nombre de séquences présentant l'acide aminé x dans la colonne et p_x la fréquence observée de cet acide aminé dans la nature.

Le paramètre de conservation évolutive empirique pour le site i (ΔG_i^{stat}) est calculé en comparant P(x) à la probabilité de la fréquence totale de x dans l'alignement :

$$\Delta G_i^{stat} = kT^* \sqrt{\sum_x \left(\ln \frac{P_i^x}{P_{MSA}^x} \right)^2}$$

Où kT^* est une unité d'énergie arbitraire, P_i^x la probabilité de trouver un acide aminé x en position i ; P_{MSA}^x la probabilité d'avoir un acide aminé x conservé dans toutes les séquences.

7.2.4. Score de la moyenne de la norme des vecteurs

La méthode de la moyenne de la norme des vecteurs a été développée au sein du laboratoire par Luc Moulinier et se base sur une représentation vectorielle à deux dimensions des acides aminés (French and Robson, 1983), que nous appelons communément la « rosace des acides aminés » (Figure 42). Cette représentation est obtenue en appliquant une technique statistique de positionnement multidimensionnel, à partir de la matrice de substitution de Dayhoff (Dayhoff *et al.*, 1972), qui permet de modéliser les proximités entre les résidus de façon à pouvoir les représenter le plus fidèlement possible dans un espace de faible dimension.

La rosace des acides aminés reflète la similarité entre les acides aminés. On y remarque, aux vues du positionnement relatif des vecteurs, que, du point de vue de la tendance à se substituer au cours de l'évolution, les propriétés les plus importantes des résidus sont leur hydrophobicité et leur volume.

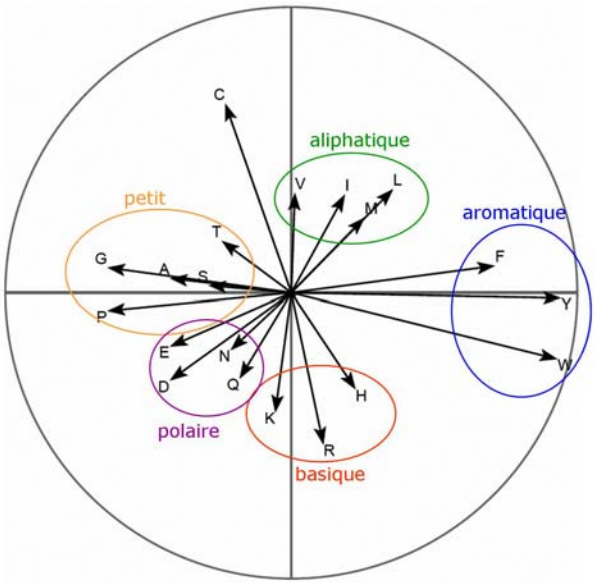


Figure 42 - « Rosace des acides aminés ». Cette présentation en deux dimensions des caractéristiques les plus importantes, d'un point de vue évolutif, des résidus permet de dépeindre de manière visuelle les relations de similarité entre résidus.

Considérons un alignement de N séquences. Dans le cadre de la méthode de la moyenne de la norme des vecteurs, le calcul de la conservation d'une colonne est réalisé de la manière suivante : les vecteurs des N résidus présents dans la colonne sont additionnés pour obtenir le vecteur représentatif de la colonne, ou vecteur moyen V_m. La norme de V_m est divisée par la somme des normes des vecteurs des résidus de la colonne (Figure 43). Le score S de conservation de la colonne est ensuite normalisé, en le multipliant par le nombre de résidus dans la colonne et en le divisant par N.

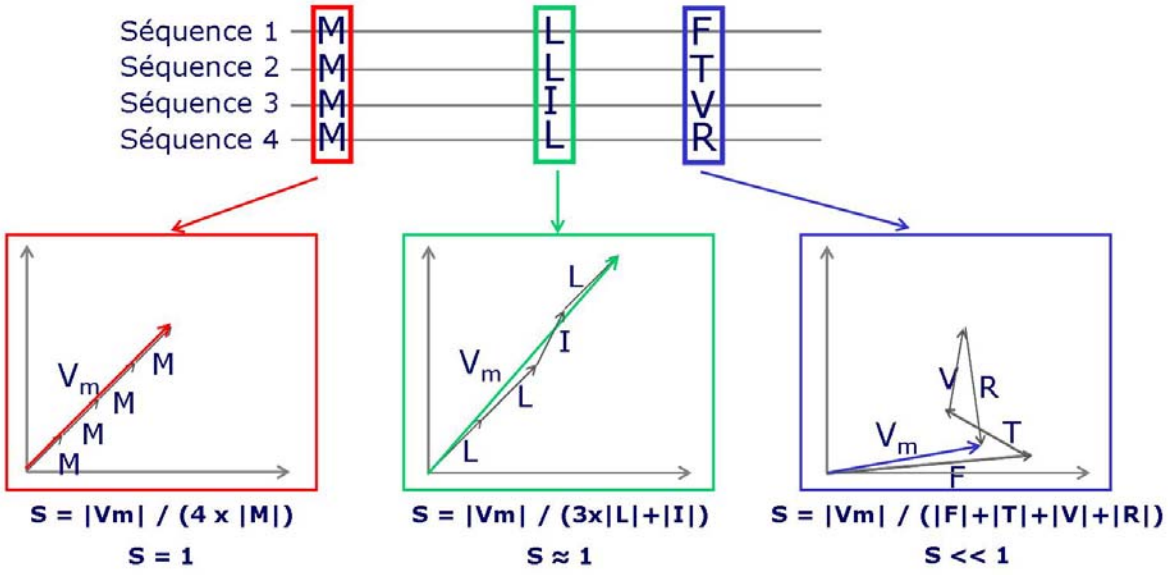


Figure 43 - Scores de conservation dans les colonnes de l'alignement, par la méthode de la norme des vecteurs moyens.

Le score de conservation S vaut 1 si les vecteurs représentant les résidus de la colonne sont tous colinéaires. Or, à la vue de la rosace des acides aminés, seuls les résidus identiques ont des vecteurs colinéaires. Si les résidus ne sont pas tous identiques au sein de la colonne, le score S sera donc inférieur à un.

Chapitre 8 - MAGOS

MAGOS, pour *Multiple AliGnment and mOdelling Server*, est un serveur Web développé en collaboration avec l'équipe de Gilbert Deléage de l'Institut de Biologie et de Chimie des Protéines (IBCP) de Lyon. MAGOS est disponible à l'adresse suivante : http://pig-pbil.ibcp.fr/cgi-bin/magos/home et a fait l'objet d'une publication dans la revue *Bioinformatics* (*voir* Chapitre 12 - *Caractérisation fonctionnelle et structurale des mutations*, p. 173).

MAGOS est né d'une volonté de proposer aux biologistes un espace de travail réunissant de manière conviviale les informations de séquence, de structure, de fonction et d'évolution relatives à une protéine d'intérêt.

8.1. Fonctionnement général de MAGOS

A partir d'une séquence protéique, MAGOS permet de construire, de manière couplée, un alignement multiple de séquences complètes annoté avec des données fonctionnelles et structurales et un modèle tridimensionnel de la protéine d'intérêt. L'interface graphique intuitive de visualisation des résultats permet l'interconnexion de la séquence protéique, au sein de l'alignement et de son modèle tridimensionnel, ainsi que la représentation des annotations sur l'alignement et sur le modèle.

Le serveur Web en tant que tel est hébergé à l'IBCP, alors que les calculs relatifs à MAGOS sont répartis entre les serveurs de calcul de l'IGBMC et de l'IBCP. Les serveurs des deux instituts communiquent via le protocole HTTP par l'intermédiaire de scripts CGI. Le serveur Web est principalement basé sur des scripts écrits en Python.

Schématiquement, MAGOS fonctionne de la manière suivante (Figure 44) : l'utilisateur soumet une séquence protéique au serveur Web installé à l'IBCP. Après vérification du format de la séquence soumise, cette dernière est transférée sur le serveur *Star* de l'IGBMC. Le lancement, sous GScope, d'une suite de programmes permet la création d'un alignement multiple annoté. Les résultats sont renvoyés sur le serveur de l'IBCP, où une nouvelle série de calculs est réalisée pour la construction d'un modèle tridimensionnel de la protéine d'intérêt. L'ensemble des résultats obtenus est rendu accessible à l'utilisateur par l'intermédiaire du serveur Web de l'IBCP.

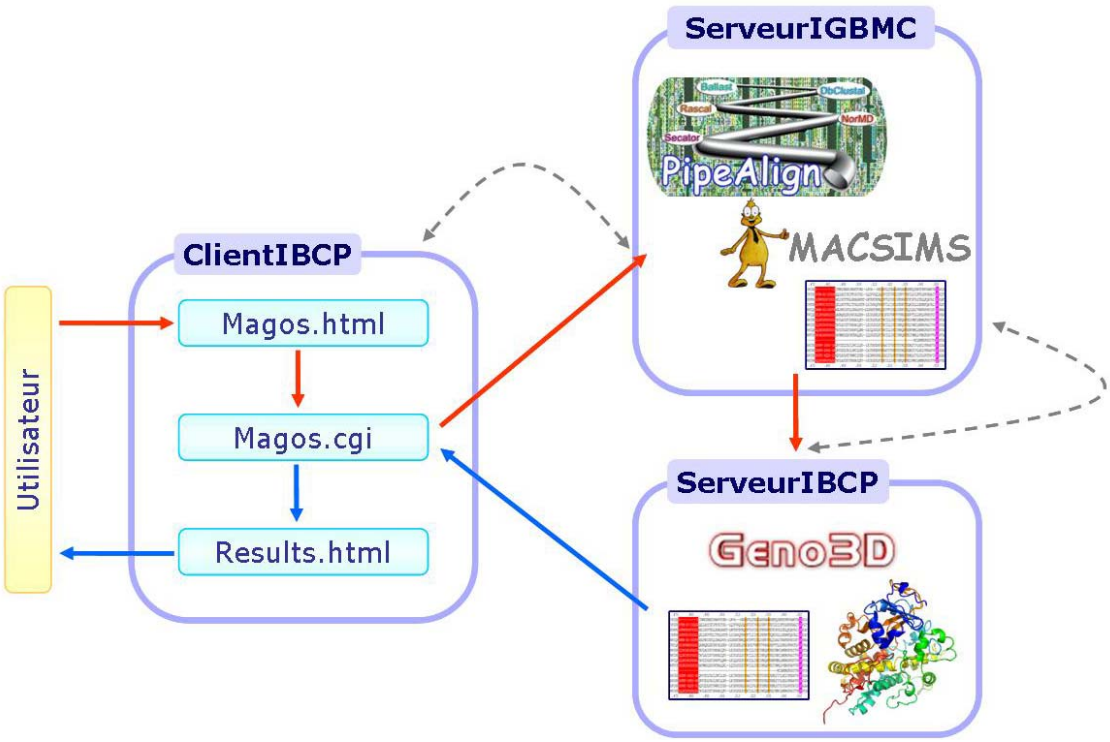


Figure 44 - Organisation du serveur MAGOS. Les flèches rouges symbolisent la soumission de « requêtes », les flèches bleues les réponses générées et les flèches grises indiquent la récupération de données nécessaires pour effectuer les calculs.

Le détail des programmes lancés sur chaque serveur sera fourni dans la section 8.3, p. 115.

8.2. Interface de soumission

L'interface de soumission de MAGOS (Figure 45) se veut délibérément très épurée et simple d'utilisation, car nous souhaitions que MAGOS se définisse comme une plateforme d'intégration pour non-experts.

L'utilisateur copie/colle sa séquence protéique d'intérêt dans l'espace prévu à cet effet.

Quelques paramètres peuvent être modifiés par l'utilisateur, comme :

- le nombre de séquences maximal alignés au sein de l'alignement multiple (par défaut : 100),
- les banques de données dans lesquelles la recherche de similarité sera effectuée (UniProt associée à la PDB ou UniRef 90).

L'utilisateur peut optionnellement préciser l'identifiant PDB et la chaîne peptidique de la structure à utiliser comme empreinte pour la construction du modèle tridimensionnel.

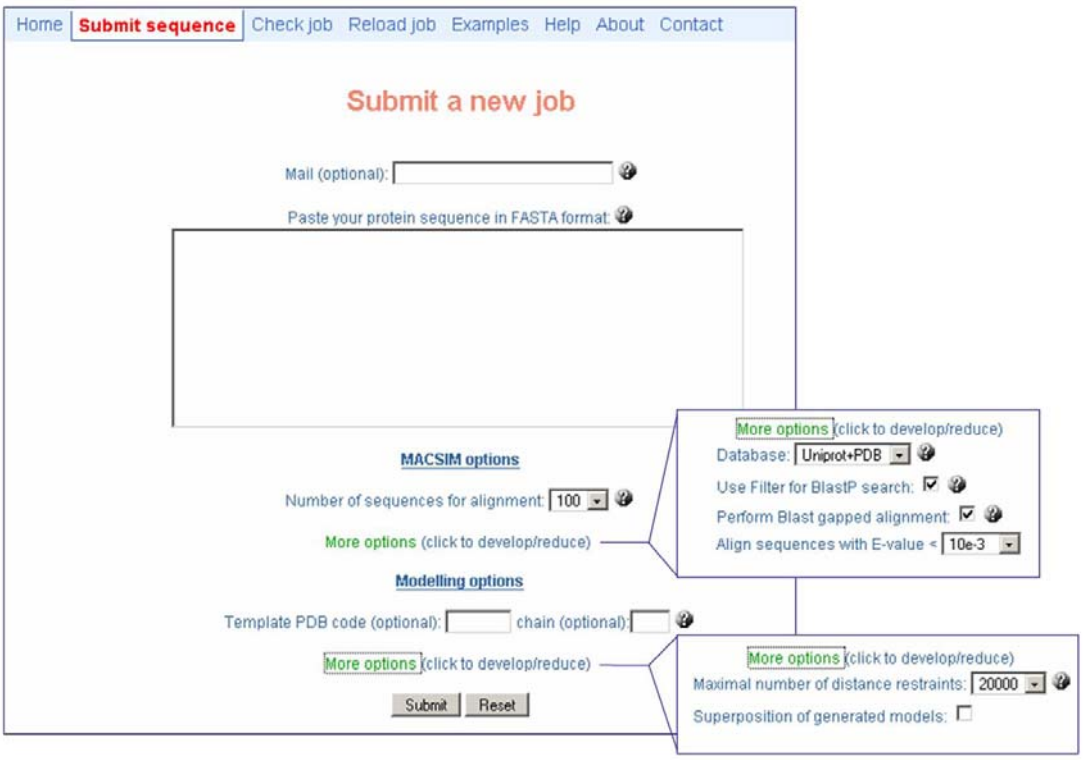


Figure 45 - Interface de soumission de MAGOS.

Il est également conseillé à l'utilisateur de renseigner son adresse *e-mail* pour être prévenu de la fin des calculs concernant sa protéine d'intérêt. S'il ne souhaite pas fournir cette information, il pourra accéder aux résultats par l'intermédiaire du numéro d'identifiant unique automatiquement assigné à sa session.

8.3. Protocole implémenté dans MAGOS

Une fois la séquence protéique soumise par l'utilisateur, un identifiant est assigné à la session de calculs, dont les résultats seront sauvegardés durant deux semaines. Cet identifiant unique permet d'accéder à tous les résultats obtenus durant cette période.

Le serveur *Star* de l'IGBMC est ensuite contacté par l'intermédiaire du protocole HTTP, et la séquence soumise par l'utilisateur rapatriée.

8.3.1. Sélection de l'empreinte pour la construction du modèle tridimensionnel

Si l'utilisateur n'a pas renseigné le champ lui permettant de préciser quelle structure doit servir d'empreinte pour la construction du modèle tridimensionnel par homologie, celle-ci sera déterminée automatiquement. Pour cela, une recherche de similarité est effectuée au

sein de la banque de structures PDB, à l'aide du programme BlastP, et la séquence présentant le meilleur score Blast est sélectionnée.

8.3.2. Construction et annotation de l'alignement multiple

La construction de l'alignement multiple est réalisée à partir de la séquence soumise par l'utilisateur, en s'appuyant sur une version ajustée de PipeAlign (*voir* section 6.3.5, p. 93 pour sa version originale).

Les modifications apportées à la version originale sont de plusieurs ordres. D'une part, l'empreinte est ajoutée à la liste des séquences à aligner si elle ne fait pas partie du jeu de séquences initialement sélectionnées. D'autre part, une fois les alignements issus des programmes DbClustal et RASCAL construits, leurs scores NorMD sont comparés afin de lancer le programme LEON sur l'alignement de meilleure qualité. Si la valeur de NorMD de l'alignement issu de LEON est supérieure ou égale à 0,3, il sera utilisé pour effectuer la suite des traitements. Le cas contraire, l'alignement présentant le score NorMD le plus élevé sera retenu.

Si l'alignement retenu est celui obtenu en sortie de LEON, on procède à une ultime vérification, pour s'assurer que la séquence empreinte, qui est souvent plus courte que la moyenne des séquences alignées, n'ait pas été retirée de l'alignement. Si tel devait être le cas, l'alignement considéré pour la suite serait celui présentant le meilleur score NorMD en sortie de DbClustal et RASCAL.

L'alignement retenu est ensuite annoté par l'intermédiaire de MACSIMS.

Un module additionnel permet de mettre en évidence, parmi les séquences alignées, les protéines impliquées dans des maladies génétiques humaines ainsi que leurs homologues. Ce module, dont les résultats sont stockés dans un fichier texte indépendant, est basé sur les analyses réalisées dans le cadre de MS2PH-db (*voir* Chapitre 9 - *MS2PH-db*, p. 123).

8.3.3. Création de l'alignement protéine d'intérêt / empreinte

La construction d'un modèle tridimensionnel par homologie est basée sur l'alignement de la séquence d'intérêt avec l'empreinte structurale. Il est primordial que l'alignement utilisé soit de très bonne qualité pour que le modèle généré soit pertinent.

L'alignement de ces deux séquences est extrait de l'alignement multiple considéré en sortie de la version ajustée de PipeAlign. Il sera potentiellement plus fiable que s'il avait été calculé de manière indépendante. En effet, la construction d'un alignement multiple permet de prendre en compte plus d'informations relatives aux membres de la famille protéique et donc d'améliorer l'ajustement de l'alignement, particulièrement au niveau des domaines de moindre conservation.

8.3.4. Transferts des fichiers

La Figure 46 présente un récapitulatif de l'ensemble des calculs effectués sur le serveur *Star* de l'IGBMC.

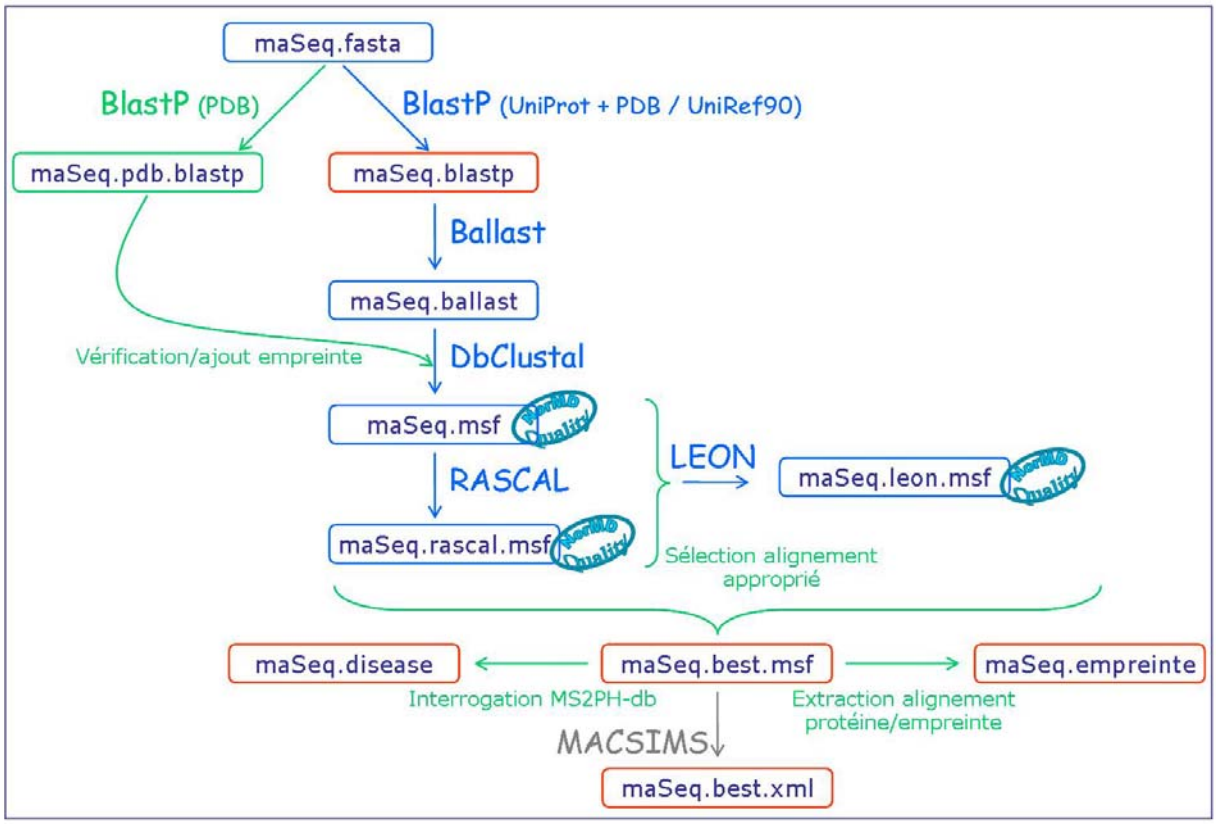


Figure 46 - Schématisation des calculs effectués sur le serveur *Star*. En bleu les étapes classiques de PipeAlign ; en vert, les ajustements apportés à PipeAlign. Les cadres rouges symbolisent les fichiers qui seront archivés afin d'être transférés sur le serveur de l'IBCP.

Un fichier d'archivage est finalement créé pour être transféré sur le serveur de l'IBCP. Celui-ci regroupe l'ensemble des fichiers qui sera mis à la disposition de l'utilisateur ainsi que ceux nécessaires à la construction du modèle tridimensionnel par homologie :

- le fichier issu de la recherche BlastP dans les banques de séquences,
- l'alignement multiple retenu en sortie de la version ajustée de PipeAlign, au format MSF,
- l'alignement annoté par MACSIMS, au format XML,
- l'alignement de la protéine d'intérêt avec l'empreinte structurale,
- le fichier de caractérisation des protéines alignées par rapport aux protéines impliquées dans les maladies génétiques humaines.

Une fois l'archive créée, le serveur de l'IBCP est contacté. Celui-ci rapatrie l'archive et effectue une ultime série de calculs.

8.3.5. Construction du modèle tridimensionnel

La protéine d'intérêt est modélisée en utilisant une version ajustée de Geno3D, qui accepte en entrée l'alignement de la protéine d'intérêt avec l'empreinte PDB calculée sur *Star*. Cet alignement est modifié pour s'adapter au format d'entrée de Geno3D.

Trois modèles tridimensionnels sont générés puis validés par alignement structural sur l'empreinte grâce au logiciel CE (Shindyalov and Bourne, 1998). Parmi les modèles ayant un RMSD (*Root Mean Square Deviation*) inférieur ou égal à 2 Å, le modèle présentant l'énergie minimale est finalement retenu. Les structures secondaires déterminées à partir du modèle par le programme DSSP (Kabsch and Sander, 1983) sont ensuite ajoutées au fichier PDB du modèle.

8.4. Interface de résultats

La page d'accueil des résultats de MAGOS (Figure 47) permet le chargement des fichiers de résultats, ainsi que l'accès à la page d'interconnexion Jmol des données. Le diagramme situé dans la partie haute de cette page représente la partie de la protéine d'intérêt qui a pu être modélisée (rectangle rouge), par rapport à sa longueur totale (trait noir).

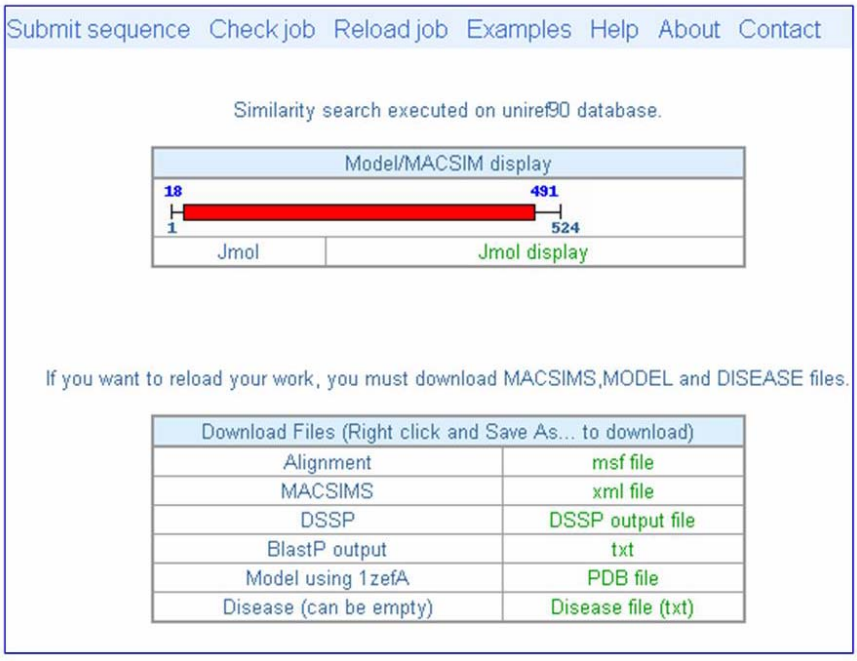


Figure 47 - Page d'accueil des résultats de MAGOS.

L'interface Jmol de résultat (Figure 48), dont l'originalité est de permettre la visualisation simultanée d'annotations fonctionnelles et structurales sur la séquence d'intérêt dans le cadre de l'alignement multiple et sur son modèle tridimensionnel, est l'atout principal de MAGOS. La génération de cette page est programmée en Python et utilise la méthodologie AJAX (*Asynchronous Javascript And XML*) pour la mise à jour dynamique de la page.

Figure 48 - Interface de résultat Jmol fournie par MAGOS.

L'interface de résultat est découpée en deux parties interconnectées.

La partie de droite présente l'alignement multiple. Le score NorMD de l'alignement est affiché au bas de la page. La séquence protéique soumise par l'utilisateur se situe toujours en première position de l'alignement et la fraction modélisée est soulignée. Par défaut, seule la première séquence de chaque sous famille est affichée (mode *collapse*), mais les séquences constituant les sous familles sont aisément accessibles (mode *uncollapse*). Les numéros d'accession des séquences alignées se trouvent à gauche de l'alignement, ce sont des éléments cliquables qui permettent l'ouverture d'une fenêtre externe (Figure 49) présentant les informations d'ordre général concernant la protéine d'intérêt. Leur couleur est fonction de l'éventuelle implication des protéines dans des maladies génétiques humaines : rouge pour les protéines directement impliquées dans une maladie, violet pour leurs homologues.

P05186	
Accession	P05186
Definition	Alkaline phosphatase tissue-nonspecific isozyme precursor EC 3.1.3.1 AP-TNAP TNSALP Alkalin
Group	4
Hydrophobicity	1.30
Length	2149
Lifedomain	Eukaryota
Nid	PPBT_HUMAN
Organism	Homo sapiens (Human)
Seq-name	P05186
Seq-type	protein
Taxid	9606
Weight	100
OMIM	146300
	171760
	241500
	241510

Figure 49 - Fenêtre externe relative aux numéros d'accession des séquences alignées, obtenue en cliquant sur ces numéros. Celle-ci contient un lien vers les banques UniProt et OMIM ainsi que des informations relatives à la protéine.

Les annotations fonctionnelles et structurales générées par MACSIMS sont accessibles par l'intermédiaire du menu *Features* et visualisables sur l'alignement par l'intermédiaire d'un code couleur.

La séquence d'intérêt présentée dans le cadre de l'alignement multiple annoté est connectée avec le modèle lui correspondant. Celui-ci est présenté dans la partie gauche, au cœur d'une fenêtre Jmol. Les annotations visualisées sur la partie modélisée de la séquence d'intérêt au sein de l'alignement sont également transposées sur le modèle, en respectant les mêmes codes couleurs. La sélection d'un résidu au niveau de la structure mettra en évidence ce résidu au sein de la séquence et vice versa.

Un certain nombre d'options de visualisation sont disponibles sous la fenêtre de visualisation du modèle, allant des options de bases (zoom, arrière-plan, couleurs, etc.) accessibles à tout un chacun, à l'utilisation de lignes de commande permettant des manipulations poussées du modèle par des utilisateurs aguerris.

L'utilisateur peut également choisir de visualiser l'alignement de la séquence d'intérêt avec l'empreinte PDB utilisée pour la construction du modèle dans la partie droite (Figure 50), à la place de l'alignement annoté.

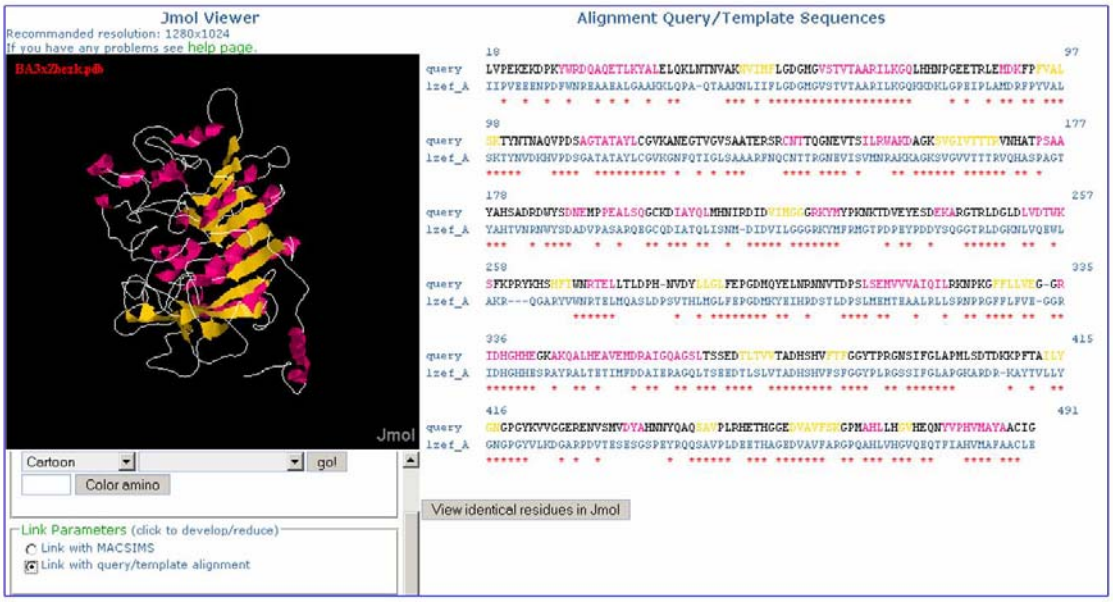


Figure 50 - Interface de résultat Jmol : alignement par paires utilisé pour la construction du modèle tridimensionnel. Les étoiles rouges symbolisent les résidus identiques entre la séquence d'intérêt et l'empreinte structurale.

Enfin, précisons qu'un mode qualifié de « sélectif » est activé par défaut, ce qui permet la mise en évidence simultanée d'un résidu sur le modèle et la partie modélisée de la protéine d'intérêt. Un mode « distance » est également disponible : il permet la sélection de deux résidus (au niveau du modèle ou de la partie modélisée de la séquence au sein de l'alignement) et l'obtention, en retour, de la distance qui sépare les deux résidus dans le modèle.

Chapitre 9 - MS2PH-db

MS2PH-db est une banque de données relationnelle, dédiée aux protéines impliquées dans les maladies génétiques humaines, au sein de laquelle les capacités d'intégration de MAGOS, présenté dans le chapitre précédent, ont notamment été mises à profit. MS2PH-db est hébergée à l'IBCP de Lyon et est publiquement accessible à l'adresse suivante : http://ms2phdb-pbil.ibcp.fr/cgi-bin/home.

9.1. Contenu de la banque

Chaque entrée de MS2PH-db est une protéine impliquée dans une maladie génétique humaine. Le protocole de détermination des entrées de la banque sera présenté dans la section 9.3, p. 124.

Pour chacune de ses entrées, MS2PH-db fournit trois types de vues : une vue évolutive, une vue structurale et une vue informationnelle. Ces vues peuvent être interconnectées par l'intermédiaire d'une interface graphique dérivée de celle de MAGOS.

La vue évolutive est illustrée par l'intermédiaire de deux alignements multiples de séquences complètes. Nous avons choisi de présenter un alignement regroupant les séquences eucaryotes homologues, ainsi qu'un alignement dédié à l'analyse structurale pour lequel la sélection des séquences est basée sur les résultats de l'étude présentée dans le Chapitre 11 - *Alignements multiples dédiés à l'analyse structurale*, p. 145.

La vue structurale est, quant à elle, illustrée par l'intermédiaire d'un ou plusieurs modèles tridimensionnels construits par homologie.

Les protocoles mis en place pour la construction de ces vues seront détaillés dans la section 9.4.1, p. 126.

Enfin, la vue informationnelle est schématisée par plusieurs types de données :

- des données que l'on peut qualifier de « généralistes », comme le nom de la protéine, le nom du gène, etc., complétées par des liens croisés vers des banques de données comme UniProt, OMIM ou encore GO,
- les annotations fonctionnelles et structurales des alignements,
- des données de mutations, associées à des données phénotypiques (limitées au nom de la maladie observée dans la version actuelle de la banque).

9.2. Structure de la banque

MS2PH-db est une base de données relationnelle gérée par le système PostgreSQL 7.3.4, dérivée du modèle de Modeome3D (Garnier *et al.*, manuscrit en préparation), un système de gestion et d'analyse de modèles tridimensionnels disponible en ligne à l'adresse suivante : http://modeome3d-pbil.ibcp.fr/.

MS2PH-db, dont le schéma relationnel est disponible en Annexe 4, p. 234, est constituée de 24 tables distinctes. La structure de la base de données est principalement centrée sur la table contenant des informations générales relatives à la protéine. Les autres tables regroupent différents niveaux d'informations, relatifs aux modèles tridimensionnels construits et à leurs caractéristiques (structures secondaires, coordonnées dans l'espace), aux mutations et phénotypes associés, etc.

Ces tables sont reliées entre elles par des clés étrangères, chaque protéine possédant une clé étrangère principale unique.

Une table indépendante contient des informations générales relatives à la banque, comme le nombre d'entrées et de modèles tridimensionnels qu'elle contient.

9.3. Identification des entrées de la banque

MS2PH-db se définit comme une infrastructure pouvant servir de base à l'analyse globale des mutations et de leurs conséquences structurales et phénotypiques et est dédiée aux protéines connues pour être impliquées dans des maladies monogéniques humaines.

La liste des protéines d'entrées de MS2PH-db a été créée en se basant sur trois banques de données publiques : la GDB, GeneCards et la HGMD (Figure 51).

La GDB et GeneCards présentent l'avantage de fournir un accès aisé à une liste de gènes impliqués dans des maladies génétiques humaines. Les données accessibles par la GDB sont extraites d'OMIM et classées par chromosome (http://www.gdb.org/gdbreports/GeneticDiseases.html), alors qu'elles sont classées par nom de gène dans GeneCards (http://www.genecards.org/cgi-bin/listdiseasecards.pl?type=full). La liste offerte par GeneCards regroupe des gènes liés, directement ou indirectement, à une maladie et sont extraites de plusieurs banques de données publiques.

La HGMD contient des données de mutations relatives à des gènes impliqués dans des maladies monogéniques humaines.

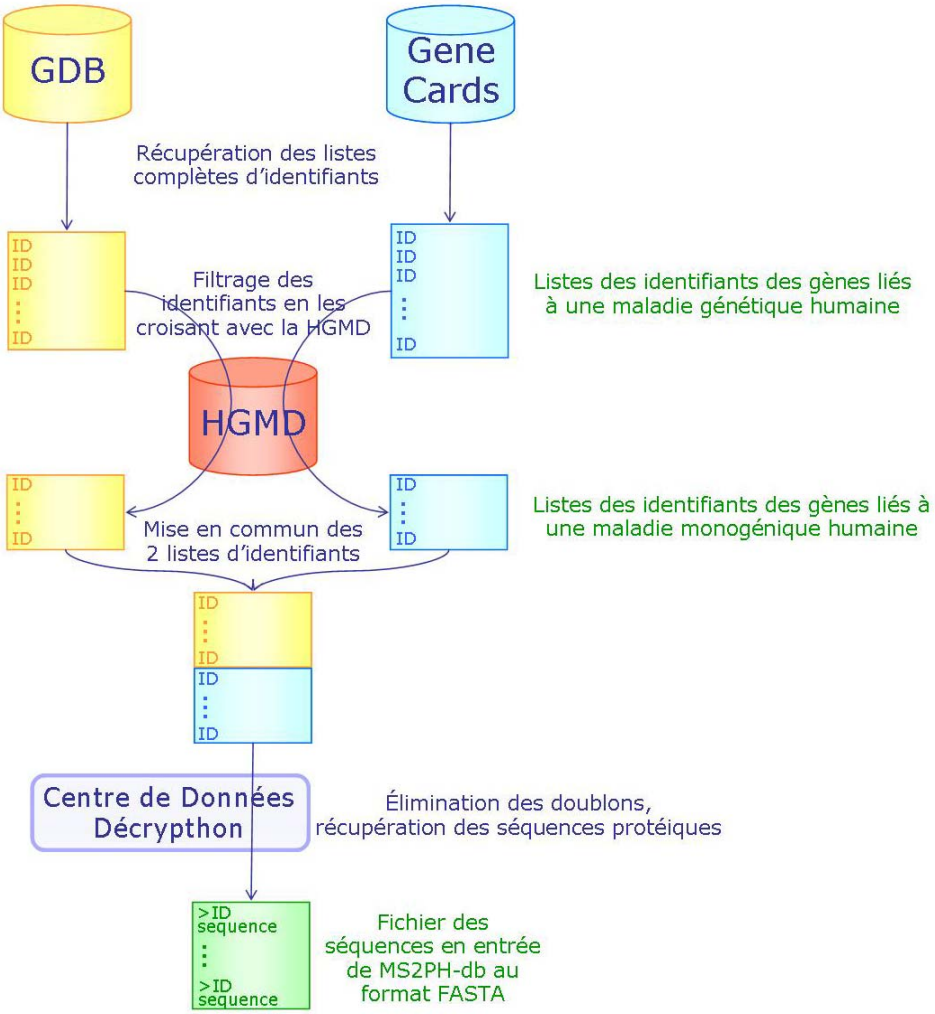


Figure 51 - Sélection des protéines en entrée de MS2PH-db. Cette procédure aboutit à la création d'un fichier unique contenant l'ensemble des séquences protéiques, au format FASTA, qui vont constituer les entrées de la banque MS2PH-db.

Afin d'être sûrs de ne sélectionner que des gènes impliqués dans des maladies monogéniques, nous avons mis en place un script qui croise les données issues de la GDB et de GeneCards avec les données de la HGMD, en prenant garde à ne pas intégrer de doublons.

Les séquences protéiques associées aux gènes finalement retenus représentent les entrées de MS2PH-db.

9.4. Calcul des données relatives à chaque entrée

Schématiquement, deux catégories de données sont regroupées au sein de MS2PH-db : (i) des données issues de diverses banques de données publiques, comme UniProt, OMIM et GO, et récupérées automatiquement à partir du numéro d'accession de chacune des entrées de la banque ; (ii) des données calculées à partir de leur séquence.

La prochaine section présentera le protocole mis en place pour le calcul des données relatives à chaque entrée de la banque, à partir des séquences protéiques.

9.4.1. Protocole général

Pour chacune des protéines en entrée de MS2PH-db, une série de calculs est initiée. Le protocole mis en place pour le calcul des données est proche de celui développé dans le cadre de MAGOS (*voir* section 8.3, p. 115).

9.4.1.1. Sélection des empreintes pour la construction des modèles tridimensionnels

La sélection des empreintes se fait automatiquement. Plusieurs empreintes non-chevauchantes peuvent potentiellement être sélectionnées pour une même séquence protéique.

La sélection des empreintes est initiée par une recherche de similarité au sein de la banque de structures PDB, à l'aide du programme BlastP. La méthodologie développée permet la sélection d'empreintes partageant au moins 25% d'identité avec la séquence d'intérêt, sachant que le pourcentage d'identité est calculé par rapport à la séquence la plus courte, sur la base d'un alignement global par paires.

Les empreintes potentielles sont ensuite analysées séquenciellement et comparées deux à deux :

- si deux empreintes se chevauchent, une seule sera sélectionnée. Cette sélection est basée, d'une part, sur le pourcentage d'identité entre l'empreinte et la protéine d'intérêt et, d'autre part, sur la longueur de l'empreinte. Si les pourcentages d'identité diffèrent de plus de 10%, l'empreinte partageant l'identité la plus forte est sélectionnée. Dans le cas contraire, l'empreinte la plus longue est choisie,
- les empreintes qui ne se chevauchent pas sont toutes deux sélectionnées.

9.4.1.2. Construction et annotation des alignements multiples

Deux alignements multiples de séquences complètes sont calculés et annotés de manière indépendante. Leur construction est basée sur la version ajustée de PipeAlign déployée dans le cadre de MAGOS (*voir* section 8.3.2, p. 116).

Les différences entre les deux alignements résident, d'une part, dans les banques utilisées pour effectuer les recherches de similarité Blast et, d'autre part, dans les méthodes de sélection des séquences à aligner.

Dans le cadre de l'alignement considéré comme l'alignement « structural », la recherche de similarité se fait dans la banque non redondante UniRef90 et la méthode d'échantillonnage

des bandelettes (*voir* section 7.1.1.1, p. 104) est appliquée pour la sélection des séquences à aligner. Le nombre maximal de séquences à aligner est fixé à 100.

Dans le cadre du second alignement, la recherche de similarité se fait simultanément dans les banques UniProt et PDB. Le filtre Eucaryote (*voir* section 7.1.2, p. 107) est appliqué sur les séquences d'UniProt, limitant la sélection des séquences à celles issues d'organismes eucaryotes. Le nombre total de séquences sélectionnées est limité à 200.

Remarquons que lors de la construction des deux alignements, les empreintes sélectionnées lors de la première étape du protocole sont ajoutées à la liste des séquences à aligner si elles ne font pas partie du jeu initial de séquences choisies.

9.4.1.3. Construction des modèles tridimensionnels

Comme dans le cadre de MAGOS, l'alignement par paires de chacune des empreintes avec la séquence d'intérêt est extrait d'un alignement multiple (l'alignement structural). Ces alignements sont ensuite soumis à Geno3D pour la construction des modèles par homologie.

Sur la base du modèle généré, l'accessibilité au solvant des résidus est calculée par DSSP pour ensuite être intégrée à la base de données.

9.4.2. Implémentation du protocole sur la grille de calcul Décrypthon

A ce jour, près de 2 000 protéines ont été identifiées comme étant impliquées dans des maladies monogéniques. La gestion de l'ensemble des données relatives à ces protéines demande de disposer d'une très grande puissance de calcul. C'est dans le contexte de développements tels que celui de MS2PH-db, que la mise à disposition d'une grille de calcul prend tout son sens.

L'ensemble du protocole qui conduit de la protéine d'intérêt à l'obtention des deux alignements multiples annotés, en passant par la sélection des empreintes et l'extraction des alignements nécessaires à la modélisation par homologie, a par conséquent été porté sur la grille de calcul universitaire Décrypthon. A l'heure actuelle (septembre 2007), la construction des modèles par homologie est toujours réalisée sur les serveurs de l'IBCP et les possibilités d'implémenter cette ultime étape sur la grille sont en cours d'évaluation.

La migration des applications sur la grille a été réalisée en collaboration avec Raphaël Bolze et Nicolas Bard de l'ENS de Lyon et avec le soutien de plusieurs consultants d'IBM. Il a nécessité l'adaptation des codes sources de la plupart des applications ainsi que la réécriture de certains éléments, afin de les rendre conformes à l'environnement des grilles de calcul.

Le projet MS2PH, dans le cadre duquel MS2PH-db a été développée, étant un projet pilote du programme Décrypthon, nous avons été les premiers utilisateurs de la grille universitaire à y porter nos applications. C'est pourquoi une longue étape de validation des résultats

obtenus sur chaque nœud de calcul a due être menée. Dans un premier temps, chacun des modules élémentaires a été validé manuellement après compilation sur le serveur de grille. Dans un second temps, et suite à la validation technique de l'application, des comparaisons portant sur des données massives, calculées à la fois au laboratoire et sur la grille, ont permis de valider la robustesse et la fiabilité des applications ainsi « gridifiées ».

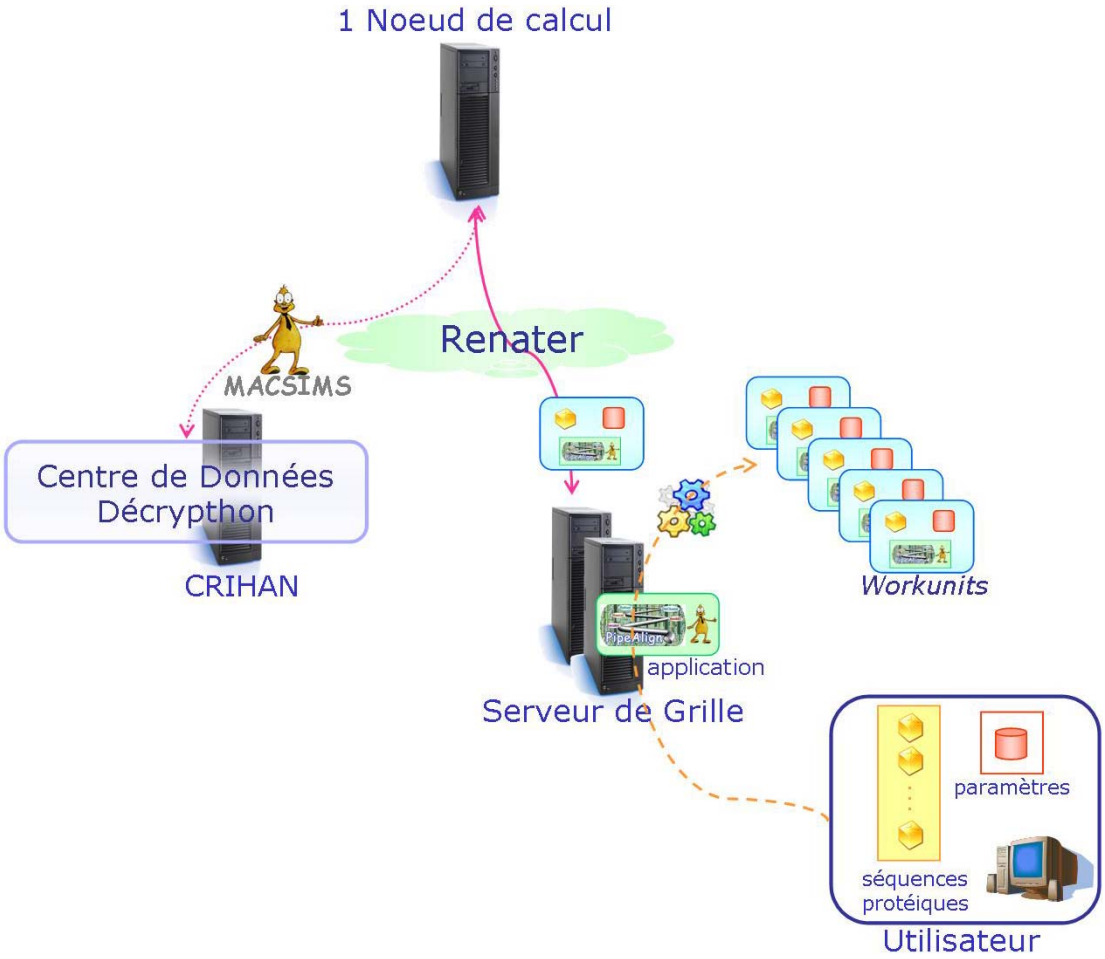


Figure 52 - Schématisation du fonctionnement de nos applications sur la grille (cas d'un seul nœud de calcul).

