UNIVERSITÉ DE STRASBOURG

Institut de Génétique et de Biologie Moléculaire et Cellulaire

Département de Biologie et de Génomique Structurale

Laboratoire de Bioinformatique et de Génomique Intégratives

Thèse présentée pour obtenir le grade de Docteur de l'Université de Strasbourg

Discipline : Sciences du vivant
Spécialité : Bioinformatique
par Nicolas GAGNIÈRE

Développement d'une suite logicielle pour l'analyse et l'annotation intégrative automatiques de transcrits et de protéines. Application aux banques d'ADNc de l'annélide polychète *Alvinella pompejana*

Soutenue publiquement le 13 octobre 2009

Membres du jury

Directeur :	LECOMPTE Odile, Maître de Conférences UDS
Co-directeur :	POCH Olivier, Directeur de Recherche CNRS
Rapporteur interne :	CAVARELLI Jean, Professeur UDS
Rapporteur externe :	MÉDIGUE Claudine, Directeur de Recherche CNRS
Rapporteur externe :	ROBINSON-RECHAVI Marc, Professeur à l'Université de Lausanne
Examinateur :	MOSZER Ivan, Chargé de Recherche à l'Institut Pasteur
Membre invité :	JOLLIVET Didier, Chargé de Recherche CNRS

Finir sa thèse, c'est comme essayer d'atteindre la vitesse de la lumière. Plus on s'en rapproche, plus cela nous demande de l'énergie…

—un nain connu très fatigué

REMERCIEMENTS

Tout d'abord, je tiens à exprimer ma profonde reconnaissance à Jean Cavarelli, Claudine Médigue, Marc Robinson-Rechavi et Ivan Moszer pour avoir accepté de juger ce travail de thèse.

Merci aussi à Didier Jollivet pour avoir accepté de venir assister à mon show, mais surtout merci d'être parti à l'aventure pour aller pêcher ces petites bêtes qui ont occupé mes jours (et hanté mes nuits ?).

Je remercie également Dino Moras et Jean-Claude Thierry de m'avoir accepté au sein du Département de Biologie et de Génomique Structurale au cours de mes nombreux stages et de ces trois dernières années.

Un immense merci à mes deux co-directeurs préférés, Odile et Olivier, qui m'ont porté, supporté, soutenu et retenu pendant cette poignée d'années, mais surtout qui ont réussi à me motiver et à me faire prendre confiance en moi à chaque moment difficile. Odile, une bonne partie de ce que je sais maintenant je te le dois, et je n'aurais jamais pu arriver au bout sans ton aide. Merci beaucoup pour ta gentillesse et ta disponibilité pour avoir corrigé mes innombrables divagations. Je suis impatient qu'on se descende quelques petits mojitos en terrasse afin de souffler un peu après ce long marathon (ou des whiskies, je ne suis pas sectaire). Olivier, tes capacités de projection dans les bois et de quasi-omniscience m'ont toujours sidéré. En les combinant à ton énergie intarissable et à tes qualités humaines, tu arrives à maintenir une équipe soudée où il fait bon vivre (et travailler un peu tout de même). Rien que pour ça, merci.

Bon, et maintenant on se concentre, il ne s'agit pas d'oublier de remercier quelqu'un parmi notre grande famille.

Tout d'abord, merci aux doyens du laboratoire. Raymond, le grand maitre de Gscope, toujours disponible, dont les nombreuses discussions m'ont beaucoup appris, et que j'ai finalement réussi à convertir à Ubuntu. Luc, toujours le mot pour rire et de bonne humeur, ça a été un plaisir de t'asticoter sur le TCL dès que l'occasion se présentait. Julie, fidèle au poste dès le lever du jour, souvent la première personne avec qui je débute ma petite tournée des popotes matinale que j'affectionne tout particulièrement. Et enfin Frédéric, l'incarnation zen de la certification ISO9001, et éminent collègue de thé et de restau U.

Viennent ensuite les habitants du couloir. Laetitia P, qui a toujours le sourire, surtout quand elle arrive pour me piquer des chewing-gums. Manu, s'il n'y avait qu'une chose à retenir, ce serait le congrès à Vienne où l'on a vraiment bien déliré.

On continue la visite avec Radwen, alias JR. Merci beaucoup à toi et à Rym de m'avoir convié en Tunisie pour votre mariage. Courage pour finir ta thèse en beauté (et en anglais, la classe…). Bon courage aussi à Émeline et Dao qui ont entamé la leur il y a peu de temps. Valentin, bonne chance pour la suite.

Il y a aussi Laetitia G, toujours prête à faire signer nos papiers aux plus hautes autorités, mais surtout une véritable touriste globetrotteuse qui vient nous faire déprimer avec sa peau toute bronzée, en plus de nous spammer à longueur de temps. Alors, pas trop de fautes dans ces remerciements ?

Arrive maintenant Sophie, la magicienne de la plate-forme, qui parvient on ne sait trop comment, à faire sortir des NullPointerException là où l'on s'y attend le moins. Laurent, le grand sage… du SAGE. Véronique, toujours la pêche, jour après jour. Stéphanie, où la fille qui murmurait aux oreilles du Solexa.

Bien entendu, il y a aussi les résidents du grand labo. Anne, qui nous quitte pour ses champignons. Ta bonne humeur et ta tasse qui traine au coin café vont me manquer. Yann, sans doute encore plus taquin que moi, toujours en quête du cristal ultime (et s'il diffracte c'est encore mieux). Hoan, notre data-manager haut-débit favori, merci de t'être souvent inquiété de l'avancement de ma thèse… et de mon après thèse. Gioia, drôle de physicienne qui fait de la bio-info, conversion réussie. Nicolas, féru de culture asiatique, souvent une anecdote marrante à raconter. Florence, un peu tête en l'air, ça m'a fait bizarre de passer cette thèse avec un ancien professeur de lycée, mais on s'y habitue. Wolfgang, qui tente tant bien que mal de mettre de l'ordre dans nos données, alors qu'on ne sait pas clairement ce que l'on cherche. Nicodème, un rire inimitable et irrésistiblement communicatif.

Et puis il ne faut certainement pas oublier Serge et Guillaume, nos ingénieurs système et réseau, qui m'ont, hélas pour eux, souvent vu pointer le bout du nez dès qu'il y avait un problème. Et Alain, McGyver en électronique, qui n'oublie jamais de passer faire un petit « Saaaalut » à notre bureau.

Et finalement, notre bureau, connu sous les noms de « bureau des djeunz » (qui ne le sont plus trop finalement) ou le « bureau des glands », c'est selon (n'est ce pas Olivier ?). Tout d'abord Guillaume, parti depuis déjà quelques mois, souvent dans le même trip qui consistait à sortir le plus d'imbécilités geekiennes à la minute, des tas de conversations stimulantes et quelquefois sérieuses. Que la force du zombie mutant Maya soit avec toi. David, ou plutôt Dave, parce que DAVID c'est de la m… Toujours prêt à aider les autres, et d'un point de vue extérieur, imperturbablement calme et posé (mais je suppose que le 2 de

tension, ça aide). Bon courage pour la fin, c'est la dernière ligne droite. Laurent-Philippe, l'intermittent du bureau. Vivement qu'on soit moins occupés pour reprendre nos parties de Magic… et pourquoi pas un petit séjour au Liban ? Et enfin Yannick-Noël, testeur et dealer officiel de séries US, promotologue à ses heures perdues (ou proctologue, j'ai comme un trou de mémoire), mais surtout expert en Web 2.0, d'un point de vue écrémage facebookien et adopteunmecquien.

Et puis il y a aussi tout les « anciens » que j'ai croisés pendant une durée plus ou moins longue. Merci à Adeline, Frédéric, Aurélie, Jean (ah ben tiens non, il revient, bon retour parmi nous), Annaïck, Ravi, Odile, et Francisco pour tout ce que vous avez pu m'apporter.

Que serait le labo sans sa ribambelle de stagiaires qui y ont séjourné, avec par ordre d'apparition : Fabrice, Xavier, Yahya, Laurent, Louise, Éveline, Némo, Léa, Bénédicte, Julien, Sophie, Benjamin, Jonathan, Tao, Ali, François, Seydou, et Enzo. Spéciale dédicace à Fabrice pour m'avoir fait découvrir X-wars, Yahya et son délicieux thé à la menthe, Némo pour ses délires jeux vidéotesques et pour m'avoir fait craquer pour une PS3, Ê-veuh, Louise et Sophie pour avoir fait remonter le niveau d'œstrogènes qui avait atteint le seuil critique dans notre bureau, et Enzo, que je n'ai pas beaucoup croisé, mais dont je ne désespère pas de pouvoir écraser un jour aux Magic. Et enfin bon courage à Benjamin et Jonathan qui vont débuter leur thèse d'ici cette année (je vous aurai pourtant prévenus !).

Je souhaite aussi remercier toute l'équipe de la plate-forme de Biologie et Génomique structurales, et plus particulièrement Didier Busso, Loubna, Matthieu, Pierre et Édouard avec qui j'ai passé deux semaines passionnantes à répliquer ces satanés clones d'Alvinella.

Je remercie aussi les équipes de la station biologique de Roscoff pour leur implication dans ce projet, et tout particulièrement Didier Jollivet, Arnaud Tanguy et Jean Mary.

Je remercie également toute l'équipe de coureurs de l'IGBMC qui m'ont accompagné, ou plutôt que j'ai tenté de suivre, pendant que je me vidais un peu l'esprit : Raymond, Alain, Dave, Nicolas, Isabelle, Nathalie, Bruno, Jean-Paul, Jean-Marie…

Je remercie tout particulièrement mes amis proches : Guillaume, Dave, Yannick, Laurent-P, Agathi, Benjamin Schwarz, Nathanaël Weill, Pierre « Pierretta » Hassenboehler, Claude Schenck et Christophe Huault. Il y aurait trop de choses à dire ici, mais merci beaucoup pour votre amitié et pour m'avoir soutenu à un moment ou un autre. J'espère pouvoir vous rendre un jour la pareille.

Merci à toute ma famille, surtout à ceux qui ont dû supporter mon sale caractère de ces derniers mois. Je pense bien sûr à ma Meuhman adorée, qui va fêter son demi-siècle pendant la semaine de ma soutenance. Le week end va être sacrément arrosé !! Sans oublier de remercier le couple infernal Mamie et Tonton boissons. Et aussi mon frangin, qui a eu le bon goût de jouer à la PS3 dans la même pièce que moi, pendant ma rédaction. Et bien croyez le ou non… ça motive à en finir au plus vite.

Pour terminer, je remercie Tricia pour m'avoir fait passer de tellement bons moments pendant nos soirées télé, et bien entendu le club havane, mon sponsor (dés)hydratant officiel.

Un grand merci à vous tous…

LISTE DES ABRÉVIATIONS

ADN	Acide désoxyribonucléique
ADNc	ADN complémentaire
AJAX	*Asynchronous JavaScript and XML*
ARN	Acide ribonucléique
ARNm	ARN messager
ARNr	ARN ribosomique
ARNt	ARN de transfert
BIPS	*BioInformatics Platform of Strasbourg*
BIRD	*Biological Integration and Retrieval Data*
BIRD-QL	*BIRD Query language*
BLAST	*Basic Local Alignment Search Tool*
BMRB	*Biological Magnetic Resonance Data Bank*
BNL	*Brookhaven National Laboratory*
CDS	*Coding sequence*
CGI	*Common Gateway Interface*
CRUD	*Create, Retrieve, Update, Delete*
CSS	*Cascading Style Sheets*
DAG	*Directed Acyclic Graph*
DAO	*Data Access Object*
DAS	*Distributed Annotation System*
DAS	*Distributed Annotation System*
DAVID	*Database for Annotation, Visualization and Integrated Discovery*
DDBJ	*DNA Data Bank of Japan*
ddNTP	di-désoxyribonucléotide triphosphate
dNTP	désoxyribonucléotide triphosphate
DPC	*Density of Points Clustering*
EBI	*European Bioinformatics Institute*
EC	*Enzyme Commission*
EGPM	Évolution et Génétique des Populations Marines
EMBL	*European Molecular Biology Laboratory*
EST	*Expressed Sequence Tag*
Gio	gibi octet (giga binaire)
GO	*Gene Ontology*
GSC	*Genomic Standards Consortium*
GSS	*Genome Survey Sequence*
HMM	*Hidden Markov Model*
HSP	*High-scoring Segment Pairs*
HTC	*High Throughput cDNA sequencing*
HTGS	*High Throughput Genomic Sequencing*
HTML	*Hypertext Markup Language*
HTTP	*Hypertext Transfer Protocol*
IGBMC	Institut de Génétique et de Biologie Moléculaire et Cellulaire
ILP	*Inductive logic programming*
IPA	*Ingenuity Pathways Analysis*
iSCSI	*Internet Small Computer System Interface*

JSON	*JavaScript Object Notation*
KDD	*Knowledge Discovery in Databases*
KEGG	*Kyoto Encyclopedia of Genes and Genomes*
KGML	*KEGG Markup Language*
LBGI	Laboratoire de Bio-informatique et Génomique Intégratives
LEON	*multiple aLignment Evaluation Of Neighbours*
LMS	*Local Maximum Segments*
MACS	*Multiple Alignment of Complete Sequences*
MACSIMS	*Multiple Alignment of Complete Sequences Information Management System*
MAO	*Multiple Alignment Ontology*
MGED	*Microarray Gene Exression Data Society*
MVC	Modèle-Vue-Contrôleur
NCBI	*National Center for Biotechnology Information*
NFS	*Network File System*
NorMD	*Normalized Mean Distance*
OBF	*Open Bioinformatics Foundation*
OBO	*Open Biomedical Ontologies*
OBO	*Open Biomedicals Ontologies*
ORF	*Open Reading Frame*
PAB	*Pyrococcus abyssi* box
pb	paires de bases
PDB	*Protein Data Bank*
PDBe	*PDB in Europe*
PDBj	*PDB of Japan*
PDO	*PHP Data Objects*
PEAR	*PHP Extension and Application Repository*
PHP	*PHP: Hypertext Preprocessor*
PIR	*Protein Information Resource*
poly(x)	région homopolymérique du nucléotide ‘x’
PPi	pyrophosphate inorganique
RAID	*Redundant Array of Independent Disks*
RASCAL	*RApid Scanning and Correction of ALignment errors*
RCSB	*Research Collaboratory for Structural Bioinformatics*
SAGE	*Serial Analysis of Gene Expression*
SCF	*Standard Chromatogram File*
SGBDR	Système de Gestion de Bases de Données Relationnelles
SIB	*Swiss Institute of Bioinformatics*
SOA	*Service Oriented Architecture*
SOAP	*Simple Object Access Protocol*
SQL	*Structured query language*
SRS	*Sequence Retrieval System*
SRS	*Sequence Retrieval System*
STRING	*Search Tool for the Retrieval of Interacting Genes/Proteins*
STS	*Sequence Tagged Site*
SVM	machine à vecteurs de support
Tcl	*Tool Command Language*
Tk	*ToolKit*

To	téra octet
TrEMBL	*Translated EMBL*
TSA	*Transcriptome Shotgun Assembly*
TSS	*Transcription start site*
UniMES	*UniProt Metagenomic and Environmental Sequences*
UniParc	*UniProt Archive*
UniProt	*Universal Protein resource*
UniProtKB	*UniProt Knowledgebase*
UniRef	*UniProt Reference clusters*
URI	*Uniform Resource Identifier*
URL	*Uniform Resource Locator*
UTR	Untranslated region
WGS	*Whole Genome Shotgun*
Wolcanno	*Web Viewer Of aLvinella cDNA ANNOtation*
WS	*Web Service*
WS	*Web Service*
WSDL	*Web Services Description Language*
wwPDB	*Worldwide PDB*
XGMML	*eXtensible Graph Markup and Modeling Language*
XML	*Extensible Markup Language*

TABLE DES MATIÈRES

TABLE DES FIGURES

TABLE DES TABLEAUX

AVANT-PROPOS

Les travaux présentés dans ce manuscrit de thèse ont été effectués au Laboratoire de Bioinformatique et Génomique Intégratives (LBGI) à l'Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), sous la direction d'Odile LECOMPTE et Olivier POCH, dans le cadre d'un projet d'annotation de banques d'ADNc de l'annélide polychète *Alvinella pompejana* (Desbruyères et Laubier, 1980).

A. pompejana est un ver de quelques centimètres vivant en colonie sur les fumeurs de sources hydrothermales. Il y est confronté à des conditions environnementales extrêmes et changeantes et est considéré comme l'animal le plus thermotolérant à ce jour (Cary *et al.*, 1998). Ce projet a été lancé par le Consortium *Alvinella* qui a été réuni à l'initiative du LBGI afin de remplir deux objectifs majeurs que sont la constitution d'une banque de référence de séquences annotées afin d'étudier à la fois les mécanismes d'adaptation aux conditions extrêmes mis en œuvre chez *Alvinella* et le phylum de Annélides, encore très peu représenté dans les banques de séquences, ainsi que de disposer d'une source de protéines et de complexes protéiques thermostables d'origine animale.

Ce manuscrit est divisé en trois parties :

La première partie présente l'état de l'art relatif aux différents thèmes abordés au cours de cette thèse. Après une brève mise en place du contexte de l'ère post-génomique (chapitre 1), le chapitre 2 définit les différentes techniques de séquençage jusqu'au séquençage moderne à haut-débit. Le chapitre 3 aborde le domaine de l'annotation de séquences et tente de dégager les problèmes majeurs liés à ce dernier. Enfin, le chapitre 4 décrit *A. pompejana* et son biotope, ainsi que les projets de séquençage ciblant cet animal.

La deuxième partie, Matériel et méthodes, décrit au travers des chapitres 5 à 9 l'environnement informatique dans lequel j'ai évolué au cours de ces dernières années, ainsi que les banques biologiques et les différents outils informatiques et bioinformatiques qui m'ont permis de mener à bien mes travaux.

La troisième et dernière partie consacrée aux résultats obtenus au cours de mes travaux, est elle-même divisée en quatre parties traitant chacune un des grands axes de ma thèse, ces axes étant intimement liés. Le premier axe décrit la conception d'un pipeline de traitement et d'analyse de données brutes de séquençage développé en premier lieu pour l'analyse des banques d'ADNc d'*Alvinella* (chapitre 10). Le deuxième axe présente le développement d'un deuxième pipeline, réalisant l'annotation intégrative automatique de séquences protéiques (chapitre 11). Un chapitre intermédiaire (chapitre 12) est consacré à la première publication réalisée dans le cadre du Consortium *Alvinella* afin d'illustrer l'utilisation des deux pipelines décrits précédemment, tout en faisant le lien avec le dernier axe qui englobe la visualisation et la recherche de données par l'intermédiaire d'interfaces Web (chapitre 13).

INTRODUCTION

1 CONTEXTE DE L'ÈRE POST-GÉNOMIQUE

Au cours de la deuxième moitié du 20ième siècle, la biologie a connu un essor sans précédant, grâce aux avancées technologiques et scientifiques dans les domaines de la biologie cellulaire et moléculaire, de la robotique, de l'imagerie et de la bioinformatique. En l'espace de 50 ans seulement, la biologie est passée de l'étude d'un seul gène à l'étude de génomes, de transcriptomes ou de protéomes d'organismes complets, avec une granularité d'analyse allant de la simple molécule isolée à l'organisme dans sa globalité. Ceci a été rendu possible par les avancées technologiques réalisées conjointement en biologie et en bioinformatique qui ont été motivées par la volonté des scientifiques de soutirer toujours davantage d'informations de notre génome.

1.1 La révolution biologique

La révolution biologique a été initiée par quatre événements majeurs : la découverte de l'ADN en tant que support de l'information génétique (Avery *et al.*, 1944)[1] et de sa structure en double hélice (WATSON et CRICK, 1953), ainsi que la mise en place du dogme central de la biologie moléculaire (CRICK, 1958) et le déchiffrage du code génétique (MATTHAEI *et al.*, 1962). Ces événements constituent les fondements de la génomique.

Le dogme central (Figure 1) est la modélisation simplifiée du flux de l'information génétique à travers différentes molécules de la cellule et se résume en trois processus :

- La réplication de l'information génétique portée par l'ADN pour donner vie à de nouvelles cellules,
- La transcription en ARN des gènes dispersés tout au long des molécules d'ADN,
- La traduction des ARN messagers (ARNm) en protéines qui vont participer à la structure et au fonctionnement de la cellule.

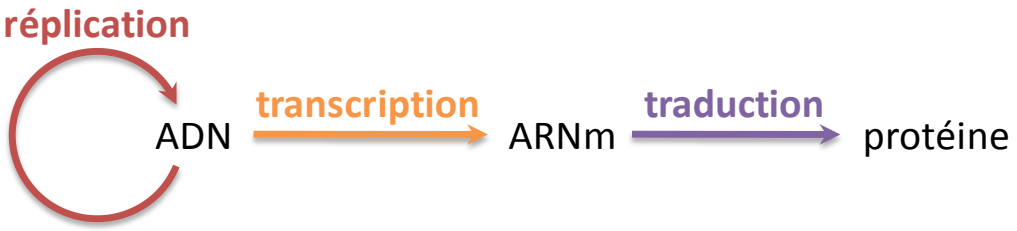


Figure 1 – Dogme central de la biologie moléculaire

[1] Les références sont au format CELL. Les références qui ne présentent pas d'année de publication correspondent à des sites Web.

Grâce à ces principes, de nouvelles techniques de biologie moléculaire ont pu être développées, et en synergie avec la montée en puissance et la disponibilité des ordinateurs, une nouvelle discipline a émergé : la bioinformatique. Le Tableau 1 regroupe les événements majeurs qui sont intervenus en biologie (incolore), en informatique (en vert) et en bioinformatique (en bleu) depuis 1944.

Tableau 1 – Chronologie non exhaustive des événements marquants survenus en biologie, en informatique et en bioinformatique
Adapté de http://bioinformatics.ws/index.php/History_of_Biology, http://bioinformatics.ws/index.php/History_of_Bioinformatics, http://en.wikipedia.org/wiki/Timeline_of_computing, http://www.thocp.net/timeline/timeline.htm, http://seqanswers.com/ et http://www.genomesonline.org

Année	***Auteurs***	***Événement***
1944	Avery	Démonstration de l'ADN en tant que support de l'information génétique (Avery *et al.*, 1944)
1946		ENIAC, premier ordinateur totalement électrique et Turing-complet
1947	Bell Labs	Invention du transistor
1951		UNIVAC, premier ordinateur commercialisé
1953	Sanger, Thompson	Détermination de la séquence des chaines A et B de l'insuline (SANGER et THOMPSON, 1953a, 1953b)
	Watson, Crick	Modèle de la structure en double hélice de l'ADN (WATSON *et al.*, 1953)
1956	IBM	Premiers disques durs commercialisés (5 Mo)
1958	Crick	Énonciation du dogme central de la biologie moléculaire (CRICK, 1958)
	Texas intruments	Réalisation du premier circuit intégré
1962	Matthaei	Déchiffrage du code génétique (MATTHAEI *et al.*, 1962)
1965	Dayhoff	Premier atlas de la séquence et de la structure des protéines (Dayhoff et National Biomedical Research Foundation., 1965)
1967	Fitch	Construction d'arbres phylogénétiques (Fitch et Margoliash, 1967)
1969	US DoD	ARPANET, premier réseau informatique permanent
	Bell Labs	Naissance du futur Unix
1970	Smith, Wilcox	Première enzyme de restriction spécifique isolée (Smith et Wilcox, 1970)
	Needleman, Wunsch	Algorithme d'alignement global optimal entre deux séquences (Needleman et Wunsch, 1970)
1971	Intel	Premier microprocesseur (Intel 4004)
1973		Annonce de la *Protein Data Bank* (PDB) (Protein Data Bank, 1973)
1974	Chou, Fasman	Algorithme de prédiction des structures secondaires des protéines (Chou et Fasman, 1974)
1977	Sanger	Méthode de séquençage par synthèse enzymatique de l'ADN (Sanger, Nicklen, *et al.*, 1977)

	Sanger, Air	Premier génome complet séquencé : phage ϕX174 (Sanger, Air, *et al.*, 1977)
	Maxam, Gilbert	Méthode de séquençage chimique de l'ADN (Maxam et Gilbert, 1977)
	Staden	Suite d'analyse de séquences d'ADN Staden
1980	EMBL	Première banque de séquences nucléiques
1981	Anderson	Séquence du génome mitochondrial humain (Anderson *et al.*, 1981)
	Smith, Waterman	Algorithme d'alignement local optimal entre deux séquences (Smith et Waterman, 1981)
	IBM	Commercialisation du premier PC (IBM PC 5150)
1982	GenBank	Banque Américaine de séquences nucléiques
		Naissance du réseau des réseaux : Internet
	Commodore	Commercialisation de mon 1er ordinateur : Commodore 64
1983	Mullis	Invention de la réaction en chaine de la polymérase (PCR) (Mullis, 1994)
1985	Lipman, Pearson	FASTA, programme de recherche de séquences par similarité (Lipman et Pearson, 1985)
	Gouy	ACNUC, programme d'interrogation des banques de séquences (Gouy *et al.*, 1985)
	Philips, Sony	Invention du CD-ROM (*Compact Disc Read Only Memory*) (650 Mo)
1986	Bairoch	SWISS-PROT, Banque de séquences protéiques
	DDBJ	Banque Japonaise de séquences nucléiques
1987	Applied Biosystems	Premier séquenceur automatique (ABI 370)
	Burke	Création du vecteur de clonage *Yeast Artificial Chromosome* (YAC) (Burke *et al.*, 1987)
	Kulesh	Apparition de la technologie des puces à ADN (Kulesh *et al.*, 1987)
1988		Lancement du projet international de séquençage du génome humain (National Research Council (U.S.)., 1988)
	Higgins, Sharp	CLUSTAL, programme d'alignement multiple (Higgins et Sharp, 1988)
1989	O'Connor	Création du vecteur de clonage *Bacterial Artificial Chromosome* (BAC) (O'Connor *et al.*, 1989)
	CERN	Invention du World Wide Web et du langage HTML
1990	Altschul	BLAST, programme de recherche de séquences par similarité (Altschul *et al.*, 1990)
1991	Adams	Création et utilisation à grande échelle du séquençage partiel d'ADNc (EST) (Adams *et al.*, 1991)
	Roberts	GRAIL, programme de localisation de gènes (Roberts, 1991)
1993	Etzold, Argos	SRS, programme d'interrogation des banques de séquences (Etzold et Argos, 1993)
	NCSA	Premier navigateur Web : Mosaic
1994	Thompson	CLUSTALW (Thompson *et al.*, 1994)
1995	Fleischmann	Premier organisme vivant séquencé : *Haemophilus*

		influenzae (Fleischmann *et al.*, 1995) (1,8 Mb)
	Fraser	Plus petit organisme vivant séquencé : *Mycoplasma genitalium* (Fraser *et al.*, 1995) (580 kb)
	Jong, Brenner	Création de la librairie biologique *open source* BioPerl
	DVD Forum	Spécifications du DVD (*Digital Versatile Disc*) (8,5 Go)
1996	Walsh, Barrell	Premier organisme eucaryote séquencé : *Saccharomyces cerevisiae* (Walsh et Barrell, 1996) (12,1 Mp)
	Affymetrix	Commercialisation de la première puce à ADN
1997	Blattner	Génome complet de *Escherichia coli* (Blattner *et al.*, 1997) (4,7 Mb)
	Altschul	Gapped BLAST et PSI-BLAST (Altschul *et al.*, 1997)
	Burge, Karlin	GenScan, prédiction de la structure complète des gènes dans l'ADN génomique humain (Burge et Karlin, 1997)
1998		Premier organisme pluricellulaire séquencé : *Caenorhabditis elegans* (CeSC, 1998) (97 Mp)
2000	Dennis, Surridge	Génome d'*Arabidopsis thaliana* (Dennis et Surridge, 2000) (100 Mp)
	Adams	Génome de *Drosophila melanogaster* (Adams *et al.*, 2000) (180 Mb)
	AMD	Premier microprocesseur x86 atteignant 1 GHz
2001	Lander	Publication préliminaire du génome humain par le *Human Genome Project* (Lander *et al.*, 2001) (2,9 Gb)
	Venter	Publication préliminaire du génome humain par *Celera Genomics* (Venter *et al.*, 2001) (2,9 Gb)
	Ensembl	Navigateur de génome Ensembl (Hubbard *et al.*, 2009)
	NCBI	Navigateur de génome du NCBI (http://www.ncbi.nlm.nih.gov/mapview/)
DÉBUT DE L'ÈRE POST-GÉNOMIQUE		
2002	Waterson	Séquence préliminaire du génome de la souris (Waterston *et al.*, 2002) (2,5 Gb)
	UCSC	Navigateur de génome de l'UCSC (Kuhn *et al.*, 2009)
2004	IHGSC	Finalisation du génome humain (IHGSC, 2004) (3,2 Gp)
	ENCODE PC	ENCODE, projet d'identification de tous les éléments fonctionnels du génome humain (ENCODE Project Consortium, 2004)
	Blu-ray Disc Association	Spécifications du Blu-ray Disc (50 Go)
2005	Roche, 454	Séquenceur automatique haut-débit de 2ème génération par pyroséquençage : GS20
	AMD	Premier processeur x86 double cœurs
2006	Folding@Home	Utilisation de la puissance des processeurs de flux des GPGPU (nouvelles cartes graphiques) des PC et du processeur Cell BE des PlayStation3 pour accélérer jusqu'à 30x les calculs de dynamique moléculaire
2007	Illumina, Solexa	Séquenceur automatique haut-débit de 2ème génération par synthèse microfluidique : Genome Analyzer

	Applied Biosystems	Séquençage automatique haut-débit de 2ème génération par ligation : système SOLiD
	Hitachi	Premier disque dur atteignant 1 To de capacité
	HVD Forum	Spécifications du HVD (Holographic Versatile Disc) (200 Go)
2008	Helicos	Séquenceur automatique de 2ème génération par synthèse sans pré-amplification
2009…		Près de 1 000 projets de séquençage de génomes eucaryotes en cours…

La bioinformatique, dont Margaret O. Dayhoff figure parmi les pionniers grâce à son atlas de la séquence et de la structure des protéines (Dayhoff *et al.*, 1965), va rapidement s'avérer indispensable, notamment dans la gestion des données, aboutissant naturellement aux premières banques de données biologiques dont la PDB (Berman *et al.*, 2000).

C'est à partir de 1977, avec l'apparition de deux nouvelles approches de séquençage de l'ADN (Maxam *et al.*, 1977; Sanger, Nicklen, *et al.*, 1977) que la bioinformatique prend réellement son envol. La production de séquences par ces méthodes est l'occasion de créer les nouvelles banques de données EMBL (Cochrane *et al.*, 2009a) et GenBank (Benson *et al.*, 2009) afin de répertorier ces séquences nucléiques, et de développer de nouveaux algorithmes permettant de traiter les données biologiques. Ces derniers ont abouti aux outils majeurs de la bioinformatique que sont FASTA (Pearson et Lipman, 1988), CLUSTALW (Thompson *et al.*, 1994), et BLAST (Altschul *et al.*, 1997).

Avec l'apparition des séquenceurs automatiques et de nouveaux outils de biologie moléculaire (Mullis, 1994; Burke *et al.*, 1987; O'Connor *et al.*, 1989; Adams *et al.*, 1991), la production des séquences s'accélère et l'on voit apparaître des projets de séquençage de génomes complets qui aboutissent vers la fin du siècle.

Après dix années d'efforts, l'arrivée des premières séquences préliminaires du génome Humain (Lander *et al.*, 2001; Venter *et al.*, 2001) marque la fin de l'ère génomique et l'entrée dans l'ère post-génomique. Cependant, il a fallu attendre jusqu'en 2004 pour obtenir de la part de l'International Human Genome Sequencing Consortium une version que l'on peut considérer comme finalisée (IHGSC, 2004).

Actuellement, grâce aux techniques de séquençage haut-débit, les projets de séquençage se sont multipliés de sorte que la communauté scientifique a accès à 1059 génomes complets et publiés (Figure 2).

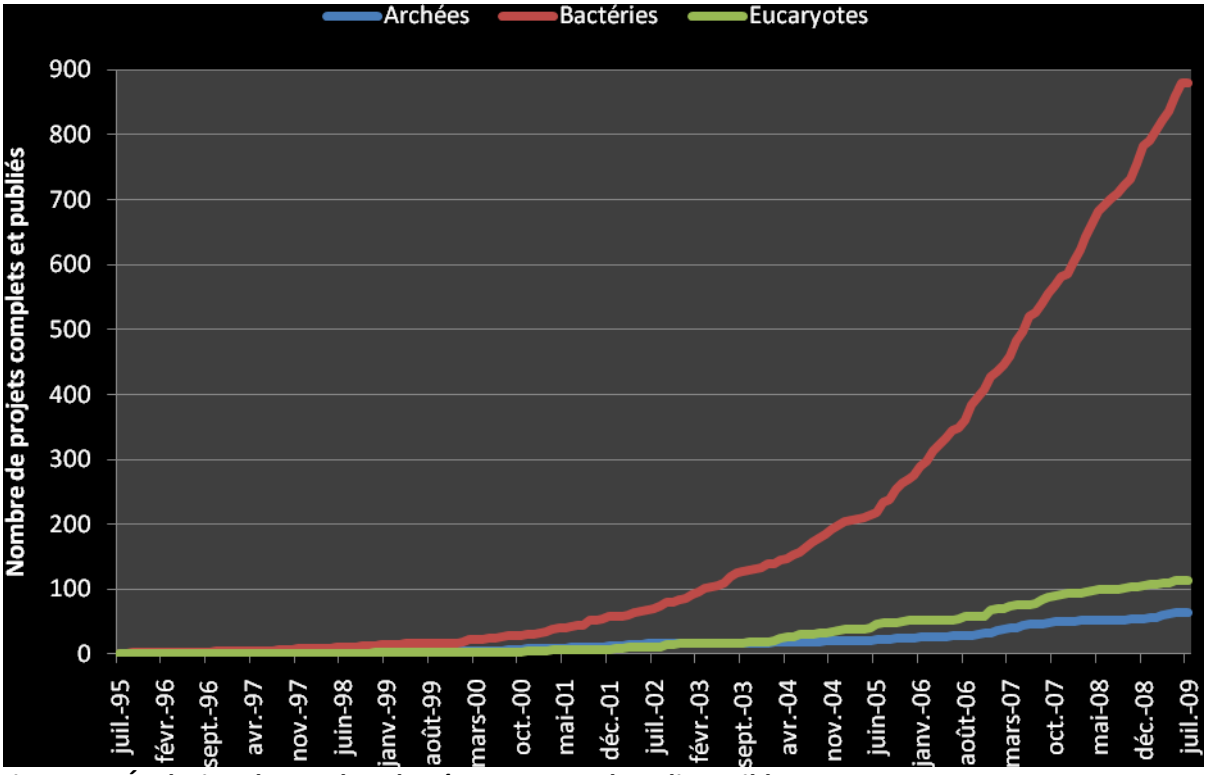


Figure 2 – Évolution du nombre de génomes complets disponibles
En juillet 2009, sont complets et disponibles 65 génomes d'Archées, 880 génomes de Bactéries et 114 génomes d'Eucaryotes. 978 génomes eucaryotes sont en cours de séquençage. Source GOLD (http://www.genomesonline.org).

1.2 Généralisation des « –omiques »

Bien évidemment, l'arrivée du génome humain et des autres génomes d'organismes modèles n'est que le début d'un monde nouveau, celui du haut-débit. Pendant que les techniques de séquençage et la puissance des logiciels bioinformatiques permettent de générer et de traiter un nombre toujours plus grand de données biologiques, l'utilisation des suffixes « –ome » et « –omique » se démocratise.

Le terme « génomique » a été introduit dès 1986 par Thomas H. Roderick (Kuska, 1998) et désigne la science dédiée à l'étude du génome, c'est-à-dire l'intégralité du matériel génétique d'un organisme. Par analogie, de nouvelles « –omiques » ont émergé, parmi lesquelles on retrouve la transcriptomique, la protéomique, l'interactomique, la métabolomique, et la fluxomique qui s'intéressent respectivement à l'étude de l'intégralité des ARNm, des protéines, des interactions protéine-protéine, des métabolites, et des flux de métabolites. Il existe encore plusieurs dizaine d'« –omiques » répertoriées (http://www.omics.org) et, cette nomenclature étant extensible à l'infini, de nouvelles « –omiques » sont créées continuellement.

Les « –omiques » ne sont qu'une classification très générale d'études expérimentales ou bioinformatiques haut-débit portant sur un même sujet et peuvent être subdivisées en axes

de recherche plus spécifiques. En effet, le terme génomique regroupe la génomique structurale (cartographie de la structure du génome ou détermination de structures 3D à l'échelle du génome), la génomique fonctionnelle (étude de la fonction des gènes et de leurs produits par des techniques *in silico*, des expériences de transcriptomique et de protéomique) et la génomique comparative (comparaison de l'organisation des gènes entre plusieurs organismes ou études phylogénétiques). De même, la transcriptomique englobe les expériences de puces à ADN, de SAGE (*Serial Analysis of Gene Expression*) (Velculescu *et al.*, 1995), le *differential display* (Liang et Pardee, 1992), ou le séquençage d'EST (Adams *et al.*, 1991) ; et la protéomique englobe les études par électrophorèse bidimensionnelle ou par spectrométrie de masse…

Alors que la génomique structurale (d'un point de vue cartographique) consiste en une vue plus ou moins figée dans le temps, les « –omiques » citées précédemment décrivent des entités dynamiques dépendantes de plusieurs paramètres temporels et expérimentaux. En effet, les données collectées vont varier en fonction de l'état physiologique de la cellule, du tissu, ou de l'organisme considéré. Cette apparente complexité supplémentaire est en fait un avantage puisqu'en jouant sur ces paramètres, il devient possible d'analyser le comportement de l'entité observée, ce qui permet de mieux comprendre le fonctionnement d'une de ses sous-parties. Cette sous-partie peut être par exemple un ou plusieurs complexes moléculaires dans le cas de l'interactomique, ou une voie métabolique dans le cas de la métabolomique.

Ce concept est poussé à son paroxysme dans le cadre de la biologie des systèmes, science qui ambitionne de modéliser un organisme complet, voire un ensemble d'organismes interagissant entre eux, en intégrant les données de plusieurs « –omiques » pour obtenir un immense réseau d'interactions intervenant entre les différents composants du système étudié (métabolites, macromolécules, complexes moléculaires, voies métaboliques, organites, cellules, tissus…). C'est par l'étude de cette vue systémique qu'il devient possible de dégager certaines propriétés du système modélisé qui ne seraient pas apparues en étudiant chaque composant isolément.

Ainsi, l'explosion des données a profondément modifié notre vision de la biologie en conduisant au recours massif à la bioinformatique et à la généralisation des approches intégratives. Dans le prochain chapitre, nous allons décrire l'acquisition de ces données en nous concentrant sur les données issues du séquençage de l'ADN.

2 SÉQUENÇAGE D'ADN

L'acquisition d'une grande quantité de séquences a permis de réaliser de nombreuses avancées dans différents domaines de la science. Bien que le séquençage de protéines permette d'accéder à des données plus informatives encore, telles que les modifications post-traductionnelles des résidus, il est beaucoup plus facile de réaliser un séquençage d'ADN. En effet, les séquençages chimiques ou enzymatiques permettent rarement d'obtenir des séquences de plus d'une centaine de résidus, la résolution des séquences en spectrométrie de masse en tandem et en résonance magnétique nucléaire se complexifie drastiquement en fonction de la longueur de la protéine, et l'obtention de cristaux diffractant à une bonne résolution en cristallographie par rayons X peut s'avérer très difficile selon la protéine.

C'est parce que le séquençage d'ADN pose moins de problèmes que les techniques ont beaucoup évolué, au point que, lorsqu'on parle de séquençage, on pense souvent au séquençage de génome. Ce chapitre va décrire l'historique de cette évolution, du séquençage manuel jusqu'aux nouveaux séquenceurs haut-débit.

2.1 Constitution de banques d'ADN

Tout projet de séquençage commence par la constitution d'une ou plusieurs banques d'ADN, bien que ça ne soit plus nécessairement vrai dans le cadre des nouvelles technologies de séquençage haut-débit (Cf. Séquençage haut-débit, page 20). Cette étape est imposée par les limites technologiques des techniques de séquençage qui ne peuvent traiter que des fragments d'une certaine taille.

Une banque d'ADN est une collection de fragments d'ADN à séquencer qui ont été intégrés au génome de cellules hôtes (généralement des microorganismes) à des fins de stockage et de réplication. L'intégration est réalisée par l'intermédiaire d'une molécule d'ADN, appelée vecteur de clonage, à l'intérieur de laquelle a été placé un fragment de l'ADN que l'on veut séquencer (appelé dans le cas présent '*insert*').

Il existe deux types de banques d'ADN : les banques d'ADN génomique dont les inserts sont issus de la fragmentation du matériel génétique initial à séquencer, et les banques d'ADNc (ADN complémentaire) dont les inserts sont des ARNm qui ont été « recopiés » en ADN sous l'effet d'une enzyme de rétrovirus, la transcriptase inverse. Les banques génomiques sont utilisées dans le cadre de séquençage de génomes, alors que les banques d'ADNc sont utilisées dans des études d'expression de gènes.

2.1.1 Vecteurs de clonage

Les vecteurs de clonage sont des constructions synthétiques qui doivent remplir plusieurs conditions :

- Ils doivent être doués de réplication autonome à l'intérieur de la cellule hôte afin d'être conservés au fil de nombreuses générations de cellules hôtes, appelées clones.
- La réplication doit être effectuée le plus fidèlement possible, c'est-à-dire ne pas introduire de mutations, afin de conserver l'insert de façon stable dans le temps.
- Ils doivent comporter des sites uniques de restriction ou de recombinaison utilisables afin d'y insérer le fragment d'ADN à cloner.
- Ils doivent aussi comporter un ou plusieurs marqueurs permettant de les identifier et de les purifier aisément.

Actuellement, il existe plus de 5 000 vecteurs différents, chacun comportant ses propres spécificités. Ce chiffre est basé sur les données publiques de la banque GenBank (Cf. GenBank, page 67). Il en existe donc certainement encore beaucoup plus, dont les séquences sont protégées par des brevets détenus par les compagnies commercialisant des kits de clonage.

Le choix de vecteurs est donc très vaste et va dépendre de la longueur des fragments que l'on veut cloner, du type d'expérience que l'on veut mener (séquençage, expression protéique…) et donc, du type de cellule hôte.

Dans le cadre du séquençage, les cellules hôtes proviennent généralement d'une souche d'*Escherichia coli* ou de *Saccharomyces cerevisiae*. Les vecteurs sont quant à eux d'origine plasmidique ou de type BAC (O'Connor *et al.*, 1989) dans le cas de cellules d'*E.coli*, et de type YAC (Burke *et al.*, 1987) dans le cas de cellules de levure.

Habituellement, les ARNm étant de petite taille, les banques d'ADNc sont construites à l'aide de plasmides qui peuvent contenir des inserts de 0,1 à 10 kpb. Selon la taille désirée pour les fragments, les banques génomiques sont construites à l'aide de plasmides, de BAC (30 à 350 kpb) ou de YAC (100 à 3 000 kpb), tout en sachant que pour un séquençage complet, il faut disposer de plusieurs banques construites à partir des fragments de taille moyenne différentes, afin de cartographier le génome plus facilement.

2.1.2 Étapes de construction

Mis à part une première étape différente entre la construction d'une banque génomique et celle d'une banque d'ADNc, les étapes suivantes sont communes aux deux.

La première étape de la construction d'une banque génomique consiste à fractionner l'ADN génomique (Figure 3). Ce fractionnement peut être réalisé par une digestion partielle par

une endonucléase ou une enzyme de restriction, mais les méthodes physiques sont préférées car elles sont plus reproductibles et les fragmentations sont plus aléatoires. Ces méthodes physiques peuvent mettre en jeu les ultrasons (sonication) (FREIFELDER et DAVISON, 1962), la nébulisation sous haute pression, ou la force de cisaillement (*DNA shearing*) (LEVINTHAL et DAVISON, 1961).

Dans le cadre de la construction d'une banque d'ADNc, aucune fragmentation n'est nécessaire. Au contraire, une attention toute particulière est apportée pour préserver les molécules d'ARN qui sont plus fragiles que l'ADN, puisqu'elles ne sont constituées que d'un seul brin. Lors de cette première étape, l'ARNm est rétro-transcrit en ADN sous l'action de la transcriptase inverse, une enzyme de rétrovirus.

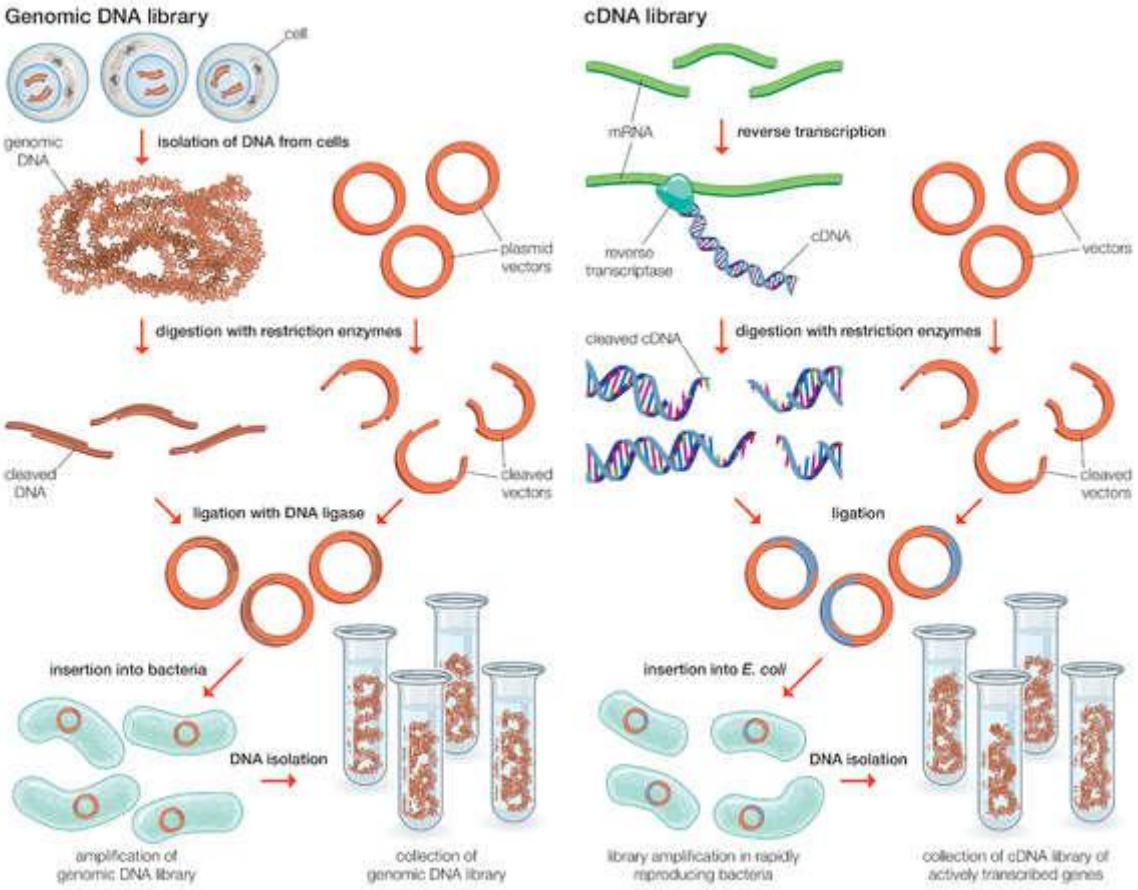


Figure 3 – Schémas simplifiés de préparation de banques d'ADN génomique et d'ADNc
Source Encyclopedia Britannica (http://www.britannica.com/)

Les étapes suivantes consistent en une électrophorèse sur gel d'agarose des fragments d'ADN ou d'ADNc afin de sélectionner et d'extraire du gel les fragments de taille désirée, puis de les intégrer au vecteur de clonage choisi. Les cellules hôtes sont ensuite transformées par l'insertion d'un vecteur à leur matériel génétique. Finalement, les cellules ayant été transformées sont cultivées et isolées en colonies bien distinctes. Les cellules ayant intégré un vecteur sont sélectionnées à l'aide d'un des marqueurs du vecteur,

généralement un gène de résistance à un antibiotique présent dans le milieu de culture, et qui empêche la multiplication des cellules non transformées. Chaque colonie est ensuite repiquée, conservée et étiquetée par un identifiant unique en vue de son séquençage.

2.2 Le séquençage classique

Il existe deux méthodes de séquençage dites « classiques », datant toutes deux de 1977 : une méthode chimique (Maxam *et al.*, 1977) et une méthode par synthèse enzymatique (Sanger, Nicklen, *et al.*, 1977).

2.2.1 Méthode chimique

La première étape de cette méthode consiste à marquer une des extrémités 5' ou 3' du fragment d'ADN à l'aide de ^{32}P. L'ADN est ensuite dénaturé et purifié pour ne conserver que le brin positif qui va être séquencé (Figure 4).

La solution obtenue contenant le brin positif marqué est répartie dans 4 milieux réactionnels dans lesquels des réactions chimiques vont spécifiquement couper le brin respectivement au niveau des bases G, (A et G), (C et T), et C. Puisque les conditions de coupure sont ajustées pour n'obtenir qu'une seule coupure par brin marqué, on obtient statistiquement des produits de réaction représentant une coupure à chacune des bases du fragment initial à séquencer.

Les produits de chaque milieu réactionnel sont ensuite séparés par électrophorèse sur un gel qui sera autoradiographié. Les plus petits produits de réactions ayant migré le plus rapidement, la séquence nucléotidique se lit de bas en haut du gel.

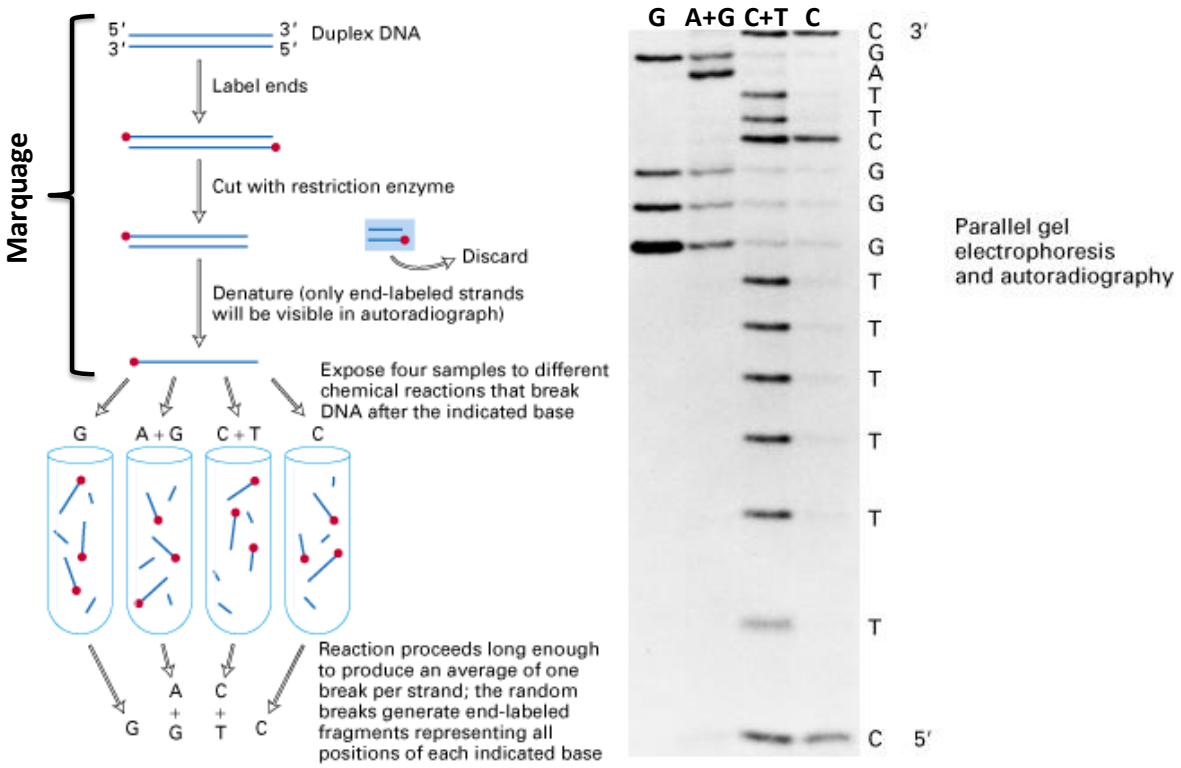


Figure 4 – Schéma récapitulatif de la méthode de séquençage de Maxam et Gilbert
Le résultat d'électrophorèse apparait à droite, avec l'interprétation de la séquence nucléotidique en regard de chaque bande. Pour la lecture des pistes où la réaction coupe deux types de résidus, la lecture est indirecte puisqu'il faut comparer les bandes de ces pistes avec celles des pistes où la réaction ne coupe qu'un type de résidu. Adapté de
http://departments.oxy.edu/biology/Stillman/bi221/092200/lecture_notes.htm

Les produits chimiques utilisés dans les milieux réactionnels lors des coupures spécifiques étant excessivement dangereux pour la santé, cette méthode a été abandonnée au profit de la méthode par synthèse enzymatique.

2.2.2 Méthode par synthèse enzymatique

Cette méthode, encore appelée méthode Sanger en raison de son inventeur, est basée sur l'activité de l'ADN polymérase ADN dépendante qui permet de polymériser un brin d'ADN complémentaire à un brin matrice, à partir d'un oligonucléotide, appelé amorce (ou *primer*) dans ce cas précis. Cette capacité est utilisée pour synthétiser un brin complémentaire, mais de façon incomplète, en arrêtant aléatoirement la réaction de manière à obtenir statistiquement des produits issus de réaction interrompue à chacune des bases du fragment à séquencer.

Le mix réactionnel est constitué du vecteur de clonage contenant le fragment à cloner, de la polymérase, des amorces et des dNTP (désoxyribonucléotides triphosphates dATP, dCTP, dGTP et dTTP). Pour arrêter aléatoirement la réaction, une faible concentration de ddNTP (di-désoxyribonucléotides triphosphates) est ajoutée. Ces ddNTP ne comportant pas de

groupement 3’-OH, ils agissent comme des terminateurs de la réaction de polymérisation en empêchant l’accomplissement d’une liaison 5'-3' phosphodiester ultérieure.

A l’instar de la méthode chimique, un marquage des produits de réaction est nécessaire pour pouvoir détecter ces derniers après leur séparation par électrophorèse sur gel. Pour ceci, il existe deux chimies : *dye primer* et *dye terminator*.

2.2.2.1 Chimie dye primer

Avec cette chimie, le marquage est effectué au niveau des amorces en 5’. Les marqueurs utilisés sont des fluorochromes dont la couleur émise est différente selon le mix réactionnel. Il est nécessaire de réaliser un total de 4 mix, chacun comportant un ddNTP différent (Figure 5). Chacun de ces mix va ainsi permettre de séquencer séparément un type de nucléotide.