Nos applications sur la grille de calcul demandent en entrée un fichier contenant l'ensemble de nos protéines d'intérêt au format FASTA. Ce fichier est pris en charge par le serveur de grille, qui va le découper en autant de tâches, couramment appelées *job*, que de séquences protéiques contenues dans le fichier.

Le serveur distribue ces tâches aux clients de la grille, représentés par les nœuds de calcul situés sur les serveurs distants, en fonction de leur disponibilité. Les tâches sont illustrées par l'envoi d'une unité de travail, couramment appelée *workunit*, qui encapsule l'ensemble des données et des applications nécessaires pour le bon déroulement des calculs (Figure 52).

Une particularité offerte par le logiciel de grille installé sur le serveur central est de permettre de ne pas faire transiter toutes les données avec chaque *workunit*. Dans le cadre de notre application, c'est le cas des banques Blast. En effet, ces dernières sont envoyées sur les serveurs distants lors de l'envoi de la première tâche et resteront en place et fonctionnelles jusqu'à la fin de la session de calculs émis. Ceci permet de réduire considérablement la quantité de données à faire transiter entre les serveurs et d'augmenter par conséquent les performances de la grille.

Précisons que le déploiement de MACSIMS sur cette grille de calcul n'a été possible que grâce à la disponibilité du Centre de Données Décrypthon, qui peut être directement interrogé par chaque nœud de calcul. En effet, il n'était pas envisageable de faire transiter directement sur les nœuds de la grille l'ensemble des banques interrogées par ce programme.

Bien que la mise à jour des données de MS2PH-db soit lancée de manière automatique, une interface de soumission a été mise à notre disposition pour nous permettre de tirer profit de ces développements dans le cadre de projets menés en parallèle au laboratoire.

9.5. Interrogation de MS2PH-db

MS2PH-db peut être interrogée par l'intermédiaire de formulaires ou d'une recherche de similarité au sein des entrées de la banque.

Trois formulaires distincts d'interrogation de MS2PH-db sont disponibles. Le premier (*Simple Search*) est destiné à la réalisation d'une recherche simple. Il comporte deux champs et un nombre limité de mots clés (numéro d'accession de la protéine, nom de la protéine, nom d'une maladie, etc.). Une aide contextuelle à la recherche a été mise en place pour ce formulaire : au moment de la frappe des premières lettres, les premiers résultats sont dynamiquement affichés afin d'aider l'utilisateur si sa recherche est assez peu précise (Figure 53).

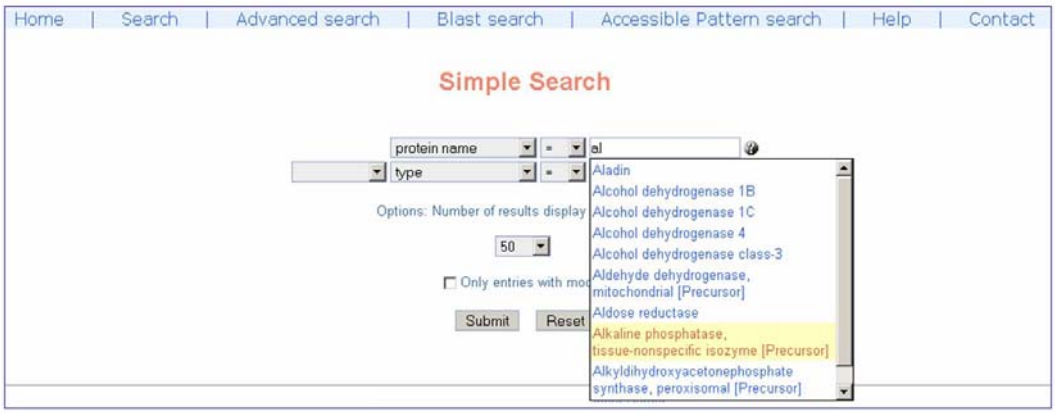


Figure 53 - Formulaire d'interrogation « simple » de MS2PH-db, intégrant une aide contextuelle à la recherche.

Le deuxième formulaire (*Advanced Search*) permet de faire une recherche avancée, en combinant jusqu'à quatre mots clés différents. En plus des mots clés disponibles dans le premier formulaire, la recherche peut s'effectuer sur la proportion de la protéine modélisée, le nombre d'hélices du modèle généré, etc…

Le dernier formulaire *(3D Accessible Pattern Search*) permet de rechercher les protéines présentant un motif de deux à dix résidus situés dans un environnement tridimensionnel restreint et dont le diamètre est ajustable par l'utilisateur. Cette recherche est basée sur un algorithme récursif de recherche de résidus accessibles au solvant proche dans l'espace. Une matrice de distances est construite à partir des résidus accessibles remplissant les conditions de l'utilisateur, puis les combinaisons potentielles sont testées pour ne garder que celles qui satisfont les contraintes de distance. L'utilisateur pourra ainsi retrouver toutes les protéines ayant un modèle possédant au moins une alanine, un acide glutamique, une glycine et une thréonine accessibles situés dans une sphère de 12 Å, en indiquant le motif « AEGT » et en précisant un diamètre de 12 Å.

L'interrogation de la banque, par l'intermédiaire d'une recherche de similarité, réalisée par BlastP, est possible grâce à la génération automatique, lors de chaque mise à jour, d'une banque de séquences au format Blast, à partir des séquences protéiques en entrée de MS2PH-db.

Toutes ces méthodes d'interrogation et d'extraction de données appellent des scripts CGI qui interrogent la base de données et retournent une page HTML comme résultat. Une aide interactive a été intégrée pour faciliter la saisie et les recherches. Elle est accessible par l'intermédiaire d'une icône représentant un point d'interrogation.

9.6. Interface de résultats

La Figure 54 présente l'interface de résultats obtenue suite à une requête.

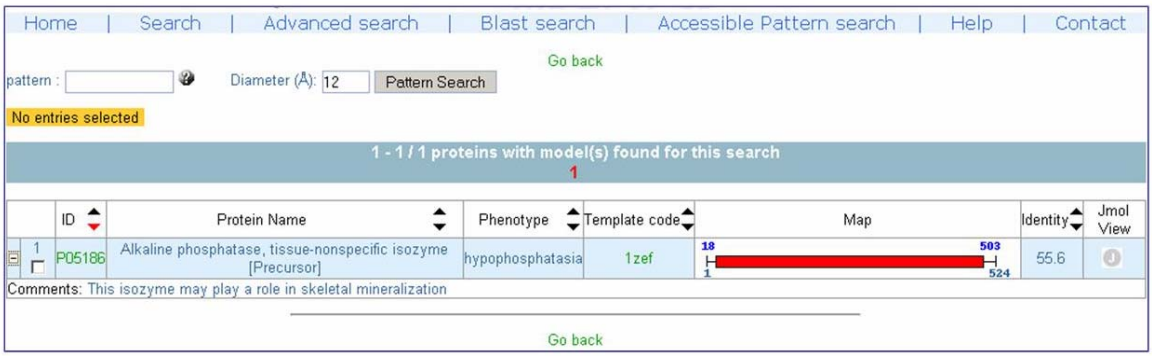
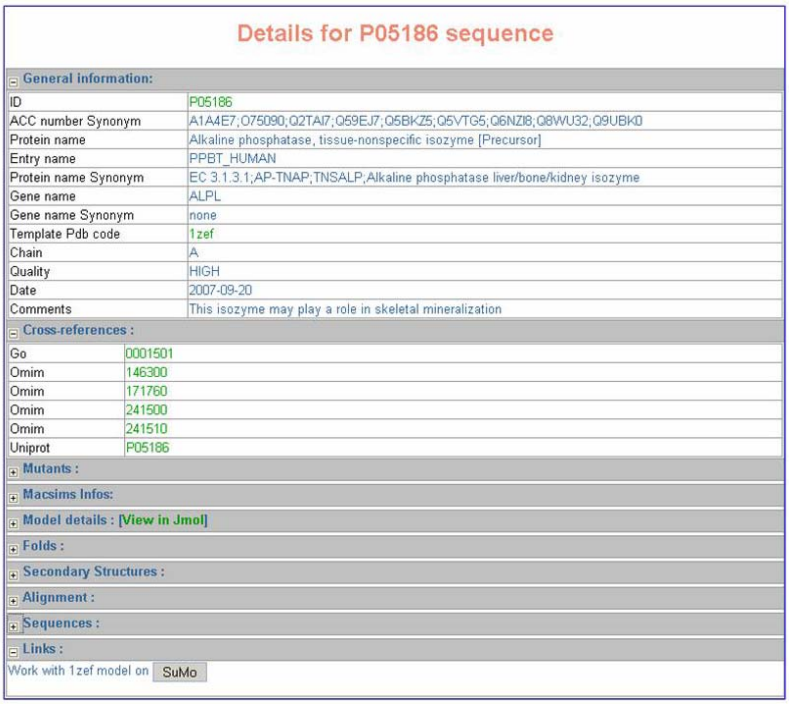


Figure 54 - Capture d'écran du résultat d'une requête simple (interrogation du champ *Protein Name* avec le terme « alkaline phosphatase »).

Le tableau de résultats présente un récapitulatif des données associées aux protéines répondant aux critères de la recherche effectuée : il contient l'identifiant de la protéine, son nom ainsi que le nom d'une maladie associée à cette protéine. Lorsqu'un modèle tridimensionnel a pu être construit, le code PDB de l'empreinte structurale utilisée pour la génération du modèle est précisé et le diagramme situé dans la colonne *Map* représente la partie de la protéine qui a pu être modélisée avec cette empreinte. Le diagramme est une zone interactive, par l'intermédiaire de laquelle il est possible de télécharger le fichier du modèle tridimensionnel au format PDB. Le pourcentage d'identité entre la protéine d'intérêt et l'empreinte structurale est précisé dans la colonne *Identity*.

Plusieurs liens sont accessibles par l'intermédiaire de cette page. Par exemple, cliquer sur le numéro d'accession d'une protéine permet d'accéder à la page d'accueil de cette dernière (Figure 55), qui peut être considérée comme une carte d'identité de la protéine d'intérêt. Celle-ci regroupe des informations généralistes dans l'onglet *General information*, des liens vers d'autres banques de données publiques comme GO et OMIM au niveau de l'onglet *Cross references*, mais aussi des informations relatives aux mutants connus et aux maladies associées dans l'onglet *Mutants* ainsi que des détails sur le modèle tridimensionnel construit (onglet *Model details*) etc.



Details for P05186 sequence

General information:

ID	P05186
ACC number Synonym	A1A4E7;O75090;Q2TAI7;Q59EJ7;Q5BKZ5;Q5VTG5;Q6NZI8;Q8WU32;Q9UBK0
Protein name	Alkaline phosphatase, tissue-nonspecific isozyme [Precursor]
Entry name	PPBT_HUMAN
Protein name Synonym	EC 3.1.3.1;AP-TNAP;TNSALP;Alkaline phosphatase liver/bone/kidney isozyme
Gene name	ALPL
Gene name Synonym	none
Template Pdb code	1zef
Chain	A
Quality	HIGH
Date	2007-09-20
Comments	This isozyme may play a role in skeletal mineralization

Cross-references :

Go	0001501
Omim	146300
Omim	171760
Omim	241500
Omim	241510
Uniprot	P05186

Mutants :

Macsims Infos:

Model details : [View in Jmol]

Folds :

Secondary Structures :

Alignment :

Sequences :

Links :

Work with 1zef model on SuMo

Figure 55 - Carte d'identité d'une protéine de la banque, la phosphatase alcaline non tissu-spécifique.

L'accès à l'interface Jmol d'interconnexion des différentes vues afférentes à la protéine peut se faire par l'intermédiaire du « J » de la colonne *Jmol View* du tableau de résultats (Figure 54)

ou par le lien *View in Jmol* de la page principale de la protéine (Figure 55). L'interface (Figure 56) est très proche de celle de MAGOS (*voir* section 8.4, p. 118) et permet l'étude conjointe des données de séquence, de structure, de fonction, d'évolution, ainsi que la localisation des mutations connues.

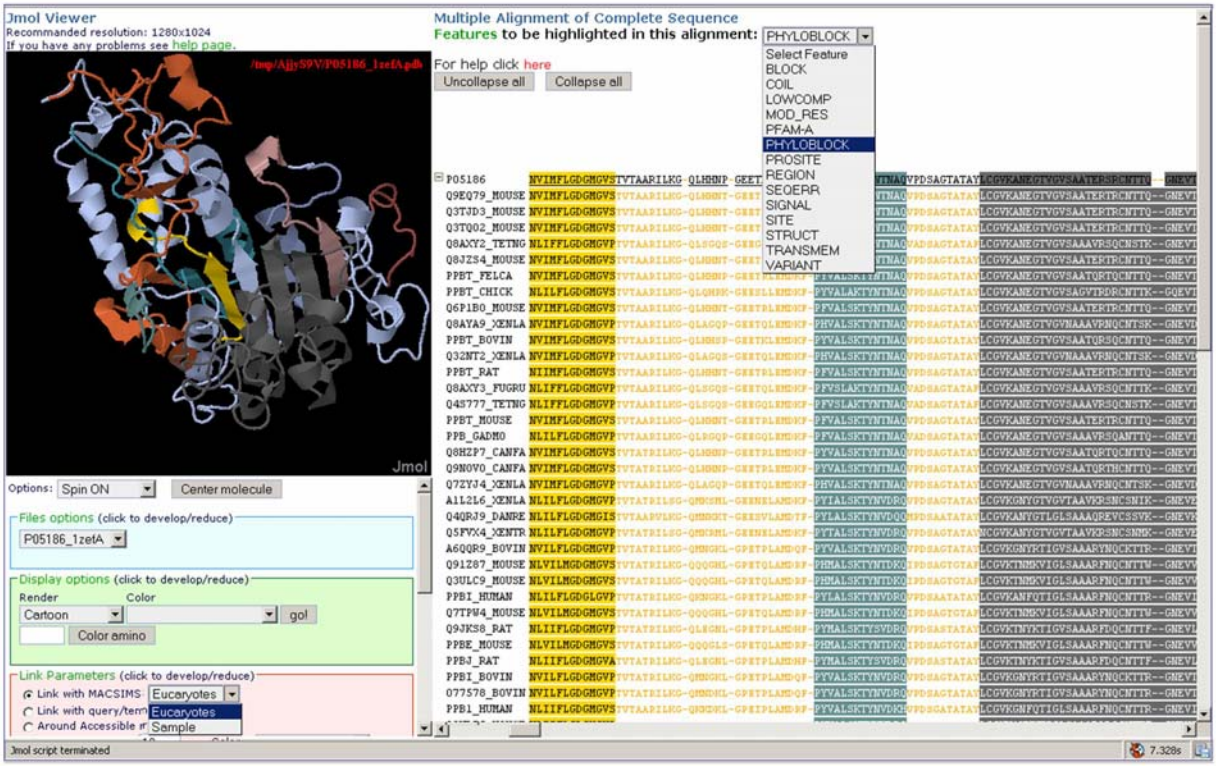


Figure 56 - Interface Jmol d'interconnexion des différentes « vues » relatives à la phosphatase alcaline non tissu-spécifique.

L'alignement multiple annoté par MACSIMS est visualisable dans la partie droite de l'interface et est connecté au modèle tridimensionnel de la protéine. L'utilisateur peut choisir de visualiser l'alignement ayant subit l'étape d'échantillonnage par la méthode des bandelettes (*Sample*), l'alignement regroupant les séquences eucaryotes homologues (*Eucaryotes*) (*voir* section 9.4.1.2, p. 126), ou l'alignement de la séquence d'intérêt avec l'empreinte PDB utilisé pour la construction du modèle.

Quelques fonctionnalités supplémentaires ont été intégrées par rapport à MAGOS (Figure 57). De par la disponibilité des données de mutation associées à des données phénotypiques, l'interface offre la possibilité de localiser les mutations faux-sens référencées dans la banque au sein du modèle tridimensionnel de la protéine ainsi qu'au niveau de l'alignement multiple annoté. Il est également possible de générer à la volée le modèle de la protéine mutée, en cliquant sur *Generate*, dans l'onglet *Mutants*. Le modèle n'est pas reconstruit dans son intégralité, mais la modification engendrée par la mutation lui est appliquée, une étape de minimisation d'énergie est ensuite réalisée à l'aide du logiciel CNS. L'onglet *Files options*

permet l'affichage, dans la fenêtre Jmol, de n'importe quel modèle tridimensionnel parmi l'ensemble des modèles générés.

Enfin, l'option *Around Accessible residues* permet de visualiser les résidus de surface situés dans un environnement structural proche d'un résidu d'intérêt.

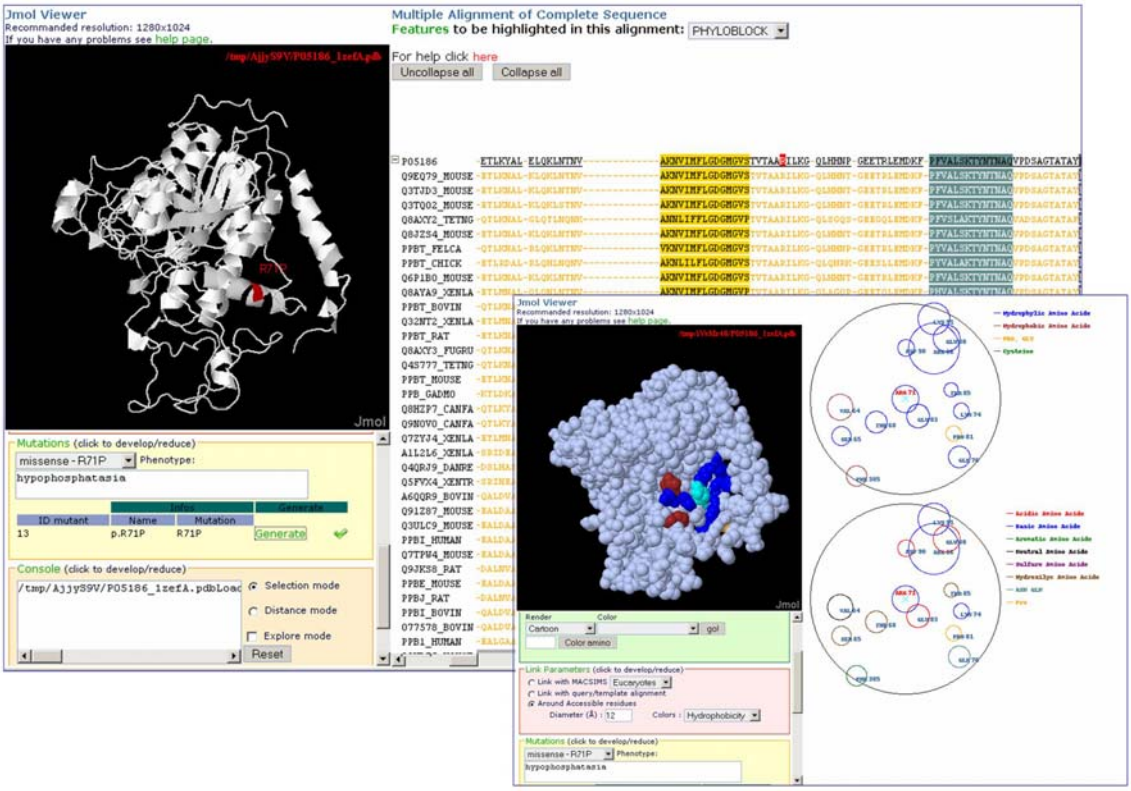


Figure 57 - Fonctionnalités supplémentaires intégrées à l'interface Jmol, par rapport à MAGOS. Visualisation de la localisation d'une mutation (à gauche) au niveau du modèle structural ainsi que de l'alignement ; visualisation des résidus structuralement proches d'un résidu d'intérêt (à droite).

9.7. Mise à jour des données

La banque de données est mise à jour de manière automatique tous les deux mois. La liste des protéines impliquées dans des maladies monogéniques humaines est dans un premier temps mise à jour. Cette étape est effectuée sur le serveur *Star* de l'IGBMC. Le fichier, qui contient l'ensemble des protéines entrées de la banque au format FASTA, est dans un second temps soumis à la grille de calcul, sur laquelle les deux alignements annotés sont construits de manière indépendante pour chaque protéine. Les modèles tridimensionnels ne sont recalculés que si les empreintes identifiées diffèrent des empreintes utilisées pour générer les modèles stockés dans la banque.

MS2PH-db est mise à jour par l'intermédiaire d'un rechargement complet de l'ensemble des données.

Résultats & Discussion

Chapitre 10 - Problématique du contenu informationnel des alignements multiples

L'analyse approfondie d'une protéine dont le dysfonctionnement est à l'origine d'une pathologie est une étape primordiale dans l'étude d'une maladie génétique. Cependant, comme nous l'avons déjà évoqué, l'ère post-génomique est caractérisée par une quantité considérable et sans cesse croissante d'informations disponibles. Cette accumulation de données, qui offre de grandes perspectives en termes de compréhension du vivant, a de profondes répercussions notamment sur la quantité, la qualité et le type d'informations à conserver ainsi que sur la manière d'aborder et de mener à bien une analyse.

Pendant longtemps, la plupart des études ont été réalisées en veillant à intégrer un niveau d'informations maximal. Est-il dorénavant indispensable, voire possible ou pertinent de traiter l'ensemble des données accessibles, quelque soit le type d'étude mené ? Cette question est au centre des débats, de manière consciente ou inconsciente, et il paraît indéniable que la réponse est négative. En effet, nous avons atteint une situation où « toute l'information » représente « trop d'informations » et tenter de la gérer dans sa globalité devient déraisonnable, non seulement en termes de temps de traitement, mais aussi en termes d'extractibilité du message approprié et ce tout particulièrement dans le cadre d'études massives.

L'organisation des données est, dans ce contexte, au cœur des préoccupations de la bioinformatique et de la biologie moderne. Les stratégies organisationnelles mises en œuvre visent à permettre au scientifique d'accéder au niveau d'information adéquat pour son étude. Cette réorganisation, qui a pour ambition de « simplifier » les données et leur accès, se révèle sous diverses formes, telles que de la mise en place de standards et d'ontologies (Rubin *et al.*, 2006) ou le développement de banques de données dédiées. Ces informations « simplifiées » permettent de rendre accessibles de nouveaux contenus informationnels. A titre d'exemples, on peut citer les banques de données CATH (Orengo *et al.*, 1997) ou SCOP (Murzin *et al.*, 1995), qui mettent au second plan l'ensemble des informations présentent dans la PDB, pour privilégier la notion du mode de repliement structural commun, ou bien encore des banques de séquences pré-traitées comme UniRef90 ou UniRef50 (Suzek *et al.*, 2007) qui, en ne retenant que des séquences variables (moins de 90% ou 50% d'identité respectivement), vont filtrer les séquences ayant faiblement évoluées, éliminant ainsi fréquemment les clades trop proches, donc faiblement informatifs.

Le corollaire de toute « simplification » étant la perte d'un certain niveau d'information, il est essentiel d'une part, de mesurer l'impact de la perte d'informations sur la résolution d'un probléme donné et d'autre part, de déterminer quels types de problèmes, une information simplifiée permet d'aborder.

La construction et l'analyse d'alignements multiples qui, en replaçant une protéine au cœur de sa famille, peuvent notamment représenter un support idéal pour l'interprétation des conséquences des mutations qui l'affecte, ne dérogent pas à ces nouvelles règles.

Nous avons été amenés, au cours de nos travaux, à mener une réflexion sur la double problématique que représentent l'alignement multiple et son contenu informationnel à l'ère post-génomique. Ceci a donné lieu à une publication, à paraître dans un livre consacré aux alignements multiples (*voir* Annexe 7, p. 239) et au développement d'une stratégie d'échantillonnage pour la construction d'alignements multiples dédiés aux analyses structurales, qui sera décrite dans le prochain chapitre (*voir* Chapitre 11 - *Alignements multiples dédiés à l'analyse structurale*, p. 145).

Dans la suite de ce chapitre, nous présenterons les points clefs de notre réflexion sur la mise en place de stratégies adaptées dans le cadre de la construction et de l'exploitation des alignements multiples de séquences protéiques. Nous reviendrons succinctement sur les rôles des alignements multiples et sur leur construction traditionnelle, puis nous aborderons les problèmes rencontrés par les biologistes dans ces démarches à l'ère post-génomique et les stratégies mises en place pour y pallier.

10.1. Alignement multiple de séquences protéiques : un outil central en biologie moderne

L'alignement multiple est un outil puissant d'analyse de familles de protéines, en particulier lorsque les séquences sont alignées sur toute leur longueur, pour obtenir un alignement multiple de séquences complètes, ou MACS, dont les rôles se sont diversifiés au cours des dernières années pour devenir un outil central en biologie moderne (Figure 58).

Le MACS permet de réaliser à la fois une analyse « horizontale » des séquences alignées, en tenant compte de leur longueur totale, et une analyse « verticale », en les intégrant dans leur contexte évolutif (Lecompte *et al.*, 2001b). L'analyse « horizontale » des séquences offre la possibilité d'étudier l'organisation en domaines des protéines et permet notamment la mise en évidence d'insertions ou de délétions. L'analyse « verticale », quant à elle, repose principalement sur l'étude de la conservation des résidus au sein des séquences alignées. Elle permet notamment la mise en évidence des résidus, des segments ou des domaines conservés dans toutes les séquences de la famille, qui représentent généralement des résidus ou des motifs ayant une importance structurale ou fonctionnelle, ainsi que des résidus ou

régions conservés spécifiquement dans une sous-famille, qui représentent des résidus potentiellement discriminants de cette dernière (del Sol Mesa *et al.*, 2003).

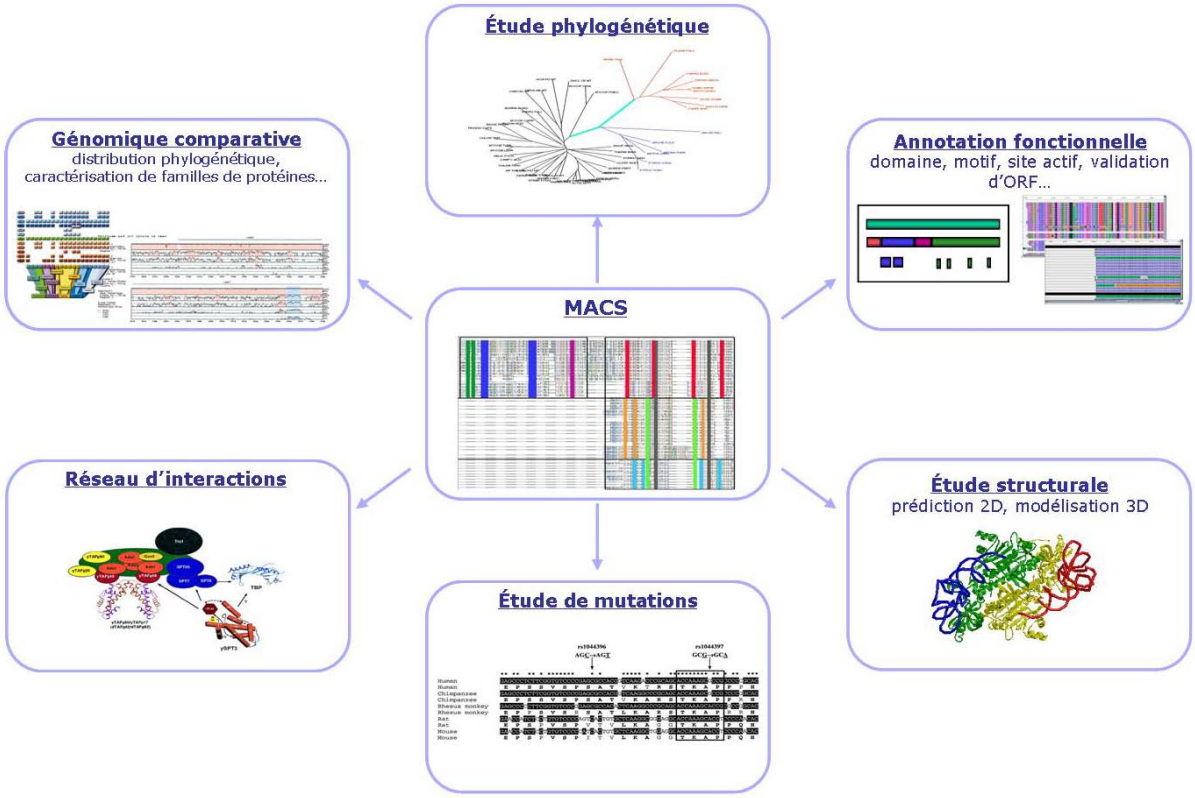


Figure 58 - Rôle central des alignements multiples de séquences protéiques en biologie moderne.

Classiquement, les rôles des alignements multiples de protéines peuvent être partagés en six thématiques majeures (Figure 58), au sein desquelles des problématiques plus spécifiques peuvent encore être distinguées, chacune demandant un certain degré de précision et l'accès à un niveau d'information propre :

- les alignements multiples ont, historiquement, été mis à profit dans la détermination des relations phylogénétiques entre organismes, par la construction d'arbres phylogénétiques (Phillips *et al.*, 2000),
- ils sont à présent couramment utilisés dans le cadre de la caractérisation fonctionnelle de protéines, principalement pour la validation de séquences protéiques, grâce à la mise en évidence d'erreurs de séquençage, de décalage du cadre de lecture, d'erreurs de prédiction de la structure du gène (codon initiateur, site d'épissage exon/intron) (Bianchetti *et al.*, 2005), ou pour la réalisation des prédictions fonctionnelles par la propagation d'informations des séquences annotées vers les séquences non annotées,

- dans un contexte structural, leur utilisation permet d'améliorer la prédiction des éléments de structures secondaires (Heringa, 2000; Lee *et al.*, 2006) et tertiaires (Moult, 2005; Yang and Honig, 2000),
- ils peuvent aussi être mis à profit dans le cadre d'analyses de co-variation de résidus qui, lorsqu'elles sont menées sur deux protéines connues pour interagir, fournissent une aide précieuse à la sélection de l'arrangement structural d'interaction correct (Pazos and Valencia, 2002),
- ils assurent également un rôle prépondérant en génomique comparative. Les protéines orthologues pouvant être discriminées des paralogues par leur intermédiaire, ils constituent un support de choix pour la construction de profils phylogénétiques valides (Pellegrini *et al.*, 1999) qui, par l'étude des co-présences/co-absences de protéines, peuvent permettre d'identifier la distribution de complexes ou de voies métaboliques dans les organismes,
- ils ont déjà révélé leur intérêt dans le développement de systèmes bioinformatiques dédiés à la prédiction des effets des mutations (Mooney, 2005) (*voir* Chapitre 5 - *Vers une approche intégrative des mutations et de leurs conséquences*, p. 67) et sont à même de fournir une prédiction des conséquences fonctionnelles et structurales des mutations sur une protéine.

On peut ainsi remarquer que les alignements multiples permettent d'adresser un grand nombre de questions biologiques. Le niveau d'informations à extraire des alignements diffère selon la problématique traitée, c'est pourquoi on peut considérer que le besoin en contenu informationnel des alignements varie.

10.2. Alignement multiple : construction traditionnelle

La construction traditionnelle d'un alignement multiple peut schématiquement être découpée en trois étapes : (i) la recherche, à partir d'une séquence d'intérêt, de séquences potentiellement homologues dans les banques de données ; (ii) la construction de l'alignement à partir des séquences mises en évidence ; (iii) la correction (automatique et/ou manuelle) de l'alignement afin d'obtenir un alignement de qualité, qui reflète les relations évolutives, structurales et fonctionnelles entre les séquences alignées.

Ces différentes étapes sont détaillées au sein de la publication (*voir* Annexe 7, p. 239).

A l'heure actuelle, le biologiste a à sa disposition de nombreux programmes de construction d'alignements multiples (revue dans (Edgar and Batzoglou, 2006)). Ces derniers sont basés sur différents types d'algorithmes (Figure 59), chacun présentant des forces et des faiblesses en fonction des séquences à aligner (Thompson *et al.*, 1999). Des développements récents, basés sur des algorithmes combinatoires (Do *et al.*, 2005; Edgar, 2004; Katoh *et al.*, 2005;

Thompson *et al.*, 2000) permettent de disposer de programmes performants, associant qualité des résultats et rapidité d'exécution.

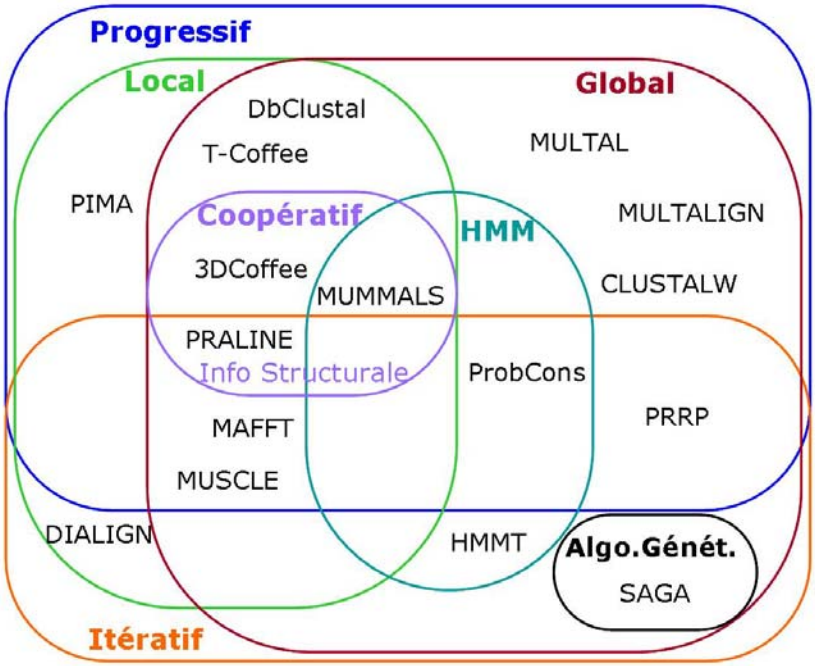


Figure 59 - Aperçu des principaux programmes d'alignements multiples disponibles, classés par types d'algorithmes. (adapté et complété à partir de (Thompson *et al.*, 1999))

Des progrès restent toutefois à réaliser pour améliorer encore la qualité des alignements multiples obtenus de manière automatique. Ceci représente un vaste sujet de recherche, mais nous ne l'approfondirons pas dans le cadre de cette thèse. Nous nous concentrerons à présent sur les problèmes rencontrés lors de la construction d'alignements multiples, liés à l'accumulation des données au sein des banques de séquences.

10.3. La construction d'alignements multiples à l'ère post-génomique

La construction d'alignements multiples se heurte aujourd'hui à un problème transversal à de nombreuses études *in silico* : la quantité d'informations disponibles au sein des banques de données, qui fait la force de l'ère post-génomique, a l'inconvénient de placer le scientifique face à une masse déraisonnable de données à traiter.

La généralisation des projets de séquençage à haut débit a entrainé une croissance exponentielle du nombre de séquences disponibles au sein des banques protéiques. Par conséquent, il est fréquent que, lors de la première étape de la construction des alignements, les recherches de similarité mettent en évidence des centaines, voire des milliers, de séquences à partir d'une protéine d'intérêt.

Deux éléments majeurs relatifs à cette quantité de séquences peuvent nuire à la significativité des résultats :

- la forte hétérogénéité dans la qualité des séquences disponibles. Elle est non seulement liée à la profondeur du séquençage des génomes (c'est-à-dire au nombre moyen de lectures utilisées pour constituer la séquence consensus) et aux erreurs qui en découlent, mais aussi aux limites des programmes de prédiction de gènes et de séquences protéiques, qui ajoutent des erreurs à chaque étape de la caractérisation d'une séquence,
- l'existence, selon les banques, de fortes redondances entre les séquences et/ou la présence de séquences partielles. En effet, certaines banques conservent, dans un souci d'exhaustivité, des séquences identiques ou quasi-identiques, issues par exemple de programmes de séquençage différents. De plus, on observe une redondance naturelle entre les séquences issues d'organismes proches ou au sein des mêmes organismes du fait de la duplication de certains gènes au cours de l'évolution.

L'abondance de séquences partielles ou de qualité douteuse représente une source de bruit non négligeable. Par ailleurs, si l'on considère que l'utilisateur aligne un nombre prédéfini de séquences (qu'il aura défini lui-même ou en fonction des limites du programme d'alignement), la redondance, ou la présence de séquences proches, peut empêcher l'accès à des informations pertinentes inscrites dans des séquences plus éloignées qui ne seront pas prises en compte.

Par conséquent, l'alignement automatique de toutes les séquences homologues disponibles n'est dorénavant plus pertinent. Un challenge majeur réside donc dans la mise en place de stratégies de construction d'alignements multiples dédiés, dont le contenu informationnel est en adéquation avec la problématique biologique traitée. L'ensemble des études ne demande pas de considérer les mêmes séquences (en termes de nombre, de diversité, etc.), ni de travailler avec un même degré de qualité des alignements.

Nous nous sommes, dans ce contexte, focalisés sur la première étape de la construction des alignements multiples, à savoir la sélection des séquences à aligner pour « simplifier » l'alignement, tout en conservant le contenu informationnel nécessaire et suffisant pour mener à bien une analyse précise.

Deux possibilités s'offrent au biologiste : une réduction *a priori* de l'espace de recherche des séquences, ou une sélection *a posteriori* des séquences détectées par la recherche de similarité.

10.3.1. Réduction *a priori* de l'espace de recherche des séquences

Selon le but de l'analyse, une réduction *a priori* de l'espace de recherche des séquences peut parfois être effectuée. Cette notion de réduction *a priori* peut être illustrée par quelques exemples :

- dans le cadre d'études phylogénétiques, il peut être intéressant de restreindre ses recherches aux séquences de certains clades ou organismes. De nombreux serveurs Blast en ligne, comme celui du NCBI (http://www.ncbi.nlm.nih.gov/BLAST/), proposent la restriction des recherches à des organismes spécifiques, comme par exemple les mammifères ou les vertébrés,
- lors de l'annotation de séquences inconnues, il est souvent préférable d'utiliser une banque non-exhaustive, mais contenant des annotations de qualité, comme Swiss-Prot, plutôt que de risquer d'inférer et de propager des annotations erronées en se fiant à des banques plus exhaustives en termes de séquences, mais de mauvaise qualité,
- pour les études structurales ou la construction de modèles par homologie, les séquences correspondant aux structures présentes dans la PDB sont souvent choisies comme références,
- en génomique comparative, la création de profils phylogénétiques n'a d'intérêt que si l'on dispose des génomes ou des protéomes entiers des organismes, ce qui permet une restriction évidente de l'espace de recherche des séquences,
- l'utilisation de banques prétraitées et non redondantes, comme les banques de la série UniRef, permet d'augmenter automatiquement la diversité au sein des alignements multiples tout en réduisant considérablement l'espace de recherche.

Ces quelques exemples sont loin de recouvrir l'ensemble des différentes réductions *a priori* possibles. Cependant, on peut remarquer qu'au-delà de l'adéquation entre l'information réduite et les objectifs de l'étude envisagée, pour s'appliquer à des analyses à haut débit, ces méthodes impliquent souvent une connaissance des banques, de leurs structures, des informations disponibles ainsi que des outils bioinformatiques nécessaires pour créer des ensembles réduits.

10.3.2. Réduction *a posteriori* des séquences détectées

Il n'est toutefois pas toujours envisageable de restreindre l'espace de recherche des séquences *a priori*. Le jeu de séquences détectées peut alors être réduit *a posteriori*.

BLAST Filter (Spalding and Lammers, 2004) et UniqueProt (Mika and Rost, 2003) ont été développés dans cette optique. BLAST Filter permet de post-traiter les résultats issus de

recherches Blast et de générer un jeu réduit de séquences, en se basant sur des filtres configurables par l'utilisateur. UniqueProt permet, quant à lui, de créer un jeu de séquences représentatif et non-biaisé, en retirant les séquences surreprésentées.

Ces deux programmes sont basés sur des données de similarité entre les séquences détectées mais nécessitent, pour être efficaces et puissants, une pré-analyse des résultats. En effet, suivant leur fonction, les gènes sont soumis à des pressions de sélection variables au cours de l'évolution, se traduisant par une large diversité de conservation entre protéines homologues : il n'est par conséquent pas possible de se fier à un pourcentage d'identité fixe pour filtrer efficacement les séquences.

10.4. Conclusions et perspectives

La construction et l'exploitation des alignements multiples de séquences protéiques ont pris une nouvelle dimension à l'ère post-génomique, grâce à la multiplicité des rôles qu'ils sont à même d'assurer, mais également de par le besoin émergeant de mise en place de stratégies de construction d'alignements adaptées à une problématique biologique précise.

Ces stratégies doivent permettre une simplification de la construction et de l'analyse des alignements, tout en préservant le contenu informationnel nécessaire à l'aboutissement des études envisagées.

Si la prise de conscience relative à ces problématiques n'est que très récente, il nous paraît évident que la mise en place de protocoles dédiés, qui intègrent une sélection pertinente des séquences en amont de la construction des alignements, deviendra incontournable pour l'exploitation efficace de ces derniers et constituera un challenge majeur dans les années à venir.

Chapitre 11 - Alignements multiples dédiés à l'analyse structurale

L'alignement multiple, en plaçant une protéine au cœur de sa famille, constitue un support idéal pour l'interprétation des conséquences des mutations qui l'affectent.

Dans le cadre du projet MS2PH, nous nous intéressons particulièrement aux conséquences fonctionnelles et structurales des mutations sur une protéine. De ce fait, nous avons souhaité mettre en place une stratégie de construction et d'exploitaion des alignements multiples, applicable à haut débit, permettant de conserver l'information structurale pertinente.

Nous présenterons dans ce chapitre une étude relative à des méthodes d'échantillonnage des résultats de Blast dans le cadre de l'analyse fonctionnelle et structurale d'une protéine, pour nous intéresser ensuite à la conservation des résidus au sein des alignements.

11.1. Echantillonnage des résultats de recherches Blast en vue de la construction d'alignements multiples dédiés à l'analyse structurale

Nous avons développé et testé plusieurs méthodes d'échantillonnage des résultats issus de recherches de similarité Blast (*voir* Publication n°1 ci-dessous). Les méthodes d'échantillonnage sont présentées en détails dans le Chapitre 7 - *Développements relatifs aux alignements multiples*, p. 104.

Trois des quatre méthodes testées, les méthodes de la moyenne (*mm*), de la dérivée seconde (*sdm*) et des bandelettes (*sm*), sélectionnent des séquences représentatives de la dispersion des *E-value*. Ces méthodes reposent sur l'hypothèse que, aux niveaux structural et fonctionnel, les séquences les plus proches ne sont pas celles qui apportent le plus d'informations pertinentes et que la diversité est généralement plus informative (Przybylski and Rost, 2002). Dans le cadre des recherches de similarité Blast, cela se traduit par le fait que les séquences détectées avec des *E-value* proches, et ce plus particulièrement dans le cadre des *E-value* très faibles, sont potentiellement très similaires. Inversement, des séquences présentant des *E-value* plus disparats représentent souvent des séquences divergentes.

La quatrième méthode, *rm*, sélectionne les séquences de manière aléatoire et a servi de témoin dans le cadre de notre étude.

Pour pouvoir être considérée comme efficace, une méthode d'échantillonnage doit être à même de réduire considérablement les jeux de séquences à aligner tout en préservant l'information structurale et fonctionnelle pertinente en leur sein. Aucune méthode générique n'étant disponible pour mesurer de manière objective l'information structurale dans son ensemble, nous avons choisi, dans le cadre de notre étude, de nous concentrer exclusivement sur les résidus annotés dans la PDB comme étant impliqués dans des sites actifs. Ces résidus sont généralement conservés au sein des familles protéiques et bien caractérisés sur un plan biochimique et structural.

Le contenu informationnel des alignements générés a ainsi été estimé en analysant le comportement, en termes de conservation, des résidus impliqués dans des sites actifs.

Pour cela, nous avons mené notre étude sur un jeu de 284 protéines extraites de la PDB, dont les séquences partagent moins de 70% d'identité entre elles et contiennent au moins un site actif annoté dans la PDB.

L'analyse du comportement des quatre méthodes d'échantillonnage des résultats des recherches Blast nous a permis de les caractériser en termes de taux de réduction et de séquences sélectionnées.

L'étude de l'impact des échantillonnages sur le potentiel informationnel des jeux de séquences sélectionnées, par l'intermédiaire de MACS, nous a permis de montrer qu'il est possible de réduire considérablement le nombre de séquences à aligner (70% en moyenne) tout en conservant l'information pertinente, relative aux sites actifs des protéines, au sein de l'alignement multiple associé.

Une stratégie optimale d'échantillonnage des séquences selon les résultats du Blast a finalement été proposée. L'application de cette stratégie permet un gain de temps au niveau de la construction des alignements et une facilité de traitement et d'interprétation des résultats par le biologiste, tout en conservant l'information structurale pertinente.

Publication n° 1

[*Signalement bibliographique ajouté par* :
SICD Strasbourg - Département de la Documentation électronique
Service des thèses électroniques]

Blast sampling for structural and functional analyses

Anne FRIEDRICH, Raymond RIPP, Nicolas GARNIER, Emmanuel BETTLER, Gilbert DELÉAGE, Olivier POCH and Luc MOULINIER

BMC Bioinformatics 2007, Vol. 8, Pages 62-78

Pages 149 - 165 :

La publication présentée ici dans la thèse est soumise à des droits détenus par un éditeur commercial.

Pour les utilisateurs ULP, il est possible de consulter cette publication sur le site de l'éditeur :

http:/dx.doi.org/10.1186/1471-2105-8-62

La version imprimée de cette thèse peut être consultée à la bibliothèque ou dans un autre établissement via une demande de prêt entre bibliothèques (PEB) auprès de nos services :

http://www-sicd.u-strasbg.fr/services/peb/

11.2. Etude de la conservation des résidus au sein des alignements multiples

L'importance d'un résidu au sein d'une protéine, pour le maintien de sa structure ou dans le cadre de sa fonction, peut généralement être évaluée par sa conservation dans un alignement multiple regroupant les séquences de sa famille protéique.

Bien que la notion de conservation d'un résidu soit tout à fait commune, le problème de son calcul est loin d'être trivial. En effet, la conservation est plus souvent décrite, par les biologistes, de manière qualitative que par des critères quantitatifs. Par conséquent, son analyse se fait généralement par l'intermédiaire d'une inspection manuelle ou semi-automatique d'un alignement particulier.