À la fin des réactions, puisque le fluorochrome de chaque mix est différent, il est possible de mélanger les produits de réactions avant de réaliser l’électrophorèse sur gel. La révélation des bandes se fait par excitation lumineuse, et la lecture des pistes est directe.

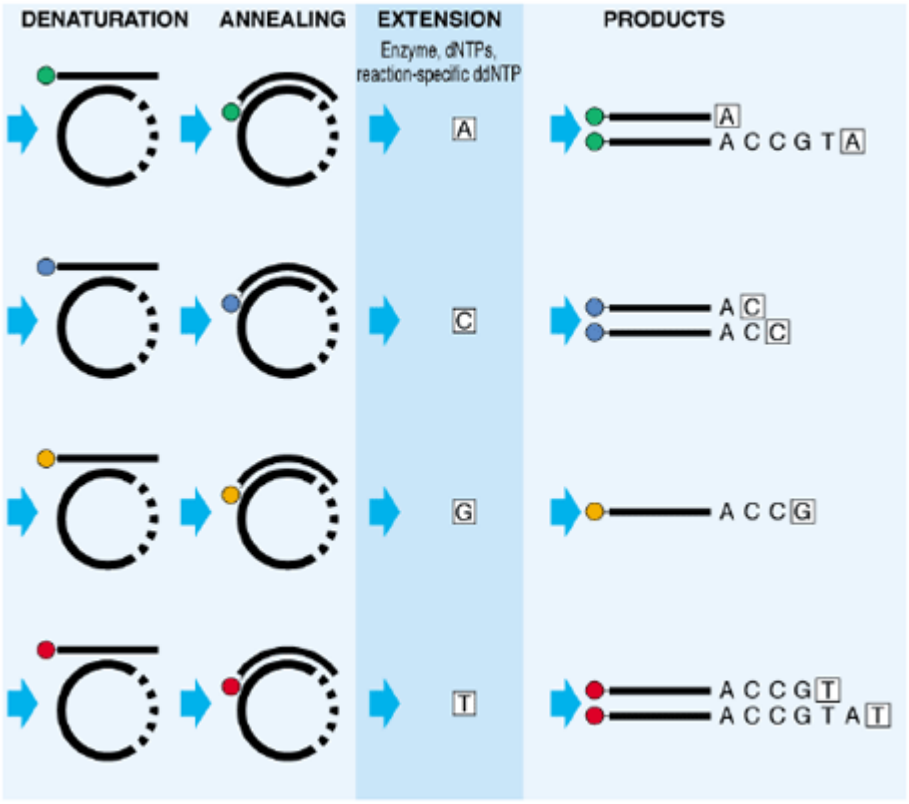


Figure 5 – Séquençage par synthèse enzymatique avec chimie *dye primer*
Chaque ligne correspond aux réactions d’un mix différent. La séquence obtenue est ACCGTAT. Source http://www.appliedbiosystems.com/

2.2.2.2 Chimie dye terminator

Cette chimie est une amélioration de la précédente. Ici ce ne sont pas les amorces qui sont marquées, mais chacun des ddNTP, toujours avec un fluorochrome différent. L'avantage évident de cette chimie est qu'un seul mix contenant tous les ddNTP est suffisant (Figure 6).

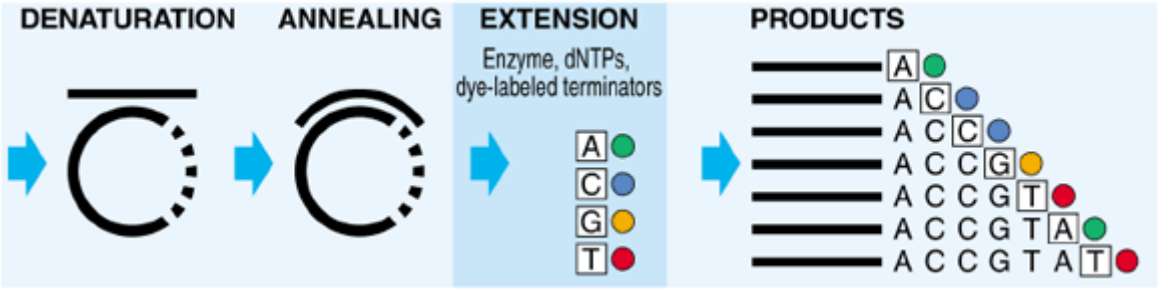


Figure 6 – Séquençage par synthèse enzymatique avec chimie dye terminator
Source http://www.appliedbiosystems.com/

2.2.3 Automatisation

La lecture manuelle de gels de séquençage est fastidieuse et sujette à de multiples erreurs d'interprétation, surtout après avoir lu les premiers milliers de bases ! C'est pourquoi cette tâche a été automatisée à l'aide des séquenceurs.

L'étape de séquençage par synthèse enzymatique est toujours nécessaire, mais cette fois cette tâche est déléguée à des robots qui la réalisent dans des microplaques de 96 ou 384 puits, correspondant à autant de clones séquencés. Ces plaques sont ensuite transférées dans un séquenceur, qui va réaliser l'électrophorèse tout en enregistrant sur ordinateur les profils d'intensités lumineuses des fluorochromes. Ces profils sont appelés chromatogrammes (ou *traces*).

Il existe deux types de séquenceurs automatiques de technologie Sanger : les séquenceurs à gel plat et les séquenceurs à capillaires. Les séquenceurs à gel plat disposent d'un gel enserré entre deux plaques de verre en haut duquel sont disposés des puits (192 sur un modèle LI-COR 4300) où sont déposées automatiquement les solutions de produits réactionnels (Figure 7a). Les séquenceurs à électrophorèse capillaire disposent quant à eux d'une série de capillaires contenant le gel de séparation (jusqu'à 96 sur un modèle ABI 3730xl) plongeant directement dans les cupules de la microplaque contenant les produits de réaction (Figure 7b). La lecture se fait, dans les deux cas, en fin de course du gel ou des capillaires par des détecteurs qui excitent les fluorochromes à l'aide de diodes laser.

La capacité journalière de traitement des séquenceurs de type Sanger varie de 1 à 2 Mpb pour des séquences d'une longueur moyenne de 1200 pb.

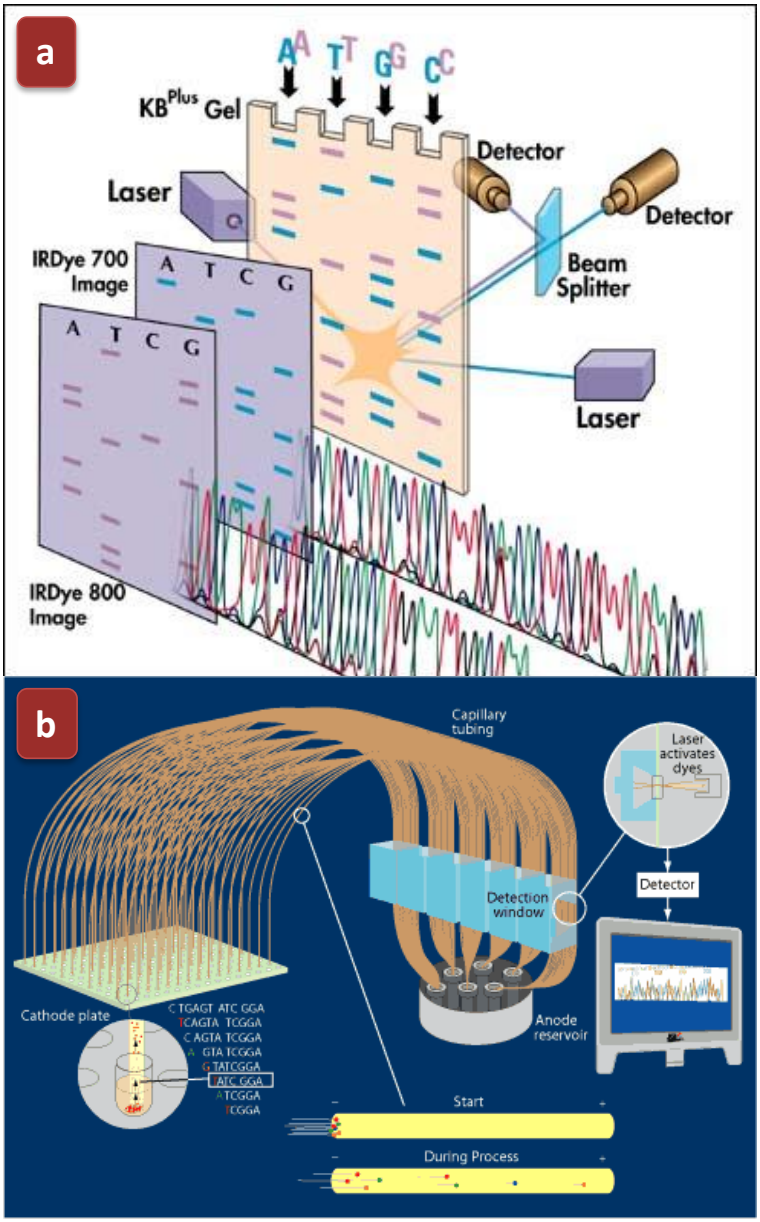


Figure 7 – Les deux technologies de séquenceurs automatiques
(a) Schéma de fonctionnement d'un séquenceur LI-COR 4300 à gel plat (http://www.licor.com/). Ici, il y a deux marqueurs différents, un par sens de séquençage (5' et 3'). (b) Schéma de fonctionnement d'un séquenceur à capillaires (source http://www.jgi.doe.gov/).

2.3 Séquençage haut-débit

La technique de séquençage Sanger a atteint ses limites en termes de rendement. Il n'est plus possible d'accélérer le séquençage qu'en multipliant le nombre de gels, de capillaires ou de machines, et la vitesse d'électrophorèse peut difficilement être accélérée sans amoindrir la qualité de séquençage. C'est pourquoi de nouvelles technologies ont vu le jour pour aboutir à des séquenceurs temps réel (« séquençage base par base ») massivement parallélisés, utilisant des techniques très diverses. Nous allons présenter les quatre

technologies disponibles actuellement, tout en sachant qu'une cinquième, dénommée séquençage par nanopores est en cours de développement (Rusk, 2009).

Le séquenceur 454 de Roche repose sur la technologie de pyroséquençage (Ronaghi *et al.*, 1998). Lors de l'élongation d'un brin par la polymérase, l'incorporation d'un dNTP relâche une molécule de pyrophosphate inorganique (PPi). C'est en mesurant la quantité de PPi relâchée dans le milieu lors de l'ajout d'un type de dNTP que s'effectue le pyroséquençage. La mesure est réalisée en ajoutant de l'ATP sulfurylase, de la luciférase, et leur substrat (adénosine 5´ phosphosulfate et luciférine) dans le milieu. Lorsqu'un PPi est relâché, l'ATP sulfurylase le convertit en ATP en présence d'adénosine 5´ phosphosulfate. Cet ATP est ensuite utilisé par la luciférase qui convertit la luciférine en oxyluciférine tout en émettant une lumière proportionnelle à la quantité d'ATP utilisée, et donc du nombre de dNTP intégrés.

Le séquenceur 454 nécessite la préparation d'une librairie de séquençage spécialisée dont voici les grandes étapes :

- Des séquences adaptateurs sont fixées à chaque extrémité des fragments d'ADN à séquencer. Un des adaptateurs contient un marqueur biotine en 5' qui sert à fixer les fragments sur des billes grâce à un système streptavidine/biotine. Les fragments d'ADN sont dénaturés afin de libérer les brins non-biotinés, pour obtenir une librairie d'ADN simple brin qui va servir de matrice au séquençage (sstDNA, *single stranded template DNA*) (Figure 8a).
- Les molécules de sstDNA sont ensuite amplifiées par emPCR (PCR en émulsion) (Figure 8b). Les sstDNA sont fixées sur des billes de manière à fixer une seule molécule de sstDNA par bille. Les billes sont ensuite émulsionnées dans une solution eau/huile dans des proportions favorisant l'inclusion d'un seul couple bille-sstDNA par gouttelette d'eau, qui va agir comme un microréacteur pour la réaction de PCR. Après plusieurs cycles de PCR, chaque gouttelette contient une bille où sont fixés tous les clones amplifiés. Un ensemble de clones issus d'une même matrice est souvent dénommé « polonie » en référence à l'amplification de toute une colonie de clones par la polymérase.

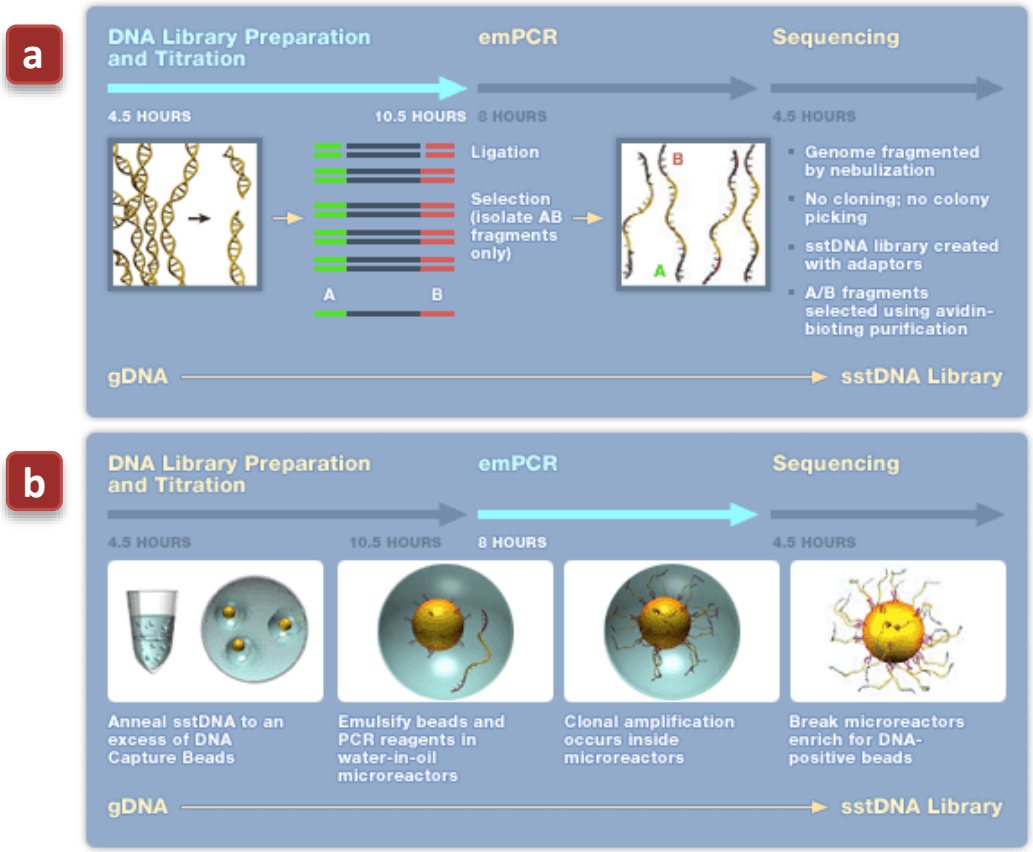


Figure 8 – Préparation des billes où sont fixées les sstDNA pour le séquenceur 454
Source http://www.454.com/

Afin de paralléliser massivement les réactions de pyroséquençage, le séquenceur 454 GS20 utilise une puce constituée de 200 000 puits d'environ 44 µm de diamètre (Figure 9). Les billes couvertes de sstDNA sont mélangées au mix de réaction contenant la polymérase et sont étalées sur la puce. Les puits de la puce sont d'une forme telle qu'une seule bille peut être déposée dans chaque puits. Les puits et les billes de sstDNA sont ensuite recouverts de billes d'enzymes immobilisées contenant la sulfurylase et la luciférase (Figure 9a). Le pyroséquençage peut alors commencer (Figure 9b). Chacune des bases A, C, G, et T est ajoutée séquentiellement sur la puce. L'image des intensités de lumière émise par la réaction intervenue dans chaque puits après l'ajout de chaque base est enregistrée. Avant l'ajout de la base suivante, les bases non intégrées par la polymérase sont dégradées par une solution d'apyrase. Ce cycle d'ajout des quatre bases est réitéré 42 fois pour obtenir des séquences d'une longueur moyenne de 100 pb pour un total de 20 millions de bases séquencées en 4h30.

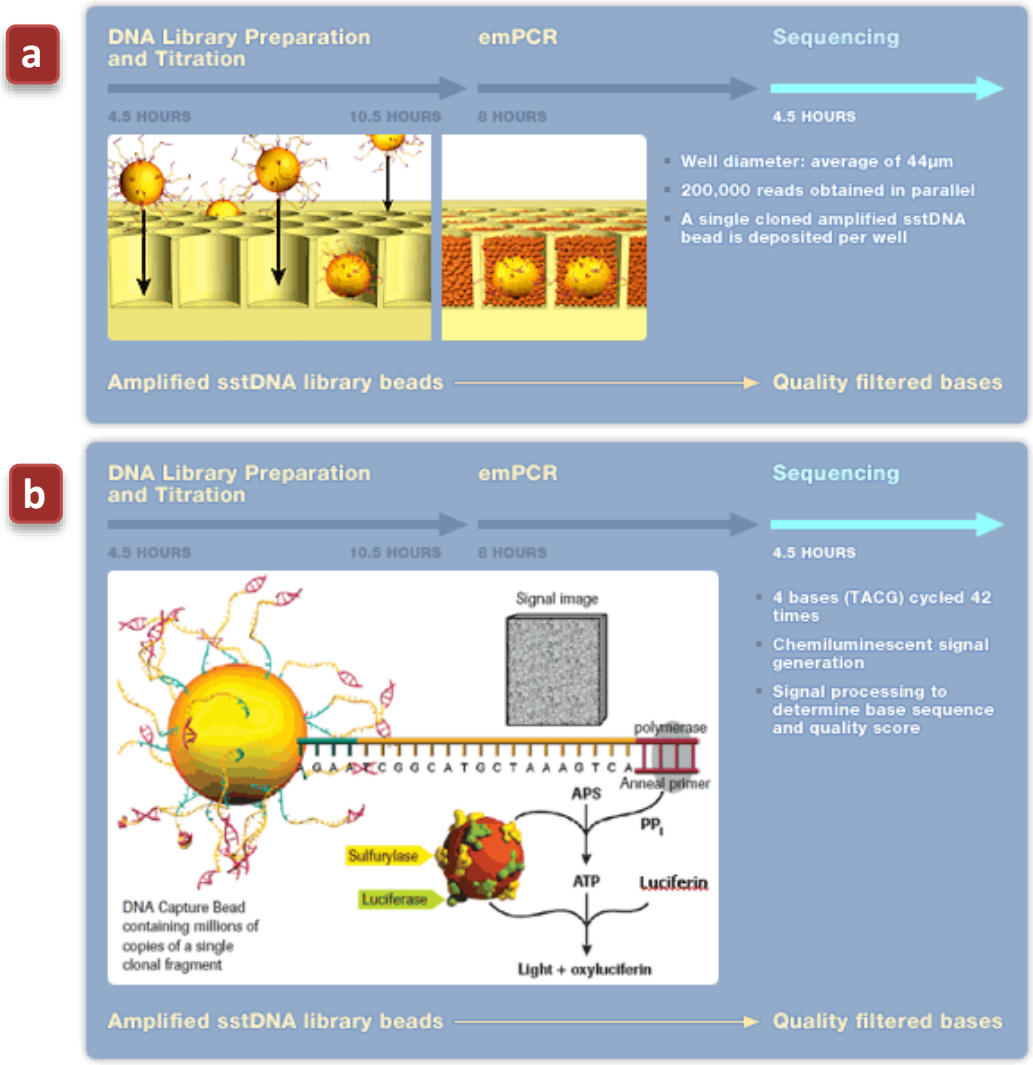


Figure 9 – Pyroséquençage sur puce du séquenceur 454
Les chiffres indiqués correspondent au modèle GS20. Source http://www.454.com/

Le dernier modèle GS FLX Titanium utilise quant à lui des puces de plus d'un million de puits permettant ainsi d'obtenir des séquences d'une longueur moyenne de 500 pb pour un total de 400 à 600 millions de bases séquencées en 10 heures. Ce chiffre permet d'estimer la capacité de séquençage à environ 1,2 milliards de bases par jour ; à comparer aux 2 millions de bases séquencées par un séquenceur classique.

2.3.1 Technologie Illumina Solexa

Tout comme le séquenceur 454, la parallélisation du séquençage du séquenceur Solexa repose sur l'utilisation d'une matrice relativement dense de polonies attachées sur une surface solide. Cependant, même si le séquençage est toujours effectué en temps réel, la technique utilisée repose sur des fluorochromes fixés sur les dNTP.

La préparation de la librairie de séquençage nécessite de lier des séquences adaptateurs aux deux extrémités de fragments d'ADN à séquencer pour pouvoir les fixer de manière

relativement espacée sur une surface plane recouverte d'amorces simple brin (Figure 10). L'amplification des fragments se fait par « amplification en pont » : l'ADN est dénaturé puis le cycle d'amplification commence par la synthèse des brins complémentaires en utilisant une amorce voisine fixée sur la surface plane. Le brin à amplifier forme ainsi un pont sur la surface. La synthèse est réalisée par la polymérase puis l'ADN est dénaturé et reste ancré sur la surface puisqu'il a été synthétisé à partir d'une amorce elle-même fixée sur cette surface. Après plusieurs cycles, une matrice de polonies ancrées sur la surface est obtenue.

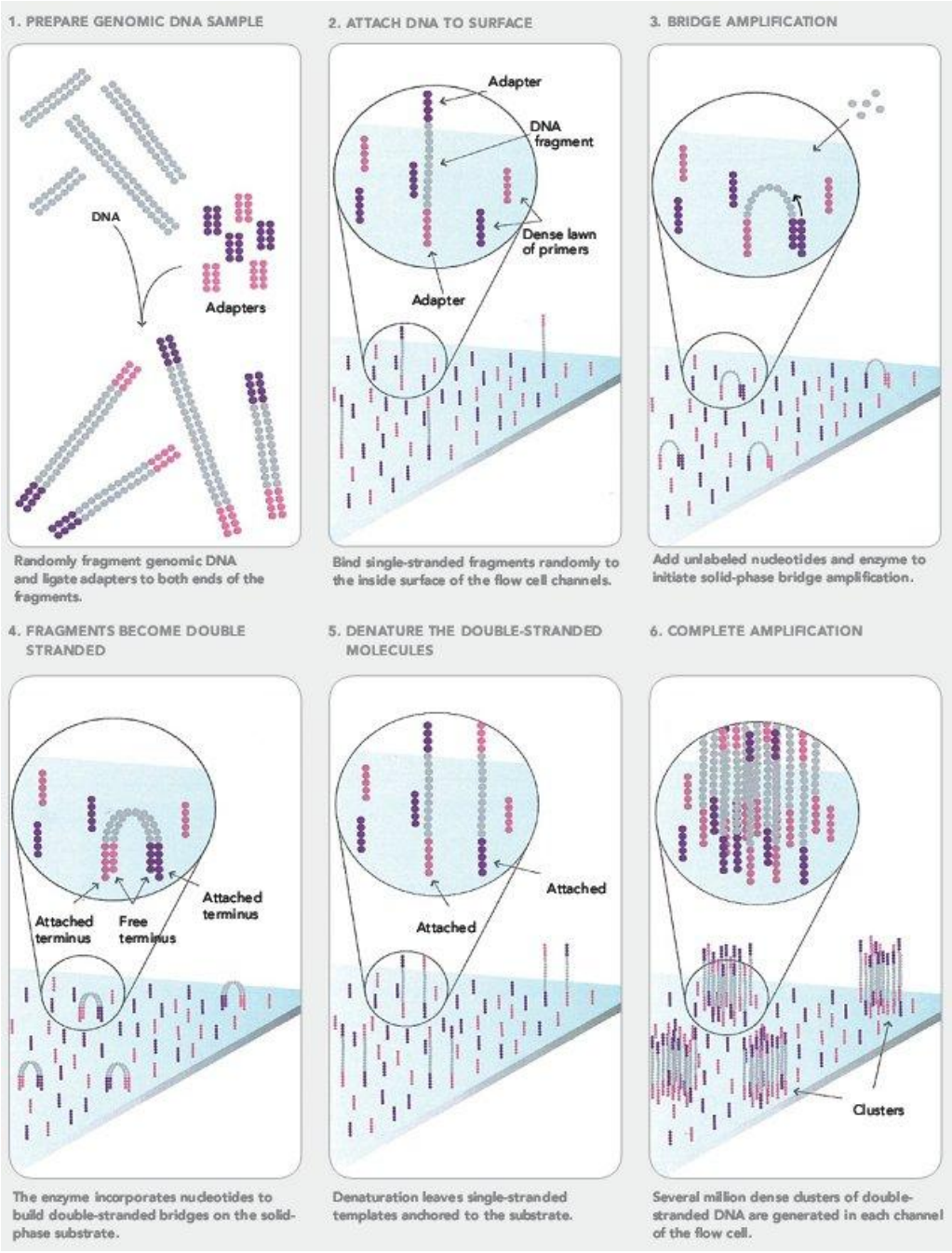


Figure 10 – Préparation de la matrice de polonies par amplification en pont du séquenceur Solexa
Source http://www.illumina.com/

Le séquençage de ces polonies est initié en incorporant un mélange des quatre dNTP terminateurs marqués (Figure 11). Ces dNTP terminateurs agissent comme des ddNTP, bloquant la synthèse par la polymérase après leur incorporation. Cependant les terminateurs utilisés par cette technologie ont la spécificité d'être réversibles, ce qui permet de reprendre la synthèse lorsqu'on le souhaite. Une fois la première série de dNTP terminateurs incorporés, une image des intensités lumineuses des différents fluorochromes est réalisée et la capacité de blocage de synthèse enzymatique des dNTP terminateurs est supprimée afin de continuer la synthèse lors du prochain cycle.

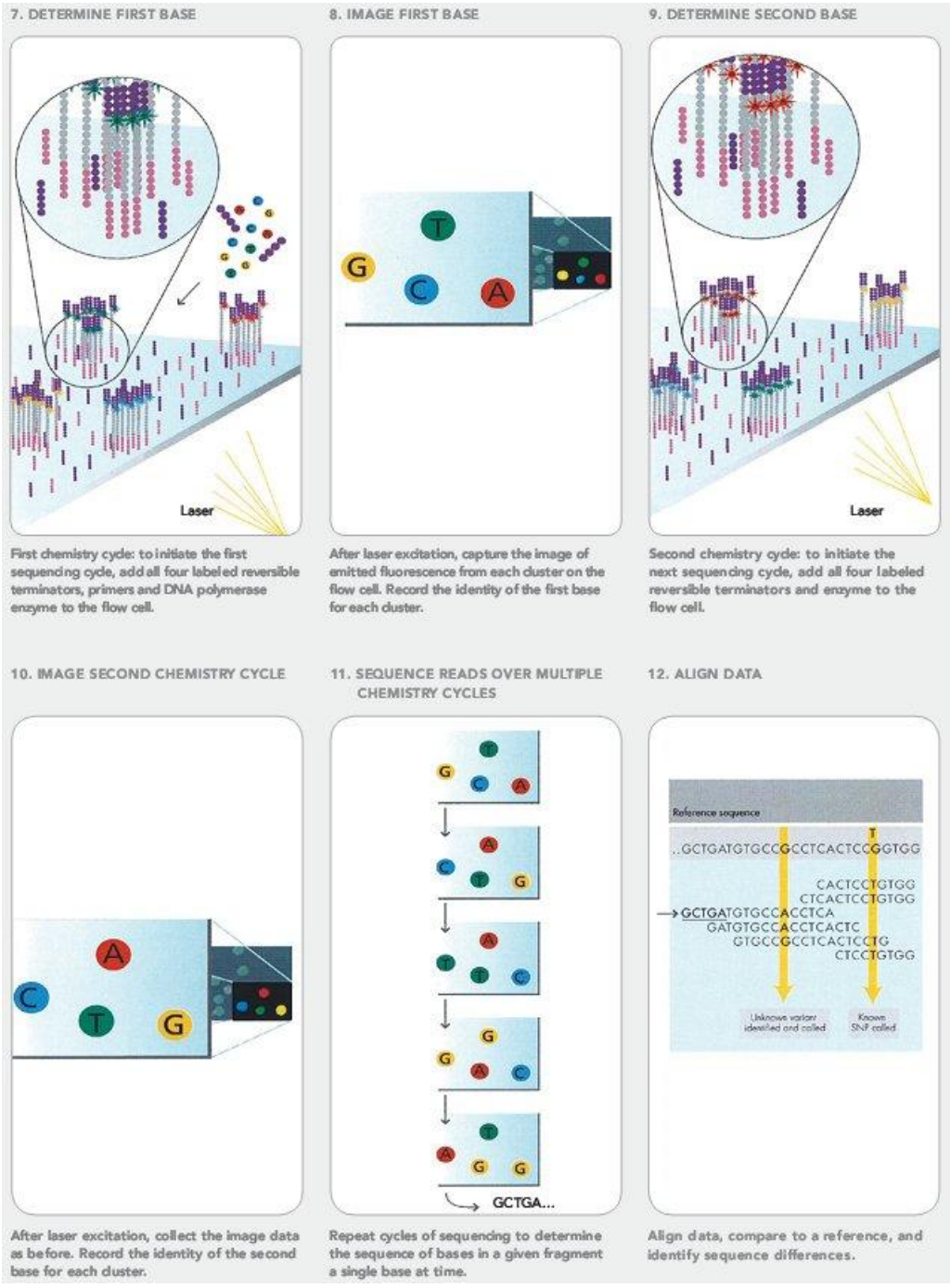


Figure 11 – Séquençage temps réel des polonies par le séquenceur Solexa
Source http://www.illumina.com/

La dernière évolution de ce système, le Solexa GA2 permet de réaliser des séquences de 50 pb pour 3 milliards de bases séquencées en environ 5 jours, soit une capacité journalière d'environ 600 millions de bases.

2.3.2 Technologie Applied Biosystems SOLiD

Le système SOLiD (*Supported Oligo Ligation Detection*) est basé sur une technologie de séquençage par ligation.

La préparation de la librairie de séquençage est réalisée de manière très similaire au séquenceur 454, de façon à obtenir des billes de polonies amplifiées par emPCR. Les extrémités 3' des sstDNA sont ensuite modifiées afin de fixer les billes sur une surface plane.

Le séquençage en lui-même est très différent des autres technologies puisqu'il ne met pas en jeu la polymérase. Ce séquençage comporte 5 cycles dont la première étape consiste à hybrider une amorce de taille « n ». Chaque cycle est lui-même constitué de 7 sous-cycles.

Un mélange d'oligonucléotides marqués par des fluorochromes est ensuite ajouté au milieu réactionnel. Ces oligonucléotides sont des 8-mers dont les positions 3 à 6 sont dégénérées, tandis que les positions 1 à 2 sont des dinucléotides spécifiques, parmi 16 possibilités différentes. Les oligonucléotides, dont les bases 1 à 2 s'hybrident avec le brin matrice, sont ligaturés à l'aide d'une ligase (Figure 12a). Les oligonucléotides non fixés sont éliminés et l'acquisition de l'image des intensités lumineuses des différents fluorochromes est réalisée. Ensuite, les trois derniers résidus des oligonucléotides sont supprimés par clivage chimique.

À la fin des 7 sous-cycles, le séquençage est réinitialisé en éliminant l'amorce et le nouveau brin ligaturé. Un nouveau cycle recommence avec une nouvelle amorce de taille « n=n-1 », décalant ainsi les bases « testées » par le séquençage par ligation. De la sorte, pour une taille séquencée de 35 pb, chaque base est séquencée 2 fois, ce qui permet de détecter d'éventuelles erreurs (Figure 12b).

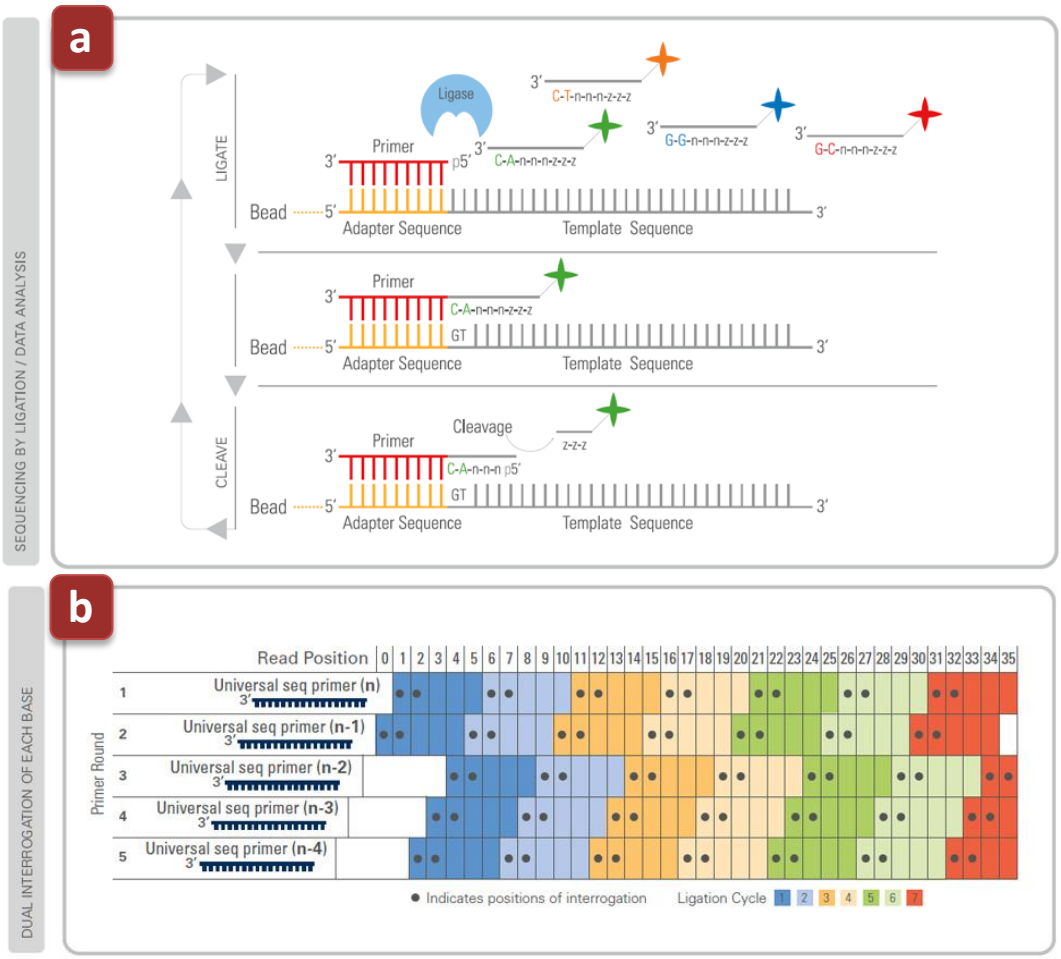


Figure 12 – Séquençage par ligation du séquenceur SOLiD
(a) Séquençage par ligation (b) Résidus séquencés à chaque cycle. Source http://www.appliedbiosystems.com/

Les oligonucléotides utilisés ne sont marqués qu'à l'aide de 4 fluorochromes différents alors qu'il existe 16 dinucléotides : 1 fluorochrome « code » donc pour 4 dinucléotides. Puisque chaque base est séquencée 2 fois, il est possible de décoder la base à l'aide des 2 couleurs émises par les fluorochromes (Figure 13).

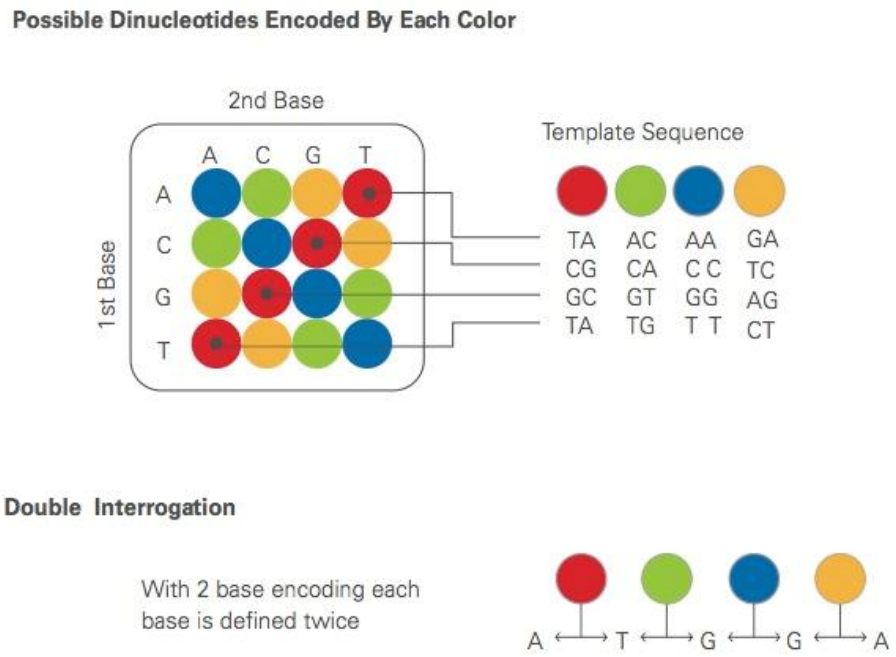


Figure 13 – Décodage de la base séquencée par un séquenceur SOLiD
Source http://www.appliedbiosystems.com/

La dernière version 3 du système SOLiD permet de séquencer jusqu'à 300 millions de séquences de 50 pb en environ 7 jours, soit une capacité journalière d'environ 2,2 milliards de bases.

2.3.3 Technologie Helicos tSMS

Le séquenceur Helicos tSMS (*true Single Molecule Sequencing*) est une amélioration de la technologie du séquenceur Illumina Solexa. Les fragments d'ADN à séquencer sont fixés sur une surface plane, mais avec une densité supérieure à celle du Solexa puisque le tSMS ne nécessite pas d'amplification, et donc pas d'espace pour la génération des polonies. Le capteur du tSMS est suffisamment performant pour enregistrer la fluorescence émise par un seul brin.

La longueur maximale des séquences lues est aussi de 50 pb, mais la densité de brins fixés permet de séquencer jusqu'à 140 millions de bases par heure, soit une capacité journalière d'environ 3,4 milliards de bases.

2.4 Bienvenue à Gattaca

La course au génome a permis le développement de technologies de séquençage de plus en plus performantes et abordables. Le Tableau 2 récapitule les performances de ces séquenceurs haut-débit comparées à celles d'un séquenceur classique.

Tableau 2 – Récapitulatif des performances des différents séquenceurs présentés

Séquenceur	*Longueur moyenne des séquences (bases)*	*Capacité journalière de séquençage (Mpb)*
Classique	1200	2
Roche 454	500	1200
Illumina Solexa	50	600
ABI SOLiD	50	2200
Helicos tSMS	50	3400

Avec l'aide de telles technologies, il devient possible de séquencer un génome humain en une journée alors qu'il a fallu près de 10 ans pour arriver à finaliser le projet génome humain. Alors que les génomes de personnalités comme Craig Venter (Levy *et al.*, 2007) et James Watson (Wheeler *et al.*, 2008) sont déjà disponibles, on évoque à présent la génomique personnelle afin de mener des études de variations génétiques de grande ampleur (projet 1000 génomes, http://www.1000genomes.org/), ou à des fins de médecine personnalisée.

Les capacités de séquençage impressionnantes de ces séquenceurs haut-débit sont tout de même à nuancer en regard du taux d'erreur de séquençage. Ces nouvelles techniques sont limitées au séquençage en une seule passe avec un taux d'erreur de 1 à 2% (Schröder *et al.*, 2009) alors que le séquençage classique permet des reséquençages hiérarchiques, abaissant le taux d'erreur à 0,1 pour 1 million (Mandel *et al.*, 2008). De plus, la plupart de ces nouvelles technologies fournissent des séquences très courtes nécessitant de lourds traitements au niveau de l'assemblage, en particulier dans le cadre de séquençage *de novo*.

Avec la multiplication des services de séquençage de génomes complets à des prix de plus en plus attractifs – Illumina pour 48 000 $ (http://www.everygenome.com/), Complete Genomics pour 5 000 $ –, la course au génome est clairement terminée. Le nouveau défi est d'être le premier à proposer le génome humain à 1 000 $…

3 ANNOTATION DE SÉQUENCES

Depuis l'avènement du haut-débit, une grande quantité d'informations a été rendue accessible, en particulier dans le domaine des séquences génomiques et des transcrits. Pour tirer pleinement profit de toutes ces séquences, il convient de les déchiffrer pour en extraire les différents composants, qu'il faudra ensuite caractériser afin de définir le plus précisément possible le, ou bien souvent les, rôles qu'ils jouent au sein d'un organisme. Tout ce processus de décryptage et de caractérisation définit le travail de l'annotation.

Ce travail peut être décomposé en deux tâches dont l'une, l'annotation structurale, consiste à identifier et à localiser les éléments génétiques en présence. La deuxième, l'annotation fonctionnelle, consiste en la détermination de la fonction de ces éléments, en particulier la fonction des gènes protéiques qui demeurent centraux dans l'annotation des génomes. Pour ces derniers, la détermination fonctionnelle se rapporte au gène dans son ensemble et à son (ses) produit(s), mais peut aussi être plus finement rapportée à certains des domaines ou résidus des protéines correspondantes, ou plus largement aux interactions dans lesquelles celles-ci sont impliquées au niveau de complexes moléculaires ou de voies métaboliques.

Le domaine de l'annotation a logiquement été profondément modifié par l'arrivée du haut-débit. Si, il y a peu de temps encore, un expert pouvait s'appuyer sur des outils bioinformatiques, des données expérimentales et de la littérature pour réaliser manuellement un travail de qualité sur une famille de gènes d'intérêt ou sur un génome de taille raisonnable, la somme actuelle des données à traiter est telle que la majorité des séquences sont annotées de manière automatisée en attendant, le cas échéant, la validation et la correction des annotations par un collège d'experts.

L'introduction dans les banques de ces séquences annotées de manière automatique induit plusieurs conséquences :

- Puisque les annotations automatiques se basent essentiellement sur les données préexistantes dans les banques, les séquences ayant une fonction spécifique ou étant divergentes des séquences préexistantes obtiendront une annotation relativement pauvre, voir nulle. L'ajout massif de séquences d'organismes divers et variés va donc enrichir les banques en séquences non annotées.
- L'annotation étant un processus relativement complexe, un algorithme d'annotation automatique peut facilement assigner une fonction à tort. La publication de cette annotation erronée dans les banques pourra alors être propagée lors de l'annotation automatique d'autres séquences, provoquant ainsi un effet boule de neige contaminant progressivement les banques de séquences.

Tout l'art de l'annotation de séquences repose donc sur la capacité à trouver le bon compromis entre la quantité de données traitées et la qualité des annotations expertisées afin de ne pas laisser s'accumuler les séquences dont la fonction est encore inconnue, sans pour autant introduire de plus en plus de bruit dans les banques.

Dans ce chapitre, nous allons décrire les principales étapes de l'annotation en présentant les diverses approches et outils utilisés, mais aussi les problèmes majeurs rencontrés régulièrement ainsi que les solutions mises en œuvre pour y remédier.

3.1 Annotation structurale

Par analogie, l'annotation structurale peut être comparée à la ponctuation d'un langage écrit. Elle permet de rendre compréhensible une suite continue de lettres en la structurant en paragraphes, en phrases et en mots. Dans le cadre de séquences nucléotidiques, l'annotation structurale permet de localiser les différents éléments génétiques qui la composent.

Pour une séquence génomique, cette localisation concerne des éléments relativement évidents tels que les gènes, les alternances introns/exons, les régions codantes (CDS, *Coding Sequence*) et les régions non traduites (UTR, *Untranslated region*), et peut être approfondie pour identifier les régions promotrices, régulatrices, les sites de début de transcription (TSS, *Transcription start site*), les sites de fixation d'histones, de réplication… Pour des séquences de transcrits, il s'agit essentiellement de localiser la région codante.

3.1.1 Prédiction de gènes

Cette localisation débute systématiquement par la prédiction des gènes qui sont les éléments fondamentaux d'un génome. La notion de gène a évolué au fil du temps et reste encore difficile à définir. Une définition acceptable serait de dire qu'un gène est une « région de séquence génomique, correspondante à une unité héréditaire, associée à des régions régulatrices, des régions transcrites et/ou d'autres régions fonctionnelles » (Pearson, 2006). Cependant la réalité est encore plus complexe, et la présence de produits fonctionnels différents partageant des régions génomiques chevauchantes redéfinit un gène en tant qu'« union de séquences génomiques codant pour un ensemble cohérent de produits fonctionnels potentiellement chevauchants » (Gerstein *et al.*, 2007). Dans le cadre de la prédiction de gènes, cette définition est nettement simplifiée puisqu'elle correspond à une région d'ADN transcrite à l'origine d'une ou plusieurs formes de variants d'épissage.

Une étape de localisation des éléments répétés doit être réalisée au préalable puisque ces éléments peuvent constituer une grande partie d'un génome, plus particulièrement chez les eucaryotes où, par exemple, ils constituent plus de 50% du génome Humain (Richard *et al.*, 2008). Ces éléments sont généralement détectés à l'aide de RepeatMasker (Smith *et al.*) qui

permet de rechercher par similarité plusieurs familles d'éléments répétés dispersés à l'aide de sa base de séquences. Il permet aussi de localiser les éléments répétés en tandem en analysant le biais en composition des séquences.

3.1.1.1 Gènes protéiques

La prédiction de gènes protéiques comprend deux grandes familles de méthodes : les méthodes *ab initio* se basant sur les propriétés intrinsèques des séquences en analysant les biais en composition entre les régions codantes et non codantes qui peuvent être complétés par la détection de différents signaux (promoteurs, sites donneurs/accepteurs, sites d'épissage, signaux de polyadénylation…), et les méthodes par similarité.

Parmi les méthodes de prédiction *ab initio*, il existe différents programmes spécialisés dans la prédiction de gènes procaryotes ou eucaryotes. Une sélection des programmes utilisés le plus couramment est présentée dans le Tableau 3.

Tableau 3 – Quelques programmes de prédiction *ab initio* de gènes protéiques
* MM : modèle de Markov, IHMM : modèle de Markov caché itératif, HMM, modèle de Markov caché, HSMM : semi modèle de Markov caché, GHMM : modèle de Markov caché généralisé, IMM : modèle de Markov interpolé, SVM : machine à vecteurs de support

Programme	***Type de gènes prédits***	***Algorithme****
GENSCAN (Burge *et al.*, 1997)	Eucaryote	MM
GeneMarkS (Besemer *et al.*, 2001)	Procaryote	IHMM
GeneMark.hmm (Lukashin et Borodovsky, 1998)		HMM
GeneMark-ES (Ter-Hovhannisyan *et al.*, 2008)	Eucaryote	HMM
GeneMark.hmm-ES (Lomsadze *et al.*, 2005)		HSMM
Glimmer (Delcher *et al.*, 2007)	Procaryote	IMM
GlimmerHMM (Majoros *et al.*, 2004)	Eucaryote	GHMM
GeneZilla/TigrScan (Majoros *et al.*, 2004)	Eucaryote	GHMM
CONTRAST (Gross *et al.*, 2007)	Eucaryote	SVM

Bien que la plupart de ces programmes utilisent maintenant des dérivés de modèles de Markov, leur spécialisation dans la recherche de gènes procaryotes ou eucaryotes est dictée par la structure même de ces gènes. En effet, les gènes protéiques procaryotes ne comportant pas de régions introniques, les modèles utilisés par les programmes sont optimisés pour détecter les régions promotrices procaryotes mieux connues et pour rechercher de longs cadres de lectures ouverts. Ces programmes arrivent ainsi à prédire les gènes procaryotes avec une sensibilité et une spécificité dépassant les 90% (Besemer *et al.*, 2001; Delcher *et al.*, 2007). Dans le cas des gènes eucaryotes, les régions promotrices sont plus variables et bien moins connues, et les régions codantes peuvent être séparées par de longues régions introniques pour former des exons relativement courts, ce qui complexifie la tâche de prédiction. Malgré cela, les programmes déployant des modèles récents qui prennent en compte de nombreux signaux (îlots CpG, sites d'épissage, sites de

polyadénylation…) permettent de détecter les gènes eucaryotes avec une spécificité pouvant atteindre 40-60% (Majoros *et al.*, 2004; Gross *et al.*, 2007).

Compte tenu des difficultés rencontrées, en particulier chez les eucaryotes, les prédictions *ab initio* sont souvent complétées, dans la mesure du possible, par des méthodes de recherche par similarité. Ces méthodes consistent à cartographier les régions transcrites à l'aide de séquences codantes. Ces séquences peuvent être des gènes déjà connus d'organismes proches, des séquences d'EST, d'ADNc ou des séquences protéiques.

La cartographie à l'aide de séquences d'EST/ADNc ou de protéines permet ainsi de valider les gènes prédits à l'aide de données expérimentales. Il existe par ailleurs plusieurs programmes qui mélangent ces deux approches pour améliorer encore les prédictions, dont CONSTRAST (Gross *et al.*, 2007) ainsi que TWINSCAN_EST et N-SCAN_EST (Wei et Brent, 2006).

3.1.1.2 Gènes non traduits

Les ARN de transfert (ARNt) sont aisément identifiables par la structure secondaire peu variable qu'ils adoptent. Le logiciel tRNAscan-SE (Lowe et Eddy, 1997) permet de les localiser en recherchant les motifs et les appariements caractéristiques de cette structure secondaire.

Les ARN ribosomaux (ARNr) sont souvent identifiés par similarité en raison de leur forte conservation. Cette caractéristique, ainsi que leur caractère universel, en a fait les séquences de références de la phylogénétique, et plusieurs bases de séquences dédiées ont été construites en ce sens : SILVA (http://www.arb-silva.de/) (Pruesse *et al.*, 2007) et le *Comparative RNA Web* (http://www.rna.ccbb.utexas.edu/) (Cannone *et al.*, 2002). Il existe cependant plusieurs programmes de localisation d'ARNr basés sur des modèles HMM comme RNAmmer (Lagesen *et al.*, 2007), ou un programme permettant de localiser des ARNr fragmentaires des données de métagénomique (Huang *et al.*, 2009).

La liste des autres ARN non codants est trop longue pour ne citer que quelques ressources. Il existe toutefois la banque Rfam (Gardner *et al.*, 2009), qui regroupe en une seule ressource les familles des ARN non codants afin de les localiser dans des séquences avec un modèle de covariance.

3.1.2 Autres éléments génétiques

Au-delà des gènes, il existe de nombreux autres éléments génétiques dans les génomes parmi lesquels on pourrait citer :

- Les éléments de régulation d'expression. Ces sites de fixation des facteurs de transcription étant de très courte taille et souvent très variables, les recherches *in silico* s'appuyant sur des approches de type motif ou profil génèrent très souvent

beaucoup de faux positifs. Pour filtrer ces derniers, on a recours à la génomique comparative avec les techniques de *phylogenetic footprinting/shadowing*. Le principe sous-jacent est que ces éléments étant indispensables au bon fonctionnement de la transcription, ils sont soumis à une pression de sélection importante et sont donc conservés entre espèces proches (Dickmeis et Müller, 2005). Les sites potentiels détectés par recherche de motifs ou de profils localisés dans des zones non conservées entre un ensemble de génomes sélectionnés pourront ainsi être éliminés.

- Des sites de fixations divers, dont les éléments de régulation, peuvent être localisés à l'aide d'expériences de ChIP-chip (*Chromatin immunoprecipitation on chip*) (Gilchrist *et al.*, 2009). La protéine d'intérêt en interaction avec l'ADN (ainsi que les autres protéines) est fixée *in vivo* de manière covalente à l'ADN à l'aide de formaldéhyde. L'ADN est ensuite extrait et fragmenté, puis les fragments fixés avec la protéine d'intérêt sont purifiés grâce à des anticorps immobilisés ciblant cette protéine. Les fragments purifiés sont alors détectés avec une puce à ADN de type '*tiling array*'. Les régions génomiques localisées sur la puce par les fragments purifiés indiquent les localisations des sites de fixations de la protéine d'intérêt. Une technique alternative plus récente, le ChIP-Sequencing met à profit les nouveaux séquenceurs haut-débit pour séquencer directement les fragments obtenus.
- Les sites de début de transcription (TSS, *Transcription start site*). Ces derniers peuvent être localisés grâce à des expériences de CAGE (*Cap analysis gene of expression*). Cette technique permet d'obtenir des marqueurs de 20 bp à partir de la partie la plus en amont des ARNm, c'est-à-dire les 20 premières bases de la région 5'UTR. Ces marqueurs qui correspondent au tout début de gènes transcrits, sont cartographiés sur la séquence génomique pour localiser les TSS (Shimokawa *et al.*, 2007).

Les approches pour localiser les différents éléments d'un génome sont donc nombreuses et font appel à des compétences multiples. Cette réalité est parfaitement illustrée par le projet ENCODE (*ENCyclopedia Of DNA Elements*) qui ambitionne d'identifier tous les éléments fonctionnels du génome humain. Un projet pilote a été mené sur 1% du génome humain (environ 30 Mpb) pour tester la faisabilité du projet global, tout en évaluant plusieurs techniques expérimentales haut-débit (Birney *et al.*, 2007). Les premiers résultats de ce pilote se sont avérés surprenants : le génome humain serait transcrit sur quasiment toute sa longueur, ce qui fait que chacune de ses bases est au moins associée à un transcrit. Parmi ces derniers, de nombreux transcrits non codants ont été découverts et localisés comme chevauchant des régions codantes pour des protéines, ou à l'intérieur de régions que l'on ne

croyait jamais exprimées, allant ainsi à l'encontre de nombreux dogmes établis depuis plusieurs décennies.

L'analyse des génomes n'a donc pas fini de nous révéler ses secrets, et une annotation structurale exhaustive du génome humain nécessitera encore de nombreuses années avant d'être finalisée.

3.1.3 Prédiction de régions codantes dans les transcrits

La prédiction des régions codantes (CDS) dans les transcrits consiste à différencier la région codante des régions non traduites (UTR) (Figure 14). La définition exacte des bornes de la région codante n'est cependant pas aussi triviale qu'il y parait en raison des erreurs de séquençages qui peuvent être nombreuses dans le cas de singletons et du caractère incomplet de nombre de séquences d'ADNc. Ces erreurs peuvent aboutir à des décalages de cadre ou à un codon stop erroné.