A l'heure de la généralisation des études massives, il devient crucial de disposer de méthodes d'estimation de la conservation des résidus fiables et applicables quelque soit l'alignement considéré. Ces méthodes devraient permettre l'identification et la classification de résidus en fonction de leur niveau et de leur type de conservation. Malgré le développement de nombreuses méthodes de quantification de cette conservation au cours des 20 dernières années, aucune n'a pour le moment été admise communément comme un standard (revue dans (Valdar, 2002)).

Au sein du laboratoire, nous utilisons de manière courante un logiciel interne, appelé OrdAlie (*voir* section 6.3.8, p. 99), pour l'analyse interactive des alignements multiples. L'étude de la conservation au sein des alignements y a historiquement été basée sur l'utilisation d'un seuil de conservation prédéterminé. Cette méthode présente cependant des limites, un seuil adapté étant théoriquement dépendant du contexte de l'alignement. En effet, d'une part, une famille de protéines peut avoir évolué plus ou moins rapidement au cours du temps : la conservation entre les séquences homologues varie donc d'une famille protéique à l'autre. D'autre part, la conservation des résidus est relative au jeu de séquences alignées par le biologiste et ne pourra pas être interprétée sur les bases d'un même seuil si l'alignement regroupe uniquement des séquences issues de primates ou s'il contient des séquences issues de l'ensemble des eucaryotes.

11.2.1. Détermination automatique de classes de conservation

Afin de s'abstenir de la définition d'un seuil lors des études de conservation, trois méthodes distinctes de calcul de la conservation au sein des colonnes d'un alignement ont été intégrées à OrdAlie : la méthode dite de la distance moyenne, la méthode dite de la norme du vecteur moyen et la méthode dite du deltaG (*voir* Chapitre 7 - *Développements relatifs aux alignements multiples*, p. 108).

La détermination des classes de conservation se fait en trois étapes (Figure 60) : (i) calcul de la conservation de chaque colonne de l'alignement; (ii) classification des scores de conservation à l'aide d'une méthode automatique pour la création de groupes de scores ; (iii) tri des groupes de scores. Ceux qui présentent les valeurs les plus élevées sont représentatifs des colonnes les mieux conservées au sein de l'alignement.

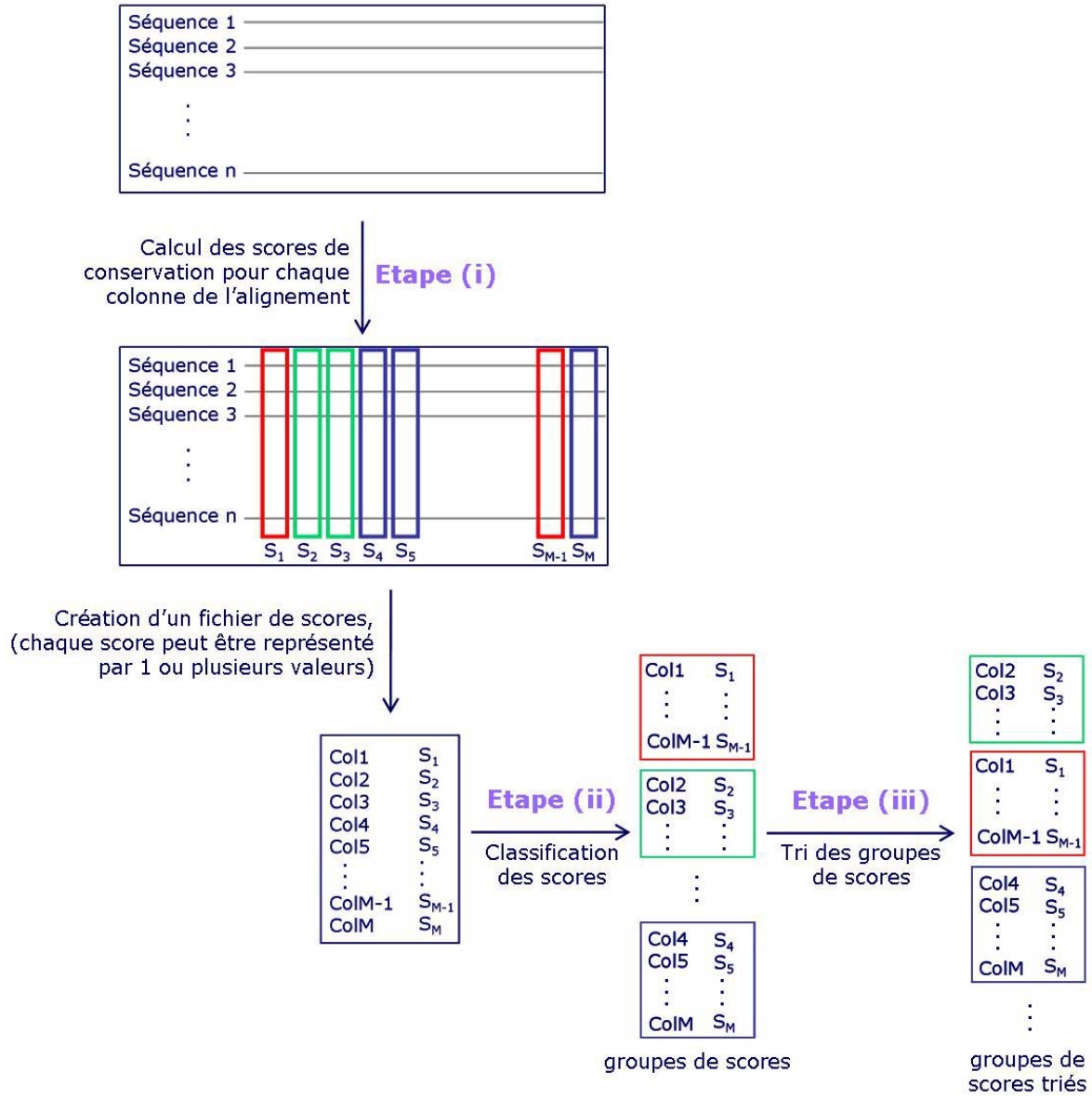


Figure 60 - Principales étapes des méthodes de typification des colonnes de conservation. Sur cet exemple, les colonnes les mieux conservées sont les colonnes vertes, puis les rouges, puis les bleues etc.

Au total, 14 variantes de ces méthodes sont disponibles :

- la conservation au sein de chaque colonne peut être calculée de sept manières différentes (étape (i)) : à partir de chacune des trois méthodes de calcul de score de

conservation de manière indépendante ainsi qu'à partir d'une combinaison des méthodes deux à deux et enfin d'une combinaison des trois méthodes,

- pour chacune des méthodes ou combinaison de méthodes testées, la classification des scores de conservation (étape (ii)) peut être réalisée par deux moyens : à l'aide de DPC (Wicker *et al.*, 2002) ou d'un programme reposant sur un algorithme de modèle de mélange de lois de Dirichlet (Sjolander *et al.*, 1996), auquel nous nous réfèrerons sous les termes de « modèle de mélange ».

11.2.2. Etude comparative des méthodes intégrées à OrdAlie

Nous avons profité des données et des résultats dont nous disposions suite à l'étude présentée dans la Publication n°1 pour tester l'efficacité des méthodes nouvellement implémentées.

Dans cette optique, nous avons considéré les 192 alignements référencés en tant que MACS_init dans la Publication n°1, section *MACS information content*.

Rappelons que ces 192 alignements ont été construits de manière indépendante, sans appliquer d'échantillonnage, en partant, pour chacun, d'une séquence protéique extraite de la PDB, contenant au moins un site actif annoté. L'ensemble des 192 séquences d'intérêt renferme au total 731 sites actifs annotés.

Nous disposons pour ces 192 alignements des résultats des études de sensibilité et de spécificité de détection des sites actifs en appliquant la méthode du seuil, défini à 80%.

La détection des sites actifs parmi les colonnes de meilleure conservation définies par chacune des 14 méthodes décrites dans la section 11.2.1 a été analysée. Nous nous sommes basés sur les mêmes paramètres statistiques que ceux présentés dans la publication, à savoir la sensibilité de la détection, sa spécificité et la valeur de *Gmean* associée.

Nous effectuerons un rappel du calcul de ces paramètres dans la section suivante.

11.2.2.1. Rappels : paramètres statistiques

Dans le cadre de notre étude, la sensibilité (Se) représente la proportion de sites actifs localisés dans des colonnes conservées parmi l'ensemble des sites actifs recensés (731). La spécificité (Sp) est représentative de la proportion de résidus non actifs qui ne sont pas situés dans des colonnes conservées, parmi l'ensemble des résidus non actifs.

Les formules qui définissent la sensibilité et la spécificité sont les suivantes :

$$Se = \frac{VP}{VP + FN} \qquad Sp = \frac{VN}{VN + FP}$$

où VP (Vrais Positifs) représente le nombre de sites actifs situés dans des colonnes de conservation ; FP (Faux Positifs) représente le nombre de résidus non actifs situés dans des colonnes de conservation ; VN (Vrais Négatifs) représente le nombre de résidus non actifs situés dans des colonnes non conservées ; FN (Faux Négatifs) représente le nombre de sites actifs situés dans des colonnes non conservées.

Comme le nombre de sites actifs ne représente qu'une faible minorité de l'ensemble des colonnes des MACS, nous travaillons avec des classes déséquilibrées. Le paramètre le mieux adapté dans ce cas pour comparer les résultats de sensibilité et de spécificité obtenus par différentes méthodes est la moyenne géométrique de la précision, ou *geometric means of accuracy* (*Gmean*), définie de la manière suivante :

$$Gmean = \sqrt{Se \times Sp}$$

11.2.2.2. Résultats de l'analyse

Dans le cadre des résultats présentés dans la Publication n°1, où la détermination de nos colonnes de conservation est basée sur la méthode des seuils, les meilleurs résultats ont été observés en appliquant un seuil de 80%. Les valeurs de sensibilité, de spécificité et de *Gmean* sont respectivement de 0,82, 0,83 et 0,83.

Les valeurs de ces mêmes paramètres, pour chacune des 14 méthodes testées, sont regroupées dans le Tableau 8. Elles ont été obtenues en considérant comme « colonnes conservées » celles contenues dans les deux meilleurs groupes de conservation.

Nous avons constaté que le fait de ne tenir compte que des colonnes du meilleur groupe de conservation est trop restrictif (données non présentées) : de manière générale, les valeurs de spécificité sont très élevées (entre 98 et 100%), traduisant que la plupart des colonnes de conservation contiennent des sites actifs, mais les valeurs de sensibilité sont relativement faibles (entre 35 et 70%), signifiant que seule une faible proportion des sites actifs sont détectés.

De manière générale, les résultats obtenus en utilisant DPC pour la classification des colonnes de conservation présentent une sensibilité de détection des sites actifs beaucoup plus faible que ceux obtenus avec le modèle de mélange. Les spécificités sont quant à elles légèrement supérieures. Ceci peut s'expliquer par la définition d'un plus grand nombre de groupes de conservation par DPC, chacun des groupes contenant moins de colonnes de conservation. Considérer deux groupes de conservation est donc trop strict pour permettre

l'application de DPC : les résidus détectés sont bien majoritairement des sites actifs, mais seule une proportion relativement faible (48 à 70%) est mise en évidence.

		N	D	dG	N-D	N-dG	D-dG	N-D-dG
Modèle de mélange	**Se**	0,86	0,76	0,81	0,79	**0,85**	0,78	0,80
	Sp	0,91	0,97	0,96	0,96	**0,95**	0,98	0,97
	Gmean	0,89	0,86	0,88	0,87	**0,90**	0,87	0,88
DPC	**Se**	0,48	0,64	0,55	0,70	0,64	0,66	0,65
	Sp	0,99	0,98	0,99	0,98	0,98	0,99	0,98
	Gmean	0,69	0,79	0,74	0,83	0,80	0,80	0,80

Tableau 8 - Sensibilité, spécificité et *Gmean* associés à chacune des 14 méthodes de caractérisation des conservations au sein des alignements multiples. N représente la méthode de la norme du vecteur moyen, D la méthode de la distance moyenne et dG la méthode du deltaG. En rouge, les valeurs associées à la méthode la plus performante.

Les résultats obtenus avec le modèle de mélange sont bien meilleurs et permettent de combiner valeurs élevées de sensibilité et de spécificité.

Les méthodes qui intègrent le calcul de la conservation par la distance moyenne, seule ou en combinaison avec une des deux autres méthodes, génèrent les résultats de sensibilité les plus faibles, inférieurs à 80%.

La méthode de détection qui associe le calcul des scores de conservation par la norme des vecteurs moyens et par le deltaG est celle qui produit les meilleurs résultats (valeurs en rouge dans le tableau), présentant un *Gmean* de 90%. La combinaison de ces deux méthodes permet de cumuler les avantages des méthodes considérées indépendamment : la méthode des normes des vecteurs moyens présente de très bons résultats de sensibilité (86%) au détriment de la spécificité (91%), alors que la méthode des deltaG présente de très bons résultats de spécificité (96%) au détriment de la sensibilité (81%). Combiner les deux méthodes permet d'augmenter la sensibilité obtenue avec la méthode du deltaG seule (de 81 à 85%), sans trop affecter sa spécificité (de 96 à 95%).

11.2.2.3. Conclusions et perspectives

Les résultats obtenus au cours de cette étude sont très encourageants et indiquent qu'il est envisageable de mettre en place des stratégies de typification de la conservation des colonnes d'un alignement multiple en s'abstenant de la détermination de seuils.

En effet, la comparaison des *Gmean* montre que l'ensemble des méthodes étudiées, basées sur le modèle de mélange pour la classification des colonnes de conservation, est plus efficace

que la méthode des seuils (entre 86 et 90% contre 83%). Ce phénomène peut s'expliquer par le fait que les groupes de conservation définis soient automatiquement adaptés au contexte de l'alignement. Ces résultats confirment, de plus, que la méthode des seuils ne devrait être employée qu'en complément d'une inspection visuelle des alignements, qu'il n'est pas possible d'appliquer dans le cadre d'études à haut débit.

Ces résultats préliminaires sont voués à être exploités dans le cadre d'une étude plus exhaustive qui sera menée au sein du laboratoire. Celle-ci sera dédiée à la mise en place de stratégies de détection de résidus particuliers (ne se limitant plus aux résidus actifs), basées sur la typification automatique des colonnes de conservation dans l'ensemble d'un alignement et/ou par sous-famille.

Chapitre 12 - Caractérisation fonctionnelle et structurale des mutations

Mener des études relatives à l'impact structural d'une mutation nécessite que celle-ci soit, dans un premier temps, repositionnée dans la structure tridimensionnelle de la protéine affectée. Pour cela, il nous a paru essentiel de disposer d'un espace de travail qui réunit les informations de séquence, de structure, de fonction et d'évolution relatives à une protéine d'intérêt.

A notre connaissance, aucun logiciel ou serveur Web ne permettait de disposer d'un tel environnement à la fin de l'année 2005, c'est pourquoi nous avons développé MAGOS, pour *Multiple AliGnment and mOdelling Server* (*voir* Publication n°2 ci-dessous), qui se présente sous la forme d'un serveur Web.

A partir d'une séquence protéique, MAGOS permet de construire automatiquement un alignement multiple de séquences complètes au sein duquel les séquences sont distribuées en sous-familles. Cet alignement est annoté avec des données fonctionnelles et structurales, récupérées au sein de banques de données publiques, ou prédites *ab initio*. De plus, un modèle tridimensionnel de la protéine est construit par homologie si la structure d'une protéine de la même famille a été résolue expérimentalement (*voir* Chapitre 8 - *MAGOS*, p. 113).

L'atout majeur de MAGOS réside dans les possibilités d'analyse intégrative offertes à l'utilisateur par l'interface graphique intuitive de visualisation des résultats. Cette dernière permet l'interconnexion de la séquence protéique au sein de l'alignement annoté et de son modèle tridimensionnel, ainsi que la représentation des informations fonctionnelles, comme par exemple la présence de sites actifs ou de liaison, au sein de l'alignement et du modèle tridimensionnel.

Nous illustrerons les capacités d'intégration de MAGOS en nous basant sur un exemple, avant de fournir quelques données statistiques relatives à l'utilisation du serveur.

Publication n° 2

[*Signalement bibliographique ajouté par* :
SICD Strasbourg - Département de la Documentation électronique
Service des thèses électroniques]

MAGOS: multiple alignment and modelling server

N. GARNIER, A. FRIEDRICH, R. BOLZE , E. BETTLER, L. MOULINIER, C. GEOURJON, J. D. THOMPSON, G. DELÉAGE and O. POCH

Bioinformatics Applications note, 2006, Vol. 22, N° 17, Pages 2164–2165

Pages 177-178 :

La publication présentée ici dans la thèse est soumise à des droits détenus par un éditeur commercial.

Pour les utilisateurs ULP, il est possible de consulter cette publication sur le site de l'éditeur :

http://dx.doi.org/10.1093/bioinformatics/btl349

La version imprimée de cette thèse peut être consultée à la bibliothèque ou dans un autre établissement via une demande de prêt entre bibliothèques (PEB) auprès de nos services :

http://www-sicd.u-strasbg.fr/services/peb/

12.1. Illustration des capacités d'intégration de MAGOS

Les capacités d'intégration de MAGOS peuvent être illustrées par l'étude de mutations ponctuelles affectant une protéine impliquée dans une maladie génétique humaine et de leurs positions relatives par rapport aux caractéristiques fonctionnelles et structurales de la protéine.

Nous présenterons dans cette section l'analyse de mutations affectant la myotubularine, protéine impliquée dans la myopathie myotubulaire liée à l'X, en nous focalisant sur ce qui est aisément réalisable par l'intermédiaire du seul serveur Web MAGOS. Nour réaliserons une étude plus complète de données de mutations dans le Chapitre 14 - *Vers une typification des mutations*, p. 203, où nous exploiterons plus profondément les méthodes de sélection des séquences pour la construction d'alignements multiples dédiés, la conservation des résidus ainsi que des données structurales supplémentaires.

12.1.1. Myotubularine

La myotubularine est une phosphatase lipidique à phosphoinositides codée par le gène MTM1, situé sur le chromosome X (Laporte *et al.*, 1996). Cette enzyme, potentiellement impliquée dans le trafic membranaire, est essentielle au maintien du muscle (Tronchere *et al.*, 2004).

La myotubularine est une protéine de 603 résidus de long, constituée de quatre domaines (d'après Swiss-Prot et (Laporte *et al.*, 2003)) : le domaine PH-GRAM, connu pour lier les phosphoinositides et rencontré dans de nombreuses protéines essentielles à la signalisation cellulaire, à l'organisation du cytosquelette et au trafic vésiculaire ; le domaine « *Myotubularin-related* », caractéristique des membres de la famille des myotubularines (au nombre de 14 chez l'homme) ; le domaine tyrosine phosphatase ou PTP pour *Protein-Tyrosin Phosphatase*, au cœur duquel est située la poche du site actif ; le domaine SID, pour *SET (Su(var), Enhancer of zeste, Trithorax) Interacting Domain*, médiateur de l'interaction de MTM1 avec MTMR12 (*myotubularin-related protein 12*) (Figure 61).

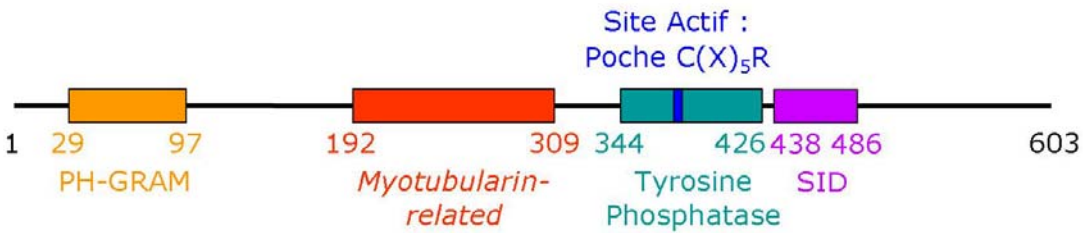


Figure 61 - Représentation schématique de la constitution en domaines de la myotubularine.

12.1.2. Myopathie myotubulaire liée à l'X

La myopathie myotubulaire liée à l'X (OMIM 310400) est une maladie musculaire congénitale récessive caractérisée, de manière générale, par une hypotonie (ou diminution pathologique de la tonicité musculaire) néonatale marquée, une faiblesse généralisée des muscles et une détresse respiratoire chez les garçons affectés.

La myopathie myotubulaire liée à l'X est due à un défaut dans le gène MTM1, qui conduit à une perte totale ou partielle de l'activité de la myotubularine et entraîne un arrêt précoce de la maturation de la fibre musculaire.

Bien que les manifestations cliniques de cette maladie soient relativement homogènes, plusieurs degrés de sévérité peuvent être distingués (Herman *et al.*, 1999) :

- le phénotype sévère se traduit par une insuffisance respiratoire chronique et l'absence de développement des capacités motrices ou par le décès du patient dans la petite enfance,
- le phénotype intermédiaire rend nécessaire une assistance respiratoire de longue durée et entraîne un retard important dans la réalisation des étapes du développement moteur,
- le phénotype modéré est quant à lui caractérisé par la nécessité d'une assistance respiratoire de courte durée après la période néonatale et un retard minime dans la réalisation des étapes du développement moteur.

Les formes sévères de la maladie, qui entraînent généralement la mort durant la première année de vie, sont les plus courantes.

Biancalana et ses collaborateurs (Biancalana *et al.*, 2003) ont rapporté que sur 328 mutations décrites dans le gène MTM1, on compte environ 30% de mutations faux-sens, 20% de mutations non-sens, 20% de petites insertions ou délétions, 20% d'altération de site d'épissage et 7% de grandes délétions.

La quasi-totalité des mutations non-sens, ainsi que celles à l'origine de défauts d'épissage, sont associées à des phénotypes sévères (McEntagart *et al.*, 2002).

12.1.3. Analyse de mutations affectant la myotubularine par MAGOS

12.1.3.1. Requête MAGOS

Nous avons soumis la séquence protéique de la myotubularine (Swiss-Prot : MTM1_HUMAN) à MAGOS. La recherche d'homologie a été réalisée dans la banque UniRef90, pour augmenter la diversité des séquences au sein de l'alignement. Ce dernier a été réalisé à partir des 100 premières séquences détectées par Blast.

Le score de qualité NorMD associé à l'alignement construit est de 0,65, ce qui correspond à un alignement de bonne qualité.

Lors du lancement de notre requête, nous n'avons pas précisé d'identifiant PDB pour la construction du modèle tridimensionnel par homologie. Plusieurs structures de protéines homologues à la myotubularine humaine ont été résolues expérimentalement et sont disponibles dans la PDB. MAGOS a sélectionné de manière automatique la structure de la *myotubularin-related protein 2* (MTMR2) humaine (PDB : 1ZVR) comme empreinte pour la construction du modèle tridimensionnel par homologie.

Cette empreinte a permis la construction d'un modèle tridimensionnel couvrant 85% de la longueur de la myotubularine (des résidus 32 à 544, sur les 603 totaux) (Figure 62). La myotubularine et la MTMR2 humaines partagent 69% d'identité, ce qui laisse penser que le modèle tridimensionnel construit est de bonne qualité.

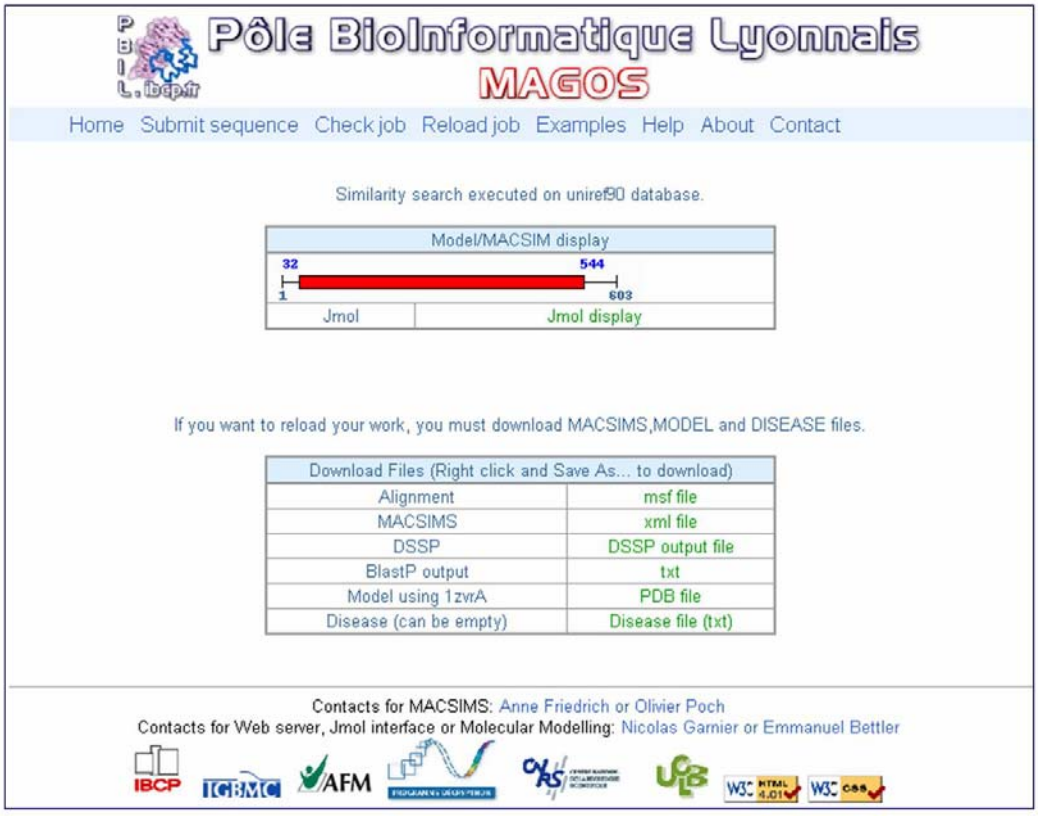


Figure 62 - Page d'accueil des résultats obtenus par MAGOS pour la myotubularine.

La Figure 63 représente l'interface de visualisation des résultats de MAGOS, dans laquelle la séquence de la myotubularine au sein de l'alignement multiple annoté est connectée avec son modèle tridimensionnel. On remarque que les domaines PH-GRAM, *myotubularin-related* et tyrosine phosphatase ont été correctement positionnés sur la séquence au sein de l'alignement et sont tous trois visualisables de manière simultanée sur le modèle

tridimensionnel. Aucune information concernant le domaine SID n'a par contre été récupérée. Ceci peut s'expliquer par le fait que la banque de données Swiss-Prot ne fournit pas d'information au sujet de ce dernier.

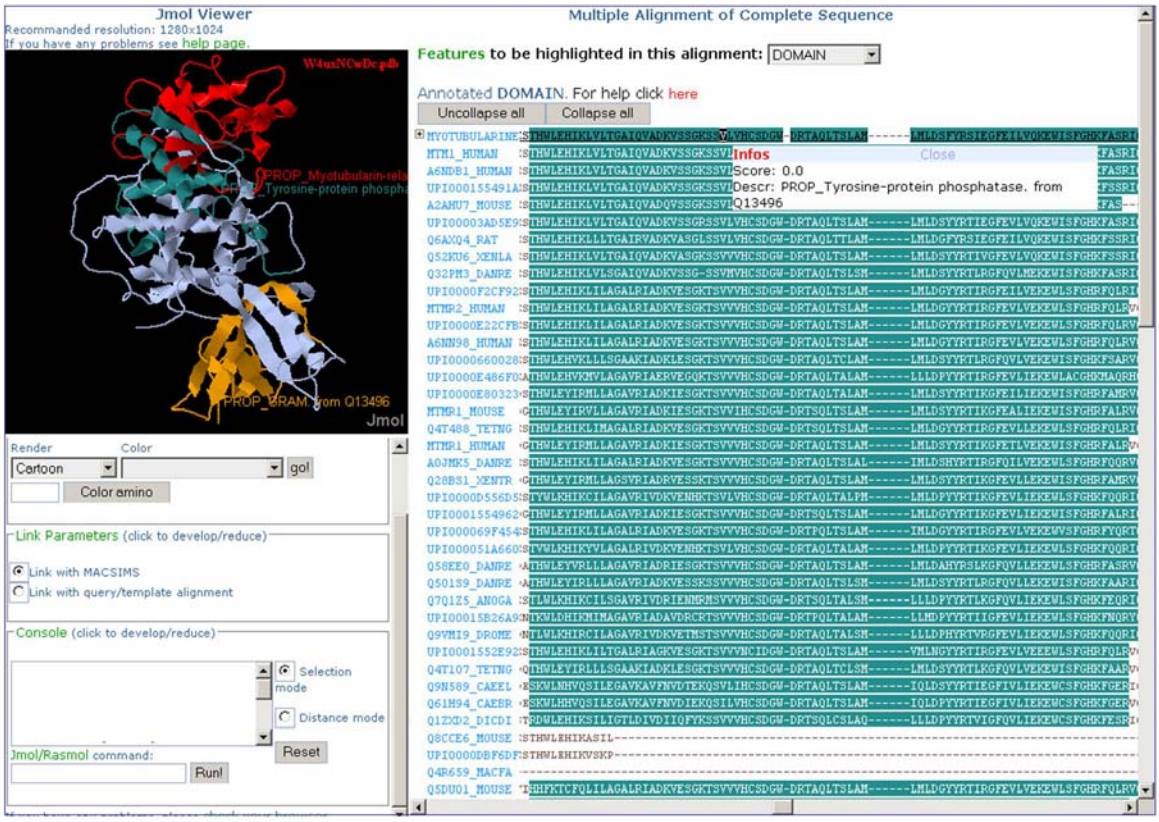


Figure 63 - Interconnexion de la séquence de la myotubularine dans l'alignement et son modèle tridimensionnel dans l'interface Jmol. Dans l'onglet de visualisation du modèle tridimensionnel, les couleurs des domaines annotés par MAGOS ont été modifiées à l'aide de lignes de commande Jmol afin de correspondre aux couleurs de la Figure 61.

Le site actif localisé en position 375 de la myotubularine est également annoté, mais la poche catalytique, formée par les résidus 375 à 381, n'est pas renseignée en tant que telle.

Notons que les informations connues par le biologiste et non retrouvées de manière automatique par MAGOS peuvent être positionnées de manière relativement aisée sur le modèle tridimensionnel à l'aide de lignes de commande Jmol.

12.1.3.2. Données de mutation relatives à la myotubularine

Nous avons récupéré des informations relatives aux mutations faux-sens affectant la myotubularine et leurs conséquences phénotypiques dans la banque Swiss-Prot (http://www.expasy.org/uniprot/Q13496). Au mois d'août 2007, 56 mutations faux-sens y étaient référencées.

Nous avons réduit notre analyse aux 27 mutations faux-sens auxquelles des phénotypes cliniques caractérisés par leur sévérité (sévère/intermédiaire/modéré) sont associés (Tableau 9).

12.1.3.3. Analyse des mutations dans l'environnement MAGOS

Mutation	Domaine	Elément de structure secondaire	Phénotype associé
p.R69C	PH-GRAM	feuillet β	modéré
p.R69S	PH-GRAM	feuillet β	sévère
p.L70F	PH-GRAM	feuillet β	modéré
p.L87P	PH-GRAM		modéré
p.P179S			modéré
p.N180K			très modéré
p.R184G		feuillet β	sévère
p.Y198N	*Myotubularin-related*		sévère
p.L202S	*Myotubularin-related*	feuillet β	sévère
p.P205L	*Myotubularin-related*	feuillet β	sévère
p.I225T	*Myotubularin-related*		modéré
p.S229P	*Myotubularin-related*	feuillet β	modéré
p.R241C	*Myotubularin-related*	feuillet β	modéré à intermédiaire
p.I264S	*Myotubularin-related*		sévère
p.M317R		hélice α	modéré
p.W346C	Tyrosine phosphatase	hélice α	modéré
p.G378R	Tyrosine phosphatase (poche du site actif)		sévère
p.A389D	Tyrosine phosphatase	hélice α	sévère
p.Y397C	Tyrosine phosphatase	hélice α	sévère
p.G402A	Tyrosine phosphatase	hélice α	modéré
p.E404K	Tyrosine phosphatase	hélice α	modéré
p.L406P	Tyrosine phosphatase	hélice α	sévère
p.R421Q	Tyrosine phosphatase	hélice α	sévère
p.L470P	(SID)	hélice α	sévère
p.N481Y	(SID)		modéré
p.W499R		hélice α	modéré
p.K510N			sévère

Tableau 9 - Récapitulatif des mutations faux-sens affectant la myotubularine, de leurs positions par rapport à des informations collectées par MAGOS et des phénotypes associés. Dans la désignation des mutations, « p. » indique que la description est faite au niveau protéique, la première lettre majuscule représente le résidu de la protéine sauvage, suivi de sa position et du résidu après mutation. Les données entre parenthèses n'ont pas été fournies par MAGOS et proviennent de la littérature.

Une étude intégrative de la position des mutations affectant la myotubularine peut être réalisée dans l'environnement MAGOS. En effet, MAGOS permet de localiser les mutations par rapport à un certain nombre d'éléments connus ou prédits pour la protéine d'intérêt, en particulier lorsqu'un modèle tridimensionnel a pu être construit (Tableau 9).

Les mutations faux-sens référencées dans la banque de données Swiss-Prot sont localisées tout le long de la myotubularine et on en rencontre dans chacun des domaines annotés de la protéine (Figure 64).

Les mutations qui affectent le domaine PH-GRAM sont associées à des phénotypes dits modérés, alors que celles localisées dans le domaine tyrosine phosphatase sont majoritairement associées à des phénotypes sévères.

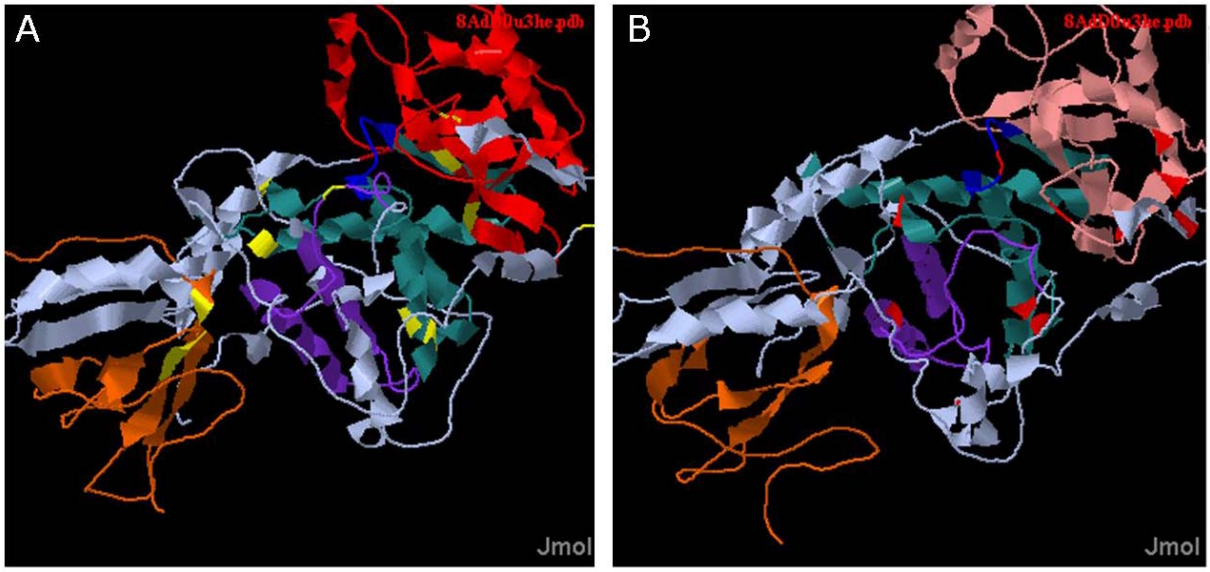


Figure 64 - Localisation des mutations sur le modèle tridimensionnel de la myotubularine. Les couleurs associées aux domaines connus correspondent à celles de la Figure 61 (ou une nuance proche pour permettre la visualisation de la localisation des mutations). (A) localisation des mutations associées à des phénotypes modérés, en jaune ; (B) localisation des mutations associées à des phénotypes sévères, en rouge.

Grâce à la disponibilité du modèle tridimensionnel, il est possible de localiser les mutations par rapport aux éléments de structures secondaires et tertiaires. 18 des 27 mutations se situent dans des éléments de structures secondaires (hélice α ou feuillet β), ces mutations étant susceptibles d'affecter la stabilité de la protéine.

La Figure 65 présente, au sein de l'interface de visualisation des résultats, les mutations localisées à proximité de la poche catalytique. Ces mutations sont visualisables d'une part sur la séquence de la myotubularine dans le contexte de l'alignement et d'autre part sur le modèle tridimensionnel.

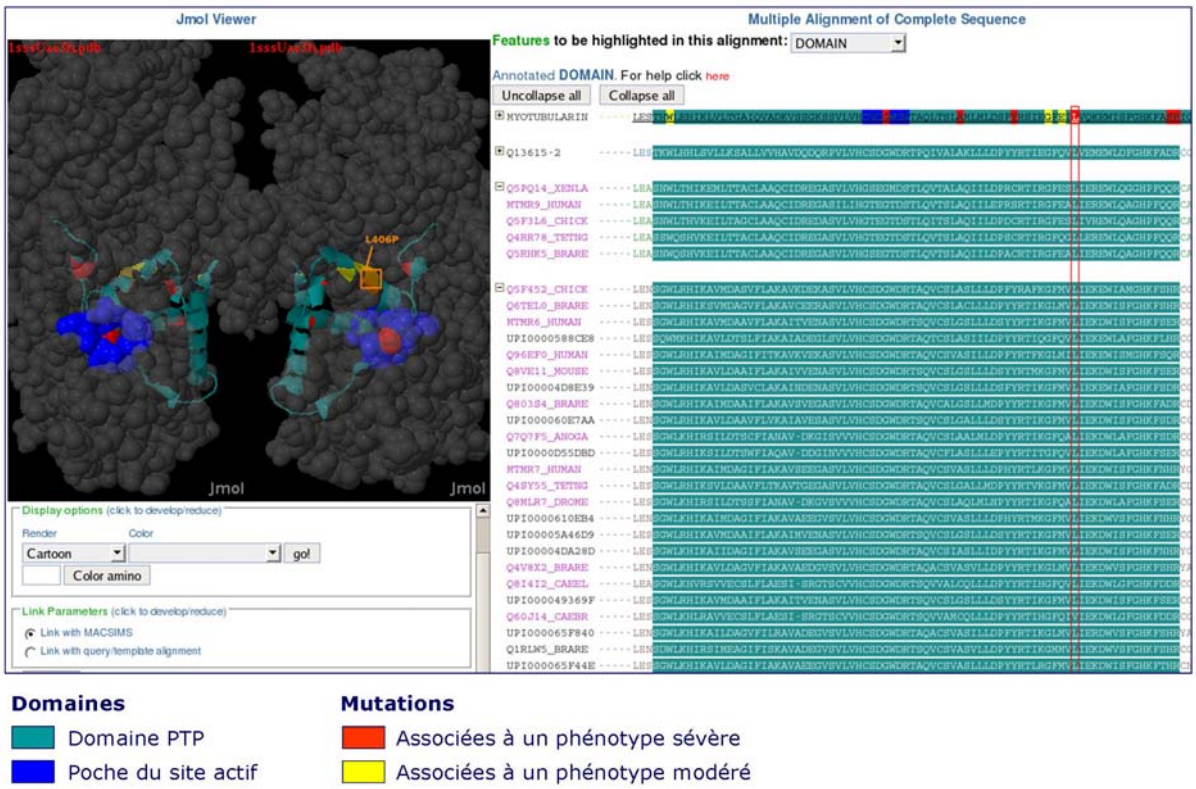


Figure 65 - Représentation des mutations situées à proximité de la poche catalytique sur l'interface de visualisation des résultats de MAGOS.

On remarque que la mutation p.G378R affecte un résidu situé au cœur de la poche du site actif et est à l'origine de l'émergence d'un phénotype sévère.

La mutation p.A389D affecte un résidu situé au sein d'une hélice α de treize résidus de long. Les deux premiers résidus de cette hélice font partie de la poche catalytique. On peut supposer que, dans le cas de cette mutation, la déstabilisation de l'hélice, et donc de la poche du site actif, est liée à l'émergence d'un phénotype sévère.

La mutation p.G402A est localisée dans une hélice α. Cette substitution est très conservative, ce qui peut sans doute expliquer le phénotype modéré associé à cette mutation.

La mutation p.L406P affecte un résidu très conservé dans l'alignement, situé au sein du domaine tyrosine phosphatase, au cœur d'une hélice constituée de dix résidus. La « cassure » de l'hélice α, engendrée par la substitution d'une leucine par une proline, est probablement à l'origine du phénotype sévère associé à cette mutation. Le même type de substitution est observé pour la mutation p.L470P, qui affecte aussi un résidu situé dans une hélice α et à laquelle un phénotype sévère est également associé.

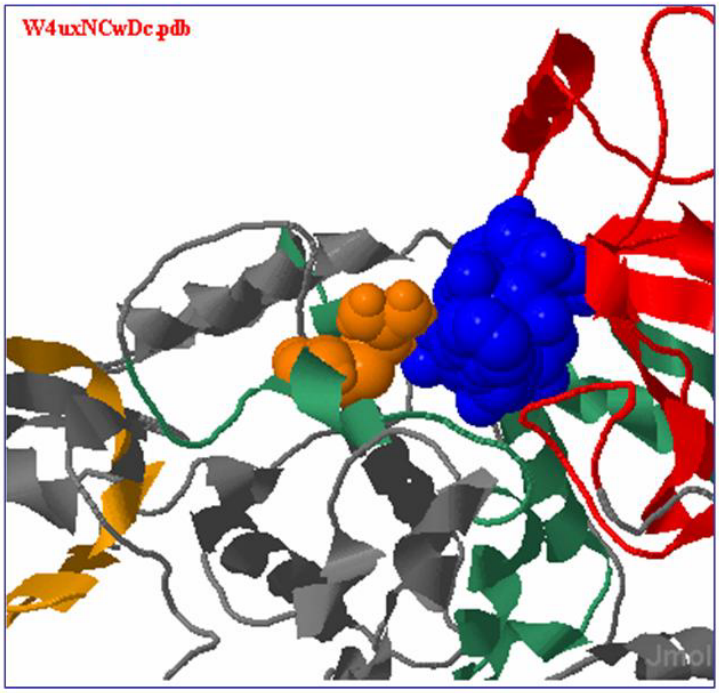


Figure 66 - Visualisation de la localisation de la mutation p.R421Q (capture d'écran de l'onglet Jmol de la page de résultats). Les couleurs des domaines respectent celles observées sur la Figure 61. La poche du site actif est colorée en bleu (représentation atomique). L'arginine 421 est visible en orange, au cœur d'une hélice. Sa chaîne latérale est en contact avec la poche du site actif.

La Figure 66 permet de localiser la mutation p.R421Q dans le modèle tridimensionnel de la myotubularine, au sein de l'onglet de visualisation Jmol. On remarque que l'arginine en position 421 de la myotubularine est située au cœur d'une petite hélice constituée de cinq résidus. La chaîne latérale de l'arginine est en contact avec la poche du site actif, ce qui permet de supposer que cette dernière joue un rôle dans la stabilisation de cette poche. La mutation de l'arginine en glutamine est associée à un phénotype sévère, qui peut potentiellement s'expliquer par la déstabilisation de la poche du site actif.

12.1.3.4. Conclusions et perspectives

L'analyse manuelle des mutations affectant la myotubularine montre que MAGOS permet d'aborder très facilement plusieurs aspects de l'étude de protéines de manière couplée. L'environnement convivial et intuitif de visualisation des résultats permet de poser les premières hypothèses de travail, qui devront être confirmées par des analyses complémentaires plus fines.

MAGOS peut ainsi être défini comme une plateforme d'intégration pour non-experts, permettant de réaliser, de manière conjointe, une analyse de séquence alimentée par des données évolutives et fonctionnelles et une analyse structurale.

12.2. Statistiques d'utilisation du serveur MAGOS

MAGOS a été mis en service en février 2006 et sa publication officielle date du mois de juillet 2006. Depuis cette date, 541 requêtes ont été soumises à MAGOS (au 3 octobre 2007), dont 97 pour lesquelles aucun modèle tridimensionnel n'a pu être construit, du fait de l'absence d'un homologue proche de la séquence d'intérêt. On compte également 22 requêtes qui n'ont pas abouti, principalement à cause de problèmes de communication entre les serveurs.

Le temps d'une requête, qui dépend principalement de la longueur de la séquence soumise et du nombre de séquences à aligner, peut varier d'environ quinze minutes à plus de trois heures.

Pour des questions pratiques et dans l'attente de son implémentation sur la grille de calcul Décrypthon, nous avons limité la génération du modèle par homologie à 600 résidus de long. En effet, la construction d'un modèle plus grand est trop gourmande en termes de ressources et de temps de calcul, ce qui est susceptible de conduire le serveur de l'IBCP de Lyon à saturation.

Précisons finalement que ce serveur Web est voué à évoluer. Nous avons prévu d'y intégrer prochainement différentes méthodes de sélection des séquences pour la construction de l'alignement multiple, comme la méthode d'échantillonnage des bandelettes et les filtres Eucaryote et Métazoaire, ainsi que des options de visualisation supplémentaires au niveau de l'interface Jmol, notamment celles déjà implémentées dans le cadre de MS2PH-db (*voir* Chapitre 9 - *MS2PH-db*, p.130).

Chapitre 13 - Mise en place des bases de l'infrastructure dédiée à l'analyse globale des mutations

L'un des objectifs principaux du projet MS2PH réside dans la mise à disposition de sources d'informations visant à faciliter la compréhension des relations qui existent, au niveau des protéines impliquées dans les maladies génétiques humaines, entre la séquence de la protéine, son évolution, sa structure tridimensionnelle et les pathologies associées.

Grâce au développement de MAGOS, nous disposons d'un environnement de travail permettant une approche interactive relative aux informations structurales dans le cadre de la pertinence évolutive des données de séquences et de fonction.

La volonté de s'appuyer sur les capacités d'intégration de MAGOS pour le développement d'une infrastructure permettant d'associer de surcroît des données génotypiques et phénotypiques nous a amené à développer MS2PH-db. MS2PH-db peut ainsi être définie comme une banque de données relationnelle, qui représente un système dédié à l'analyse globale des mutations et de leurs conséquences structurales et phénotypiques dans le cadre des maladies monogéniques humaines.

Nous évoquerons dans ce chapitre les difficultés que représentent la détermination et la collecte de données génotypiques et phénotypiques répondant à nos besoins. Nous présenterons ensuite MS2PH-db, en nous appuyant notamment sur l'exemple d'une protéine de la banque, la myotubularine. Nous terminerons ce chapitre en démontrant l'intérêt de disposer d'une infrastructure informatique puissante comme celle fournie par le programme Décrypthon dans le cadre de développements d'envergure comme MS2PH-db.

13.1. Données génotypiques et phénotypiques

L'une des principales difficultés rencontrées au cours du déploiement de MS2PH-db a été la détermination des données génotypiques et phénotypiques que nous pourrions intégrer dans notre système.

Nous avons considéré les banques de données de variations génétiques disponibles à travers le Web, afin de nous familiariser avec les données disponibles et d'être à même de définir les solutions qui répondraient au mieux à nos ambitions.