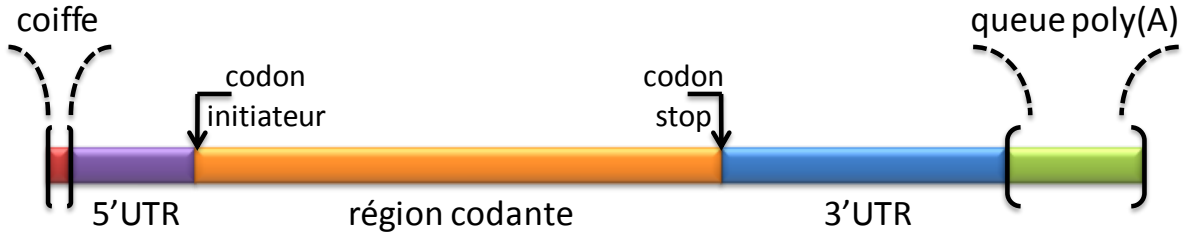


Figure 14 – Structure simplifiée d'une séquence d'ARNm
La coiffe et la queue poly(A) sont des structures uniquement présentes chez les ARNm eucaryotes et sont ajoutées lors de l'étape de maturation de l'ARN pré-messager. La région codante est la partie de d'ARNm qui est traduite en protéine et débute par un codon initiateur, se termine par un codon stop. Les régions UTR ne sont pas traduites mais servent principalement à réguler la traduction de l'ARNm, à stabiliser et empêcher sa dégradation (de même que la coiffe et la queue poly(A)).

La détection de la région codante est généralement effectuée en modélisant le biais d'utilisation en hexanucléotides aussi bien pour les régions codantes que pour les régions 5' et 3'UTR. Les logiciels les plus répandus utilisent des modèles qui tiennent compte, non seulement de ces biais, mais aussi des sites de transitions entre les différentes régions (site initiateur et codon stop). Ces modèles sont aussi généralement tolérants à l'erreur, afin de modéliser, voir de corriger, les erreurs de séquençage.

ESTScan2 (Lottaz *et al.*, 2003) et FrameDP (Gouzy *et al.*, 2009) emploient cette technique à l'aide de modèles HMM ; DIANA-EST (Hatzigeorgiou *et al.*, 2001) fait de même à l'aide de réseaux neuronaux et DECODER (Fukunishi et Hayashizaki, 2001) utilise une fonction de vraisemblance.

Cependant, du dire même de leurs auteurs, la prédiction exacte des sites de début et de fin de CDS reste délicate (environ 60% de prédictions correctes pour ESTScan2 (Lottaz *et al.*, 2003)). C'est pourquoi il existe des logiciels qui combinent ces méthodes de prédiction avec

des méthodes par similarité de séquence, dont le logiciel prot4EST (Wasmuth et Blaxter, 2004) qui s'appuie sur les prédictions d'ESTScan2, DECODER et des prédictions par similarité par BLASTX.

3.2 Annotation fonctionnelle

L'annotation fonctionnelle, en reprenant notre analogie littéraire, consiste à définir avec une précision plus ou moins grande, le sens des mots structurant un langage. Du point de vue biologique, cela consiste à attribuer une fonction biologique à chacun des gènes ou produits de gènes d'un génome ou d'une collection de transcrits. Cette définition soulève un point crucial de l'annotation : qu'est-ce que la fonction d'un gène ?

3.2.1 Qu'est-ce que la « fonction » d'un gène ?

La définition de la fonction biologique d'un gène reste vague et n'a jamais été clairement définie (Friedberg, 2006). Le sens biologique de la fonction est grandement dépendant du point de vue avec lequel on la décrit. En prenant par exemple le cas d'école des kinases, on peut définir leur fonction biochimique par le transfert d'un groupement phosphate sur un groupement hydroxyle. Mais intégrées dans un contexte cellulaire, les kinases sont impliquées dans la régulation de nombreuses voies métaboliques. De même, dans un contexte médical, une déficience fonctionnelle d'une kinase spécifique peut très bien entrainer l'apparition d'une maladie.

Un gène n'a donc pas une fonction, mais des fonctions caractérisées différemment selon l'intérêt qui est porté au moment de l'annotation. C'est pourquoi, dans le contexte de l'annotation, le terme « fonction » englobe un ensemble de fonctions moléculaires, de localisations cellulaires, de domaines fonctionnels, de voies métaboliques, de signaux de localisation, ou toute autre caractéristique que l'on peut rattacher à un gène ou à son produit.

Tout comme dans le cadre de l'annotation structurale, les approches utilisées en annotation fonctionnelle sont très variées, et nous n'allons parcourir ici que les méthodes les plus courantes, regroupées selon le type de fonction recherchée.

3.2.2 Annotation de définition fonctionnelle

L'attribution d'une définition synthétique constitue une étape clé d'un travail d'annotation fonctionnelle. Bien que, comme cela a été décrit ci-dessus, cette définition soit souvent fortement dépendante du contexte et ne permette pas de décrire entièrement le gène ou son produit, elle permet souvent de se forger rapidement une idée globale de leur fonction réelle.

Typiquement, cette attribution est réalisée en transférant la fonction de séquences similaires détectées à l'aide des programmes BLAST ou PSI-BLAST (Altschul *et al.*, 1997), PSI-BLAST étant une version itérative de BLAST permettant de détecter des similarités entre protéines plus éloignées. Cette famille de programmes effectue les recherches à l'aide d'alignements deux-à-deux, locaux et non optimaux, entre la séquence recherchée et les séquences d'une ou plusieurs banques. Les résultats de recherches (*hits*) étant scorés, ce transfert peut s'effectuer sur la base de critères simples, comme par exemple transférer la fonction de la séquence présentant le meilleur score ou la meilleure *E-value*, ou des critères plus complexes combinant des seuils d'*E-value*, la longueur des segments alignés, et l'analyse des séquences sélectionnées pour en extraire la meilleure définition. Ces méthodes par similarité simple ont l'avantage d'être très rapides mais peuvent engendrer des erreurs de prédiction (Sjölander, 2004) et nécessitent leur validation par un expert.

Un moyen plus élégant d'attribuer une définition fonctionnelle passe par exemple par l'utilisation de la banque TIGRGAMs (Selengut *et al.*, 2007). TIGRFAMs est une banque construite manuellement par des experts. Elle regroupe une collection d'alignements de séquences multiples de familles ou de sous-familles protéiques partageant une même fonction moléculaire. Ces alignements multiples sont utilisés pour générer des modèles HMM qui permettent de classifier une séquence d'intérêt à l'intérieur d'une famille ou sous-famille et ainsi d'attribuer la fonction correspondante. La banque HAMAP (Gattiker *et al.*, 2003) utilise une approche similaire pour annoter une protéine à l'intérieur de ses familles d'orthologues bactériens. Cependant, le profil HMM peut être complété par un jeu de règles pouvant moduler la fonction attribuée. Par exemple en fonction de l'appartenance au domaine Bactérie ou Archée, les règles de la famille MF_01224 de la banque HAMAP attribuent respectivement les définitions '*Molybdenum cofactor biosynthesis protein C*' ou '*Probable molybdenum cofactor biosynthesis protein C*'.

Lorsqu'une séquence de gène ne peut être annotée sur la base de similarité de séquence, des approches de génomique comparative permettent d'obtenir une piste approximative sur sa fonction potentielle :

- La méthode des gènes voisins exploite l'organisation de gènes procaryotes en opérons. Les gènes d'un opéron sont co-transcrits et codent généralement pour des protéines qui interagissent entre elles (physiquement et/ou fonctionnellement) au sein d'un complexe macromoléculaire, d'une voie métabolique... On peut ainsi inférer un lien fonctionnel entre les membres d'un opéron (Overbeek *et al.*, 1999).
- Dans un même ordre d'idée, la méthode « Pierre de Rosette » part de l'observation que dans certains organismes, des protéines interagissant entre elles sont codées par

des gènes bien distincts, alors que leurs orthologues peuvent être fusionnés en une seule protéine dans d'autres organismes, ce qui démontre la nécessité d'une co-régulation et donc une fonction proche (Marcotte et Marcotte, 2002).

- Enfin, la méthode du profil phylogénétique implique d'établir les profils de présence/absence d'orthologues de plusieurs génomes, puis de les partitionner afin d'extraire les gènes qui au fil de l'évolution ont été acquis ou perdus de manière simultanée, indiquant ainsi un lien fonctionnel potentiel (Ranea *et al.*, 2007).

3.2.3 Annotation de domaines protéiques

La recherche de domaines fonctionnels peut être tout aussi informative qu'une définition fonctionnelle et présente l'avantage d'être plus sensible. En effet, les protéines qui partagent une fonction commune peuvent diverger largement en termes de séquence mais vont conserver les motifs de séquence ou de structure nécessaires à leur cette fonction.

PROSITE (Hulo *et al.*, 2008) et Pfam-A (Finn *et al.*, 2008) figurent parmi les banques de familles et de domaines protéiques généralistes les mieux fournies et leur construction manuelle par des experts leur confère une grande qualité. Ces deux banques fournissent des signatures sous forme de motifs de séquences, de profils (matrices positions spécifiques) ou de modèles HMM permettant d'identifier les domaines incriminés.

PROSITE et Pfam ont été intégrés à la banque InterPro (Hunter *et al.*, 2009) qui regroupe plusieurs autres banques de motifs dont TIGRFAMs. Le logiciel InterProScan (Mulder et Apweiler, 2007) permet de balayer facilement un grand nombre de séquences afin d'y rechercher les domaines des différentes banques intégrées par InterPro.

3.2.4 Annotation Enzyme Commission

La classification Enzyme Commission (EC) (Webb et International Union of Biochemistry and Molecular Biology., 1992) permet de classer sous une forme hiérarchique les fonctions moléculaires catalysées par les enzymes en fonction du type de réaction catalysée et du substrat transformé. Chaque niveau de la hiérarchie correspond à un niveau de précision de plus en plus élevé dans la description des réactions, dont le quatrième niveau permet d'obtenir le nom complet de l'enzyme (Cf. Numéros *Enzyme Commission* : ENZYME, page 75).

Si en apparence une réaction catalysée par une enzyme peut sembler vague, celle-ci prend tout son sens dans le contexte des voies métaboliques qui ont été extraites de la littérature pour être décrites par les banques KEGG PATHWAY (Kanehisa *et al.*, 2008), MetaCyc (Caspi *et al.*, 2008) ou Reactome (Matthews *et al.*, 2009a). En effet, les numéros EC associés à chaque description d'enzyme sont souvent utilisés afin de cartographier les gènes d'un

organisme sur l'ensemble des voies métaboliques connues afin de mieux comprendre sa physiologie.

L'attribution de numéro EC peut être effectué sur la base d'un transfert d'annotation (Tian et Skolnick, 2003) mais il existe aussi la banque de profils PRIAM (Claudel-Renard *et al.*, 2003) et le logiciel en ligne EFICAz (Arakaki *et al.*, 2009). Les profils de la banque PRIAM ont été réalisés à partir des domaines conservés des familles d'enzyme extraites de la banque ENZYME (Bairoch, 2000) et peuvent être rapatriés en local pour effectuer les prédictions. EFICAz, quant à lui, utilise les résidus discriminants des familles d'enzymes pour combiner des modèles HMM et SVM. Il peut aussi utiliser des modèles provenant de domaines Pfam et PROSITE afin d'obtenir une plus grande spécificité.

3.2.5 Annotation Gene Ontology

La Gene Ontology (GO) (Ashburner *et al.*, 2000) est une ontologie, c'est-à-dire un ensemble de termes d'un vocabulaire contrôlé qui sont reliés entre eux par des relations, destiné à modéliser les informations d'un domaine particulier, sous forme d'un graphe orienté acyclique (DAG, *Directed Acyclic Graph*).

GO est constituée de trois vocabulaires distincts décrivant les fonctions moléculaires, les processus biologiques et les localisations cellulaires des gènes et de leurs produits (Cf. Gene Ontology, page 74). Elle a été conçue à des fins de standardisation pour faciliter l'extraction de connaissances des informations d'annotation de manière informatisée (Cf. Standardisation des données : les ontologies, page 43). En effet, les définitions fonctionnelles des banques de séquences étant décrites en langage naturel sous forme libre, on peut retrouver une même séquence sous les termes de '*Glycyl-tRNA synthetase*', '*Glycine-tRNA synthetase*', '*Glycine--tRNA ligase*' ou tout simplement '*GlyRS*'.

La Gene Ontology figure parmi les ontologies les plus utilisées en biologie, ce qui a favorisé son acceptation et son intégration par les banques majeures de données biologiques. Le projet GOA (GO Annotation) (Barrell *et al.*, 2009) dont le but est de réaliser l'annotation GO manuelle et automatique sur les entrées de la banque UniprotKB (Boeckmann *et al.*, 2003), illustre parfaitement le succès de GO. Ces annotations sont accompagnées d'un code (*evidence code*) indiquant la source de l'annotation afin d'accorder plus ou moins d'importance à l'association gène-annotation GO considérée (Tableau 4).

Tableau 4 – Liste des codes indiquant la source d'annotation GO
Adapté de http://www.geneontology.org/GO.evidence.shtml

Code	***Signification***	***Description***
Codes de preuves expérimentales		
EXP	Inferred from Experiment	Type expérimental non spécifié
IDA	Inferred from Direct Assay	Tests fonctionnels directs
IPI	Inferred from Physical Interaction	Tests d'interaction protéine-protéine/ADN
IMP	Inferred from Mutant Phenotype	Tests de comparaisons sauvages/mutants
IGI	Inferred from Genetic Interaction	Tests génétiques
IEP	Inferred from Expression Pattern	Tests de niveaux d'expression
Codes de preuves d'analyse informatique		
ISS	Inferred from Sequence or Structural Similarity	Par similarité de séquence ou de structure
ISO	Inferred from Sequence Orthology	Par orthologie de séquence
ISA	Inferred from Sequence Alignment	Par alignement multiple de séquences
ISM	Inferred from Sequence Model	Par prédiction d'annotation structurale
IGC	Inferred from Genomic Context	Par contexte génétique
RCA	Inferred from Reviewed Computational Analysis	Par analyses statistiques de grands jeux de données
Codes de preuves provenant de la littérature		
TAS	Traceable Author Statement	Preuve trouvée et justifiée par une citation
NAS	Non-traceable Author Statement	Preuve trouvée sans justification par une citation
Codes de preuves de révision par un expert		
IC	Inferred by Curator	Annotation trouvée mais non justifiable à l'aide des autres codes
ND	No biological Data available	Recherches effectuées mais aucune annotation trouvée
Code de preuve d'annotation automatique		
IEA	Inferred from Electronic Annotation	Tout type d'annotation automatique

Outre les annotations manuelles de haute qualité, le Consortium GOA fournit des tables de conversions permettant de tisser des liens entre les entrées de plusieurs banques de données biologiques et leur annotation GO. Ces tables de conversions sont réalisées manuellement par les propres membres des différentes banques de données biologiques composant le Consortium GOA et sont ensuite utilisées dans l'annotation automatique GOA et redistribuées pour être utilisées publiquement. Ainsi l'annotation GO peut très facilement être réalisée à partir de mots clés de la banque UniprotKB (table Keywords2Go), de numéro EC (EC2Go), de domaines InterPro (InterPro2Go), Pfam (Pfam2GO), de familles HAMAP (HAMAP2GO), Rfam (Rfam2GO)…

Il existe aussi des outils pour réaliser automatiquement cette tâche à partir d'une séquence protéique non préalablement annotée. Parmi ceux-ci, on peut citer blast2GO (Götz *et al.*, 2008) et GOAnno (Chalmel *et al.*, 2005). Blast2GO réalise une recherche BLAST de la protéine d'intérêt avant de récupérer l'annotation GO des *hits* qui ont été sélectionnés en fonction d'un seuil d'*E-value* et de longueur d'alignement. Un score est ensuite calculé pour chaque terme GO récupéré afin de déterminer si le transfert d'annotation est possible. GOAnno est présenté dans le chapitre GOAnno, page 94.

3.2.6 Annotation de différents signaux

Certains signaux, notamment des signaux de localisation, sont aisément identifiables sur la base de la séquence protéique et peuvent être très utiles pour orienter l'annotation de séquences. Encore une fois, les différents logiciels et méthodes sont nombreux, mais on peut citer parmi les plus utilisés :

- TMHMM (Krogh *et al.*, 2001) : permet de prédire avec précision les hélices transmembranaires des séquences protéiques à l'aide d'un modèle HMM tenant compte du cœur et des extrémités de l'hélice, des domaines globulaires et des régions près de la membrane,
- SignalP (Bendtsen *et al.*, 2004) : combine un réseau de neurones et un modèle HMM pour prédire les sites de coupure des peptides signaux, indiquant que la protéine en question est potentiellement sécrétée,
- PSORTb (Gardy *et al.*, 2005) : combine plusieurs classificateurs pour localiser les protéines bactériennes au niveau des composants cellulaires. Une version existe aussi pour les protéines eucaryotes,
- predictNLS (Cokol *et al.*, 2000) : prédit les signaux de localisation nucléaire (NLS) à l'aide de sa base de motifs construite à partir de la littérature.

3.3 Du chaos vers la standardisation

Le monde de l'annotation est vaste et passionnant, et chaque nouveau génome apporte son lot de découvertes tout en ouvrant la porte à de nouvelles questions soulevées à la vue de ces résultats. Cependant cette vision idyllique est quelque peu entachée par le problème majeur apporté par le haut-débit, et qui paradoxalement fait sa force : l'énorme quantité de données générées.

Ainsi, sous l'effet de masse, de nombreuses séquences sont encore en attente d'annotation. Pire encore, les objectifs de certains groupes de travail se limitent à fournir une vague définition comportant un numéro de clone (par exemple '*BY621188 RIKEN full-length enriched, visual cortex Mus musculus cDNA clone K330305I17 3', mRNA sequence*').

Un autre effet secondaire de la multiplication des génomes, et donc des groupes de travail, est l'apparition d'une pléthore de protocoles d'annotation, souvent peu détaillés au travers de la littérature, apparaissant ainsi comme des « boites noires » dont il est difficile d'estimer la qualité. Ce travail d'estimation est d'autant plus difficile à réaliser que la diversité des outils et des méthodes employés au niveau de ces protocoles d'annotation est grande. Lors de l'écriture de ce manuscrit, près de 90 000 articles indexés par Medline comportent les termes '*annotation*' ou '*prediction*' au sein de leur résumé, dont environ 9 000 rien qu'au cours de l'année 2008. Ainsi, même si la moitié de ces articles seulement traitent effectivement de méthodes de prédiction ou d'annotation, il est devenu humainement impossible de parcourir une telle quantité de littérature.

C'est en partant de ce constat préoccupant que des consortiums constitués des grands acteurs du domaine se sont formés afin d'élaborer des standards biologiques permettant de favoriser l'accès et l'échange d'informations. Parmi les pionniers de la fin des années 1990, on retrouve le *Gene Ontology Consortium* (GO Consortium) (Ashburner *et al.*, 2000) et la *Microarray Gene Exression Data Society* (MGED) (Ball et Brazma, 2006). Depuis, de nouveaux Consortiums majeurs continuent d'apparaitre, tels que l'*Open Biomedicals Ontologies Consortium* (OBO) (Smith *et al.*, 2007) et le *Genomic Standards Consortium* (GSC) (Field, Garrity, Sansone, *et al.*, 2008), ce qui démontre la prise de conscience des enjeux encourus et de la nécessité de standards en biologie. Cette volonté de standardisation est très bien illustrée par une édition spéciale du journal OMICS qui invite différentes communautés à participer à l'élaboration et à l'adoption de standards biologiques (Field et Sansone, 2006).

Dans cette fin de chapitre, nous allons décrire quelques standards majeurs issus de ces différentes initiatives qui sont applicables au domaine de l'annotation.

3.3.1 Standardisation des données : les ontologies

L'utilité des ontologies a été brièvement décrite dans le chapitre Annotation Gene Ontology, page 40. Même si elle reste l'exemple type de l'intégration des ontologies en biologie, GO n'est pas la seule des ontologies biologiques. Le projet OBO (http://obofoundry.org/) est un dépôt d'ontologies biomédicales qui coordonne près de 100 projets. Pour être retenue et répertoriée, une ontologie doit respecter un certain nombre de critères définis par la communauté afin de maintenir un niveau de qualité élevé.

Parmi les ontologies les plus utilisées et les plus pertinentes par rapport à ce manuscrit, on retrouve bien évidemment GO, mais aussi la *Sequence Ontology* (SO) (Eilbeck *et al.*, 2005), la *Protein Feature Ontology* (Reeves *et al.*, 2008) qui est maintenant intégrée à SO en tant que *Sequence Ontology Feature Annotation* (SOFA), et la Protein Ontology (PRO) (Natale *et al.*, 2007).

SO a été développée pour homogénéiser le vocabulaire et les liens entre les différents concepts utilisés en annotation structurale. SOFA, l'ontologie sœur de SO, permet quant à elle de décrire les annotations fonctionnelles de séquences, à un niveau global (termes non-positionnels, toute la séquence est concernée) ou à un niveau local (termes positionnels, régions ou résidus de la séquence). L'ontologie PRO est une ontologie ambitieuse de haut-niveau qui regroupe plusieurs ontologies (GO, SO, Protein Modification Ontology – http://psidev.sourceforge.net/mod/, et Disease Ontology – http://diseaseontology.sourceforge.net/) grâce à ses propres termes. L'objectif est de modéliser tous les aspects des protéines en termes de familles de gènes, de leurs produits et des différents isoformes protéiques.

La force des ontologies réside dans leur capacité à intégrer les différentes informations. Ainsi, lorsqu'on annote une séquence en tant qu'ARN de transfert, il suffit de la marquer du code '*tRNA*' de la SO (SO:0000253) pour qu'une personne, ou un logiciel informatique, n'ayant aucune notion dans le domaine, puisse consulter cette ontologie et savoir que cette séquence fait aussi partie de la famille des ARN non codants (*ncRNA*) (Figure 15). D'une part, cela lève toute ambigüité sur les différents termes et leurs relations et, d'autre part, cela permet à un logiciel de traiter efficacement des informations à l'aide d'algorithmes de parcours de graphes, sans avoir préalablement programmé explicitement la nature des informations manipulées.

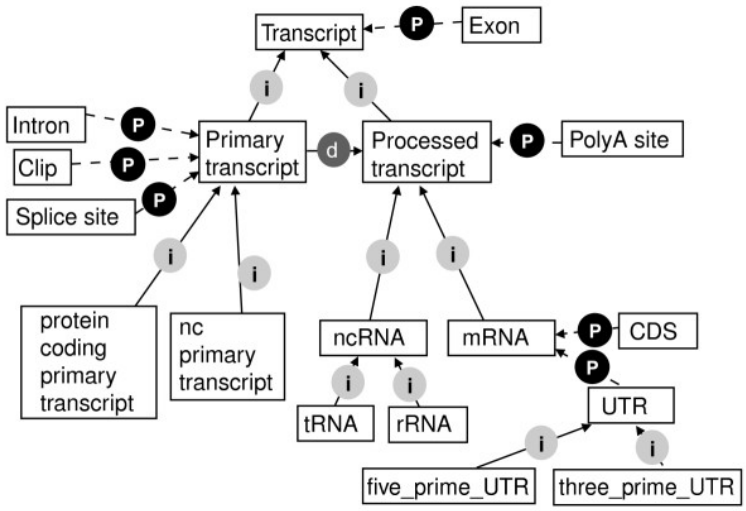


Figure 15 – Extrait du modèle de la Sequence Ontology enraciné à partir du terme '*Transcript*'
Les termes apparaissent dans les boites. Les relations marquées d'un « i » sont de type '*kind_of*' (type de), d'un « P » de type '*part_of*' (partie de), et celle d'un « d » est de type '*derives_from*' (dérive de). La lecture se fait dans le sens des flèches représentant les relations : par exemple un '*tRNA*' est un type de '*ncRNA*', et une région '*CDS*' fait partie d'un '*mRNA*'. Source (Eilbeck *et al.*, 2005).

Les différentes ontologies disponibles sur le site OBO Foundry peuvent être téléchargées dans différents formats de fichiers visualisables et éditables graphiquement grâce aux logiciels Protégé (Rubin *et al.*, 2007) ou OBO-Edit (Day-Richter *et al.*, 2007). La visualisation sous forme arborescente de ces ontologies peut aussi être effectuée à l'aide de l'*Ontology Lookup Service* (http://www.ebi.ac.uk/ontology-lookup/).

3.3.2 Standardisation des formats d'échanges

La standardisation des données n'a de sens que dans le cadre de l'échange de données par l'intermédiaire de formats de fichiers ou de protocoles de transferts standardisés. En effet, l'organisation et la structuration des données sont essentielles afin de pouvoir retrouver l'information souhaitée.

La standardisation de formats d'échanges de séquences et de leur annotation existe depuis la naissance des toutes premières banques biologiques, où l'acceptation de nouvelles données n'est valable qu'après une vérification plus ou moins stricte de leur conformité au format demandé. Ces formats sont très largement documentés et disponibles sur les sites Web de chaque banque de séquences publiques.

L'accès à ces banques est public et leurs données sont librement téléchargeables. Il y a cependant au moins deux raisons valables de vouloir échanger des données sous d'autres formes que celles des banques publiques :

- Pour une raison de réannotation de séquences. En effet, dans les banques publiques, les dépositaires de séquences en deviennent les propriétaires, et eux seuls peuvent en modifier le contenu. Les banques de séquences manuellement entretenues par des experts et à libre accès, telles que UniprotKB/Swissprot (Boeckmann *et al.*, 2003) ou RefSeq (Pruitt *et al.*, 2007), proposent bien la correction d'annotations, mais le passage obligatoire par la vérification effectuée par ces même experts peut ralentir drastiquement le processus de correction. En échangeant les nouvelles annotations dans un format adéquat, il devient alors possible de partager rapidement son travail en attendant la mise à jour des banques par les experts.
- Pour une raison de coûts matériels. Avec l'apparition de techniques de haut-débit, la taille des banques s'est accrue de manière exponentielle. La mise à jour des banques exige maintenant un espace disque important (plusieurs centaines de giga-octets) et bien plus en terme de bande passante réseau. En décentralisant les données sur plusieurs sites et en ne distribuant que les données nécessaires, les investissements en temps et en argent peuvent être réduits.

3.3.2.1 Le format GFF

Il existe depuis le milieu des années 1990 un format d'échange standard des données d'annotations : le format GFF (Generic Feature Format) (http://www.sanger.ac.uk/Software/formats/GFF/). Ce format permet de décrire sous forme de texte plat des annotations en se rapportant à une séquence de référence indiquée par son numéro d'accès et la source de sa banque de données (GenBank, EMBL, UniProt…) (Annexe 3). Avant sa version 3 (GFF3, http://www.sequenceontology.org/gff3.shtml), la description des annotations était libre et ne comportait aucune restriction. Le format GFF3 est maintenant géré par la Sequence Ontology et se conforme à cette dernière pour décrire ces annotations.

Les banques thématiques telles que WormBase (Bieri *et al.*, 2007), FlyBase (Tweedie *et al.*, 2009) ou SGD (Hong *et al.*, 2008), ainsi que les navigateurs de génomes tels que ceux d'Ensembl (Hubbard *et al.*, 2009) et de l'UCSC (Kuhn *et al.*, 2009) permettent l'échange de données au format GFF et une représentation graphique des données qu'il contient (Figure 16). Dans cette représentation, chaque séquence ou annotation de séquence peut être placée sur une « piste » (ou ligne) différente, pour éviter les superpositions d'annotations qui rendraient difficile la visualisation. Le résultat obtenu est un alignement visuel d'annotations multiples. La banque UniProt propose aussi d'exporter ses annotations directement dans ce format.

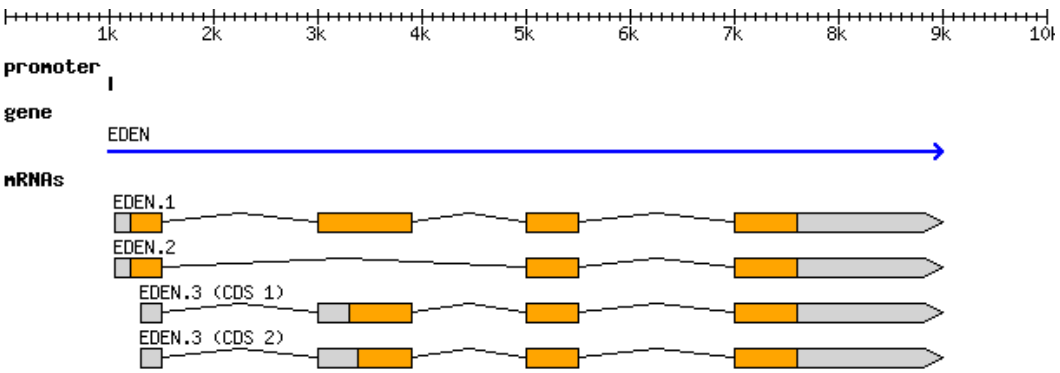


Figure 16 – Représentation sous forme de pistes des données d'annotation de l'Annexe 3
Sous le gène « EDEN » figurent les 3 variants d'épissage d'ARNm. Le troisième variant est représenté deux fois car il comporte 2 prédictions différentes de région codante (CDS, en orange). Source http://www.sanger.ac.uk/Software/formats/GFF/GFF_Spec.shtml

3.3.2.2 Protocole Distributed Annotation System

Le protocole *Distributed Annotation System* (DAS) (Jenkinson *et al.*, 2008) est la transcription plus évoluée du format GFF sous forme de Services Web (WS, *Web Service*). DAS permet de distribuer des annotations à la demande de manière totalement décentralisée. Ce système est composé d'un ensemble de serveurs maintenus par les entités voulant proposer leurs

annotations. Par exemple, les banques Ensembl, UCSC, WormBase et FlyBase disposent chacune d'un serveur DAS pour distribuer leurs annotations. Pour se faire connaître par les clients DAS, les serveurs DAS doivent s'enregistrer auprès de l'annuaire du système DAS : le *DAS registration server* (http://www.dasregistry.org/). À l'heure actuelle, 633 serveurs DAS sont enregistrés.

Avant d'interroger le système, un client DAS doit tout d'abord obtenir auprès de l'annuaire la liste des différents serveurs. Une fois la liste obtenue, le client peut interroger les serveurs qu'il désire pour rechercher les annotations d'une séquence. Les spécifications du protocole sont disponibles à l'adresse http://www.biodas.org/.

Plusieurs systèmes proposent d'utiliser le protocole DAS : le visionneur d'alignement Jalview (Waterhouse *et al.*, 2009), le navigateur de génomes Ensembl, ou la librairie de programmation BioJava (http://biojava.org). Dasty2 (Jimenez *et al.*, 2008) est un client DAS en ligne permettant de visualiser les annotations d'une séquence de manière interactive (Figure 17).

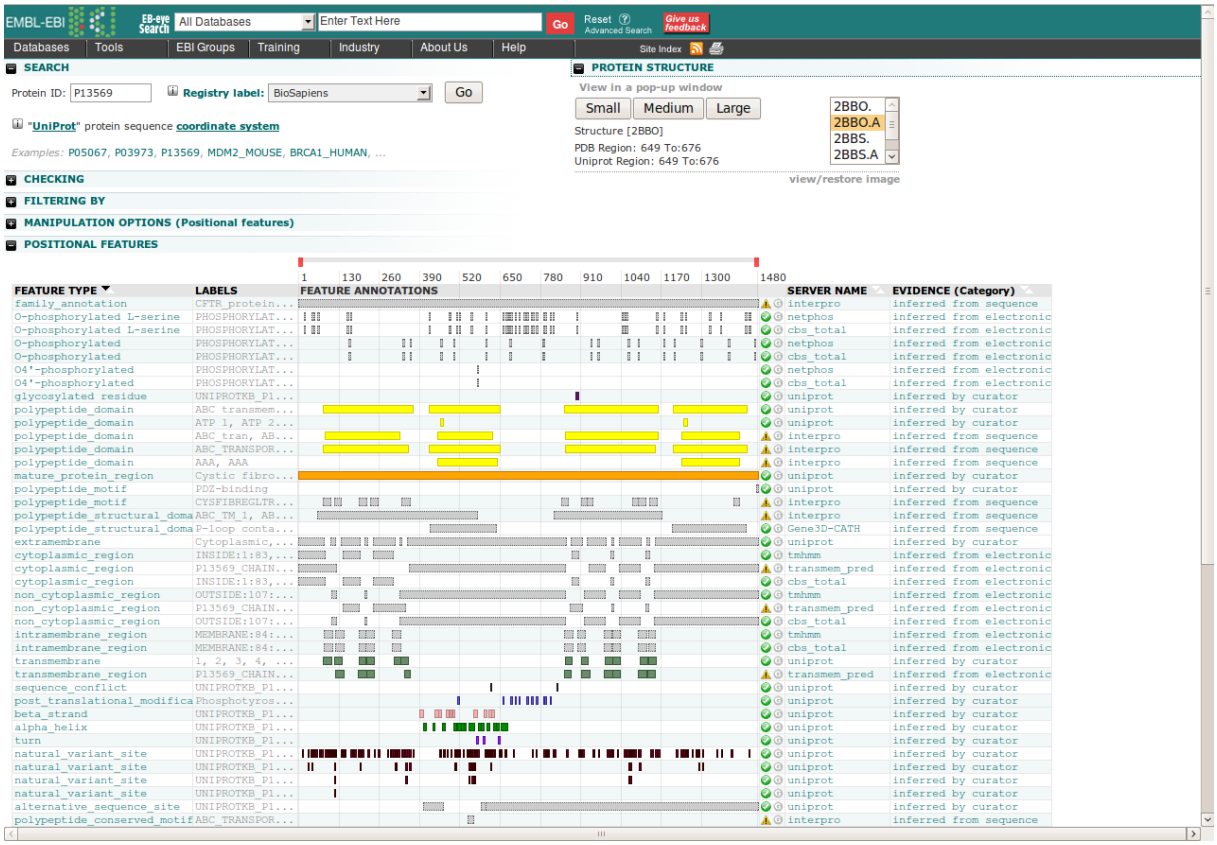


Figure 17 – Visualisation des annotations de la protéine P13569 sous le client DAS Dasty2
Chaque piste représente un couple de type d'annotations et d'un serveur correspondant. Les annotations affichées peuvent être sélectionnées en déroulant le menu '*Filtering by*'. L'affichage des pistes peut être personnalisé à l'aide du menu déroulant 'Manipulation options'. En bas de l'interface (non représentées dans cette figure), apparaissent la table des annotations globales (qui concernent toute la séquence), ainsi que la séquence en elle-même.

Le système DAS permet donc de récupérer rapidement les informations d'annotations d'une ou plusieurs séquences sans avoir à télécharger une grande quantité de données. Il permet aussi aux laboratoires qui ont déployé un serveur DAS, de proposer leurs propres annotations pour des séquences déjà annotées, permettant ainsi la comparaison des annotations avec celles d'autres groupes.

3.3.3 Standardisation des protocoles opératoires et des métadonnées

Dans beaucoup de domaines expérimentaux, les différents travaux sont régis par des protocoles opératoires et des conditions bien définis (SOP, *Standard Operating Protocol*). Ces SOP permettent d'obtenir des résultats reproductibles, et assurent un suivi de ce qui doit être fait. Malheureusement, dans le domaine de l'annotation, il est souvent difficile de distinguer ce que les différents groupes d'annotations réalisent, et il existe peu de ressources décrivant les bonnes pratiques d'annotation, obligeant les nouveaux venus à réinventer la roue continuellement (Angiuoli *et al.*, 2008).

Pour essayer d'endiguer en partie ce phénomène, le Consortium GSC tente de promouvoir l'utilisation de standard d'informations minimales (MI, *Minimum Information Standards*) dans le domaine de la génomique. Ces différents MI se présentent sous la forme d'une liste de contrôle, indiquant les informations expérimentales essentielles ou optionnelles à fournir, lorsqu'on désire publier des résultats.

Pour mettre en avant ces standards, le GSC a cofondé le portail MIBBI (*Minimum Information for Biological and Biomedical Investigations*) (http://www.mibbi.org/) qui recense tous les MI des biosciences. Parmi ceux-ci, le standard MIAME (*Minimum Information About a Microarray Experiment*) fait figure d'exemple puisque une grande partie des journaux scientifiques requièrent que les expériences basées sur des données de puce à ADN remplissent les critères énoncés par ce standard avant toute publication (Taylor *et al.*, 2008).

Le *Minimum Information about a (Meta)Genome Sequence* (MIGS/MIMS) (Field, Garrity, Gray, *et al.*, 2008) est le standard MI publié par le GSC pour réaliser un séquençage de génome. Ce standard comporte pour l'instant 4 points principaux (Annexe 4) :

- La soumission des chromatogrammes à une banque publique,
- La spécification du type d'organisme séquencé (eucaryote, bactérie…),
- Les informations du projet qui comprennent le nom du projet, des informations sur l'habitat de l'organisme et sur la séquence nucléique séquencée,
- Les informations sur le séquençage en lui même.

Ce standard est relativement court car le GSC souhaite dans un premier temps favoriser son adoption avant de l'étoffer. Bien que cela puisse être étonnant, ces informations pourtant capitales ne sont pas systématiquement fournies lors du dépôt d'une séquence génomique, soulignant tout le chemin restant à parcourir dans le domaine de la standardisation.

4 ALVINELLA POMPEJANA

Alvinella pompejana (Desbruyères *et al.*, 1980), appelé communément « ver de Pompéi », est un Annélide polychète marin tubicole, endémique des sources hydrothermales distribuées le long de la ride Est-Pacifique (de 21°N jusqu'à 32°S) (Hurtado *et al.*, 2004) (Figure 18). Le nom d'*Alvinella pompejana* provient du nom du submersible '*DSV Alvin'* (*Deep submergence vehicle*) utilisé lors de sa découverte en 1980, et à une analogie entre les sources hydrothermales qu'il colonise (fumeurs noirs) et l'éruption du mont Vésuve qui a détruit la cité romaine de Pompéi en 79 après JC. La colonisation de ces fumeurs noirs où règnent des conditions extrêmes (température, anoxie, pH, métaux lourds…) fait d'*A. pompejana* l'un des métazoaires les plus thermotolérants connus à ce jour (Cary *et al.*, 1998).

Figure 18 – Carte indiquant la ride Est-Pacifique et le Rift des Galápagos
Les noms des différents plateaux océaniques sont indiqués en rouge. Quelques sites de sources hydrothermales sont indiqués en noir. Adapté de http://www.divediscover.whoi.edu/hottopics/aha-pic1.html

Dans ce chapitre, nous allons présenter *Alvinella* en décrivant brièvement le phylum des Annélides et sa position au sein des métazoaires puis nous présenterons l'habitat, la morphologie et la physiologie d'*Alvinella*. Enfin, nous terminerons cette introduction par la présentation du projet *Alvinella* et de ses enjeux.

4.1 Les Annélides

4.1.1 Position phylogénétique des Annélides

Selon la phylogénie traditionnelle des Métazoaires basée sur les caractères morphologiques et embryologiques (Figure 19a), une première démarcation sépare les Spongiaires des Eumétazoaires puis les Eumétazoaires sont divisés en diploblastiques (dont les Cnidaires) et en triploblastiques ou bilatériens. Chez les bilatériens émergent successivement les acœlomates (dont les Plathelminthes), les pseudocœlomates (comprenant les Nématodes) puis les cœlomates. Les cœlomates regroupent les Protostomiens (dont les Annélides, Mollusques et Arthropodes) et les Deutérostomiens (Chordés et Echinodermes principalement).

La phylogénie moléculaire a bien confirmé le caractère monophylétique des Deutérostomes mais a abouti à des résultats divers et contradictoires pour les autres Bilatériens. La reconstruction d'arbres phylogénétiques est en effet très sensible aux taux d'évolution des branches respectives : les espèces à évolution rapide ont tendance à être repoussées vers la racine de l'arbre. Un important travail d'échantillonnage a été mené pour disposer dans chaque groupe d'organismes d'espèces à évolution lente (Aguinaldo *et al.*, 1997). Cette étude a abouti à une nouvelle phylogénie (Figure 19b) qui sépare les Deutérostomiens des Protostomiens, ces derniers regroupant aussi bien des acœlomates que des cœlomates. Les Protostomiens sont eux-mêmes divisés en Lophotrochozoaires (Annélides, Mollusques, Plathelminthes) et Ecdysozoaires (Nématodes, Arthropodes). Les premiers sont généralement caractérisés par un stade larvaire particulier, la trochophore, ou la présence d'un lophophore (couronne de tentacules ciliés) tandis que les seconds se développent en effectuant une ou plusieurs mues cuticulaires.

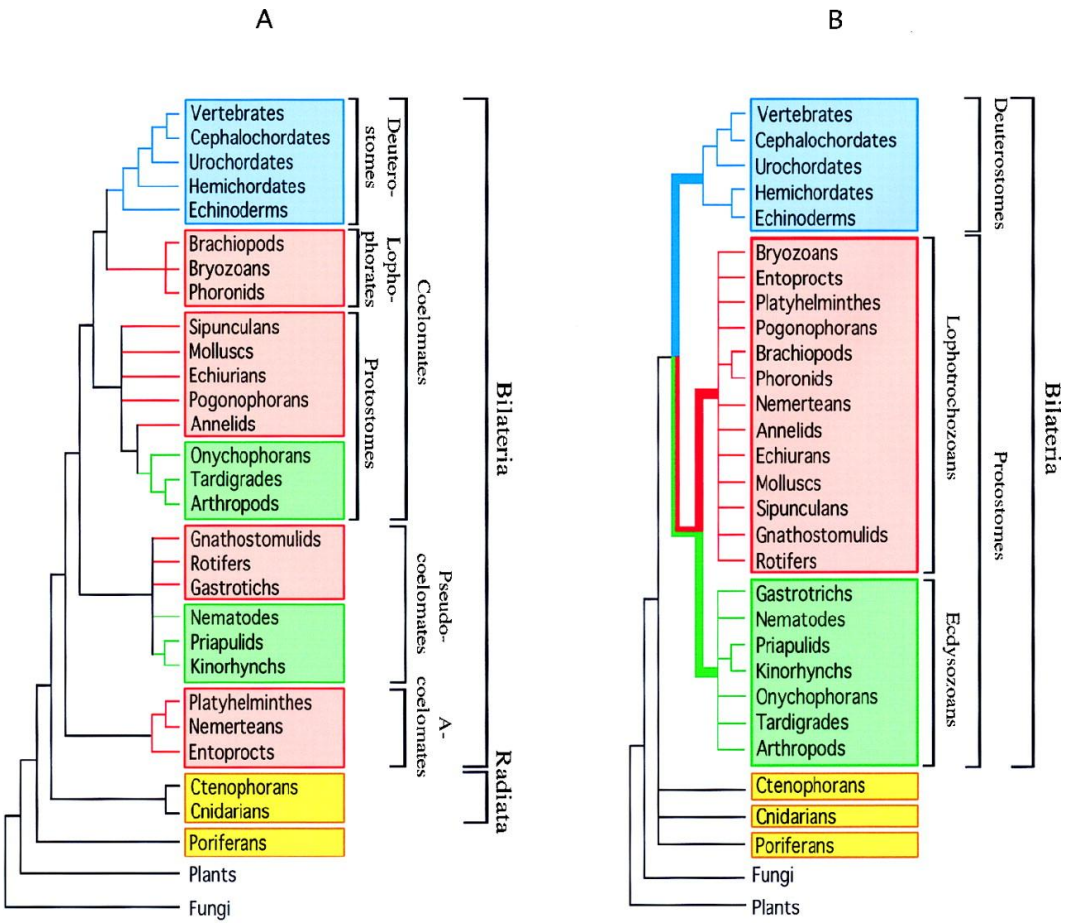


Figure 19 – Phylogénies des métazoaires
(A) Phylogénie traditionnelle basée sur l'observation de la morphologie et les caractères phénotypiques. (B) Nouvelle phylogénie moléculaire. Extrait de (Adoutte *et al.*, 2000)

Dans le contexte de cette nouvelle classification, les Lophotrochozoaires font figure de parents pauvres à l'ère post-génomique, avec environ 113 000 séquences protéiques (7 000 pour les Annélides), contre plus d'un million chez les Ecdysozoaires. Il n'existe pas de génomes complets de Lophotrochozoaires alors que les premières séquences génomiques des Ecdysozoaires (*Caenorhabditis elegans*, *Drosophila melanogaster*) sont disponibles depuis plusieurs années et continuent de se multiplier. L'ensemble de séquences protéiques le plus complet chez les Lophotrochozoaires est, à l'heure actuelle, celui du Plathelminthe parasite *Schistosoma japonicum*.

4.1.2 Principales caractéristiques des Annélides

Les Annélides sont des vers segmentés. Le premier segment antérieur appelé prostomium se distingue des autres segments par la présence du cerveau et des organes sensoriels, alors que le segment suivant, le péristomium contient la bouche de l'animal. Le pygidium, qui est le dernier segment postérieur est précédé de la zone de croissance d'où naissent les nouveaux segments. Outre leur corps segmenté, les Annélides se distinguent aussi des

autres phylums de vers (Nématodes et Plathelminthes) par la présence d'un véritable cœlome.

Au sein des Annélides, on distingue trois groupes principaux : les Polychètes essentiellement marins dont les segments centraux sont recouverts de soies, les Oligochètes (généralement limicoles ou terricoles comme le lombric) qui portent quelques soies et les Achètes dépourvus de soies que l'on retrouve surtout en eau douce et qui comprennent, par exemple, les sangsues. *A. pompejana* appartient, comme la néréis ou l'arénicole, à la classe des Polychètes, de loin la plus étendue.

4.2 Le ver de Pompéi

4.2.1 Habitat et biotope des sources hydrothermales

Les sources hydrothermales, ou fumeurs, sont des évents situés près des dorsales océaniques, d'où s'échappent des fluides surchauffés par le magma du noyau terrestre. Ces fumeurs sont la résultante de la tectonique des plaques, qui par des phénomènes d'accrétion et d'extension, génère des fissures dans la croûte océanique par où s'infiltre l'eau de mer sous l'effet de la pression. Surchauffée à environ 400°C par son approche du magma du noyau terrestre, et chargée en gaz acidifiants (H_2S, CH_4, CO, CO_2, H_2) et en métaux (Si, Mn, Fe, Zn) par son contact avec les roches basaltiques, cette eau remonte en surface de la croûte océanique. Au contact de l'eau glaciale du plancher océanique, les sulfures métalliques précipitent en formant des cheminées pouvant atteindre plusieurs dizaines de mètres de hauteur (Figure 20). Les fluides continuant d'être expulsés, ils entretiennent ensuite la croissance des cheminées et maintiennent un gradient chimique et un gradient de température plus ou moins variables en périphérie de ces cheminées.

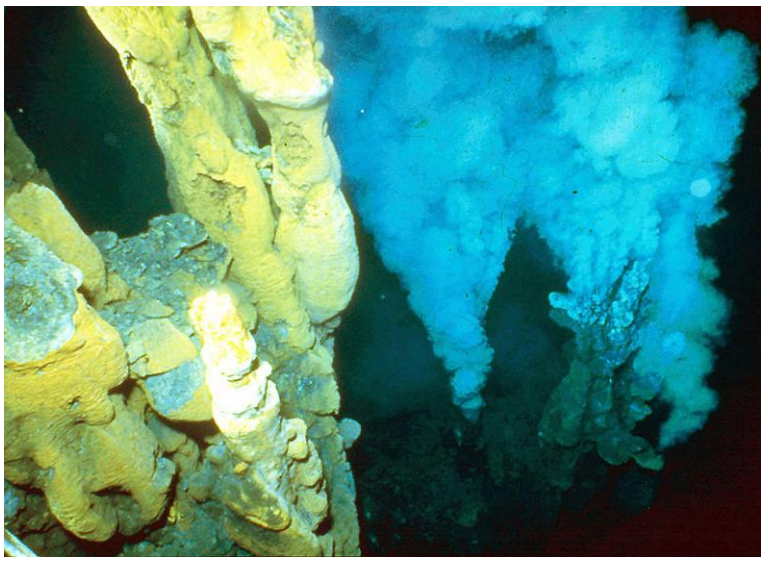


Figure 20 – Monts hydrothermaux laissant s'échapper des précipités métalliques
Source http://www.science.psu.edu/alert/Fisher9-2004.htm

C'est en 1977, le long du rift des Galápagos, à une profondeur de 2 500m, que furent explorés pour la première fois les monts hydrothermaux (Corliss et Ballard, 1977). Ce milieu était précédemment qualifié d'abiotique par référence aux conditions environnementales extrêmes y régnant, mais cette première exploration a permis de mettre en lumière tout un oasis de vie évoluant dans un écosystème périphérique aux fumeurs.

En absence totale de lumière, cet écosystème est organisé en premier lieu autour d'une communauté de bactéries chimiosynthétiques qui convertissent les composés chimiques évacués dans les « fumées » des sources hydrothermales en énergie et en composés organiques (Reysenbach, 2001). Ces bactéries forment ainsi le premier maillon de la chaîne trophique. Le second maillon est constitué des consommateurs primaires. Il consiste en une importante communauté composée de seulement quelques espèces animales portant des symbiotes nourriciers bactériens, telles que les vestimentifères *Riftia pachyptila* et *Ridgeia piscesae*, le gastéropode *Alviniconcha hessleri*, et les mollusques bivalves *Bathymodiolus puteoserpentis*, ou des animaux portant des épibiotes telles que *Alvinella pompejana* ou la crevette *Rimicaris exoculata* (Jollivet, 1996). Plusieurs communautés de consommateurs secondaires (poissons, crustacés, pieuvres…) terminent la chaîne trophique en se nourrissant de ces consommateurs primaires. En parallèle de cette chaîne, plusieurs espèces d'animaux filtreurs (éponges, anémones…) se développent grâce aux retombées des sources hydrothermales.

La présence des sources hydrothermales permet donc de maintenir en vie tout un écosystème très diversifié, au milieu d'un vaste désert d'eau glacée.

4.2.2 Morphologie et physiologie d'*Alvinella pompejana*

Alvinella pompejana fait partie de la famille des Alvinellidés (Desbruyères et Laubier, 1986) qui comporte deux genres : le genre *Alvinella*, composé de 2 espèces (*caudata* et *pompejana*) et le genre *Paralvinella* composé de 9 espèces (*bactericola*, *dela*, *fijiensis*, *grasslei*, *hessleri*, *palmiformis*, *pandorae*, *sulfincola* et *unidentata*).

Les Alvinellidés sont des polychètes tubicoles qui forment des massifs très denses le long des parois des cheminées hydrothermales (Figure 21a). Ils disposent d'un système circulatoire très évolué où les échanges gazeux sont assurés par quatre paires de branchies dorsales saturées en mitochondries (Figure 21b), et où le transport des gaz est effectué par des hémoglobines extracellulaires et par les érythrocytes du liquide cœlomique (Desbruyères *et al.*, 1998).

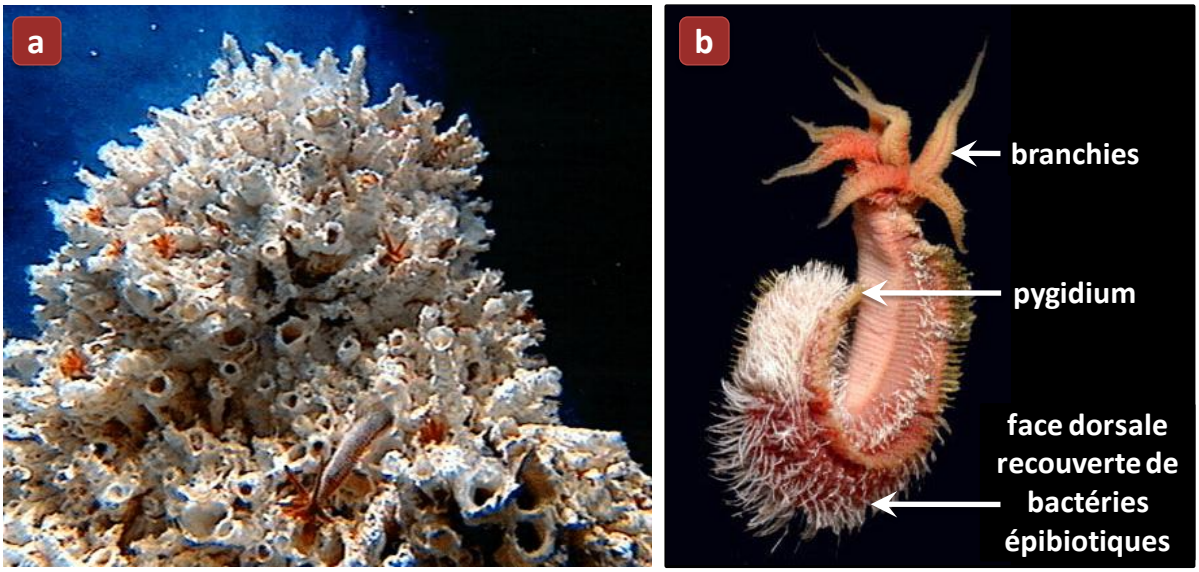


Figure 21 – Colonie d'*A. pompejana* et individu isolé
(a) Colonie d'*A. pompejana* ayant formé des tubes le long d'une cheminée hydrothermale : on peut observer dans le bas de l'image un individu sortant de son tube. (b) Un individu isolé. Images Ifremer.

À l'âge adulte, *A. pompejana* comporte de 80 à 100 courts segments pour atteindre une longueur d'une dizaine de centimètres pour un diamètre de 12 mm. Comme tous les Alvinellidés, le mode de nutrition consiste à manger les dépôts bactériens. Cependant, le genre *Alvinella* nécessite une symbiose obligatoire avec des bactéries épibiotiques localisées sur la face dorsale de l'animal. La nature de cette relation symbiotique n'a pas encore été clairement définie (Campbell *et al.*, 2003).

A. pompejana fait partie des premières espèces à coloniser les cheminées nouvellement formées (Desbruyères *et al.*, 1998). Bien qu'elle puisse supporter sporadiquement des basses températures, *A. pompejana* est l'espèce du genre *Alvinella* qui vit au plus près des cheminées. Ses branchies sont orientées vers l'extérieur de son tube tandis que le pygidium est situé dans la partie du tube reposant sur la paroi du fumeur. *Alvinella* supporterait ainsi un gradient de température allant de 15°C à l'extérieur du tube, jusqu'à 80°C au plus près de la paroi de la cheminée, avec des pics de température pouvant dépasser 100°C (Cary *et al.*, 1998; Le Bris et Gaill, 2007). De telles conditions de température ont motivé plusieurs études sur la thermotolérance des protéines d'*A. pompejana* (Tableau 5).

Tableau 5 – Thermotolérance chez *A. pompejana* (d'après (Chevaldonné *et al.*, 2000))

Paramètres mesurés	*Température maximum*
Respiration mitochondriale	49°C
Dissociation de l'hémoglobine	50°C
Cinétique de la malate dehydrogenases cytosolique	31°C
Stabilité thermique de l'aspartate-amino transferase	61°C
Stabilité thermique de la glucose-6-phosphate isomerase	52°C
Dénaturation de l'ARNr	87°C
Dénaturation du collagène cuticulaire	45°C
Dénaturation du collagène interstitiel	46°C

Plus récemment, une analyse comparative sur le facteur d'épissage U2AF65 d'*A. pompejana* a démontré un gain de stabilité de 6°C par rapport à son homologue humain (Henscheid *et al.*, 2005).