13.1.1. Banques de données de mutations

Comme nous l'avons déjà évoqué dans le Chapitre 3 - *Les maladies génétiques humaines*, p. 47, on recense quatre types de banques de données spécialisées dans la centralisation des variations génétiques : les banques de données centrales, les banques de données spécifiques d'un locus (ou LSDB), les banques de polymorphismes et les banques nationales de mutations.

Les conséquences des variations génétiques ne sont pas renseignées dans les banques de polymorphismes ; les banques nationales de mutations sont, quant à elles, dédiées aux analyses de populations. C'est pourquoi nous nous sommes exclusivement intéressés aux banques centrales et aux LSDB.

Rappelons que, schématiquement, la philosophie des banques de données centrales consiste en la collecte de toutes les données de mutations ayant trait à l'ensemble des gènes auxquels une maladie génétique est associée. Les données sont, pour la plupart, extraites de la littérature et présentées avec un minimum de détails, notamment en ce qui concerne des descriptions phénotypiques, généralement restreintes à un nom de maladie, éventuellement associé à un degré de sévérité.

Les LSDB se concentrent quant à elles sur les variations affectant un gène (ou une famille de gènes) particulier. Elles sont gérées par des chercheurs ayant une expertise dans ce gène ou les maladies qui y sont associées. Le principal intérêt de ces banques est la qualité des données et des annotations qui y sont collectées. Selon les banques, les descriptions génotypiques et phénotypiques peuvent atteindre des degrés de détails très élevés (description des deux allèles du patient, plusieurs niveaux de description phénotypique, etc.). Près de la moitié des données contenues dans les LSDB sont des données non publiées (Cotton, 2000), saisies directement par les expérimentateurs.

Ces deux types de banques de données sont complémentaires et chacune bénéficie de l'existence de l'autre. Les banques centrales présentent l'avantage de fournir un format homogène pour l'ensemble des gènes, mais le manque de détails les rend assez peu intéressantes pour répondre aux besoins de la génétique médicale. Les LSDB contiennent un niveau de détails bien plus approprié mais ces banques sont souvent développées à façon, pour répondre aux besoins spécifiques des cliniciens et/ou des chercheurs. Un manque d'homogénéité entre les LSDB est par conséquent à déplorer.

13.1.2. Dans un monde idéal…

Dans le cadre de la mise en place de notre infrastructure, il nous faudrait pouvoir disposer, pour chacune des protéines impliquées dans une maladie monogénique humaine, d'un accès à une LSDB qui contiendrait :

- une description génotypique précise, présentant pour chaque mutation des données relatives à la zygotie de l'individu la portant et une description du second allèle en cas d'hétérozygotie,
- une description phénotypique homogène et précise, regroupant plusieurs niveaux de précision (molécule, cellule, tissu, organisme).

De plus, chacune des banques devrait disposer d'une API (pour *Application Programming Interface*), c'est-à-dire une interface permettant une interrogation aisée de la base ou, à défaut, un site FTP (pour *File Transfer Protocol*), pour pouvoir rapatrier les données et les stocker localement.

Enfin, l'ensemble des LSDB devrait être homogène, que ce soit au niveau de leur structure, de la qualité de leur contenu ou des descriptions génotypiques et phénotypiques reposant sur des standards, voire des ontologies, adaptés.

13.1.3. En réalité…

A ce jour (octobre 2007), on compte près de 700 LSDB référencées et accessibles par le Web (http://www.hgvs.org/dblist/glsdb.html). Nombreuses sont celles qui présentent pas de descriptions génotypiques et/ou phénotypiques très détaillées et une incroyable hétérogénéité entre les différentes banques est à déplorer (Claustres *et al.*, 2002).

En l'état actuel des choses, il n'est, par conséquent, pas envisageable d'exploiter l'ensemble des LSDB disponibles de manières automatique et massive.

13.1.4. Un espoir : vers une mise en place de standards

L'étude des variations génétiques chez l'homme et la compréhension de leurs conséquences sont considérées comme un challenge majeur dans le domaine des sciences du vivant (Hudson, 2003). La situation relative aux banques de variations, décrite comme une « cacophonie » de banques diverses et variées (editorial, 2005), ne permet cependant pas de mettre au point des solutions efficaces allant dans ce sens. Une prise de conscience de la communauté scientifique s'est opérée ces dernières années (Cotton, 2000) et a été à l'origine de plusieurs initiatives, qui se sont notamment concrétisées par le développement de standards.

Une nomenclature officielle (den Dunnen and Antonarakis, 2000; Horaitis and Cotton, 2004) relative à la description des mutations (http://www.hgvs.org/mutnomen/) a été élaborée et est distribuée par la HGVS (pour *Human Genome Variation Society*). Elle permet de décrire sans ambiguïté les variations génétiques au niveau de l'ADN, de l'ARN et de la protéine.

Des recommandations concernant le contenu minimal des LSDB (Scriver *et al.*, 1999) ont également été introduites et des systèmes de génération de LSDB ont été développés, comme

UMD (Beroud *et al.*, 2000), LOVD (Fokkema *et al.*, 2005) et MUTbase (Riikonen and Vihinen, 1999). L'idée qui sous-tend ces systèmes est de permettre aux cliniciens d'installer aisément un programme flexible, ne nécessitant pas de support informatique particulier, et prêt à accueillir leurs données (mutations, phénotypes, patients). Des fonctionnalités de base pour la manipulation des données y sont intégrées et une architecture modulaire permet une extension aisée autour du cœur du service. La généralisation de l'utilisation de ces systèmes devrait, d'une part, favoriser l'uniformisation des LSDB et, d'autre part, améliorer la qualité des données diffusées. En effet, ils intègrent un certain nombre d'outils permettant, par exemple, la vérification de la cohérence des mutations saisies (Claustres *et al.*, 2002).

Le système UMD propose des solutions de visualisation des données, ayant pout but de faciliter l'analyse des relations qui lient le génotype au phénotype (Figure 67), ainsi que des aides au développement de solutions thérapeutiques (Beroud *et al.*, 2005).

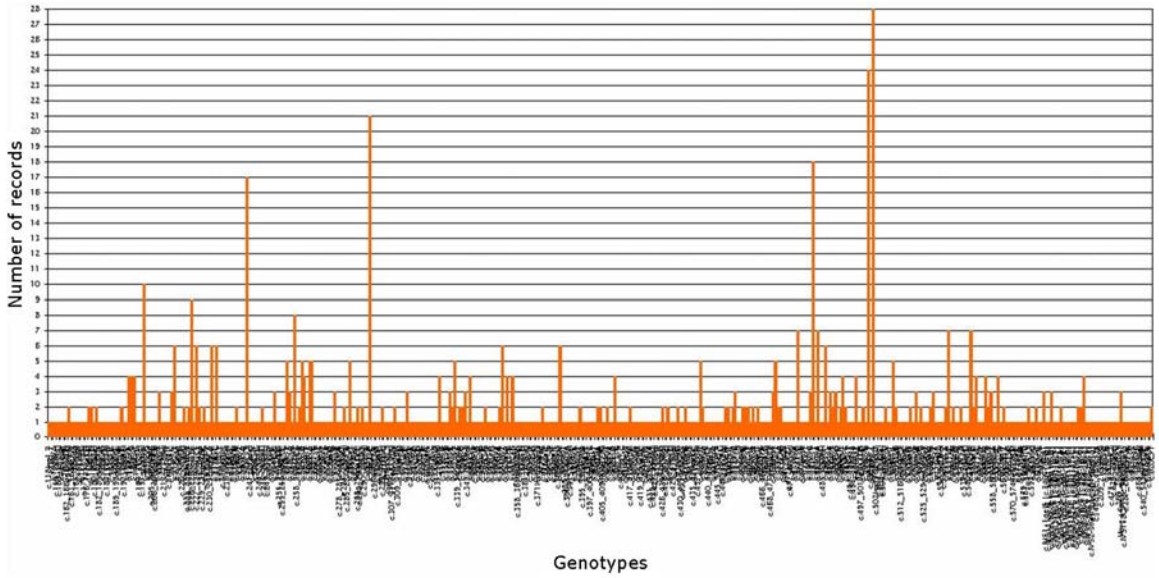


Figure 67 - Fréquence des mutations observées le long d'un gène impliqué dans le syndrome de von Hippel Lindau. Ce graphique est extrait d'une banque de données UMD, accessible à l'adresse suivante : http://www.umd.be:2020/.

Force est de constater qu'à l'heure actuelle, en dépit des efforts déployés en termes de simplicité de manipulation et de convivialité, seule une faible proportion des LSDB est construite à partir de ces systèmes. De plus, il est important de noter que de nombreuses banques développées et maintenues au sein de divers laboratoires restent « privées » et ne sont pas diffusées à travers le Web. Une explication de ce phénomène réside dans la difficulté à gérer la confidentialité des données relatives aux patients, qui soulève des questions d'éthique.

En outre, bien que les efforts réalisés pour la standardisation des données de variations et les solutions de stockage soient conséquents et reposent sur un travail de longue haleine, ils

semblent « minimes » par rapport à celui que représente la mise en place d'une standardisation de l'information phénotypique (editorial, 2005). En effet, pour être efficacement interprétées et liées à des données génétiques, les données phénotypiques ne peuvent plus être réduites à du texte libre. Le développement et l'utilisation d'ontologies relatives aux phénotypes, reposant sur l'analyse complète des phénotypes, sont devenus résolument indispensables pour l'harmonisation des descriptions (Gerlai, 2002; Scriver, 2004). Un projet centré sur le phénome humain, le *Human Phenome Project* (Freimer and Sabatti, 2003), a notamment vu le jour il y a quelques années et a pour objectifs de clarifier ce qui constitue un phénotype et de délimiter les composants phénotypiques du phénome.

Enfin, un projet récent semble particulièrement intéressant dans le cadre de nos développements : le projet du variome humain, ou *Human Variome Project* (editorial, 2007; Ring *et al.*, 2006), dont l'une des missions principales est de favoriser l'unification des données relatives aux variations génétiques humaines et de comprendre leurs impacts phénotypiques, afin de faciliter l'étude des relations génotype/phénotype.

13.1.5. Notre choix

La situation actuelle ne nous permet pas d'envisager, dans l'immédiat, l'alimentation de notre système avec des données issues de LSDB pour chaque protéine impliquée dans une maladie monogénique humaine.

Nous avons par conséquent choisi, afin de couvrir le spectre le plus large possible d'entrées de notre système, d'intégrer dans un premier temps des données issues de banques de données centrales.

Les trois principales banques centrales que nous avons considérées sont OMIM (Hamosh *et al.*, 2005), la HGMD (Stenson *et al.*, 2003) et Swiss-Prot (Boeckmann *et al.*, 2003) (*voir* section 3.4.1, p. 49).

La structure d'OMIM, qui repose essentiellement sur du texte libre, interdit son exploitation automatique sans un outil puissant de fouille de données, dont nous ne disposons pas encore au sein du laboratoire.

La HGMD est sans conteste la banque centrale la plus complète mais des droits d'auteurs très stricts régissent la manipulation de cette banque, qu'il n'est permis d'interroger que manuellement, après authentification. Le chargement automatique et la diffusion des données ne sont pas autorisés. Nous sommes actuellement en contact avec les gestionnaires de cette banque, dans l'optique d'établir une collaboration, qui nous permettrait de disposer des données et éventuellement de les diffuser sous certaines conditions.

Tant que les modalités de la collaboration ne seront pas officiellement fixées, nous nous limiterons à l'intégration, dans notre système, des données de mutations relatives aux

protéines impliquées dans les maladies génétiques humaines répertoriées dans Swiss-Prot. Ces dernières ne représentent cependant qu'une quantité relativement faible comparativement aux données accessibles par la HGMD.

Par la suite, nous intégrerons, au fur et à mesure, des données issues de LSDB, afin d'apporter plus de détails pour un certain nombre de protéines d'intérêt. Ces intégrations se feront en fonction de nos besoins ou sur demande. Nous intègrerons en priorité des données de LSDB construites à l'aide d'outils génériques comme UMD, LOVD ou MUTbase.

13.2. MS2PH-db

Les bases de l'infrastructure dédiée à l'analyse globale des mutations et de leurs conséquences structurales et phénotypiques, dans le cadre des protéines impliquées dans les maladies monogéniques humaines, ont pris la forme d'une banque de données relationnelle, MS2PH-db. Celle-ci fournit, pour chacune de ses entrées, c'est-à-dire pour chaque protéine impliquée dans une maladie monogénique humaine, trois types de vues : une vue évolutive par l'intermédiaire d'alignements multiples ; une vue structurale par l'intermédiaire d'un ou plusieurs modèles tridimensionnels construits par homologie ; une vue informationnelle par l'intermédiaire des données généralistes, fonctionnelles, structurales et « cliniques » (mutations et phénotypes associés). Ces trois types de vues peuvent être interconnectés par l'intermédiaire d'une interface graphique conviviale, dérivée de celle de MAGOS.

Les détails concernant la structure de la banque, ainsi que la manière dont les données sont récupérées et calculées, sont décrits dans le Chapitre 9 - *MS2PH-db*, p. 123. Le système d'interrogation intégré à MS2PH-db y est également abordé dans ses grandes lignes, c'est pourquoi nous n'y reviendrons pas dans ce chapitre.

Nous présenterons succinctement dans cette section les données auxquelles MS2PH-db permet d'accéder, en nous concentrant sur une protéine d'intérêt avec laquelle nous avons eu l'occasion de nous familiariser au préalable, la myotubularine (*voir* section 12.1.1, p. 179), avant de dévoiler quelques chiffres relatifs à MS2PH-db.

13.2.1. La myotubularine : une entrée de MS2PH-db

13.2.1.1. Carte d'identité de la myotubularine : vue informationnelle

La carte d'identité de la myotubularine (Figure 68) regroupe un certain nombre d'informations générales comme son identifiant UniProt (Q13496), son nom, le nom du gène qui la code (MTM1), ainsi que des liens croisés vers d'autres banques publiques, comme GO, OMIM et UniProt (onglet *Cross-references*) (Figure 68 A).

L'onglet *Mutants* permet un accès aisé aux données de mutations connues, récupérées au sein de Swiss-Prot (Figure 68 E) : les 58 mutations faux-sens référencées dans Swiss-Prot (au

10 octobre 2007) y sont consultables. La position de la mutation dans la séquence, ainsi que l'acide aminé qu'on y trouve dans la protéine sauvage et l'acide aminé muté y sont indiqués. Chaque mutation est associée au nom de la maladie qu'elle engendre, en l'occurrence la XLMTM, pour *X-linked myotubular myopathy*. Des informations supplémentaires relatives à la maladie sont aisément accessibles en utilisant le lien vers la banque OMIM (onglet *Cross-references*).

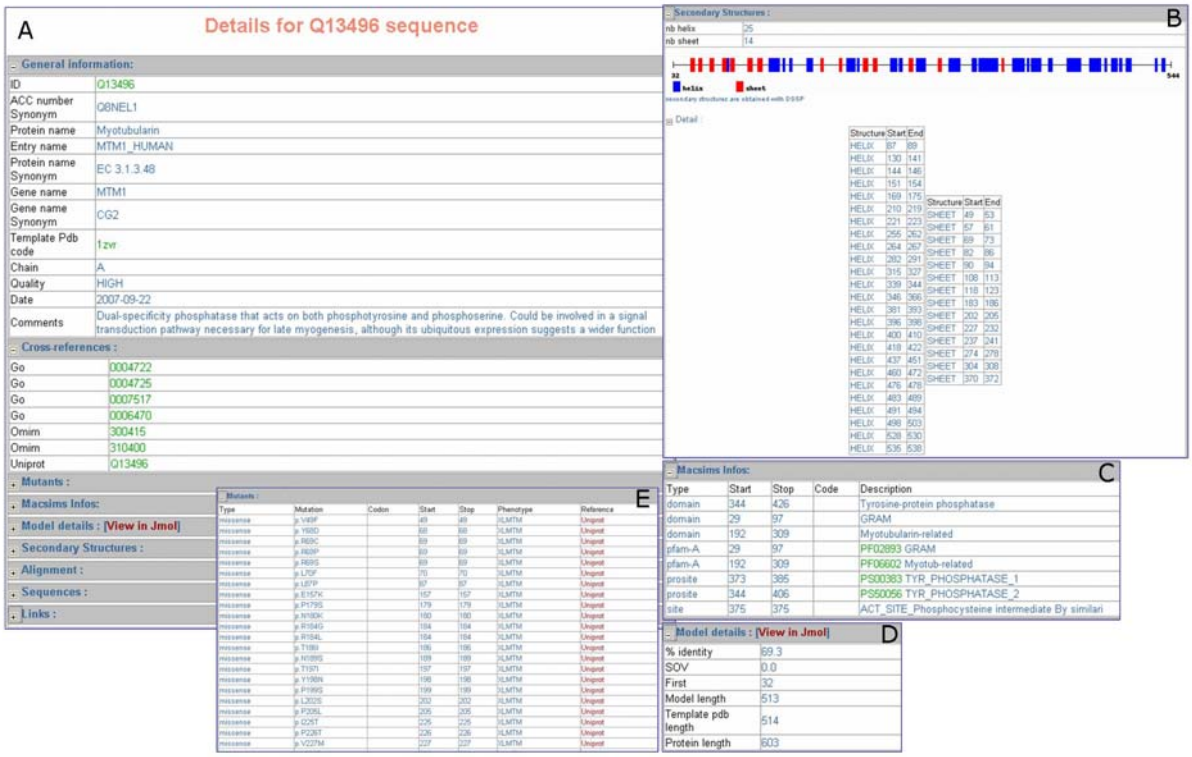


Figure 68 - Carte d'identité de la myotubularine dans MS2PH-db.

Un récapitulatif des données pré-calculées complète cette carte d'identité. Des annotations relatives aux domaines et sites fonctionnels associées à la séquence de la myotubularine par MACSIMS sont disponibles (onglet *Macsims Infos*, Figure 68 C). MACSIMS récupère des informations relatives aux domaines et aux sites fonctionnels dans plusieurs banques de données : on peut remarquer que les positions des domaines PH-GRAM, *Myotubularin-related* et tyrosine phosphatase y sont assignées de manière correcte (*voir* Figure 61).

Enfin, des données relatives au modèle structural construit sont disponibles. On retrouve, au niveau de l'onglet *General Information*, l'identifiant PDB de la structure utilisée comme empreinte (PDB : 1ZVR, chaîne A), ainsi que des informations relatives au pourcentage d'identité que partagent la myotubularine et l'empreinte, à la longueur du modèle et à celle de la protéine entière (onglet *Model details*, Figure 68 D). L'alignement utilisé pour la modélisation par homologie peut également être visualisé (onglet *Alignment*). La séquence de la myotubularine partage près de 70% d'identité avec celle de l'empreinte sélectionnée ;

l'alignement par paires utilisé pour la construction du modèle ne contient pas de grandes insertions ou délétions et semble de très bonne qualité ; le modèle construit couvre une grande partie de la protéine (85%). L'ensemble de ces données permet d'indiquer que la qualité du modèle tridimensionnel construit est bonne.

Les structures secondaires assignées par le logiciel DSSP sont également disponibles (onglet *Secondary Structures*, Figure 68 B) de manière visuelle, sous la forme d'un diagramme schématisant la succession d'hélices et de feuillets, ainsi que sous la forme d'un tableau qui reprend les positions de début et de fin de chaque élément de structure secondaire. On peut remarquer une concentration de feuillets dans la partie N-terminale (au niveau du domaine PH-GRAM notamment) de la myotubularine et une majorité d'hélices dans la partie C-terminale.

13.2.1.2. Interconnexion des vues informationnelles, structurales et évolutives

Le lien *View in Jmol*, accessible par l'onglet *Model details*, permet d'accéder à l'interface d'interconnexion des vues informationnelles, structurales et évolutives relatives à la myotubularine. Cette interface est dérivée de celle mise en place dans le cadre de MAGOS.

Par défaut, l'alignement multiple annoté des séquences eucaryotes homologues de la myotubularine est affiché dans la partie droite de l'interface et est connecté au modèle tridimensionnel visualisable dans la partie de gauche. L'ensemble des annotations assigné à la séquence de la myotubularine par MACSIMS, ainsi que les régions et/ou domaines les mieux conservés, sont visualisables de manière simultanée sur la séquence au sein de l'alignement multiple et au niveau du modèle tridimensionnel.

L'alignement des séquences homologues retrouvées dans la banque UniRef90 et sélectionnées par la méthode d'échantillonnage des bandelettes (*Sample*), ainsi que l'alignement de la séquence d'intérêt avec l'empreinte PDB utilisé pour la construction du modèle, peuvent optionnellement être affichés dans l'onglet de droite, en modifiant les paramètres de l'onglet *Link Parameters*, situé dans la partie de gauche de l'interface.

D'après les scores NorMD associés aux deux alignements multiples (0,70 pour l'alignement des séquences eucaryotes et 0,57 pour l'alignement des séquences échantillonnées), ces derniers sont de relativement bonne qualité.

En outre, l'interface mise en place dans le cadre de MS2PH-db intègre des fonctionnalités supplémentaires par rapport à MAGOS. En effet, l'onglet *Mutations* (Figure 69) permet de localiser la position des mutations engendrant l'émergence d'une maladie sur le modèle structuralet sur la séquence de la myotubularine au sein de l'alignement multiple. Le modèle de chacune des protéines mutées peut être généré à la volée, en cliquant sur *Generate*. Ceci offre la possibilité à l'utilisateur de visualiser le modèle qu'il souhaite au niveau de l'onglet Jmol parmi l'ensemble des modèles qui ont été générés.

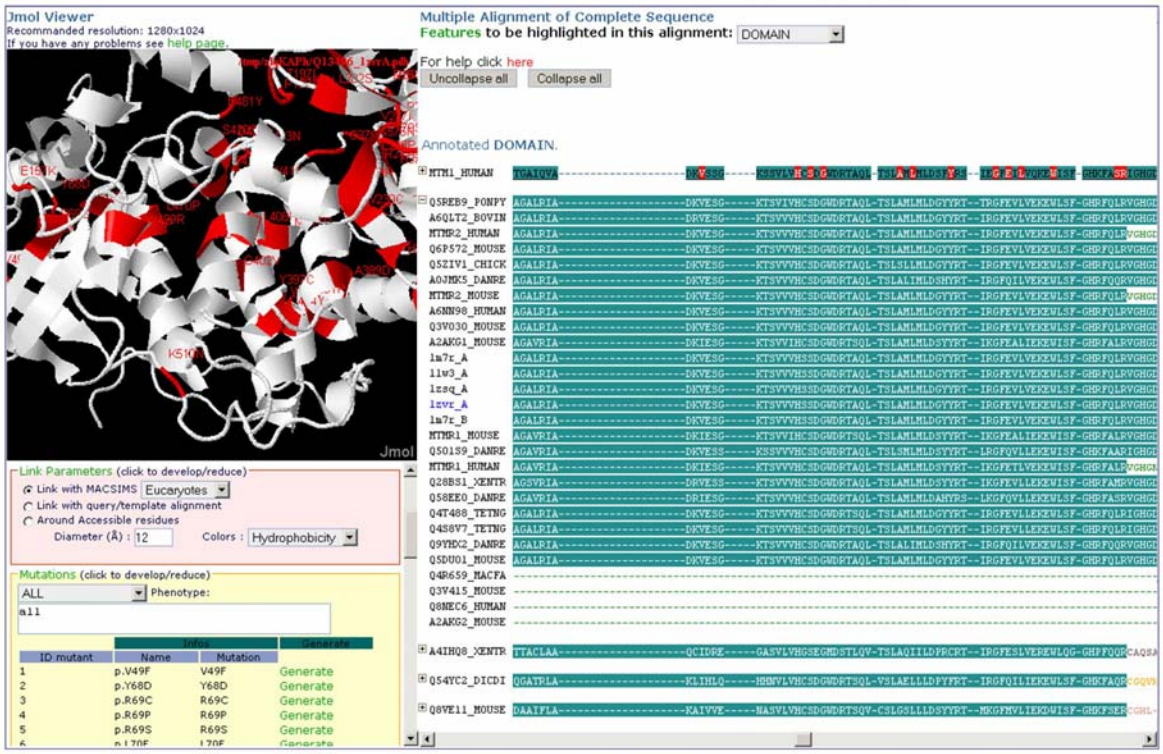


Figure 69 - Localisation de l'ensemble des mutations affectant la myotubularine et liées à l'émergence d'une maladie. Les positions des mutations sont cartographiées sur le modèle structural de la protéine ainsi que sur la séquence de la myotubularine dans le cadre de l'alignement multiple.

L'option *Around Accessible residues*, accessible par l'onglet *Link Parameters*, permet à l'utilisateur de visualiser les résidus de surface situés dans un environnement structural proche d'un résidu d'intérêt.

La Figure 70 représente les résidus situés à moins de 12 Å de l'arginine située en position 421 de la myotubularine. La modification de cette arginine en glutamine est à l'origine de l'émergence d'une myopathie myotubulaire (XLMTM). On remarque que cette arginine se situe à moins de 12 Å de l'arginine en position 377 de la séquence, qui est située au cœur de la poche catalytique de la myotubularine. On peut penser que la mutation de l'arginine en glutamine est à l'origine d'une déstabilisation de la poche du site actif, qui expliquerait l'émergence de la maladie, comme nous avions déjà pu l'identifier lors de l'analyse de cette mutation par l'intermédiaire de MAGOS (*voir* section 12.1.3.3, p. 183).

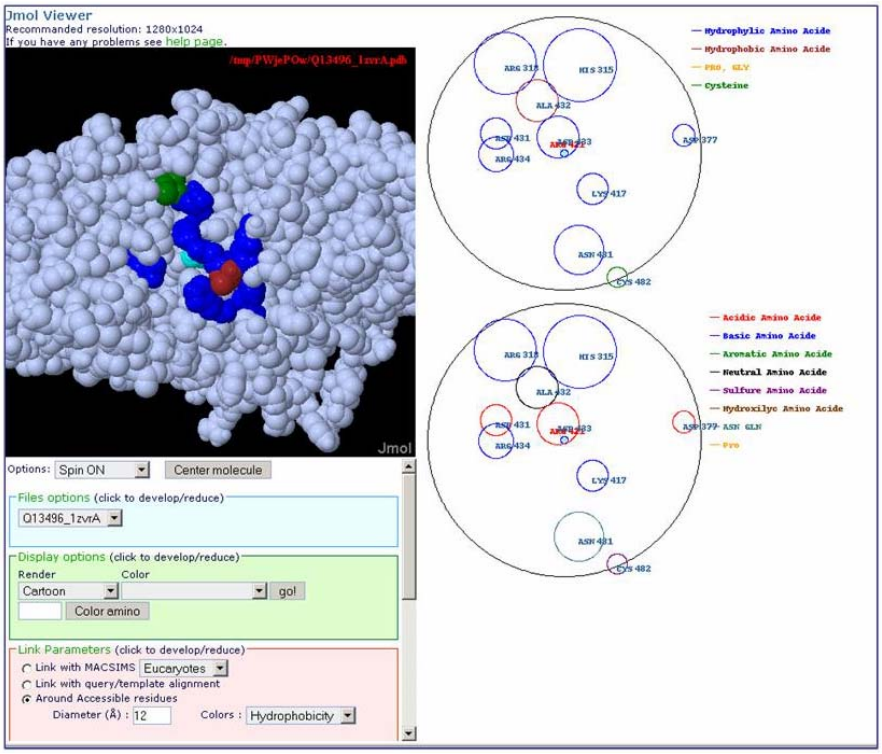


Figure 70 - Visualisation des résidus situés à moins de 12 Å de l'arginine en position 421 de la séquence de la myotubularine.

Ainsi, l'ensemble des données fournies pour la myotubularine par MS2PH-db et leur interconnexion *via* un environnement de travail convivial permet d'aborder aisément différents aspects de cette protéine, tant au niveau de sa séquence, de sa structure, de sa fonction que de son évolution, et de les lier avec des données génotypiques et phénotypiques.

13.2.2. MS2PH-db en quelques chiffres

Au mois d'octobre 2007, MS2PH-db regroupait des données relatives à 1 915 protéines impliquées dans des maladies monogéniques humaines.

Pour chacune de ces protéines, un alignement multiple annoté regroupant les séquences eucaryotes homologues et un autre construit en utilisant la méthode d'échantillonnage des bandelettes (*voir* Chapitre 11 - *Alignements multiples dédiés à l'analyse structurale*, p. 145) sont disponibles.

1 037 modèles tridimensionnels, relatifs à 1 013 protéines (990 protéines avec un modèle associé, 22 protéines avec deux modèles associés et une protéine avec trois modèles associés), ont pu être construits.

Les données de mutations sont au nombre de 16 783 et sont relatives à 3 234 phénotypes distincts. Notons que si nous pouvions bénéficier de l'accès aux données de la HGMD, en se

limitant aux données publiquement accessibles, nous disposerions de plus de 30 000 mutations faux-sens ou non-sens. C'est pourquoi il est essentiel pour nous que les termes de la collaboration avec la HGMD soient définis au plus vite.

13.3. L'absolue nécessité de l'infrastructure Décrypthon dans le cadre de MS2PH-db

Dans le cadre du développement de MS2PH-db, il était absolument primordial de disposer d'une grande puissance informatique, notamment pour les calculs relatifs aux protéines en entrée de la banque. C'est dans ce contexte que nous avons pu mettre à profit l'infrastructure déployée dans le cadre du programme Décrypthon, qui s'appuie sur une grille de calcul universitaire, constituée de supercalculateurs, ainsi que sur une grille dite « d'internautes », constituée d'ordinateurs personnels.

Les calculs nécessaires à la construction des alignements multiples étant, d'une part, fortement couplés entre eux et demandant, d'autre part, l'échange de données de taille importante et à une fréquence élevée, la grille de calcul universitaire s'est rapidement révélée être mieux adaptée à nos besoins que la grille d'internautes.

13.3.1. Déploiement de la grille de calcul universitaire Décrypthon

En tant qu'acteurs d'un projet pilote du programme Décrypthon, nous avons participé activement aux discussions relatives à l'implémentation de la grille de calcul universitaire Décrypthon. Les experts d'IBM en charge du déploiement de cette infrastructure ont pris en compte les caractéristiques des programmes que nous souhaitions y installer, notamment en termes de transfert de données, pour le choix du logiciel de pilotage de la grille.

L'architecture de la grille de calcul universitaire Décrypthon, qui regroupe 112 nœuds de calcul, est décrite plus en détails dans la section 6.1.2.1, p. 81.

13.3.2. Implémentation des outils sur la grille de calcul Décrypthon

Le portage de nos applications sur la grille de calcul universitaire s'est déroulé en deux étapes : nous avons implémenté, dans un premier temps (juin 2005), la suite logicielle PipeAlign (*voir* section 6.3.5, p. 93), puis, dans un second temps (mai 2007), l'ensemble du protocole mis en place dans le cadre de MS2PH-db (*voir* section 9.4.1, p. 126) qui conduit, à partir d'une protéine d'intérêt, à l'obtention de deux alignements multiples distincts annotés, en passant par la sélection d'empreintes et l'extraction des alignements nécessaires à l'étape de modélisation par homologie.

Ce portage, fruit d'une collaboration avec des ingénieurs de l'ENS de Lyon, a nécessité la mise en conformité des applications à l'environnement de grille, ainsi qu'une étape de

vérification des résultats. Précisons que, le développement et les tests minutieux de l'ensemble des applications portées sur la grille, effectués sous GScope (*voir* section 6.2.1, p. 82), ainsi que la réactivité et le professionnalisme de nos collaborateurs ont permis une réalisation très efficace de ces étapes de transfert.

13.3.3. Intérêt de la grille Décrypthon dans le cadre de MS2PH-db

C'est dans le cadre de développements de grande ampleur, qui demandent l'intégration d'une vaste quantité de données hétérogènes, des calculs coûteux et une mise à jour régulière, que la grille de calcul universitaire Décrypthon prend tout son sens.

La mise à jour des données de MS2PH-db est programmée tous les deux mois et implique notamment la reconstruction des deux alignements multiples annotés pour chaque protéine entrée de MS2PH-db, soit près de 2 000 protéines.

Le Tableau 10 présente une estimation des temps de calcul nécessaires pour la construction de ces alignements à l'IGBMC (en utilisant deux processeurs du serveur *Star* à temps plein) et sur la grille de calcul universitaire. Le temps de calcul estimé pour une protéine représente une moyenne que nous avons obtenue en réalisant les calculs pour 100 protéines. On remarque que le temps de traitement d'une protéine est le même, que les calculs soient effectués au laboratoire ou sur la grille. L'utilisation de la grille de calcul devient pertinente quand de nombreuses constructions sont à réaliser. On observe que pour l'ensemble des 1 915 protéines de MS2PH-db, le temps de calcul nécessaire en utilisant la grille de calcul est de trois jours, alors qu'il est estimé à quasiment un mois si on réalise ces calculs au laboratoire, en monopolisant deux processeurs. Précisons que les temps de calcul estimés sur la grille tiennent compte du fait que nous ne sommes pas les seuls utilisateurs de la grille (d'autres applications y étant implémentées), ainsi que des aléas informatiques, qui impliquent que l'ensemble des processeurs ne soit pas disponible à tout moment.

	1 protéine	MS2PH-db (~ 2 000 protéines)
IGBMC (2 processeurs)	20 minutes	28 jours
Grille de calcul universitaire Décrypthon	20 minutes	3 jours

Tableau 10 - Temps de calcul nécessaires pour la construction de deux alignements multiples annotés.

La grille de calcul Décrypthon joue ainsi un rôle clé au cœur de la banque de données MS2PH-db, dont le développement et surtout la mise à jour régulière ne sont envisageables que grâce à sa disponibilité.

13.4. Conclusions et perspectives

Par le développement de MS2PH-db, nous avons mis en place les bases de l'infrastructure dédiée à l'analyse globale des mutations, dans le cadre des maladies monogéniques humaines. Le déploiement de cette banque de données repose sur la puissance de la grille de calcul universitaire Décrypthon et permet d'accéder, de manière aisée et conviviale, à différents niveaux d'informations relatifs à chaque protéine impliquée dans une maladie monogénique humaine. Plusieurs aspects de l'étude de ces protéines (séquence, structure, fonction et évolution) peuvent être abordés de manière couplée et les mutations à l'origine de l'émergence d'une pathologie peuvent y être confrontées.

MS2PH-db ne représente cependant que le socle de l'infrastructure globale visant à mieux comprendre les relations qui lient génotypes et phénotypes dans le cadre des maladies monogéniques humaines. Cette banque est par conséquent vouée à évoluer rapidement.

Il nous faudra donc être à même d'alimenter notre système avec des données génotypiques et phénotypiques plus complètes et plus précises.

Enfin, nous souhaitons rendre accessible un ensemble supplémentaire de données pré-calculées, pour chacune des protéines en entrée de la banque, qui nous semblera pertinent dans le cadre de l'élucidation de la relation qui lie le phénotype de la maladie au génotype d'un individu. Ces données seront identifiées grâce à des études que nous avons d'ores et déjà entamées et qui seront abordées dans le prochain chapitre.

Chapitre 14 - Vers une typification des mutations

Le développement de solutions permettant de comprendre et de prédire les effets des variations génétiques sur le phénotype d'un individu constitue un enjeu primordial de la recherche biomédicale.

Nous avons pu mettre en place, au cours de nos travaux, les bases d'une plateforme consacrée à l'analyse et à la compréhension des mutations affectant les protéines impliquées dans les maladies monogéniques humaines et de leurs conséquences. Cependant, le chemin restant à parcourir pour arriver au déploiement d'un système générique de prédiction des effets d'une mutation sur le phénotype clinique est encore extrêmement long.

Nous souhaitons néanmoins clore cette partie *Résultats & Discussion* en ouvrant la voie vers la mise en place d'un système expert de compréhension et de prédiction des effets des mutations, qui s'appuie sur des protéines pour lesquelles les relations génotype/phénotype ont, au moins en partie, pu être établies. Ce chapitre se veut un chapitre exploratoire. Il a pour ambition de mettre en place des éléments de typification des mutations dont l'effet est connu, pour envisager le déploiement d'un système de prédiction automatique.

Après un bref rappel des méthodes de prédiction des effets des variations génétiques déjà disponibles et des ambitions du projet MS2PH, nous présenterons les relations qui lient le génotype au phénotype dans le cadre de deux maladies monogéniques humaines. Nous verrons comment il est possible d'envisager la typification des positions des mutations connues et nous nous concentrerons sur les leçons à tirer de telles études ainsi que sur une série de points clés qui devront être pris en compte pour le déploiement du système expert.

14.1. Prédiction des conséquences des mutations

14.1.1. Discrimination des mutations neutres/délétères

Des outils dédiés à la prédiction des effets des mutations, introduits dans le Chapitre 5 - *Vers une approche intégrative des mutations et de leurs conséquences*, p. 68, ont d'ores et déjà été développés, à l'image de SIFT (Ng and Henikoff, 2003), PolyPhen (Ramensky *et al.*, 2002) et SNAP (Bromberg and Rost, 2007).

Les méthodes de prédiction implémentées sont, pour la plupart, spécifiques des mutations ponctuelles faux-sens et limitées à la discrimination des mutations délétères (avec un effet majeur sur la protéine) par rapport aux mutations neutres. Ces méthodes sont toutes basées

sur des données de séquences, et plus particulièrement de conservation des résidus au cours de l'évolution, qui peuvent, selon les cas, être complétées par des données fonctionnelles et/ou structurales. Dans leur philosophie, les outils disponibles partent du principe que chaque position d'une protéine est essentielle et ils visent à déterminer, pour chacune d'elles, les substitutions tolérées.

D'après les résultats de sensibilité et de spécificité de détection des mutations délétères (*voir* Tableau 6, p. 72), ces méthodes de prédiction permettent d'obtenir des résultats satisfaisants. Cependant, certains points restent problématiques :

- d'une part, les résultats de spécificité sont biaisés par les jeux de données testés et notamment par le manque de données relatives aux mutations neutres, pourtant essentielles pour tester la robustesse des programmes (Care et al., 2007). Pour pallier ce manque, des jeux de données artificielles ont été créés en postulant que les substitutions observées chez un organisme proche, comme la souris, seront tolérées, donc neutres, chez l'homme. Cependant, ces données artificielles sont elles-mêmes biaisées, puisqu'il est raisonnable de penser qu'un résidu variable entre l'homme et la souris n'a pas subi une forte pression de sélection au cours de l'évolution, or c'est l'un des critères essentiels des méthodes de prédiction des effets des mutations ; il semble par conséquent indéniable que ces substitutions soient prédites comme tolérées. L'utilisation d'un jeu de données créé de cette manière donne obligatoirement des résultats de spécificité très élevés ;
- d'autre part, un sytème de prédiction à deux états (neutre/délétère) des effets d'une mutation sur une protéine est difficile à exploiter dans un contexte phénotypique. En effet, si les conclusions relatives à une mutation neutre semblent évidentes, étudier ou prédire les relations entre une mutation « délétère » et le phénotype clinique d'un individu est autrement plus compliqué et nécessite de prendre en compte de nombreuses données supplémentaires.

14.1.2. Prédiction d'un phénotype clinique à partir du génotype d'un individu

L'ambition du projet MS2PH consiste, à terme, en la mise à disposition d'un système générique de prédiction des effets délétères d'une mutation qui ne se limiterait pas aux conséquences moléculaires, mais permettrait d'extrapoler ces effets au phénotype clinique de l'individu, en fournissant les informations nécessaires pour lier un génotype à un phénotype clinique. Dans le cadre d'une telle démarche, seules les maladies monogéniques peuvent être considérées, puisque, schématiquement, les défauts dans un seul gène sont à l'origine de l'émergence d'un phénotype clinique. Rappelons que des facteurs génétiques et environnementaux peuvent également influencer l'expression du phénotype (*voir*

Chapitre 3 - *Les maladies génétiques humaines*, p. 36), mais la complexité de ces relations et le manque fréquent d'informations à leur sujet nous empêchent, pour l'instant, de les prendre en compte.

Dans le cadre de nombreuses maladies monogéniques, des phénotypes distincts émergent suite à des mutations affectant la même protéine. Ces phénotypes sont généralement caractérisés par des sévérités de symptômes différentes, mais peuvent aussi définir deux maladies distinctes, à l'exemple des défauts dans la dystrophine pouvant être associés à l'émergence d'une myopathie de Duchenne ou d'une myopathie de Becker (Muntoni *et al.*, 2003). L'émergence de phénotypes distincts à partir de mutations affectant une même protéine implique que deux mutations délétères n'auront pas forcément le même effet sur cette protéine. Ces effets dépendront du type de mutation qui affecte la protéine (mutation non-sens, faux-sens, etc.), ainsi que de sa position au sein de cette dernière. Il n'est, par conséquent pas possible, dans le cadre de ce projet, de se limiter à la discrimination des mutations délétères par rapport aux mutations neutres. Il est, au contraire, nécessaire d'être capable de discriminer les mutations responsables de phénotypes distincts dans différents contextes liés au génotype et, notamment, à l'homozygotie ou à l'hétérozygotie, à l'organisation de la protéine, à l'existence de variants d'épissage, à l'expression tissulaire, etc.

Dans l'optique de commencer à baliser le chemin vers la mise en place d'un système expert de compréhension et de prédiction des effets des mutations, nous avons choisi de nous baser sur deux « systèmes » bien documentés, que nous présenterons dans la prochaine section.

14.2. Présentation des deux « systèmes » étudiés

Dans le cadre des maladies monogéniques, deux cas doivent être clairement distingués : les maladies génétiques dont le gène responsable est porté par un chromosome autosome et celles dont le gène responsable est porté par un chromosome sexuel. Dans le cas des maladies autosomiques, la présence de deux allèles et les relations qu'ils partagent sont à l'origine de l'émergence d'un phénotype clinique, tandis que dans celui des maladies liées aux chromosomes sexuels, le lien entre la présence d'une mutation et le phénotype clinique est plus direct, en particulier chez les garçons, du fait de l'absence d'un second allèle.

Ainsi, nous avons choisi de nous intéresser, d'une part, à une maladie génétique autosomique, l'hypophosphatasie et, d'autre part, à une maladie dont le gène responsable est porté par le chromosome X, la myopathie myotubulaire liée à l'X.

14.2.1. Système 1 : l'hypophosphatasie

14.2.1.1. Maladie et phénotypes cliniques

L'hypophosphatasie (OMIM 241500, 241510, 146300) est une maladie héréditaire rare, due au déficit ou à l'absence d'activité de la phosphatase alcaline non tissu-spécifique (TNSALP). Elle se traduit par un défaut de minéralisation dentaire et/ou osseuse (Mornet and Simon-Bouy, 2004). Cette maladie présente une variabilité clinique très importante et cinq formes sont usuellement distinguées, en fonction de l'âge d'apparition des symptômes (Whyte, 1994) :

- la forme néonatale est caractérisée par des malformations du squelette apparaissant *in utero*, généralement incompatibles avec la vie,
- la forme infantile apparaît vers l'âge de six mois et se traduit par des malformations osseuses et un rachitisme important, elle est souvent létale,
- la forme juvénile est très hétérogène et apparaît en moyenne vers l'âge de cinq ans ; les malformations osseuses sont généralement mineures, la maladie peut conduire à une chute prématurée des dents, associée parfois à un rachitisme modéré,
- la forme adulte se manifeste par une chute prématurée des dents associée à des fractures récurrentes,
- la forme odontologique entraîne la chute prématurée des dents.

Par souci de simplification, nous avons regroupé, dans le cadre de notre analyse, ces différentes formes en deux grandes classes : les formes sévères (néonatale et infantile), qui sont transmises de manière récessive et les formes plus modérées (juvénile, adulte et odontologique), qui peuvent être transmises de manière récessive ou dominante selon le cas.

14.2.1.2. Phosphatase alcaline non tissu-spécifique

La phosphatase alcaline non tissu-spécifique (TNSALP) est une enzyme de 524 acides aminés (Swiss-Prot : PPBT_HUMAN), présente dans la plupart des tissus, qui fonctionne sous forme dimérique dans le sérum. Elle est codée par le gène ALPL, pour *ALkaline Phosphatase Liver form*, localisé sur le bras court du chromosome 1.

La TNSALP appartient à la famille des phosphatases alcalines humaines, qui regroupe également trois isoenzymes tissu-spécifiques (placentaire, intestinale et des cellules germinales), partageant 90 à 98% d'identité entre elles. La TNSALP partage, quant à elle, environ 50% d'identité avec chacune des trois phosphatases alcalines tissu-spécifiques.

La fonction de la TNSALP consiste, *a priori*, à hydrolyser le pyrophosphate inorganique, un inhibiteur de la minéralisation, afin d'en maintenir une concentration adéquate et d'assurer une minéralisation des os contrôlée (Llinas *et al.*, 2006).

14.2.1.3. Relation génotype/phénotype

L'hypophosphatasie est génétiquement très hétérogène : plus de 190 mutations sont répertoriées à ce jour, dont la grande majorité (78,5%) est représentée par des mutations faux-sens (http://www.sesep.uvsq.fr/database_hypo/Mutation.html).

Plusieurs études relatives à la relation qui lie le génotype au phénotype des patients atteints d'hypophosphatasie ont été menées, en combinant des données relatives à l'activité résiduelle de la protéine et des données structurales (Brun-Heath *et al.*, 2005; Herasse *et al.*, 2003; Mornet *et al.*, 2001; Spentchian *et al.*, 2006). Ces relations sont à présent relativement bien comprises.

14.2.1.3.1. Transmission sur le mode dominant

La transmission de certaines mutations sur un mode dominant s'explique par un effet dominant négatif (*voir* section 3.1.2.3, p. 38). Dans ce cas, le monomère sauvage est inhibé par le monomère muté dans le cadre de l'hétérodimère sauvage/mutant, conséquence directe des propriétés allostériques de l'enzyme (Lia-Baldini *et al.*, 2001).

14.2.1.3.2. Transmission sur le mode récessif

Une bonne corrélation a pu être établie entre la sévérité de la maladie et l'activité enzymatique *in vitro* des protéines mutées. La grande hétérogénéité phénotypique observée dans l'hypophosphatasie est principalement liée à la variabilité de cette activité enzymatique, elle-même causée par des mutations faux-sens affectant différents résidus de la protéine.

Les patients présentant un phénotype modéré portent au moins une mutation dans la protéine qui exhibe une activité résiduelle significative, alors que ceux présentant un phénotype sévère portent des mutations qui aboutissent, majoritairement, à des protéines ne possèdant pas d'activité enzymatique (Zurutuza *et al.*, 1999).

Ainsi, schématiquement, les mutations peuvent également être classées en deux groupes : les mutations sévères, correspondant à l'absence d'activité enzymatique et les mutations modérées, qui exhibent une activité résiduelle significative. Les mutations entraînant la production d'une protéine tronquée sont systématiquement sévères. Les mutations faux-sens, quant à elles, peuvent conférer à la protéine une activité enzymatique résiduelle plus ou moins importante.

14.2.2. Système 2 : la myopathie myotubulaire liée à l'X

La myopathie myotubulaire liée à l'X (OMIM 310400) est une maladie musculaire congénitale récessive due à un défaut dans le gène MTM1, situé sur le chromosome X, qui code pour la myotubularine. Ce système ayant été décrit au préalable (*voir* section 12.1.1, p. 179), nous n'entrerons pas dans les détails ici. Rappelons simplement que la maladie présente plusieurs degrés de sévérité à l'origine de la définition de trois phénotypes cliniques : un phénotype sévère, un phénotype intermédiaire et un phénotype modéré.

Cette maladie étant récessive, elle affecte principalement les garçons, chez qui l'émergence du phénotype clinique est directement liée à la présence d'un allèle muté sur le chromosme X. Par conséquent, la sévérité d'une mutation peut directement être estimée d'après les symptômes observés et *vice versa*.

L'ensemble des mutations entraînant la production d'une protéine tronquée (à l'exception d'une mutation non-sens affectant l'extrémité C-terminale de la protéine) est associé à un phénotype sévère. Les mutations faux-sens peuvent, quant à elles, être associées aux différents degrés de sévérité de la maladie.

14.3. Analyses préliminaires relatives aux deux systèmes

D'après les sytèmes présentés, il semble raisonnable de classer les mutations en trois groupes : les mutations neutres, qui n'auront pas d'effet sur la protéine, les mutations modérées et les mutations sévères.

Le manque de données relatives aux mutations neutres nous a conduits à ne pas en tenir compte dans le cadre de cette étude. En effet, seules trois substitutions faux-sens affectant la phosphatase alcaline non tissu-spécifique (p.R152H, p.Y263H et p.V522A) sont, à ce jour, connues pour être neutres et nous n'avons pas trouvé de substitution dans la myotubularine annotée comme neutre (p.E410K est référencée dans la banque dbSNP (Sherry *et al.*, 1999), mais aucun lien avec le phénotype n'est disponible).

Nous nous sommes donc focalisés sur les substitutions faux-sens délétères, auxquelles un degré de sévérité est associé, afin de voir s'il est possible de typifier, en première approximation, les positions de ces substitutions, pour les corréler avec la sévérité observée, en nous basant sur les outils bioinformatiques à notre disposition.

14.3.1. Données génotypiques et phénotypiques

Pour mener à bien une telle analyse, nous devons disposer de données génotypiques et phénotypiques précises, les informations recueillies dans MS2PH-db étant, à l'heure actuelle, insuffisantes.

14.3.1.1. Données de mutations relatives à la TNSALP

Nous avons extrait automatiquement les données de mutations de la TNSALP d'une banque de données spécifique de l'hypophosphatasie (http://www.sesep.uvsq.fr/Database.html). Nous avons complété ces données par des informations provenant de la littérature (Mornet *et al.*, 2001; Zurutuza *et al.*, 1999), notamment pour l'assignation d'un degré de sévérité.

Nous avons pu recueillir 87 positions distinctes pour lesquelles la sévérité d'une substitution est connue : 76 positions sont associées à une mutation sévère, 11 à une mutation modérée.

14.3.1.2. Données de mutations relatives à la myotubularine

Pour la myotubularine, nous avons complété les données disponibles dans Swiss-Prot par des données issues de la littérature (Biancalana *et al.*, 2003; Laporte *et al.*, 2000; Tsai *et al.*, 2005). La majorité des substitutions faux-sens affectant la myotubularine étant liée à un phénotype sévère, nous avons choisi de regrouper, en une seule classe, les mutations associées à un phénotype intermédiaire et celles associées à un phénotype modéré.

Nous avons ainsi pu mettre en évidence 41 positions distinctes pour lesquelles la sévérité d'une substitution est connue : 25 positions sont associées à une mutation sévère et 16 à une mutation modérée.

14.3.2. Prédictions des effets délétères/neutres des substitutions

Nous avons testé les prédictions obtenues grâce aux serveurs SIFT (Ng and Henikoff, 2003), PolyPhen (Ramensky *et al.*, 2002) et SNAP (Bromberg and Rost, 2007), en soumettant à chacun l'ensemble des substitutions délétères dont nous disposions pour nos deux protéines d'intérêt.

La Figure 71 présente les résultats obtenus par les trois serveurs. On remarque qu'ils fournissent tous les trois de meilleurs résultats dans le cadre de la myotubularine (en moyenne 86% des substitutions sont prédites comme délétères contre une moyenne de 74% pour la TNSALP). De plus, on constate que, pour les deux protéines, les résultats obtenus avec SIFT et SNAP sont très proches et meilleurs que ceux obtenus avec PolyPhen, malgré la disponibilité de structures proches pour nos deux protéines d'intérêt.

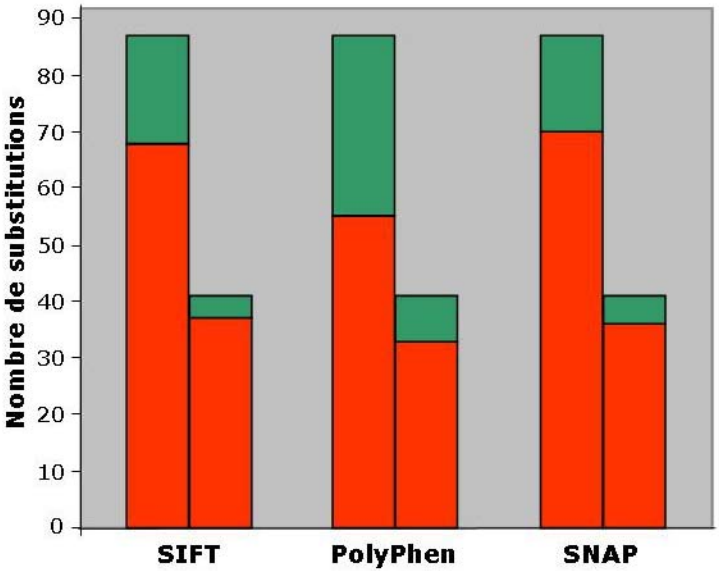


Figure 71 - Prédiction des effets délétères de substitutions par SIFT, PolyPhen et SNAP. En rouge, les substitutions prédites comme délétères, en vert celle prédites comme neutres. Pour chaque serveur, on retrouve à gauche, les prédictions obtenues pour les substitutions affectant la TNSALP et à droite, celles pour les substitutions affectant la myotubularine.

D'après ces résultats, il semblerait que ces outils de prédiction fournissent des résultats relativement robustes, en particulier SIFT et SNAP. Afin d'illustrer la difficulté d'interprétation de ces résultats, la Figure 72 présente les prédictions issues de SIFT pour l'ensemble des substitutions possibles dans chacune de nos protéines d'intérêt. On constate que sur les 9 956 substitutions envisageables le long de la TNSALP (524 positions, 19 substitutions possibles pour chacune), 6 605 sont prédites comme non-tolérées, soit les deux-tiers. Pour la myotubularine, sur les 11 457 substitutions possibles (603 positions, 19 substitutions à chaque position), 7495 sont prédites comme non-tolérées, soit la même proportion que pour la TNSALP.

Ces résultats sont difficilement exploitables par le biologiste. Prenons l'exemple des substitutions p.I490F, associée à une effet modéré, et p.P108L, associée à un effet sévère. SIFT prédit ces deux mutations comme délétères, sachant qu'à chacune de ces positions, 2 substitutions sont prédites commes tolérées et 17 comme non-tolérées. Les résultats expérimentaux ont montré que la protéine portant la mutation p.I490F conserve une activité résiduelle significative (37% de l'activité de la protéine sauvage) alors que la protéine portant la mutation p.P108L présente une activité résiduelle négligeable (0,6% de l'activité de la protéine sauvage). SIFT n'offre pourtant aucun élément permettant de différencier les effets de ces substitutions sur la protéine et masque ainsi complètement leurs effets potentiels, ce qui illustre bien la difficulté liée à l'interprétation des résultats issus de ce serveur. Devant la complexité de la situation, le biologiste ne pourra pas se servir de ces résultats pour établir une interprétation précise.

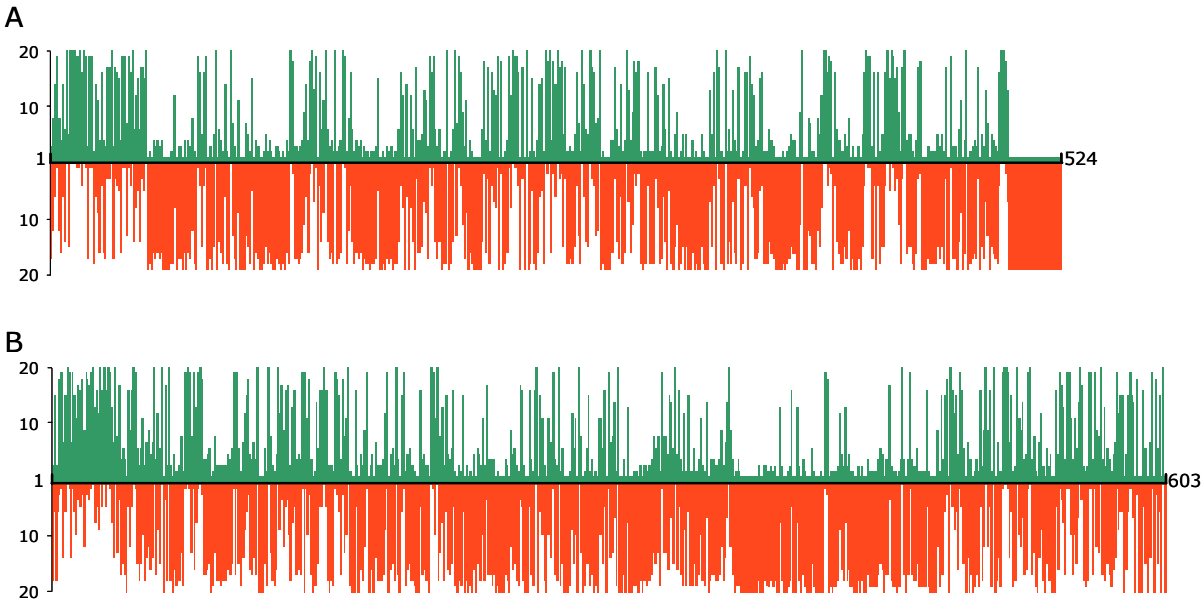


Figure 72 - Prédictions des effets des substitutions faux-sens par SIFT, en chaque position de la TNSALP (A) et de la myotubularine (B). En vert, le nombre de substitutions prédites comme tolérées à chaque position et en rouge le nombre de substitutions prédites comme non-tolérées.

14.3.3. Vers une typification automatisée des positions délétères

Dans le cadre de notre analyse, nous avons « mimé » une démarche automatisée afin de typifier les positions des mutations connues et de voir s'il est possible de discriminer, en première approche, les positions auxquelles des mutations sévères ou modérées sont associées.

La démarche que nous avons suivie se base schématiquement sur deux types de données : des données « connues » et des données « calculées ». Les données « connues » concernent aussi bien les données génotypiques et phénotypiques que les données récupérées automatiquement par MACSIMS (Thompson *et al.*, 2006), notamment celles relatives au découpage en domaines fonctionnels de nos protéines d'intérêt. Les données « calculées » correspondent, quant à elles, à la conservation des résidus au cours de l'évolution et aux modèles tridimensionnels générés pour chacune de nos protéines d'intérêt, un certain nombre de ces données étant disponibles au sein de MS2PH-db.

14.3.3.1. Données fonctionnelles

Nous avons récupéré un certain nombre de données fonctionnelles relatives à nos protéines d'intérêt dans MS2PH-db (Figure 73).

Aucune information relative au découpage en domaines fonctionnels de la TNSALP n'est disponible, néanmoins, 16 résidus représentant des sites « remarquables » ont été mis en évidence par MACSIMS (Figure 73 A) : 1 résidu du site actif, 9 résidus impliqués dans une

liaison métallique, qui participent à la catalyse et 6 sites de glycosylation. On remarque que les positions des mutations sont réparties assez uniformément le long de la protéine et que les positions situées à proximité des résidus participant de la catalyse sont systématiquement associées à des mutations sévères.

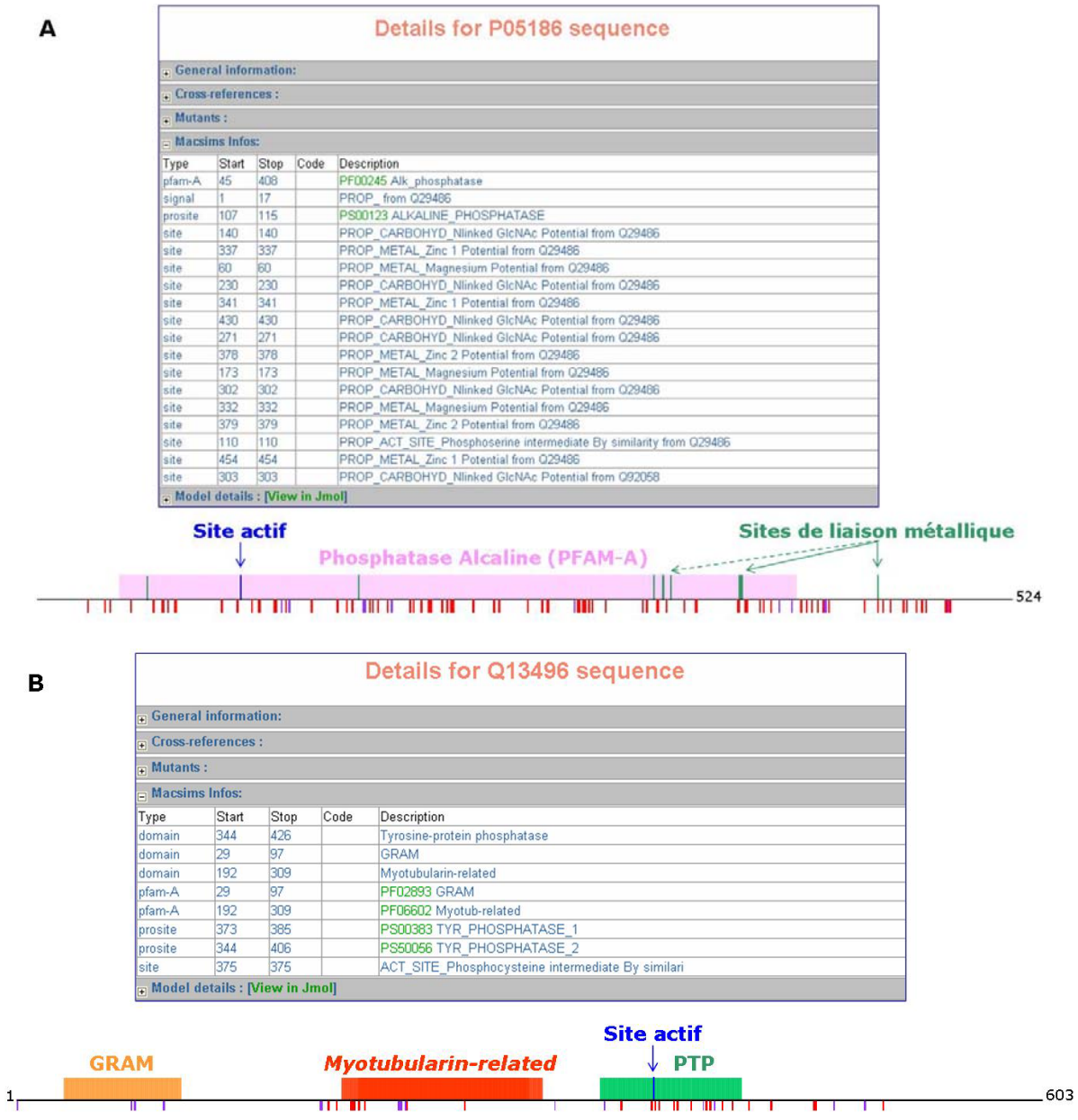


Figure 73 - Données fonctionnelles relatives à la TNSALP (A) et à la myotubularine (B) (extraites de MS2PH-db). Le schéma linéaire de chacune des protéines présente, en plus de l'emplacement des domaines fonctionnels annotés et des sites « remarquables », la localisation des substitutions dont l'effet est connu (en rouge, substitutions ayant un effet sévère ; en violet, substitutions ayant un effet modéré).

Trois domaines fonctionnels distincts ont été mis en évidence au sein de la myotubularine (Figure 73 B) : le domaine GRAM, le domaine *myotubularin-related* et le domaine tyrosine

phosphatase (PTP) qui contient le site actif. Les positions des mutations sont réparties le long de la protéine et on remarque que les mutations situées dans ou à proximité du domaine GRAM sont associées à un effet modéré, alors qu'on retrouve un enrichissement en mutations sévères dans le domaine PTP, en particulier à proximité du site actif.

14.3.3.2. Données évolutives

Nous avons récupéré, pour chacune de nos protéines d'intérêt, les deux alignements multiples fournis par MS2PH-db (*voir* section 9.4.1.2, p. 126), à savoir l'alignement des séquences eucaryotes homologues ainsi que l'alignement construit en appliquant la méthode d'échantillonnage des bandelettes après une recherche de similarité dans la banque UniRef90 (Suzek *et al.*, 2007). Ces alignements sont de bonne qualité, les scores NorMD variants entre 0,6 et 0,9.

Nous avons appliqué, sur chacun de nos alignements, les sept méthodes de détermination des classes de conservation, décrites dans la section 11.2.1, p. 167, basées sur la classification des scores de conservation par la méthode du modèle de mélange. Nous ne nous sommes cette fois plus limité à la détermination de deux classes, notre but étant de voir quelle méthode nous permet de mettre en évidence, au sein des classes de conservation, un grand nombre de positions mutées, en limitant le nombre total de colonnes.

Pour nos deux protéines d'intérêt, les meilleurs résultats ont été obtenus avec l'alignement des séquences homologues eucaryotes et la méthode de détermination des classes de conservation qui associe la méthode de la norme du vecteur moyen et celle du deltaG (*voir* Chapitre 7 - *Développements relatifs aux alignements multiples*, p. 108) (Tableau 11).

TNSALP			
	Nombre de colonnes	Mutation sévère	Mutation modérée
Classe 1	68	16	1
Classe 2	88	21	2
Classe 3	98	18	3
Classe 4	49	9	0
Classe 5	82	5	2
Classe 6	42	2	2
Classe 7	72	4	0
Classe 8	25	1	1

Myotubularine			
	Nombre de colonnes	Mutation sévère	Mutation modérée
Classe 1	88	15	5
Classe 2	121	6	6
Classe 3	210	3	4
Classe 4	151	1	0
Classe 5	32	0	1

Tableau 11 - Classes de conservation mises en évidence au sein des alignements multiples des séquences eucaryotes homologues de la TNSALP (à gauche) et de la myotubularine (à droite) et répartition des positions mutées dans chacune de ces classes. Les classes sont présentées par ordre décroissant de conservation.

Le nombre de classes de conservation est déterminé automatiquement par la méthode de classification des colonnes et est entièrement dépendant de l'alignement multiple, c'est pourquoi on n'obtient pas le même nombre de classes avec l'alignement de la TNSALP et celui de la myotubularine.

La Figure 74 présente la répartition des classes de conservation le long de nos séquences d'intérêt. On y constate que les résidus « remarquables » se situent dans des zones de forte conservation. Pour ce qui est de la myotubularine, on peut voir que le domaine PTP semble particulièrement bien conservé au cours de l'évolution, contrairement au domaine GRAM, qui a, *a priori*, subi une pression de sélection moins forte.

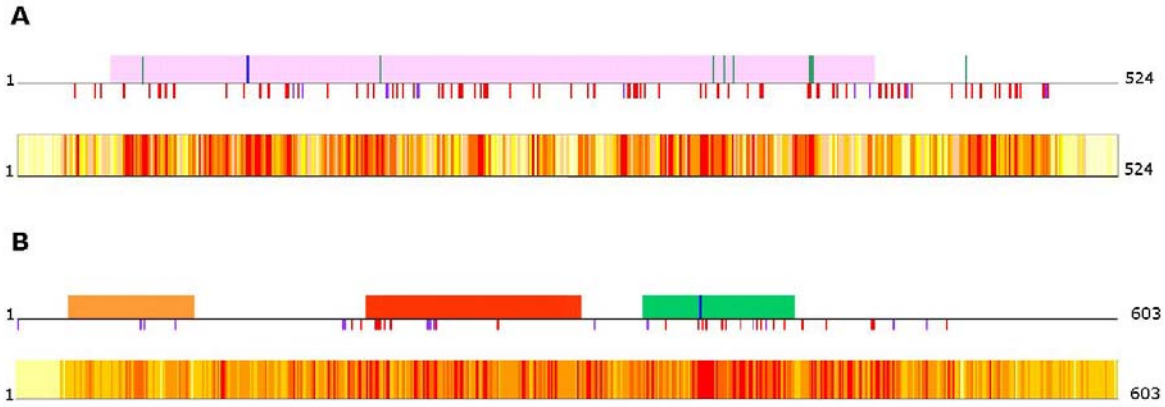


Figure 74 - Conservation le long de la TNSALP (A) et de la myotubularine (B). Le spectre de la conservation est visualisable pour chacune des protéines sous sa schématisation en domaines fonctionnels : du rouge au jaune pâle, des résidus les mieux conservés aux résidus les moins bien conservés.

Les positions des mutations se répartissent dans l'ensemble des classes de conservation, qu'elles soient sévères ou modérées. On peut néanmoins constater que, pour la TNSALP, les positions de 55 des 76 mutations sévères et de 6 des 11 modérées sont regroupées dans les trois classes les mieux conservées. De même, pour la myotubularine, 21 des 25 positions associées à des mutations sévères et 11 des 16 modérées sont contenues dans les deux meilleures classes de conservation. Les positions délétères semblent ainsi avoir une tendance à la conservation.

On peut enfin remarquer, pour les deux protéines, un enrichissement en substitutions sévères dans la classe 1, mais force est de constater que la conservation, même hiérarchisée, ne suffit pas à elle seule pour discriminer les positions sévères et modérées.

14.3.3.3. Données structurales

Nous avons récupéré, pour nos deux protéines d'intérêt, les modèles tridimensionnels générés dans MS2PH-db.

Le modèle structural de la TNSALP a été construit en utilisant comme empreinte la structure de la phosphatase alcaline placentaire (PDB : 1ZEF). Ces deux séquences partagent 57% d'identité et le modèle tridimensionnel couvre 93% de la longueur de la TNSALP (des résidus 18 à 503, sur les 524 totaux).

Le modèle structural de la myotubularine a été construit en utilisant comme empreinte la structure de la *myotubularin-related protein 2* (MTMR2) humaine (PDB : 1ZVR). Ces deux séquences partagent 69% d'identité et le modèle tridimensionnel couvre 85% de la longueur de la myotubularine (des résidus 32 à 544 sur 603).

Les substitutions semblent réparties sur l'ensemble du modèle structural pour la TNSALP, sans distinction évidente entre les mutations sévères et modérées (Figure 75 A). Pour la myotubularine, on observe quelques zones enrichies en mutations sévères, notamment au niveau d'un regroupement de quatre hélices α, qui semblent former le cœur du repliement (Figure 75 B), alors que les mutations modérées se retrouvent principalement en périphérie.

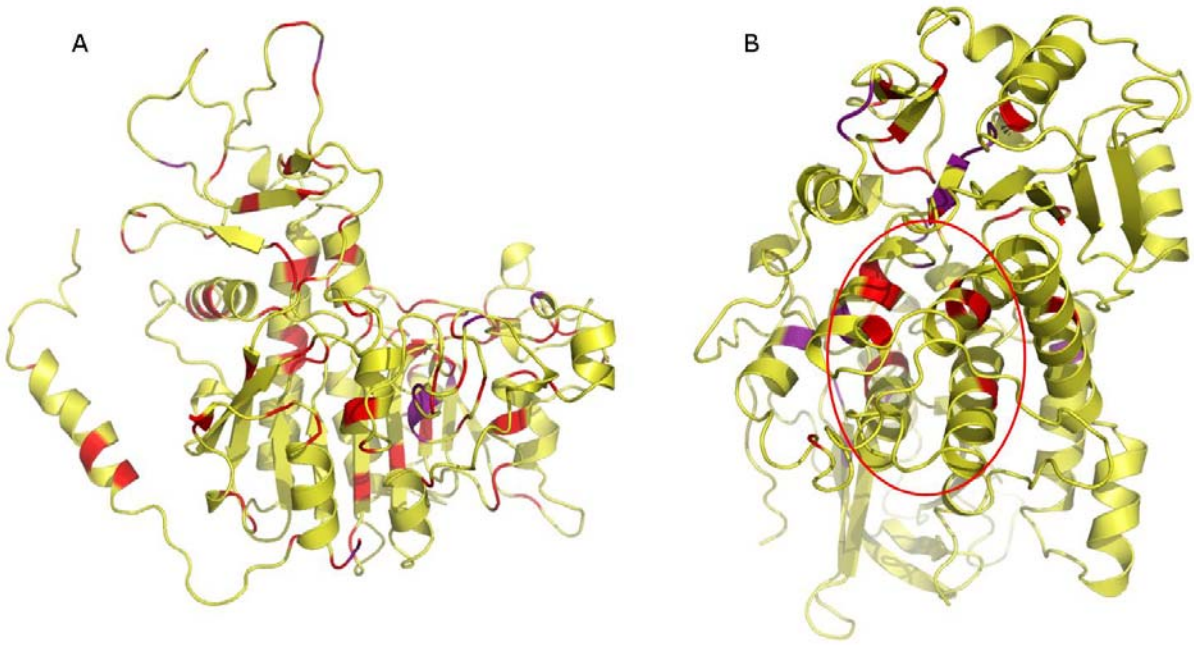


Figure 75 - Localisation des mutations sur les modèles structuraux de la TNSALP (A) et de la myotubularine (B). En rouge, les positions des mutations sévères ; en violet les positions des mutations modérées. Le cercle rouge indique une zone enrichie en mutations sévères sur le modèle structural de la myotubularine.

D'après la littérature (Mornet *et al.*, 2001), le modèle tridimensionnel de la TNSALP peut être divisé en trois domaines structuraux : le domaine appelé « domaine en couronne », le domaine de liaison au calcium et le reste de la protéine. Seul le domaine en couronne est prédit de manière automatique par le programme *Domain3D* (Taylor, 1999).

Sur les 87 mutations étudiées, 13 sont localisées dans le domaine en couronne (10 sévères, 3 modérées) et 14 dans le domaine de liaison au calcium (13 sévères, 1 modéré), qui présente un enrichissement en mutations sévères, attestant de l'importance de ce domaine au sein de la protéine (Figure 76).

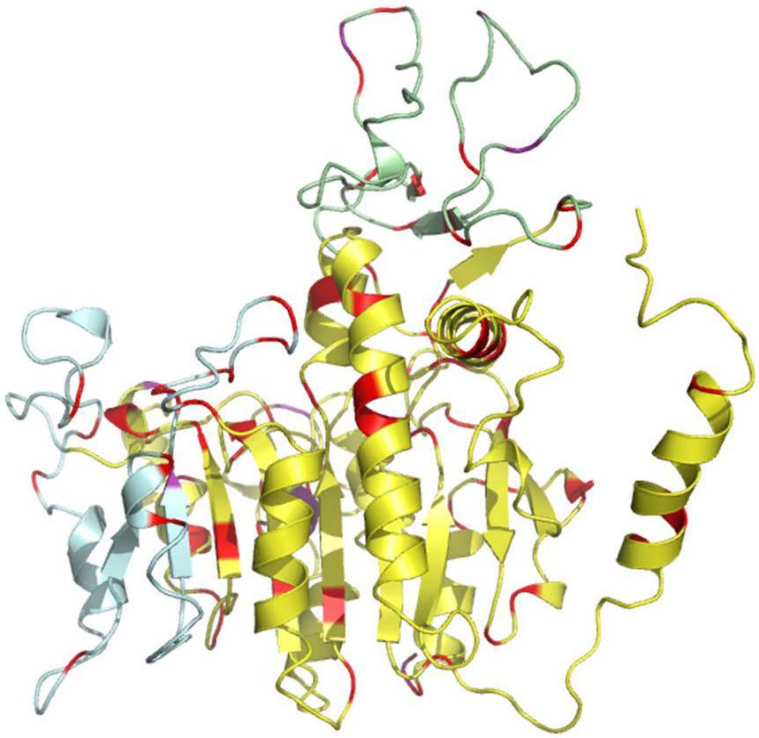


Figure 76 - Découpage du modèle tridimensionnel de la TNSALP en domaines structuraux et localisation des mutations. En vert, le domaine en couronne ; en bleu le domaine de liaison au calcium ; en rouge les positions des mutations sévères ; en violet les positions des mutations modérées.

Les positions des éléments de structures secondaires ont également été extraites de MS2PH-db. La TNSALP est constituée de 20 hélices α et de 11 feuillets β, représentant respectivement 143 et 55 résidus. La myotubularine est, quant à elle, constituée de 25 hélices α et de 14 feuillets β, représentant respectivement 191 et 69 résidus.

Pour la TNSALP, 19 mutations sont localisées au sein d'une hélice α (16 sévères et 3 modérées) et 10 au sein d'un feuillet β (9 sévères, 1 modérée). Pour la myotubularine, 13 mutations sont localisées au sein d'une hélice α (8 sévères et 5 modérées) et 7 au sein d'un feuillet β (4 sévères, 3 modérées).

La localisation de la position de la mutation au sein d'un élément de structure secondaire ne représente donc pas un critère discriminant en tant que tel. La Figure 77 nous montre néanmoins que, pour la myotubularine, les quatre hélices α pointées dans la Figure 75

comme enrichies en mutations sévères (hélices α 14, 16, 18 et 19) se situent dans des zones de forte conservation.

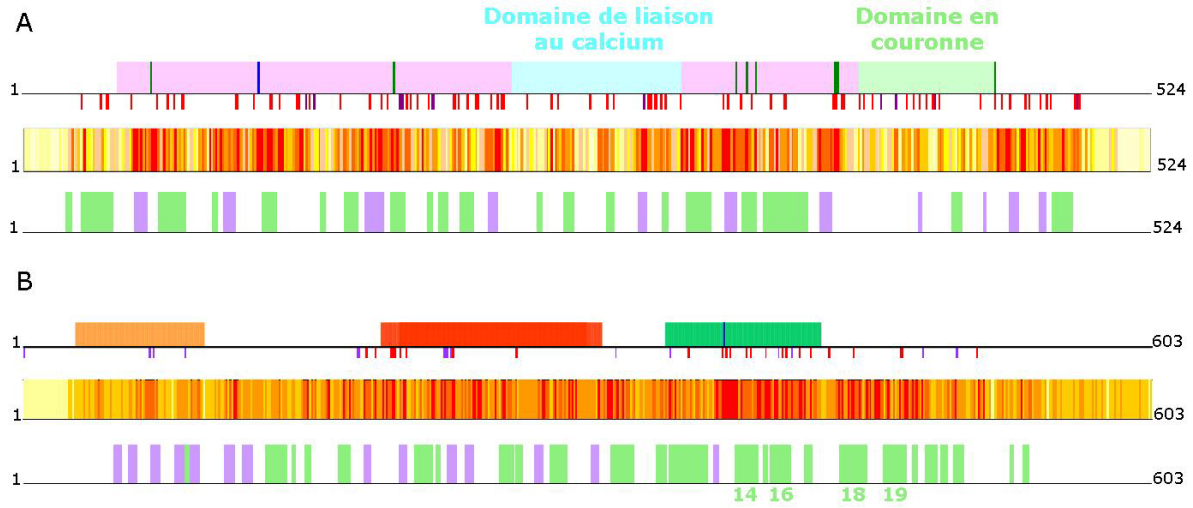


Figure 77 - Eléments de structure secondaire le long de la TNSALP (A) et de la myotubularine (B). Les hélices α sont représentées en vert clair, les feuillets β en violet.

Nous avons par la suite évalué l'accessibilité à la surface des résidus de chacun des modèles, par l'intermédiaire du programme NACCESS (Hubbard and Thornton, 1993). Nous avons classé les résidus en deux groupes : les résidus enfouis, pour lesquels moins de 5% de la surface de référence est accessible et les résidus « de surface » (Miller *et al.*, 1987).

Pour le modèle de la TNSALP, 161 résidus ont été prédits comme enfouis et 328 comme étant de surface. Pour le modèle de la myotubularine, 152 résidus ont été prédits comme enfouis et 361 comme étant de surface. Pour la TNSALP, 36 mutations sévères sont enfouies, 40 sont à la surface ; 5 mutations modérées sont enfouies, 6 sont à la surface. De même, pour la myotubularine, 14 mutations sévères sont enfouies, 11 sont à la surface alors que 11 mutations modérées sont enfouies et 4 à la surface. L'accessibilité à la surface des résidus ne représente par conséquent pas non plus un critère discriminant en tant que tel.

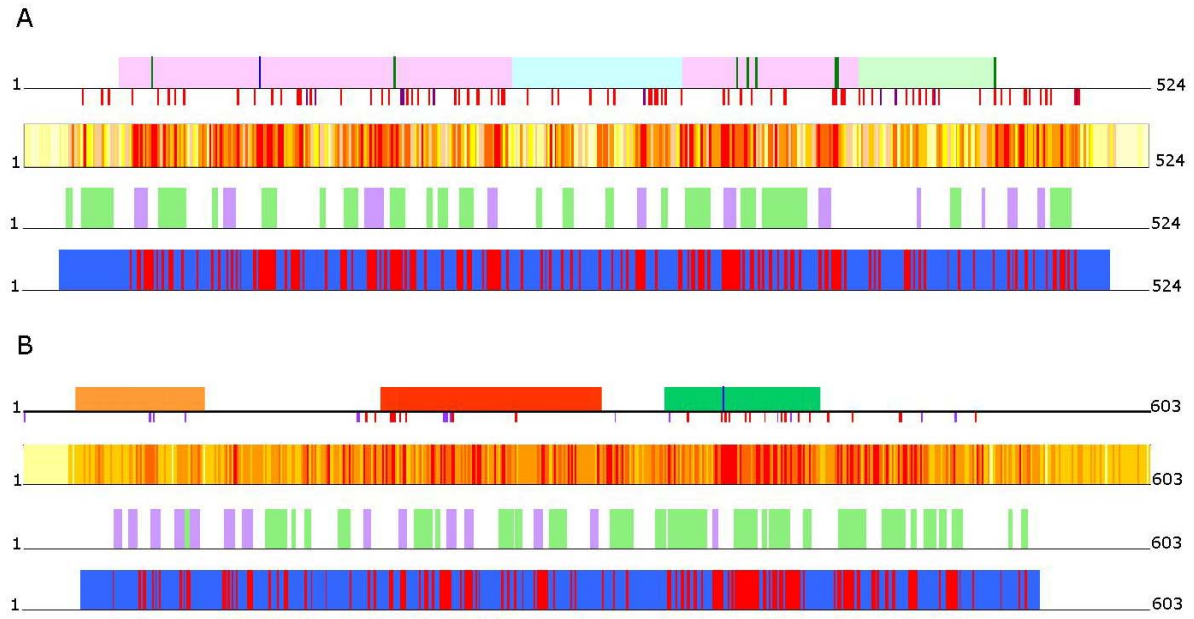


Figure 78 - Acessibilité à la surface des résidus constituant la TNSALP (A) et la myotubularine (B). En rouge les résidus prédits comme étant enfouis, en bleu ceux prédits comme étant accessibles au solvant.

On peut toutefois remarquer (Figure 78) que les hélices 14, 16, 18 et 19 du modèle de la myotubularine se situent principalement dans le cœur de la protéine.

Enfin, nous avons mis en evidence les positions mutées situées à proximité de sites « remarquables », grâce à la disponibilité des modèles structuraux (Figure 73). On constate (Figure 79) que, pour la TNSALP tout comme pour la myotubularine, les trois mutations situées à moins de 6 Å des sites actifs sont sévères. Pour la TNSALP, toutes les mutations situées à moins de 6 Å des sites de liaison métallique sont sévères, à l'exception de deux substitutions très conservatives (p.A176T et p.A177T).

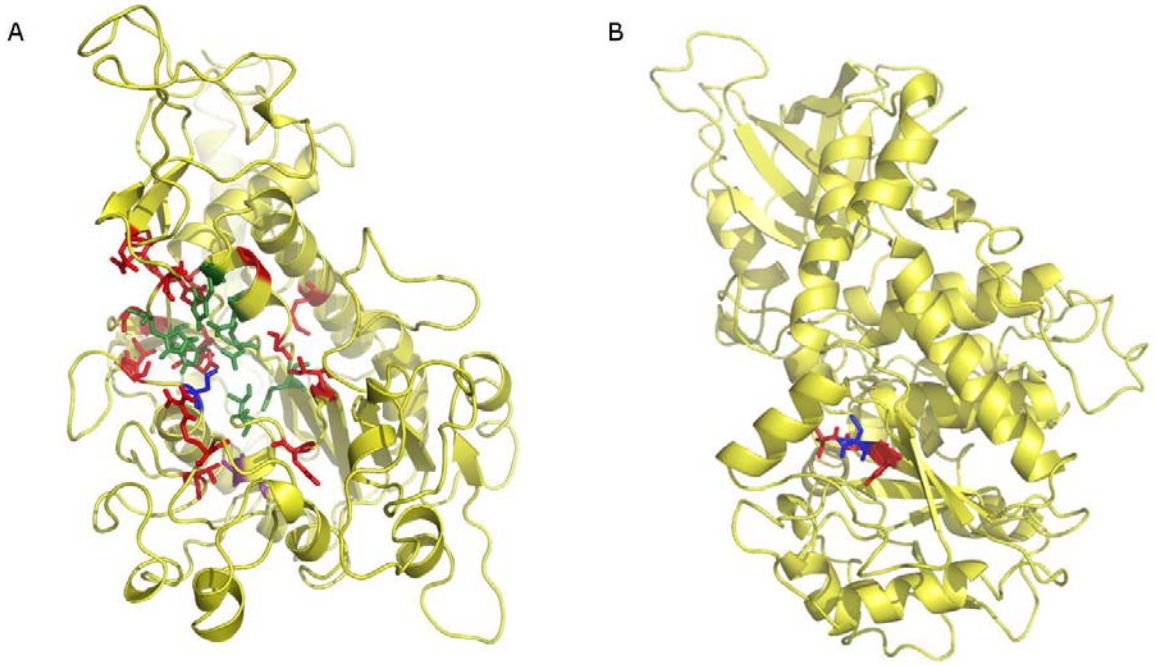


Figure 79 - Proximité des mutations par rapport aux sites « remarquables » pour la TNSALP (A) et la myotubularine (B). En bleu, le site actif ; en vert les sites de liaison métallique ; en rouge, la localisation des mutations sévères ; en violet, la localisation des mutations modérées.

Ainsi, l'ensemble des données que nous avons traité de manière indépendante ne fournit pas suffisamment d'informations pour permettre une discrimination efficace des positions auxquelles des mutations délétères ou modérées sont associées.

Le cumul de ces données semble plus informatif et quelques règles simples commencent à émerger : la proximité d'un site impliqué dans la catalyse entraîne une plus forte concentration en mutations sévères. Des éléments de structure secondaire enfouis et bien conservés au cours de l'évolution semblent également être de bons candidats pour caractériser des mutations sévères. On remarque aussi, en ce qui concerne la myotubularine, que selon le domaine fonctionnel affecté, les conséquences des mutations diffèrent : elles sont systématiquement modérées quand elles affectent le domaine GRAM, situé dans la partie N-terminale de la protéine, qui est faiblement conservé au cours de l'évolution.

La localisation de la mutation n'est cependant pas le seul paramètre à typifier. En effet, la substitution en elle-même est essentielle.

14.3.3.4. Caractérisation des substitutions

Nous avons enfin caractérisé les substitutions en nous basant sur une classification des acides aminés, dérivée de la « rosace des acides aminés », présentée dans la Figure 42, p. 111.

Schématiquement, six groupes d'acides aminés en résultent : les acides aminés à petit volume (groupe 1 : P, A, G, S et T), les acides aminés polaires (groupe 2 : D, E, N et Q), les

acides aminés basiques (groupe 3 : K, R et H), les acides aminés aromatiques (groupe 4 : F, Y et W), les acides aminés aliphatiques (groupe 5 : V, I, L et M) et la cystéine C qui est l'unique représentant du groupe 6.

Dans le cas de la TNSALP, on observe que 50 des 76 substitutions entraînant un effet sévère sont le résultat d'un changement de groupe tandis que seulement 6 des 11 substitutions à effet modéré correspondent à un changement de groupe.

Dans le cadre de la myotubularine, la situation est beaucoup plus tranchée puisque toutes les substitutions associées à un effet sévère entraînent un changement de groupe, ainsi que 10 des 15 modérées.

Il semble donc clair que la nature de la modification, et notamment le caractère « drastique » de la mutation (abordé, dans cette étude, par le critère de mutations entraînant un changement de groupe physico-chimique des acides aminés), peut constituer un outil précieux de typification. Cependant, là encore, la typification de la substitution seule n'est pas suffisante et nécessitera des études plus poussées impliquant, sans doute, une meilleure caractérisation de l'environnement structural de la mutation et une analyse des différentes liaisons interatomiques.

Pour certaines positions, des substitutions associées à différents effets sont observées, comme l'illustrent les mutations p.H171Y et p.H171R qui affectent la TNSALP et qui ont respectivement un effet sévère et un effet modéré. La substitution p.H171Y engendre le changement d'un résidu basique par un résidu aromatique, alors que p.H171R est très conservative, ce qui peut expliquer la différence d'effet. Le raisonnement ne peut cependant pas toujours s'appliquer comme l'illustrent les substitutions p.M62I et p.M62L, qui correspondent toutes deux à des mutations n'entraînant pas de changement de groupe et qui sont pourtant à l'origine d'effets sévère (p.M62I) et modéré (p.M62L).

14.3.3.5. Effets cliniques liés à un génotype

Comme nous l'avons déjà évoqué, dans le cadre de la myotubularine, la sévérité du phénotype clinique peut être directement estimée à partir de la sévérité d'une mutation. Dans le cadre de la TNSALP, le problème est bien plus compliqué, l'émergence d'un phénotype clinique étant liée à la présence de deux allèles et aux relations qu'ils partagent, d'autant plus que les deux modes de transmission (récessif et dominant) ont été observés pour l'hypophosphatasie.

La banque de données spécifique de l'hypophosphatasie sur laquelle nous nous sommes basés (http://www.sesep.uvsq.fr/Database.html) contient des données relatives à 72 individus hétérozygotes doubles faux-sens, 21 individus homozygotes faux-sens et 6 individus portant un allèle faux-sens dominant.

De manière surprenante, 20 des 21 individus homozygotes déclarent un phénotype sévère. Ceci laisse supposer que l'homozygotie présente une tendance à la sévérité. Au regard des positions de ces mutations, qui se répartissent dans toutes les régions de la TNSALP, tant du point de vue de la séquence que de la structure, il est difficile d'expliquer ce phénomène. Une hypothèse reposerait sur l'organisation dimérique de l'enzyme et l'existence d'un mécanisme d'allostérie (Hoylaerts *et al.*, 1997), qui pourraient entraîner une amplification systématique de l'effet d'une mutation, même modérée, si cette dernière est présente dans les deux monomères. Ce phénomène d'amplification n'interviendrait pas dans le cas des individus hétérozygotes, qui présentent des phénotypes sévères ou modérés.

Le positionnement des mutations dominantes sur le modèle du dimère (Figure 80) nous montre qu'à l'exception d'une mutation située dans le domaine de liaison au calcium, toutes sont localisées au niveau de l'interface d'homodimérisation. Ceci laisse supposer que la dominance peut être liée à l'allostérie : la présence d'une mutation à l'interface serait à même d'entrainer des perturbations dans la voie de communication entre les deux monomères et d'affecter par conséquent l'activité de l'enzyme.

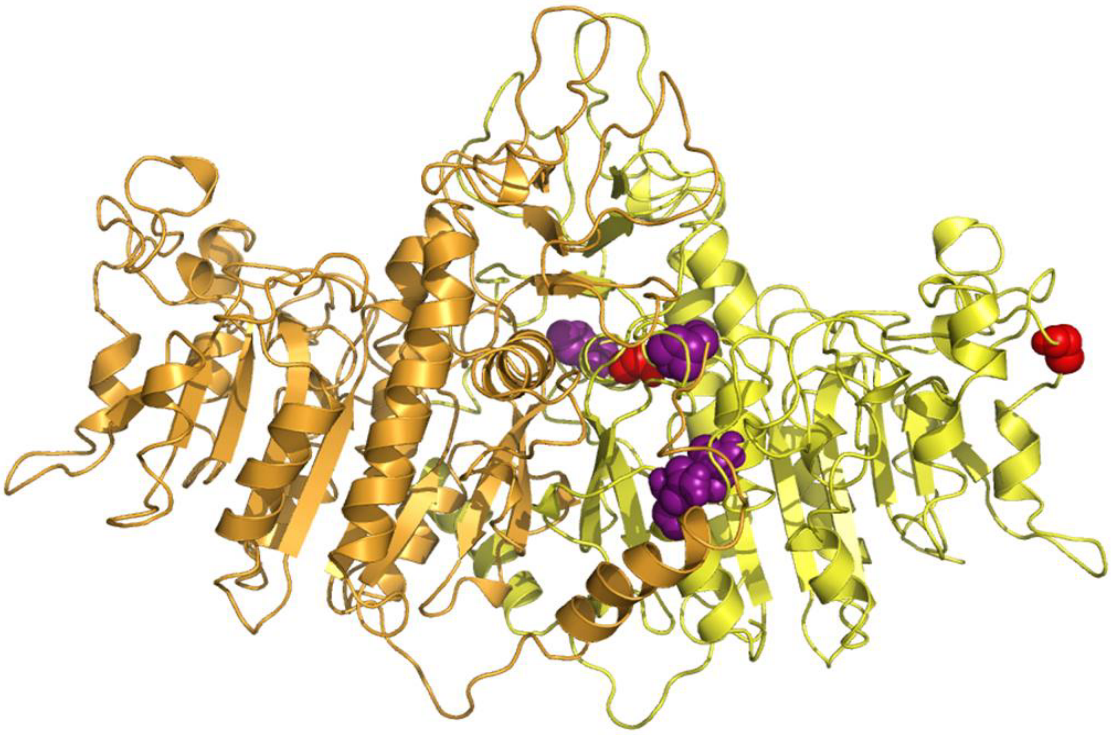


Figure 80 - Localisation des mutations dominantes dans la TNSALP, présentée sous sa forme dimérique.

Ainsi, ces observations nous montrent que les propriétés d'oligomérisation de la protéine ont une influence primordiale sur la manière d'exploiter les données génotypiques, notamment les données d'homo/hétérozygotie.

14.3.4. Vers la mise en place d'un système expert

Les tableaux en Annexes 5 et 6 récapitulent l'ensemble des données présentées dans ce chapitre.

Notre analyse préliminaire a permis de mettre en évidence de nombreuses difficultés liées à la typification des mutations et à la prédiction des phénotypes qui y sont associés. Nous avons pu constater qu'aucun des paramètres testés ne permet, à lui seul, une discrimination des mutations ayant un effet sévère de celles ayant un effet modéré. Ceci nous conforte dans l'idée qu'il est nécessaire de cumuler plusieurs niveaux d'informations pour comprendre au mieux les effets d'une mutation. Quelques règles ont néanmoins pu être mises en évidence, comme l'importance de la proximité d'un résidu impliqué dans la catalyse, ainsi que l'association de données de conservation, de structures secondaires et d'accessibilité à la surface qui semblent fournir une bonne indication de la sévérité des mutations. Le découpage en domaines de la protéine d'intérêt et la caractérisation de la substitution en tant que telle paraissent également essentiels à une meilleure compréhension des effets des mutations.

Pour avancer dans la mise en place d'un système expert de typification automatique des mutations et de prédiction de leurs effets, nous devrons, d'une part, multiplier les systèmes étudiés et, d'autre part, ajouter des niveaux d'informations supplémentaires, comme des données d'oligomérisation et d'interactomique. Ces dernières représentent, à notre sens, des sources d'informations particulièrement intéressantes qui nous permettront notamment de mieux caractériser les surfaces des protéines d'intérêt.