4.3 Projet et Consortium *Alvinella*

La thermostabilité des protéines d'*A. pompejana* a suscité l'intérêt du LBGI et, plus largement, du Département de Biologie et Génomique Structurales. Un gain de thermostabilité, même modeste, peut être décisif pour l'obtention de structures tridimensionnelles comme l'ont démontré les travaux réalisés tout récemment sur la superoxide dismutase d'*Alvinella* (Shin *et al.*, 2009). Ainsi, afin de disposer d'une source de protéines et de complexes protéiques thermostables d'origine animale et d'explorer le phylum des Annélides encore peu représenté dans les banques de séquences, un projet de séquençage d'ADNc d'*Alvinella pompejana* a été lancé à l'initiative du LBGI. Ce projet a débuté par la création d'un Consortium Européen regroupant 12 équipes (Annexe 5). À la suite de cette création, la proposition du projet de construction de banques d'ADNc et de séquençage de 200 000 lectures a été acceptée par le Genoscope – CNS (http://www.genoscope.cns.fr/) en 2004. En 2005, *A. pompejana* a été retenu comme eucaryote modèle dans le projet Européen SPINE2-Complexes (*Structural Proteomics in Europe 2 – Complexes*, http://www.spine2.eu/), dont le but est l'obtention massive de structures de complexes protéine-protéine ou protéine-nucléique par des techniques haut-débit.

Il est à noter qu'un projet de séquençage équivalent a été initié par S. Craig Cary (Université du Delaware), qui a obtenu de la part du DOE Joint Genome Institute le séquençage de 140 000 clones d'*Alvinella pompejana* en 2005 (http://www.jgi.doe.gov/sequencing/why/3135.html). Les séquences des ESTs ont été déposées sans annotation dans les banques publiques début 2009.

4.3.1 Matériel et banques disponibles

La construction des banques d'ADNc et le séquençage ont été réalisés au Genoscope – CNS. Les échantillons d'individus d'*Alvinella* ont été collectés spécialement durant la campagne océanographique BioSpeedo (http://www.sc.ucl.ac.be/nemo/Biospeedo.htm) qui s'est déroulée le long de la ride Est-Pacifique en 2004.

À l'heure actuelle, 5 banques ont été construites et séquencées :

- La première banque a été réalisée à partir de plusieurs individus entiers.
- Trois banques tissu-spécifiques ont été réalisées à partir de branchies, du tissu ventral, et de pygidium provenant de plusieurs adultes. La banque pygidium comporte le pygidium proprement dit (le dernier segment postérieur de l'animal) mais aussi la zone de croissance. Par la suite, cette banque sera simplement appelée « pygidium ».
- Une dernière banque composée d'un seul individu entier. Cette banque n'a pas été traitée dans le cadre de cette thèse mais son analyse est envisagée, notamment pour étudier le polymorphisme interindividuel (SNP, Single Nucleotide Polymorphism) à l'aide de logiciel tels que PolyBayes (Marth *et al.*, 1999) ou autoSNP (BARKER *et al.*, 2003).

Un total d'environ 120 000 séquences a été pour l'instant obtenu à partir de ces différentes banques. La construction d'une ou plusieurs banques normalisées était prévue afin d'élargir l'éventail des transcrits mais les différentes tentatives conduites au Genoscope se sont avérées infructueuses.

4.3.2 Construction des banques

La construction des banques d'ADNc pleine longueur a été réalisée préférentiellement grâce à la technique de clonage Oligo-Capping (Maruyama et Sugano, 1994). Cependant des rendements trop faibles de clonage ont été obtenus avec les banques d'individus entiers et de pygidium. Ces banques ont donc finalement été clonées à l'aide du kit commercial CloneMiner™ (http://www.invitrogen.com/).

4.3.2.1 Technique Oligo-Capping

Cette technique clone préférentiellement les ARNm de pleine longueur en les sélectionnant par leur coiffe en 5' et leur queue poly(A) en 3'.

La coiffe des ARNm est modifiée en trois étapes. L'action de la phosphatase alcaline bactérienne (BAP) permet d'hydrolyser en 5' les phosphates des ARNm dégradés. Ensuite, le pyrophosphatase acide du tabac (TAP) hydrolyse la structure de la coiffe en laissant le

phosphate en 5'. Enfin, l'ARN ligase du bactériophage T4 fixe spécifiquement l'oligonucléotide de clonage en 5' grâce au phosphate restant.

Le reste du clonage est réalisé par RT-PCR (PCR avec rétrotranscription) à l'aide d'oligonucléotides spécifiques de l'oligonucléotide en 5' et d'oligonucléotides poly(T) pour sélectionner la queue poly(A). Les extrémités des fragments amplifiés sont ensuite digérées par l'enzyme de restriction SfiI, puis les fragments sont intégrés au vecteur pME18s-FL3.

4.3.2.2 Technique CloneMiner

Ce kit permet la construction de librairies d'ADNc par recombinaison (système Gateway®, Invitrogen). Les ARNm sont purifiés par leur queue poly(A) après RT-PCR à l'aide d'une amorce poly(T)-attB2-biotine et d'un système biotine-streptavidine. Suite à cette purification, un adaptateur attB1 est fixé en 5' des fragments amplifiés. Le système Gateway permet ensuite de recombiner ces fragments dans le vecteur d'entrée pDONR222 à l'aide des sites attB1 et attB2 fixés aux fragments, et des sites attP1 et attP2 du vecteur d'entrée.

MATÉRIEL ET MÉTHODES

Les travaux présentés dans ce manuscrit s'inscrivent dans le cadre général du traitement de données à haut-débit, dont l'information est enrichie et mise à jour quotidiennement. Dans ce contexte, l'analyse bioinformatique des données nécessite de faire appel à des banques de données régulièrement mises à jour, d'automatiser les traitements à l'aide d'un ensemble de logiciels génériques ou spécialisés, ainsi que d'organiser efficacement les résultats obtenus.

Ces différents aspects vont faire l'object des chapitres suivants, où nous allons décrire l'ensemble des ressources qui ont été mises en œuvre afin de mener à bien nos travaux.

Après un bref survol des ressources et équipement informatiques dont nous disposions (chapitre 5), nous présenterons les banques de données biologiques généralistes ou spécialisées qui ont été utilisées, ainsi que les outils permettant de les interroger efficacement (chapitre 6). Ensuite, les langages et outils purement informatiques exploités dans l'élaboration des programmes présentés dans ce manuscrit seront décrits brièvement (chapitre 7). Au cours du chapitre 8, les outils bioinformatiques utilisés seront présentés en deux parties : les outils permettant de traiter les données de séquençage, et ceux permettant de réaliser des annotations de séquences. Enfin, nous terminerons cette partie en présentant Gscope (R. Ripp, manuscrit en préparation), la plate-forme de génomique déployée au LBGI, qui a été exploitée quotidiennement pour réaliser les tâches bioinformatiques « classiques » et pour automatiser les traitements de données de séquençage et d'annotation décrits dans ce manuscrit (chapitre 9).

5 RESSOURCES ET ÉQUIPEMENTS INFORMATIQUES

La gestion et l'analyse de telles quantités de données nécessitent une puissance de calcul et une capacité de stockage adaptées. Au cours de nos travaux, nous avons pu profiter de l'infrastructure existante au LBGI, et des services proposés par la Plate-forme de BioInformatique de Strasbourg (BIPS). Nous allons décrire ces différentes ressources ainsi qu'« Alnitak », le serveur de base de données et de sites Web du LBGI dont j'ai eu partiellement la charge. Les banques de séquences seront décrites dans le chapitre suivant.

5.1 La Plate-forme de BioInformatique de Strasbourg

BIPS (*BioInformatics Platform of Strasbourg*) (http://bips.u-strasbg.fr/) est la plate-forme à haut-débit pour la génomique comparative et structurale de l'IGBMC. Elle appartient au Réseau National des plates-formes Bioinformatiques (ReNaBi). Elle a été identifiée plate-forme nationale RIO (Réseau Inter-Organismes) en 2003 et fait partie du Génopôle Grand-Est « du Gène au Médicament ». En 2005, elle devient la première plate-forme française de bioinformatique à être certifiée norme ISO9001:2000.

Son rôle consiste à déployer, maintenir et mettre à disposition tout un ensemble de ressources bioinformatiques matérielles et logicielles, ainsi qu'à mettre en œuvre et partager son expertise et ses compétences pour développer des solutions innovantes et diffuser son savoir-faire. C'est donc elle qui, entre autres, a la charge de maintenir à jour les logiciels et les banques de données biologiques sur les serveurs de calculs de l'institut.

Ces serveurs, dédiés aux calculs intensifs, forment une grappe de calcul dénommée « Star », et qui est composée de huit nœuds d'architecture x64 Opteron™ baptisés « star1 » à « star8 ». Deux nœuds sont des *Sun Fire V40z* quadri-processeurs dotés de 32 Gio de mémoire et pilotés par le système d'exploitation *Solaris 10*, et sont en partie dédiés à la gestion de la grappe et au partage des disques de stockage *via* le protocole NFS (*Network File System*). Les 6 autres nœuds sont pilotés par le système d'exploitation *Red Hat Enterprise Linux ES 5* et sont constitués de 3 autres *Sun Fire V40z* quadri-processeurs avec 16 Gio de mémoire, 2 *Sun Fire X4100 M2* bi-processeurs bi-cœurs et d'un *Supermicro H8DMU+* bi-processeurs quadri-cœurs. Chacun de ces 6 nœuds est doté de 16 Gio de mémoire à l'exception du dernier qui en comporte 32 Gio.

Le serveur de disques dispose quant à lui d'une capacité de 12 To de stockage RAID5 (*Redundant Array of Independent Disks*) et est architecturé autour d'un *Sun Fire V480* sous *Solaris 9*.

5.2 Alnitak : serveur de base de données et de sites Web du LBGI

Afin d'héberger dans de bonnes conditions de rapidité et de réactivité les différents sites Web et bases de données relationnelles du LBGI, le laboratoire a récemment fait l'acquisition d'un serveur dédié supplémentaire dénommé « alnitak » dont j'ai partiellement assuré l'administration. Il s'agit d'un *Sun Fire X4200 M2* bi-processeurs bi-cœurs doté de 8 Gio de mémoire et fonctionnant sous système d'exploitation *Ubuntu 8.04 LTS*. Cette machine accède directement par protocole iSCSI (*Internet Small Computer System Interface*) à un serveur de disques dédié disposant de 2 To de stockage RAID1.

6 BANQUES DE DONNÉES BIOLOGIQUES

Les banques de données biologiques sont l'outil de base du bioinformaticien pour réaliser son travail d'analyse. Dans cette section, nous allons décrire précisément les principales banques de données utilisées au cours de nos travaux de thèse.

6.1 Banques de données généralistes

C'est sous ce terme que l'on désigne les banques de référence qui regroupent l'ensemble des données de séquences et de structures.

6.1.1 GenBank

GenBank (Benson *et al.*, 2009) a été créée au début des années 1980 au *Los Alamos National Laboratory*. Cette banque est maintenant maintenue et distribuée par le NCBI (*National Center for Biotechnology Information*) et regroupe l'intégralité des séquences nucléotidiques publiquement disponibles. Elle fait partie de l'*International Nucleotide Sequence Database Collaboration* qui vise à unifier les données par un échange quotidien entre les trois banques nucléotidiques majeures que sont GenBank, l'EMBL Nucleotide Sequence Database et la DDBJ (*DNA Data Bank of Japan*). Cette structure d'échange, ainsi que le libre accès à la soumission de séquences, ont permis une croissance quasi-exponentielle du nombre de séquences dans la banque (Figure 22).

Cependant, de cette exhaustivité résulte une grande redondance de séquences dans la banque. Pour palier à ce problème, la banque de données RefSeq (Pruitt *et al.*, 2007) a été développée par le NCBI. L'objectif est d'élaborer un jeu non-redondant de séquences nucléotidiques et protéiques composé de séquences représentatives choisies parmi des groupes de séquences hautement similaires de la banque GenBank.

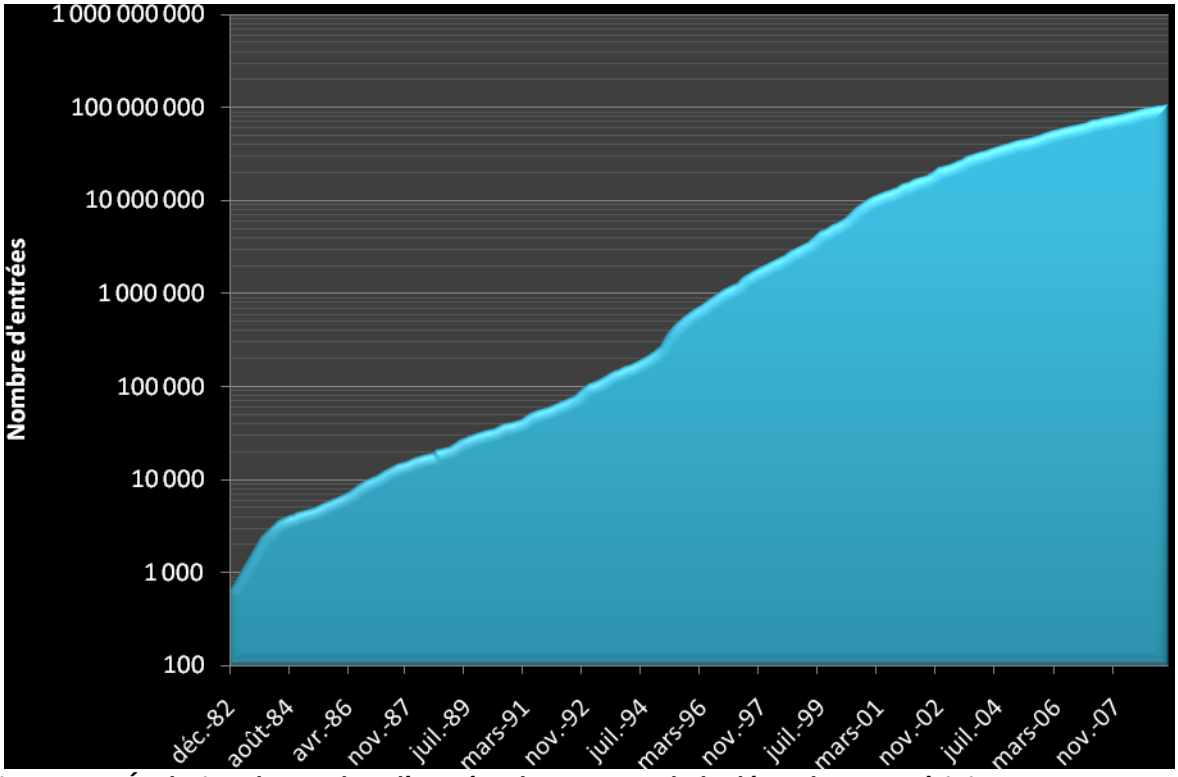


Figure 22 – Évolution du nombre d'entrées dans GenBank de décembre 1982 à juin 2009.
Données extraites de **ftp://ftp.ncbi.nih.gov/genbank/gbrel.txt** (GenBank version 172).

Initialement, GenBank fut subdivisée en fichiers correspondant plus ou moins à des groupes taxonomiques. Cependant, face à l'afflux de données fortement hétérogènes, de nouvelles divisions basées sur la nature des données ont été créées récemment (Tableau 6).

Tableau 6 – Divisions GenBank actuellement existantes

Divisions taxonomiques	***Description***
BCT	Procaryotes
ENV	Échantillons environnementaux/métagénomes
INV	Invertébrés
MAM	Autres mammifères
PHG	Phages
PLN	Plantes
PRI	Primates
ROD	Rongeurs
SYN	Séquences synthétiques/chimériques
UNA	Séquences inconnues
VRL	Virus
VRT	Autres vertébrés
Divisions fonctionnelles	
CON	« constructions »
EST	*Expressed Sequence Tag*
GSS	*Genome Survey Sequence*
HTC	*High Throughput cDNA sequencing*
HTG	HTGS (*High Throughput Genomic Sequencing*)
PAT	Séquences brevetées
STS	*Sequence Tagged Site*
TPA	Annotation par une tierce personne
TSA	*Transcriptome Shotgun Assembly*
WGS	*Whole Genome Shotgun*

6.1.2 UniProt

UniProt (*Universal Protein resource*) (The UniProt Consortium, 2009) est une collaboration initiée en 2002 entre l'*European Bioinformatics Institute* (EBI), le *Swiss Institute of Bioinformatics* (SIB) et le *Protein Information Resource* (PIR). Cette ressource constitue actuellement le catalogue de séquences et d'annotations protéiques le plus complet.

UniProt comporte quatre composants principaux :

- UniProtKB (*UniProt Knowledgebase*) est la banque de séquences et d'annotations protéiques maintenue et distribuée par le consortium UniProt.
- UniRef (*UniProt Reference clusters*) rassemble en groupes de 100%, 90% et 50% de similarité de séquence les séquences d'un même organisme provenant d'UniProtKB. Cette banque permet d'accélérer et affiner les recherches de similarité en masquant une grande partie de la redondance.

- UniMES (*UniProt Metagenomic and Environmental Sequences*) est une banque spécialement dédiée aux données de métagénomique et aux séquences environnementales.
- UniParc (*UniProt Archive*) rassemble en entrées uniques les séquences protéiques d'un grand nombre de banques publiques de séquences (UniProtKB, Ensembl, PDB, RefSeq, …) étant 100% identiques, quel que soit l'organisme, et conserve l'intégralité de l'historique des numéros d'accès vers ces différentes banques publiques.

UniProtKB constitue la réunification en une seule ressource des banques Swiss-Prot et TrEMBL (Boeckmann *et al.*, 2003).

Swiss-prot a été développée et maintenue conjointement par le SIB et l'EBI de 1986 à 2003. Cette banque de données se veut la référence en matière de séquences protéiques annotées. Elle présente une particularité évidente parmi les banques de séquences généralistes. En effet, en privilégiant la qualité et la richesse des annotations par rapport à l'exhaustivité de sa collection de séquences, elle offre une redondance minimale, les différentes versions d'une même entrée étant fusionnées. Pour ce faire, l'entrée de chaque protéine 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 intégrées à cette banque proviennent essentiellement de la traduction des gènes annotés de la banque *EMBL Nucleotide Sequence Database* (Cochrane *et al.*, 2009b), mais proviennent aussi d'autres banques de séquences protéiques, de la littérature scientifique, ainsi que de soumissions directes. Malgré tous les bénéfices que peuvent procurer cette politique de développement, le flot grandissant de nouvelles séquences issues de projets de séquençage de génomes joue en la défaveur de Swiss-Prot. En effet, il en résulte une progression lente du nombre d'entrées, même si l'on observe une accélération du rythme de croissance courant 2005 (Figure 23).

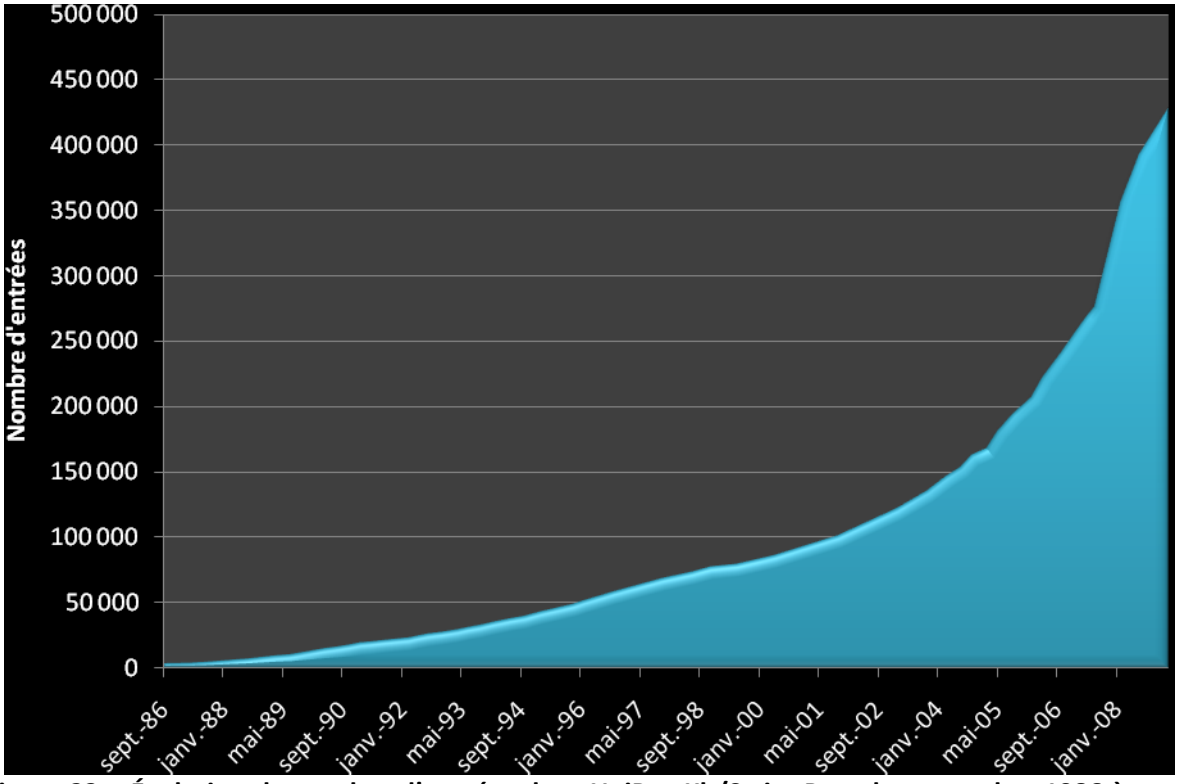


Figure 23 – Évolution du nombre d'entrées dans UniProtKb/Swiss-Prot de septembre 1986 à mars 2009.
Données extraites de ftp://ftp.uniprot.org/pub/databases/uniprot/current_release/knowledgebase/complete/docs/relnotes.htm (UniprotKb version 15.0).

Introduite en 1996 pour complémenter Swiss-Prot, TrEMBL (*Translated EMBL*) consiste en la traduction de toutes les parties codantes des séquences nucléotidiques issues de la banque EMBL. TrEMBL constitue ainsi un dépôt exhaustif, mais notablement redondant, de séquences non validées et faiblement annotées. Ces séquences sont ensuite destinées à être examinées par les annotateurs de Swiss-Prot pour y être intégrées. La croissance de TrEMBL est littéralement exponentielle et correspond aux flux de données générés par les projets de séquençage de génomes (Figure 24).

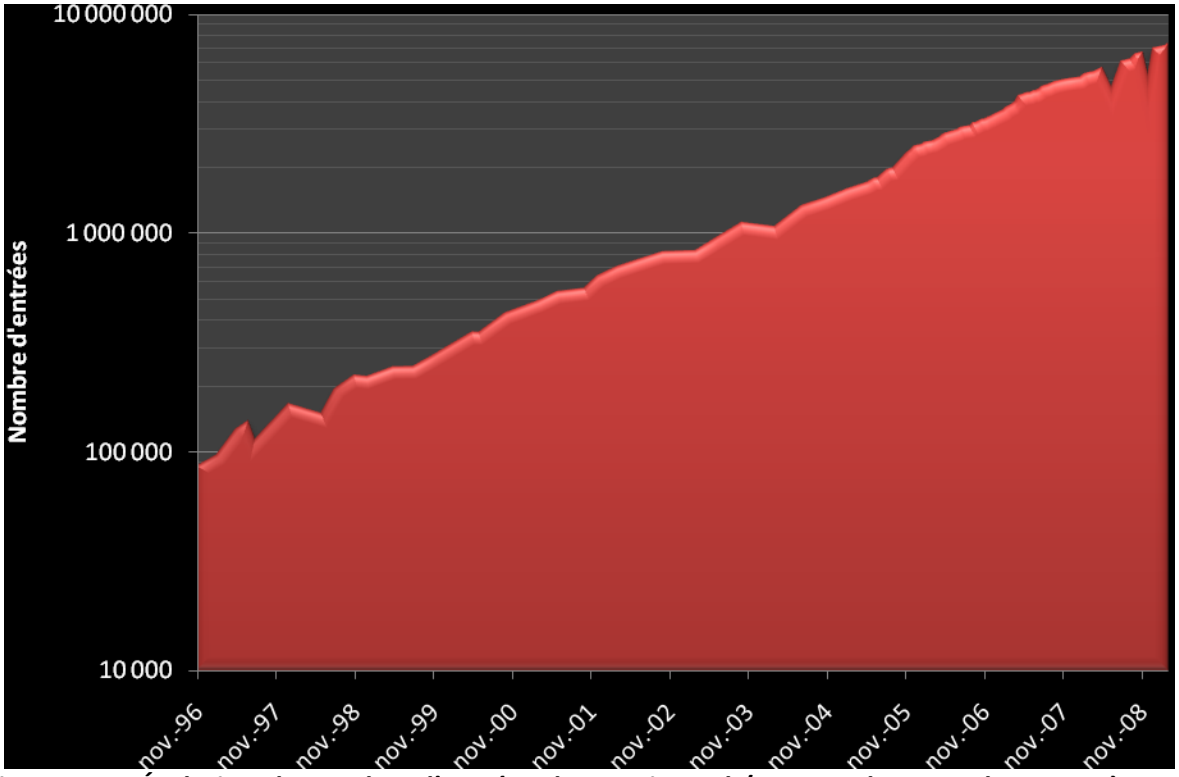


Figure 24 – Évolution du nombre d'entrées dans UniProtKb/TrEMBL de novembre 1996 à mars 2009.
Données compilées à partir de http://www.ebi.ac.uk/Information/News/news.html

6.1.3 PDB

La PDB (*Protein Data Bank*) (Berman *et al.*, 2000) est la principale banque internationale de structures tridimensionnelles de macromolécules biologiques. Cette banque a été établie dès 1971 au *Brookhaven National Laboratory* (BNL) à partir de 7 structures. Depuis 1998, elle a été transférée au *Research Collaboratory for Structural Bioinformatics* (RCSB). En 2003, l'organisation wwPDB (*Worldwide PDB*), constituée du RCSB, du groupe EBI-PDBe (*PDB in Europe*), de la PDBj (*PDB of Japan*) et de la BMRB (*Biological Magnetic Resonance Data Bank*), a été fondée pour superviser la distribution de la PDB.

À date du 30 juin 2009, la PDB comporte un total de 58 588 structures très majoritairement résolues par cristallographie aux rayons X (86%), mais on y retrouve aussi des structures obtenues par Résonnance Magnétique Nucléaire (RMN), par microscopie électronique ou d'autres méthodes hybrides. Les structures protéiques constituent l'essentiel des entrées (92%), le reste se partageant entre les structures d'acides nucléiques et les complexes protéine-acide nucléique.

En plus des coordonnées atomiques, des références croisées de banques de données et des références bibliographiques, chaque entrée comporte les structures primaires et

secondaires des molécules considérées, ainsi que les détails des expériences (conditions de cristallisation, collecte des données, résolution de structure…).

Bien que le nombre de structures de macromolécules biologiques soit très inférieur à celui des séquences, celui-ci croît actuellement à une vitesse comparable à celle observée pour les séquences protéiques il y a quelques années (Figure 25), notamment grâce aux projets massifs de génomique structurale de la *Protein Structure Initiative* (Berman *et al.*, 2009; Nair *et al.*, 2009)

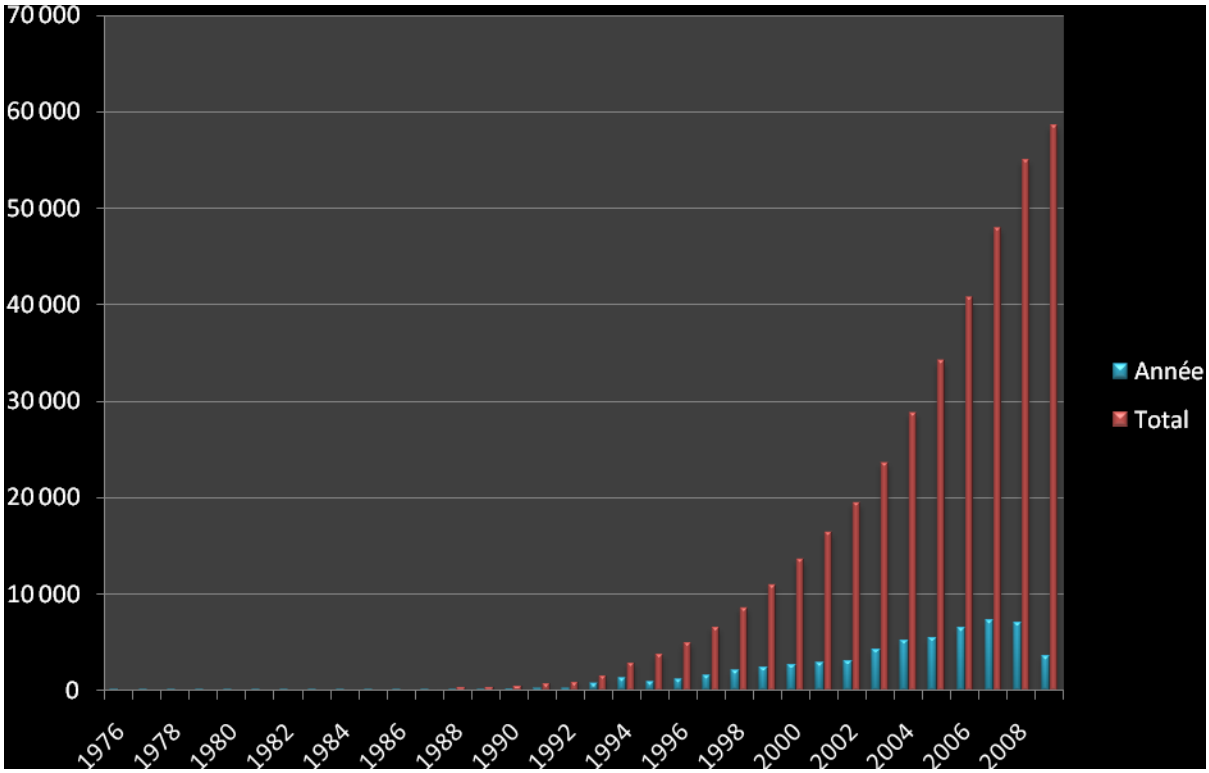


Figure 25 – Évolution du nombre d'entrées de la PDB de 1976 à juin 2009.
Les barres bleues illustrent le taux de croissance annuelle et les barres rouges la croissance cumulée. Données adaptées de http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total

On peut cependant noter une redondance importante dans la PDB, puisque plusieurs structures tridimensionnelles peuvent correspondre à la même séquence, selon les conditions d'obtention de la structure, la présence ou non de ligand, l'existence de mutations...

6.2 Banques de données spécialisées

6.2.1 Vecteurs de clonage : Univec

UniVec (Kitts *et al.*) est une banque distribuée par le NCBI qui regroupe les séquences nucléotidiques des vecteurs de clonage utilisés en biologie moléculaire. Sont aussi présentes les séquences des adaptateurs (*linkers)* et des amorces (*primers*) couramment utilisés en

techniques de clonage. Ainsi, UniVec permet, par comparaison, d'identifier et d'éliminer les séquences contaminantes d'origine vectorielle éventuellement présentes à l'intérieur des données brutes issues de séquençage automatique.

La force d'UniVec est d'avoir été optimisée pour effectuer rapidement et efficacement les recherches de séquences contaminantes (*screening*). En effet, les vecteurs de clonages étant régulièrement construits à partir des mêmes motifs (vecteurs de base, cassettes de clonage, gènes de résistance…), l'ensemble des séquences de ces vecteurs est extrêmement redondant. Dans UniVec, cette redondance est réduite grâce à l'élimination des segments de vecteurs déjà présents à l'intérieur d'autres vecteurs. La taille d'UniVec en paires de bases (pb) est ainsi réduite à 20% de la taille initiale (Tableau 7), ce qui induit une vitesse de recherche théorique 5 fois supérieure. De plus, les vecteurs de clonages étant majoritairement circulaires, les 49 premières bases de chaque séquence « linéarisée » ont été copiées en fin de séquence afin que les algorithmes de recherche par similarité ne gérant pas ce cas puissent détecter des séquences au point de jonction.

Tableau 7 – Statistiques de composition de la banque UniVec (version 5.1)
Données adaptées de http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html

	Nombre	***Taille (pb)***
Séquences représentées	1 464	4 955 915
Segments représentés	2 943	672 866
Pourcentage de redondance initiale	86%	4 283 049

6.2.2 Gene Ontology

La banque *Gene Ontology* (GO) (Ashburner *et al.*, 2000) a été créée en 1998 en tant que projet collaboratif pour normaliser les noms de produits de gènes entre trois banques de données d'organismes modèles : *FlyBase* (drosophile) (Tweedie *et al.*, 2009), *Saccharomyces Genome Database* (Hong *et al.*, 2008) et la *Mouse Genome Database* (Bult *et al.*, 2008). Depuis lors, le *GO consortium* s'est étendu à une vingtaine de banques majeures incluant des génomes de plantes, d'animaux et de microorganismes, et GO est à présent sans doute l'ontologie la plus connue et la plus utilisée en biologie.

GO est composée de trois hiérarchies de vocabulaires contrôlés englobant les termes représentant la fonction moléculaire d'un produit de gène (*molecular function*), le processus biologique dans lequel il est impliqué (*biological process*) et sa localisation cellulaire (*cellular component*). Chacune des trois hiérarchies se présente sous la forme d'un graphe orienté acyclique (DAG, *Directed Acyclic Graph*) où les termes peuvent être associés à plusieurs termes parents à l'aide de relations de type *is_a* (« est un ») ou *part_of* (« fait partie de »). Il en résulte que plus l'on descend dans la hiérarchie, plus les termes sont spécialisés et précis (Figure 26). A chaque terme est associé un numéro d'accès ainsi qu'une définition détaillée.

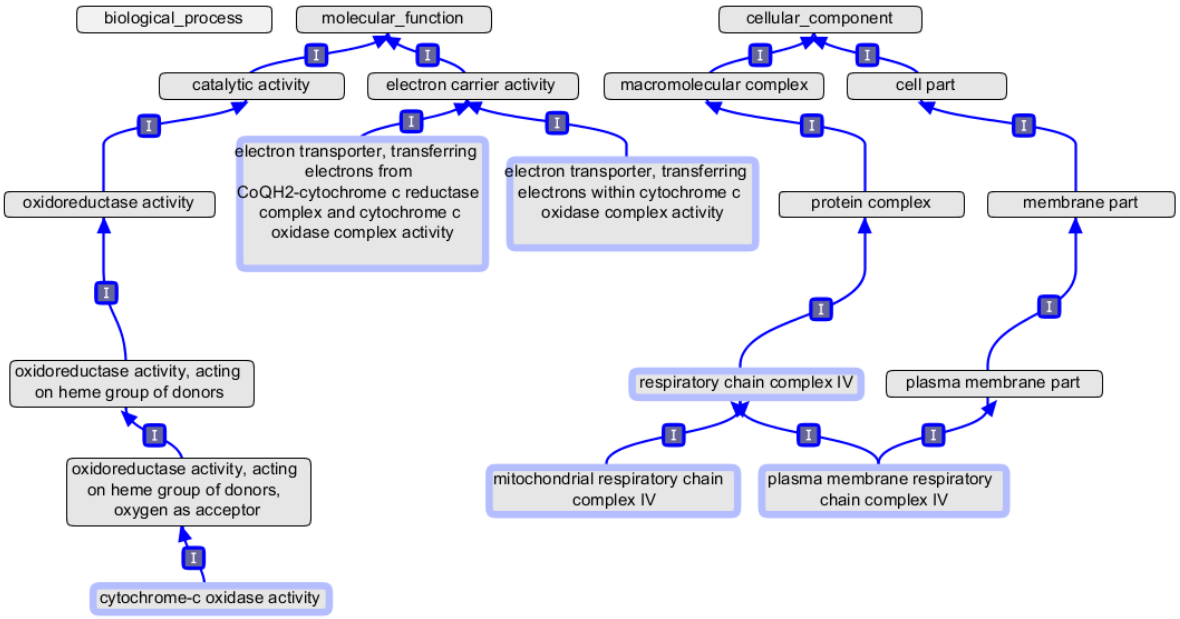


Figure 26 – Exemple de graphes des termes GO en rapport avec la cytochrome-c oxydase.
Les termes encadrés en bleu sont ceux qui ont été sélectionnés par la recherche des mots « cytochrome c oxidase » à l'intérieur du champ nom ou définition. Ici, tous les termes sont reliés à l'aide de relation de type *is_a*. Par manque de place, les termes de la hiérarchie des processus biologiques ont été masqués. Image réalisée à l'aide du logiciel OBO-Edit 2 (Day-Richter *et al.*, 2007).

La version d'avril 2009 de GO comprend un total de 27 078 termes applicables à un grand nombre d'espèces dont 8 516 termes décrivant les fonctions moléculaires, 16 216 les processus biologiques et 2 318 caractérisant les localisations cellulaires.

6.2.3 Numéros *Enzyme Commission* : ENZYME

ENZYME (Bairoch, 2000) est une banque de données distribuée par le SIB et qui décrit chaque numéro EC (*Enzyme Commission*) existant. Ces numéros EC permettent d'identifier de manière précise les enzymes et constituent donc un outil indispensable à l'annotation fonctionnelle. Ils proviennent d'une classification numérique qui a été initiée en 1955, puis publié en 1961 par l'*International Union of Biochemistry* désormais rebaptisée *International Union of Biochemistry and Molecular Biology*. A chaque numéro EC correspond un nom de classe d'enzymes caractérisée par le type de réaction enzymatique qu'elles catalysent. La sixième et dernière révision actuelle de cette norme date de 1992 et décrit 3 196 enzymes (Webb *et al.*, 1992).

Un numéro EC est composé de quatre nombres séparés par un point et est souvent précédé du préfixe « EC ». Chacun de ces nombres désigne une caractéristique de l'enzyme avec un niveau de précision croissant de gauche à droite (Tableau 8).

Tableau 8 – Décomposition du numéro EC de la glucose-6-phosphate isomérase

Numéro EC	*Niveau de précision*	*Description*
5.-.-.-	Réaction catalysée	Isomérase
5.3.-.-	Substrat générique	Oxidoreductase intramoléculaire
5.3.1.-	Substrat spécifique	Interconversion d'aldoses et de kétoses
5.3.1.9	Nom complet	glucose-6-phosphate isomérase

Chaque entrée d'ENZYME regroupe le numéro EC, le nom officiel, les noms alternatifs, la réaction catalysée, les cofacteurs ainsi que des références croisées vers la banque Swiss-Prot.

6.2.4 Profils et motifs de familles protéiques : InterPro

InterPro (Hunter *et al.*, 2009) est une banque intégrative, 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 et repose sur un consortium collaboratif entre plusieurs banques de données de signatures de familles protéiques. InterPro regroupe les banques de domaines protéiques Pfam (Finn *et al.*, 2008), ProDom (Bru *et al.*, 2005), SMART (Letunic *et al.*, 2009) et TIGRFAMs (Haft *et al.*, 2003), et les banques de motifs PROSITE (Hulo *et al.*, 2008) et PRINTS (Attwood *et al.*, 2003).

La méthode de recherche d'une signature dans une séquence protéique diffère selon chaque banque de données (motifs de type expressions régulières, profils HMM – *Hidden Markov Model*, Modèle de Markov caché –…). 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.

6.2.5 Voies métaboliques : KEGG PATHWAY

KEGG (*Kyoto Encyclopedia of Genes and Genomes*) (Kanehisa *et al.*, 2008) est une collection de banques de données initiée en 1995 par le *GenomeNet Database Service* du *Kanehisa Laboratory* à l'université de Kyoto dans le cadre de l'*Human Genome Program* lancé par le Ministère de l'éducation, la culture, des sports, des sciences et des technologies japonais. KEGG se définit comme une base de données orientée vers la biologie des systèmes permettant la compréhension des mécanismes et des concepts fonctionnels complexes de la cellule ou d'un organisme à partir de son génome.

KEGG est constituée de quatre banques principales connectées entre elles et jouant le rôle de points d'entrée pouvant aboutir aux autres sous-banques qui les composent :

- PATHWAY est une base de connaissances de voies métaboliques dessinées manuellement et de voies non-métaboliques générées automatiquement,

- BRITE est l'ontologie de tous les concepts et connaissances présents dans KEGG,
- GENES est un catalogue de gènes de plusieurs génomes complets,
- LIGAND est un catalogue de substances chimiques et de réactions qui interviennent dans le domaine de la vie.

Chacune des banques de KEGG est conçue pour être représentée sous forme de graphes dont les entrées sont les nœuds et les relations biologiques les arêtes.

PATHWAY est sans doute le point central de la banque KEGG. Ses entrées sont organisées en une hiérarchie à deux niveaux reflétant la résolution de chaque voie (Tableau 9). Pour chaque entrée de PATHWAY, il existe une voie de référence comportant tous les objets du graphe possible, ainsi qu'une voie pour chacun des organismes placés dans cette voie. Le graphe de chaque voie de PATHWAY peut être décomposé en deux sous-graphes : un graphe d'interactions protéine-protéine et un graphe de réactions enzymatiques.

Tableau 9 – Hiérarchie des voies de KEGG PATHWAY

Premier niveau	***Second niveau***
Métabolisme	Métabolisme des carbohydrates Métabolisme de l'énergie Métabolisme des lipides Métabolisme des nucléotides Métabolisme des aminoacides Métabolisme d'autres aminoacides Métabolisme et biosynthèse de glycanes Biosynthèse de polyketides et de peptides non-ribosomaux Métabolisme des cofacteurs et des vitamines Biosynthèse de métabolites secondaires Biodégradation et métabolisme de xénobiotiques Récapitulatifs
Traitement de l'information génétique	Transcription Traduction Repliement, adressage moléculaire et dégradation Réplication et réparation
Traitement de l'information environnementale	Transport membranaire Transduction du signal Molécules signal et interaction
Processus cellulaires	Transport et catabolisme Motilité cellulaire Croissance et mort cellulaire Communication cellulaire Système circulatoire Système endocrinien Système immunitaire Système nerveux Système sensoriel Développement Comportement
Maladies humaines	Cancers Maladies immunitaires Maladies neurodégénératives Troubles métaboliques Maladies infectieuses
Développement de molécules actives	Antibiotiques Antinéoplasiques Psychotropes Autres drogues Classification structurale basés sur la cible Classification structurale basés sur le squelette

La version 51 de juillet 2009 de KEGG comporte 332 voies de référence.

6.2.6 Interactions protéine-protéine : STRING

STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) (Jensen *et al.*, 2009) est une base de données d'interactions protéine-protéine prédites ou démontrées expérimentalement chez un certain nombre d'organismes. Il peut s'agir d'interactions directes (physiques) ou indirectes (fonctionnelles). Elle est distribuée depuis 2000 par l'EMBL et est développée conjointement avec le SIB et l'université de Zurich.

Les données d'interactions sont entièrement précalculées à partir d'un grand nombre de sources de données :

- Importation des banques de données d'interactions protéiques directes IntAct (Kerrien *et al.*, 2007), BioGRID (Breitkreutz *et al.*, 2008), HPRD (Prasad *et al.*, 2009), MINT (Chatr-aryamontri *et al.*, 2007), DIP (Salwinski *et al.*, 2004), GO (Ashburner *et al.*, 2000), et BIND (Alfarano *et al.*, 2005),
- Importation des banques de voies métaboliques KEGG (Kanehisa *et al.*, 2008), Reactome (Matthews *et al.*, 2009b), PID (Schaefer *et al.*, 2009) et EcoCyc (Keseler *et al.*, 2009),
- Extraction de données à partir de bases de connaissances PubMed, OMIM (Hamosh *et al.*, 2005), FlyBase (Tweedie *et al.*, 2009), et SGD (Hong *et al.*, 2008) et recherche de cooccurrences de noms de gènes.

Les interactions importées sont ensuite complémentées par des méthodes de prédictions basées sur le contexte génomique des gènes ou leur profil de expression :

- Conservation de gènes dans un voisinage proche : une suite de gènes conservée, entre plusieurs génomes procaryotes, à l'intérieur d'une distance chromosomique compatible avec un mécanisme de régulation commun peut indiquer une relation fonctionnelle.
- Fusion de gènes : la fusion de plusieurs gènes dans certains génomes indique potentiellement une étroite relation fonctionnelle entre ces gènes.
- Cooccurrence de gènes : les familles de gènes partageant un même profil phylogénétique de présence/absence dans plusieurs génomes sont considérées comme ayant une relation fonctionnelle.
- Coexpression : des associations fonctionnelles sont recherchées à l'intérieur des données d'expression de puces à ADN à l'aide du logiciel ArrayProspector (Jensen *et al.*, 2004).

À chaque interaction importée ou prédite, STRING associe un score de confiance compris entre 0 et 1. Ces scores sont dérivés des scores d'un jeu de données de référence composé

d'associations réelles et bien connues extraites à partir des groupements fonctionnels maintenus par KEGG (Kanehisa *et al.*, 2008). Pour chaque paire de protéines, il est possible de calculer un score d'interaction « combiné » reflétant le ou les types d'interactions que l'on souhaite, ou non, prendre en compte lors de la construction et de la visualisation d'une partie du graphe d'interactions fourni par le site Web de STRING (Figure 27).

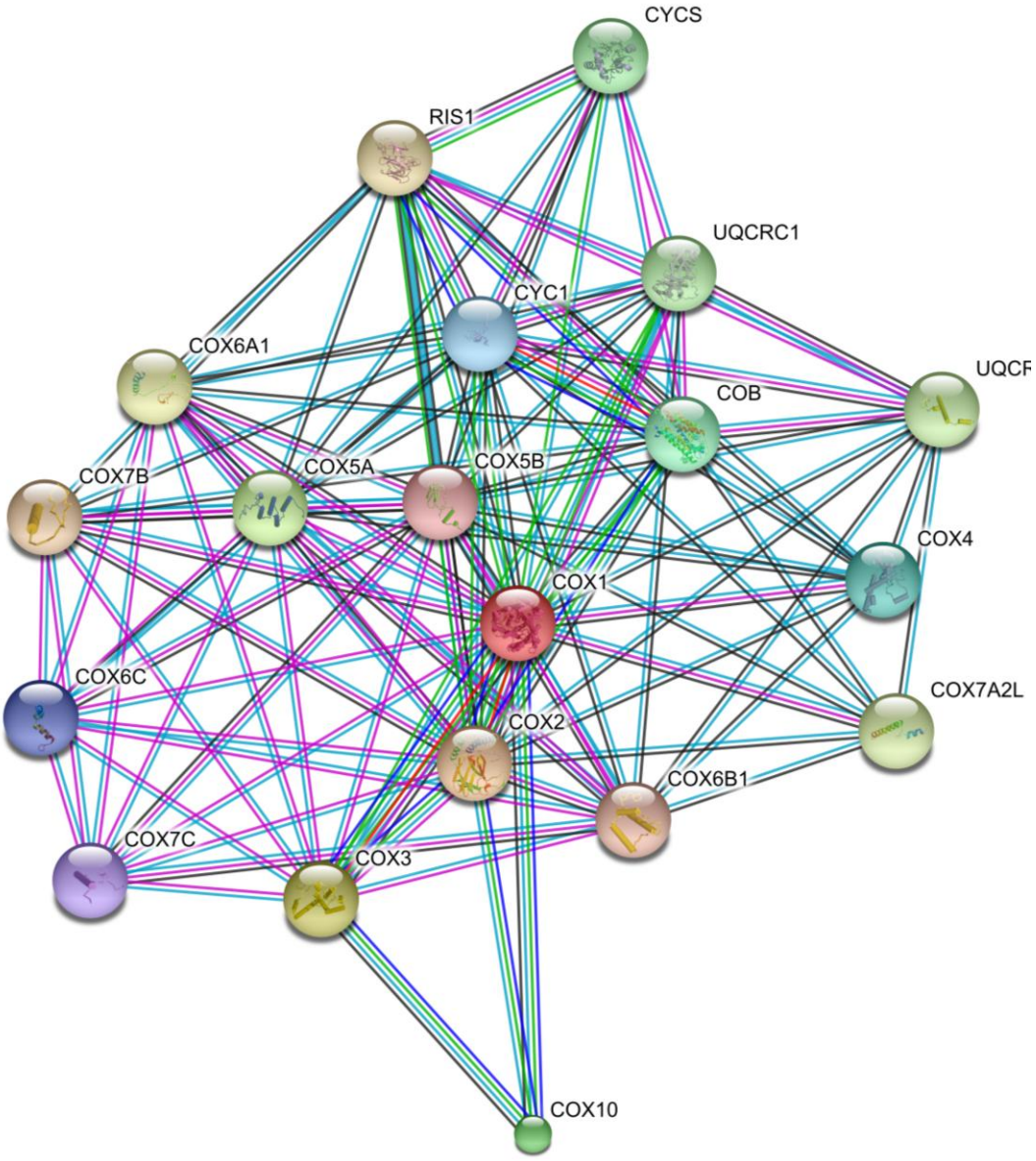


Figure 27 – Visualisation d'un sous-graphe STRING rassemblant les interactants de COX1
Seuls les interactants au premier degré de la cytochrome oxidase 1 humaine sont représentés (seuil de confiance ≥ à 0,900). Chaque nœud du graphe représente une protéine et chaque arête une interaction. Les couleurs des arêtes dépendent de la nature des données d'interaction (fusion de gènes, coexpression…). Image générée à partir du site http://string.embl.de/.

Enfin, les interactions connues ou prédites de STRING sont transférées entre organismes selon le principe des intérologues. Deux méthodes de prédiction d'orthologie sont utilisées :

- La première repose sur les groupes d'orthologues fournis par la banque COG (Tatusov *et al.*, 2003),
- La seconde utilise un algorithme de recherche des orthologues potentiels, inspiré des algorithmes de type INPARANOID (Remm *et al.*, 2001) et de type COG. Dans un premier temps, chacune des protéines contenues dans STRING est comparée à l'ensemble des autres protéines existantes. Les séquences protéiques très proches à l'intérieur d'un même génome sont regroupées en groupes d'in-paralogues (gènes dupliqués après un évènement de spéciation). Finalement, l'orthologie est détectée en comparant les groupes d'in-paralogues de plusieurs espèces, pour être joints en triangles de meilleurs hits réciproques, à la manière de COG.

La version 8.1 actuelle de STRING contient environ 2,5 millions de protéines provenant de 630 espèces.

6.3 Interrogation des banques

Les banques de données biologiques sont le plus souvent distribuées sous forme de fichiers plats. L'information à l'intérieur de ces fichiers est structurée de manière à être facilement lisible et modifiable par un être humain. Cependant l'organisation séquentielle de ces données ralentit la recherche d'informations que l'on souhaite extraire ou consulter, et ce d'autant plus que le volume de ces banques est imposant. Il est ainsi nécessaire d'utiliser des outils de recherche adaptés pour mener à bien des études à haut-débit.

Il existe principalement deux catégories d'outils de recherche : d'une part, les outils de recherche par similarité de séquence, qui indexent uniquement les données de séquence et permettent de retrouver les séquences similaires d'une séquence fournie en tant que cible de recherche et, d'autre part, les outils de recherche textuelle, qui indexent tout, ou partie, des informations et permettent de réaliser des recherches par mots-clés.

6.3.1 Interrogation par similarité : BLAST

La recherche par similarité d'une séquence inconnue dans une banque en vue de la caractériser rapidement par l'intermédiaire de séquences proches déjà annotées constitue une approche de base indispensable en bioinformatique.

Pour ces recherches, nous avons utilisé la famille d'outils BLAST (*Basic Local Alignment Search Tool*) (Altschul *et al.*, 1997) fournie par le NCBI. BLAST est un algorithme heuristique qui permet de retrouver très rapidement un ensemble de séquences proches d'une séquence d'intérêt en procédant à des alignements 2-à-2 non optimaux.

À ces alignements est rattachée une valeur d'espérance mathématique (*E-value* ou *expect*) mesurant la signification biologique de ces alignements par comparaison à des alignements

générés à partir de séquences aléatoires ayant même longueur et même composition que la séquence requête. Plus la valeur d'*expect* est proche de 0, plus un alignement est significatif. De manière empirique, une séquence ayant une *E-value* associée ≤ 0,001 présente généralement une similarité significative avec la séquence d'intérêt.

6.3.2 Interrogation textuelle : SRS

Le moteur d'indexation et de recherche SRS (*Sequence Retrieval System*) (Etzold *et al.*, 1993) a été développé à l'EMBL, puis racheté par *LION bioscience* et est maintenant détenu par *biowisdom* (http://www.biowisdom.com/).

Il est installé localement à l'IGBMC, dans sa dernière version 8.3, et est interrogeable par l'intermédiaire de son interface de recherche Web (Figure 28) ainsi que par son programme d'interrogation en ligne de commande « *getz* ».

Ces deux méthodes d'interrogation permettent de réaliser des requêtes simples ainsi que des requêtes croisées sur les banques indexées par SRS. Pour effectuer des requêtes complexes impossibles à réaliser à l'aide du langage de requête de SRS, il est possible de passer par le langage de programmation *Icarus* (*Interpreter of commands and recursive syntax*) qui régit la mécanique interne de SRS. *Icarus* permet une manipulation directe et beaucoup plus rapide des index de SRS. L'interface de programmation Java *SRS WSObjects* qui permet aussi l'interrogation directe du cœur de SRS est actuellement en test au BIPS et sera bientôt disponible.

Figure 28 – Interface Web de recherche de SRS 8.3 sur le serveur du BIPS

6.3.3 BIRD/BIRD-QL

BIRD (*Biological Integration and Retrieval Data*) (Nguyen *et al.*, 2008) est le système de recherche textuelle développé actuellement au LBGI. BIRD a été intégré et utilisé dans nos développements en parallèle ou en complément à SRS selon la situation.

Le système BIRD ne se veut pas limité à la gestion de banques de données et à la récupération de données, mais va aussi permettre dans un futur proche, l'extraction automatique de connaissances (KDD, *Knowledge Discovery in Databases*), ainsi que l'élaboration de jeux de règles de décision grâce aux techniques de la programmation logique inductive (ILP, *Inductive logic programming*).

Le système BIRD repose sur un cœur *IBM WebSphere* (http://www.ibm.com/), qui contient, entre autres, le serveur de données *IBM DB2* et le serveur de fédération de bases de données *IBM Websphere Federation Server*. BIRD dispose ainsi d'un accès rapide à l'information et peut interroger indifféremment des bases de données locales et distantes sous la forme d'une seule base de données virtuelle.

L'interrogation des banques par BIRD peut s'effectuer soit par l'intermédiaire d'une interface Web, soit par *Web Service* (WS), ce qui permet une interrogation automatisée à

partir de n'importe quel langage de programmation supportant le protocole de transfert HTTP (*Hypertext Transfer Protocol*).

La récupération de données passe par un langage de requête propre à BIRD et appelé BIRD-QL (*BIRD Query language*). Ce langage dérivé du langage d'interrogation de bases de données SQL (*Structured query language*) et présenté dans un format proche du format EMBL, permet aux bioinformaticiens de formuler facilement des requêtes élaborées en masquant la complexité de l'architecture sous-jacente (Figure 29).

```
ID   * DB Uniprot
WH   DE contains "cytochrome"
WH   DE contains not "fragment"
WH   OX is 6376
FD   AC,ID,DE,SQ
FM   FASTA
//
```

Figure 29 – Exemple de requête BIRD-QL
Demande de toutes les séquences de la banque Uniprot dont la description comporte le mot « cytochrome » mais pas le mot « fragment » et provenant de l'espèce *Alvinella pompejana* (NCBI taxonomy ID de 6376). Les résultats sont retournés au format FASTA et comportent les numéros d'accès, l'identifiant, la description et la séquence.

7 OUTILS INFORMATIQUES

Dans cette section, nous présenterons les deux langages de programmation qui ont principalement été utilisés dans les développements bioinformatiques et les développements Web réalisés au cours de mes travaux de thèse. Ensuite, nous décrirons le système de gestion de base de données relationnelles (SGBDR) mis en oeuvre dans la conception de la base de données d'assemblage (Cf. Base de données d'assemblage, page 124).

7.1 Langages de programmation

Afin de se reposer sur l'infrastructure déjà largement testée et éprouvée par le laboratoire, de maintenir une cohérence avec les développements déjà réalisés avant mon arrivée au laboratoire, et de faciliter ainsi, l'intégration et la maintenance des nouvelles réalisations présentées dans ce manuscrit, il a rapidement été décidé d'adopter les deux principaux langages de programmation utilisés par les membres de l'équipe : le Tcl/Tk et le PHP.

7.1.1 Tcl/Tk

Le Tcl (*Tool Command Language*) et son extension graphique Tk (*ToolKit*) (Ousterhout, 1994) est un langage de script, c'est-à-dire interprété et non compilé, qui a été créé en 1988 par John Ousterhout à l'université de Californie, Berkeley.

Tcl/Tk a l'avantage d'être très facile à prendre en main puisque le langage de base ne comporte qu'un jeu relativement restreint de commandes qui peuvent être aisément complétées grâce aux nombreuses extensions disponibles, notamment la librairie standard *tcllib* (http://tcllib.sourceforge.net/). Tcl/Tk est aussi reconnu pour être performant dans la manipulation de données textuelles, ce qui le rend tout à fait adapté à une utilisation bioinformatique.

Le couple Tcl/Tk a été utilisé au cours de cette thèse dans sa version 8.5 pour réaliser toutes les opérations et les interfaces graphiques des programmes d'analyse, de traitement et d'annotation de séquences, ainsi que pour le chargement de base de données.

7.1.2 PHP

PHP (*PHP: Hypertext Preprocessor*) (PHP) est un langage de script interprété initialement conçu par Rasmus Lerdorf en 1995 et destiné à être utilisé dans la génération dynamique de pages Web. Il a depuis été repris et intégralement réécrit en 1998 par Andi Gutmans et Zeev Suraski, cofondateurs de *Zend Technologies* qui maintient le cœur de PHP et qui commercialise le support et des outils de développement pour ce langage.

PHP est devenu extrêmement populaire dans la création de sites Web pour sa simplicité de mise en œuvre et ses centaines d'extensions disponibles en standard ou dans les dépôts

PEAR (*PHP Extension and Application Repository*). Cette popularité est illustrée par les 20 millions de domaines fonctionnant en PHP (http://www.php.net/usage.php).

La version de PHP qui a été utilisée est la version 5.2, déployée en tant que module du serveur Web Apache 2.2.

7.2 PostgreSQL

La gestion et la recherche au travers d'un grand nombre de données peuvent être simplifiées par l'utilisation d'un système de gestion de bases de données relationnelles (SGBDR). Cela permet de structurer l'information et d'y accéder dans des temps réduits.

PostgreSQL (PostgreSQL) est un SGBDR dont le développement a été initié en 1985 par Michael Stonebraker à Berkeley. PostgreSQL supporte presque intégralement le dernier standard en date SQL:2003 (*Structured query language*, norme de 2003), ce qui en fait un outil de qualité professionnelle.

Au moment où la question du choix de SGBDR s'est posée, nous hésitions entre l'utilisation de PostgreSQL et celle de MySQL (MySQL), un autre SGBDR légèrement plus rapide. Cependant PostgreSQL était déjà utilisé dans la conception de la banque de données de transcriptomique rétinienne RETINOBASE (Cf. Annexe 2, page 191) et s'est avéré très stable. Comme nous avons préféré miser sur cette stabilité, notre choix s'est porté sur PostgreSQL dans sa version 8.3 pour stocker et gérer une partie des données traitées et générées au cours de cette thèse.

8 OUTILS BIOINFORMATIQUES

Au cours de cette thèse, de nombreux programmes du laboratoire ou développés par des laboratoires tiers ont été utilisés. Schématiquement, ils peuvent être regroupés en deux catégories : les outils de traitement des ESTs et les outils d'annotation de protéines.

8.1 Outils liés aux traitements des EST

Les outils qui vont être présentés ici ont été intégrés au pipeline de traitement et d'analyse des données de séquençage (Cf. Traitement et analyse des données brutes de séquençage, page 109). Ils permettent le traitement et le nettoyage des données fournies par des centres de séquençage, pour obtenir au final des séquences d'EST de bonne qualité qui seront annotées structuralement.

8.1.1 Suite logicielle Staden

Le *Staden package* (Staden *et al.*, 2000) est une suite logicielle regroupant plusieurs programmes dédiés à l'assemblage de séquences nucléotidiques, l'édition d'assemblages, et l'analyse de séquences nucléotidiques et protéiques. Cette suite est développée depuis 1977 par le groupe de Roger Staden au *MRC Laboratory of Molecular Biology* à Cambridge.

C'est notamment grâce aux recherches de ce groupe qu'est né le format de fichier SCF (*Standard Chromatogram File*) (Dear et Staden, 1992). Ce format est très bien documenté et permet de stocker les profils chromatographiques des séquenceurs automatiques de manière standard. Il est reconnu et utilisé par un grand nombre de logiciels du domaine.

Les chromatogrammes stockés dans des formats propriétaires (ABI, ALF, ESD, …) peuvent être convertis au format SCF grâce au programme *makeSCF* de la suite Staden.

8.1.2 Phred

Phred (Ewing *et al.*, 1998; Ewing et Green, 1998) fait partie intégrante de la suite Phred/Phrap/Consed destinée à l'assemblage de génomes et développée par le laboratoire de Phil Green du *Genome Science Department* de l'université de Washington. Phred est le programme standard *de facto* réalisant l'opération de '*base-calling*' qui consiste à assigner une base azotée (*a*, *c*, *g*, *t*, ou *n* si indétermination) à chaque pic d'intensité d'un chromatogramme. L'opération de *base-calling* est complétée par l'attribution d'une valeur de qualité à chaque base assignée (Figure 30). Cette valeur représente l'inverse du taux d'erreur d'attribution d'une mauvaise base pour un pic donné. Plus cette valeur est élevée, meilleure est la prédiction de base.

$$q = -10 \times \log_{10}(p)$$

Figure 30 – Formule de calcul d'une valeur de qualité par le logiciel Phred
q représente la valeur de qualité et *p* représente la probabilité d'erreur de base calling, *q* pouvant varier de 0 à 97.

8.1.3 Cross_match

Cross_match (Phil Green, non publié) est un programme qui compare très rapidement un jeu de séquences de référence à un (ou plusieurs) autre(s) jeu(x) de séquences et qui permet de masquer, dans le jeu de référence, les segments de séquences retrouvés dans le jeu contre lequel il a été comparé.

Cross_match a été utilisé dans le but de masquer les séquences contaminantes apparaissant à l'intérieur d'ESTs par comparaison contre la banque Univec.

8.1.4 Cap3

Cap3 (Huang et Madan, 1999) est un programme réalisant l'assemblage de séquences nucléotidiques.

Le processus d'assemblage consiste à réunir et aligner des fragments de séquences d'ADN ou d'ADNc sous contraintes du sens de séquençage (5' ou 3') de chaque fragment, des valeurs de qualité de chaque base, et d'un certain nombre de critères à remplir au niveau des parties chevauchantes. Pour chaque alignement, appelé dans ce cas précis « contig », une séquence consensus est calculée, en tenant compte des valeurs de qualité de chaque base, pour déterminer une « séquence moyenne » représentant le plus fidèlement possible le contig. Les séquences n'ayant pu être intégrées à aucun contig sont appelées singletons (*singlets*) (Figure 31).

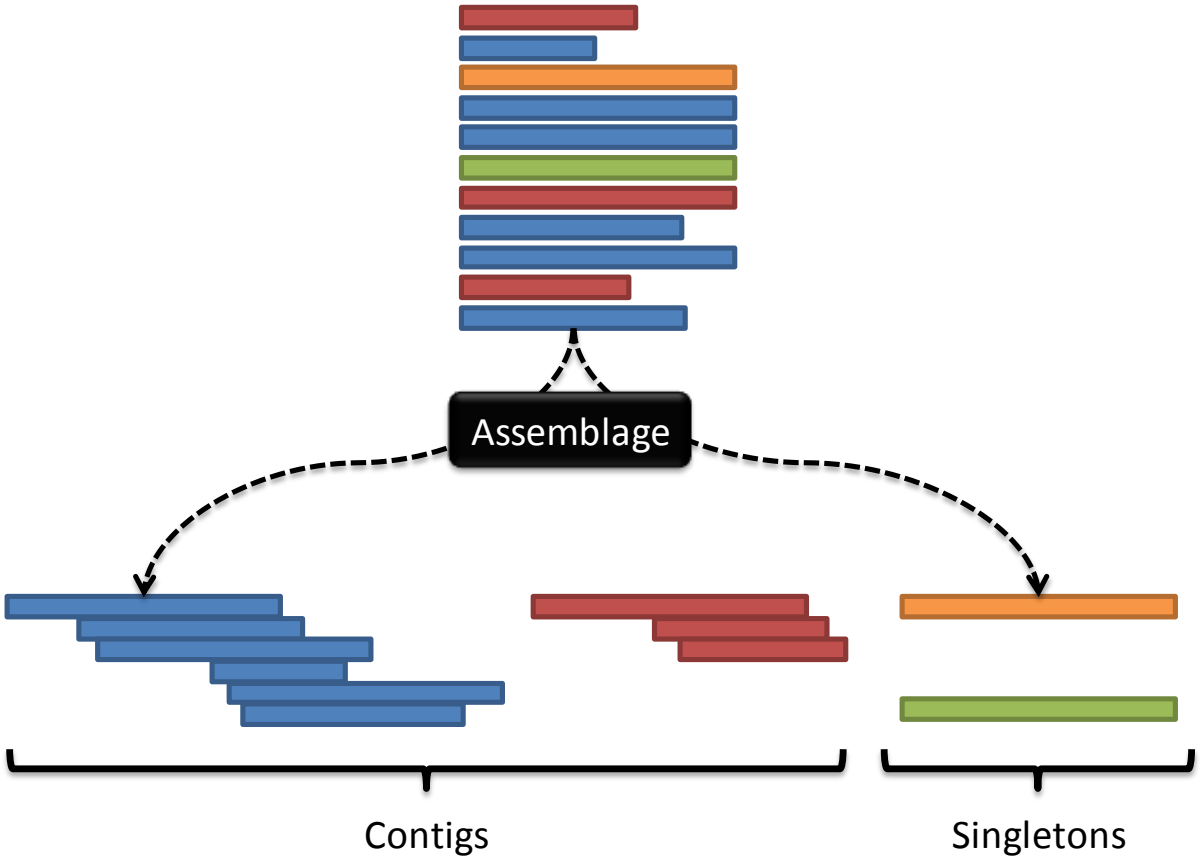


Figure 31 – Assemblage de séquences
Les séquences de la même couleur ont des parties chevauchantes et peuvent donc être alignées entre elles dans un même contig.

Ce procédé permet d'obtenir des séquences de plus grande taille et de bien meilleure qualité. Idéalement, avec une couverture de séquençage suffisante, le nombre final de séquences obtenu doit correspondre, dans le cas de fragments génomiques, au nombre de chromosomes de l'organisme séquencé, ou dans le cas d'EST, au nombre de transcrits différents présents lors de la construction de la librairie de séquençage du tissu/organe/organisme.

8.1.5 Consed

Consed (Gordon *et al.*, 1998; Gordon, 2003) est un logiciel graphique dédié à la visualisation, à l'édition et à la finition (*finishing*) d'assemblages.

Grâce à son mode de coloration en dégradé de gris, basé sur les valeurs de qualité de bases et la signalisation en rouge de conflits avec la séquence consensus, la détection de problèmes d'assemblage est rapide et simplifiée. Une fois les problèmes identifiés, il est possible de déplacer les séquences d'un contig à l'autre et de lancer un réassemblage d'un ou plusieurs contigs.

Consed dispose aussi d'une fonction de finition semi-automatique '*autofinish*' (Gordon *et al.*, 2001), qui dans le cas de banques génomiques, propose les clones à reséquencer pour combler des zones manquantes ou corriger des zones de mauvaise qualité.

8.1.6 tRNAscan-SE

tRNAscan-SE (Lowe *et al.*, 1997) est un programme qui détecte les ARNt à l'intérieur de séquences nucléotidiques d'organismes procaryotes ou eucaryotes.

tRNAscan-SE combine deux algorithmes heuristiques pour obtenir une recherche exhaustive. L'un est basé sur la recherche d'occurrences du signal 'B box' de la boucle T et tente la construction de la structure secondaire canonique en forme de trèfle des ARNt autour de chaque occurrence. L'autre algorithme se base uniquement sur la recherche de motifs 'A Box', 'B Box' et sur la distance entre le signal de terminaison de la polymérase III et du motif 'B Box'. Les gènes potentiels ainsi détectés sont ensuite validés ou rejetés par un troisième algorithme, Covels, basé sur des modèles de covariance.

8.1.7 ESTScan2

ESTScan2 (Iseli *et al.*, 1999; Lottaz *et al.*, 2003) est utilisé pour détecter les régions codantes des séquences d'ARNm, même en cas d'erreurs de séquençage ou de décalage de cadre de lecture (*frameshift*).

ESTScan2 utilise un modèle HMM qui représente le biais d'utilisation en hexanucléotides aussi bien pour les régions codantes que pour les régions 5' et 3' UTR. Ce modèle permet de détecter les débuts et fins de régions codantes, ainsi que de possibles erreurs de séquences et propose une version corrigée de la séquence par des insertions ou des délétions. Une traduction protéique de la région codante de la séquence corrigée peut aussi être obtenue en sortie de ce programme.

8.2 Outils liés à l'annotation automatique de protéines

Une grande partie des programmes que nous avons utilisés dans le cadre de l'annotation automatique de protéines est basée sur l'analyse et l'exploitation d'alignements multiples de séquences complètes (MACS, *Multiple Alignment of Complete Sequences*) (Lecompte, Thompson, *et al.*, 2001).

Dans cette section seront décrits les outils utilisés pour la construction de MACS de bonne qualité, leur exploitation, et leur analyse. 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.

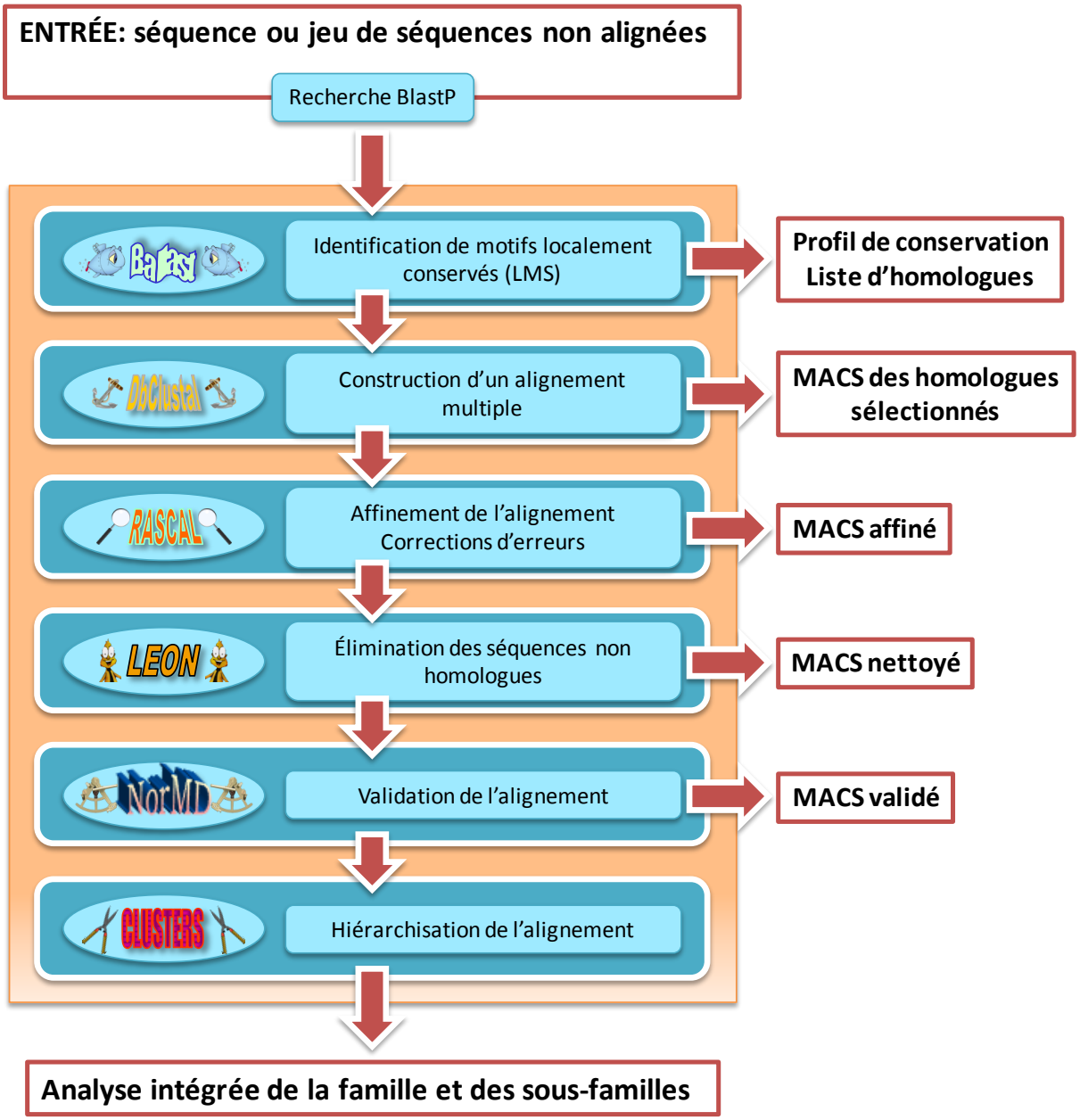


Figure 32 – Aperçu de la cascade de programmes constituant PipeAlign
Adaptée de (Plewniak *et al.*, 2003)

8.2.1 PipeAlign : un outil d'analyse de familles protéiques

PipeAlign (Plewniak *et al.*, 2003) est un outil d'analyse de famille de protéines qui a été développé au sein du laboratoire. Il 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 séquences protéiques à l'aide du logiciel BlastP. Les programmes composant PipeAlign vont être présentés succinctement dans les sections suivantes.

8.2.1.1 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 inversement proportionnelle à son *E-value*. Le profil est ensuite lissé. La dérivée seconde du profil fournit des pics définissant les segments de conservation maximale, encore appelés LMS (*Local Maximum Segments*). Les LMS 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.

8.2.1.2 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. Il se base sur l'algorithme du très populaire ClustalW (Thompson *et al.*, 1994), programme d'alignement multiple basé sur l'algorithme d'alignement global développé par Needleman et Wunsch (Needleman *et al.*, 1970). Même si ClustalW reste très utilisé, des études ont mis en évidence les inconvénients d'une méthode basée uniquement sur l'alignement global (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 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.

8.2.1.3 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 (*RApid Scanning and Correction of ALignment errors*) (Thompson *et al.*, 2003) 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 identifiées. 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 d'affinement plus performante.

8.2.1.4 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 consiste à détecter, au sein du MACS, les séquences n'appartenant pas à la famille d'intérêt. Le programme LEON (*multiple aLignment*

Evaluation Of Neighbours) (Thompson *et al.*, 2004), 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 théoriquement d'un MACS de bonne qualité qui ne contient que des séquences partageant au moins une région homologue à la séquence initiale.

8.2.1.5 NorMD : évaluation de la qualité d'un MACS

NorMD (*Normalized Mean Distance*) (Thompson *et al.*, 2001) 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… 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.

8.2.1.6 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 (*Density of Points Clustering*) (Wicker *et al.*, 2002), permettent la classification, ou hiérarchisation, des séquences d'un alignement multiple dans des sous-groupes potentiellement fonctionnels ou taxonomiques, le nombre de sous-groupes créés étant déterminé de façon automatique par ces programmes.

PipeAlign est mis à disposition de la communauté scientifique par l'intermédiaire du site http://bips.u-strasbg.fr/PipeAlign/.

8.2.2 Annotation automatique de protéines

8.2.2.1 GOAnno

GOAnno (Chalmel *et al.*, 2005) permet l'annotation Gene Ontology automatique de séquences protéiques inconnues. GOAnno profite du contexte évolutif apporté par les MACS hiérarchisés pour récupérer les termes GO de la sous-famille de séquences à laquelle appartient la séquence inconnue à annoter afin de propager ces termes à cette dernière.

L'algorithme de GOAnno est suffisamment sélectif pour ne propager que les termes GO qui apparaissent un nombre raisonnable de fois, tout en évitant de propager les termes trop spécialisés potentiellement inapplicables à la totalité de la sous-famille.

GOAnno est accessible à l'adresse http://bips.u-strasbg.fr/GOAnno/, et est disponible en tant qu'exécutable en ligne de commande.

8.2.2.2 MACSIMS

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

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 les prédictions *ab initio* basées sur les séquences. Une étape de validation croisée des informations recueillies, s'appuyant sur les données de conservation au sein de l'alignement multiple hiérarchisé, et la cohérence 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 33.

Les informations relatives aux séquences de l'alignement sont collectées au sein des banques de données publiques UniProt, PDB et InterPro, et sont de natures très diverses (domaines Pfam-A et Prosite, données taxonomiques, numéros EC, sites actifs ou de liaison, résidus modifiés…). 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.

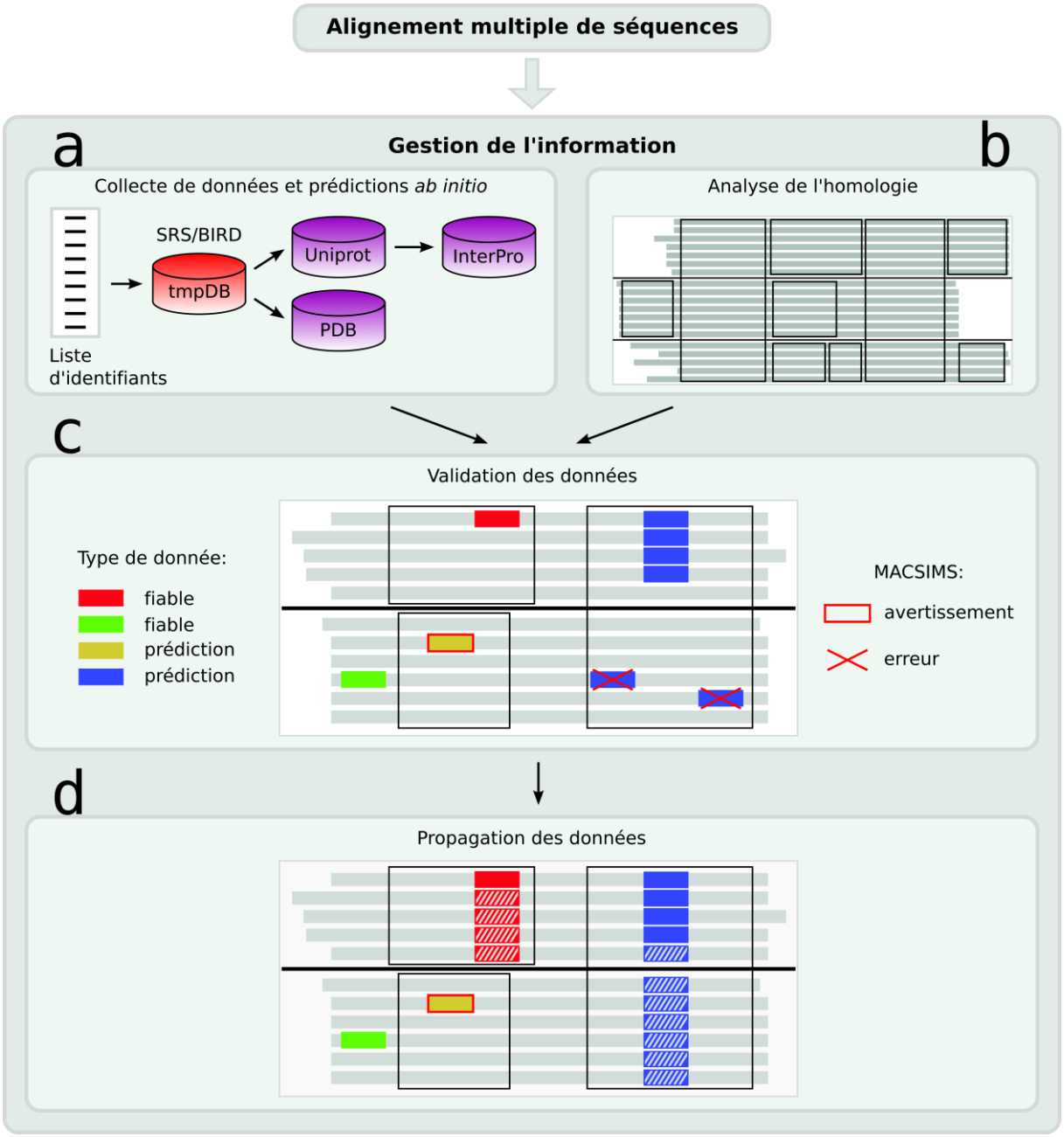


Figure 33 – Les quatre grandes étapes de MACSIMS
(a) Pour chaque séquence de l'alignement, des informations sont collectées au sein de banques de données publiques, et certaines caractéristiques sont prédites *ab initio*. (b) Les blocs de conservation de l'alignement sont déterminés. (c) Les données fiables sont validées, les autres sont éliminées. (d) Les données fiables sont transférées des séquences annotées aux séquences inconnues. Figure adaptée de (Thompson *et al.*, 2006).

Les informations collectées ou générées par MACSIMS sont sauvegardées dans un format XML (*Extensible Markup Language*), un langage standardisé de structuration des données, permettant une lecture et une exploitation informatique des données simplifiées. MACSIMS prend en charge la conversion de son format XML dans un format compatible avec l'éditeur d'alignements multiples Jalview (Waterhouse *et al.*, 2009) qui permet la visualisation simple

et conviviale, des alignements annotés sous forme d'application ou d'applet Web (Figure 34).

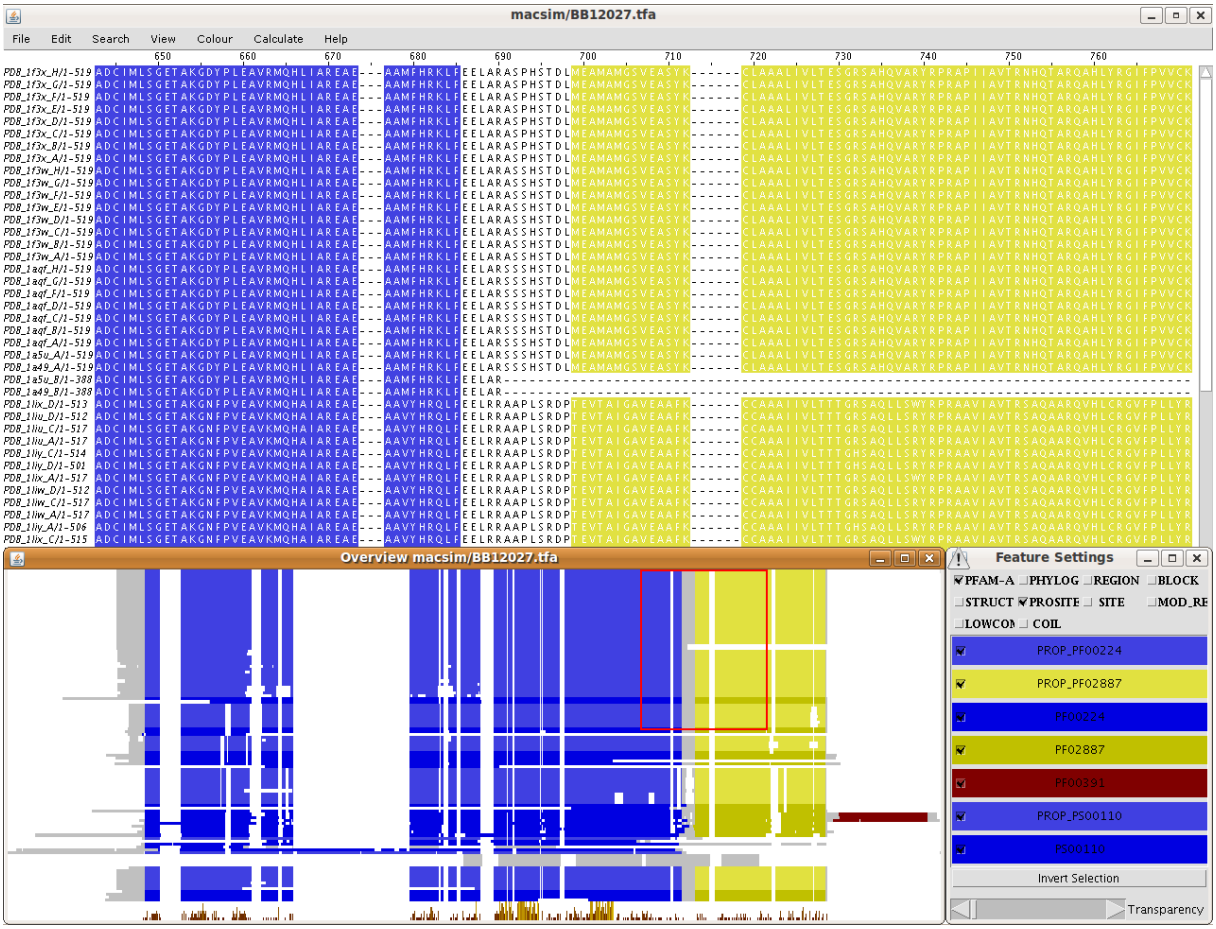


Figure 34 – Alignement multiple annoté par MACSIMS et visualisé dans Jalview
L'alignement multiple annoté apparait dans la fenêtre principale. Un aperçu de l'alignement complet est visible en bas à gauche. La fenêtre de sélection des informations à afficher qui apparait en bas à droite permet de choisir les données à visualiser sur l'alignement. Les différentes annotations propagées par MACSIMS sont préfixées de « PROP_ ».

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. Il est disponible à l'adresse http://bips.u-strasbg.fr/MACSIMS/.

8.2.3 Analyse quantitative de l'annotation

8.2.3.1 Outil d'annotation fonctionnelle du logiciel DAVID

DAVID (*Database for Annotation, Visualization and Integrated Discovery*) (Dennis *et al.*, 2003; Huang *et al.*, 2009) regroupe un ensemble d'outils Web destinés à l'annotation fonctionnelle d'ensemble de gènes à l'aide de sa propre banque de données, *DAVID knowledgebase*. Cette dernière intègre les identifiants de gènes ou protéines de plusieurs

espèces, ainsi que leur annotation, à partir d'une grande variété de banque de données publiques (NCBI, PIR, SWISS-PROT, GO, OMIM, PubMed, KEGG, BIOCARTA, AffyMetrix, TIGR, Pfam, BIND, MINT, DIP…).

Les outils fournis par DAVID analysent des listes de gènes fournies par l'utilisateur et sont disponibles à l'adresse http://david.abcc.ncifcrf.gov/. Ils comprennent l'outil d'annotation fonctionnelle (analyse de l'enrichissement en catégories fonctionnelles, cartographie sur les voies métaboliques, résumé d'annotation sous forme de graphiques…), l'outil de classification fonctionnelle de gènes (regroupement de gènes ayant une annotation fonctionnelle similaire), et l'outil de conversion d'identifiants.

Dans nos études, nous avons utilisé l'outil d'analyse des enrichissements fonctionnels dans plusieurs listes de gènes.

8.2.3.2 Ingenuity Pathway Analysis

Ingenuity Pathways Analysis (IPA) (Ingenuity® Systems, http://www.ingenuity.com) est une suite applicative Web commerciale permettant de modéliser et d'analyser des systèmes biologiques et chimiques complexes.

IPA propose différents types d'analyses :

- '*IPA Core Analysis*' : interprétation de jeux de données dans le contexte de processus biologiques, de voies métaboliques et de réseaux moléculaires
- '*IPA-Metabolomics Analysis*' : analyse de la physiologie et du métabolisme cellulaire
- '*IPA-Tox Analysis*' : analyse de la toxicité de composés chimiques dans le contexte de phénotypes et de pathologies cliniques
- '*IPA-Biomarker Analysis*' permet de prioriser les molécules actives en fonction de leurs effets sur des maladies dans le cadre de candidats de médicaments

Nos listes de gènes ont été analysées en parallèle du logiciel DAVID avec *IPA-Core Analysis*.

9 GSCOPE : PLATE-FORME DE GÉNOMIQUE DU LABORATOIRE

Gscope (Raymond Ripp, manuscrit en préparation) est une plate-forme de génomique dédiée à l'étude, la visualisation, la gestion et le traitement automatisé de données massives. Gscope a été initialement développé pour les besoins de l'annotation du génome de *Pyrococcus abyssi* (Lecompte, Ripp, *et al.*, 2001).

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 le pipeline de traitement de données de séquençage (Cf. Traitement et analyse des données brutes de séquençage, page 109) et le pipeline d'annotation intégrative de séquences protéiques (Cf. Annotation intégrative automatique de séquences protéiques, page 135).

Gscope va être ici présenté du point de vue historique avec l'annotation de *Pyrococcus abyssi* afin de mieux comprendre sa philosophie, tout en sachant que la plupart des concepts qui vont être présentés sont transposables pour des organismes eucaryotes.

9.1 Architecture générale

La philosophie de Gscope repose sur le concept de projet, constitué initialement à partir d'une séquence génomique ou d'une collection homogène de séquences de gènes, d'ADNc ou de protéines. Chaque séquence, lors de son intégration au projet, se voit attribuer un numéro d'accès séquentiel spécifique au projet.

Un projet Gscope étant structuré par une hiérarchie de dossiers et de fichiers plats bien définie, ce numéro permet de maintenir le lien entre le fichier de la séquence initiale et les différents fichiers de traitements et d'analyses qui lui ont été appliqués (Figure 35). Cet ensemble de fichiers constitue les informations relatives à ce qu'on pourrait appeler par abus de langage une ORF (*Open Reading Frame*), qui dans Gscope est historiquement dénommé PAB (pour *Pyrococcus abyssi* box), et qui est l'élément de base sur lequel on peut appliquer des traitements puis les visualiser.

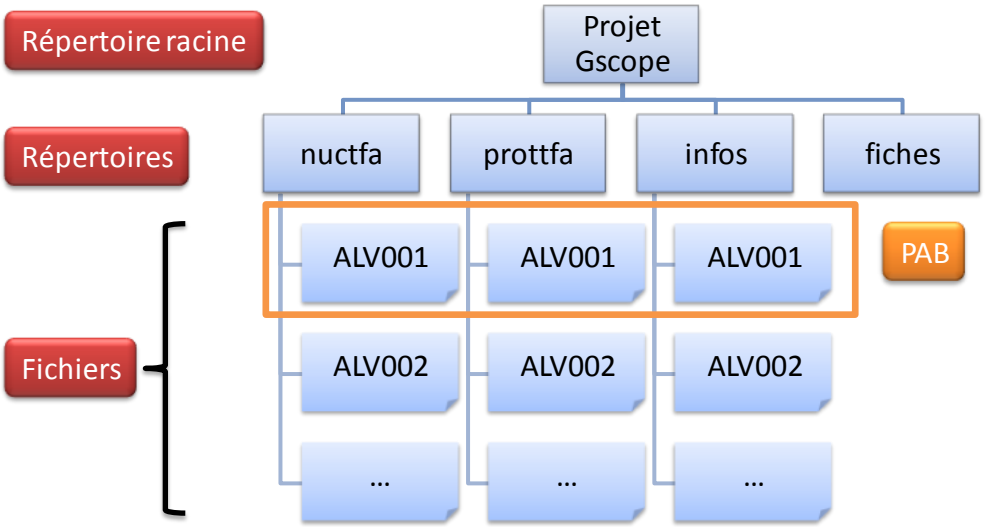


Figure 35 – Structure minimale simplifiée d'un projet Gscope.
Les répertoires « nuctfa » et « prottfa » contiennent respectivement les fichiers des séquences nucléotidiques et protéiques au format FASTA. Le répertoire « infos » renferme des fichiers plats de propriétés de type clé/valeur qui décrivent des annotations spécifiques à un PAB (composition en GC, description, organisme d'où provient la séquence…). Le répertoire « fiches » peut contenir des fichiers de résultats qui concernent tous les PAB, plutôt que de générer un fichier de résultat par PAB. Cette structure est dynamique : par exemple, Gscope peut créer un répertoire « blastn » pour y placer les fichiers de résultats d'une recherche BlastN.

Gscope est intégralement conçu en langage Tcl/Tk. Il dispose d'une interface graphique permettant de visualiser facilement et rapidement les PAB et leur annotation à l'aide d'une vue sous forme linéaire (et circulaire dans le cas de génomes bactériens) (Figure 36).

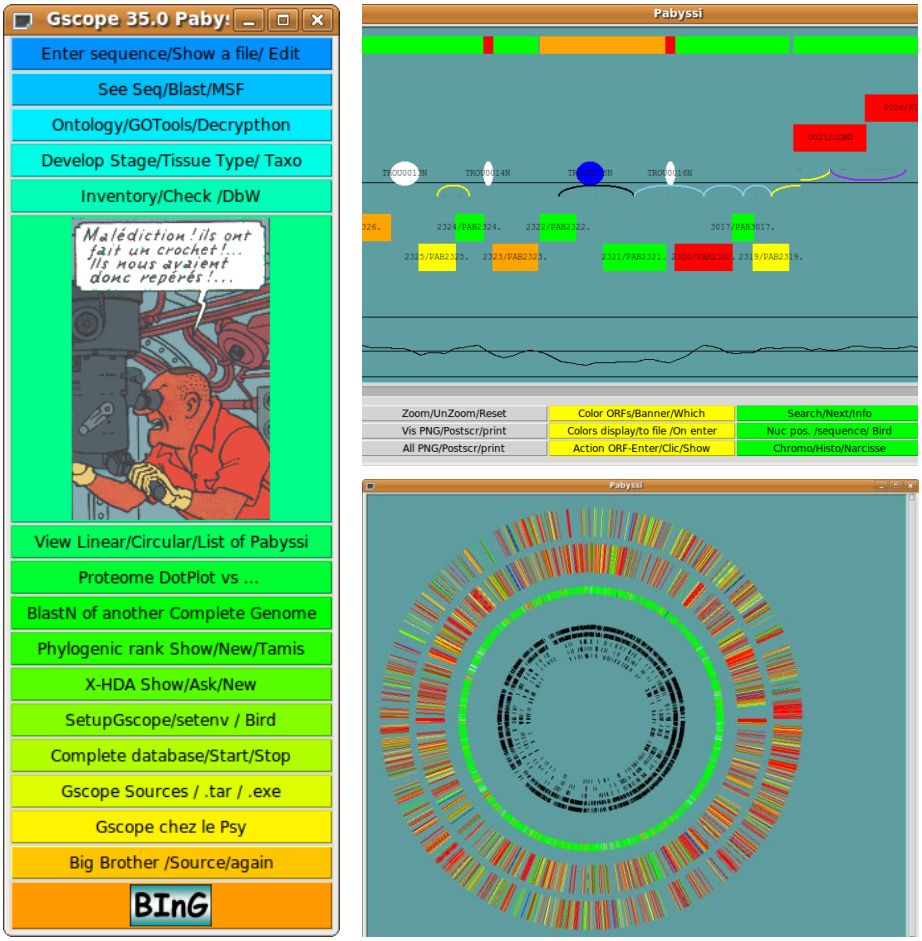


Figure 36 – Interface graphique de Gscope et aperçus du génome annoté de *P.abyssi*
La fenêtre principale (à gauche) permet, parmi un choix de plusieurs vues graphiques, d'accéder aux aperçus linéaire (en haut à droite) ou circulaire (en bas à droite) des PAB du projet en cours.

En sélectionnant un PAB, une interface apparait pour permettre de visualiser et modifier son annotation et donne accès à tout un panel de traitements classiques applicables aux séquences de n'importe quel projet (BLAST, alignement multiple…) (Figure 37).

Figure 37 – Visualisation de l'annotation d'un PAB sous Gscope.

9.2 Interfaces supplémentaires

Gscope dispose aussi d'une interface en ligne de commande pour traiter les données par lot qui permet de lancer des traitements classiques, ou des traitements spécifiques non accessibles par l'intermédiaire de l'interface graphique, sur tout ou partie des PAB d'un projet. Les traitements lourds et longs peuvent ainsi être lancés sur plusieurs serveurs en même temps pour répartir la charge et accélérer la vitesse de traitement.

Chaque projet est aussi accessible par l'intermédiaire d'une interface Web CGI (*Common Gateway Interface*) pour mettre les résultats à la disposition de collaborateurs externes au laboratoire (Figure 38).

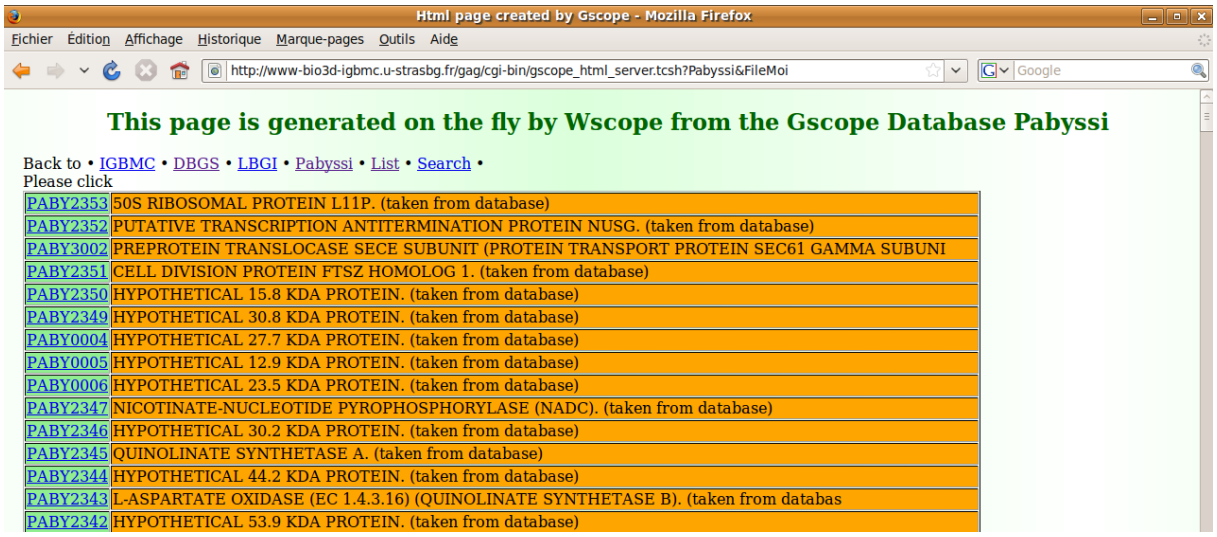


Figure 38 – Interface Web de Gscope

9.3 Et le café ?

Gscope est une véritable boite à outils qui a évolué pour intégrer de nombreux programmes externes et une grande quantité de sources de données par l'intermédiaire de requêtes SRS, Bird, MySQL, PostgreSQL, HTTP et de requêtes Gscope inter-projets.

Son extension est assurée par son architecture modulaire qui permet aux différents membres du laboratoire de rajouter leurs propres modules d'analyses adaptés à leur domaine d'étude.

De par son ouverture, Gscope permet de couvrir les besoins nécessaires à la conduite de nombreux types d'analyses, de l'annotation de génome à l'analyse de données de transcriptomique Affymetrix, en passant par la conception d'oligonucléotides de puces à ADN ou de mutagénèse.

RÉSULTATS ET DISCUSSIONS

Les résultats obtenus au cours de cette thèse ont été dissociés en 4 chapitres organisés selon le schéma suivant (à l'exception notable du chapitre 12) : chaque début de chapitre va décrire les stratégies employées dans l'élaboration des protocoles ou logiciels qui ont permis le traitement des banques d'ADNc d'*Alvinella* à différents niveaux. Leur utilisation sera ensuite illustrée dans le cadre du projet *Alvinella* et de collaborations.

Le premier chapitre (chapitre 10) décrira le protocole mis en place dans une logique haut-débit et d'automatisation pour traiter les données brutes de séquençage et aboutir à des séquences d'ADNc exploitables, c'est-à-dire annotées structuralement et traduites en séquences protéiques le cas échéant. Le chapitre suivant (chapitre 11) traitera de l'annotation intégrative de séquences protéiques, toujours en suivant cette logique de haut-débit et d'automatisation. Le chapitre 12 introduira la première publication issue du projet *Alvinella* et en présentera les résultats majeurs. Enfin, le dernier chapitre de résultats (chapitre 13) va illustrer deux approches qui ont été implémentées pour partager et visualiser les résultats des traitements précédents.

10 TRAITEMENT ET ANALYSE DES DONNÉES BRUTES DE SÉQUENÇAGE

Pour être exploitables en tant que séquences nucléotidiques, les données provenant de séquenceurs automatiques doivent être pré-traitées. Ces traitements permettent, entre autres, d'éliminer les séquences ou fragments de séquences contaminants issus de données artéfactuelles attribuables aux techniques de clonage et de séquençage, et de maximiser le rapport signal/bruit afin de ne conserver que les données de bonne qualité.

La quantité de données de séquençage à gérer impose le déploiement d'un système automatique de traitement et d'analyse très souvent relié à un système d'annotation fonctionnelle automatique des séquences. Bien qu'il en existe un large panel, ces systèmes ne sont souvent publiquement accessibles que par l'intermédiaire d'une interface Web à l'instar des systèmes ESTExplorer (Nagaraj, Deshpande, *et al.*, 2007), ESTpass (Lee *et al.*, 2007), ESTPiper (Tang *et al.*, 2009) ou PESTAS (Nam *et al.*, 2009). Ce type de systèmes est très avantageux puisqu'il ne nécessite ni le déploiement ni la maintenance de toute une infrastructure matérielle et logicielle souvent lourde et coûteuse. Cependant, même avec un niveau de paramétrage relativement élevé, ces systèmes ne peuvent se permettre de proposer une liste exhaustive de traitements et d'analyses pour construire un protocole de traitement à façon. De plus, il peut subsister un problème de droits quant à la soumission de séquences potentiellement brevetables ou brevetées à un organisme tiers.

Parallèlement, il existe des systèmes publiquement disponibles que l'on peut installer localement. TGICL (Pertea *et al.*, 2003), ESTAP (Mao *et al.*, 2003), PartiGene (Parkinson *et al.*, 2004) ou MAKER (Cantarel *et al.*, 2008) en sont quelques rares exemples. Malheureusement, de tels systèmes nécessitent généralement l'installation d'un grand nombre de logiciels externes qui ne sont pas nécessairement fournis, et qui peuvent requérir des coûts de licences supplémentaires. Enfin, la maîtrise du paramétrage de ces systèmes en vue de leur adaptation aux problèmes considérés demande un investissement non négligeable.

C'est pour ces différentes raisons que nous avons décidé de développer notre propre pipeline de traitement de séquences d'ADNc dans le but premier d'analyser les données issues du projet de séquençage massif du transcriptome d'*Alvinella pompejana*. Il intègre des programmes externes et des modules développés au cours de cette thèse. Le pipeline développé en Tcl/Tk a été intégré à Gscope et utilisé pour le traitement de plusieurs projets ADNc.

Nous allons détailler dans ce chapitre la structure et le fonctionnement de ce pipeline (Figure 39) tout en décrivant les problèmes rencontrés lors de son élaboration, puis nous présenterons les résultats obtenus lors du traitement des séquences d'*A. pompejana*.

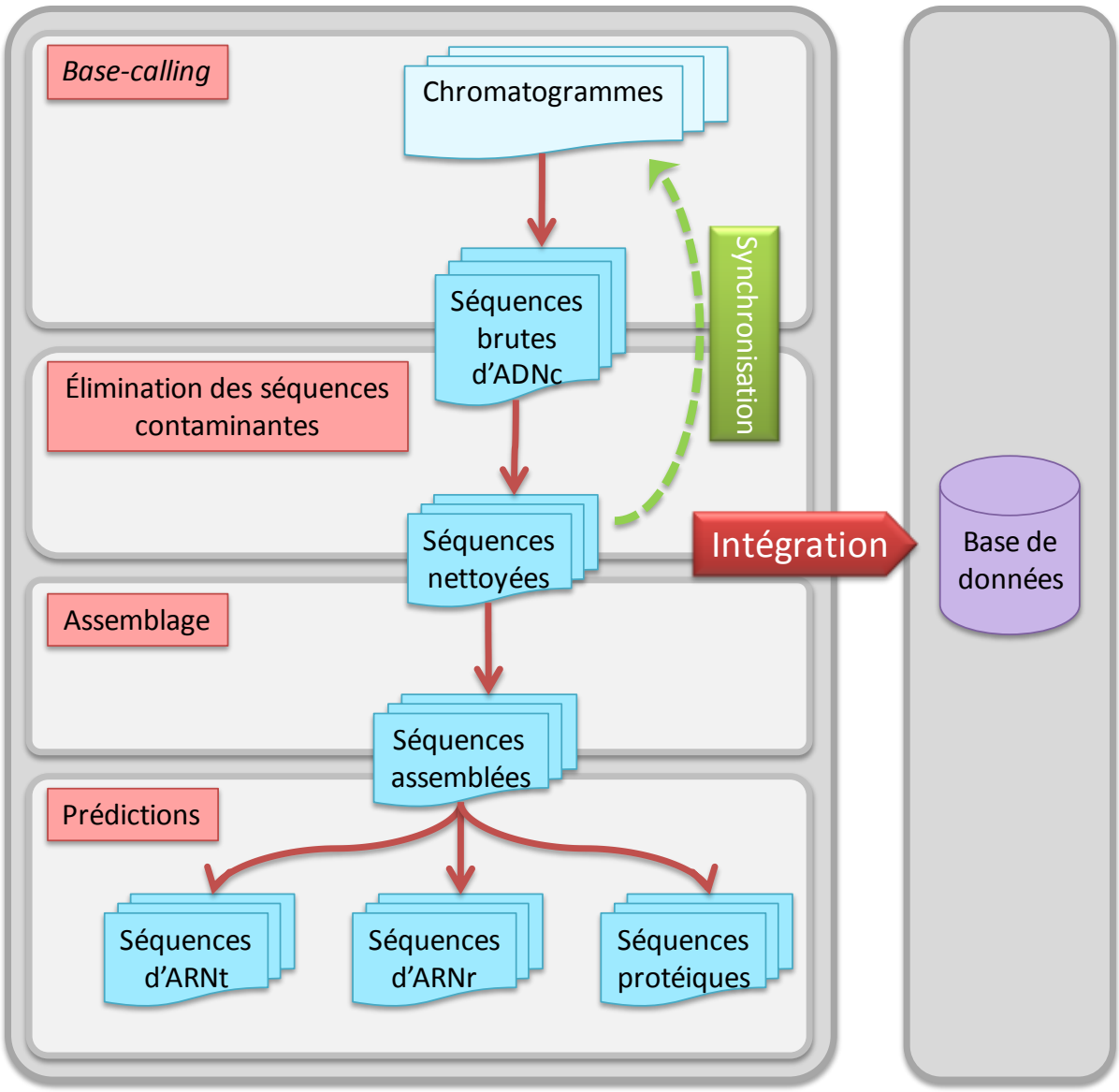


Figure 39 – Schéma général du pipeline de traitement des ADNc

Notre pipeline prend en entrée les chromatogrammes pour les convertir en séquences qui vont être « nettoyées » afin éviter tout problème lors de l'étape d'assemblage. Durant cette étape, la modification de données nécessite une synchronisation en amont. Les séquences assemblées sont ensuite analysées pour y détecter les séquences d'ARNt et d'ARNr et pour prédire des séquences protéiques à partir des séquences restantes. Durant chaque étape, toute information pertinente est intégrée à une base de données dédiée (Cf. Organisation et gestion des données, page 122).

10.1 Conversion des données de séquençage

Cette étape appelée *base-calling* permet de convertir les chromatogrammes, qui sont des fichiers binaires renfermant principalement les profils d'intensités de luminescence

échantillonnés par un séquenceur automatique, en fichiers plats contenant des données de séquences et de valeurs de qualité.

Pour une raison de praticité, le pipeline décrit ici ne gère que les chromatogrammes au format standardisé SCF (Cf. Du besoin de synchroniser les données, page 117). Les chromatogrammes apparaissant dans un format propriétaire devront donc préalablement être convertis dans ce format à l'aide du programme *makeSCF* de la suite Staden.

Le *base-calling* est délégué au programme Phred. Les séquences brutes ainsi obtenues ont une longueur moyenne variant entre 600 et 1 400 pb dans le cas d'un séquençage classique de type Sanger, avec un taux d'erreur de *base-calling* estimé de 1 à 5% (Aaronson *et al.*, 1996; Richterich, 1998).

La qualité des pics n'est pas homogène tout au long d'un chromatogramme. Au début de chaque chromatogramme, les pics d'intensités présentent de nombreux chevauchements (Figure 40a) dus aux propriétés physico-chimiques des marqueurs fluorescents qui influent sur la vitesse de migration des fragments synthétisés de taille très courte. Par la suite, on observe une diminution progressive du signal qui correspond à un épuisement des didésoxyribonucléotides marqués dans le mix de synthèse enzymatique (Figure 40c).

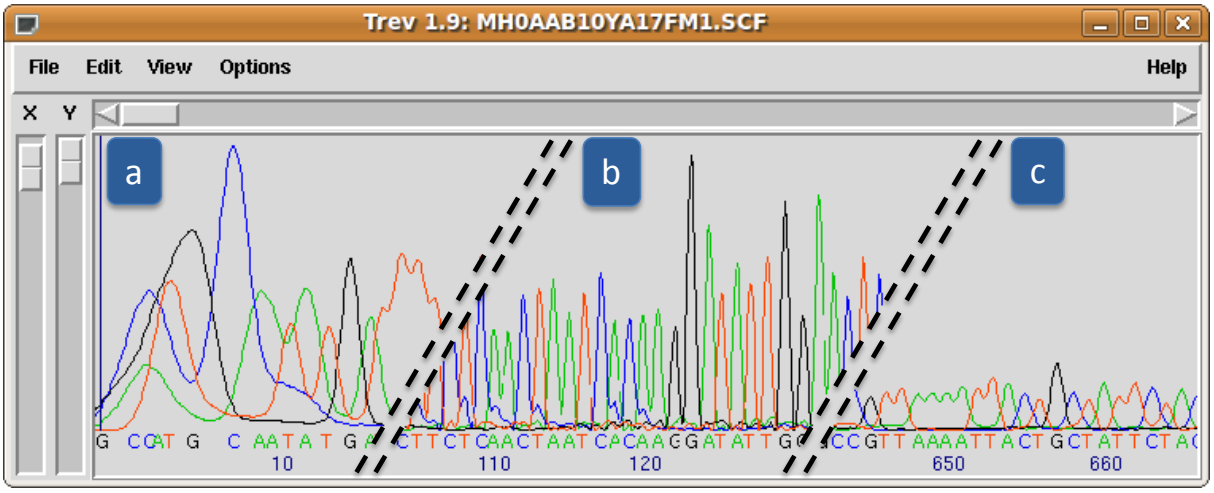


Figure 40 – Visualisation d'un chromatogramme
(a) Début comportant des pics chevauchants (b) Section de bonne qualité (c) Le signal devient faible en fin de chromatogramme. Visualisation sous le logiciel Trev du package Staden.