Les difficultés liées à la typification des mutations au niveau protéique sont attribuables à plusieurs paramètres, notamment l'ensemble des facteurs environnementaux qui influence l'expression d'un phénotype. Cependant, il faut aussi garder à l'esprit qu'une modification délétère dans la région codante d'un gène peut ne pas affecter directement la structure ou la fonction d'une protéine mais avoir pour effet, par exemple, de modifier la dynamique du temps de demi-vie de l'ARNm ou d'être à l'origine de problèmes traductionnels. Une réflexion s'impose par conséquent sur les objectifs du système expert à mettre en place. Peut-être serait-il plus judicieux, pour la communauté scientifique, de disposer d'un système à même de fournir l'accès à un ensemble d'informations pertinentes, couplant des données de séquence, de structure, de fonction, etc. à des données cliniques pour que les spécialistes d'une protéine ou d'une maladie puissent les confronter à leurs connaissances pour l'évaluation finale des conséquences des mutations. Le pouvoir prédictif de ce système devrait peut-être se focaliser plus particulièrement sur une élimination robuste de possibilités, en montrant par exemple qu'aucun effet structural et/ou fonctionnel n'est associé à une mutation et que si effet délétère il y a, celui-ci n'est pas lié à une altération des fonctions de la protéine mais plutôt à un défaut dans sa production.

Conclusions & Perspectives

Chapitre 15 - Conclusions et perspectives

Les travaux que nous avons réalisés au cours de cette thèse ont contribué à mettre en évidence les forces et les faiblesses des analyses et des développements bioinformatiques relatifs à la gestion et à l'analyse des données, dans le contexte d'approches à haut débit. Nous avons également montré l'importance de la bioinformatique dans des études appliquées à la recherche biologique et biomédicale. Dans le cadre du projet MS2PH, ces démonstrations ont été réalisées en abordant, de manière globale, les mutations affectant les protéines impliquées dans les maladies monogéniques humaines.

Les outils que nous avons conçus ont pour objectif de permettre une meilleure compréhension des relations qui lient un génotype aux phénotypes moléculaires et cliniques associés, au travers d'approches combinant plusieurs niveaux d'informations.

Dans un premier temps, le développement du serveur Web MAGOS nous a permis de fournir un espace de travail au sein duquel il est possible d'analyser, de manière couplée, les données de séquence, de structure, de fonction et d'évolution relatives à une protéine d'intérêt. Dans un second temps, nous avons mis à profit les capacités intégratives de ce serveur dans le cadre du développement de la banque de données MS2PH-db, qui associe, en plus des données citées ci-dessus, des données « cliniques » (mutations et phénotypes associés). MS2PH-db, en regroupant une multitude d'informations prétraitées relatives à l'ensemble des protéines impliquées dans les maladies monogéniques humaines, constitue une bonne illustration de la puissance des études bioinformatiques à haut débit. Le développement et la mise à jour de cette banque de données reposent sur la grille de calcul universitaire Décrypthon, au déploiement de laquelle nous avons participé. Notons que celle-ci représente l'une des premières grilles européennes couplant puissance de calcul et fédération de banques de données, qui soit dédiée aux études appliquées dans le domaine des sciences du vivant. La généralisation des traitements bioinformatiques à haut débit, qui nécessitent l'intégration de connaissances toujours plus nombreuses et hétérogènes et dont les besoins en termes de calcul deviennent de plus en plus lourds, implique que l'utilisation des grilles de calcul dans ce domaine n'en est qu'à ses débuts et qu'elle est promise à un bel avenir.

MS2PH-db permet d'ores et déjà d'aborder plusieurs niveaux d'informations relatives à une protéine (séquence, structure, fonction, évolution), couplés à des données génotypiques et phénotypiques. Cette banque de données ne constitue cependant que le socle de l'infrastructure finale qui sera déployée dans le cadre du projet MS2PH et est, de ce fait,

vouée à évoluer rapidement. Des données supplémentaires seront intégrées, dans l'optique de mettre à la disposition des utilisateurs l'information la plus complète et la plus pertinente possible pour estimer au plus juste l'impact phénotypique d'une mutation d'intérêt. Ces données seront déterminées sur la base des analyses effectuées sur un nombre restreint de protéines bien documentées. Les résultats préliminaires obtenus sur les systèmes de la myotubularine et de la phosphatase alcaline non tissu-spécifique nous confortent dans l'idée qu'il est primordial de conjuguer plusieurs niveaux d'informations pour atteindre un degré de compréhension plus fin du lien existant entre les variations génétiques et leurs conséquences phénotypiques. De plus, des modèles de données phénotypiques sophistiqués seront nécessaires pour renforcer les analyses informatiques et devront faire un usage rigoureux d'ontologies phénotypiques exhaustives. Nous envisageons également le développement d'une fonction de score, qui intégrera des données évolutives, fonctionnelles et structurales, pour la prédiction automatique des conséquences structurales et phénotypiques d'une mutation.

Par ailleurs, nous avons pu constater, au cours de nos travaux, que le biologiste se retrouve fréquemment confronté à deux problèmes récurrents et antinomiques, à savoir, d'une part, une abondance de données biologiques à gérer, notamment pour les données de séquences et, d'autre part, un manque d'informations précises et homogènes concernant des données plus spécialisées.

L'abondance de données est d'autant plus problématique que leur qualité est tout à fait variable : leur production se base couramment sur des technologies et sur des études à haut débit (séquençage, annotation structurale ou fonctionnelle, etc.) au cours desquelles les erreurs générées peuvent s'accumuler, jusqu'à masquer le signal biologique. Une nouvelle réflexion doit d'être menée sur le traitement des données en amont de leur analyse, la gestion de l'ensemble des informations disponibles étant devenue contre-productive. Il est dorénavant primordial d'établir des stratégies de tri des informations non seulement pour que le biologiste puisse restreindre son espace de recherche à des données de bonne qualité et pertinentes dans le cadre de ses études, mais aussi pour que le message biologique recherché ne soit pas noyé. Ces idées sont transversales à l'ensemble des études menées en sciences du vivant et nous nous sommes concentrés sur cette problématique dans le cadre de la construction d'alignements multiples de séquences protéiques. Nous avons développé une stratégie de sélection des séquences à aligner permettant de maintenir l'information structurale pertinente au sein des alignements ; il est cependant clair que d'autres stratégies de réduction de l'information initiale sont à imaginer pour aborder des problématiques plus complexes.

A l'inverse, on déplore une pénurie d'informations précises et homogènes concernant des données plus spécifiques, comme les données génotypiques et phénotypiques. Les

connaissances sont dispersées au travers de publications scientifiques ou au sein de laboratoires spécialisés et un effort doit être fait pour les homogénéiser et les rendre accessibles. L'utilisation d'ontologies spécifiques représente, a priori, une réponse appropriée qui permettra une « simplification » de l'information pour l'homme, tout en la rendant plus précise et, surtout, plus exploitable par l'outil informatique.

La compréhension automatique des relations existant entre le génotype d'un individu et son phénotype restera sans doute longtemps au-delà des moyens d'investigations de la communauté scientifique. Il est cependant évident qu'à long terme, le système expert envisagé, pour allier efficacité et robustesse, devra être à même de fournir un panel d'informations relatives aux protéines d'intérêt couvrant l'ensemble des « omiques », allant de la transcriptomique à l'interactomique, en passant par la protéomique et la métabolomique. L'intégration de l'ensemble de ces données permettra de connecter de manière plus efficace les différences génotypiques aux conséquences phénotypiques, et participera sans doute à une meilleure compréhension du lien génotype/phénotype et à la mise en place d'une médecine plus individualisée.

Annexes

Annexe 1 - Exemple d'alignement multiple au format FASTA

```
>Q1HE14
MALSSQIWAACLLLLLLLASLTSGSVFPQQTGQLAELQPQDRAGAR-ASWMPMFQ--RRR
RRDTHFPICIFCCGCCHRSKCGMCCKT
>P81172
MALSSQIWAACLLLLLLLASLTSGSVFPQQTGQLAELQPQDRAGAR-ASWMPMFQ--RRR
RRDTHFPICIFCCGCCHRSKCGMCCKT
>Q5NVR8
MALSSQIWAACLLLLLLLASLTSGSVFPQQTGQLAELQPQDRAGAR-AGWTPMLQ--RRR
RRDTHFPIYIFCCGCCHRSKCGMCCKT
>Q2NKT0
MALNTQIRATCLLLLVLLS-LTSGSVLPPQTRQLTDLQTKDTAGAA-AGLTPVLQ--RRR
R-DTHFPICIFCCGCCRKGTCGMCCRT
>Q8MJ80
MALSVQIRAACLLLLLLVS-LTAGSVLPSQTRQLTDLRTQDTAGAT-AGLTPVAQ--RLR
R-DTHFPICIFCCGCCRKAICGMCCKT
>Q5U9D2
MALSTRIQAACLLLLLLAS-VASVSVLPHQTGQLTDLRAQDTAGAE-AGLQPTLQLRRLR
RRDTHFPICIFCCGCCKTPKCGLCCKT
>Q49S92
MALSSQTQAACLLLLLLAS-VASVSVLPHQTGQLTDLRAQDTAGAE-AGLQPTLQLRRLR
RRDTHFPICIFCCGCCKTPKCGLCCIT
>Q9EQ21
MALSTRTQAACLLLLLLAS-LSSTTYLHQQMRQTTELQPLHGEESR-ADIAIPMQ--KRR
KRDTNFPICIFCCKCCNNSQCGICCKT
>Q80T19
MALSTRTQAACLLLLLLAS-LSSTTYLQQQMRQTTELQPLHGEESR-ADIAIPMQ--KRR
KRDINFPICRFCCQCCNKPSCGICCEE
>Q99MH3
MALSTRIQAACLLLLLLAS-LSSGAYLRQQTRQTTALQPWHGAESKTDDSALLML--KRR
KRDTNFPICLFCCKCCKNSSCGLCCIT
```

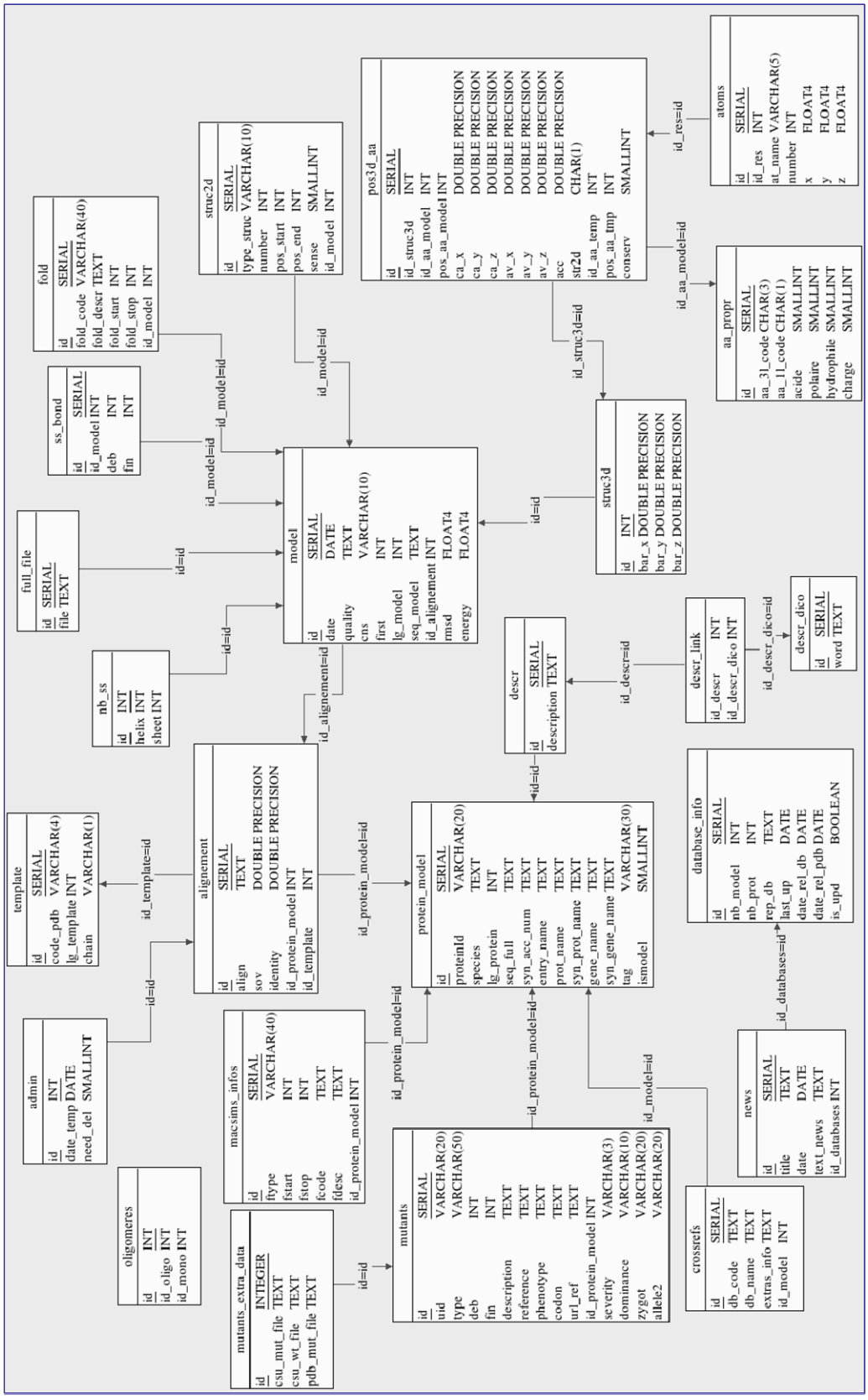
Annexe 2 - Exemple d'alignement multiple au format MSF (*Multiple Sequence Format*)

```
   MSF:   87  Type: P    Check:  8156   ..

 Name: Q1HE14 oo  Len:   87  Check:  5298  Weight:  6.0
 Name: P81172 oo  Len:   87  Check:  5298  Weight:  6.0
 Name: Q5NVR8 oo  Len:   87  Check:  5655  Weight:  6.0
 Name: Q2NKT0 oo  Len:   87  Check:  5566  Weight:  13.0
 Name: Q8MJ80 oo  Len:   87  Check:  5057  Weight:  13.0
 Name: Q5U9D2 oo  Len:   87  Check:  7777  Weight:  8.0
 Name: Q49S92 oo  Len:   87  Check:  7785  Weight:  9.0
 Name: Q9EQ21 oo  Len:   87  Check:  4818  Weight:  10.0
 Name: Q80T19 oo  Len:   87  Check:  4505  Weight:  15.0
 Name: Q99MH3 oo  Len:   87  Check:  6397  Weight:  14.0

//


Q1HE14          MALSSQIWAA CLLLLLLLAS LTSGSVFPQQ TGQLAELQPQ DRAGAR.ASW
P81172          MALSSQIWAA CLLLLLLLAS LTSGSVFPQQ TGQLAELQPQ DRAGAR.ASW
Q5NVR8          MALSSQIWAA CLLLLLLLAS LTSGSVFPQQ TGQLAELQPQ DRAGAR.AGW
Q2NKT0          MALNTQIRAT CLLLLVLLS. LTSGSVLPPQ TRQLTDLQTK DTAGAA.AGL
Q8MJ80          MALSVQIRAA CLLLLLLVS. LTAGSVLPSQ TRQLTDLRTQ DTAGAT.AGL
Q5U9D2          MALSTRIQAA CLLLLLLAS. VASVSVLPHQ TGQLTDLRAQ DTAGAE.AGL
Q49S92          MALSSQTQAA CLLLLLLAS. VASVSVLPHQ TGQLTDLRAQ DTAGAE.AGL
Q9EQ21          MALSTRTQAA CLLLLLLAS. LSSTTYLHQQ MRQTTELQPL HGEESR.ADI
Q80T19          MALSTRTQAA CLLLLLLAS. LSSTTYLQQQ MRQTTELQPL HGEESR.ADI
Q99MH3          MALSTRIQAA CLLLLLLAS. LSSGAYLRQQ TRQTTALQPW HGAESKTDDS


Q1HE14          MPMFQ..RRR RRDTHFPICI FCCGCCHRSK CGMCCKT
P81172          MPMFQ..RRR RRDTHFPICI FCCGCCHRSK CGMCCKT
Q5NVR8          TPMLQ..RRR RRDTHFPIYI FCCGCCHRSK CGMCCKT
Q2NKT0          TPVLQ..RRR R.DTHFPICI FCCGCCRKGT CGMCCRT
Q8MJ80          TPVAQ..RLR R.DTHFPICI FCCGCCRKAI CGMCCKT
Q5U9D2          QPTLQLRRLR RRDTHFPICI FCCGCCKTPK CGLCCKT
Q49S92          QPTLQLRRLR RRDTHFPICI FCCGCCKTPK CGLCCIT
Q9EQ21          AIPMQ..KRR KRDTNFPICI FCCKCCNNSQ CGICCKT
Q80T19          AIPMQ..KRR KRDINFPICR FCCQCCNKPS CGICCEE
Q99MH3          ALLML..KRR KRDTNFPICL FCCKCCKNSS CGLCCIT
```

Annexe 3 - La matrice de substitution BLOSUM 62. (extrait de carverlab.org/testing/epp.shtml)

	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	
C	9																				C
S	-1	4																			S
T	-1	1	5																		T
P	-3	-1	-1	7																	P
A	0	1	0	-1	4																A
G	-3	0	-2	-2	0	6															G
N	-3	1	0	-2	-2	0	6														N
D	-3	0	-1	-1	-2	-1	1	6													D
E	-4	0	-1	-1	-1	-2	0	2	5												E
Q	-3	0	-1	-1	-1	-2	0	0	2	5											Q
H	-3	-1	-2	-2	-2	-2	1	-1	0	0	8										H
R	-3	-1	-1	-2	-1	-2	0	-2	0	1	0	5									R
K	-3	0	-1	-1	-1	-2	0	-1	1	1	-1	2	5								K
M	-1	-1	-1	-2	-1	-3	-2	-3	-2	0	-2	-1	-1	5							M
I	-1	-2	-1	-3	-1	-4	-3	-3	-3	-3	-3	-3	-3	1	4						I
L	-1	-2	-1	-3	-1	-4	-3	-4	-3	-2	-3	-2	-2	2	2	4					L
V	-1	-2	0	-2	0	-3	-3	-3	-2	-2	-3	-3	-2	1	3	1	4				V
F	-2	-2	-2	-4	-2	-3	-3	-3	-3	-3	-1	-3	-3	0	0	0	-1	6			F
Y	-2	-2	-2	-3	-2	-3	-2	-3	-2	-1	2	-2	-2	-1	-1	-1	-1	3	7		Y
W	-2	-3	-2	-4	-3	-2	-4	-4	-3	-2	-2	-3	-3	-1	-3	-2	-3	1	2	11	W
	C	S	T	P	A	G	N	D	E	Q	H	R	K	M	I	L	V	F	Y	W	

Annexe 4 - Modèle logique de la banque de données MS2PH-db.



Annexe 5 - Mutations relatives à la TNSALP

Mutation	Sévérité	génotype	Classe de conservation				Accessi-bilité	Struct. Second.	A moins de 6 Å de ...									Dom. structural
			1	2	3	4			S 110	D 60	D 378	H 379	T 173	E 332	D 337	H 341	H 454	
p.N478I	sévère	dominant					surface											
p.P108L	sévère	dominant		X			cœur		X		X	X					X	
p.A116T	sévère	dominant	X				cœur	Hélice										CAbinding
p.G249V	sévère	dominant		X			surface											Couronne
p.V424M	modéré	homozygote					surface											
p.G334D	sévère	homozygote		X			cœur			X			X	X	X			
p.N170D	sévère	homozygote		X			cœur						X					
p.D189E	sévère	homozygote			X		surface	Hélice										
p.D203E	sévère	homozygote		X			cœur											CAbinding
p.L299P	sévère	homozygote					surface	Hélice										
p.L37P	sévère	homozygote				X	surface	Hélice										Couronne
p.G420S	sévère	homozygote				X	cœur											
p.A179T	sévère	homozygote		X			cœur											
p.A468T	sévère	homozygote					cœur											
p.A377V	sévère	homozygote		X			cœur				X	X						CAbinding
p.D306V	sévère	homozygote					cœur											
p.R136C/ p.R136H	modéré	hétérozygote					surface											
p.S17F	modéré	hétérozygote					surface											
p.D406G	modéré	hétérozygote		X			surface											Couronne
p.A399S/ p.A399V	modéré	hétérozygote					cœur											Couronne
p.A176T	modéré	hétérozygote				X	cœur	Hélice					X					
p.A177T	modéré	hétérozygote					cœur	Hélice					X	X				
p.A132V/ p.A132T	modéré	hétérozygote		X			surface											
p.V423A	sévère	hétérozygote					surface											Couronne
p.T148A/ p.T148I	sévère	hétérozygote					surface											
p.Y28C	sévère	hétérozygote		X			surface	Hélice										
p.G426C/ p.G426D	sévère	hétérozygote					surface											Couronne
p.R391C/ p.R391H	sévère	hétérozygote			X		surface											Couronne
p.R71C/ p.R71S/ p.R71P/ p.R71H	sévère	hétérozygote	X				surface	Hélice										

			Classe de conservation						A moins de 6 Å de ...									
Mutations	Sévérité	génotype	1	2	3	4	Accessi-bilité	Struct. Second.	S 110	D 60	D 378	H 379	T 173	E 332	D 337	H 341	H 454	Dom. structural
p.E354D	sévère	hétérozygote				X	surface	Hélice										
p.N211D	sévère	hétérozygote				X	surface											
p.Y297D	sévère	hétérozygote		X			surface											CAbinding
p.K224E	sévère	hétérozygote					surface											
p.I212F	sévère	hétérozygote					surface											
p.E235G	sévère	hétérozygote					surface											CAbinding
p.V382I	sévère	hétérozygote				X	surface										X	
p.E291K	sévère	hétérozygote		X			surface											CAbinding
p.E452K	sévère	hétérozygote				X	cœur								X	X	X	
p.F328L	sévère	hétérozygote	X				cœur	Feuillet										
p.L414M	sévère	hétérozygote					surface											Couronne
p.T167M	sévère	hétérozygote	X				cœur	Feuillet		X			X	X				
p.T68M	sévère	hétérozygote		X			surface	Hélice										
p.T100M	sévère	hétérozygote	X				cœur				X	X						
p.V128M	sévère	hétérozygote		X			cœur						X					
p.V459M/ p.V459L	sévère	hétérozygote	X				cœur	Feuillet			X	X						
p.L275P	sévère	hétérozygote		X			cœur	Hélice										CAbinding
p.Q207P	sévère	hétérozygote		X			surface	Hélice										
p.G326R	sévère	hétérozygote	X				cœur											
p.G491R	sévère	hétérozygote		X			surface											
p.G129R	sévère	hétérozygote	X				cœur											
p.H381R	sévère	hétérozygote			X		cœur			X		X			X	X	X	
p.K264R	sévère	hétérozygote					surface											CAbinding
p.G339R/ p.G339E	sévère	hétérozygote		X			cœur	Hélice							X	X		
p.G473S	sévère	hétérozygote	X				surface	Feuillet										
p.G75S	sévère	hétérozygote		X			cœur	Hélice										
p.R246S	sévère	hétérozygote	X				cœur											CAbinding
p.A348T	sévère	hétérozygote		X			cœur	Hélice										
p.I218T	sévère	hétérozygote		X			cœur	Feuillet										
p.I395T	sévère	hétérozygote			X		cœur											Couronne
p.M355T	sévère	hétérozygote		X			cœur	Hélice						X				
p.P292T	sévère	hétérozygote				X	surface											CAbinding
p.M295T/ p.M295I/ p.M295V	sévère	hétérozygote		X			cœur											CAbinding
p.A40V	sévère	hétérozygote					surface	Hélice										

			Classe de conservation						A moins de 6 Å de ...									
Mutations	**Sévérité**	**génotype**	**1**	**2**	**3**	**4**	**Accessi-bilité**	**Struct. Second.**	**S 110**	**D 60**	**D 378**	**H 379**	**T 173**	**E 332**	**D 337**	**H 341**	**H 454**	**Dom. structural**
p.D378V	sévère	hétérozygote	X				cœur		X	X	X	X	X					
p.G221V	sévère	hétérozygote	X				cœur	Feuillet										
p.A51V/ p.A51S	sévère	hétérozygote		X			cœur											
p.G162V/ p.G162S	sévère	hétérozygote	X				cœur	Feuillet										
p.D294Y/ p.D294A	sévère	hétérozygote				X	surface											CAbinding
p.T411A	sévère	hétérozygote					cœur											Couronne
p.T389I	sévère	hétérozygote				X	surface											Couronne
p.S181L	sévère	hétérozygote					cœur											
p.G120R	sévère	hétérozygote	X				surface											
p.C489S	sévère	hétérozygote		X			surface											
p.N417S	sévère	hétérozygote		X			surface	Feuillet									X	Couronne
p.R184W	sévère	hétérozygote				X	surface		X								X	
p.C201Y	sévère	hétérozygote		X			surface											
p.G63V/ p.G63R	sévère	dom/homo		X			cœur				X	X						
p.E476K/ p.E476G	sévère	hetero/homo	X				surface	Feuillet										
p.T134H/ p.T134N	sévère	hetero/homo					surface											
p.G455D/ p.G455S	sévère	homo/hétéro		X			surface					X					X	
p.R272H/ p.R272L	sévère	homo/hétéro		X			surface	Hélice										CAbinding
p.E191K/ p.E191G	modéré	hétérozygote				X	surface	Hélice										
p.I490F	modéré	?					cœur											
p.L289F	modéré	?	X				cœur	Feuillet										CAbinding
p.P466L	sévère	?	X				surface											
p.S445P	sévère	?				X	surface											Couronne
p.R223W	sévère	?		X			cœur											

Annexe 6 - Mutations relatives à la myotubularine

		Classes de conservation					
Mutation	**Sévérité**	**1**	**2**	**Accessibilité**	**Structure Secondaire**	**A moins de 6 Å de C375**	**Domaine fonctionnel**
p.E404K	modéré			cœur	Hélice		PTP
p.I225T	modéré		X	cœur			myo-related
p.L70F	modéré		X	cœur	Feuillet		GRAM
p.L87P	modéré			cœur	Hélice		GRAM
p.M317R	modéré		X	cœur	Hélice		
p.M1V	modéré			-			
p.N180K	modéré			surface			
p.N481Y	modéré	X		surface			
p.P179S	modéré			surface			
p.P226T	modéré	X		cœur			myo-related
p.S229P	modéré		X	cœur	Feuillet		myo-related
p.V227M	modéré		X	cœur	Feuillet		myo-related
p.W411C	modéré	X		cœur			PTP
p.W499R	modéré	X		cœur	Hélice		
p.W346S/ p.W346C	modéré	X		cœur	Hélice		
p.Y68D	modéré		X	surface			GRAM
p.A389D	sévère		X	cœur	Hélice		PTP
p.C444Y	sévère	X		cœur	Hélice		
p.D431N	sévère			surface			
p.G356E	sévère			cœur			PTP
p.G378E/ p.G378R	sévère	X		cœur		X	
p.G415V	sévère	X		cœur			PTP
p.H374D	sévère	X		cœur		X	
p.H469P	sévère	X		surface	Hélice		
p.I264S	sévère		X	cœur	Hélice		myo-related
p.K510N	sévère			surface			
p.L406P	sévère	X		cœur	Hélice		PTP
p.L470P	sévère			surface	Hélice		
p.L202S	sévère		X	surface	Feuillet		myo-related
p.N189S	sévère	X		cœur			
p.P199L	sévère	X		surface			myo-related
p.P205L	sévère	X		cœur	Feuillet		myo-related
p.Q408P	sévère	X		surface			PTP
p.R184G/ p.R184L	sévère		X	surface	Feuillet		
p.R421Q	sévère	X		surface	Hélice		PTP
p.S387F	sévère		X	cœur			PTP
p.S376N	sévère	X		surface		X	
p.T197I	sévère		X	cœur			myo-related
p.W230C/ p.W230R	sévère	X		surface	Feuillet		myo-related
p.Y397C	sévère	X		cœur	Hélice		PTP
p.Y198N	sévère	X		cœur			myo-related
p.N189S	sévère	X		cœur			
p.P199L	sévère	X		surface			myo-related
p.P205L	sévère	X		cœur	Feuillet		myo-related

Annexe 7 - Publication n° 3

Friedrich A, Poch O, Moulinier L: Strategies for efficient exploitation of the informational content of protein multiple alignments. dans *Multiple Sequence Alignment*, éditeur M.S. Rosenberg (à paraître en 2008).

Strategies for efficient exploitation of the informational content of protein multiple alignments

Anne Friedrich, Olivier Poch and Luc Moulinier

Département de Biologie et Génomique Structurales,
Institut de Génétique et de Biologie Moléculaire et Cellulaire,
(CNRS/INSERM/ULP),
BP 10142,
67404 Illkirch Cedex,
France

1 Introduction

Life can now be considered as a complex system in which molecular agents are interconnected in space and time. The information for life is mainly stored and organized by stretched chains of chemical building blocks: 4 nucleotides for genes and, after gene translation, 20 amino acids are used to form proteins. Despite the recent developments highlighting the existence of codes and higher orders of organisation at the genome level[1, 2], proteins still represent the major mediator for life information management. Various levels of organisation can be considered. First, the primary structure describes the number and arrangement of amino acids characterizing a particular protein. The secondary structure then represents the assemblage of the polypeptide backbone into local regions of alpha-helices, beta-sheets, coils and turns. The tertiary structure refers to the entire 3-dimensional structure of the protein, and finally the quaternary structure describes interactions between separate polypeptide chains, called subunits, arranged in protein complexes. Furthermore, proteins and macromolecular complexes are interconnected in specialized cellular pathways or processes to perform most of the cellular functions and events. These processes are strongly regulated at the gene, protein and macromolecular levels by expression, regulation or chemical modification controls which are highly dependent on a number of different factors, such as the cellular localization, tissue type, developmental stage or environment.

These interwoven informational levels are clearly essential for the understanding of the functions of a gene or protein system. The widespread use of high-throughput technologies, the so called “omics” (e.g. structural genomics, transcriptomics, interactomics…) has not only increased the total number of DNA and protein sequences available but has also given us access to many information levels that were not accessible before.

In this context, sequence comparisons facilitate the identification of both major and minor genetic events (e.g. adaptation; selection; speciation; genetic drift; flows, duplications, mutations or recombination of gene and genome) and play a central role in biological studies by introducing the evolutionary dimension to the analysis of the complex gene and protein relationships[3]. Sequence comparison is therefore one of the most common tasks in bioinformatics, increasing our understanding of the relationships between sequence/structure/function and evolution. The purpose of multiple sequence comparisons, or alignments is to elucidate the complex relationships that exist between a set of related sequences, in terms of conserved (evolutionarily fixed) residues, substitutions, insertion or deletion events.

Sequence alignments can be classified in two main categories: pairwise alignments i.e. the alignment of two sequences and multiple alignments i.e. from three to thousands sequences. For both categories, the sequences can be aligned across their entire length (global alignment) or only across the most similar segments of the sequences (local alignment). Pairwise alignments are commonly used in database search programs[4, 5] in order to detect sequences that have evolved from the same ancestor i.e. homologous sequences and MSA are mainly used to study the effect of evolution across the sequence family and have a wide range of applications.

Here, we will concentrate on the multiple sequence alignment (MSA) of complete proteins, which represents an ideal workbench to study all the information related to a set of

homologous sequences. Indeed, by placing the sequence in the context of its overall family, it permits not only a “horizontal” analysis of an event (residue, sequence motif, domain insertion or fusion, etc) along the complete length of the proteins but also a “vertical” view of its evolution among different clades or organisms. In systems-level studies, proteins are considered as interacting multi-dimensional entities and in this case, multiple sequence alignment can also contribute to our understanding of the majority of these dimensions.

In this chapter, we will first give an overview of the central role of multiple sequence alignments and their main applications in current biological studies. We will then discuss the problem of how to efficiently construct MSA depending on the biological question being studied. Finally we will consider how the standard applications can be integrated for wider exploitation.

2 MSA : central role in protein analysis

Since their introduction in the early seventies[6], sequence comparisons and alignments have been used in a wide range of biological applications and multiple alignments now play a fundamental role in most of the computational methods used in genomic or proteomic projects. This central role is further emphasized by the numerous studies and analysis that are directly or indirectly associated to multiple sequence alignment (MSA) and especially, to multiple alignment of complete sequences, ranging from evolutionary, functional or structural studies up to comparative genomics (Figure 1).

2.1 Evolutionary studies

One of the earliest applications of multiple sequence alignments was to define the phylogenetic relationships between organisms[7] through phylogenetic tree construction. Briefly, the objective is to reflect the evolutionary relationships between each organism in a tree and the methods for calculating phylogenetic trees fall into two general categories[8]: distance-matrix methods, like UPGMA and neighbour-joining, also known as clustering or algorithmic methods which provide an estimation of the distance between each organism by computing each pairwise sequence distance, and discrete data methods such as parsimony, maximum-likelihood or Bayesian methods, also known as tree searching methods.

Strikingly, it has been shown that variation on resulting phylogeny is more dependent on the mode of alignment than on the method of phylogenetic reconstruction[9]. This emphasizes the fact that generating high-quality sequence alignments is clearly a critical step on the path from raw sequence to phylogeny. Multiple alignments are also essential for the estimation of the reliability of each branch of the constructed tree using bootstrapping methods which involve re-sampling of the alignment columns and subsequent tree reconstruction.

The information from several multiple alignments can be used for the estimation of major genetic events and even, of the Tree of Life[10], by concatenating various alignments of distinct protein families [11, 12]. The concatenation increases the number of informative sites for the phylogeny reconstruction and overcomes the discrepancies introduced by lateral gene transfers or by the existence of divergent evolutionary rates. Another commonly used

approach to combine phylogenetic information is based on multiple, independently reconstructed trees[13, 14]. These trees can be calculated using different types of multiple alignments involving protein, DNA or RNA sequences.

Phylogenetic studies can also be exploited for the reconstruction of ancestral protein sequences, which may improve the sensitivity of remote homolog searches and the prediction of functional sites. The ancestral reconstruction has to take into account the observed variation of evolutionary rates between specific positions of the alignment that more precisely describes the evolution of protein families[15]. Finally, proteins may evolve at different rates and evolutionary studies can be used to determine family specific evolutionary rates. Indeed, the mutational process which acts on a protein is sometimes subject to a specific selective pressure and MSA represents the basis for the estimation of this evolutionary rate. As an example, the analysis of Luz and Vingron [16] revealed that the evolution of indispensable proteins is constrained by selection while extra-cellular proteins are evolving faster.

2.2 Functional assignments

In the context of protein functional characterisation, the first levels of analysis cover the detection of sequence errors, the identification of functional domains and the study of domain organization.

In most genome annotation projects, the standard approach to characterize a novel protein is to search the sequence databases for homologs and to propagate the structural/functional annotation from the known to the unknown protein. Most automatic annotation methods use information from the top best hits in database searches, but it is now well documented that best hits do not necessarily represent the closest sequence relatives[17], thereby casting doubt on the reliability of this approach, which leads to a certain number of annotation errors[18]. Two main error types can be met[19]: under-prediction, when a function is not transferred although it should be, for example when all the top blast hits are uncharacterized or over-prediction, when similarity is restricted to a region of the sequence and all annotations are transferred. Another approach is to look for similarities to known domains in pre-compiled databases, such as Interpro[20]. These databases contain representations such as profiles or HMM's of individual protein domains based on multiple alignments of known sequences.

Recent developments in database search methods have exploited multiple alignments to detect more distant sequences[21-23]: database similarity searches are performed with a profile constructed by considering the whole protein family. This allows a more sensitive detection of new family members.

MSA can alleviate some of the problems due to sequencing errors[24] or on the presence of annotation errors in the source database which are frequently ignored in automatic annotation systems . By integrating a sequence in the context of its close relatives, multiple alignment of complete sequences allows the detection of inconsistencies in open reading frame determination. The vALId[25] program has been developed to take advantage of this feature of MSA. It warns about the occurrence of suspicious insertions, deletions and divergent segments absent of close relatives, and proposes corrections based on mRNA or genome-based translation of alternative frame-shifted regions.

Wrong annotation[26] due to human or automatic assignments can be introduced in the public sequence databases and some studies have shown that once an erroneous annotation is introduced into a database, it tends to propagate via automatic annotation inference methods that are based on sequence similarity[27]. Therefore, existing functional assignments need to be cross-validated before they can be used as a basis for reliable annotations,. This has lead to the introduction of knowledge-based systems for automated multiple alignment annotation, e.g. [28]. The major advantage of these integrated systems is that the sequence information mined from the public databases can be cross-validated within the MSA, in order to distinguish between reliable, consistent information and spurious predictions. The validated data then provides a basis for the accurate propagation of information from known to unknown sequences.

MSA can also be used to study the domain organization of proteins[29]: in the course of evolution, the fusion of functional/structural domains and the splitting of multi-domain proteins have occurred frequently and as a result, some protein families present a complex mosaic picture. Furthermore, insights into cellular localization information can be obtained with the prediction of transmembrane helices for example, for which MSA improves the reliability[30].

MSA also reveal evolutionary constraints existing at particular sequence positions that conserve residues or their physico-chemical properties and thus allows a straightforward detection of potential functional sites (e.g. those involved in catalytic sites, protein interactions, ligand binding) that characterize a protein family or sub-family[31, 32]. Methods for the prediction of these functional sites have been developed that use as the primary indicator of potential sites the conservation of amino acid in MSA (e.g. [33, 34]). Other prediction methods exploit structural information but it has now become clear that neither sequence nor structure alone is sufficient for accurate predictions. Efforts are now being concentrated on the combined use of both sequence conservation and structural information (e.g. [35, 36]).

2.3 Structural studies

MSA play an important role in a number of aspects of the characterisation of protein structures, in particular in the improvement of secondary and tertiary structure prediction, aimed at estimating the role a residue plays in a protein structure (e.g. helix or strand; buried or exposed).

The current most successful secondary structure prediction methods all employ positional information from multiple alignments[37-39]. Prediction based on MSA allows the increase of the prediction accuracy from about 60%[40] to 75%[41, 42] compared to prediction based on single sequences. Indeed, the use of aligned sequences allows better application of the propensities of particular residues for particular secondary structures and improved identification of patterns of hydrophobicity[43]. The rationale behind such improvements is that the pattern of substitutions observed in a column directly reflects the type of constraints imposed on that position in the course of evolution.

Protein tertiary structures are classified into families based on a limited set of folding patterns, defined as the three-dimensional fitting of the secondary structure elements (see SCOP[44] and CATH[45]). In general, proteins sharing the same function have similar

structure and functional sites i.e. the residues that are involved in the protein molecular function, maintain identical structural positions. Such observations became the foundation of homology modelling, which is the most accurate *in silico* method for determining the tertiary structure of an unknown protein. Sequence similarity between proteins usually indicates a structural resemblance, and accurate sequence alignments provide a practical approach for structure modelling, when a 3D structural prototype is available. MODELLER[46], Geno3D[47], SWISS-MODEL[48], 3D-JIGSAW[49] are accurate tools for protein modelling. For models based on distant evolutionary relationships, it has been shown that multiple sequence alignments often improve the accuracy of the structural prediction[50].

2.4 Mutation studies

A considerable effort is now underway to relate human phenotypes to variation at the DNA level. Most human genetic variation is represented by single nucleotide polymorphisms (SNPs) and many of them are believed to cause phenotypic differences between individuals[51]. One of the main goals of SNP research is therefore to understand the genetics of human phenotype variation and especially the genetic basis of complex diseases, thus providing a basis for assessing susceptibility to diseases and designing individual therapy. Whereas a large number of SNPs may be functionally neutral, others may have deleterious effects on the regulation or the functional activity of specific gene products. Non-synonymous single-nucleotide polymorphisms (nsSNPs) that lead to an amino acid change in the protein product are of particular interest because they account for nearly half of the known genetic variations related to human inherited disease[52]. There is a surprisingly high fraction of nsSNPs that affect the structure and probably the functional sites of proteins[53], which implies a probably negative effect on the associated phenotype.

Given the huge number of nsSNPs already discovered[54, 55], it has become imperative to predict the phenotype of a nsSNP *in silico*. Computational tools are therefore being developed e.g. nsSNPAnalyzer[56], SIFT[57], PolyPhen[51], which use structural information or evolutionary information from MSA to predict a nsSNP's phenotypic effect and to identify disease-associated nsSNPs. Recent studies have shown that combining information obtained from multiple sequence alignment and three-dimensional protein structure can increase the prediction accuracy[56, 58].

2.5 Interaction networks

Biological entities are more and more considered as part of complex networks of interactions. Powerful experimental techniques, such as the yeast two-hybrid system or tandem-affinity purification and mass spectrometry, are used to determine protein-protein interactions systematically and protein interaction maps have been built for several organisms such as *S. cerevisiae[59]*, *E. coli[60]* and *D. melanogaster[61, 62]* . In parallel with these experimental developments, a number of computational techniques, that give rise to the term comparative interactomics[63], have been designed for predicting protein interactions. The basic assumption of these techniques is that interactions may be conserved during evolution and can thus be inferred by studying the orthology of proteins. The "Interaction Domain Profile Pairs" (IDPP, [64]) approach uses a high quality reference protein interaction map to predict an interaction map in another organism. This method combines sequence similarity searches and alignments with clustering based on interaction patterns and interaction domain

information. Some other methods explore genome organization: one implies that proteins that are consistently present or absent in different proteome sets are likely to interact functionally[65]. The Rosetta stone method[66] is based on the observation that some interacting proteins have homologues in other organisms fused into a single protein chain. This method has recently been improved using MSA information and global measures of hydrophobic core formation[67]. Pairs of interacting proteins have also been predicted using a measure of the similarity between phylogenetic trees of protein families[68]. This method was adapted to consider the multi-domain nature of proteins by breaking the sequence into a set of segments of predetermined size and constructing a separate profile for each segment[69]. Another approach involves quantifying the degree of co-variation (based on the MSA) between residues from pairs of interacting proteins (correlated mutations), known as the "*in silico* two-hybrid" method. For some proteins that are known to interact, correlated mutations have been demonstrated to be able to select the correct structural arrangement of two proteins based on the accumulation of signals in the proximity of interacting surfaces[70]. This relationship between correlated residues and interacting surfaces has been extended to the prediction of interacting protein pairs based on the differential accumulation of correlated mutations between the interacting partners (inter-protein correlated mutations) and within the individual proteins (intra-protein correlated mutations)[71]. Comparison of predicted data with experimental data has demonstrated that the predictive power is increased when several independent sources of data and different algorithms are combined[72].

2.6 Comparative genomics

Comparative genomics is a discipline based on whole genome comparison which has become more and more informative as genomic and proteomic sequence data accumulate[73]. This discipline is generally based on DNA sequences but is clearly essential in the context of gene and protein characterization, mainly at the evolutionary and functional levels. At the end of 2006, over 1000 genomes (from bacteria, archaea and eukaryota, as well as viruses and organelles) are either complete or being determined. In this era of complete genome sequences, it has become possible to perform comparative multiple sequence analysis at the genome level and also at a proteome level. A number of software tools have been developed for use in comparative genomics, in order to explore the similarities and differences between genomes at different levels. Because of the volume and nature of the data involved, almost all the visualization tools in this field use a web interface to access large databases of pre-computed sequence comparisons and annotations, e.g. VISTA[74], Ensembl[75], UCSC[76]. Platforms dedicated to comparative genomics have also been made available such as Phytome[77], which provides online resources that can be applied to functional plant genomics, molecular breeding and evolutionary studies. Phytome contains predicted protein sequences, protein family assignments, multiple sequence alignments, phylogenies and functional annotations for proteins from a large, phylogenetically diverse set of plant taxa.

At the protein analysis level, comparative genomics take essentially a medium-resolution view. By identifying all the known proteins from one genome and finding their protein homologues, if they exist, in another genome, one can determine which proteins have been conserved between species, and which are unique. Studies are mainly based on the determination of protein presence/absence, in order to deduce phylogenetic profiles. A comparison of the genomes of yeast, worms and flies revealed that these eukaryotes encode many of the same proteins, but different gene families are expanded in each genome[78]. Genes and gene products that are responsible for species-specific phenotypes can also be

identified with this method. By comparing the *Helicobacter pylori* genome with its closest relatives for which complete genome sequences are available, the pathogenic *Haemophilus influenzae* and the benign *Escherichia coli* K12[79], it has been possible to identify a set of species-specific genes that are responsible for pathogen-specific features (Figure 2).

Determining protein functions from genomic sequences can also be achieved using the phylogenetic profile method[65]. This method is based on the assumption that proteins that function together in a pathway or structural complex, i.e. functionally linked proteins, are likely to evolve in a correlated fashion. The phylogenetic profile of a protein describes the presence or absence of homologs in organisms. Proteins that make up multimeric structural complexes are likely to have similar profiles. Also, proteins that are known to participate in given biochemical pathways are likely to have close phylogenetic profiles. Comparing profiles is therefore a useful tool for identifying the macromolecular complex or pathway in which a protein participates. It is possible to assign the function of uncharacterized proteins by examining the function of proteins with identical phylogenetic profiles.

3 Efficient construction of MSA

Genomics and proteomics technologies, together with the new system biology strategies, have led to a paradigm shift in bioinformatics. The traditional reductionist approach has been replaced by a more global, integrated view. Multiple sequence alignment is one example where a shift of thinking or focus is now leading to the development of new methods.

The construction of a multiple sequence alignment traditionally begins with a sequence database search to find homologous sequences. These sequences are then aligned with appropriate software and automatically and/or manually refined to obtain a high quality multiple alignment. With the exponential increase in the size of the sequence databases, it is no longer possible to align all the homologous sequences available, and therefore biologists have to develop novel strategies to select the pertinent sequences, according to their analysis goals.

3.1 Detection of similar sequences

The alignment methods only make sense if they are assumed to be dealing with homologous sequences, i.e. sequences sharing a common ancestor. Several programs such as Blast[22] and Fasta[5] search for local similarity between the query and database sequences, in order to detect homologous sequences. These similarities might lead to evolutionary clues about the structure and/or function of the query sequence. Each comparison is given a score reflecting the degree of similarity between the query and the sequence being compared: the higher the score, the greater the degree of similarity. In both Blast and Fasta, in addition to the alignment scores, the significance of each alignment is computed as a p-value or an E-value, based on the alignment scores expected by chance in the total sequence space. Although these methods have proven their ability to detect closely related sequences (those with sequence identity larger than 30%), their performance decreases quickly when one tries to find more distant homologs[80].

For remotely related proteins, conservation is weak and is often restricted to short segments of the sequences. To overcome this limitation, researchers have developed various homology search methods based on the common features of a group of related sequences. These methods, e.g. HMMER[23], based on hidden markov models (HMM), and PSI-BLAST[22] exploit multiple alignments to detect more and more distant homologues. With PSI-BLAST, a profile is automatically constructed from a multiple alignment of the highest scoring hits in an initial BLAST search. Highly conserved positions receive high scores and weakly conserved positions receive scores near zero and the profile is used to perform a second BLAST search (etc.). The results of each iteration are used to refine the profile, which increased the sensitivity for homologues search.