Ces chevauchements de pics et le signal quasi-nul respectivement en début et en fin de chromatogrammes impliquent des extrémités de séquences brutes de mauvaise qualité informationnelle. Pour éliminer ces extrémités, nous avons utilisé l'option '*-trim_alt*' de Phred (Figure 41). Cette option active un algorithme qui détecte puis élimine le plus grand fragment de séquence à chaque extrémité ayant un taux d'erreur supérieur à un certain seuil. Nous avons utilisé le seuil par défaut qui est fixé à 5 % (valeur de qualité ≤ 13).

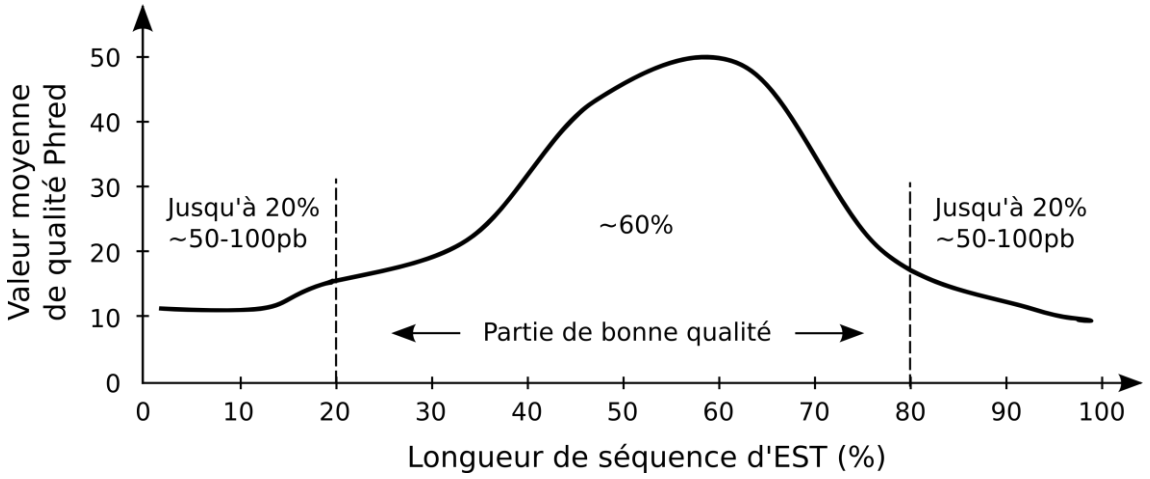


Figure 41 – Valeurs de qualité Phred en fonction de la position sur la séquence brute
Les extrémités de mauvaise qualité ont généralement une valeur de qualité moyenne < 13 Adapté de (Nagaraj, Gasser, *et al.*, 2007)

Pour chaque chromatogramme traité par Phred, on obtient un fichier de type PHD formé d'un entête refermant des informations sur le chromatogramme suivi d'une section de résultats de *base-calling* dont chaque ligne comporte la base attribuée, la valeur de qualité et la position sur le chromatogramme (Figure 42).

```
BEGIN_SEQUENCE MH0AAB12YL19FM1.SCF

BEGIN_COMMENT

CHROMAT_FILE: MH0AAB12YL19FM1.SCF
ABI_THUMBPRINT: 0
PHRED_VERSION: 0.020425.c
CALL_METHOD: phred
QUALITY_LEVELS: 99
TIME: Mon Mar 21 16:39:36 2005
TRACE_ARRAY_MIN_INDEX: 0
TRACE_ARRAY_MAX_INDEX: 16303
TRIM: 21 640 0.0500
TRACE_PEAK_AREA_RATIO: 0.0112
CHEM: term
DYE: big

END_COMMENT

BEGIN_DNA
g 10 13
c 10 23
c 9 34
c 10 45
                .
                .
                .
c 3 16287
END_DNA

END_SEQUENCE
```

Figure 42 – Extrait d'un fichier PHD créé par Phred
Le bloc d'en-tête se situe entre les lignes *BEGIN_COMMENT* et *END_COMMENT*. Les résultats de *base-calling* se situent entre les lignes *BEGIN_DNA* et *END_DNA*. Pour chaque ligne du bloc de *base-calling* figurent le type de base, la valeur de qualité et la position d'échantillonnage sur le chromatogramme.

Tous les fichiers PHD obtenus sont enfin regroupés en deux fichiers de séquences et de valeurs de qualité au format FASTA qui seront utilisables par un programme d'assemblage. Cette étape est réalisée à l'aide du programme phd2fasta livré avec la suite Phred/Phrap/Consed.

10.2 Prétraitements des séquences brutes

L'élimination des extrémités de mauvaise qualité par le logiciel Phred n'est pas suffisante pour obtenir un assemblage de bonne qualité. D'autres facteurs relatifs aux techniques de biologie moléculaire ou à la nature même du matériel séquencé peuvent intervenir et sont traités par le pipeline pour « nettoyer » les séquences brutes.

10.2.1 Élimination des séquences contaminantes

Lors de l'étape de séquençage, des oligonucléotides de séquençage sont choisis pour s'hybrider spécifiquement sur le vecteur de clonage, quelques paires de bases en amont et en aval de l'insert. De ce fait, ces paires de bases sont aussi séquencées et apparaissent aux extrémités des séquences brutes.

De plus, il arrive fréquemment que lors de l'intégration des inserts dans les vecteurs de clonage, des vecteurs se referment sur eux-mêmes pendant l'étape de ligation (clonage traditionnel par enzymes de restriction) ou qu'ils ne subissent pas de recombinaison (technologie de type Gateway). Une fraction des données de séquençage comportera donc la séquence de ces vecteurs vides.

Ces séquences de vecteurs sont détectées, puis masquées, à l'intérieur du fichier de séquences FASTA (les résidus à masquer sont remplacés par des caractères « X ») grâce au logiciel Cross_match qui recherche par similarité les séquences de la banque de vecteurs Univec.

10.2.2 Traitement des séquences masquées

Afin d'éliminer les séquences non informatives ou insuffisamment informatives, c'est-à-dire les séquences de vecteurs vides (totalement masquées par Cross_match) ou les séquences trop courtes (initialement ou après masquage par Cross_match), un filtre sur la longueur des séquences est appliqué. Ce traitement permet d'éviter de conserver des séquences non exploitables qui alourdiraient inutilement le processus d'assemblage.

Ce filtre détecte les segments contigus de résidus non masqués et sélectionne le plus grand pour chaque séquence. Si ce dernier présente une longueur ≥ 100 pb, il sera conservé. Ce seuil nous a paru un bon compromis pour préserver un maximum d'informations tout en évitant de polluer les étapes ultérieures.

10.2.3 Traitement des queues polyadénylées et des régions homopolymériques

Les ARNm, et donc par extension, les ADNc, comprennent naturellement une queue 3' polyadénylée (poly(A)) rajoutée lors de l'étape de maturation des ARN pré-messagers chez les eucaryotes.

Lors de nos premiers tests d'assemblage de séquences, nous avons remarqué que certaines séquences pouvaient comporter une région poly(A) dans leur extrémité 5'. Ces régions 5' poly(A) peuvent être alignées avec la queue 3' poly(A) d'autres séquences lors de l'assemblage, entrainant ainsi l'apparition de contigs chimériques (Figure 43a).

Bien que nous n'ayons reçu que des clones séquencés dans le sens 5'→3', quelques séquences ont été, à juste raison, complémentées puis alignées contre plusieurs autres séquences. L'existence de ces séquences peut être expliquée par un problème lors de l'orientation de l'insert dans le vecteur pendant l'étape de clonage. Or, certaines de ces séquences comportaient initialement une queue 3' poly(T) qui a été transformée par complémentarité en région 5' poly(A), entrainant le même cas que précédemment.

Enfin, est aussi apparue l'existence de séquences chimériques comportant une région poly(A) en leur milieu, et pouvant provenir de la recombinaison de deux inserts dans un même vecteur pendant l'étape de clonage. Ces séquences ont, elles aussi, abouti à la formation de contigs chimériques (Figure 43b).

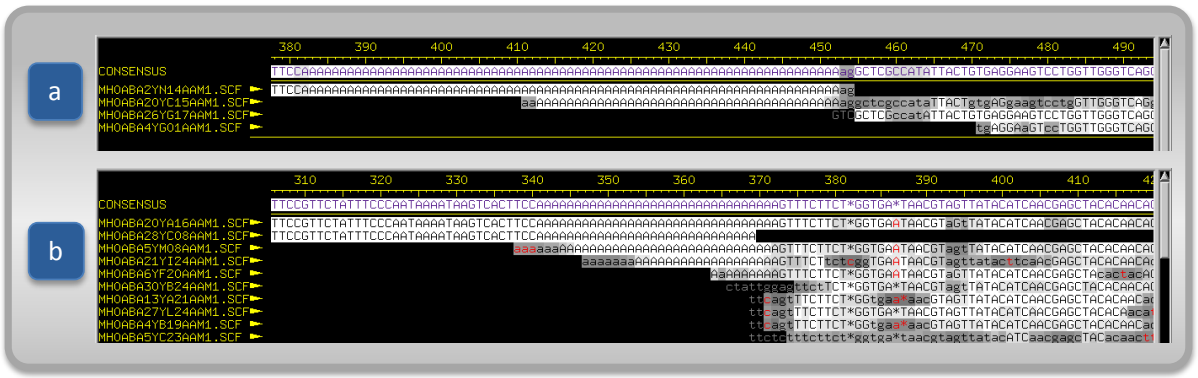


Figure 43 – Problèmes d'assemblage avant nettoyage des séquences brutes
(a) Contig chimérique provoqué par l'alignement d'une région 5' poly5(A) avec une queue poly(A) d'une autre séquence (b) Contig chimérique formé autour d'une séquence chimérique.

Pour éviter l'apparition de ce type de problèmes lors de l'assemblage, les séquences subissent un traitement en quatre étapes (Figure 44) :

- Les régions poly(A) et poly(T) sont détectées sur chaque séquence à l'aide d'une méthode à fenêtre glissante de 20/25 pb. En d'autres termes, les résidus 1 à 25 de la séquence sont marqués comme poly(x) si au moins 20 résidus sont un nucléotide 'x'. On continue de la même manière avec les résidus de 2 à 26, puis 3 à 27, et ainsi de suite.

- Chaque séquence est décomposée arbitrairement en 3 zones. Les 50 premiers résidus forment la zone 5', les 50 derniers résidus la zone 3' et les résidus restants la zone centrale. Les séquences contenant une région poly(A) localisée exclusivement dans leur zone centrale sont considérées comme des séquences chimériques et seront exclues de l'assemblage.

- Les régions poly(A) dont au moins un résidu est localisé dans la zone 5' sont masquées. De même, les régions poly(T) dont au moins un résidu est localisé dans la

zone 3' sont considérées comme une probable queue 3' poly(A) après complémentation et sont masquées.

- Les résidus situés en aval des régions 3' poly(A) sont éliminés pour obtenir des queues 3' poly(A) bien nettes. La même opération est réalisée pour les résidus en amont des régions 5' poly(T) dans le cas où la séquence doit être complémentée lors de l'assemblage.

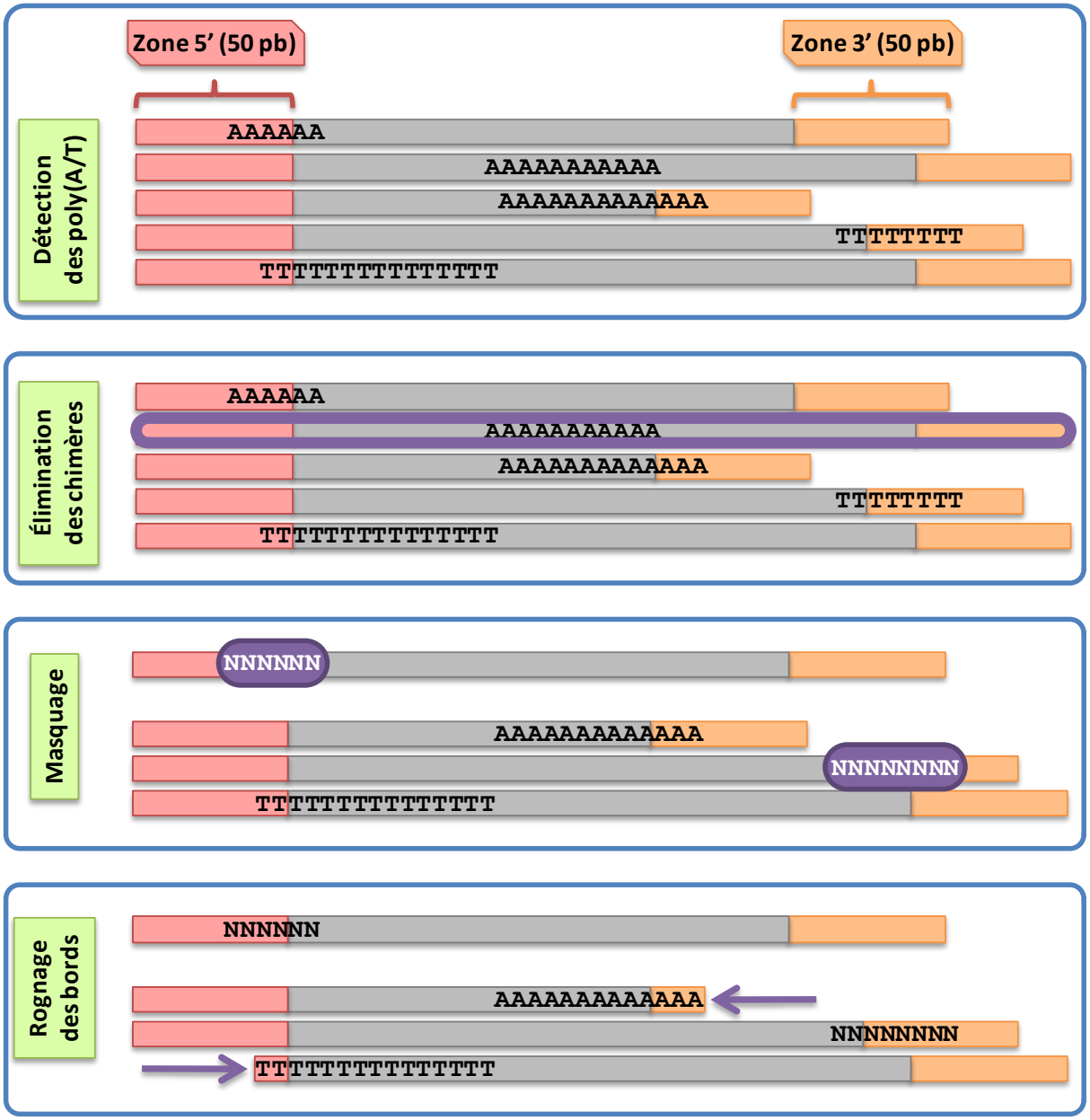


Figure 44 – Les 4 étapes du processus de nettoyage des séquences brutes d'ADNc

A la fin de cette étape, les séquences sont considérées comme « propres » et prêtes à être assemblées.

10.3 Du besoin de synchroniser les données

Tous les traitements précédents sont appliqués au fichier FASTA de séquences. Dès lors, les modifications effectuées entrainent une désynchronisation entre le fichier séquence et le fichier FASTA des valeurs de qualité (décalage de valeurs, valeurs superflues). Il en est de même, pour les informations de séquence et qualité contenues dans chaque fichier PHD issu de Phred, ainsi que pour les informations de séquence que peuvent contenir les chromatogrammes au format SCF.

Afin d'assurer l'homogénéité des données indispensable à l'assemblage et à la visualisation simultanée des séquences et des chromatogrammes (Cf. Visualisation de chromatogramme, page 165), nous répercutons en cascade chaque modification effectuée sur le fichier de séquence. Cette répercussion s'effectue dans l'ordre fichier de séquences→fichier de qualité→fichier PHD→chromatogramme SCF.

10.4 Assemblage

Le but de l'assemblage est d'augmenter la qualité et la taille des séquences en les regroupant par similarité puis en les alignant. Dans le cas précis des banques d'ADNc, l'assemblage permet également de réduire la redondance des transcrits d'un même gène. En effet, dans une banque non normalisée, un produit de gène surexprimé peut être cloné et séquencé de très nombreuses fois. Cela a été particulièrement notable dans le cas de l'analyse des banques d'ADNc d'*A. pompejana*, qui comportent un très grand nombre de transcrits de gènes liés au stress oxydatif (Cf. Publication 1, page 153).

Initialement, nous avons testé le logiciel Phrap (Phil Green, non publié) afin de réaliser nos assemblages. Malheureusement, la qualité des séquences consensus produites était relativement mauvaise. Après quelques investigations, il s'est avéré que Phrap utilise un algorithme de construction de consensus par mosaïque, c'est-à-dire que pour chaque colonne des alignements multiples produits, la base consensuelle choisie correspondra à la base de la colonne ayant la meilleure valeur de qualité, quelque soit le nombre de bases différentes en désaccord avec ce choix.

Notre choix sur le programme d'assemblage s'est alors porté sur Cap3, qui est une des références les plus utilisées dans ce domaine. Cap3 se démarque des autres ténors du genre par sa sensibilité à regrouper correctement les transcrits d'un même gène et à la fidélité des séquences consensus qu'il produit (Liang *et al.*, 2000).

En utilisant les paramètres par défaut, la qualité des séquences consensus était correcte mais celle des contigs n'était pas assez satisfaisante. En effet, de longues portions aux extrémités de certaines séquences n'étaient pas alignées et certaines parties alignées comportaient trop de désaccords. Afin d'obtenir des contigs de haute qualité, tout en évitant le regroupement de variants d'épissage ou des transcrits de différents paralogues,

des paramètres d'assemblage relativement stricts ont été fixés en augmentant la stringence des critères suivants :

- Les régions chevauchantes doivent comporter au moins 90 % d'identité (contre 80 % par défaut)
- Un maximum de 30 pb peuvent être ignorées à chaque extrémité de séquence lors de l'alignement (contre 250 pb par défaut)

A la suite de l'assemblage, les séquences consensus des contigs (alignements de séquences regroupées) et les singletons (séquences restées isolées) sont analysées par la dernière partie du pipeline.

10.5 Analyses des séquences d'ADNc

Les séquences issues de l'étape d'assemblage doivent être caractérisées afin de localiser la région codante et les extrémités 5' et 3' UTR, la région codante étant ensuite traduite en séquence protéique. Il convient au préalable de s'assurer qu'il s'agit effectivement bien de séquences d'ADNc et non de séquences provenant d'ARN non-codants.

10.5.1 Détection d'ARN non-codants

La recherche d'ARN non-codants a été limitée aux ARNt et ARNr. Bien que la présence d'autres types d'ARN ne soit pas à écarter, leur prédiction est relativement difficile et cette recherche n'a pas été intégrée à notre protocole automatique.

10.5.1.1 Détection des ARNt

La recherche des séquences d'ARNt est accomplie à l'aide du logiciel tRNAscan-SE. Ce programme réalise des prédictions fiables et les séquences détectées sont annotées en tant qu'ARNt et ne seront donc plus traitées par la suite.

10.5.1.2 Détection des ARNr

La famille des gènes des ARN ribosomaux figure comme la famille la plus conservée en termes de séquence et de structure, de par les très fortes contraintes imposées en tant que composants principaux de la structure, mais aussi de la fonction, du ribosome. Cette conservation caractéristique, combinée à sa nature ubiquitaire l'a promue au rang de standard *de facto* en phylogénétique (Woese, 2000; Koonin, 2003).

Grâce à cette conservation, il est relativement aisé d'identifier des séquences d'ARNr par une simple recherche par similarité. Cependant, afin d'améliorer la spécificité des résultats, une banque BLAST dédiée est créée à partir d'un jeu de séquences extrait de GenBank à l'aide de requêtes SRS. Ce jeu est composé de l'ensemble des séquences nucléiques appartenant au même embranchement que l'organisme étudié. Dans le cas d'*A. pompejana*,

l'embranchement des Annélides ne comportant pas encore de génome complet, le jeu de données a été complété par les séquences de Mollusques, taxon proche de celui des Annélides.

La détection s'effectue sur la base d'une recherche BlastN (recherche BLAST d'une séquence nucléique à l'intérieur d'une banque de séquences nucléiques) filtrée pour ne garder que les résultats avec une E-value ≤ 1^{e-10}, un alignement ≥ 100 pb avec un pourcentage d'identité ≥ 50 %. Si la définition d'au moins un des résultats filtrés comporte les termes 'rRNA' ou 'ribosomal RNA', ainsi que le type de sous-unité (sous-unités 5S, 23S et 16S procaryotes ou 5S, 5.8S, 28S et 18S eucaryotes), la séquence d'intérêt est marquée en tant qu'ARNr du type adéquat.

10.5.2 Prédiction de séquences codantes

La localisation des éléments génétiques présents sur un ADNc (Cf. Prédiction de régions codantes dans les transcrits, page 36) consiste essentiellement, en pratique, à détecter la région codante. Cette détection est évidemment grandement facilitée par rapport à une recherche de région codante sur une séquence génomique eucaryote, compte tenu de l'absence de régions introniques.

Pour ce faire, nous avons combiné une approche originale de prédiction par similarité à une approche de prédiction *ab initio*. Cette approche combinée offre l'avantage d'améliorer le nombre de séquences codantes détectées, en particulier dans le cas d'organismes appartenant à des taxons encore peu représentés dans les banques, comme *A. pompejana*. En effet, la prédiction par similarité ne nécessite pas de connaissances préalables sur les séquences codantes et les séquences UTR de l'organisme considéré, contrairement aux méthodes *ab initio* qui s'appuient sur un jeu d'apprentissage. Les séquences caractérisées par similarité peuvent être utilisées, le cas échéant, comme jeu d'apprentissage par les méthodes *ab initio* et ainsi permettre de détecter des séquences codantes ne présentant pas de similarité avec les séquences protéiques répertoriées dans les banques.

10.5.2.1 Détection des séquences codantes par similarité

Cette méthode se base sur la recherche de similarité entre les séquences d'ADNc (après traduction dans les six cadres de lecture) et une banque de séquences protéiques. Le principe est d'identifier une (ou plusieurs) zone(s) de la séquence présentant une conservation significative avec une protéine connue et d'en déduire le cadre de lecture (ou les cadres dans le cas d'erreurs de séquençage aboutissant à des décalages). La zone identifiée est alors prolongée en amont et en aval pour tenter d'obtenir la séquence codante complète. Nous allons maintenant détailler le principe de cette méthode qui est illustrée par la Figure 45.

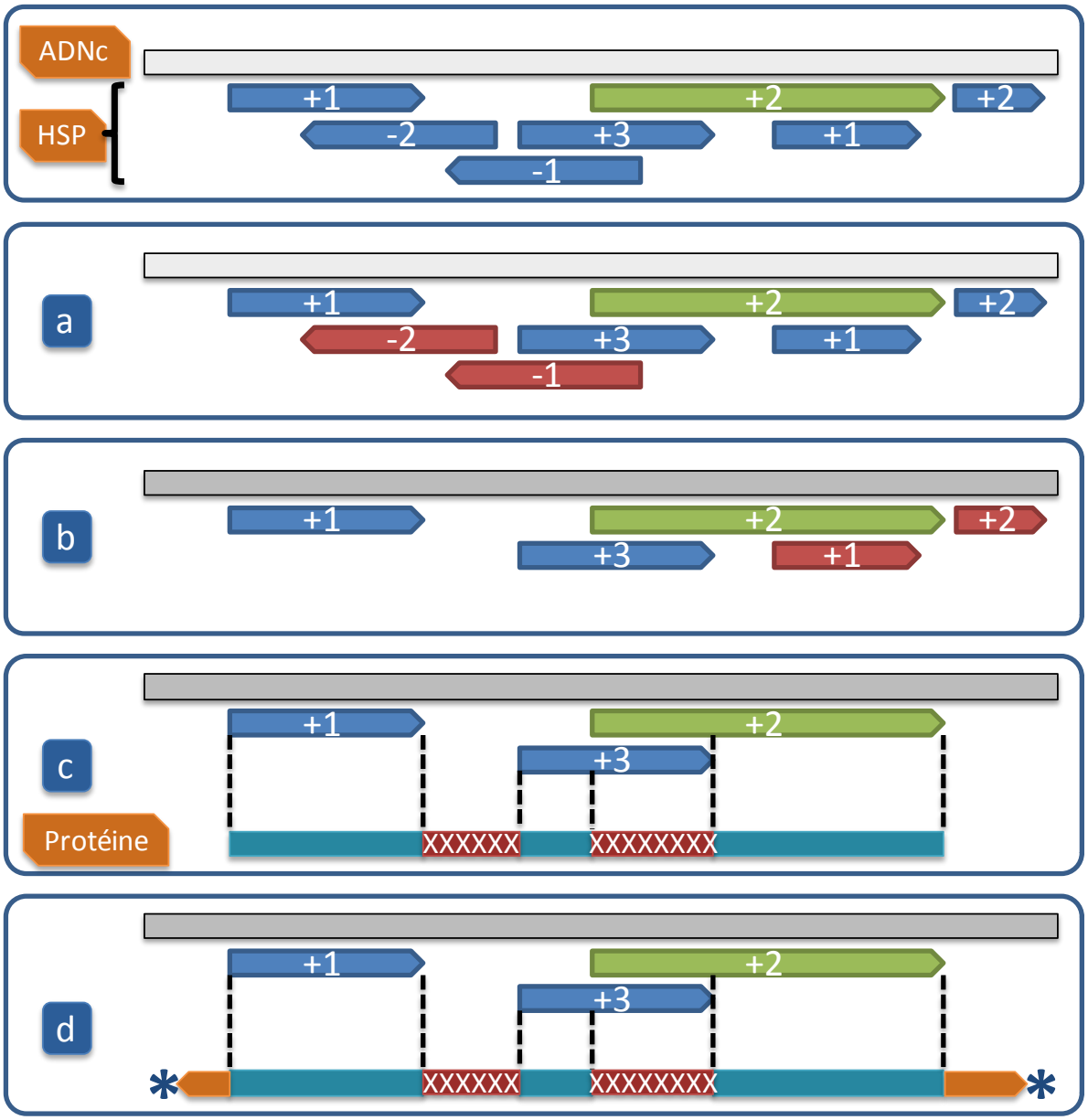


Figure 45 – Prédiction de séquence protéique à partir d'un ADNc et d'un BlastX
Le meilleur HSP du premier hit du BlastX apparaît en vert. (a) Élimination des HSP alignés sur le brin opposé au meilleur HSP (b) Élimination des HSP totalement recouverts ou suivant un HSP dans le même cadre de lecture (c) Traduction des parties non chevauchantes des HSP et complétion à l'aide de segments de 'X' (d) Extension des extrémités jusqu'à atteindre un codon STOP ou la fin de l'ADNc.

Pour chaque séquence à analyser, une recherche BlastX est effectuée sur la banque UniProtKB et seul le premier '*hit*' est considéré, à condition d'avoir obtenu une *E-value* ≤ 1^{e-3}. Le premier alignement deux-à-deux (HSP, *High-scoring Segment Pairs*), qui constitue le meilleur des HSP du premier hit (en termes de score, et généralement de longueur), sert de référence pour la construction de la séquence protéique :

- détermination de l'orientation : le sens du cadre de lecture dans lequel a été traduite la séquence nucléique de référence (« brin + » ou « brin - ») va définir l'orientation

dans laquelle va se faire la traduction protéique. Les HSP n'ayant pas la même orientation sont ignorées pour la suite du traitement.

- détermination du code génétique : les protéines mitochondriales sont traduites sur la base d'un code génétique différent du code standard. Afin de détecter l'origine de la séquence et d'adapter le code génétique en conséquence, la présence de codons stop alignés contre des résidus tryptophane est recherchée. En effet, le codon stop UGA « opale » du code génétique standard code pour le résidu tryptophane dans le code génétique mitochondrial.

Les HSP dans l'orientation définie ci-dessus sont triés par « ordre croissant » à l'aide des coordonnées de la séquence de référence. Cet ordre est évidemment inversé lorsque la traduction sera effectuée sur le brin -. Ces HSP sont ensuite parcourus et filtrés afin d'éliminer :

- les HSP dont les coordonnées sont totalement incluses dans d'autres HSP (totalement recouverts par des HSP plus longs). Ceci permet de favoriser les HSP les plus longs.
- les HSP successifs ayant le même cadre de lecture.

Une fois les HSP filtrés, un ensemble de séquences peptiques est traduit à partir de la séquence nucléique de référence grâce aux coordonnées et aux cadres de lecture extraits de ces HSP. La séquence protéique est créée en « assemblant » ces séquences peptidiques de telle sorte que :

- les segments de séquences peptidiques se chevauchant sont remplacés par un masque de taille équivalente.
- les segments utilisés pour faire la jonction entre plusieurs peptides sont composés d'une séquence de caractères 'X' de taille *ad hoc*.

Les extrémités N-terminales et C-terminales de la protéine nouvellement créée sont ensuite allongées en traduisant la séquence nucléique de référence jusqu'à parvenir à un codon STOP ou à la fin de la séquence nucléique.

Enfin, si un codon stop est trouvé lors de l'élongation de l'extrémité N-terminale, on peut supposer que la traduction est allée en amont du codon initiateur. Dans ce cas, l'extrémité N-terminale est corrigée par coupure au niveau du premier résidu méthionine trouvé (correspondant probablement au codon initiateur), le cas échéant.

10.5.2.2 Prédiction *ab initio*

Cette prédiction est réalisée uniquement sur les séquences d'ADNc qui n'ont pas abouti à la création d'une séquence protéique lors de l'étape précédente, suite à une similarité trop faible ou inexistante avec les séquences de la banque UniProtKB.

La prédiction *ab initio* des régions codantes est réalisée à l'aide du modèle de Markov du logiciel ESTScan2 (Lottaz *et al.*, 2003). Nous avons choisi ce logiciel pour sa capacité à corriger les *frameshifts* dus aux erreurs de séquençage, tout en sachant que nous n'avons pas pu obtenir facilement des copies des logiciels DIANA-EST (Hatzigeorgiou *et al.*, 2001) et DECODER (Fukunishi *et al.*, 2001) qui présentent des fonctionnalités équivalentes.

Chaque modèle HMM de ESTScan2 nécessite d'être paramétré par une matrice de transitions qui définit les probabilités de passer d'un état du modèle à un autre. L'utilisateur du pipeline doit donc choisir la matrice adaptée à l'organisme étudié parmi les matrices prédéfinies. Si aucune matrice n'est adaptée, un jeu de données d'entrainement doit être constitué pour le programme '*build_model*' fourni avec ESTScan2. Ce programme nécessite en entrée un fichier au format FASTA de séquences d'EST dont chaque ligne d'en-tête comporte les coordonnées de début et de fin de la région codante. Il en résulte une matrice de transitions directement utilisable pour prédire des régions codantes avec ESTScan2. Nous reviendrons ultérieurement sur la matrice utilisée dans le cadre de l'étude d'*A. pompejana* (Cf. Création du modèle pour *Alvinella pompejana*, page 129).

10.6 Organisation et gestion des données

Les données brutes de séquençage et les données générées par le pipeline décrit ci-dessus peuvent représenter des volumes considérables, qu'il est nécessaire d'organiser et de gérer de manière rationnelle pour une exploitation optimale. La gestion des données est d'autant plus cruciale que ces dernières arrivent de manière progressive. En effet, lors de projets de séquençage de grande envergure, les séquences sont généralement obtenues par lots, séparés dans le temps pour des raisons liées aussi bien aux disponibilités et priorités du centre de séquençage qu'aux différentes banques étudiées qui peuvent être de qualités ou d'origines tissulaires très diverses. Cette stratégie par étapes permet d'obtenir un aperçu de la banque de clones étudiée, de sa qualité et de la qualité du séquençage qui en résultera, et permet d'orienter le programme de séquençage en fonction de la nature du projet (génome, ESTs…), de la couverture désirée, de la diversité des séquences des banques disponibles…

Cela a été le cas pour les différentes banques de clones d'*A. pompejana* qui ont chacune été séquencées en plusieurs lots par le Genoscope.

L'arrivée séquentielle de données a donc nécessité une organisation qui permette non seulement l'analyse, puis l'annotation de chaque banque considérée isolément, mais également l'intégration et l'analyse globale de toutes les séquences dans un projet commun préservant l'origine et les informations liées à chaque banque, chaque lot et chaque séquence. Pour cela, nous avons utilisé :

- l'infrastructure GScope : chaque lot de séquences constitue un nouveau projet GScope, permettant ainsi de bénéficier de l'infrastructure GScope dans le cadre de

l'analyse des données. L'ensemble des séquences est réuni dans un projet GScope global. Des projets GScope complémentaires, réunissant par exemple l'ensemble des séquences issues d'une même banque, peuvent être créés.

- une base de données relationnelle permettant le suivi et l'intégration de ces données.

10.6.1 Projets Gscope

Les projets Gscope ont été structurés en respectant l'organisation classique de Gscope (Cf. Architecture générale, page 99). La structure minimale a cependant été complétée pour tenir compte des spécificités d'un projet d'ADNc. Nous avons en particulier veillé à ce que les résultats d'assemblage soient visualisables à l'aide du logiciel Consed, permettant ainsi une vérification aisée de chaque assemblage en cas de mise en évidence d'un problème à ce niveau.

Lors de son lancement, Consed vérifie l'existence de la présence de trois répertoires nommés « chromat_dir », « phd_dir » et « edit_dir » contenant respectivement les chromatogrammes, les fichiers de sortie de Phred et les résultats d'assemblage. Ces trois répertoires ont donc été créés avec les chromatogrammes placés à l'endroit adéquat (Figure 46).

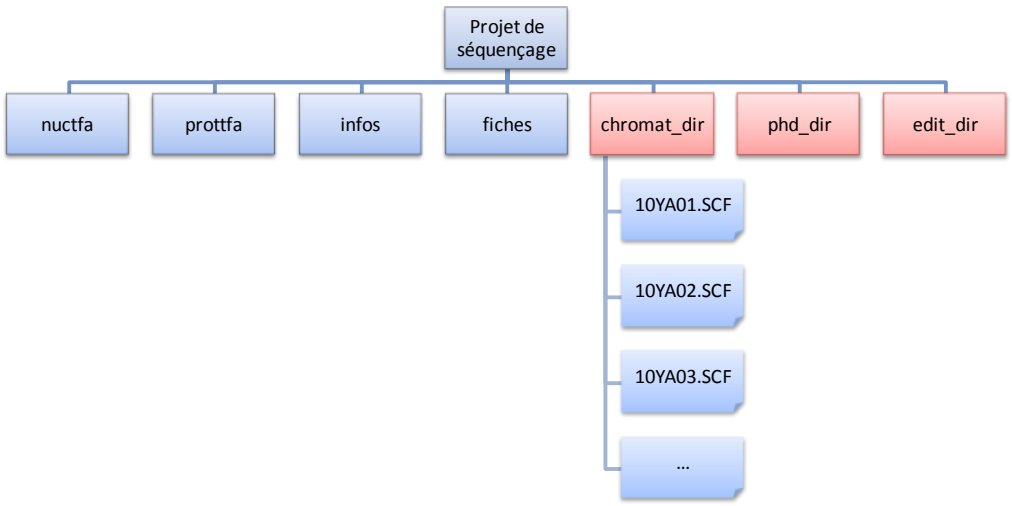


Figure 46 – Structure d'un projet de séquençage Gscope
Les répertoires spécifiques à ce type de projet apparaissent en rouge.

Une fois l'assemblage terminé, le pipeline ajoute les séquences assemblées (les séquences consensus des contigs et les singletons) au projet afin d'obtenir un numéro d'accès. Enfin, la prédiction de séquences d'ARNt, d'ARNr et de protéines est lancée sur chaque séquence nouvellement ajoutée (Figure 47).

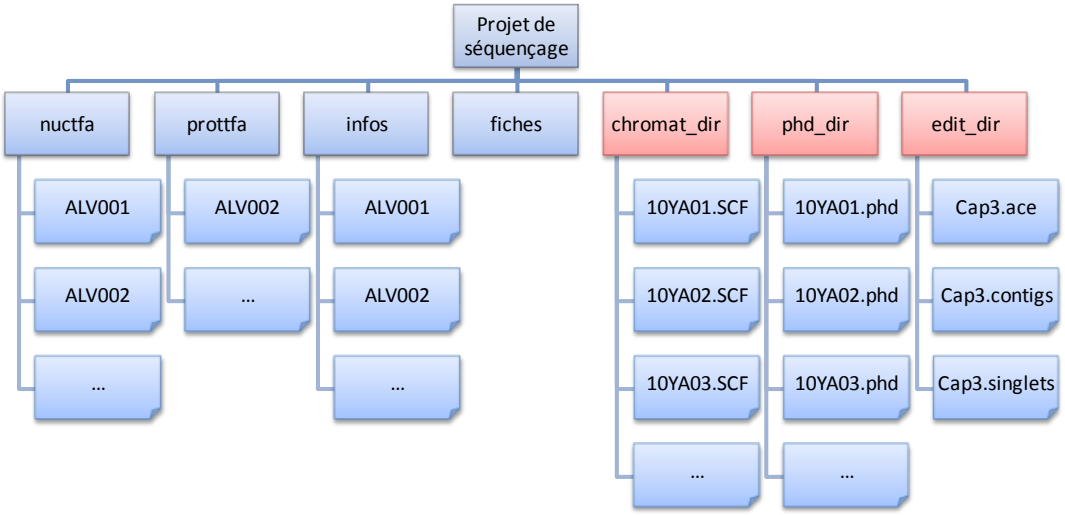


Figure 47 – Structure d'un projet de séquençage à la fin du pipeline d'analyse
Les fichiers de *base-calling* et d'assemblage sont maintenant présents. Les séquences assemblées ont été créées sous forme de séquences nucléiques Gscope. Les informations de prédictions d'ARNt et d'ARNr figurent dans les « infos ». Les séquences protéiques prédites sont regroupées dans « prottfa ».

La fusion de projets Gscope de séquençage a été rendue possible afin de regrouper différents lots dans un même projet pour assembler et analyser conjointement les séquences provenant d'une même banque de clones. De même, une fusion de l'intégralité des lots de séquençage a été réalisée afin de mener une étude globale sur les séquences d'*A. pompejana* (Cf. Nettoyage et assemblage, page 127).

10.6.2 Base de données d'assemblage

La répartition des données de séquençage à l'intérieur d'une multitude de projets Gscope peut devenir rapidement lourde à gérer. De plus, chaque réassemblage engendre obligatoirement un réarrangement des ADNc et donc la perte d'une partie du travail d'annotation qui a été réalisé en aval d'un précédent assemblage. En effet, un ajout de séquences (ou un retrait) peut entrainer la fusion ou la séparation d'anciens contigs, ou permettre l'assemblage d'anciens singletons. Les séquences assemblées ne seront donc plus forcément identiques aux précédentes, sans compter la présence de nouvelles séquences.

Pour répondre à ces contraintes, nous avons entrepris de construire une base de données relationnelle permettant :

- D'intégrer l'ensemble des projets de séquençage,
- De maintenir le lien entre les séquences d'ADNc et leur banque d'origine,
- De conserver les paramètres d'assemblage,

- De maintenir l'historique des différents assemblages d'une même banque de clones,
- De permettre l'intégration de l'annotation des séquences protéiques (Cf. Annotation intégrative automatique de séquences protéiques, page 135).

Cette base de données a été réalisée sous le système de base de données PostgreSQL et est représentée dans le diagramme entité-relation suivant (Figure 48).

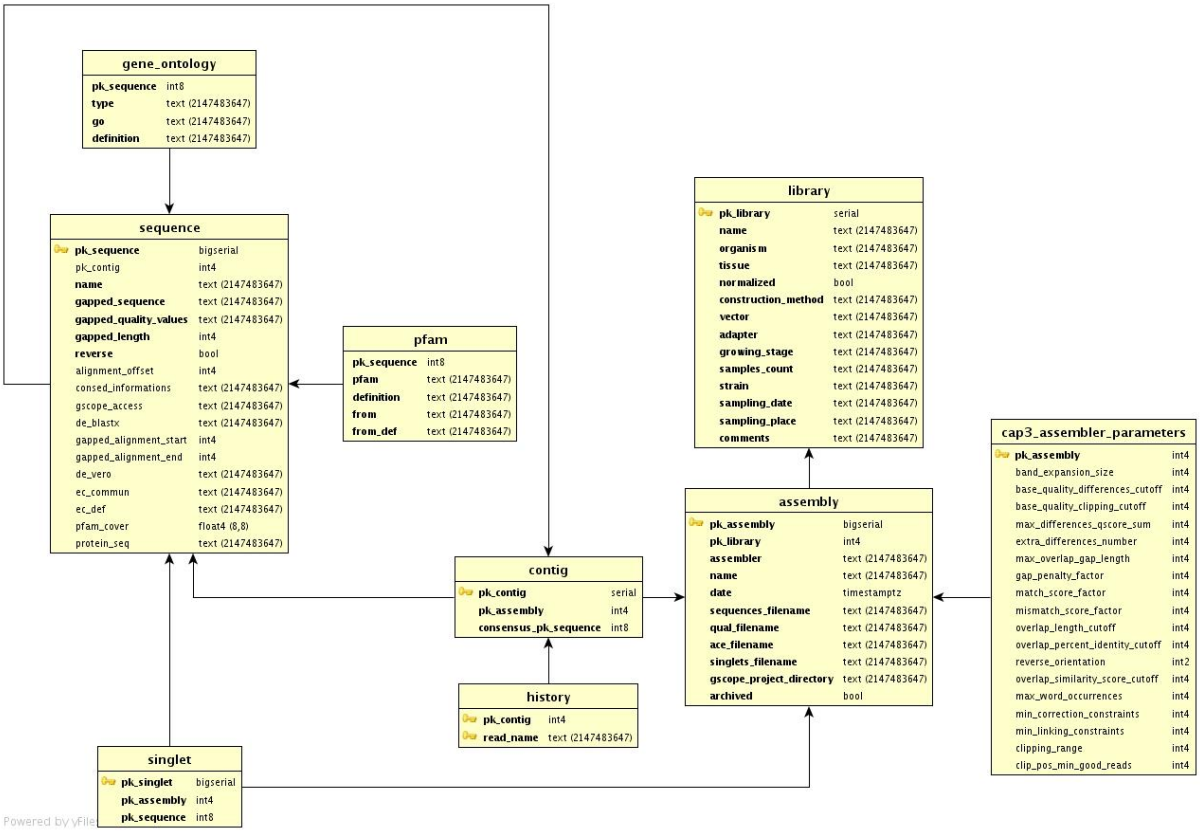


Figure 48 – Diagramme ER de la base de données relationnelle de données d'assemblage
Diagramme réalisé à l'aide de DbVisualizer (http://www.dbvis.com/).

Cette base de données est actuellement relativement simple et comporte deux tables principales :

1. la table « assembly » qui regroupe ou fait le lien entre les informations sur le projet Gscope, les fichiers d'entrée et de sortie, les paramètres d'assemblage ainsi que la banque sur laquelle l'assemblage a été réalisé,
2. la table « sequence » qui regroupe les informations de séquence d'ADNc dans le contexte d'un contig, c'est-à-dire alignée et accompagnée de ses valeurs de qualité. Elle permet de stocker les séquences consensus de contigs, les séquences composant les contigs, ainsi que les singletons. Cette table comporte également une partie des informations d'annotations.

La table « history » sert à l'archivage d'assemblage lorsque l'on souhaite économiser de l'espace disque. Les informations relatives aux séquences appartenant à un contig sont alors déplacées de la table « sequence » vers la table « history » où ne sont conservés que les noms de séquence. Les séquences de consensus et de singletons ne sont pas concernées par cet archivage.

10.6.2.1 Intégration des données

Pour permettre l'intégration des données générées par le pipeline dans la base de données d'assemblage, un module de remplissage a été développé pour Gscope. Ce module est constitué de deux interfaces.

La première permet d'ajouter et de modifier une banque de clones dans la base. La seconde permet le remplissage effectif de la base en intégrant un projet Gscope après avoir fourni quelques informations concernant l'assemblage (Figure 49).

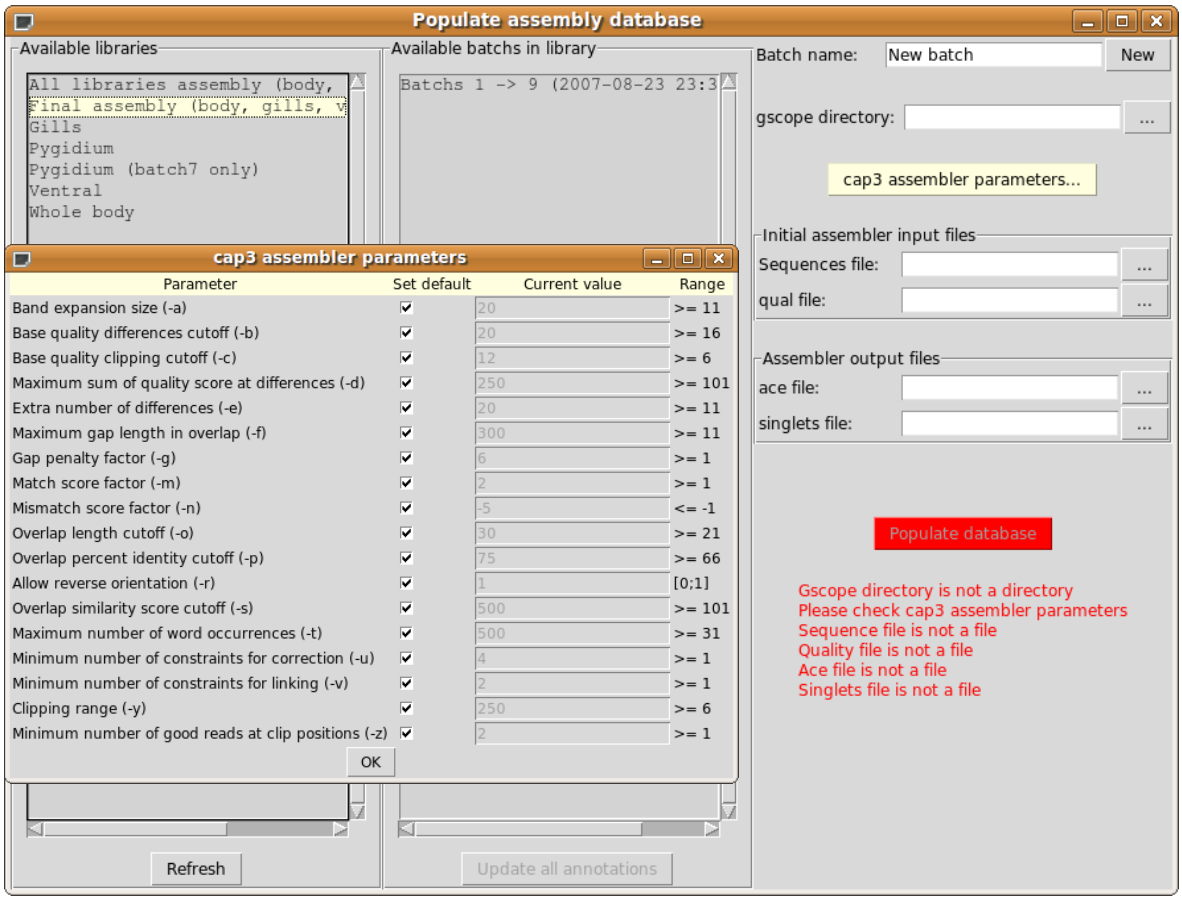


Figure 49 – Interface Gscope de remplissage de la base de données d'assemblage

La fenêtre principale comporte la liste des différentes librairies (à gauche) et la liste des différents assemblages effectués pour chaque librairie (au milieu). Le formulaire de droite permet d'intégrer facilement des données en sélectionnant quatre différents fichiers. Avant toute intégration, la validité des fichiers est vérifiée et tout problème est signalé en rouge. La fenêtre de paramètres d'assemblage permet de spécifier ceux qui ont été effectivement utilisés.

10.7 Analyse des banques d'ADNc d'*A. pompejana*

Le pipeline de traitements et le système de gestion des données présentés ont été utilisés pour analyser quatre banques non normalisées d'ADNc dans le cadre du projet *Alvinella*. Ces banques, construites et séquencées au Genoscope, recouvrent des réalités biologiques très différentes :

- une première banque a été réalisée à partir de plusieurs individus entiers. Le tégument dorsal colonisé par des épibiontes bactériens a été préalablement retiré afin d'éviter les contaminations.
- trois autres banques ont été réalisées à partir d'un seul type de tissu, à savoir les branchies, le tissu ventral et le pygidium.

10.7.1 Nettoyage et assemblage

Le pipeline de traitement des données de séquençage a été utilisé pour nettoyer et assembler un total de 100 177 séquences brutes. Ces traitements ont été appliqués pour chaque banque séparément et un assemblage global de l'ensemble des séquences a été réalisé (Tableau 10).

Tableau 10 – Bilan du nettoyage et de l'assemblage des banques d'ADNc d'*A. pompejana*

Banques	CloneMiner		Oligo-capping		Assemblage global
	Animal entier	Pygidium	Tissu ventral	Branchies	
Chromatogrammes	20 549	36 648	16 411	26 569	100 177
Séquences nettoyées	19 739 (96%)	25 419 (69%)	12 871 (78%)	18 105 (68%)	76 134 (76%)
Nombre de 3' poly(A) (%)	1 599 (8%)	3 156 (12%)	5 465 (43%)	1 467 (8%)	11 687 (15%)
Longueur moyenne (pb)	633	610	720	776	674
Assemblage					
Contigs	1 365	2 327	917	1 193	4 993
Singletons	4 060	6 355	1 914	2 567	10 865
Longueur moyenne des contigs (pb)	993	951	852	931	1 017
Redondance (%)	73	66	78	79	79

Les traitements et assemblages séparés nous ont permis de dégager certaines spécificités des banques ou des méthodes de clonage. Ainsi, il apparaît que la méthode de clonage CloneMiner sélectionne préférentiellement de plus petit transcrits que la méthode par Oligo-capping. Le fort taux de queues poly(A) de la banque de tissu ventral, couplé à la méthode par Oligo-capping qui sélectionne préférentiellement les transcrits par leur coiffe et leur queue poly(A), semble indiquer que les transcrits de cette banque sont en moyenne plus courts que les autres banques ou qu'ils sont dans un état plus dégradé. Enfin, le faible taux de redondance (Figure 50) en séquences de la banque provenant du pygidium en

regard de son grand nombre de séquences, suggère que ce tissu exprime une grande diversité de transcrits.

$$Redondance = 1 - \frac{Nb.\ de\ contig\ s + Nb.\ de\ single\ tons}{Nb.\ de\ séquence\ \ nettoyées}$$

Figure 50 – Formule de calcul de la redondance d'une banque

L'assemblage global à partir de la totalité des 76 134 séquences nettoyées à permis obtenir 4 993 contigs et 10 865 singletons pour un total de 15 858 séquences qui ont ensuite été analysées par les autres modules du pipeline.

10.7.2 Caractérisation des séquences d'ADNc

Parmi les 15 858 séquences issues de l'assemblage, 6 ont été détectées en tant qu'ARNt par tRNAscan-SE et 66 autres ont été détectées en tant que séquences d'ARNr par notre méthode avec similarité. Le reste des séquences a été soumis à notre approche combinée de recherche de séquences codantes.

10.7.2.1 Prédiction des séquences codantes par similarité

La prédiction de séquences protéiques par similarité a permis de créer 7 353 séquences. Elle n'a donc été applicable qu'à la moitié des séquences issues de l'assemblage. Ce faible ratio peut s'expliquer par :

- la faible représentation voire la quasi-absence de séquences d'Annélides et de Mollusques (taxon très proche) dans les banques protéiques. Les protéines spécifiques à ces organismes ne peuvent être détectées par cette première méthode.
- le caractère incomplet de nombreuses séquences d'ADNc. Certaines séquences d'ADNc incomplètes peuvent correspondre uniquement ou essentiellement à une extrémité UTR.

Parmi ces 7 353 séquences protéiques, 2 507 sont prédites complètes (Figure 51), c'est-à-dire comportant un codon initiateur et un codon stop.

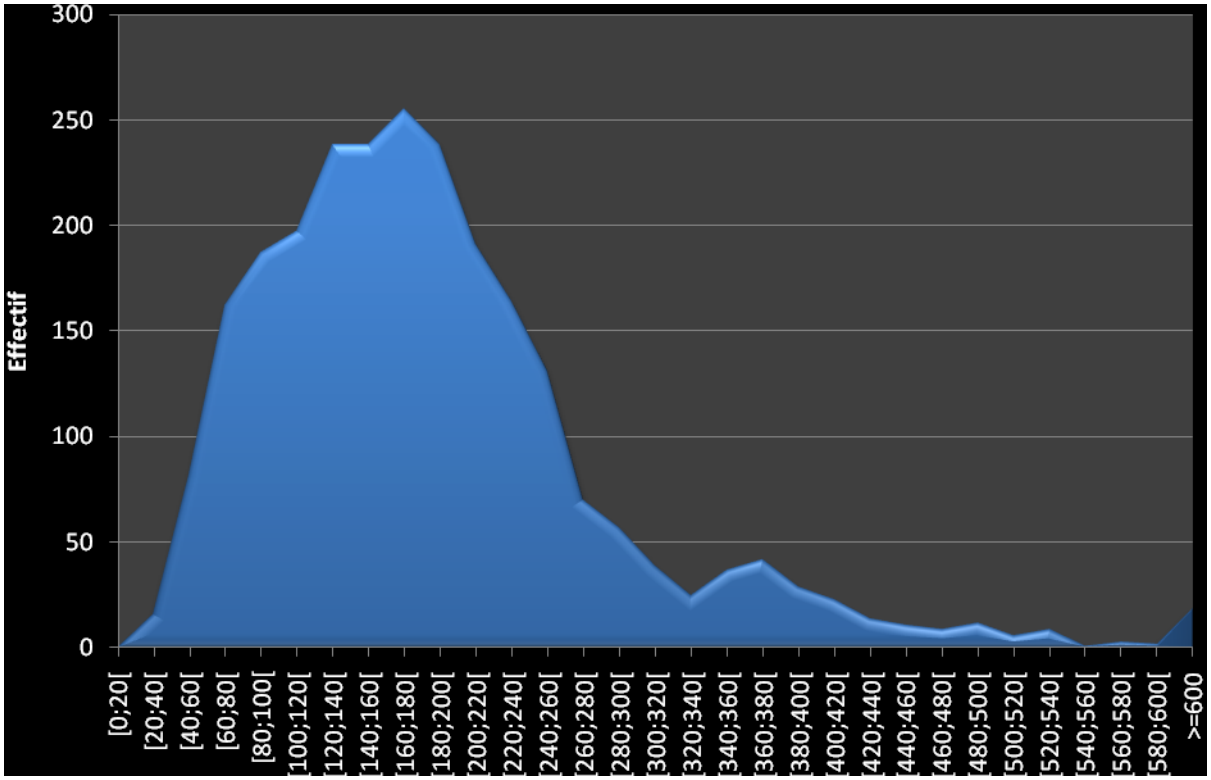


Figure 51 – Histogramme de la longueur en acides aminés des 2 507 séquences protéiques « complètes » prédites par similarité chez *A. pompejana*.
L'axe des abscisses représente les classes de longueurs, et celui des ordonnées l'effectif de séquences par classe.