3.2 Alignment software

Once the homologous sequences to be aligned have been determined, the biologist has to choose an appropriate alignment software. Three main considerations have to be taken into account in choosing a program: biological accuracy, execution time and memory usage, biological accuracy being the most important concern. It has been shown that the best choice of an alignment program depends on the sequences set to be aligned and that no single 'best' program exists[81-83]: no automatic alignment method can be expected to produce biologically meaningful alignments whatever the sequences to be aligned. There is therefore a continuous effort to improve the biological accuracy of multiple sequence alignments (MSA) tools. Until recently the choice for building MSAs was limited to a handful of packages but a recent increase in genomic data has fuelled the development of many novel methods arguably more accurate and faster than the older ones[83].

The next section briefly describes the evolution of multiple alignment methods from their beginnings in the seventies to the recent introduction of new integrative and co-operative strategies (Figure 3).

Although a dynamic programming algorithm exists which guarantees a mathematically optimal alignment, the method is limited to a small number of short sequences since the computing power required for larger alignments becomes too prohibitive[84]. To overcome this problem, various heuristic approaches have been developed leading to a huge quantity of programs using fundamentally different strategies based on very different algorithms.

Traditionally the most popular method has been the progressive alignments[85] which exploit the fact that homologous sequences are evolutionarily related. A multiple alignment is built up gradually by aligning the closest sequences first and successively adding in the more distant ones. PIMA[86], MULTALIGN[87], MULTAL[88] and CLUSTALW[89] are for example based on this method, and mainly differ in the method used to determine the order of alignment of the sequences. Progressive methods have the great advantage of speed and simplicity combined with reasonable sensitivity, even if it is by nature a heuristic that does not guarantee any level of optimization.

Later, numerous alignment algorithms were developed which applied iterative strategies, introducing the capacity to refine previously aligned sequences at each iteration with the goal of improving the overall alignment quality, using an objective function (see *MSA quality*). A local alignment approach is implemented in the DIALIGN[90] program to construct multiple alignments based on segment-to-segment comparison rather than the residue-to-residue

comparison used previously. The PRRP[91] program optimises a progressive, global alignment by iteratively dividing the sequences into two groups, which are subsequently realigned using a global group-to-group alignment algorithm. SAGA[92] uses a genetic algorithm to select from an evolving population the alignment which optimises the COFFEE[93] objective function (see *MSA quality*). The program HMMT[23] uses a simulated annealing method to maximise the probability that an HMM represents the sequences to be aligned. Iterative methods often produced more accurate alignments, although these methods are generally slower than the progressive ones[81].

The complexity of the multiple alignment challenge has lead to the combination of different sequence-based alignment algorithms and even, more recently, to the incorporation of biological information other than the sequence itself [94]. DbClustal[95] for example uses information about locally conserved segments detected during a database search to guide or "anchor" a global multiple alignment. T-Coffee[96] uses information from a pre-compiled library of pairwise alignments (local, global and structural alignments) and constructs a multiple alignment with the highest possible level of consistency with the alignments within the library. MAFFT[97] and MUSCLE[98] both use fast algorithms to detect local sequence similarities, then use a progressive alignment algorithm together with an iterative refinement step to produce an accurate alignment of large sets of sequences. ProbCons[99] uses HMM-derived posterior probabilities and three-way alignment consistency in a global, progressive alignment, together with an iterative refinement step. The most recent multiple alignment methods that are now being developed can be considered as knowledge-based strategies : 3DCoffee[100] integrates structural information, PRALINE[101] exploits protein secondary structure information either from 3D structures or from computational predictions, MUMMALS[102] uses HMM structural information (Figure 3).

Other recent developments address the problems associated with sequences having distinct domain organisation, such as RAlign[103], POA[104] and AliWABA[105] that employ partially ordered graphs. These three programs use local algorithms suitable for multi-domain proteins that may contain repeated or shuffled elements.

Despite the enormous progress achieved, current implementations of multiple alignment algorithms are still heuristic, and it is necessary to examine the alignment to check that there are no obvious errors. It is therefore common practice to use automatic refinement programs such as RASCAL[106] or Refiner[107] and then manually improve the alignments, using an alignment editor such as Jalview[108], SEAVIEW[109], ANTHEPROT[110] or CINEMA[111].

3.3 MSA quality

Since multiple alignments are usually employed at the beginning of a series of bioinformatics and/or biological studies, they have to be of high-quality. Errors in the alignment will lead to further errors in the subsequent analyses and might generate misleading patterns and result in false hypotheses. It is therefore critical to be able to distinguish 'good' alignments from bad ones. A good alignment is one that corresponds to the biologically correct alignment, accurately reflecting the evolutionary, structural and functional relationships between the sequences.

A number of studies have been realized to determine which alignment programs work best in which situations. These studies are usually based on benchmarks that have been specifically designed for this purpose (cf chapter 'Constructing alignment benchmarks'), taking the form of databases of precompiled alignments to which the alignments generated by test algorithms are compared.

None of the current implementations of MSA guarantees a full alignment optimization. Given a particular set of sequences, an objective score that describes the optimal or "biologically correct" multiple alignment is needed. Sub-optimal or incorrect alignments would then score less than this maximal score. Such measures, also known as objective functions, are currently used to evaluate and compare multiple alignments from different sources and to detect low-quality alignments. They are also used in iterative alignment methods to improve the alignment by seeking to maximize the objective function. One of the first scoring systems was the Sum-of-Pairs score[84], based on pairwise scores. Pairwise scores are also used in the COFFEE objective function[93] which reflects the level of consistency between a multiple sequence alignment and a library containing pairwise alignments of the same sequences. Column statistics have also been used to score alignments. One approach uses an information content statistic (relative entropy)[112], but it considers only the frequencies of identical residues in each column and does not take into account similarities between residues. Another column-based measure, norMD[113] is based on the Mean Distance (MD) column scores implemented in ClustalX[114] and its main advantage is to incorporate *ab initio* sequence information, such as the number, length and similarity of the sequences to be aligned. Other developments measuring local reliability or column conservation have also been recently reported (e.g [115-117]).

3.4 Sequence selection for more accurate MSA

We are currently facing an explosion of the data available in databases, due to the generalization of high-throughput experimental technologies: genome sequencing as well as structural genomics or transcriptomics have increased the number of DNA and protein sequences available as well as the rate at which genome and sequence products are being functionally characterized.

Some biological sequence databases are highly redundant, for two main reasons:

- some databases keep redundant sequences with many identical and nearly identical sequences,
- natural sequences often have high sequence identities due to gene duplication.

As a direct consequence of the recent database growth, homology searches frequently lead to the identification of hundreds to thousands of potential homologues for a single query sequence. Dealing with so much data can be detrimental in terms of computational and human analysis time: indeed for the majority of alignment programs, the number of sequences to be aligned is limited and even if all sequences can be aligned, the more sequences an alignment contains, the more difficult it would be for the biologist to analyse it. Moreover, this vast quantity of available sequences can also be detrimental in terms of the accuracy of the significance of the results: apart from inherent database noise linked to sequencing or intron/exon prediction errors, problems, such as redundancy or the presence of partial sequences, may represent a significant source of noise, depending on the biological question under study. Redundancy is now a major problem, notably if we consider that aligning a

defined number of sequences according to user-defined or program limitation may hinder access to significant information present in the additional non-aligned sequences.

It is now essential to develop new complementary approaches for an efficient exploitation of this large sequence space. As a consequence, the current challenges are shifting from the multiple alignment to the choice of sequences to align that will yield the most biologically correct and informative alignments. Even though the alignment-based methods are ubiquitous and have been much discussed [94, 118], very little has been written about the actual selection of sequences the alignment is built on. Strategies for sequence selection are essentially based on the assumption that, at the structural and functional level, closely related sequences may not add relevant information, whereas diversity is usually more informative[119].

It is clearly advantageous in terms of processing time to reduce the sequence space at the earliest possible stage of an analysis, i.e. during the initial homology search step. There are clearly two possibilities in this case: an *a priori* reduction of the sequence database search space and an *a posteriori* sampling of the sequences detected by the database search program. The *a priori* reduction of the search space can sometimes easily be done by choosing an appropriate sequence database to run the homology searches according to the biological question under study:

- for comparative genomic studies, the search space is usually restricted to complete genome or proteome databases

- for some evolutionary studies, the sequence search space can be restricted to specific clades, such as mammals or vertebrates, and numerous Blast server, like the Blast NCBI site propose the restriction to specific organisms

- to annotate an unknown sequence, biologists can choose to work with a reduced database with higher-quality annotation such as SWISS-PROT[120], a curated protein sequence database containing a minimal level of redundancy

- for the majority of structural or functional assignments and in the case of information propagation, pre-processed non-redundant databases, where sequences are clustered by means of their percentage of identity, such as the UniRef database series[121] can be used

- for protein homology modelling, the database search can be done in the Protein Data Bank (PDB)[122], a protein 3D structure database which includes structures experimentally determined by X-ray crystallography or by NMR

Obviously, the set of sequences detected by homology searches can also be *a posteriori* sampled. The current most widely used methods for detecting homologous sequences is the Blast suite of programs and software for Blast results post-treatment have been developed. BLAST Filter[123] generates smaller sequence sets by filtering the Blast results: the filter is based on 15 user-configurable predefined rules. UniqueProt[124] creates a representative and unbiased dataset by removing over-represented sequences: the user has to define an identity threshold above which sequences will be clustered. These two programs are based on sequence similarity criteria and need at least a brief analysis of the Blast search results to be powerful.

The MyHits[125] web server, which is an integrated service dedicated to the annotation of protein sequences and to the analysis of their domains and signatures, also provides a tool for the automated clustering of identical and highly similar protein sequences. Typically, a representative subset out of a large set of sequences is extracted, based on a taxonomic filter and a procedure to group the matched sequences with an identity level equal or superior to a user-defined threshold.

A recent study has concentrated on how far the homologous sequences detected by BlastP can be reduced while maintaining the relevant information concerning the active site residues as annotated in the PDB structural database[126]. This study has demonstrated that, on average, it is sufficient to align 30% of the detected sequences to efficiently maintain the relevant functional information when more than 100 sequences are detected, but the sequence selection cannot be performed randomly. The most suitable method tested is based on the BlastP E-values and allows the conservation of the potential structural and functional information in the sampled set while restricting the alignment computation time to a reasonable limit.

The PhyloGenie[127] suite of programs, dedicated to automated phylogenetic studies, integrates the java program Blammer which consists of 5 Blast post-processing steps that convert sets of high-scoring segment pairs (HSPs) to multiple alignments and perform a sequence selection optimization.

In some specific cases, a filtering step can be integrated once the MSA has been computed to ease and optimize its analysis. QR factorization[128] produces minimally redundant sets of protein sequences, based on column pivoting of a matrix encoding the sequence alignment. This software has been designed to choose automatically a linearly independent subset of sequences and has been integrated in MultiSeq[129], a unified bioinformatics analysis environment that allows one to organize, display, align and analyze both sequence and structure data for proteins and nucleic acids.

Mihalek[130] proposes, in the case of proteins with known structure, a heuristic Metropolis Monte Carlo strategy to select sequences from a large set of homologues, in order to improve detection of functional surfaces. The optimization is based on the clustering of residues which are under increased evolutionary pressure, according to the sample of sequences under consideration.

4 MSA integrative exploitation tools

Bioinformatics has to respond to the challenges represented by the vast amounts of heterogeneous data available in the databases. Studying sequence/structure/function/evolution relationships within a protein family can be achieved by integrating the results of many different analyses in the context of multiple sequence alignments.

Thanks to the increasing adoption of common data standards and exchange formats, a number of systems have been described recently that allow the integration and visualization of heterogeneous information in the context of a family of proteins. Pfaat[131] is a semi-automatic protein family annotation tool incorporated in a multiple alignment viewer. The application merges display features such as dendrograms, secondary and tertiary protein structure with SRS[132] (Sequence Retrieval System), subgroup comparison, and extensive

user-annotation capabilities. MyHits[125], an integrated web server dedicated to the annotation of protein sequences based on MSA, helps to the analysis of the protein domains and signatures. A dictionary driven annotation has also been developed, based on Bio-dictionary[133], a collection of amino acid patterns that completely covers the natural sequence space and can capture functional and structural signals that have been reused during evolution, within and across protein families.

A tool combining knowledge-based methods with complementary *ab initio* sequence based predictions for protein family analysis has also been recently developed. MACSIMS[28] integrates different types of data in the framework of the multiple alignment taking advantage of the recently developed Multiple Alignment Ontology (MAO)[134]. A wide range of information, from taxonomic data and functional descriptions to individual sequence features, such as structural domains and active site residues, is mined from the public databases using SRS. MACSIMS also includes a number of algorithms developed for reliable data validation, consensus predictions and rational propagation of information from the known to the unknown sequences. Jalview[108] has been adapted to be able to visualize all the mined and predicted information in MACSIMS. Another tool, OrdAlie (Ordered Alignment Information Explorer, manuscript in preparation) is available that is designed to allow interactive analysis and exploration of protein sequence, structure, function and evolution relationships. With OrdAlie, the multiple sequence alignment is displayed in a graphical window, together with user-selected sequence features. Sequences can be clustered automatically into sub-families and a detailed, hierarchical analysis of residue conservation can be performed at the family or sub-family level. Conserved residues can also be visualized in the context of their 3D structural environment, using the RASMOL structure viewer[135].

Several recent programs also introduced functions to simultaneously visualize multiple alignments and other information, mainly structural, such as VISSA[136], ViTO[137] and Friend[138]. These bioinformatics applications are mainly designed for simultaneous analysis and visualization of multiple structures and sequences, for refinement of sequence/structure alignment for homology modelling and to assist researchers handle many practical issues involving simultaneous analysis of both sequence alignments and 3D structures, for example, checking what the conservation of residues around the active site is.

A recent web server, MAGOS[139] integrates, starting with a single query sequence, the construction of a multiple alignment, its annotation with MACSIMS and the construction of a model of the query sequence. The model and the query protein in the context of the annotated multiple alignment are then interconnected via a user-friendly interface. MAGOS data integration capabilities can be illustrated by the analysis of the known point mutations and their relative position according to functional and structural features of a protein implicated in genetic disease, like Myotubularin (Figure 4). Myotubularin is implicated in the X-linked myotubular Myopathy, a severe muscular disease characterized by a marked hypotonia and generalized muscle weakness. Missense mutations which directly affect residues implicated in the active site pocket are usually associated with severe phenotypes (early postnatal death). The L406P mutation affects a highly conserved residue located in the tyrosine-protein phosphatase domain, located in the middle of a 10 residue helix. The severe phenotype associated with this mutation may be linked to the helix break due to the presence of the proline.

5 Conclusions and perspectives

Bioinformatics techniques are constantly evolving in response to the challenges presented by the mass of heterogeneous and complex information made available thanks to the new high throughput technologies. In the context of sequence-based biology, there is no doubt that multiple alignment of protein sequences will remain a central application in the foreseeable future.

The multiple sequence alignment construction process can be divided into five crucial steps (Figure 5). However, it is not possible to define a single process that is effective in all study cases and efficient MSA construction should take the biological question under study into account in order to choose an appropriate method for sequence selection, as well as the most suitable alignment program.

Protein multiple alignment tools are constantly evolving and improving in terms of scalability and accuracy. It is now clear that primary sequence data alone are not sufficient to construct accurate multiple alignments. Improvements in multiple alignment construction are likely to come by combining sequence alignment with other information, such as known structures, results of database searches for local similarities, experimental data and, in general, anything that may come from expression data and proteomic analysis. As mentioned in the *Alignment software* section, recently developed software already integrate structural and/or local similarity information and usually increase the quality of the resulting alignment compared to methods based solely on primary sequences. Moreover, with the exponential increase of the amount of sequence data available, it becomes fundamental to develop strategies aimed at reducing the number of sequences to be aligned without losing any relevant information.

As multiple sequence alignments provide an ideal environment for the integration of structural and functional information in the context of the protein family, new information management systems will continue to develop for the collection, validation and analysis of the vast amount of heterogeneous data available. A major challenge will be the selection of the most descriptive and useful information and its presentation in a suitable format for the biologist. The ideal tool will allow the validation, visualization, integration and interpretation in a biological context, of the vast amount of diverse data generated by the application of proteomic and genomic discovery science tools.

References

1. Segal E, Fondufe-Mittendorf Y, Chen L, Thastrom A, Field Y, Moore IK, Wang JP, Widom J: **A genomic code for nucleosome positioning**. *Nature* 2006, **442**(7104):772-778.
2. Kepes F: **Periodic epi-organization of the yeast genome revealed by the distribution of promoter sites**. *J Mol Biol* 2003, **329**(5):859-865.
3. Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O: **Multiple alignment of complete sequences (MACS) in the post-genomic era**. *Gene* 2001, **270**(1-2):17-30.
4. Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ: **Basic local alignment search tool**. *J Mol Biol* 1990, **215**(3):403-410.
5. Pearson WR, Lipman DJ: **Improved tools for biological sequence comparison**. *Proc Natl Acad Sci U S A* 1988, **85**(8):2444-2448.
6. Needleman SB, Wunsch CD: **A general method applicable to the search for similarities in the amino acid sequence of two proteins**. *J Mol Biol* 1970, **48**(3):443-453.
7. Phillips A, Janies D, Wheeler W: **Multiple sequence alignment in phylogenetic analysis**. *Mol Phylogenet Evol* 2000, **16**(3):317-330.
8. Page RDM, Holmes EC: **Molecular Evolution: a Phylogenetic Approach**: Blackwell Science; 1998.
9. Morrison DA, Ellis JT: **Effects of nucleotide sequence alignment on phylogeny estimation: a case study of 18S rDNAs of apicomplexa**. *Mol Biol Evol* 1997, **14**(4):428-441.
10. Wolf YI, Rogozin IB, Grishin NV, Koonin EV: **Genome trees and the tree of life**. *Trends Genet* 2002, **18**(9):472-479.
11. Brown JR, Douady CJ, Italia MJ, Marshall WE, Stanhope MJ: **Universal trees based on large combined protein sequence data sets**. *Nat Genet* 2001, **28**(3):281-285.
12. Clarke GD, Beiko RG, Ragan MA, Charlebois RL: **Inferring genome trees by using a filter to eliminate phylogenetically discordant sequences and a distance matrix based on mean normalized BLASTP scores**. *J Bacteriol* 2002, **184**(8):2072-2080.
13. Sicheritz-Ponten T, Andersson SG: **A phylogenomic approach to microbial evolution**. *Nucleic Acids Res* 2001, **29**(2):545-552.
14. Wolf YI, Rogozin IB, Grishin NV, Tatusov RL, Koonin EV: **Genome trees constructed using five different approaches suggest new major bacterial clades**. *BMC Evol Biol* 2001, **1**:8.
15. Cai W, Pei J, Grishin NV: **Reconstruction of ancestral protein sequences and its applications**. *BMC Evol Biol* 2004, **4**:33.
16. Luz H, Vingron M: **Family specific rates of protein evolution**. *Bioinformatics* 2006, **22**(10):1166-1171.
17. Koski LB, Golding GB: **The closest BLAST hit is often not the nearest neighbor**. *J Mol Evol* 2001, **52**(6):540-542.
18. Bork P, Koonin EV: **Predicting functions from protein sequences--where are the bottlenecks?** *Nat Genet* 1998, **18**(4):313-318.
19. Sasson O, Kaplan N, Linial M: **Functional annotation prediction: all for one and one for all**. *Protein Sci* 2006, **15**(6):1557-1562.
20. Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bradley P, Bork P, Bucher P, Cerutti L *et al*: **InterPro, progress and status in 2005**. *Nucleic Acids Res* 2005, **33**(Database issue):D201-205.

21. Yona G, Levitt M: **Within the twilight zone: a sensitive profile-profile comparison tool based on information theory**. *J Mol Biol* 2002, **315**(5):1257-1275.
22. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: a new generation of protein database search programs**. *Nucleic Acids Res* 1997, **25**(17):3389-3402.
23. Eddy SR: **Profile hidden Markov models**. *Bioinformatics* 1998, **14**(9):755-763.
24. Dandekar T, Huynen M, Regula JT, Ueberle B, Zimmermann CU, Andrade MA, Doerks T, Sanchez-Pulido L, Snel B, Suyama M *et al*: **Re-annotating the Mycoplasma pneumoniae genome sequence: adding value, function and reading frames**. *Nucleic Acids Res* 2000, **28**(17):3278-3288.
25. Bianchetti L, Thompson JD, Lecompte O, Plewniak F, Poch O: **vALId: validation of protein sequence quality based on multiple alignment data**. *J Bioinform Comput Biol* 2005, **3**(4):929-947.
26. Devos D, Valencia A: **Intrinsic errors in genome annotation**. *Trends Genet* 2001, **17**(8):429-431.
27. Linial M: **How incorrect annotations evolve--the case of short ORFs**. *Trends Biotechnol* 2003, **21**(7):298-300.
28. Thompson JD, Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O: **MACSIMS : Multiple Alignment of Complete Sequences Information Management System**. *BMC Bioinformatics* 2006, **7**(1):318.
29. Sonnhammer EL, Kahn D: **Modular arrangement of proteins as inferred from analysis of homology**. *Protein Sci* 1994, **3**(3):482-492.
30. Chen CP, Kernytsky A, Rost B: **Transmembrane helix predictions revisited**. *Protein Sci* 2002, **11**(12):2774-2791.
31. Ison JC, Blades MJ, Bleasby AJ, Daniel SC, Parish JH, Findlay JB: **Key residues approach to the definition of protein families and analysis of sparse family signatures**. *Proteins* 2000, **40**(2):330-341.
32. Nevill-Manning CG, Wu TD, Brutlag DL: **Highly specific protein sequence motifs for genome analysis**. *Proc Natl Acad Sci U S A* 1998, **95**(11):5865-5871.
33. Lichtarge O, Bourne HR, Cohen FE: **An evolutionary trace method defines binding surfaces common to protein families**. *J Mol Biol* 1996, **257**(2):342-358.
34. Valdar WS: **Scoring residue conservation**. *Proteins* 2002, **48**(2):227-241.
35. Cheng G, Qian B, Samudrala R, Baker D: **Improvement in protein functional site prediction by distinguishing structural and functional constraints on protein family evolution using computational design**. *Nucleic Acids Res* 2005, **33**(18):5861-5867.
36. Chelliah V, Chen L, Blundell TL, Lovell SC: **Distinguishing structural and functional restraints in evolution in order to identify interaction sites**. *J Mol Biol* 2004, **342**(5):1487-1504.
37. Lee S, Lee BC, Kim D: **Prediction of protein secondary structure content using amino acid composition and evolutionary information**. *Proteins* 2006, **62**(4):1107-1114.
38. Rost B: **Review: protein secondary structure prediction continues to rise**. *J Struct Biol* 2001, **134**(2-3):204-218.
39. Heringa J: **Computational methods for protein secondary structure prediction using multiple sequence alignments**. *Curr Protein Pept Sci* 2000, **1**(3):273-301.
40. Garnier J, Gibrat JF, Robson B: **GOR method for predicting protein secondary structure from amino acid sequence**. *Methods Enzymol* 1996, **266**:540-553.
41. Jones DT: **Protein secondary structure prediction based on position-specific scoring matrices**. *J Mol Biol* 1999, **292**(2):195-202.

42. Rost B, Sander C, Schneider R: **PHD--an automatic mail server for protein secondary structure prediction**. *Comput Appl Biosci* 1994, **10**(1):53-60.
43. King RD, Sternberg MJ: **Identification and application of the concepts important for accurate and reliable protein secondary structure prediction**. *Protein Sci* 1996, **5**(11):2298-2310.
44. Lo Conte L, Brenner SE, Hubbard TJ, Chothia C, Murzin AG: **SCOP database in 2002: refinements accommodate structural genomics**. *Nucleic Acids Res* 2002, **30**(1):264-267.
45. Pearl FM, Bennett CF, Bray JE, Harrison AP, Martin N, Shepherd A, Sillitoe I, Thornton J, Orengo CA: **The CATH database: an extended protein family resource for structural and functional genomics**. *Nucleic Acids Res* 2003, **31**(1):452-455.
46. Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M: **Evaluation of comparative protein modeling by MODELLER**. *Proteins* 1995, **23**(3):318-326.
47. Combet C, Jambon M, Deleage G, Geourjon C: **Geno3D: automatic comparative molecular modelling of protein**. *Bioinformatics* 2002, **18**(1):213-214.
48. Schwede T, Kopp J, Guex N, Peitsch MC: **SWISS-MODEL: An automated protein homology-modeling server**. *Nucleic Acids Res* 2003, **31**(13):3381-3385.
49. Bates PA, Kelley LA, MacCallum RM, Sternberg MJ: **Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM**. *Proteins* 2001, **Suppl 5**:39-46.
50. Moult J: **A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction**. *Curr Opin Struct Biol* 2005, **15**(3):285-289.
51. Ramensky V, Bork P, Sunyaev S: **Human non-synonymous SNPs: server and survey**. *Nucleic Acids Res* 2002, **30**(17):3894-3900.
52. Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN: **Human Gene Mutation Database (HGMD): 2003 update**. *Hum Mutat* 2003, **21**(6):577-581.
53. Sunyaev S, Ramensky V, Bork P: **Towards a structural basis of human non-synonymous single nucleotide polymorphisms**. *Trends Genet* 2000, **16**(5):198-200.
54. Fredman D, Munns G, Rios D, Sjoholm F, Siegfried M, Lenhard B, Lehvaslaiho H, Brookes AJ: **HGVbase: a curated resource describing human DNA variation and phenotype relationships**. *Nucleic Acids Res* 2004, **32**(Database issue):D516-519.
55. Irizarry K, Kustanovich V, Li C, Brown N, Nelson S, Wong W, Lee CJ: **Genome-wide analysis of single-nucleotide polymorphisms in human expressed sequences**. *Nat Genet* 2000, **26**(2):233-236.
56. Bao L, Zhou M, Cui Y: **nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W480-482.
57. Ng PC, Henikoff S: **SIFT: Predicting amino acid changes that affect protein function**. *Nucleic Acids Res* 2003, **31**(13):3812-3814.
58. Saunders CT, Baker D: **Evaluation of structural and evolutionary contributions to deleterious mutation prediction**. *J Mol Biol* 2002, **322**(4):891-901.
59. Uetz P, Giot L, Cagney G, Mansfield TA, Judson RS, Knight JR, Lockshon D, Narayan V, Srinivasan M, Pochart P *et al*: **A comprehensive analysis of protein-protein interactions in Saccharomyces cerevisiae**. *Nature* 2000, **403**(6770):623-627.
60. Butland G, Peregrin-Alvarez JM, Li J, Yang W, Yang X, Canadien V, Starostine A, Richards D, Beattie B, Krogan N *et al*: **Interaction network containing conserved and essential protein complexes in Escherichia coli**. *Nature* 2005, **433**(7025):531-537.

61. Giot L, Bader JS, Brouwer C, Chaudhuri A, Kuang B, Li Y, Hao YL, Ooi CE, Godwin B, Vitols E *et al*: **A protein interaction map of Drosophila melanogaster**. *Science* 2003, **302**(5651):1727-1736.
62. Formstecher E, Aresta S, Collura V, Hamburger A, Meil A, Trehin A, Reverdy C, Betin V, Maire S, Brun C *et al*: **Protein interaction mapping: a Drosophila case study**. *Genome Res* 2005, **15**(3):376-384.
63. Cesareni G, Ceol A, Gavrila C, Palazzi LM, Persico M, Schneider MV: **Comparative interactomics**. *FEBS Lett* 2005, **579**(8):1828-1833.
64. Wojcik J, Schachter V: **Protein-protein interaction map inference using interacting domain profile pairs**. *Bioinformatics* 2001, **17 Suppl 1**:S296-305.
65. Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO: **Assigning protein functions by comparative genome analysis: protein phylogenetic profiles**. *Proc Natl Acad Sci U S A* 1999, **96**(8):4285-4288.
66. Marcotte EM, Pellegrini M, Ng HL, Rice DW, Yeates TO, Eisenberg D: **Detecting protein function and protein-protein interactions from genome sequences**. *Science* 1999, **285**(5428):751-753.
67. Bonneau R, Strauss CE, Baker D: **Improving the performance of Rosetta using multiple sequence alignment information and global measures of hydrophobic core formation**. *Proteins* 2001, **43**(1):1-11.
68. Pazos F, Valencia A: **Similarity of phylogenetic trees as indicator of protein-protein interaction**. *Protein Eng* 2001, **14**(9):609-614.
69. Kim Y, Subramaniam S: **Locally defined protein phylogenetic profiles reveal previously missed protein interactions and functional relationships**. *Proteins* 2006, **62**(4):1115-1124.
70. Pazos F, Valencia A: **In silico two-hybrid system for the selection of physically interacting protein pairs**. *Proteins* 2002, **47**(2):219-227.
71. Pazos F, Helmer-Citterich M, Ausiello G, Valencia A: **Correlated mutations contain information about protein-protein interaction**. *J Mol Biol* 1997, **271**(4):511-523.
72. Eisenberg D, Marcotte EM, Xenarios I, Yeates TO: **Protein function in the post-genomic era**. *Nature* 2000, **405**(6788):823-826.
73. Hardison RC: **Comparative genomics**. *PLoS Biol* 2003, **1**(2):E58.
74. Frazer KA, Pachter L, Poliakov A, Rubin EM, Dubchak I: **VISTA: computational tools for comparative genomics**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W273-279.
75. Curwen V, Eyras E, Andrews TD, Clarke L, Mongin E, Searle SM, Clamp M: **The Ensembl automatic gene annotation system**. *Genome Res* 2004, **14**(5):942-950.
76. Hsu F, Pringle TH, Kuhn RM, Karolchik D, Diekhans M, Haussler D, Kent WJ: **The UCSC Proteome Browser**. *Nucleic Acids Res* 2005, **33**(Database issue):D454-458.
77. Hartmann S, Lu D, Phillips J, Vision TJ: **Phytome: a platform for plant comparative genomics**. *Nucleic Acids Res* 2006, **34**(Database issue):D724-730.
78. Rubin GM, Yandell MD, Wortman JR, Gabor Miklos GL, Nelson CR, Hariharan IK, Fortini ME, Li PW, Apweiler R, Fleischmann W *et al*: **Comparative genomics of the eukaryotes**. *Science* 2000, **287**(5461):2204-2215.
79. Huynen M, Dandekar T, Bork P: **Differential genome analysis applied to the species-specific features of Helicobacter pylori**. *FEBS Lett* 1998, **426**(1):1-5.
80. Brenner SE, Chothia C, Hubbard TJ: **Assessing sequence comparison methods with reliable structurally identified distant evolutionary relationships**. *Proc Natl Acad Sci U S A* 1998, **95**(11):6073-6078.
81. Thompson JD, Plewniak F, Poch O: **A comprehensive comparison of multiple sequence alignment programs**. *Nucleic Acids Res* 1999, **27**(13):2682-2690.

82. Lassmann T, Sonnhammer EL: **Quality assessment of multiple alignment programs**. *FEBS Lett* 2002, **529**(1):126-130.
83. Nuin PA, Wang Z, Tillier ER: **The accuracy of several multiple sequence alignment programs for proteins**. *BMC Bioinformatics* 2006, **7**:471.
84. Carrillo H, Lipman D: **The multiple sequence alignment problem in biology**. *SIAM J Appl Math* 1988, **48**(5):1073-1082.
85. Feng DF, Doolittle RF: **Progressive sequence alignment as a prerequisite to correct phylogenetic trees**. *J Mol Evol* 1987, **25**(4):351-360.
86. Smith RF, Smith TF: **Pattern-induced multi-sequence alignment (PIMA) algorithm employing secondary structure-dependent gap penalties for use in comparative protein modelling**. *Protein Eng* 1992, **5**(1):35-41.
87. Barton GJ, Sternberg MJ: **A strategy for the rapid multiple alignment of protein sequences. Confidence levels from tertiary structure comparisons**. *J Mol Biol* 1987, **198**(2):327-337.
88. Taylor WR: **Multiple sequence alignment by a pairwise algorithm**. *Comput Appl Biosci* 1987, **3**(2):81-87.
89. Thompson JD, Higgins DG, Gibson TJ: **CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice**. *Nucleic Acids Res* 1994, **22**(22):4673-4680.
90. Morgenstern B, Atchley WR, Hahn K, Dress A: **Segment-based scores for pairwise and multiple sequence alignments**. *Proc Int Conf Intell Syst Mol Biol* 1998, **6**:115-121.
91. Gotoh O: **Significant improvement in accuracy of multiple protein sequence alignments by iterative refinement as assessed by reference to structural alignments**. *J Mol Biol* 1996, **264**(4):823-838.
92. Notredame C, Higgins DG: **SAGA: sequence alignment by genetic algorithm**. *Nucleic Acids Res* 1996, **24**(8):1515-1524.
93. Notredame C, Holm L, Higgins DG: **COFFEE: an objective function for multiple sequence alignments**. *Bioinformatics* 1998, **14**(5):407-422.
94. Edgar RC, Batzoglou S: **Multiple sequence alignment**. *Curr Opin Struct Biol* 2006, **16**(3):368-373.
95. Thompson JD, Plewniak F, Thierry J, Poch O: **DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches**. *Nucleic Acids Res* 2000, **28**(15):2919-2926.
96. Notredame C, Higgins DG, Heringa J: **T-Coffee: A novel method for fast and accurate multiple sequence alignment**. *J Mol Biol* 2000, **302**(1):205-217.
97. Katoh K, Kuma K, Toh H, Miyata T: **MAFFT version 5: improvement in accuracy of multiple sequence alignment**. *Nucleic Acids Res* 2005, **33**(2):511-518.
98. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity**. *BMC Bioinformatics* 2004, **5**:113.
99. Do CB, Mahabhashyam MS, Brudno M, Batzoglou S: **ProbCons: Probabilistic consistency-based multiple sequence alignment**. *Genome Res* 2005, **15**(2):330-340.
100. O'Sullivan O, Suhre K, Abergel C, Higgins DG, Notredame C: **3DCoffee: combining protein sequences and structures within multiple sequence alignments**. *J Mol Biol* 2004, **340**(2):385-395.
101. Simossis VA, Heringa J: **PRALINE: a multiple sequence alignment toolbox that integrates homology-extended and secondary structure information**. *Nucleic Acids Res* 2005, **33**(Web Server issue):W289-294.

102. Pei J, Grishin NV: **MUMMALS: multiple sequence alignment improved by using hidden Markov models with local structural information**. *Nucleic Acids Res* 2006.
103. Sammeth M, Heringa J: **Global multiple-sequence alignment with repeats**. *Proteins* 2006, **64**(1):263-274.
104. Lee C, Grasso C, Sharlow MF: **Multiple sequence alignment using partial order graphs**. *Bioinformatics* 2002, **18**(3):452-464.
105. Jones NC, Zhi D, Raphael BJ: **AliWABA: alignment on the web through an A-Bruijn approach**. *Nucleic Acids Res* 2006, **34**(Web Server issue):W613-616.
106. Thompson JD, Thierry JC, Poch O: **RASCAL: rapid scanning and correction of multiple sequence alignments**. *Bioinformatics* 2003, **19**(9):1155-1161.
107. Chakrabarti S, Lanczycki CJ, Panchenko AR, Przytycka TM, Thiessen PA, Bryant SH: **Refining multiple sequence alignments with conserved core regions**. *Nucleic Acids Res* 2006, **34**(9):2598-2606.
108. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor**. *Bioinformatics* 2004, **20**(3):426-427.
109. Galtier N, Gouy M, Gautier C: **SEAVIEW and PHYLO_WIN: two graphic tools for sequence alignment and molecular phylogeny**. *Comput Appl Biosci* 1996, **12**(6):543-548.
110. Deleage G, Combet C, Blanchet C, Geourjon C: **ANTHEPROT: an integrated protein sequence analysis software with client/server capabilities**. *Comput Biol Med* 2001, **31**(4):259-267.
111. Parry-Smith DJ, Payne AW, Michie AD, Attwood TK: **CINEMA--a novel colour INteractive editor for multiple alignments**. *Gene* 1998, **221**(1):GC57-63.
112. Hertz GZ, Stormo GD: **Identifying DNA and protein patterns with statistically significant alignments of multiple sequences**. *Bioinformatics* 1999, **15**(7-8):563-577.
113. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O: **Towards a reliable objective function for multiple sequence alignments**. *J Mol Biol* 2001, **314**(4):937-951.
114. Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG: **The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools**. *Nucleic Acids Res* 1997, **25**(24):4876-4882.
115. Cline M, Hughey R, Karplus K: **Predicting reliable regions in protein sequence alignments**. *Bioinformatics* 2002, **18**(2):306-314.
116. Pei J, Grishin NV: **AL2CO: calculation of positional conservation in a protein sequence alignment**. *Bioinformatics* 2001, **17**(8):700-712.
117. Schlosshauer M, Ohlsson M: **A novel approach to local reliability of sequence alignments**. *Bioinformatics* 2002, **18**(6):847-854.
118. Notredame C: **Recent progress in multiple sequence alignment: a survey**. *Pharmacogenomics* 2002, **3**(1):131-144.
119. Przybylski D, Rost B: **Alignments grow, secondary structure prediction improves**. *Proteins* 2002, **46**(2):197-205.
120. Bairoch A, Apweiler R: **The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000**. *Nucleic Acids Res* 2000, **28**(1):45-48.
121. Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R *et al*: **The Universal Protein Resource (UniProt): an expanding universe of protein information**. *Nucleic Acids Res* 2006, **34**(Database issue):D187-191.
122. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE: **The Protein Data Bank**. *Nucleic Acids Res* 2000, **28**(1):235-242.

123. Spalding JB, Lammers PJ: **BLAST Filter and GraphAlign: rule-based formation and analysis of sets of related DNA and protein sequences**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W26-32.
124. Mika S, Rost B: **UniqueProt: Creating representative protein sequence sets**. *Nucleic Acids Res* 2003, **31**(13):3789-3791.
125. Pagni M, Ioannidis V, Cerutti L, Zahn-Zabal M, Jongeneel CV, Falquet L: **MyHits: a new interactive resource for protein annotation and domain identification**. *Nucleic Acids Res* 2004, **32**(Web Server issue):W332-335.
126. Friedrich A, Ripp R, Garnier N, Bettler E, Deleage G, Poch O, Moulinier L: **Blast sampling for structural and functional analyses** *BMC Bioinformatics* 2007, 8:62.
127. Frickey T, Lupas AN: **PhyloGenie: automated phylome generation and analysis**. *Nucleic Acids Res* 2004, **32**(17):5231-5238.
128. O'Donoghue P, Luthey-Schulten Z: **Evolutionary profiles derived from the QR factorization of multiple structural alignments gives an economy of information**. *J Mol Biol* 2005, **346**(3):875-894.
129. Roberts E, Eargle J, Wright D, Luthey-Schulten Z: **MultiSeq: Unifying sequence and structure data for evolutionary analysis**. *BMC Bioinformatics* 2006, **7**(1):382.
130. Mihalek I, Res I, Lichtarge O: **A structure and evolution-guided Monte Carlo sequence selection strategy for multiple alignment-based analysis of proteins**. *Bioinformatics* 2006, **22**(2):149-156.
131. Johnson JM, Mason K, Moallemi C, Xi H, Somaroo S, Huang ES: **Protein family annotation in a multiple alignment viewer**. *Bioinformatics* 2003, **19**(4):544-545.
132. Etzold T, Ulyanov A, Argos P: **SRS: information retrieval system for molecular biology data banks**. *Methods Enzymol* 1996, **266**:114-128.
133. Rigoutsos I, Huynh T, Floratos A, Parida L, Platt D: **Dictionary-driven protein annotation**. *Nucleic Acids Res* 2002, **30**(17):3901-3916.
134. Thompson JD, Holbrook SR, Katoh K, Koehl P, Moras D, Westhof E, Poch O: **MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences**. *Nucleic Acids Res* 2005, **33**(13):4164-4171.
135. Sayle RA, Milner-White EJ: **RASMOL: biomolecular graphics for all**. *Trends Biochem Sci* 1995, **20**(9):374.
136. Li W, Godzik A: **VISSA: a program to visualize structural features from structure sequence alignment**. *Bioinformatics* 2006, **22**(7):887-888.
137. Catherinot V, Labesse G: **ViTO: tool for refinement of protein sequence-structure alignments**. *Bioinformatics* 2004, **20**(18):3694-3696.
138. Abyzov A, Errami M, Leslin CM, Ilyin VA: **Friend, an integrated analytical front-end application for bioinformatics**. *Bioinformatics* 2005, **21**(18):3677-3678.
139. Garnier N, Friedrich A, Bolze R, Bettler E, Moulinier L, Geourjon C, Thompson JD, Deleage G, Poch O: **MAGOS: multiple alignment and modelling server**. *Bioinformatics* 2006, **22**(17):2164-2165.

Figures

Figure 1. The central role of multiple protein sequence alignments in biology.
The main applications directly related to hierarchical Multiple Sequence Alignment (MSA) are represented. These applications are mostly inter-related.

Figure 2. Representation of the comparison of the H. pylori, H. influenzae and E. coli genomes.
The total number of proteins for each species is shown between brackets. 425 orthologs are shared by all three species. The 17 orthologs shared by H. influenzae and H. pylori are responsible for pathogen-specific features.

Figure 3. Overview of different alignment algorithms.

Figure 4. Illustration of MAGOS capabilities.
A. Myotubularin domain organization: the protein, which is 603 residue long, is divided in 3 distinct domains: the PH-GRAM domain from residue 29 to 97, the myotubularin-related domain from residue 192 to 309 and the tyrosine-protein phosphatase domain from residue 344 to 426 which encompasses the active site pocket C(X)5R in black.
B. MAGOS results interface: the annotated alignment is shown on the right, the first sequence being the myotubularin which is interconnected to its model on the left side. The tyrosine-protein phosphatase domain is highlighted in dark grey in the alignment and on the model. The Leucine situated in position 406 in the myotubularin sequence is also highlighted both on the myotubularin sequence in the context of the MSA and on the constructed 3D model.

Figure 5. MSA efficient construction.
Overview of the stages where the user can intervene to reduce the size of his alignment and/or increase its quality.

Figure 1.

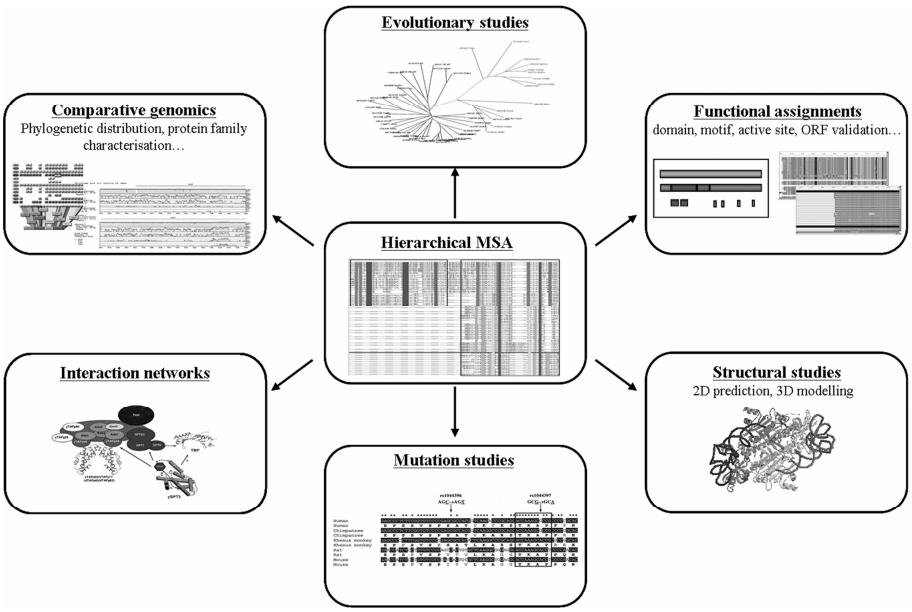


Figure 2.

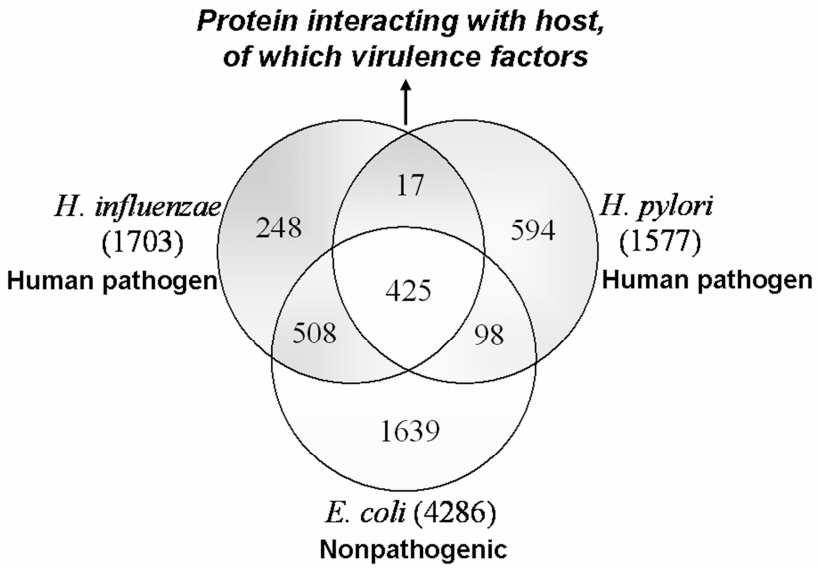


Figure 3.

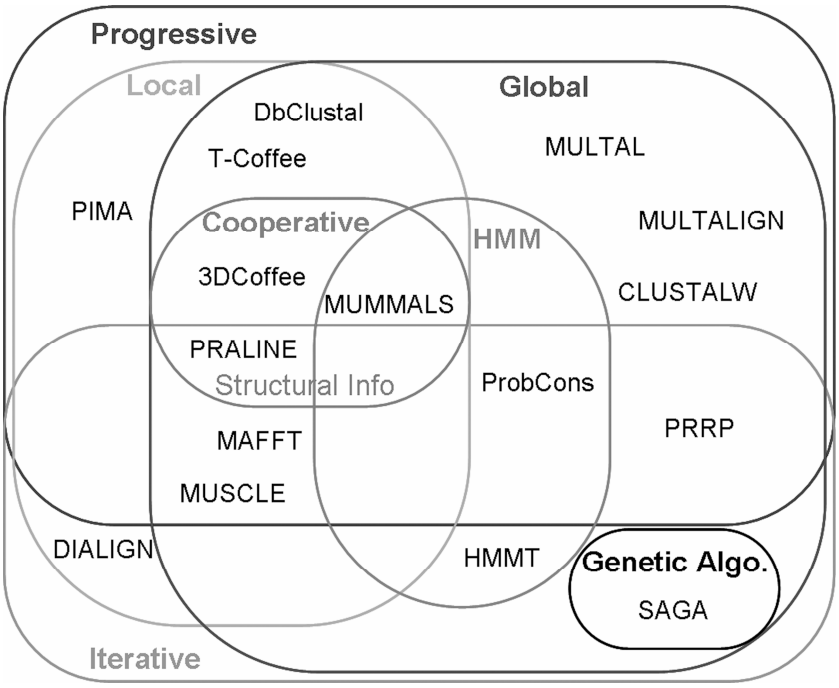


Figure 4A.

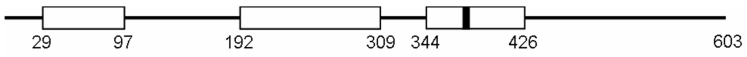


Figure 4B.

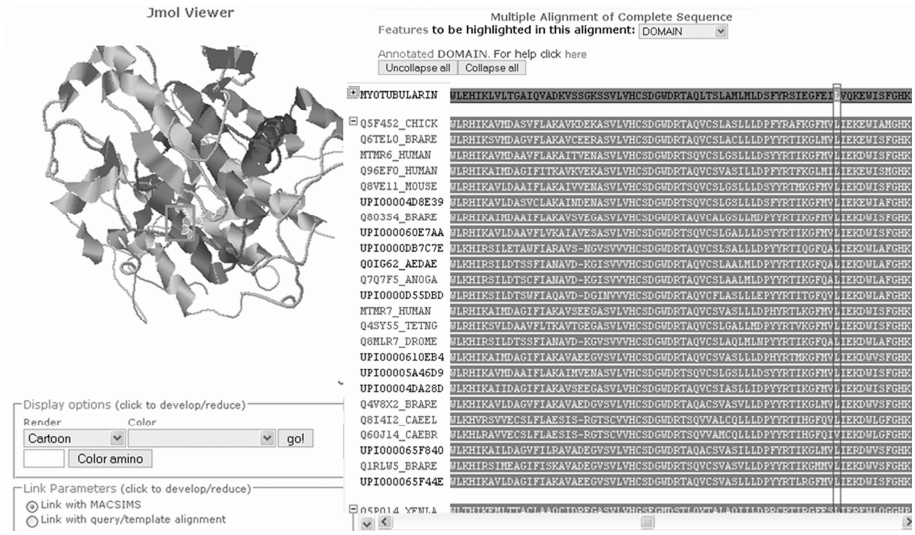
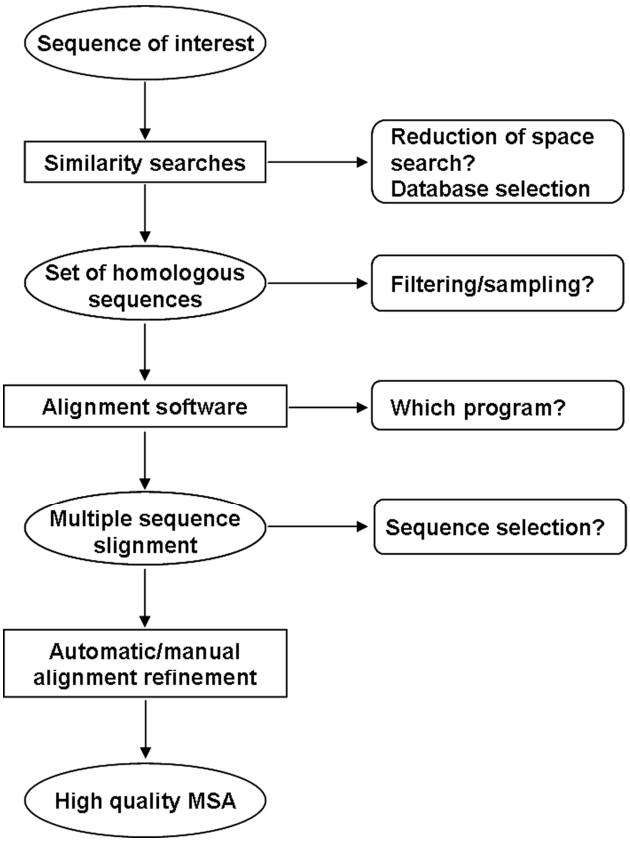


Figure 5.


Sequence of interest
Similarity searches
Reduction of space search?
Database selection
Set of homologous sequences
Filtering/sampling?
Alignment software
Which program?
Multiple sequence slignment
Sequence selection?
Automatic/manual alignment refinement
High quality MSA

Références Bibliographiques

Aartsma-Rus A, Van Deutekom JC, Fokkema IF, Van Ommen GJ, Den Dunnen JT. 2006. Entries in the Leiden Duchenne muscular dystrophy mutation database: an overview of mutation types and paradoxical cases that confirm the reading-frame rule. Muscle Nerve 34(2):135-44.

Adams MD, Celniker SE, Holt RA, Evans CA, Gocayne JD, Amanatides PG, Scherer SE, Li PW, Hoskins RA, Galle RF and others. 2000. The genome sequence of Drosophila melanogaster. Science 287(5461):2185-95.

Adams MD, Kelley JM, Gocayne JD, Dubnick M, Polymeropoulos MH, Xiao H, Merril CR, Wu A, Olde B, Moreno RF and others. 1991. Complementary DNA sequencing: expressed sequence tags and human genome project. Science 252(5013):1651-6.

Altschul SF, Gish W, Miller W, Myers EW, Lipman DJ. 1990. Basic local alignment search tool. J Mol Biol 215(3):403-10.

Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res 25(17):3389-402.

Anderson S, Bankier AT, Barrell BG, de Bruijn MH, Coulson AR, Drouin J, Eperon IC, Nierlich DP, Roe BA, Sanger F and others. 1981. Sequence and organization of the human mitochondrial genome. Nature 290(5806):457-65.

Anfinsen CB, Redfield RR. 1956. Protein structure in relation to function and biosynthesis. Adv Protein Chem 11:1-100.

Antonarakis SE, Krawczak M, Cooper DN. 2000. Disease-causing mutations in the human genome. Eur J Pediatr 159 Suppl 3:S173-8.

Antonarakis SE, Krawczak M, Cooper DN. 2001. The nature and mechanisms of human gene mutation. In: Scriver CR, Beaudet AL, Sly WS, Valle D, editors. The metabolic and molecular bases of inherited disease. New York: McGraw-Hill. p 259-91.

Ashburner M, Ball CA, Blake JA, Botstein D, Butler H, Cherry JM, Davis AP, Dolinski K, Dwight SS, Eppig JT and others. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat Genet 25(1):25-9.

Attwood TK. 2002. The PRINTS database: a resource for identification of protein families. Brief Bioinform 3(3):252-63.

August JT, Shapiro L, Eoyang L. 1965. Replication of Rna Viruses I. Characterization of a Viral Rna-Dependent Rna Polymerase. J Mol Biol 11:257-71.

Avery OT, MacLeod CM, McCart M. 1944. Studies on the chemical nature of the substance inducing transformation of the pneumococcal types: induction of transformation by a desoxyribonucleic acid fraction isolated from pneumococcus type III. J Exp Med 79:137.

Ayme S, Urbero B, Oziel D, Lecouturier E, Biscarat AC. 1998. [Information on rare diseases: the Orphanet project]. Rev Med Interne 19 Suppl 3:376S-377S.

Baltimore D. 1970. RNA-dependent DNA polymerase in virions of RNA tumour viruses. Nature 226(5252):1209-11.

Bao L, Cui Y. 2005. Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics 21(10):2185-90.

Bateman A, Coin L, Durbin R, Finn RD, Hollich V, Griffiths-Jones S, Khanna A, Marshall M, Moxon S, Sonnhammer EL and others. 2004. The Pfam protein families database. Nucleic Acids Res 32(Database issue):D138-41.

Bates PA, Kelley LA, MacCallum RM, Sternberg MJ. 2001. Enhancement of protein modeling by human intervention in applying the automatic programs 3D-JIGSAW and 3D-PSSM. Proteins Suppl 5:39-46.

Berger B, Wilson DB, Wolf E, Tonchev T, Milla M, Kim PS. 1995. Predicting coiled coils by use of pairwise residue correlations. Proc Natl Acad Sci U S A 92(18):8259-63.

Berman HM, Westbrook JD. 2004. The impact of structural genomics on the protein data bank. Am J Pharmacogenomics 4(4):247-52.

Beroud C, Collod-Beroud G, Boileau C, Soussi T, Junien C. 2000. UMD (Universal mutation database): a generic software to build and analyze locus-specific databases. Hum Mutat 15(1):86-94.

Beroud C, Hamroun D, Collod-Beroud G, Boileau C, Soussi T, Claustres M. 2005. UMD (Universal Mutation Database): 2005 update. Hum Mutat 26(3):184-91.

Biancalana V, Caron O, Gallati S, Baas F, Kress W, Novelli G, D'Apice MR, Lagier-Tourenne C, Buj-Bello A, Romero NB and others. 2003. Characterisation of mutations in 77 patients with X-linked myotubular myopathy, including a family with a very mild phenotype. Hum Genet 112(2):135-42.

Bianchetti L, Thompson JD, Lecompte O, Plewniak F, Poch O. 2005. vALId: validation of protein sequence quality based on multiple alignment data. J Bioinform Comput Biol 3(4):929-47.

Boeckmann B, Bairoch A, Apweiler R, Blatter MC, Estreicher A, Gasteiger E, Martin MJ, Michoud K, O'Donovan C, Phan I and others. 2003. The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res 31(1):365-70.

Boguski MS, Lowe TM, Tolstoshev CM. 1993. dbEST--database for "expressed sequence tags". Nat Genet 4(4):332-3.

Bonneau R, Baker D. 2001. Ab initio protein structure prediction: progress and prospects. Annu Rev Biophys Biomol Struct 30:173-89.

Bootsma D, Hoeijmakers JH. 1991. The genetic basis of xeroderma pigmentosum. Ann Genet 34(3-4):143-50.

Bork P, Copley R. 2001. The draft sequences. Filling in the gaps. Nature 409(6822):818-20.

Bowie JU, Reidhaar-Olson JF, Lim WA, Sauer RT. 1990. Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science 247(4948):1306-10.

Bromberg Y, Rost B. 2007. SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res 35(11):3823-35.

Bru C, Courcelle E, Carrere S, Beausse Y, Dalmar S, Kahn D. 2005. The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res 33(Database issue):D212-5.

Brun-Heath I, Taillandier A, Serre JL, Mornet E. 2005. Characterization of 11 novel mutations in the tissue non-specific alkaline phosphatase gene responsible for hypophosphatasia and genotype-phenotype correlations. Mol Genet Metab 84(3):273-7.

Brunger AT, Adams PD, Clore GM, DeLano WL, Gros P, Grosse-Kunstleve RW, Jiang JS, Kuszewski J, Nilges M, Pannu NS and others. 1998. Crystallography & NMR system: A new software suite for macromolecular structure determination. Acta Crystallogr D Biol Crystallogr 54(Pt 5):905-21.

Bryant SH, Lawrence CE. 1993. An empirical energy function for threading protein sequence through the folding motif. Proteins 16(1):92-112.

Care MA, Needham CJ, Bulpitt AJ, Westhead DR. 2007. Deleterious SNP prediction: be mindful of your training data! Bioinformatics 23(6):664-72.

Cartegni L, Chew SL, Krainer AR. 2002. Listening to silence and understanding nonsense: exonic mutations that affect splicing. Nat Rev Genet 3(4):285-98.

Casali P, Pal Z, Xu Z, Zan H. 2006. DNA repair in antibody somatic hypermutation. Trends Immunol 27(7):313-21.

Chakravarti A. 2001. To a future of genetic medicine. Nature 409(6822):822-3.
Chasman D, Adams RM. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol 307(2):683-706.
Chen CP, Kernytsky A, Rost B. 2002. Transmembrane helix predictions revisited. Protein Sci 11(12):2774-91.
Chou PY, Fasman GD. 1974. Prediction of protein conformation. Biochemistry 13(2):222-45.
Chuzhanova NA, Anassis EJ, Ball EV, Krawczak M, Cooper DN. 2003. Meta-analysis of indels causing human genetic disease: mechanisms of mutagenesis and the role of local DNA sequence complexity. Hum Mutat 21(1):28-44.
Clamp M, Cuff J, Searle SM, Barton GJ. 2004. The Jalview Java alignment editor. Bioinformatics 20(3):426-7.
Claustres M, Horaitis O, Vanevski M, Cotton RG. 2002. Time for a unified system of mutation description and reporting: a review of locus-specific mutation databases. Genome Res 12(5):680-8.
Cohen D, Chumakov I, Weissenbach J. 1993. A first-generation physical map of the human genome. Nature 366(6456):698-701.
Cohen JC, Kiss RS, Pertsemlidis A, Marcel YL, McPherson R, Hobbs HH. 2004. Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science 305(5685):869-72.
Collins FS, Brooks LD, Chakravarti A. 1998. A DNA polymorphism discovery resource for research on human genetic variation. Genome Res 8(12):1229-31.
Collins FS, Green ED, Guttmacher AE, Guyer MS. 2003. A vision for the future of genomics research. Nature 422(6934):835-47.
Combet C, Jambon M, Deleage G, Geourjon C. 2002. Geno3D: automatic comparative molecular modelling of protein. Bioinformatics 18(1):213-4.
Consortium1 n. 2007. The Universal Protein Resource (UniProt). Nucleic Acids Res 35(Database issue):D193-7.
Consortium TIH. 2003. The International HapMap Project. Nature 426(6968):789-96.
Cooper DN, Krawczak M. 1989. Cytosine methylation and the fate of CpG dinucleotides in vertebrate genomes. Hum Genet 83(2):181-8.
Cotton RG. 2000. Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. Hum Mutat 15(1):4-6.
Craig DW, Stephan DA. 2005. Applications of whole-genome high-density SNP genotyping. Expert Rev Mol Diagn 5(2):159-70.
Crick FH. 1958. On protein synthesis. Symp Soc Exp Biol 12:138-63.
Culbertson MR. 1999. RNA surveillance. Unforeseen consequences for gene expression, inherited genetic disorders and cancer. Trends Genet 15(2):74-80.
Dayhoff MO. 1965. Computer aids to protein sequence determination. J Theor Biol 8(1):97-112.
Dayhoff MO, Eck RV, Chang MA, Sochard MR. 1965. Atlas of Protein Sequence and Structure. Silver Spring, MD.
Dayhoff MO, Eck RV, Park CM. 1972. A model of evolutionary change in proteins. In: Foundation NBR, editor. Atlas of protein sequence and structure. Washington DC. p 89-99.
de Wind N, Hays JB. 2001. Mismatch repair: praying for genome stability. Curr Biol 11(14):R545-8.
Deininger PL, Batzer MA. 1999. Alu repeats and human disease. Mol Genet Metab 67(3):183-93.

del Sol Mesa A, Pazos F, Valencia A. 2003. Automatic methods for predicting functionally important residues. J Mol Biol 326(4):1289-302.
den Dunnen JT, Antonarakis SE. 2000. Mutation nomenclature extensions and suggestions to describe complex mutations: a discussion. Hum Mutat 15(1):7-12.
Do CB, Mahabhashyam MS, Brudno M, Batzoglou S. 2005. ProbCons: Probabilistic consistency-based multiple sequence alignment. Genome Res 15(2):330-40.
Drake JW. 1991. Spontaneous mutation. Annu Rev Genet 25:125-46.
Edgar RC. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113.
Edgar RC, Batzoglou S. 2006. Multiple sequence alignment. Curr Opin Struct Biol 16(3):368-73.
editorial. 2005. WayStation to HUGOBase. Nat Genet 37(8):783.
editorial. 2007. What is the human variome project? Nat Genet 39(4):423.
Ensenauer RE, Michels VV, Reinke SS. 2005. Genetic testing: practical, ethical, and counseling considerations. Mayo Clin Proc 80(1):63-73.
Errami M, Geourjon C, Deleage G. 2003. Detection of unrelated proteins in sequences multiple alignments by using predicted secondary structures. Bioinformatics 19(4):506-12.
Etzold T, Argos P. 1993. SRS--an indexing and retrieval tool for flat file data libraries. Comput Appl Biosci 9(1):49-57.
Fairbrother WG, Holste D, Burge CB, Sharp PA. 2004. Single nucleotide polymorphism-based validation of exonic splicing enhancers. PLoS Biol 2(9):E268.
Fernie BA, Hobart MJ. 1997. An unusual combined insertion/deletion polymorphism in intron 10 of the human complement C6 gene. Hum Genet 100(1):104-8.
Ferrer-Costa C, Orozco M, de la Cruz X. 2002. Characterization of disease-associated single amino acid polymorphisms in terms of sequence and structure properties. J Mol Biol 315(4):771-86.
Fitch WM. 1970. Distinguishing homologous from analogous proteins. Syst Zool 19(2):99-113.
Fitch WM, Margoliash E. 1967. Construction of phylogenetic trees. Science 155(760):279-84.
Fleischmann RD, Adams MD, White O, Clayton RA, Kirkness EF, Kerlavage AR, Bult CJ, Tomb JF, Dougherty BA, Merrick JM and others. 1995. Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science 269(5223):496-512.
Fokkema IF, den Dunnen JT, Taschner PE. 2005. LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat 26(2):63-8.
Fredman D, Siegfried M, Yuan YP, Bork P, Lehvaslaiho H, Brookes AJ. 2002. HGVbase: a human sequence variation database emphasizing data quality and a broad spectrum of data sources. Nucleic Acids Res 30(1):387-91.
Freimer N, Sabatti C. 2003. The human phenome project. Nat Genet 34(1):15-21.
French S, Robson B. 1983. What is a conservative substitution? Journal of molecular evolution 19(2):171-5.
Frishman D, Argos P. 1995. Knowledge-based protein secondary structure assignment. Proteins 23(4):566-79.
Garnier J, Osguthorpe DJ, Robson B. 1978. Analysis of the accuracy and implications of simple methods for predicting the secondary structure of globular proteins. J Mol Biol 120(1):97-120.
Gerlai R. 2002. Phenomics: fiction or the future? Trends Neurosci 25(10):506-9.

Gerstein MB, Bruce C, Rozowsky JS, Zheng D, Du J, Korbel JO, Emanuelsson O, Zhang ZD, Weissman S, Snyder M. 2007. What is a gene, post-ENCODE? History and updated definition. Genome Res 17(6):669-81.

Giardine B, van Baal S, Kaimakis P, Riemer C, Miller W, Samara M, Kollia P, Anagnou NP, Chui DH, Wajcman H and others. 2007. HbVar database of human hemoglobin variants and thalassemia mutations: 2007 update. Hum Mutat 28(2):206.

Gouy M, Milleret F, Mugnier C, Jacobzone M, Gautier C. 1984. ACNUC: a nucleic acid sequence data base and analysis system. Nucleic Acids Res 12(1 Pt 1):121-7.

Graveley BR. 2001. Alternative splicing: increasing diversity in the proteomic world. Trends Genet 17(2):100-7.

Hamosh A, Scott AF, Amberger JS, Bocchini CA, McKusick VA. 2005. Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res 33(Database issue):D514-7.

Hawley S, Mori CA. 2004. The human genome: a user's guide (2nd Ed.).

Henikoff S, Henikoff JG. 1992. Amino acid substitution matrices from protein blocks. Proc Natl Acad Sci U S A 89(22):10915-9.

Hentze MW, Kulozik AE. 1999. A perfect message: RNA surveillance and nonsense-mediated decay. Cell 96(3):307-10.

Herasse M, Spentchian M, Taillandier A, Keppler-Noreuil K, Fliorito AN, Bergoffen J, Wallerstein R, Muti C, Simon-Bouy B, Mornet E. 2003. Molecular study of three cases of odontohypophosphatasia resulting from heterozygosity for mutations in the tissue non-specific alkaline phosphatase gene. J Med Genet 40(8):605-9.

Heringa J. 2000. Computational methods for protein secondary structure prediction using multiple sequence alignments. Curr Protein Pept Sci 1(3):273-301.

Herman GE, Finegold M, Zhao W, de Gouyon B, Metzenberg A. 1999. Medical complications in long-term survivors with X-linked myotubular myopathy. J Pediatr 134(2):206-14.

Hermjakob H. 2006. The HUPO Proteomics Standards Initiative - Overcoming the Fragmentation of Proteomics Data. Proteomics 6 Suppl 2:34-8.

Higgins DG, Sharp PM. 1988. CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene 73(1):237-44.

Hobbs HH, Brown MS, Goldstein JL. 1992. Molecular genetics of the LDL receptor gene in familial hypercholesterolemia. Hum Mutat 1(6):445-66.

Hoffman EP, Brown RH, Jr., Kunkel LM. 1987. Dystrophin: the protein product of the Duchenne muscular dystrophy locus. Cell 51(6):919-28.

Holm L, Sander C. 1994. The FSSP database of structurally aligned protein fold families. Nucleic Acids Res 22(17):3600-9.

Horaitis O, Cotton RG. 2004. The challenge of documenting mutation across the genome: the human genome variation society approach. Hum Mutat 23(5):447-52.

Hoylaerts MF, Manes T, Millan JL. 1997. Mammalian alkaline phosphatases are allosteric enzymes. J Biol Chem 272(36):22781-7.

Hubbard ST, Thornton JM. 1993. NACCESS, computer program. Department Biochemistry and Molecular Biology, University College, London.

Hudson TJ. 2003. Wanted: regulatory SNPs. Nat Genet 33(4):439-40.

Hulo N, Bairoch A, Bulliard V, Cerutti L, De Castro E, Langendijk-Genevaux PS, Pagni M, Sigrist CJ. 2006. The PROSITE database. Nucleic Acids Res 34(Database issue):D227-30.

Ideker T, Valencia A. 2006. Bioinformatics in the human interactome project. Bioinformatics 22(24):2973-4.

IHGSC. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-45.
Jimenez-Sanchez G, Childs B, Valle D. 2001. Human disease genes. Nature 409(6822):853-5.
Jirtle RL, Skinner MK. 2007. Environmental epigenomics and disease susceptibility. Nat Rev Genet 8(4):253-62.
Johnson JM, Castle J, Garrett-Engele P, Kan Z, Loerch PM, Armour CD, Santos R, Schadt EE, Stoughton R, Shoemaker DD. 2003. Genome-wide survey of human alternative pre-mRNA splicing with exon junction microarrays. Science 302(5653):2141-4.
Jones DT. 1999. Protein secondary structure prediction based on position-specific scoring matrices. J Mol Biol 292(2):195-202.
Kabsch W, Sander C. 1983. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers 22(12):2577-637.
Kampa D, Cheng J, Kapranov P, Yamanaka M, Brubaker S, Cawley S, Drenkow J, Piccolboni A, Bekiranov S, Helt G and others. 2004. Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res 14(3):331-42.
Katoh K, Kuma K, Toh H, Miyata T. 2005. MAFFT version 5: improvement in accuracy of multiple sequence alignment. Nucleic Acids Res 33(2):511-8.
Kawabata T, Ota M, Nishikawa K. 1999. The Protein Mutant Database. Nucleic Acids Res 27(1):355-7.
Keen J, Lester D, Inglehearn C, Curtis A, Bhattacharya S. 1991. Rapid detection of single base mismatches as heteroduplexes on Hydrolink gels. Trends Genet 7(1):5.
Kondrashov AS, Sunyaev S, Kondrashov FA. 2002. Dobzhansky-Muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 99(23):14878-83.
Kouranov A, Xie L, de la Cruz J, Chen L, Westbrook J, Bourne PE, Berman HM. 2006. The RCSB PDB information portal for structural genomics. Nucleic Acids Res 34(Database issue):D302-5.
Kozlowski P, Krzyzosiak WJ. 2001. Combined SSCP/duplex analysis by capillary electrophoresis for more efficient mutation detection. Nucleic Acids Res 29(14):E71.
Krawczak M, Ball EV, Cooper DN. 1998. Neighboring-nucleotide effects on the rates of germ-line single-base-pair substitution in human genes. Am J Hum Genet 63(2):474-88.
Krawczak M, Thomas NS, Hundrieser B, Mort M, Wittig M, Hampe J, Cooper DN. 2007. Single base-pair substitutions in exon-intron junctions of human genes: nature, distribution, and consequences for mRNA splicing. Hum Mutat 28(2):150-8.
Kunkel TA. 2004. DNA replication fidelity. J Biol Chem 279(17):16895-8.
Lamesch P, Li N, Milstein S, Fan C, Hao T, Szabo G, Hu Z, Venkatesan K, Bethel G, Martin P and others. 2007. hORFeome v3.1: a resource of human open reading frames representing over 10,000 human genes. Genomics 89(3):307-15.
Lander ES, Linton LM, Birren B, Nusbaum C, Zody MC, Baldwin J, Devon K, Dewar K, Doyle M, FitzHugh W and others. 2001. Initial sequencing and analysis of the human genome. Nature 409(6822):860-921.
Laporte J, Bedez F, Bolino A, Mandel JL. 2003. Myotubularins, a large disease-associated family of cooperating catalytically active and inactive phosphoinositides phosphatases. Hum Mol Genet 12 Spec No 2:R285-92.
Laporte J, Biancalana V, Tanner SM, Kress W, Schneider V, Wallgren-Pettersson C, Herger F, Buj-Bello A, Blondeau F, Liechti-Gallati S and others. 2000. MTM1 mutations in X-linked myotubular myopathy. Hum Mutat 15(5):393-409.

Laporte J, Hu LJ, Kretz C, Mandel JL, Kioschis P, Coy JF, Klauck SM, Poustka A, Dahl N. 1996. A gene mutated in X-linked myotubular myopathy defines a new putative tyrosine phosphatase family conserved in yeast. Nat Genet 13(2):175-82.

Lecompte O, Ripp R, Puzos-Barbe V, Duprat S, Heilig R, Dietrich J, Thierry JC, Poch O. 2001a. Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res 11(6):981-93.

Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. 2001b. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene 270(1-2):17-30.

Lee S, Lee BC, Kim D. 2006. Prediction of protein secondary structure content using amino acid composition and evolutionary information. Proteins 62(4):1107-14.

Lejeune J, Turpin R, Gautier M. 1959. [Mongolism; a chromosomal disease (trisomy).]. Bull Acad Natl Med 143(11-12):256-65.

Letourneau IJ, Deeley RG, Cole SP. 2005. Functional characterization of non-synonymous single nucleotide polymorphisms in the gene encoding human multidrug resistance protein 1 (MRP1/ABCC1). Pharmacogenet Genomics 15(9):647-57.

Letunic I, Copley RR, Pils B, Pinkert S, Schultz J, Bork P. 2006. SMART 5: domains in the context of genomes and networks. Nucleic Acids Res 34(Database issue):D257-60.

Lia-Baldini AS, Muller F, Taillandier A, Gibrat JF, Mouchard M, Robin B, Simon-Bouy B, Serre JL, Aylsworth AS, Bieth E and others. 2001. A molecular approach to dominance in hypophosphatasia. Hum Genet 109(1):99-108.

Linding R, Russell RB, Neduva V, Gibson TJ. 2003. GlobPlot: Exploring protein sequences for globularity and disorder. Nucleic Acids Res 31(13):3701-8.

Lipman DJ, Pearson WR. 1985. Rapid and sensitive protein similarity searches. Science 227(4693):1435-41.

Llinas P, Masella M, Stigbrand T, Menez A, Stura EA, Le Du MH. 2006. Structural studies of human alkaline phosphatase in complex with strontium: implication for its secondary effect in bones. Protein Sci 15(7):1691-700.

Lockless SW, Ranganathan R. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286(5438):295-9.

Lupas A, Van Dyke M, Stock J. 1991. Predicting coiled coils from protein sequences. Science 252(5010):1162-4.

Makalowski W. 2000. Genomic scrap yard: how genomes utilize all that junk. Gene 259(1-2):61-7.

Makalowski W. 2003. Genomics. Not junk after all. Science 300(5623):1246-7.

Maki H. 2002. Origins of spontaneous mutations: specificity and directionality of base-substitution, frameshift, and sequence-substitution mutageneses. Annu Rev Genet 36:279-303.

Matthaei JH, Jones OW, Martin RG, Nirenberg MW. 1962. Characteristics and composition of RNA coding units. Proc Natl Acad Sci U S A 48:666-77.

Mattick JS, Makunin IV. 2006. Non-coding RNA. Hum Mol Genet 15 Spec No 1:R17-29.

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

McEntagart M, Parsons G, Buj-Bello A, Biancalana V, Fenton I, Little M, Krawczak M, Thomas N, Herman G, Clarke A and others. 2002. Genotype-phenotype correlations in X-linked myotubular myopathy. Neuromuscul Disord 12(10):939-46.

McKusick VA. 1998. Mendelian Inheritance in Man. A Catalog of Human Genes and Genetic Disorders. 12th edition: Johns Hopkins University Press.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80(4):588-604.

McKusick VA, Ruddle FH. 1987. Toward a complete map of the human genome. Genomics 1(2):103-6.
Mendel G. 1866. Versuche über pflanzen-hybriden. Verhandlungen des Naturforschenden Vereins 4:3-47.
Meyers LA. 2005. Constraints on variation from genotype through phenotype to fitness. In: B. H, K. HB, editors. Variation, a central concept in biology. p 87-111.
Mighell AJ, Smith NR, Robinson PA, Markham AF. 2000. Vertebrate pseudogenes. FEBS Lett 468(2-3):109-14.
Mika S, Rost B. 2003. UniqueProt: Creating representative protein sequence sets. Nucleic Acids Res 31(13):3789-91.
Miller S, Janin J, Lesk AM, Chothia C. 1987. Interior and surface of monomeric proteins. J Mol Biol 196(3):641-56.
Modrek B, Lee C. 2002. A genomic view of alternative splicing. Nat Genet 30(1):13-9.
Monaco AP, Neve RL, Colletti-Feener C, Bertelson CJ, Kurnit DM, Kunkel LM. 1986. Isolation of candidate cDNAs for portions of the Duchenne muscular dystrophy gene. Nature 323(6089):646-50.
Mooney S. 2005. Bioinformatics approaches and resources for single nucleotide polymorphism functional analysis. Brief Bioinform 6(1):44-56.
Mornet E, Simon-Bouy B. 2004. [Genetics of hypophosphatasia]. Arch Pediatr 11(5):444-8.
Mornet E, Stura E, Lia-Baldini AS, Stigbrand T, Menez A, Le Du MH. 2001. Structural evidence for a functional role of human tissue nonspecific alkaline phosphatase in bone mineralization. J Biol Chem 276(33):31171-8.
Morton NE. 1955. Sequential tests for the detection of linkage. Am J Hum Genet 7(3):277-318.
Moult J. 2005. A decade of CASP: progress, bottlenecks and prognosis in protein structure prediction. Curr Opin Struct Biol 15(3):285-9.
Mulder NJ, Apweiler R, Attwood TK, Bairoch A, Bateman A, Binns D, Bork P, Buillard V, Cerutti L, Copley R and others. 2007. New developments in the InterPro database. Nucleic Acids Res 35(Database issue):D224-8.
Muntoni F, Torelli S, Ferlini A. 2003. Dystrophin and mutations: one gene, several proteins, multiple phenotypes. Lancet Neurol 2(12):731-40.
Murzin AG, Brenner SE, Hubbard T, Chothia C. 1995. SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol 247(4):536-40.
Needleman SB, Wunsch CD. 1970. A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol 48(3):443-53.
Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812-4.
Olivieri NF. 1999. The beta-thalassemias. N Engl J Med 341(2):99-109.
Orengo CA, Michie AD, Jones S, Jones DT, Swindells MB, Thornton JM. 1997. CATH--a hierarchic classification of protein domain structures. Structure 5(8):1093-108.
Orita M, Iwahana H, Kanazawa H, Hayashi K, Sekiya T. 1989. Detection of polymorphisms of human DNA by gel electrophoresis as single-strand conformation polymorphisms. Proc Natl Acad Sci U S A 86(8):2766-70.
Pasternak JJ. 1999. An introduction to human molecular genetics : mechanisms of inheritated diseases: Fitzgerald Science Press.
Pastinen T, Sladek R, Gurd S, Sammak A, Ge B, Lepage P, Lavergne K, Villeneuve A, Gaudin T, Brandstrom H and others. 2004. A survey of genetic and epigenetic variation affecting human gene expression. Physiol Genomics 16(2):184-93.

Pazos F, Valencia A. 2002. In silico two-hybrid system for the selection of physically interacting protein pairs. Proteins 47(2):219-27.
Pearson H. 2006. Genetics: what is a gene? Nature 441(7092):398-401.
Pearson PL. 1991. The genome data base (GDB)--a human gene mapping repository. Nucleic Acids Res 19 Suppl:2237-9.
Pearson WR, Lipman DJ. 1988. Improved tools for biological sequence comparison. Proc Natl Acad Sci U S A 85(8):2444-8.
Peitsch MC, Schwede T, Guex N. 2000. Automated protein modelling--the proteome in 3D. Pharmacogenomics 1(3):257-66.
Pellegrini M, Marcotte EM, Thompson MJ, Eisenberg D, Yeates TO. 1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. Proc Natl Acad Sci U S A 96(8):4285-8.
Pennisi E. 2000. Human Genome Project. And the gene number is...? Science 288(5469):1146-7.
Pennisi E. 2007. Genetics. Working the (gene count) numbers: finally, a firm answer? Science 316(5828):1113.
Peri S, Navarro JD, Amanchy R, Kristiansen TZ, Jonnalagadda CK, Surendranath V, Niranjan V, Muthusamy B, Gandhi TK, Gronborg M and others. 2003. Development of human protein reference database as an initial platform for approaching systems biology in humans. Genome Res 13(10):2363-71.
Phillips A, Janies D, Wheeler W. 2000. Multiple sequence alignment in phylogenetic analysis. Mol Phylogenet Evol 16(3):317-30.
Plewniak F, Bianchetti L, Brelivet Y, Carles A, Chalmel F, Lecompte O, Mochel T, Moulinier L, Muller A, Muller J and others. 2003. PipeAlign: A new toolkit for protein family analysis. Nucleic Acids Res 31(13):3829-32.
Plewniak F, Thompson JD, Poch O. 2000. Ballast: blast post-processing based on locally conserved segments. Bioinformatics 16(9):750-9.
Prilusky J, Felder CE, Zeev-Ben-Mordehai T, Rydberg EH, Man O, Beckmann JS, Silman I, Sussman JL. 2005. FoldIndex: a simple tool to predict whether a given protein sequence is intrinsically unfolded. Bioinformatics 21(16):3435-8.
Pruess M, Fleischmann W, Kanapin A, Karavidopoulou Y, Kersey P, Kriventseva E, Mittard V, Mulder N, Phan I, Servant F and others. 2003. The Proteome Analysis database: a tool for the in silico analysis of whole proteomes. Nucleic Acids Res 31(1):414-7.
Przybylski D, Rost B. 2002. Alignments grow, secondary structure prediction improves. Proteins 46(2):197-205.
Ramensky V, Bork P, Sunyaev S. 2002. Human non-synonymous SNPs: server and survey. Nucleic Acids Res 30(17):3894-900.
Rebhan M, Chalifa-Caspi V, Prilusky J, Lancet D. 1997. GeneCards: integrating information about genes, proteins and diseases. Trends Genet 13(4):163.
Riikonen P, Vihinen M. 1999. MUTbase: maintenance and analysis of distributed mutation databases. Bioinformatics 15(10):852-9.
Ring HZ, Kwok PY, Cotton RG. 2006. Human Variome Project: an international collaboration to catalogue human genetic variation. Pharmacogenomics 7(7):969-72.
Ross J. 1995. mRNA stability in mammalian cells. Microbiol Rev 59(3):423-50.
Rost B. 1996. PHD: predicting one-dimensional protein structure by profile-based neural networks. Methods Enzymol 266:525-39.
Rost B. 2001. Review: protein secondary structure prediction continues to rise. J Struct Biol 134(2-3):204-18.
Rost B, Liu J. 2003. The PredictProtein server. Nucleic Acids Res 31(13):3300-4.

Rubin DL, Lewis SE, Mungall CJ, Misra S, Westerfield M, Ashburner M, Sim I, Chute CG, Solbrig H, Storey MA and others. 2006. National Center for Biomedical Ontology: advancing biomedicine through structured organization of scientific knowledge. Omics 10(2):185-98.

Salamov AA, Solovyev VV. 1997. Protein secondary structure prediction using local alignments. J Mol Biol 268(1):31-6.

Sali A, Potterton L, Yuan F, van Vlijmen H, Karplus M. 1995. Evaluation of comparative protein modeling by MODELLER. Proteins 23(3):318-26.

Sander C, Schneider R. 1991. Database of homology-derived protein structures and the structural meaning of sequence alignment. Proteins 9(1):56-68.

Sanger F, Coulson AR, Friedmann T, Air GM, Barrell BG, Brown NL, Fiddes JC, Hutchison CA, 3rd, Slocombe PM, Smith M. 1978. The nucleotide sequence of bacteriophage phiX174. J Mol Biol 125(2):225-46.

Sanger F, Nicklen S, Coulson AR. 1977. DNA sequencing with chain-terminating inhibitors. Proc Natl Acad Sci U S A 74(12):5463-7.

Sanger F, Thompson EO. 1953a. The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem J 53(3):353-66.

Sanger F, Thompson EO. 1953b. The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. Biochem J 53(3):366-74.

Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322(4):891-901.

Sayle RA, Milner-White EJ. 1995. RASMOL: biomolecular graphics for all. Trends Biochem Sci 20(9):374.

Schwede T, Kopp J, Guex N, Peitsch MC. 2003. SWISS-MODEL: An automated protein homology-modeling server. Nucleic Acids Res 31(13):3381-5.

Scriver CR. 2004. After the genome--the phenome? J Inherit Metab Dis 27(3):305-17.

Scriver CR, Nowacki PM, Lehvaslaiho H. 1999. Guidelines and recommendations for content, structure, and deployment of mutation databases. Hum Mutat 13(5):344-50.

Sherry ST, Ward M, Sirotkin K. 1999. dbSNP-database for single nucleotide polymorphisms and other classes of minor genetic variation. Genome Res 9(8):677-9.

Shindyalov IN, Bourne PE. 1998. Protein structure alignment by incremental combinatorial extension (CE) of the optimal path. Protein Eng 11(9):739-47.

Sjolander K, Karplus K, Brown M, Hughey R, Krogh A, Mian IS, Haussler D. 1996. Dirichlet mixtures: a method for improved detection of weak but significant protein sequence homology. Comput Appl Biosci 12(4):327-45.

Smith TF, Waterman MS. 1981. Identification of common molecular subsequences. J Mol Biol 147(1):195-7.

Spalding JB, Lammers PJ. 2004. BLAST Filter and GraphAlign: rule-based formation and analysis of sets of related DNA and protein sequences. Nucleic Acids Res 32(Web Server issue):W26-32.

Spentchian M, Brun-Heath I, Taillandier A, Fauvert D, Serre JL, Simon-Bouy B, Carvalho F, Grochova I, Mehta SG, Muller G and others. 2006. Characterization of missense mutations and large deletions in the ALPL gene by sequencing and quantitative multiplex PCR of short fragments. Genet Test 10(4):252-7.

Stenson PD, Ball EV, Mort M, Phillips AD, Shiel JA, Thomas NS, Abeysinghe S, Krawczak M, Cooper DN. 2003. Human Gene Mutation Database (HGMD): 2003 update. Hum Mutat 21(6):577-81.

Steward RE, MacArthur MW, Laskowski RA, Thornton JM. 2003. Molecular basis of inherited diseases: a structural perspective. Trends Genet 19(9):505-13.

Stitziel NO, Tseng YY, Pervouchine D, Goddeau D, Kasif S, Liang J. 2003. Structural location of disease-associated single-nucleotide polymorphisms. J Mol Biol 327(5):1021-30.

Sunyaev SR, Eisenhaber F, Rodchenkov IV, Eisenhaber B, Tumanyan VG, Kuznetsov EN. 1999. PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng 12(5):387-94.

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

Tarczy-Hornoch P, Shannon P, Baskin P, Espeseth M, Pagon RA. 2000. GeneClinics: a hybrid text/data electronic publishing model using XML applied to clinical genetic testing. J Am Med Inform Assoc 7(3):267-76.

Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN and others. 2003. The COG database: an updated version includes eukaryotes. BMC Bioinformatics 4:41.

Taylor WR. 1986. The classification of amino acid conservation. J Theor Biol 119(2):205-18.

Taylor WR. 1999. Protein structural domain identification. Protein Eng 12(3):203-16.

Temin HM, Mizutani S. 1970. RNA-dependent DNA polymerase in virions of Rous sarcoma virus. Nature 226(5252):1211-3.

Thompson JD, Gibson TJ, Plewniak F, Jeanmougin F, Higgins DG. 1997. The CLUSTAL_X windows interface: flexible strategies for multiple sequence alignment aided by quality analysis tools. Nucleic Acids Res 25(24):4876-82.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res 22(22):4673-80.

Thompson JD, Holbrook SR, Katoh K, Koehl P, Moras D, Westhof E, Poch O. 2005. MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Res 33(13):4164-71.

Thompson JD, Muller A, Waterhouse A, Procter J, Barton GJ, Plewniak F, Poch O. 2006. MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics 7:318.

Thompson JD, Plewniak F, Poch O. 1999. A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 27(13):2682-90.

Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O. 2001. Towards a reliable objective function for multiple sequence alignments. J Mol Biol 314(4):937-51.

Thompson JD, Plewniak F, Thierry J, Poch O. 2000. DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res 28(15):2919-26.

Thompson JD, Prigent V, Poch O. 2004. LEON: multiple aLignment Evaluation Of Neighbours. Nucleic Acids Res 32(4):1298-307.

Thompson JD, Thierry JC, Poch O. 2003. RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 19(9):1155-61.

Tijo TH, Levan A. 1956. The chromosome number of man. Hereditas 42:1-6.

Tronchere H, Laporte J, Pendaries C, Chaussade C, Liaubet L, Pirola L, Mandel JL, Payrastre B. 2004. Production of phosphatidylinositol 5-phosphate by the phosphoinositide 3-phosphatase myotubularin in mammalian cells. J Biol Chem 279(8):7304-12.

Tsai TC, Horinouchi H, Noguchi S, Minami N, Murayama K, Hayashi YK, Nonaka I, Nishino I. 2005. Characterization of MTM1 mutations in 31 Japanese families with myotubular myopathy, including a patient carrying 240 kb deletion in Xq28 without male hypogenitalism. Neuromuscul Disord 15(3):245-52.

Umar A, Kunkel TA. 1996. DNA-replication fidelity, mismatch repair and genome instability in cancer cells. Eur J Biochem 238(2):297-307.
Valdar WS. 2002. Scoring residue conservation. Proteins 48(2):227-41.
Vanin EF. 1985. Processed pseudogenes: characteristics and evolution. Annu Rev Genet 19:253-72.
Vapnik VN. 1998. Statistical Learning Theory. New-York: Wiley-Interscience.
Venter JC, Adams MD, Myers EW, Li PW, Mural RJ, Sutton GG, Smith HO, Yandell M, Evans CA, Holt RA and others. 2001. The sequence of the human genome. Science 291(5507):1304-51.
Wakeley J. 1994. Substitution-rate variation among sites and the estimation of transition bias. Mol Biol Evol 11(3):436-42.
Wang Y, Liu CL, Storey JD, Tibshirani RJ, Herschlag D, Brown PO. 2002. Precision and functional specificity in mRNA decay. Proc Natl Acad Sci U S A 99(9):5860-5.
Ward JJ, McGuffin LJ, Bryson K, Buxton BF, Jones DT. 2004. The DISOPRED server for the prediction of protein disorder. Bioinformatics 20(13):2138-9.
Watson JD, Crick FH. 1953. Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 171(4356):737-8.
Whyte MP. 1994. Hypophosphatasia and the role of alkaline phosphatase in skeletal mineralization. Endocr Rev 15(4):439-61.
Wicker N, Dembele D, Raffelsberger W, Poch O. 2002. Density of points clustering, application to transcriptomic data analysis. Nucleic Acids Res 30(18):3992-4000.
Wicker N, Perrin GR, Thierry JC, Poch O. 2001. Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol Biol Evol 18(8):1435-41.
Wittkopp PJ, Haerum BK, Clark AG. 2004. Evolutionary changes in cis and trans gene regulation. Nature 430(6995):85-8.
Woese CR, Kandler O, Wheelis ML. 1990. Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya. Proc Natl Acad Sci U S A 87(12):4576-9.
Wolf E, Kim PS, Berger B. 1997. MultiCoil: a program for predicting two- and three-stranded coiled coils. Protein Sci 6(6):1179-89.
Wright PE, Dyson HJ. 1999. Intrinsically unstructured proteins: re-assessing the protein structure-function paradigm. J Mol Biol 293(2):321-31.
Wu CH, Apweiler R, Bairoch A, Natale DA, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R and others. 2006. The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res 34(Database issue):D187-91.
Yang AS, Honig B. 2000. An integrated approach to the analysis and modeling of protein sequences and structures. III. A comparative study of sequence conservation in protein structural families using multiple structural alignments. J Mol Biol 301(3):691-711.
Zurutuza L, Muller F, Gibrat JF, Taillandier A, Simon-Bouy B, Serre JL, Mornet E. 1999. Correlations of genotype and phenotype in hypophosphatasia. Hum Mol Genet 8(6):1039-46.

Résumé

Le travail présenté dans cette thèse concerne l'étude, par des approches bioinformatiques à haut débit, des relations qui lient le phénotype d'un individu à son génotype, dans le cadre du projet MS2PH (« de la Mutation Structurale aux Phénotypes des Pathologies Humaines »).

Nous nous sommes tout d'abord concentrés sur la caractérisation des mutations dans un contexte évolutif et avons développé une stratégie optimale de construction d'alignements multiples de séquences protéiques dédiés aux analyses structurales. Puis, la caractérisation des mutations dans un contexte fonctionnel et structural a été abordée en déployant un serveur Web, MAGOS, réunissant les informations de séquence, structure, fonction et évolution relatives à une protéine d'intérêt. Enfin, nous nous sommes appuyés sur les capacités intégratives de MAGOS, ainsi que sur la puissance de la grille de calcul universitaire Décrypthon (www.decrypthon.fr) pour développer MS2PH-db, une banque de données dédiée aux protéines impliquées dans les maladies monogéniques humaines, qui intègre de plus des données « cliniques » (mutations et phénotypes associés).

Les règles d'association de base entre séquence, mutations, impacts structuraux et phénotypes cliniques associés sont finalement discutées, en se basant sur l'étude de deux protéines dont la relation génotype/phénotype est en partie élucidée. Ces études ouvrent la voie vers la mise en place d'un système générique de prédiction du lien entre mutation et phénotype clinique.

Abstract

This thesis work focuses on high throughput applications of bioinformatics methodologies to study genotype/phenotype correlations in the context of the MS2PH project ("from Structural Mutation to Human Pathologies Phenotypes").

In an initial step, we concentrated on the characterization of mutations in an evolutionary context and we have developed an optimal strategy for the construction of multiple alignments of protein sequences dedicated to structural analysis. Next, the characterization of the mutations in a functional and structural context was addressed by the development of MAGOS, a Web server which allows the coupling of the sequence, structural, functional and evolutionary information related to a given protein of interest. Finally, we exploited the integrative capabilities of MAGOS, as well as the power of the Decrypthon computation grid (www.decrypthon.fr) to develop MS2PH-db, a database dedicated to proteins involved in human monogenic diseases, which also integrates clinical data (mutations and associated phenotypes).

In the final section, the basic associative rules between sequence, mutation, structural impact and clinical phenotypes are discussed, illustrated by the analysis of two well-studied proteins for which the genotype/phenotype correlations are partially elucidated. These studies open up the way to the development of a system dedicated to the prediction of the link between mutations and clinical phenotypes.

Mots clés/Keywords

Maladie génétique humaine/human genetic disease, génotype/genotype, phénotype/phenotype, analyse de séquence/sequence analysis, analyse de structure/structural analysis, grille de calcul/grid computing, MS2PH, MS2PH-db, MAGOS.

Laboratoire/Laboratory

Laboratoire de Bioinformatique et de Génomique Intégratives (UMR7104), IGBMC, 1 rue Laurent Fries, 67404 Illkirch-Graffenstaden cedex, France.