10.7.2.2 Prédiction des séquences codantes par ESTScan2

10.7.2.2.1 Création du modèle pour *Alvinella pompejana*

La prédiction des séquences codantes *ab initio* nécessitant un modèle adapté, nous avons, dans un premier temps, constitué un jeu de données d'entraînement à l'aide des séquences d'*Alvinella* caractérisées par similarité. Nous nous sommes basés pour cela sur les 2 507 séquences d'ADNc comportant une région codante complète décrites ci-dessus.

Pour évaluer la pertinence de notre modèle, un jeu de 100 séquences parmi les séquences complètes prédites par similarité a été constitué aléatoirement. Les ADNc de ces séquences ont été traités par ESTScan2 avec le modèle d'*A. pompejana* pour obtenir 100 nouvelles séquences protéiques prédites *ab initio*. À titre de comparaison, nous avons parallèlement utilisé le modèle humain fourni avec ESTScan2, qui est construit à partir de plus de 14 000 séquences d'ADNc humain des banques EMBL et RefSeq.

Les 100 séquences protéiques générées par chacun des deux modèles ont été comparées à la banque UniprotKB par BlastP, et leur meilleur hit a été comparé au meilleur hit du BlastX de référence de l'ADNc sur la banque UniprotKB.

Une séquence protéique prédite par ESTScan2 est considérée comme valide lorsque l'identifiant du meilleur hit correspond à l'identifiant du meilleur hit de référence. Lorsque la protéine prédite est valide dans le cas des deux modèles, l'*E-value* de leur meilleur hit a été comparée (Tableau 11).

Tableau 11 – Résultats d'évaluation des modèles de prédiction d'ESTScan2

Modèle	***A.pompejana***	***H.sapiens***
Protéines valides	47	57
Meilleure *E-value*	9	32

Contrairement à ce que l'on pourrait attendre à première vue, le modèle humain donne de meilleurs résultats que le modèle spécialement conçu à partir des séquences d'*A. pompejana* puisqu'il permet de prédire 10% de protéines valides supplémentaires. De plus, lorsque les deux modèles prédisent une séquence protéique valide à partir du même ADNc (dans 41 cas), l'*E-value* est souvent bien meilleure pour le modèle Humain.

Plusieurs hypothèses peuvent justifier l'échec du modèle *A. pompejana* : (1) un nombre insuffisant de séquences pour permettre l'entraînement d'un modèle de Markov caché, (2) un jeu d'entraînement peu diversifié en terme de composition, ce biais pouvant être associé à la méthode de prédiction par similarité, (3) un biais en composition des régions UTR et codantes insuffisant pour permettre au modèle de bien différencier les transitions.

La troisième hypothèse peut être rapidement écartée puisque la composition moyenne en GC des régions 5'UTR, codante et 3'UTR des séquences utilisées pour l'entraînement du modèle est respectivement de 40,1 %, 44,9 % et 33,5 %.

Dans l'impossibilité de pouvoir avoir confiance dans le modèle réalisé pour *A. pompejana*, ou d'obtenir de nouvelles séquences pour le rendre plus robuste, nous nous sommes dirigés vers l'optimisation de l'utilisation du modèle humain sur les séquences d'*Alvinella*.

10.7.2.2.2 Ajustement du score pour le modèle humain

À chaque séquence prédite par ESTScan2 est associé un score de prédiction. Pour déterminer un score au dessus duquel les séquences prédites à l'aide du modèle humain ont une qualité acceptable, l'intégralité des séquences d'ADNc d'*A.pompejana* a été traitée par ce logiciel. Les séquences ont ensuite été réparties en deux groupes : les séquences avec une similarité suffisante pour avoir créé une séquence protéique par similarité, et les séquences sans similarité. Ces deux catégories ont été reportées sur un graphique cumulatif du nombre de séquences protéiques prédites par ESTScan2 en fonction de leur score (Figure 52).

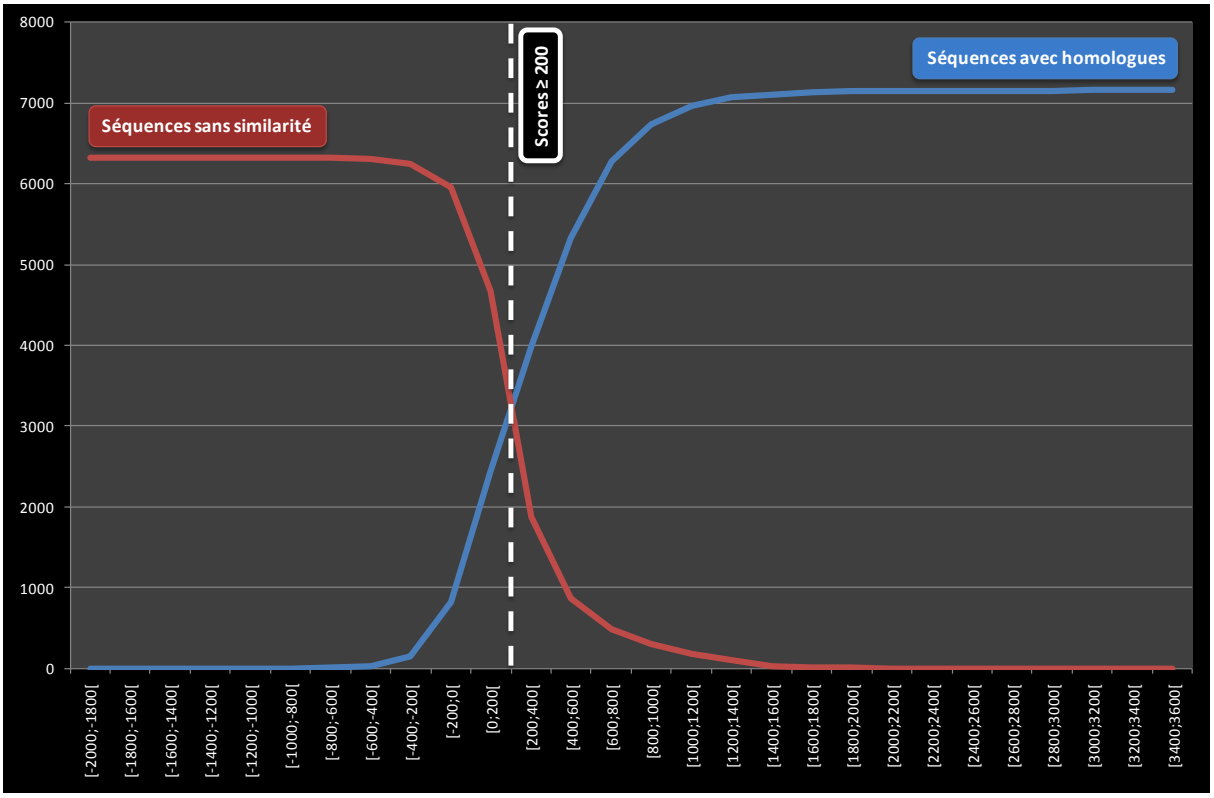


Figure 52 – Graphiques cumulatifs du nombre de protéines prédites par ESTScan en fonction du score de prédiction.
En bleu, les séquences ayant une similarité significative dans la banque UniprotKB. En rouge, les séquences n'ayant pas de similarité. La courbe rouge a été transformée en courbe soustractive pour mieux apprécier l'intersection des deux courbes.

Il est raisonnable de prendre comme seuil, le point d'intersection des deux courbes (environ 200). On obtient alors une spécificité de 70% et une sensibilité de 66% (Figure 53).

$$Sensibilit\acute{e} = \frac{VP}{VP + FN} \qquad Spécificit\acute{e} = \frac{VN}{VN + FP}$$

Figure 53 – Rappel des formules de calcul de Sensibilité et de Spécificité
Soit VP le nombre de vrais positifs (nombre de séquences avec homologie ayant un score ≥ 200), FN de faux négatifs (nombre de séquences avec homologie ayant un score < 200), VN de vrai négatif (nombre de séquences sans similarité ayant un score < 200), et FP de faux positifs (nombre de séquences sans similarité ayant un score ≥ 200)

Parmi les 8 505 séquences sans similarité significative, 1 868 ont atteint la valeur seuil de 200 pour le score d'ESTScan2, augmentant le total de séquences protéiques à 9 221 séquences. Les 1 868 séquences « orphelines » constituent des cibles prioritaires pour de futures études expérimentales. En effet, elles correspondent potentiellement à des gènes spécifiques d'*Alvinella* ou d'Annélides.

10.8 Discussion et perspectives

Par le développement de ce pipeline, nous avons mis en place les bases de l'infrastructure qui nous a servi à étudier les ADNc d'*Alvinella pompejana*.

Cette architecture a été suffisamment souple pour traiter d'autres banques d'ADNc d'organismes et de tissus eucaryotes dans le cadre de projets collaboratifs :

- *Bathymodiolus azoricus* – collaboration avec A. Tanguy, Station Biologique de Roscoff : *B. azoricus* est un mollusque bivalve rencontré, comme *A. pompejana*, dans des sites hydrothermaux mais dans des zones caractérisées par un apport hydrothermal réduit et des températures comprises entre 4 et 25°C. 3 732 EST ont été séquencés puis traités par notre pipeline. Le séquençage de ces EST s'inscrit dans le cadre d'une étude plus vaste sur la recherche de corrélation entre les variations des paramètres physico-chimiques (température, sulfures, pH et métaux lourds) et la régulation du transcriptome.
- *Branchipolynoe seepensis* – collaboration avec A. Tanguy, Station Biologique de Roscoff : *B. seepensis* est un polychète symbiotique obligatoire de bathymodioles, notamment *B. azoricus*. Nous avons traité 1 809 EST.
- *Paralvinella grasslei* – collaboration avec D. Jollivet, Station Biologique de Roscoff. Comme *A. pompejana*, *P. grasslei* appartient à la famille des Alvinellidés et est inféodé aux sites hydrothermaux mais colonise des zones moins chaudes que ce dernier. 6 752 EST ont été traités.
- rétine de souris (*Mus musculus*), de rat (*Rattus norvegicus*) et de poulet (*Gallus gallus*) - collaboration avec T. Léveillard (Institut de la Vision, Paris) dans le cadre d'études sur la dégénérescence rétinienne au profit du projet européen EVI-GENORET (http://www.evi-genoret.org/). Respectivement 43 476, 32 012 et 1 055 ESTs ont été traités. Quelques petites adaptations ont été nécessaires pour traiter ces banques. En effet, les séquences proviennent de deux centres de séquençage distincts (Génoscope – CNS et Celera Genomics), et dans un des cas, seules les séquences brutes nous ont été fournies. Le pipeline a donc été configuré pour attribuer des « pseudos » valeurs de qualité (15 par défaut) si ces données sont manquantes, afin de permettre l'assemblage.

L'utilisation du pipeline dans ces différents projets, et tout particulièrement dans le projet *A. pompejana*, nous a amenés à envisager certaines améliorations. Ainsi, les paramètres très stricts d'assemblage appliqués par défaut ont l'avantage de ne rassembler en contig que les copies d'un transcrit unique, sans amalgamer des variants d'épissage ou des transcrits de

gènes paralogues. En contrepartie, il est difficile d'estimer le nombre de gènes séquencés. Idéalement, cette estimation pourrait se faire par localisation des transcrits sur le génome quand celui-ci est disponible. Un module de localisation génomique (externe ou interne) pourrait ainsi à l'avenir compléter notre pipeline. Dans le cadre du projet *A. pompejana*, en l'absence de génome séquencé, nous avons appliqué une méthode de regroupement par BlastN (Cf. Publication 1, page 153).

Par ailleurs, des améliorations pourraient être apportées à la détection des séquences d'ARN non codants. Un banc d'essai des méthodes de prédiction d'ARN non codants relativement récent suggère que la détection d'ARNr par homologie demande une optimisation des matrices de score pour être réellement efficace (Freyhult *et al.*, 2007). Une alternative intéressante pourrait être apportée par le logiciel RNAmmer (Lagesen *et al.*, 2007) qui utilise un modèle de Markov pour détecter les ARNr.

Enfin, le pipeline pourrait être adapté au traitement de banques d'ADNc procaryotes. Pour cela, une étape de prédiction de séquences protéiques spécifique devrait être développée, puisque par exemple les ARNm polycistroniques ne sont pas gérés.

11 ANNOTATION INTÉGRATIVE AUTOMATIQUE DE SÉQUENCES PROTÉIQUES

Les premières étapes d'une annotation fonctionnelle de séquences provenant d'un projet à haut-débit consistent en l'application de protocoles automatisés afin de définir le répertoire de séquences. Idéalement, ces annotations devront être expertisées manuellement par la suite.

Tout comme les pipelines de traitement de données brutes de séquençage, il existe des pipelines d'annotation disponibles en tant que services à l'instar de MaGe (Vallenet *et al.*, 2006), AGMIAL (Bryson *et al.*, 2006), RAST (Aziz *et al.*, 2008), PIPA (Yu *et al.*, 2008), et IMG ER (Markowitz *et al.*, 2009), ou pouvant être téléchargés et installés localement à l'instar de Manatee (Manatee), DIYA (Stewart *et al.*, 2009), et GRC (Warren et Setubal, 2009).

Ces protocoles automatisés combinent plusieurs méthodes ou logiciels pour obtenir l'annotation la plus exhaustive possible. Ces méthodes peuvent se baser sur la composition de la séquence (méthodes *ab initio*) ou son contexte génomique (inférence fonctionnelle) pour déterminer certaines annotations. Cependant, la plupart transfèrent des connaissances par similarité en se reposant essentiellement sur BLAST. Les méthodes de transfert par similarité sont grandement dépendantes de la distance évolutive qui sépare les organismes entre lesquels sont transférées les annotations. En effet, plus cette distance est grande, plus la probabilité que deux séquences hautement similaires ne partagent pas la même fonction augmente (Rost, 2002). Il est donc de la plus haute importance de maintenir les séquences dans leur contexte évolutif lorsque le transfert d'annotations intervient.

C'est dans ce contexte que nous avons développé un pipeline d'annotation automatique dont l'originalité réside dans l'utilisation d'alignements multiples de séquences complètes hiérarchisés (MACS). Ce pipeline a été développé et intégré à Gscope dans le cadre de la réannotation du génome de *Mycobacterium smegmatis* (Odile Lecompte, manuscrit en préparation) et de l'annotation des protéines prédites à partir des ADNc d'*A. pompejana*, en collaboration avec Emmanuel Perrodou au LBGI.

Ce premier niveau d'annotation, qui peut être utilisé pour l'annotation des protéines de n'importe quel organisme, a été complété par une annotation de second niveau visant à positionner plus précisément les protéines d'un organisme, en l'occurrence *Alvinella,* au sein de réseaux de voies métaboliques, telles que définies dans la banque KEGG PATHWAY, ou de réseaux d'interaction protéine-protéine disponibles dans la banque STRING.

Dans ce chapitre sera décrite la méthodologie employée par ce pipeline qui est schématisée dans la Figure 54 ainsi que les résultats obtenus lors de l'annotation des protéines d'*A. pompejana* et de *M. smegmatis*.

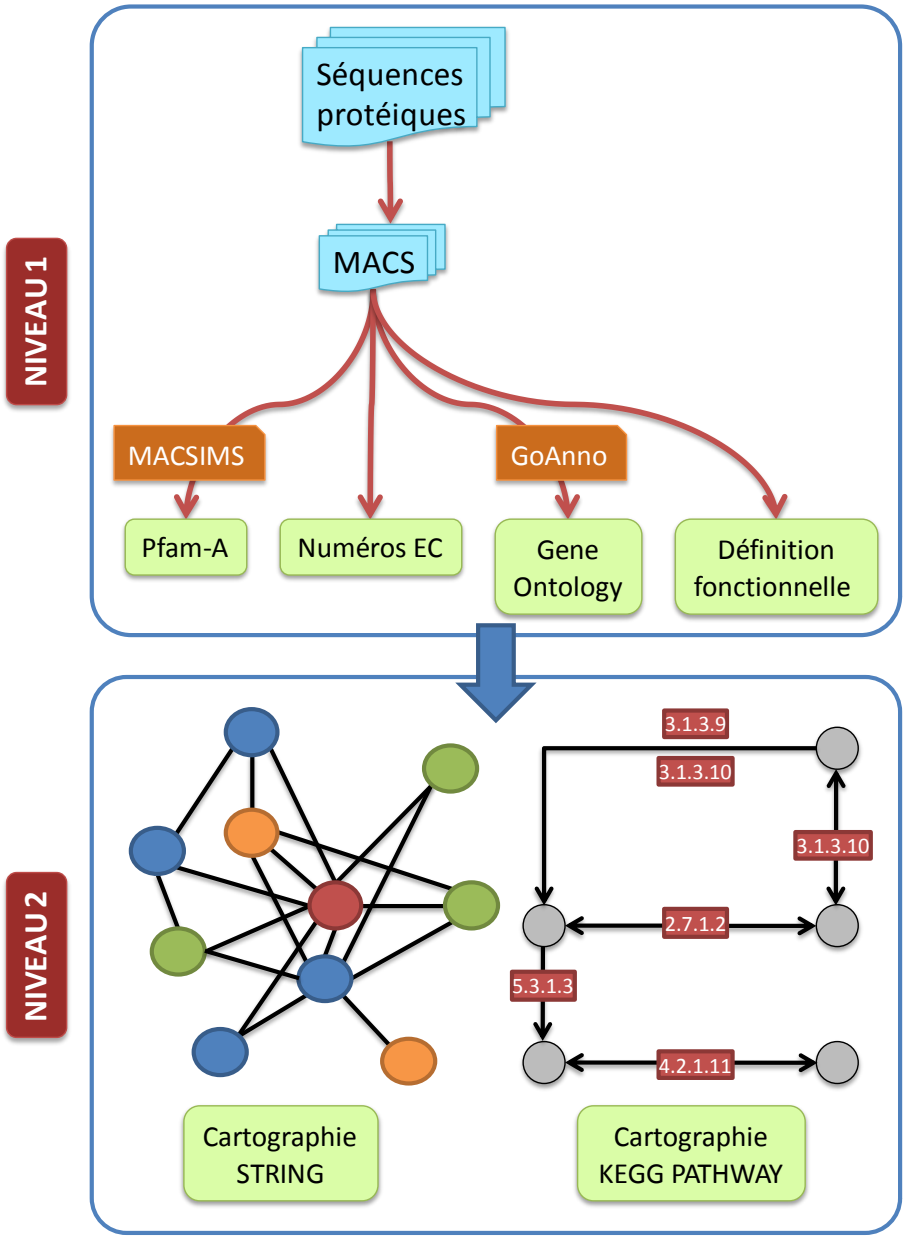


Figure 54 – Schéma général du pipeline d'annotation intégrative de séquences protéiques

11.1 Premier niveau d'annotation : annotation linéaire

Ce niveau d'annotation permet de caractériser fonctionnellement chaque protéine prise individuellement (Cf. Annotation fonctionnelle, page 37). Au cours de cette annotation, on attribue à la protéine : une définition textuelle, un ensemble de termes Gene Ontology, un numéro Enzyme Commission, et une annotation de domaines Pfam-A.

Pour prendre en compte le contexte évolutif de la séquence d'intérêt, l'attribution de ces annotations est contrainte par un ensemble de règles appliquées à l'analyse de la conservation de régions ou de domaines préservés à l'intérieur de la famille, ou de la sous-famille, au sein du MACS.

L'annotation débute par la construction, pour chacune des protéines, d'un MACS à l'aide de la cascade PipeAlign. Cet alignement multiple de séquences complètes va servir de base à chacune des étapes du processus d'annotation.

PipeAlign a été paramétré pour réaliser sa recherche BlastP sur la banque UniProtKB, et le nombre de séquences composant le MACS a été limité à 250 dans le cadre d'*Alvinella* et à 500 dans le cas de *M. smegmatis* compte tenu de la redondance observée pour les procaryotes. Ces valeurs sont un bon compromis entre la vitesse de construction et la teneur en contenu informationnel.

11.1.1 Extraction et intégration des données des programmes externes

En complément de PipeAlign, le protocole d'annotation fait appel à deux programmes externes: MACSIMS, qui réalise un certain nombre d'annotations structurales et fonctionnelles, notamment l'annotation des domaines Pfam-A, ainsi que GoAnno, qui réalise l'annotation Gene Ontology.

Ces deux programmes fournissent chacun un fichier de sortie XML et les résultats d'annotation sont extraits et compilés par notre pipeline dans un même fichier du répertoire « infos » de Gscope.

11.1.2 Attribution de définition fonctionnelle

La nature d'une définition fonctionnelle est très hétérogène. En effet, selon le contexte dans lequel s'est placé son auteur, elle peut décrire une réalité biochimique, physiologique, développementale, phénotypique, ou autre…

Dans ce contexte, il est difficile, parmi un jeu de définitions, de choisir automatiquement à l'aide de règles simples laquelle est la plus pertinente.

Cependant, dans un article de E.M. Marcotte (Marcotte *et al.*, 2001), une méthode a été décrite pour « scorer » les résumés d'articles de Medline par une approche Bayésienne en fonction des fréquences d'apparition de mots discriminants. Les mots discriminants sont identifiables par leur tendance à apparaitre peu fréquemment ou très fréquemment à l'intérieur d'un jeu de résumés d'articles d'intérêt et obtiendront un score $\ln(p) < -13$. Le score qui définit un résumé d'article est tout simplement la somme des scores des mots discriminants qu'il contient. Plus la valeur est négative, meilleur est la pertinence du résumé.

En collaboration avec Véronique Geoffroy du LBGI, nous avons adapté cette méthode afin de « scorer », puis choisir, la meilleure définition fonctionnelle parmi les définitions provenant de l'ensemble des séquences de la sous-famille (telle que définie sur la base du MACS) contenant la séquence d'intérêt.

L'ensemble des définitions est décomposé en une liste de mots qui sont « scorés » selon la formule suivante (Figure 55) :

$$\ln(p_i) = -n \times \frac{E_i}{N} + d_i \times \ln\left(n \times \frac{E_i}{N}\right) - \ln(d_i!)$$

Figure 55 – Calcul du score ln(*p*) du mot discriminant « *i* »

Soit *n* le nombre total de mots composant la liste de mots fournie, *N* le nombre total des mots composant toutes les définitions fonctionnelles d'UniprotKB/SwissProt, E_i le nombre d'entrées UniprotKB/SwissProt dont la définition comporte au moins une fois le mot « *i* », et d_i le nombre d'occurrences du mot « *i* » dans la liste de mots fournie.

Le score de chaque définition est ensuite calculé en sommant les scores des mots discriminants (score < -13) qu'elle comporte. Les définitions sont ensuite triées par ordre décroissant en privilégiant les définitions provenant d'UniprotKB/SwissProt à celles d'UniprotKB/TrEMBL qui sont de moins bonne qualité, car annotées automatiquement.

Pour ce faire, lors du tri des comparaisons deux-à-deux des scores, lorsque le ratio des deux scores est ≥ 80%, une définition provenant d'UniprotKB/SwissProt est promue par rapport à une définition provenant d'UniprotKB/TrEMBL.

De la même façon, en cas « d'égalité », nous privilégions les définitions les plus courtes. Ainsi, le niveau de précision de la définition qui sera attribuée ne sera pas exagéré. Par exemple, la définition '*Ribonucleotide reductase small subunit*' est privilégiée par rapport à '*Ribonucleoside-diphosphate reductase small chain (EC 1.17.4.1) (Ribonucleotide reductase small subunit) (Ribonucleotide reductase M2 subunit)*'.

Enfin, nous avons compilé une liste de mots non informatifs (fragment, putative, hypothetical, like, related…), dont le score sera systématiquement fixé à 0 pour ne pas biaiser le résultat final.

11.1.3 Assignation de numéro EC

Lors de l'extraction de données, MACSIMS récupère et fourni les numéros EC à partir de la banque Uniprot ou PDB, mais ne propage malheureusement pas cette information à travers d'autres séquences.

Pour tenter d'attribuer un numéro EC à l'enzyme d'intérêt, nous avons développé une procédure qui produit un numéro EC consensuel à partir de la liste de numéros EC provenant de l'annotation des séquences de la sous-famille de la protéine d'intérêt.

Son principe est de rechercher un numéro EC dont la fréquence dans la liste fournie est supérieure ou égale à une certaine valeur seuil. Plusieurs tests ont été effectués sur les jeux de données *Alvinella* et *Mycobacterium* et la valeur de 70% a été retenue empiriquenement par défaut. Si un candidat remplit ces conditions, ce numéro sera attribué à la protéine d'intérêt. Dans le cas contraire, cette recherche est réitérée sur les 3 premiers chiffres de chaque numéro EC (EC de niveau 3), et ainsi de suite jusqu'au niveau 1, jusqu'à trouver le cas échéant un numéro EC consensuel (Figure 56). Ce numéro est ensuite inscrit dans le fichier adéquat du répertoire « infos » de Gscope.

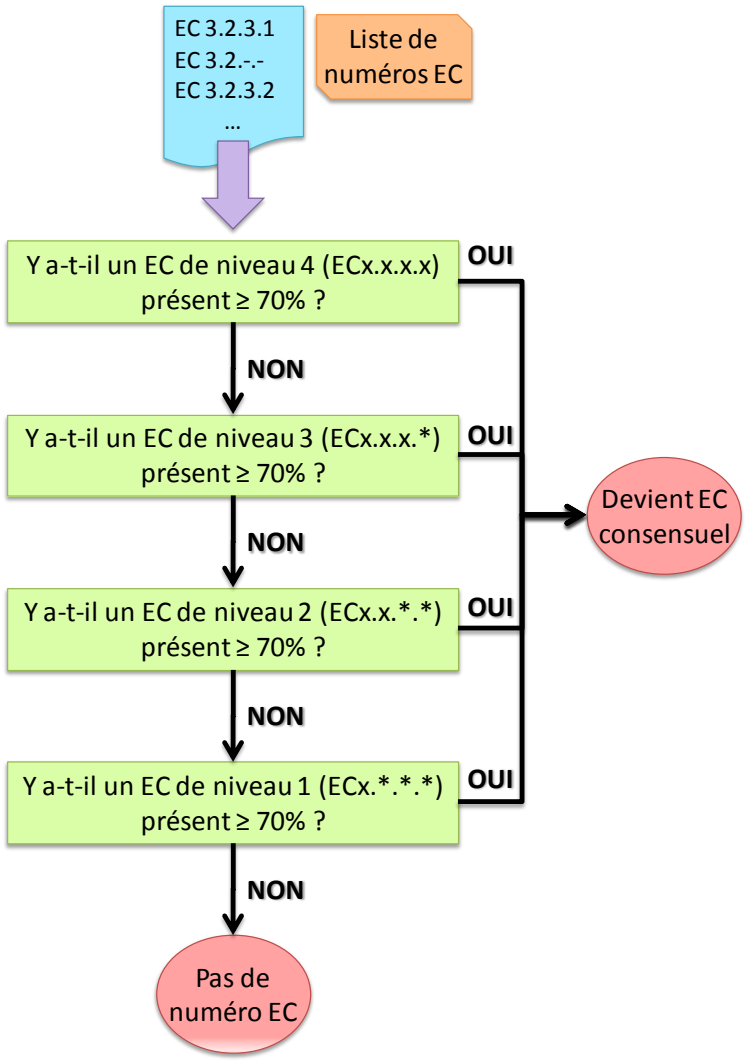


Figure 56 – Diagramme de décision permettant d'assigner un numéro EC consensuel

11.2 Deuxième niveau d'annotation : localisation à l'intérieur de réseaux

Nous avons ensuite développé des modules permettant de localiser les protéines annotées par les programmes décrits ci-dessus à l'intérieur des réseaux de voies métaboliques et d'interaction protéine-protéine KEGG PATHWAY et STRING respectivement. Ce deuxième niveau d'annotation a été appliqué aux protéines d'*A. pompejana*.

11.2.1 Réseau de voies métaboliques KEGG PATHWAY

La localisation est effectuée sur les voies métaboliques de référence de KEGG à l'aide des EC qui ont été assignés par le protocole d'annotation de premier niveau.

11.2.1.1 Structure de la banque de données KEGG

Les voies métaboliques de la banque KEGG sont distribuées sous forme d'une multitude de fichiers répartis en des répertoires figurant l'organisme qui a été cartographié sur les voies. Par exemple, le répertoire « hsa » correspond à l'humain (*Homo sapiens*), « mmu » correspond à la souris (*Mus musculus*)… Ces noms de répertoires correspondent au préfixe du numéro d'accès d'une cartographie de voie métabolique, le suffixe étant un nombre identifiant la voie métabolique en question (00010 pour la glycolyse/gluconéogenèse, 00020 pour le cycle de Krebs, …).

Le répertoire qui nous intéresse ici est le répertoire « map » qui contient les voies de référence, c'est-à-dire les voies où figurent l'intégralité des produits et des réactions enzymatiques connus pour une voie donnée, indépendamment de l'organisme.

Chaque voie est constituée de plusieurs fichiers :

- Une image de la voie établie par un expert à partir de données de la littérature,
- Un fichier contenant la liste des enzymes figurant dans cette voie (numéros EC),
- Un fichier contenant la liste des réactions qui interviennent dans cette voie (numéros d'accès KEGG LIGAND),
- Un fichier contenant la liste des orthologues cartographiés dans cette voie (numéros d'accès KEGG GENES),
- Des fichiers de coordonnées indiquant les positions relativement à l'image, des enzymes, réactions et orthologues.

11.2.1.2 Cartographie automatique sur les voies métaboliques

Pour réaliser la localisation automatique, nous avons développé un programme qui place une liste de numéros EC dans toutes les voies de référence.

Une protéine est localisée dans une voie à l'aide des fichiers de numéros EC et de coordonnées. Si son numéro EC est au moins de même niveau que l'un des numéros EC de la voie, et que tous les chiffres de même niveau concordent, la protéine est placée dans la voie. Par exemple, une protéine ayant le numéro EC 3.2.1.-, sera placée dans une voie contenant une enzyme ayant le numéro EC 3.2.1.-, EC 3.2.-.- ou EC 3.-.-.-, mais pas EC 3.2.1.26.

En retour, les fichiers du répertoire « infos » de Gscope sont actualisés par la liste de voies métaboliques dans lesquelles apparaît chaque protéine, et les images des voies

métaboliques sont modifiées automatiquement pour mettre en valeur les enzymes qui ont été cartographiées chez l'organisme d'intérêt (Figure 57).

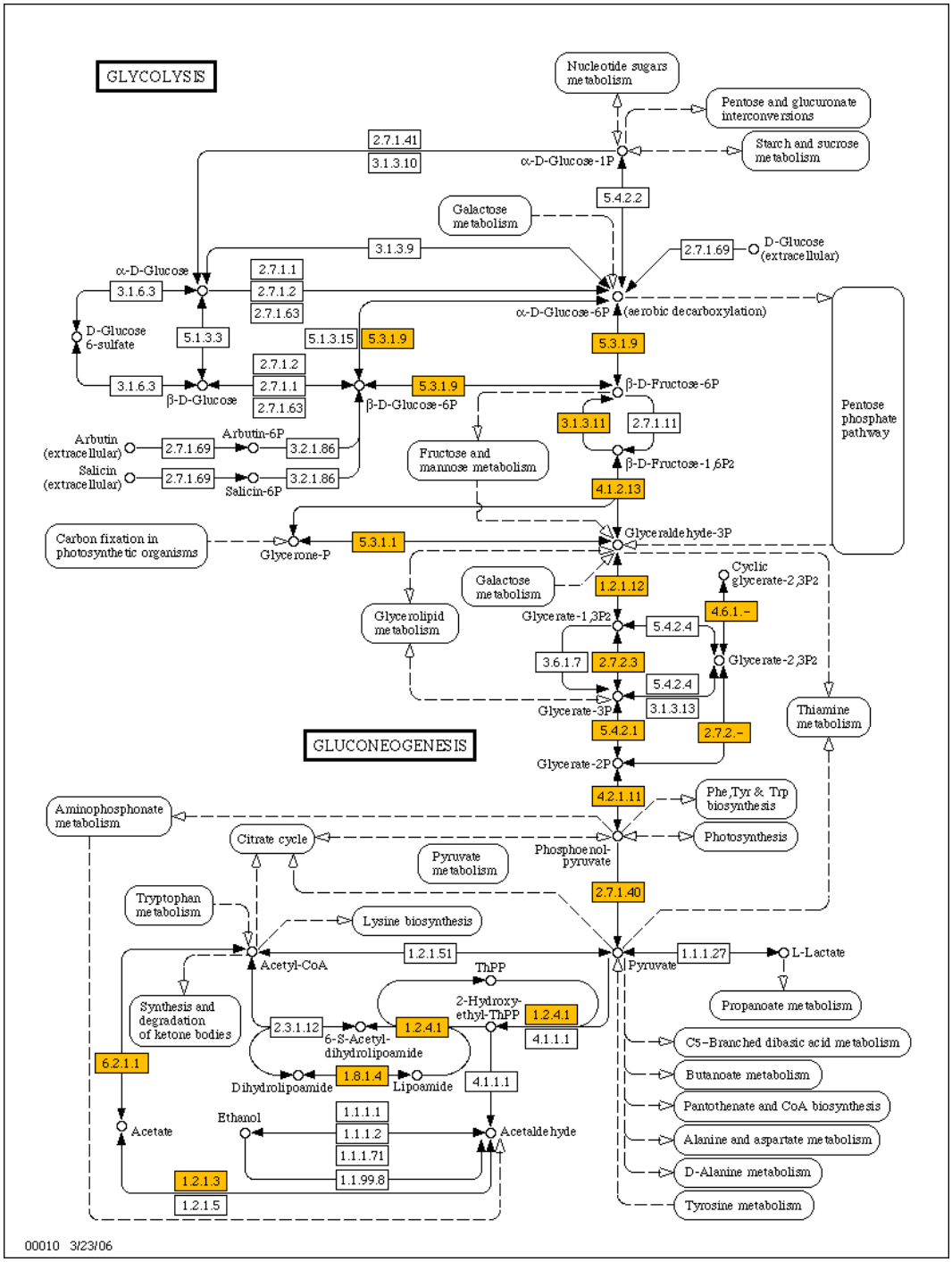


Figure 57 – Exemple de voie métabolique cartographiée automatiquement
Ici la voie de la glycolyse/gluconéogenèse. Les numéros EC coloriés en orange sont ceux qui ont été retrouvés chez *Alvinella*.

Ce programme a été réalisé en PHP, dans une version en ligne de commande, et pourra être prochainement l'intégré au site Web '*Alvinella pompejana* cDNA project' et ainsi permettre

la génération dynamique des images de voies métaboliques (Cf. Interface d'accès Web, page 157).

11.2.2 Réseau d'interactions protéine-protéine STRING

Les réseaux d'interaction protéine-protéine sont bien sûr propres à chaque organisme. Cependant le transfert d'interactions connues selon le principe des interologues permet d'obtenir une vision approximative des interactions potentielles pour les organismes où les données sont rares ou inexistantes. Nous avons donc développé des modules permettant de localiser un ensemble de protéines dans les réseaux d'interaction de la banque STRING correspondants à l'organisme considéré ou à un organisme « proche ». Une approche originale a ensuite été conçue afin d'obtenir des sous-graphes plus facilement exploitables pour analyser les complexes.

Dans le cadre de l'étude d'*Alvinella*, nous disposions localement de la banque STRING dans sa version 7.1 qui intègre 373 espèces. En raison de la proximité observée entre les protéines d'*A. pompejana* et des Chordés (Cf. Publication 1, page 153) et de la richesse des données concernant l'interactome humain, la localisation des protéines d'*A. pompejana* a été effectuée sur les réseaux d'interactions humains comme décrits ci-dessous mais l'approche est évidemment applicable à d'autres organismes.

11.2.2.1 Construction du graphe d'interactions

Pour des raisons de performance, et pour éviter de surcharger le serveur PostgreSQL, le graphe d'interactions humain est extrait de la banque STRING, reconstitué en mémoire et seules les informations nécessaires sont conservées (Figure 58).

Les 22 218 protéines du réseau humain sont extraites et vont représenter les nœuds du graphe. Chacun de ces nœuds va contenir l'identifiant numérique unique de la protéine (clé primaire de la table 'proteins' de STRING) afin de pouvoir consulter très rapidement la base de données au besoin, ainsi que le nom (celui qui est préféré par STRING) et la définition fonctionnelle de la protéine.

Les interactions dont le score combiné est ≥ 0,900 (score de grande fiabilité selon les auteurs) sont ensuite sélectionnées et recrées en tant qu'arêtes du graphe en mémoire. Ainsi, nous obtenons un graphe dont les interactions sont relativement fiables.

Enfin, les nœuds de protéines qui ne sont pas reliées à d'autres nœuds, car n'étant pas impliqués dans des interactions suffisamment fiables, sont retirés du graphe. Ainsi, le graphe final humain comporte 7 451 nœuds.

À ce niveau, la localisation des protéines d'*Alvinella* pourrait déjà être réalisée. Cependant, un graphe aussi imposant rend difficile le repérage et l'étude de complexes d'interactions. Le

graphe a donc été décomposé en sous-graphes pour obtenir des complexes d'interactions plus réduits humainement exploitables.

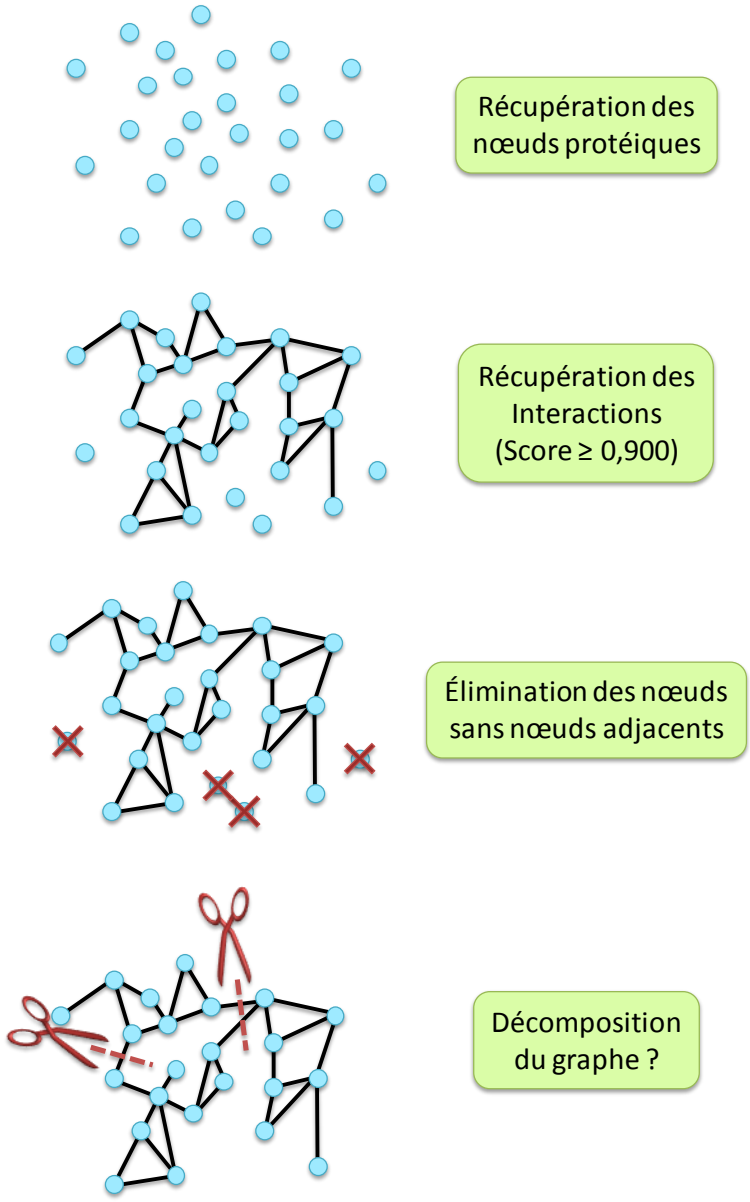


Figure 58 – Construction du graphe d'interactions STRING en mémoire

11.2.2.2 Découpage du graphe d'interactions en sous-graphes

Le graphe actuellement en mémoire est un « embrouillamini » d'interactions du fait de l'existence de protéines impliquées dans plusieurs complexes formant des « ponts » réunissant ces différents complexes d'interactions.

Une méthode pour découper le graphe serait de détecter ces protéines et de briser ces ponts pour obtenir des sous-graphes. Cependant ce type d'approche est difficile à mettre en œuvre pour obtenir des résultats cohérents. C'est pourquoi nous avons plutôt opté pour une stratégie inverse de type '*bottom-up*' : nous développé une approche originale qui consiste

de débuter à partir des plus petits sous-graphes possibles, c'est-à-dire une protéine et ses interactants directs, puis de les fusionner petit à petit selon certains critères.

Ces critères vont principalement jouer sur la densité des sous-graphes (Figure 59). La densité d'un graphe se définit par la propriété que son nombre d'arêtes soit proche du nombre maximal théorique d'arêtes. En calculant le ratio nombre d'arêtes/nombre maximal, on obtient une valeur située entre 0 et 1, la valeur 1 symbolisant un graphe très dense. À l'inverse, un graphe pour lequel cette valeur se rapproche de 0 est très peu dense, et aura tendance à former ces fameux « ponts », indésirables lors d'une fusion.

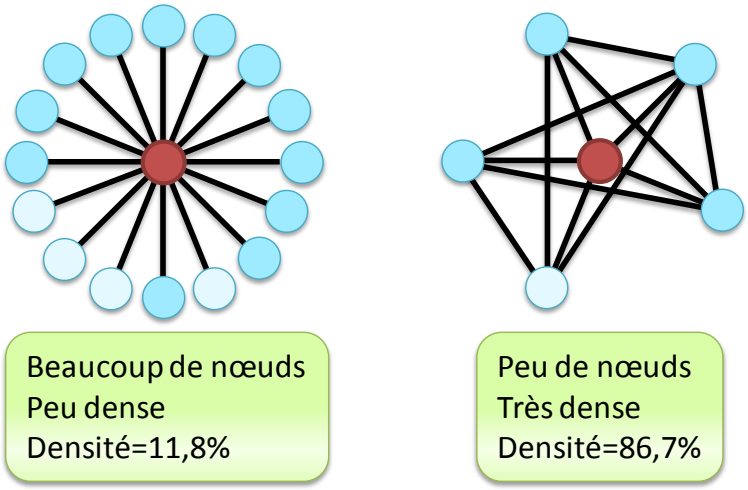


Figure 59 – Illustration de la densité de deux graphes d'interactants directs

L'idée sous-jacente est de fusionner préférentiellement les sous-graphes denses fortement apparentés (minimum de nœuds en commun et maximum d'arêtes) tout en évitant de fusionner des sous-graphes peu denses. La fusion de deux sous-graphes est opérée si, et seulement si, les deux critères suivants sont respectés (Figure 60, Figure 61) :

$$Sfusion_{A,B} = 1 - (1 - Rcommun) \times (1 - Rdensité_A) \times (1 - Rdensité_B)$$

$$Rcommun = \frac{Noeuds_{commun}}{Noeuds_A}$$

$Sfusion_{A,B} \geq 0{,}7$
$Rcommun \geq 50\%$
$Nœuds_A \leq Nœuds_B$

Figure 60 – Critères de fusion entre deux sous-graphes STRING

Avec $Sfusion_{A,B}$ le score de fusion entre le sous-graphe A et le sous-graphe B, *Rcommun* le ratio du nombre de nœuds en commun entre A et B sur le nombre de nœuds de A, A étant le plus petit des deux sous-graphes.

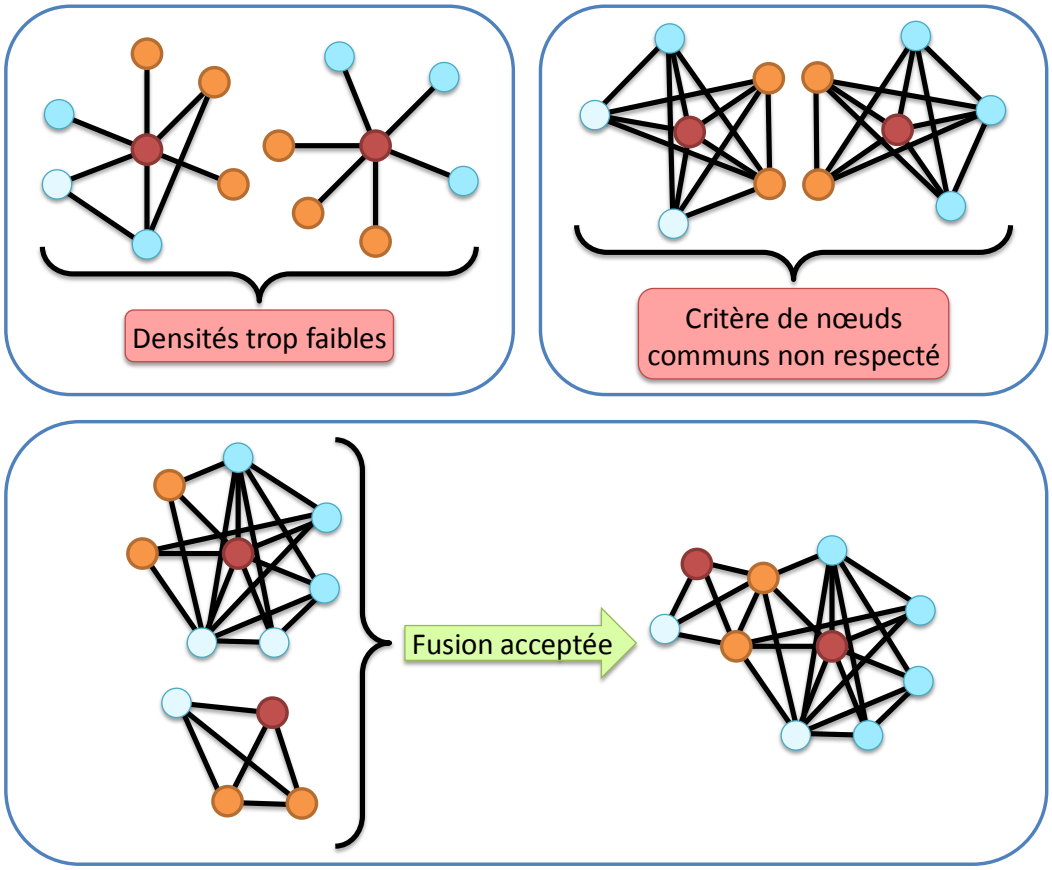


Figure 61 – Fusion de sous-graphes STRING
Les nœuds rouges représentent une protéine de STRING, les nœuds bleus et orange ses interactants directs, les nœuds orange étant aussi les nœuds communs entre deux sous-graphes.

Les tentatives de fusions sont réalisées itérativement deux-à-deux jusqu'à ce que plus aucune fusion ne soit possible. Cette méthode a permis de décomposer le graphe initial en 385 sous-graphes dont la taille varie entre 2 et 3025 nœuds. Les plus gros sous-graphes contiennent les protéines des gènes constitutifs (actine, facteurs de transcriptions, protéines ribosomales, …) qui sont intimement impliqués dans un grand nombre d'interactions.

> « *Mais alors Jamy, c'est un peu comme quand les gouttelettes d'huile de ma vinaigrette se recollent ensemble pour former des plus grosses gouttelettes ?!!* » — Fred (C'est pas sorcier, France 3)

11.2.2.3 Localisation automatique dans les sous-graphes

Les protéines d'*A. pompejana* ont été localisées à l'intérieur de chaque sous-graphe de STRING.

Pour chaque protéine, le MACS est parcouru pour récupérer le meilleur homologue humain parmi la sous-famille de la protéine d'intérêt. À défaut, le meilleur homologue humain est recherché parmi les séquences du BLASTP utilisées pour construire le MACS.

Grâce aux informations récupérées pour le meilleur homologue trouvé, la protéine correspondante est recherchée, dans l'ordre suivant, à l'aide de l'identifiant Uniprot, des numéros d'accès Uniprot, Ensembl, RefSeq, Genome Reviews (Sterk *et al.*, 2006) et enfin du nom et des synonymes de gènes.

Si la protéine correspondante n'est pas trouvée, un BlastP de la protéine d'intérêt est réalisé sur la banque des séquences protéiques humaines extraites de STRING, créée au besoin à la volée. Sous contrainte d'avoir une E-value ≤ 1^{e-5}, le premier '*hit*' est sélectionné comme la protéine STRING correspondante.

Lorsque les protéines d'*Alvinella* ont été cartographiées dans les sous-graphes de STRING, un fichier au format XGMML (*eXtensible Graph Markup and Modeling Language*) est créé pour chaque sous-graphe. Ce format permet de décrire des graphes pour pouvoir les visualiser graphiquement à l'aide de différents logiciels, dont Cytoscape (Shannon *et al.*, 2003). Un code couleur a été utilisé pour différencier les nœuds où une séquence protéique d'*Alvinella* a été localisée (Figure 62).

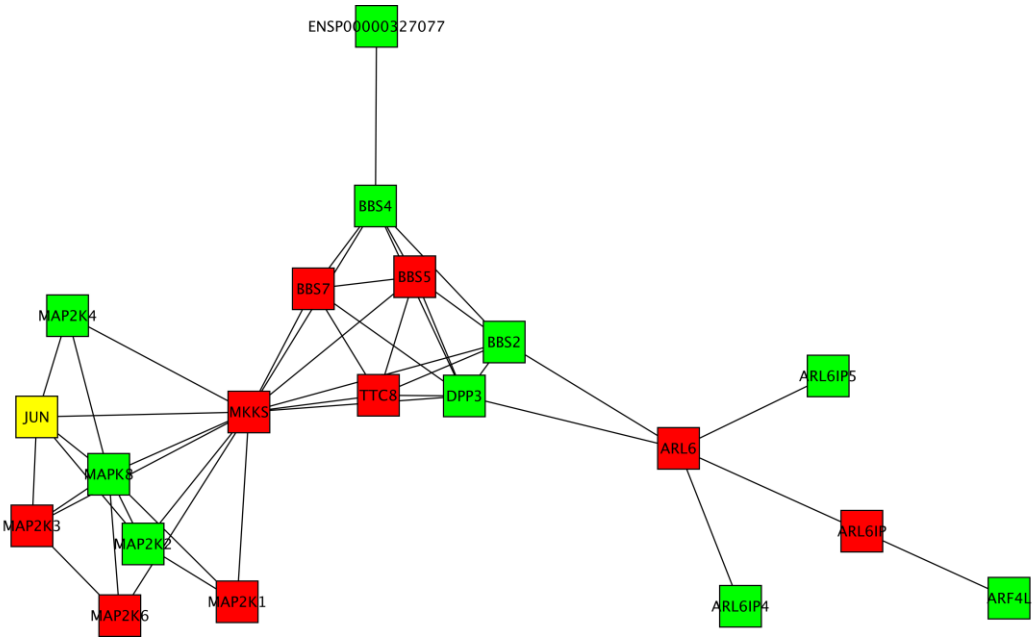


Figure 62 – Sous-graphe STRING des protéines humaines impliquées dans le syndrome de Bardet-Biedl cartographié avec les protéines d'*Alvinella pompejana*
Les nœuds rouges représentent les séquences d'*Alvinella* localisées à l'aide du meilleur homologue humain de la sous-famille du MACS, les nœuds jaunes à l'aide du meilleur homologue du MACS ou de la recherche BlastP. Les nœuds verts indiquent qu'aucune protéine d'*Alvinella* n'a été localisée. Figure réalisée avec Cytoscape.

11.3 Mise à jour des données dans la base de données d'assemblage

Alors que toutes les données d'annotations ont été regroupées à l'intérieur des fichiers « infos » de Gscope, les données de la base de données d'assemblage sont mises à jour à

l'aide de ces fichiers et de l'interface de remplissage Gscope précédemment décrite (Cf. Intégration des données, page 126), en lançant la fonction '*Update all annotations*'.

11.4 Annotation du transcriptome d'*A. pompejana*

Les séquences protéiques d'*A. pompejana* obtenues par le pipeline de traitements des ADNc (Cf. Prédiction des séquences codantes par similarité, page 128) ont été soumises au protocole d'annotation. Notre pipeline se basant sur l'homologie, l'annotation n'a concerné que les protéines présentant une similarité significative, c'est-à-dire 7 353 protéines. Le bilan de l'annotation intégrative apparait dans le Tableau 12.

Tableau 12 – Bilan de l'annotation des 7 353 séquences protéiques ayant une similarité significative

Annotation	Nombre d'éléments
Niveau 1	**Protéines**
Protéines ayant une similarité significative	7 353 (100%)
Domaines Pfam-A	4 767 (65%)
Gene Ontology	5 949 (81%)
Processus biologique	5 072 (69%)
Composant cellulaire	4 530 (62%)
Function moléculaire	5 601 (76%)
Définition fonctionnelle	4 611 (63%)
Enzyme Comission	1 243 (17%)
Niveau 4 (X.X.X.X)	1 180 (16%)
Protéines annotées	6 252 (85%)
Niveau 2	**Réseaux**
KEGG PATHWAY	345
Localisés	202
Couverts ≥50%	82
Sous-graphes STRING	385
Localisés	264
Couverts ≥50%	63

Un total de 6 252 séquences protéiques, soit 85% des séquences initiales traitées par le pipeline d'annotation intégrative, ont été annotées par au moins un domaine Pfam-A, ou un terme GO, ou une définition fonctionnelle, ou un numéro EC. La localisation des protéines d'*Alvinella* sur les réseaux KEGG PATHWAY et STRING a permis de couvrir à plus de la moitié, respectivement 40% et 68% de ces réseaux. Par ailleurs, près de la moitié des séquences protéiques d'*Alvinella* ont été localisées sur le réseau humain de STRING. Ces chiffres sont très encourageants quant à la qualité des banques d'ADNc ainsi qu'à la performance du pipeline d'annotation.

11.5 Réannotation du génome de *Mycobacterium smegmatis*

Mycobacterium smegmatis est une bactérie apparentée à de nombreuses espèces pathogènes dont *Mycobacterium tuberculosis*, responsable de la tuberculose, *Mycobacterium leprae*, à l'origine de la lèpre, ou encore *Mycobacterium ulcerans* qui provoque l'ulcère de Buruli. De par sa croissance rapide et son caractère non pathogène, *M. smegmatis* constitue un système modèle pour l'étude des Mycobactéries (Reyrat et Kahn, 2001). Elle possède en outre le plus grand génome (6 988 kb) parmi les nombreux génomes de *Mycobacterium* actuellement disponibles.

Le travail de réannotation des protéines auquel nous avons été associés s'inscrit dans le cadre d'un projet collaboratif réunissant l'équipe « Analyse génétique de l'enveloppe mycobactérienne » dirigée par JM Reyrat (Hôpital Necker, Paris), le Laboratoire de Spectrométrie de Masse Bio-Organique dirigé par A. Van Dorsselaer (ECPM, Strasbourg) et le LBGI. L'objectif était de contribuer à l'amélioration de la séquence et des annotations du génome de la souche *M. smegmatis* MC2 155, séquencé et annoté par le TIGR en associant l'expérience du LBGI en matière d'annotation, l'expertise biologique de JM Reyrat et de ses collaborateurs et le savoir-faire en protéomique du laboratoire de A. Van Dorsselaer.

Les travaux réalisés ont d'abord porté sur l'amélioration de la séquence génomique par la détection de séquences codantes interrompues pouvant correspondre à des erreurs de séquençage (Perrodou *et al.*, 2006) et le re-séquençage des régions incriminées (Deshayes *et al.*, 2007). Une nouvelle approche de protéomique orientée vers la correction ou validation des codons initiateurs dans les génomes procaryotes a ensuite été développée et appliquée au génome de *M. smegmatis* (Gallien *et al.*, 2009).

La ré-annotation fonctionnelle des 6 964 protéines obtenues a ensuite été réalisée grâce à au premier niveau d'annotation de notre pipeline. Le bilan d'annotation (Tableau 13) montre que le nombre de séquences annotées (88%) n'est pas significativement différent de celui de l'annotation de séquences d'Alvinella (85%). Les différences notables se trouvent au niveau des annotations en domaines Pfam-A, définition fonctionnelle et numéros EC, où les chiffres pour *M.smegmatis* sont systématiquement supérieurs. Ces meilleurs résultats peuvent être expliqués par la présence de plusieurs génomes de *Mycobacterium* dans les banques de séquences, dont celui de *M.tuberculosis* abondamment annoté. Quant au faible pourcentage de termes GO de type composant cellulaire, il s'explique par la nature procaryote de l'organisme.

Tableau 13 – Bilan de ré-annotation des 6 694 gènes protéiques de *M.smegmatis*

Annotation de niveau 1	Protéines	
Protéines ayant une similarité significative	6 492	(100%)
Domaines Pfam-A	5 128	(79%)
Gene Ontology	5 182	(80%)
Processus biologique	4 474	(69%)
Composant cellulaire	2 393	(37%)
Function moléculaire	4 940	(76%)
Définition fonctionnelle	5 307	(82%)
Enzyme Comission	2 032	(31%)
Niveau 4 (X.X.X.X)	1 855	(29%)
Protéines annotées	5 714	(88%)

Le génome ré-annoté a été déposé à Genbank sous le numéro CP001663. Un groupe d'experts a été réuni par notre collaborateur, JM Reyrat, pour exploiter cette annotation et une publication est actuellement en cours de préparation. A terme, le génome réannoté sera également accessible sur le site GenoList de l'Institut Pasteur (http://genolist.pasteur.fr/).

11.6 Discussion et perspectives

Par l'utilisation du MACS en tant que base de notre protocole d'annotation, nous avons réussi à construire un pipeline universel à même d'annoter de façon robuste environ 85% des séquences fournies et adapté aussi bien à l'annotation intégrative de bactéries qu'à celle d'eucaryotes supérieurs. En le complétant par un deuxième niveau d'annotation avec une logique non plus linéaire, mais avec une logique de réseau, nous avons pu évaluer le répertoire de gènes des banques d'ADNc d'*A.pompejana* qui ont révélé une large gamme de fonctions.

Ce pipeline comporte cependant certaines faiblesses. On note par exemple que la proportion de protéines recevant un numéro EC est modeste chez *Alvinella* (environ 15%). Cela s'explique en partie par les règles de décisions adoptées pour l'assignation d'un numéro EC consensuel. Ces dernières sont simples et ne tiennent pas compte du nombre de séquences que comporte la sous-famille de la protéine annotée. Cette limitation est imposée par la faible quantité de séquences qui disposent d'annotation EC parmi les séquences similaires à *Alvinella* (moins de 10% après une rapide analyse effectuée sur un échantillon de 1 000 MACS annotés par MACSIMS). Il a donc été envisagé d'évaluer la performance des profils des familles d'enzymes de la banque PRIAM (Claudel-Renard *et al.*, 2003) afin d'améliorer la sensibilité de notre protocole d'annotation.

L'estimation du nombre de voies métaboliques complètes KEGG dans le cadre de la validation de la qualité fonctionnelle des banques d'*A. pompejana* devrait elle aussi être

améliorée. Cette estimation a été établie par le pourcentage de numéros EC présents chez *Alvinella* par rapport à ceux présents dans chaque voie. Ceci représente un estimateur médiocre puisqu'il ne prend pas en compte le fait qu'une voie métabolique peut être fonctionnelle même en l'absence des certaines enzymes, du fait d'enzymes ou de réactions alternatives pouvant aboutir au même produit chimique. Pour améliorer ceci, nous nous sommes tournés vers le format KGML (*KEGG Markup Language*) de la banque KEGG PATHWAY. KGML est une représentation XML de KEGG PATHWAY sous forme de graphe. Malheureusement, les données contenues dans les fichiers KGML sont plutôt conçues pour représenter graphiquement les voies métaboliques et ne permettent pas facilement le parcours automatisé de la voie en question. De plus, toutes les voies métaboliques ne sont pas disponibles dans ce format et il est apparu que des informations essentielles sont absentes aléatoirement dans plusieurs fichiers (nom de la voie métabolique, enzyme alternative manquante pour une réaction…). Nous sommes donc retournés à la première méthode, en attendant de remédier à ces problèmes.

Concernant l'utilisation des données STRING, nous nous sommes heurtés au problème de calcul du score combiné permettant d'évaluer la fiabilité d'une interaction. En effet, en plus du score combiné directement pré-calculé, la base de données STRING contient les scores détaillés de chaque type de source d'interaction (co-expression, fusion, données importées…). Ces informations sont fournies pour permettre de recalculer le score combiné d'une interaction en omettant une ou plusieurs sources que l'on ne veut pas considérer. Nous avons remarqué que le score de fouille des résumés de Pubmed grossissait souvent artificiellement le score combiné d'une interaction. En effet, la présence combinée de deux noms de protéines dans le même résumé valide une interaction, en ignorant la valeur syntaxique des phrases. Nous avons donc tenté de re-calculer les scores combinés sans prendre en compte le score de fouille de données Pubmed lors de la création des sous-graphes STRING. Il s'est malheureusement avéré que les données permettant de recalculer ce score combiné, spécifiées au travers des différentes publications de STRING, sont incomplètes. Nous avons signalé ce problème aux personnes en charge de la banque STRING et nous re-créerons les sous-graphes sans prendre en compte la fouille de données Pubmed à la lumière de leurs nouvelles informations.

Enfin, le problème de fond reste l'estimation de la qualité d'une annotation. Ainsi, lors de la vérification manuelle de la qualité de l'annotation, certaines incohérences ont été décelées entre les différents types d'annotation (GO, Pfam, numéros EC, définition fonctionnelle) d'une même séquence. Il faudrait développer une fonction de score objective permettant d'évaluer le niveau de confiance des différentes annotations attachées à une protéine donnée. Cela reste toutefois un vaste problème qui nécessitera en premier lieu l'élaboration

d'un jeu neutre de séquences et d'annotations de référence puisqu'il n'en existe pas encore à l'heure actuelle (Godzik *et al.*, 2007).

12 PUBLICATION 1 : *ALVINELLA POMPEJANA*

Le développement des deux pipelines automatisés décrits précédemment a été réalisé en première intention pour étudier et exploiter les banques de séquences d'ADNc d'*A. pompejana*, même s'ils ont d'ores et déjà été utilisés dans d'autres cadres. Les résultats de ce travail d'annotation au sens large sont à la base de plusieurs projets expérimentaux entrepris au LBGI, au sein du consortium *Alvinella* et dans le cadre du projet européen SPINE2 (Cf. Conclusions et perspectives, page 181).

L'annotation a aussi constitué le point de départ de plusieurs études *in silico* visant à en apprendre davantage sur la biologie de ce ver insolite. Les résultats de l'annotation et de ces études ont fait l'objet d'un manuscrit présenté ci-après, intitulé '*Insights into metazoan evolution from Alvinella cDNAs*' qui va être soumis à *Genome Research* très prochainement. Nous ne reviendrons pas ici sur les résultats de l'annotation elle-même mais nous allons présenter brièvement les résultats des études ultérieures.

Dans un premier temps, nous nous sommes attachés à rechercher les gènes les plus fortement exprimés chez *Alvinella*. Dans la mesure où les banques d'ADNc ne sont pas normalisées, nous nous sommes appuyés sur le nombre de transcrits rattachés à chaque gène. Il nous a cependant fallu au préalable regrouper en cluster les contigs très similaires, susceptibles de correspondre à des isoformes d'un même gène. En parallèle, nous avons recherché le nombre de transcrits rattachés à un domaine Pfam-A particulier. Les deux approches ont clairement révélé que les transcrits les plus abondants sont liés aux mécanismes de la respiration, du stress oxydatif, de la détoxification et de la défense contre les chocs thermiques. Cette surexpression est très certainement liée à l'adaptation d'*Alvinella* à son milieu dont les conditions et niveaux de température, d'anoxie et de métaux lourds varient constamment et rapidement. Parmi les gènes fortement exprimés figurent aussi des gènes de fonction inconnue, ne présentant pas d'homologie avec les protéines actuellement répertoriées dans les banques. Il est alors tentant de penser que ces gènes pourraient eux aussi appartenir à « l'arsenal adaptatif » d'*Alvinella*. Il s'avère qu'une étude de protéomique comparative (J. Mary, manuscrit en préparation) a révélé que l'un d'entre eux est effectivement différentiellement exprimé en fonction du niveau d'oxygène du milieu. Ces gènes inconnus et fortement exprimés constituent donc des cibles prioritaires pour de futures études fonctionnelles sur les mécanismes d'adaptation chez *Alvinella*. Le résultat de protéomique semble confirmer également, de manière indirecte, le bien-fondé de notre méthode de prédiction des séquences codantes puisqu'il valide l'existence d'une protéine sans homologue prédite par notre pipeline.

Nous avons ensuite voulu identifier les gènes exprimés spécifiquement, ou tout au moins différentiellement, entre les trois tissus que sont les branchies (partie antérieure de

l'animal), le tissu ventral (partie médiane) et le pygidium (partie postérieure). Cette étude a été réalisée sur la base des séquences constituant les clusters mentionnés précédemment. Seuls les clusters contenant des EST provenant uniquement d'un même tissu ont été considérés. Nous avons éliminé les clusters comportant des EST de la banque issue d'animaux entiers, ce qui diminue fortement le nombre de candidats. Dans un même souci de rigueur, nous avons fixé à 6 l'effectif minimal d'un cluster, réduisant encore drastiquement le set retenu. Une proportion importante de ces clusters correspond à des gènes sans homologue et donc, sans fonction connue. Les autres clusters ont été analysés à l'aide de DAVID et d'Ingenuity IPA. Même si la base de données de DAVID n'a pas été récemment mise à jour (janvier 2008), DAVID a été préféré à Ingenuity IPA du fait de la classification trop généraliste de ce dernier. Malgré le nombre réduit de clusters annotés retenus pour chaque banque tissu-spécifique, notre étude a tout de même mis en évidence un faible enrichissement de fonctions en lien avec chaque tissu. Ainsi, la banque d'ADNc de branchies comporte un net enrichissement en fonctions d'antioxydation, de suppression des radicaux libres et d'oxydoréduction. La banque de tissu ventral comporte un enrichissement en protéines liées aux mécanismes de défense antibactérienne, ce qui pourrait refléter l'existence de systèmes pour protéger son épithélium des bactéries qui le recouvre. La banque de pygidium démontre un enrichissement en fonctions de différenciation et de prolifération cellulaire, le pygidium étant la zone où s'établit la croissance de l'animal (Paulus et Müller, 2006). En revanche, nous n'avons pas décelé d'enrichissement significatif en gènes liés à l'adaptation aux températures élevées dans la banque pygidium. Un tel enrichissement aurait milité en faveur de température du milieu environnant plus élevée au niveau du pygidium que des branchies, hypothèse controversée au sein de la communauté des spécialistes de la faune hydrothermale (Didier Jollivet, communication personnelle). La proportion important de gènes de fonction inconnue dans chacun des pools tissus-spécifiques ne nous permet cependant pas de réfuter l'existence d'un tel gradient de température.

Enfin, nous avons réalisé plusieurs analyses phylogénomiques à partir des gènes d'*Alvinella* afin d'en apprendre davantage sur l'évolution des Annélides au sein des Métazoaires. Une première étude massive a porté sur les meilleurs *hits* BlastP des protéines d'*A. pompejana*. Les résultats sont assez inattendus puisque *Alvinella* semble être plus proche des Deutérostomiens (49%) et des Cnidaires (12%) plutôt que proche des Protostomiens (25%) auxquels cet organisme appartient. Cette proximité se retrouve dans l'analyse des alignements multiples d'un jeu de 556 familles d'orthologues qui fait apparaître une forte conservation de séquences entre *Alvinella*, les Deutérostomiens et les Cnidaires. Une étude phylogénétique sur un jeu de 76 protéines ribosomales dont les alignements multiples ont été concaténés montre que cette conservation est liée à une vitesse d'évolution lente chez

les Annélides et les Mollusques comparée aux autres espèces de Protostomiens considérées, à savoir des représentants des Ecdysozoaires et des Plathelminthes.

La proximité entre *Alvinella* et les Deutérostomiens s'explique aussi par son répertoire de gènes puisque 203 familles de protéines d'*Alvinella* sont spécifiques aux Deutérostomiens, contre seulement 135 familles spécifiques aux Protostomiens. La présence chez *Alvinella* de ces gènes considérés comme spécifiques aux Deutérostomiens, suggère que le dernier ancêtre commun des Bilatériens possédait un répertoire de gènes beaucoup plus étendu que ce à quoi on pouvait s'attendre. Ces résultats majeurs obtenus chez un représentant des Lophotrochozoaires s'inscrivent dans la tendance actuelle des résultats de génomique comparative obtenus chez les Métazoaires qui révèlent la complexité insoupçonnée des ancêtres des animaux.

Cette proximité entre *A. pompejana* et les Vertébrés à la fois en termes de séquences et de répertoire de gènes confortent *A. pompejana* comme une source idéale de protéines thermostables proches des protéines humaines. Elle pose aussi la question des bases moléculaires de la thermostabilité chez *Alvinella*. L'étude comparative de la composition des protéines chez Alvinella et différentes espèces animales mésophiles réalisée par notre collaborateur, Didier Jollivet, donne un premier élément de réponse puisqu'elle fait apparaître un léger enrichissement en résidus chargés positivement chez *Alvinella*.

Insights into metazoan evolution from *Alvinella pompejana* cDNAs

Nicolas Gagnière[1], Didier Jollivet[2,3], Isabelle Boutet[2,3], Yann Brélivet[1], Didier Busso[1], Corinne Da Silva[4], Françoise Gaill[5], Dominique Higuet[6], Stéphane Hourdez[2,3], Bernard Knoops[7], François Lallier[2,3], Emmanuelle Leize-Wagner[8], Jean Mary[2,3], Dino Moras[1], Emmanuel Perrodou[1], Jean-François Rees[7], Béatrice Segurens[4], Bruce Shillito[6], Arnaud Tanguy[2,3], Jean-Claude Thierry[1], Jean Weissenbach[4], Patrick Wincker[4], Franck Zal[2,3], Olivier Poch[1], Odile Lecompte[1].

[1]Department of Structural Biology and Genomics, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC), F-67400 Illkirch, France; INSERM, U596, F-67400 Illkirch, France; CNRS, UMR7104, F-67400 Illkirch, France; Faculté des Sciences de la Vie, Université de Strasbourg, F-67000 Strasbourg, France;
[2] CNRS, UMR 7144, Adaptation et Diversité en Milieu Marin, Station Biologique de Roscoff, 29682, Roscoff, France;
[3] UPMC Université Paris 6, Station Biologique de Roscoff, 29682, Roscoff, France
[4] Genoscope - Centre National de Séquençage, 2 rue Gaston Crémieux CP5706 91057 Evry cedex
[5] CNRS Institut Ecologie et Environnement (INEE), 3 rue Michel-Ange, 75794, Paris cedex 16, France;
[6] UPMC Université Paris 6, UMR 7138, Systématique, Adaptation et Evolution, Campus de Jussieu, 75005 Paris, France;
[7] Université Catholique de Louvain, Laboratoire de Biologie Cellulaire, Institut des Sciences de la vie, Croix du sud 5, B-1348, Louvain-la-neuve, Belgique;
[8] UMR 7177 CNRS-UDS, LDSM2 Institut de Chimie de Strasbourg, 1 rue Blaise Pascal –BP 296 R8, 67008 Strasbourg cedex, France;

Corresponding author
E-mail lecompte@igbmc.fr; fax 33-3-88-65-32-76.

Running title: *Alvinella pompejana* cDNAs

Keywords: Annelids, cDNA, Metazoa, thermostability, evolution

Abstract

Alvinella pompejana is a representative of Annelids, a key phylum for evo-devo studies that is still poorly studied at the sequence level. *A. pompejana* inhabits deep-sea hydrothermal vents and is currently known as the most thermotolerant Eukaryote on Earth. It represents an outstanding model organism for studying adaptation to harsh physicochemical conditions and for isolating stable macromolecules resistant to high temperatures. We have constructed and analysed four full length enriched cDNA libraries to investigate the biology and evolution of this intriguing animal. Analysis of more than 100,000 reads leads to 15,858 non redundant transcript sequences and to 9,221 protein sequences. Our annotation reveals a good coverage of most animal pathways and networks with an enrichment of transcripts involved in oxidative stress resistance, detoxification, anti-bacterial defence, and heat shock protection. *Alvinella* proteins show a slow evolutionary rate and a high similarity with proteins from Vertebrates compared to proteins from Arthropods or Nematodes. Their composition shows enrichment in positively charged amino acids that might contribute to their thermostability. The gene content of *Alvinella* reveals that an important pool of genes previously considered specific to Vertebrates or Deuterostomes were in fact already present in the last common ancestor of the Bilaterian animals but have been secondarily lost in model invertebrates such as *Drosophila* or *Caenorhabditis*. This pool is enriched in glycoproteins that play a key role in intercellular communication, hormonal regulation and immunity. Our study sheds light on the complexity of animal ancestors and confirms *Alvinella* as an ideal model for studying Vertebrate targets.

Supplemental material is available online at www.genome.org. All data described are stored in the *Alvinella* online database available at http://alvinella.igbmc.fr/Alvinella/. The sequences have been submitted to the EST section of the EMBL database under accession numbers FP489021 to FP539727 and FP539730 to FP565142.

Introduction

Annelids, commonly known as segmented worms, are typical triploblastic and coelomate animals belonging, together with numerous invertebrates, to Protostomes. Annelids, and especially polychaetous annelids, are considered to be ideal systems for understanding evolution and development in animals (for recent reviews, see (De Robertis 2008; McDougall et al. 2008)). Fossil records (Morris 1998), as well as comparative morphology studies (Arendt and Nubler-Jung 1994), suggested that the urbilaterian (the last common ancestor of bilateral symmetric animals) may have resembled annelids. If such an assumption is difficult to verify, it is widely accepted that polychaetes exhibit many ancestral traits in their body plan and embryonic development (Irvine and Martindale 1996).

Despite this long history as evo-devo model organisms, polychaete annelids, and more generally Lophotrochozoan representatives, are still poorly represented in sequence databases as sequencing projects have mainly focused on Deuterostomes (Chordates and Echinoderms) and Ecdysozoa, i.e. molting Protostomes that include arthropods and nematodes (Aguinaldo et al. 1997). Enlargement and diversification of the sequencing project panel is playing a decisive role in obtaining a more realistic picture of animal evolution. For instance, the analysis of genomic loci in the marine polychaete *Platynereis dumerilii* has revealed the intron-rich nature of annelid genes (Raible et al. 2005). More recently, the genome of a bilaterian sister group, the cnidarian sea anemone *Nematostella vectensis,* has turned out to be more complex than expected, with a gene repertoire, exon-intron structure, and large-scale gene linkage more similar to vertebrates than to flies or nematodes (Putnam et al. 2007).

Among polychaete annelids, *Alvinella pompejana* (Desbruyeres and Laubier 1980), the “Pompeii worm”, has attracted attention since it is currently considered as the most thermotolerant eukaryote on Earth, withstanding the largest known chemical and thermal ranges (from 5 to 105°C) (Chevaldonné et al. 1992; Cary et al. 1998). This tube-dwelling worm forms dense colonies on the surface of hydrothermal chimneys and can withstand long periods of hypo/anoxia and long phases of exposure to hydrogen sulphides (Le Bris and Gaill 2007). *A. pompejana* specifically inhabits chimney walls of hydrothermal vents on the East Pacific Rise (Desbruyeres et al. 1998; Pradillon et al. 2005). It often co-occurs with *Alvinella caudata*, a very closely related species, and can be found in variable proportions according to the chemical conditions. The chimney walls are characterised by high flows of vent fluid, and therefore the highest temperatures for vent metazoans (temperatures usually

range between 25 and 60˚C, with exceptional bursts up to 105˚C (Chevaldonné et al. 1992; Cary et al. 1998; Le Bris and Gaill 2007)), as well as high concentrations of potentially toxic compounds (e.g. H_2S). The thermotolerance of alvinellid worms has been confirmed by laboratory observations of *Paralvinella sulfincola* thermotaxis (Girguis and Lee 2006). To survive, *Alvinella* has developed numerous adaptations at the physiological and molecular levels, such as an increase in the thermostability of proteins and protein complexes (Dahlhoff and Somero 1991; Jollivet et al. 1995; Burjanadze 2000; Sicot et al. 2000; Piccino et al. 2004; Henscheid et al. 2005). As such, *A. pompejana* constitutes a precious source of thermostable proteins and macromolecular complexes of eukaryotic origin for the biochemical, biophysical or structural characterisation of proteins of fundamental or biomedical relevance. It has been selected as a model organism for structural studies by the Structural Proteomics IN Europe 2 (SPINE2) initiative. The pertinence of the model is confirmed by the recent study of the superstable superoxide dismutase recombinant protein (Shin et al. 2009). The crystal structure at 0.99 Å resolution reveals anchoring interaction motifs in loops and termini, accounting for the enhanced stability of the *A. pompejana* protein compared to its human homolog.

Here, we report the construction, sequencing and analysis of four full-length enriched cDNA libraries of *A. pompejana*. One of these libraries has been constructed from complete animals after removal of the dorsal tegument, which harbours an episymbiont community of Epsilonproteobacteria (Grzymski et al. 2008). The other three libraries have been prepared from distinct tissues: the gills located at the anterior part of the animal and oriented towards the outside of the tube, the ventral tissues and the posterior region. This latter includes the pygidium and the subterminal growth zone of the animal where cell proliferation takes place. These three tissues, radically different with respect to their physiological role, have been chosen in order to improve the transcriptome coverage. They also represent samples along the antero-posterior axis of *A. pompejana* since it has been reported that the body of the animal experiences a temperature gradient, the posterior part of *A. pompejana* being exposed to higher temperatures (Cary et al. 1998; Le Bris et al. 2005).

From these libraries, approximately 100,000 sequences were obtained, filtered and assembled, resulting in 4,993 contigs and 10,865 singletons (15,858 unigenes). Analysis of the cDNA sequences allowed the determination of 9,221 protein sequences that were annotated using a new integrative functional annotation pipeline. With regard to the pathways and interaction networks mapped by our sequences, the four cDNA libraries provided a good coverage of the *A. pompejana* gene repertoire with a clear enrichment in transcripts related to respiration, oxidative stress resistance, detoxification, anti-bacterial defence and heat shock protection. Compared to other available

metazoan sequences, *A. pompejana* proteins are enriched in the positively charged residues lysine and arginine. Such a composition bias may contribute to the thermostability enhancement of the Pompeii worm proteins. From an evolutionary perspective, our analysis reveals striking similarities between Annelids and Deuterostomes, both in terms of sequence conservation and gene repertoires that raise interesting issues on the nature of the Bilaterian ancestor. The sequences, assembly and annotation are accessible through a user-friendly web site. They represent a significant contribution to the successful exploitation of *A. pompejana* proteins as valuable models for human protein targets, and will hopefully stimulate future research on metazoan evolution and adaptation.

Results

cDNA libraries and ESTs assembly

Four non normalized libraries enriched in full length cDNA have been constructed, with an average insert size of 2.5-3 Kb. The first library has been prepared using whole adult individuals while the others have been constructed from dissected tissues: gills, ventral tissue, and pygidium. A total of 100,177 chromatograms were processed by our semi-automated assembly pipeline (see Methods). 76% of the initial raw sequences fulfil our quality criteria, and have a mean length of 674 bp (Table 1). The percentage of sequences exhibiting a poly(A) tail ranges between 8 and 12% in the different libraries, except in the ventral tissue library in which 42% of the sequences contain a poly(A) tail.

The assembly of the 76,134 selected sequences (12.4 Mb) was performed by CAP3 using stringent parameters to avoid misassembly problems such as creation of chimeric contigs or combination of paralogs. We performed a global assembly of all sequences as well as a separate assembly for each library in order to estimate their respective redundancy (Table 1). Libraries generated by the CloneMiner™ cloning method exhibit a lower level of redundancy that could be partly explained by the reduced size of the corresponding sequences. However, the significantly low level of redundancy (65%) observed for the pygidium library may also reflect a higher diversity of gene expression in this tissue.

The global assembly yielded 4,993 contigs and 10,865 singlets (15,858 unigene sequences) with a redundancy of 79%. Contig size ranges from 2 to 7,845 reads with a mean of 13 and a median of 3 (see Supplemental Fig. S1). The average length of contigs is 1,017 bp. Taking into account our conservative approach, each contig and singlet of this assembly ideally represents a unique version of an expressed gene, *i.e.* paralogs, divergent alleles or splicing variants should not coalesce into the same contig. We performed an additional step of clustering using BLASTN (Altschul et al. 1997) to join alleles or close isoforms and to estimate the number of different genes. This resulted in 14,682 non redundant sequences (573 clusters, 4,251 contigs and 9,858 singlets).

cDNA characterization

The 15,858 unique sequences from the global assembly were analysed using two independent methods (ESTScan and a similarity-based approach) to determine the boundaries of the CoDing Sequences (CDS) and UnTranslated Regions (UTR). 7,353 CDS were created by the similarity method. Among the 8,505 remaining sequences without significant homology in public protein databases,

1,868 achieved the ESTScan cut-off score (see Methods) leading to a total of 9,221 coding regions (including 2,932 complete CDS), 7,247 5'UTR and 7,465 3'UTR.

We used the 2932 cDNA with a complete CDS as a reference set to determine cDNA features. The 935 complete 3' UTR of the reference set exhibit an average length of 319nt, in agreement with the mean length of 344nt calculated for Annelid complete 3'UTR. The mean length of 5'UTR (including partial and complete UTR) is 138nt. The complete CDS lengths range from 90nt to 2,694nt with a mean of 540nt (see the distribution of the corresponding protein lengths in Supplemental Fig. S2). As previously observed in eukaryotic mRNAs (Pesole et al. 2000; Zhang et al. 2004; Bechtel et al. 2008), the mean GC content in *A. pompejana* is higher for the 5'UTR (45.7%) than for the 3'UTR (39.7% without poly(A) tail) and is comparable to the value observed in Annelids (43.7% and 34.1% respectively, data compiled from UTRdb (Mignone et al. 2005). The mean GC content in CDS is 46.2%.

To investigate the evolutionary model driving the GC content in *A. pompejana*, an in-depth GC analysis was performed on the 84 almost complete mRNA coding for ribosomal proteins, including 15 genes expressed at low levels associated with the mitochondrial ribosome, and a set of genes with mid-to-high expression associated with the nuclear ribosome. The analysis shows that the GC3 content of CDS (0.481 ± 0.075) was significantly higher than the GC content of CDS (0.422 ± 0.029) and both UTR regions (0.389 ± 0.051; pairwise t-test, $p< 0.0001$). Unexpectedly, the GC3 was found to be constant regardless of the length of the coding regions ($F= 0.92$, p-value = 0.341), although a significant positive relationship exists between the cDNA length and its level of transcription ($F= 18.18$, p-value = $5\ 10^{-5}$) when estimated from the number of gene repeats in cDNA libraries. This number ranges from one copy (mt S proteins) to 223 copies (P0). Another striking result was the non linear evolution of GC3 content of ribosomal protein transcripts with the level of gene expression. Both GC3 (CDS) and GC (UTRs) contents rapidly increased with the number of copies until they reached a plateau at a threshold value of c.a. 25-30 copies (see Supplemental Figure S3). The GC3 and GC content asymptotic values of CDS and UTR were close to 0.55 and 0.40, respectively. The correlation was significant (coding regions: $F = 9.41$, p-value = 0.003), although it is no longer significant when the mt ribosomal-protein genes are removed from the dataset. Analysis of codon usage in the ribosomal set revealed that eight of the most frequent codons are terminated by C or G (Phe, Leu, Tyr, His, Gln, Asn, Lys and Glu), seven by A or T (Ser, Pro, Thr, Asp, Cys, Arg, Gly) and three by two equally-frequent codons (Ile, Val and Ala). The frequency of 11 codons (TAC, TCG, CGT, CAC, CGC, CTG, ATC, ACG, AAG, GCC, GAG) increases with the level of expression. Among these favoured codons, 10 end with G or C.

Integrative functional annotation

The 7,353 protein sequences from the global assembly that exhibit homology in the protein databases were analysed using a new integrative functional annotation pipeline (Gagnière *et al*., manuscript in preparation). The originality of this pipeline lies in the exploitation of the evolutionary context of the protein sequences based on a clustered Multiple Alignment of Complete Sequences (MACS) (Lecompte et al. 2001). The sequence clustering is performed by MACSIMS (MACS Information Management System) (Thompson et al. 2006). The sequences are then analysed in the framework of the overall family and subfamily, allowing a reliable propagation of sequence annotations in conserved regions.

The Pfam-A (Finn et al. 2008) annotations related to domains and protein families are provided by MACSIMS. Gene Ontology (GO) annotations are added by GOAnno (Chalmel et al. 2005). Then, a functional definition and an Enzyme Commission (EC) number (if applicable) are assigned by in-house programs that process the annotations of the subfamily members and generate consensual annotations. Thanks to this novel pipeline, 6,252 (85.0%) of the 7,353 protein sequences with homology were annotated with either a text mining definition, an EC number, Pfam-A domains or a Gene Ontology term (Table 2). 4,767 sequences (64.8%) have at least one Pfam-A annotation and 1,607 different Pfam domains are represented in our databases. 5,949 sequences (80.9%) were assigned at least one GO annotation. The repartition of *A. pompejana* annotated proteins in GO classes is shown in Supplemental Fig. S4. A textual definition has been assigned to 4,611 (62.7%) sequences. Only 1,243 sequences (16.9%) have an EC number attributed, but most of them reach level 4 of the EC classification and 487 distinct EC numbers are represented.

This primary annotation has been used to map the *A. pompejana* query proteins to the networks of the KEGG and EMBL STRING databases. The reconstructed metabolic pathways and interaction networks constitute an integrative second level annotation which is essential to the study of the biological processes at work in *A. pompejana.* Among the 345 metabolic reference pathways of the KEGG database (release 48.0), 202 pathways have been populated by at least one *A. pompejana* protein, and 82 pathways (40.6%) have been completed at a level greater than 50% (in terms of number of distinct enzymes). Since the reference pathways are highly redundant (several enzymes can catalyse the same reaction), the *A. pompejana* cDNA appear to provide a good coverage of a large panel of metabolic pathways. Common pathways such as glycolysis, gluconeogenesis, citrate cycle, purine and pyrimidine metabolism are functionally complete or almost complete. More

specific pathways are also well represented, such as androgen and estrogen metabolism or steroid biosynthesis, as illustrated in Supplemental Fig. S5.

3,582 *A. pompejana* proteins (48.1%) have also been mapped to human networks in the STRING database, which were first cut into smaller sub-networks (see methods). 68.8% of these human sub-networks were populated by at least one *A. pompejana* homolog. Supplemental Fig. S6 shows such a mapped sub-network, mainly representing proteins involved in the Bardet-Biedl syndrome. As previously observed with the KEGG pathways, the STRING network mapping emphasizes the broad coverage of our cDNA databases and provides an integrative framework for the interpretation of *A. pompejana*'s proteome features.

Gene expression level

To obtain an overview of the relative expression levels of *A. pompejana* genes, we used two complementary indicators: the size of the cDNA clusters (in terms of number of reads) (Table 3) and the abundance of transcripts corresponding to a PFAM-A domain (Figure 1). Domains involved in protein-protein or protein-DNA/RNA interactions are particularly abundant (calcium-binding domain EF-hand, WD-40 repeat, RNA recognition motif RRM_1, ankyrin…), as frequently observed in studies of eukaryotic transcriptomes. The highly expressed genes also include genes encoding extracellular structural proteins, such as collagen, as well as cytoskeleton proteins (actin, myosin, tropomyosin, calponin, tubulin, troponin). However, both approaches reveal that the most abundant transcripts are clearly linked to respiration, oxidative stress and/or detoxification: cytochrome c oxidase, cytochrome b, intracellular and extracellular globins, mitochondrial substrate/solute carrier, redoxins, glutathione peroxidase, rhodanese. Transcripts involved in anti-bacterial defence and heat shock protection are also prevalent, although to a lesser extent.

In addition to these known proteins, the highly expressed genes include several "hypothetical proteins" (Table 3). They exhibit sequence segments with a biased residue composition and have no significant similarity with known proteins, except in some cases for low complexity regions. One of these specific proteins, TERA02189, belongs to a family of proteins that are highly conserved in *A. pompejana*. A comparative proteomics study (J. Mary et al., submitted manuscript) has revealed that three members (TERA02082, TERA02935 and TERA08242) of this family are differentially expressed depending on the oxygen concentration. This oxygen-responsive gene may be involved in response to hypoxia or oxidative stress. No experimental clues are available for the other hypothetical proteins listed in Table 3 but, considering the functional profile of the known genes, they may also include additional novel proteins involved in oxidative stress resistance, detoxification or heat response.

Tissue differential expression

We investigated the differential gene expression among tissues by comparing the three tissue-specific libraries. Clusters containing sequences from the whole animal library were removed from the analysis. 273, 192 and 796 clusters are exclusively found in the gill, ventral tissue, and pygidium libraries, respectively. According to these results, the pygidium library is once again distinguished by its high degree of gene diversity. As genes expressed at very low levels can be accidentally sampled in a cDNA library, we excluded clusters containing less than 6 sequences from the subsequent analysis. After this harsh selection, a set enriched in tissue-specific transcripts was observed for each tissue: 66 clusters for the gills, 55 for the ventral tissue and 98 for the pygidium. These transcripts may mainly correspond to tissue-specific genes or divergent isoforms. Very few of them are annotated (Supplemental Table S1) which may be explained by their tissue specialisation.

The set of 14 annotated transcripts found exclusively in the gill library shows a slight enrichment in antioxidation and free radical removal (P-value=0.0025) and in oxidoreductase activity (P-value=0.039). It includes notably a peroxiredoxin, a polyamine oxidase, an inositol oxygenase as well as an aquaporin-8 homolog, a fibrinogen-like protein, a cathepsin D and a pterin-4-alpha-carbinolamine dehydratase.

For the ventral tissue, 4 of the 8 annotated transcripts may correspond to specific isoforms of basic functions. The other 4 proteins are involved in defence against pathogens. TERA02620 contains a trefoil domain found in a wide variety of extracellular eukaryotic proteins, notably in proteins involved in defence against microbial infection that protect the epithelia from the external environment (Hauser et al. 1990). TERA02630 is homologous to the Niemann Pick type C2 protein. It possesses a lipid-recognition domain that has been reported to be involved in the recognition of pathogens (Ichikawa et al. 1998; Ao et al. 2008). TERA02465 and TERA03170 both belong to the whey acidic protein (WAP) family that contains proteins involved in homeostatic control of inflammation and antibacterial/antifungal activity (see (Bingle and Vyakarnam 2008)).

Among the 7 annotated transcripts exclusively found in the pygidium library, 4 are involved in cell proliferation and differentiation: a protein kinase C (PKC), a matrilin, a calcium and integrin binding protein 1 (CIB1) and a Ras-like GTP binding protein RhoA homolog. In human, the PKC isoenzymes are key signalling components involved in the regulation of normal cell proliferation, differentiation, polarity and survival (Fields et al. 2007). CIB1 induces cell migration in mammals by enhancing focal adhesion formation and thus cell spreading (Naik and Naik 2003). Matrilins mediate cell attachment through the formation of filamentous networks (Mann et al. 2007) and may serve as markers for

cellular differentiation (Deak et al. 1999). RhoA is involved in regulating the assembly of focal adhesions and actin stress fibers.

Phylogeny and molecular evolution

To address the evolutionary relationships between *A. pompejana* and other animals, we first investigated the phylogenetic position of *A. pompejana* using a pool of 76 ribosomal proteins. A phylogenetic tree reconstruction was performed on the concatenation of the corresponding multiple alignments using MrBayes (Figure 2). The obtained topology is in agreement with the phylogenetic tree proposed by Aguinaldo *et al.* with the exception of *Schistosoma japonicum* (Platyhelminthes) that clustered within the Ecdysozoa represented by Arthropods and Nematodes, instead of Lophotrochozoa (represented by Molluscs and Annelids). Compared to other Protostomes used in this study, the annelids *A. pompejana* and *Lumbricus rubellus* and the mollusc *Argopecten irradians* show relatively small branch lengths indicating a slow evolutionary rate in these lineages. This is particularly true when compared to the parasitic worms *Schistosoma* or *Caenorhabditis,* which are rapidly evolving species. This might lead to a long-branch attraction artefact between these two lineages.

Differences in the rate of evolution are reflected in the percent identity between orthologs observed in a pool of 556 unambiguous ortholog families (86,727 positions without gaps) conserved in 6 major Metazoa lineages (Table 4). *A. pompejana*, *Homo sapiens* and *N. vectensis* exhibit high sequence conservation relative to *S. japonicum* or *C. elegans* while *D. melanogaster* appears intermediate.

Despite this slow evolutionary rate, *A. pompejana* proteins show a biased composition compared to their orthologs from 5 major Metazoa lineages (Vertebrates, Arthropods, Nematodes, Platyhelminthes, Cnidaria) (Figure 3). The amino acid composition differed significantly among taxa (Homogeneity statistic: $\chi 2$ = 0.01153 G = 0.01153). *A. pompejana* exhibits the highest proportion of charged amino acids (nearly 25.5%): a characteristic also shared by the cnidarian *N. vectensis* (25.4%). This is mainly due to an increase of the positively-charged amino acids lysine and arginine (12.6%).

Comparative genomics

We then classified all the available repertoire of *A. pompejana* proteins (7,353 protein sequences with a homolog in the databases) by analysing the taxonomic distribution of the highest scoring BLAST hit. Thirteen large taxonomical groups covering the whole tree of life were considered: Deuterostomia, Arthropoda, Nematoda, Platyhelminthes, Cnidaria, “Other Metazoa”,

Choanoflagellida, Protists, Fungi, Viridiplantae, “Other Eukaryota”, Prokaryotes and Viruses. The overall taxonomic distribution (Figure 4) is coherent, with more than 90% of the best hits belonging to Metazoa. For the remaining 10%, most of the similarities are very weak and cannot be used to draw any significant conclusion, with the notable exception of the Prokaryotic group. In this latter set, some cases of strong conservation between *A. pompejana* and prokaryotic proteins are present and may be linked to contamination of *A. pompejana* cDNA librairies by bacterial DNA from the episymbiont community found on its dorsal surface.

Within Metazoa, the taxonomic distribution is rather surprising since approximately 50% and 12% of the *A. pompejana* proteins appear closer to Deuterostomes and Cnidaria sequences, respectively. Only one quarter of the proteins are closer to Protostomes (Arthropods, Nematodes and Platyhelminthes) and almost no relationship appears between *A. pompejana* and Platyhelminthes, although Annelids and Platyhelminthes both belong to Lophotrochozoa. These unexpected relationships, in particular the apparent proximity between *A. pompejana*, Deuterostomes and Cnidaria, may be explained by the slow evolutionary rate observed in *A. pompejana* (see above) but could also be accounted for by differential gene losses/acquisitions in animals.

We explored the latter option by determining the sets of *A. pompejana* proteins whose orthologs are found exclusively in some given taxa. Only 135 protein families present in *A. pompejana* are specific to Protostomes, with only 13 conserved in all the Arthropod, Nematode and Platyhelminthe representative species. In contrast, 203 *A. pompejana* proteins belong to families or superfamilies specific to Deuterostomes. This Deuterostomia-set is significantly enriched in glycoproteins (34 proteins out of 203) that play a key role in many biological processes, in particular intercellular communication and adhesiveness, hormonal regulation, or immunity. For instance, the protein TERA08399 belongs to the superfamily of secreted cysteine rich factors and its N-terminal domain sequence exhibits the idiosyncratic features of the IGFBP (Insulin-Like Growth Factor Binding Protein) family reported to be vertebrate-specific (Vilmos et al. 2001). Another noteworthy result is the enrichment in proteins containing an epidermal growth factor (EGF)-like domain that is frequently found in the extracellular part of membrane-bound proteins or in proteins known to be secreted. In addition, the Deuterostomia-set is enriched in proteins involved in the I-kappaB kinase/NF-kappaB cascade or in death-domain containing proteins that can be involved in the regulation of apoptosis and inflammation or linked to innate immunity. This includes close homologs of the CRADD and DEDD/DEDD2 protein families (TERA03000 and TERA04373, respectively) that play a role in the stress-induced apoptosis signalling pathway and are important mediators for death receptors (Alcivar

et al. 2003; Tinel and Tschopp 2004). If we exclude the possibility of horizontal gene transfer, these different examples suggest that important functions previously considered as specific novelties of Deuterostomes were in fact already present in the Bilaterian ancestor and were subsequently lost in Ecdysozoa and Platyhelminthe model species.

In addition to this Deuterostomia-set, 147 *A. pompejana* protein families are specifically present in both Deuterostomes and Cnidaria, while 32 are specifically found in both Cnidaria and at least one Protostome. These 147 families present in the last common ancestor of the Eumetazoa may also have been lost in the Ecdysozoa and Platyhelminthe representatives. Interestingly, this set exhibits an enrichment in the selenium binding function. Notably, it includes the homolog of the selenoprotein N involved in the regulation of oxidative stress and calcium homeostasis (Lescure et al. 2009) as well as the homolog of an iodothyronine deiodinase that participates in thyroid hormone metabolism.

Differential gene losses in Ecdysozoa are also observed for more ancestral genes. For instance, the Pompeii worm possesses homologs of the component of the phagocytic NADPH oxidase (Nox) (gp91phox and p22phox) and of some of its regulatory proteins (p47phox, p67phox) that play a critical role in innate immunity of Deuterostomes. p47phox and p22phox genes are present in the Cnidaria *N. vectensis* and the unicellular choanoflagellate *M. brevicollis,* but are absent in several lineages of ecdysozoans including *Drosophila* and *Caenorhabditis*.

A. pompejana database and website

The data are stored in a relational database that maintains fine grained information about (i) the library origin of each clone, (ii) the nucleic sequences and their phred quality values, (iii) the different assemblies and their associated parameters, (iv) the predicted protein sequences and (v) all the results of the annotation process.

We developed the *A. pompejana* cDNA website in order to offer an ergonomic interface for database querying and data visualisation (Figure 5). Two modules are available for querying the database: a homology search module using NCBI BLAST and a module for performing full-text search module of all annotations. The search results offer the possibility to select a subset of relevant data for display. Different views are available: (1) nucleic cDNA sequence (2) EST trace and six-frame translation (3) contig schematic representation (4) Consed-like (Gordon 2003) contig alignment view (5) MACSIMS annotated protein alignment with customisable features display (6) integrative view of the

annotation process (text mining definition, EC number, Gene Ontology, Pfam-A domains, KEGG pathways mapping, EMBL STRING mapping).

Discussion

A large pool of annotated annelid sequences

The construction and sequencing of four non normalized cDNA libraries from *A. pompejana* resulted in 15,858 unique cDNA sequences and 9,221 protein sequences annotated using an original integrative annotation protocol. Special attention was paid to making the sequences and annotation features (definitions, Pfam-A domain, EC numbers, GO terms) accessible through a user-friendly web interface, allowing an intuitive enquiry of the underlying database, as well as convivial data visualisation through dedicated tools, ranging from trace visualization to the interactive display of the alignment of *A. pompejana* protein sequences and their families.

In the absence of a complete annotated genome of a close relative of *A. pompejana*, we cannot estimate the coverage of our cDNA libraries. Such an estimation would be further complicated by the diversity of genome size reported in polychaete annelids. The size ranges from 58 Mb in *Dinophilus gyrociliatus* to 7 Gb in *Nephtys incisa* (Gregory, T.R. (2009) Animal Genome Size Database. http://www.genomesize.com.), *A. pompejana* being intermediate with an estimated size of 675 Mb (Bonnivard et al. 2009) for 32 chromosome pairs (Dixon et al. 2009). However, with 202 of 345 KEGG metabolic pathways populated by at least one protein from *A. pompejana*, a large panel of protein functions is represented in our database. Our analysis also revealed an important pool of specific genes including highly expressed genes that may participate in *A. pompejana* adaptation as well as tissue-specific genes. Considering the fact that only 3,218 annelid protein entries are available in the Uniprot database at the time of writing, our pool of 9,221 protein sequences should be valuable to the wider scientific community wishing to understand metazoan evolution and for future annotation of genome sequences from annelids and related phyla.

A. pompejana genes under a neutral evolutionary process?

The analysis of the GC content of transcripts encoding ribosomal proteins and its variations both across synonymous and non-synonymous sites and between coding and UTR regions revealed interesting features in *A. pompejana*. The GC3 of ribosomal genes was much higher than the overall GC of both UTR and coding regions, and favored codons are predominantly G/C-ended. Classically, biases in synonymous codon usage are explained by two alternative but non exclusive models: a neutral mutational-bias and a selective model (Duret 2002). The mutational-bias model's expectation corresponds to a positive relationship between the base composition of synonymous sites and their neighboring silent sites (i.e. UTR and/or introns). In agreement with this model, we found a positive

correlation between the GC3 and the GC(UTR), suggesting that both GC classes are evolving in the same way. The selective model postulates a co-evolution between synonymous codon usage and the abundance of tRNA to optimize the translation efficiency (notion of 'optimal' codons). According to Eyre-Walker (Eyre-Walker 1996), selection maximizes the speed of the translation and minimizes the costs of proofreading, resulting in a codon usage correlated with the expression level and mRNA length. Such correlations have been observed in *Drosophila* and *Caenorhabditis* (Duret and Mouchiroud 1999) but not in Vertebrates (Duret 2002). In *A. pompejana*, there is no correlation between the GC3 and the level of gene expression within the set of nuclear ribosomal genes used. Even if differences occurred when considering the level of ribosomal gene expression, this was mainly due to the presence of two distinct sets of ribosomal protein genes (i.e. mitochondrial ribosome versus nuclear ribosome). Additionally, no correlation was found between GC3 and cDNA length. Thus, there is no evidence for a selective process acting on silent sites although extended analyses on GC content bias over the genome and the whole transcriptome (dos Reis and Wernisch 2009) are clearly necessary to validate the predominance of the neutral model in *A. pompejana*.

Adaptation to hypoxia, oxidative stress and heavy metals

As an endemic species of the hydrothermal vent ecosystem colonizing the chimney walls, *A. pompejana* has to deal with very variable conditions that are the result of a chaotic mixing of vent fluid (350°C, anoxic, CO_2- and sulphide-rich) and deep-sea water (2°C, mildly hypoxic). In this environment, oxygen and CO_2 concentrations, pH and sulphide levels vary quickly and over a wide range. These challenging environmental conditions are clearly reflected in the highly expressed gene pool detected in our study of non-normalized libraries, that mainly includes genes involved in oxygen homeostasis, oxidative stress resistance and detoxification. Among these genes, the most important fraction corresponds to proteins from the respiratory chain and three main types of hemoglobins (Hb) reported in Alvinellidae, namely a non-circulating cytoplasmic globin, the extracellular giant annelid hexagonal bilayer HBL-Hb of the vascular system and the circulating intracellular Hb found in the coelomic fluid (for a review, see (Hourdez and Weber 2005)). The abundance of Hbs in *A. pompejana* and their high oxygen affinities (Hourdez et al. 2000) may be determinant in the respiratory adaptation to hypoxic/anoxic environments. Interestingly, the set of highly expressed genes includes an *A. pompejana* specific family encoding oxygen responsive proteins (Mary et al., submitted manuscript). The molecular function of these proteins is unknown but they constitute potential candidates that may contribute to oxygen homeostasis.

Despite the hypoxia encountered in its environment, the Pompeii worm can be subject to exogenous oxidative stress (Marie et al. 2006). High levels of ferrous iron and sulphide have been reported to favour the formation of reactive sulphide species (RSS), an analog to ROS. This is in agreement with the high level of expression observed for the major antioxidative enzymes *in A. pompejana*: Mn and Cu/Zn superoxide dismutases, peroxiredoxins, glutathion peroxidases (GPX), thioredoxin. Interestingly, no catalase cDNA was detected in our set, suggesting that H_2O_2 formation may not be the most common mechanism of detoxification. An earlier study (Marie et al. 2006) also questioned the presence of catalase in *Paralvinella grasslei*, a close relative of *A. pompejana*. The authors suggested that alternative H_2O_2-scavengers, such as antioxidant osmolytes or other enzymes might replace the catalase activity. Indeed, considering the diversity and level of expression of glutathione peroxidases and peroxiredoxins in *A. pompejana*, SOD-derived H_2O_2 could be degraded by peroxidases rather than by catalases as suggested by Dixon and colleagues (Dixon et al. 2002).

In addition to hypoxia and oxidative stress, *A. pompejana* has to face large amounts of heavy metals. Invertebrates possess a variety of cellular detoxification pathways that reduce the concentrations of potentially toxic metals circulating in the blood (reviewed in (Ahearn et al. 2004)). These pathways include metal binding by cysteine-rich proteins known as metallothioneins followed by their elimination through the lysosomal endomembrane system. We detected a single EST coding for a metallothionein-like protein in our library suggesting that involvement of metallothioneins is not the major detoxication process. However, we cannot exclude the possibility that the pool of highly expressed genes of unknown function contains genes coding for new metallothionein-like proteins, since the pool of unknown genes appears enriched in cysteine-rich proteins. Another alternative for heavy metal detoxication is the intracellular sequestration in specific vacuoles producing solid granules (Ahearn et al. 2004). This would be in agreement with the presence of arsenic, zinc and copper detected in *A. pompejana* epidermal cells (Gaill et al. 1984) and the production of a large amount of iron-containing granules by *A. pompejana* mucocytes (Vovelle and Gaill 1986; Gaill and Hunt 1987). Finally, rhodanese also appear preponderant among highly expressed genes. Rhodanese can perform a variety of roles (reviewed in (Cipollone et al. 2007)), including the modulation of general detoxification processes and the maintenance of redox homeostasis.

Thermo-adaptive features in amino-acid composition

As the most thermotolerant eukaryote known to date, the Pompeii worm clearly provides a unique model for the study of adaptation to high temperature in this domain of life. Its thermal regime generally fluctuates between 25 and 60˚C, with exceptional bursts up to 105˚C (Chevaldonné et al.

1992; Le Bris and Gaill 2007). These high and variable temperatures require adaptations at the physiological and molecular levels, even though we are far from the optimal temperature range reported in hyperthermophilic prokaryotes. At the molecular level, several studies have revealed the higher thermostability of *A. pompejana* proteins and complexes compared to their orthologs from other eukaryotes (Dahlhoff and Somero 1991; Jollivet et al. 1995; Burjanadze 2000; Sicot et al. 2000; Piccino et al. 2004; Henscheid et al. 2005; Shin et al. 2009). Our analysis of *A. pompejana* protein sequences revealed a significant increase of polar charged residues compared to their eukaryotic orthologs. This excess of charged residues may enhance protein stability in thermophilic eukaryotes, notably by increasing salt-bridges. This would be in keeping with the structural analysis of the superoxide dismutase of *A. pompejana* (Shin et al. 2009) that suggests that extra salt-bridged interactions may be involved in the superstability of this protein.

Many comparative sequence and tertiary structure studies have been undertaken in an attempt to understand the molecular basis of protein thermostability in thermophilic and hyperthermophilic prokaryotes (Vogt et al. 1997; Haney et al. 1999; Szilagyi and Zavodszky 2000; Nishio et al. 2003; Berezovsky and Shakhnovich 2005; Robinson-Rechavi et al. 2006). From these studies, thermoadaptive molecular features appear to be multiple and variable among prokaryotes. Berezovsky and Shakhnovich (Berezovsky and Shakhnovich 2005) suggested two distinct evolutionary strategies to conciliate these conflicting observations. In prokaryotes with an ancestral thermophilic character (i.e. Archaea such as *Pyrococcus*), proteins may be significantly more compact and more hydrophobic than their mesophilic counterparts. Conversely, organisms that recently colonized a hot environment such as the bacteria *Thermotoga maritima,* may have evolved under a more ‘‘sequence-based’’ mechanism of thermostability. In this latter case, a few charged amino acid replacements or amino acid deletions increased occurrences of hydrogen bonds and inter-subunit electrostatic interactions or decreased the length of surface loops respectively. The enrichment in charged residues detected in *A. pompejana* suggests that the sequence-based mechanism of thermostability reported to hold in *Thermotoga* may also apply to *A. pompejana*. However, the molecular basis of eukaryotic thermotolerance is probably more complex than that of the bacterial/archaeal domains, and additional studies are needed to decipher the molecular basis of thermostability in *A. pompejana*. These may include massive structural comparisons between *A. pompejana* proteins and mesophilic homologs, as well as in-depth comparisons of amino acid compositions between close relatives of *A. pompejana* since thermoadaptive features are often masked by a background of evolutionary sequence divergence.

Proximity between Annelids and Vertebrates

Our analysis reveals that the slow evolutionary rate previously observed in diverse species of the polychaete lineage (Aguinaldo et al. 1997; Raible et al. 2005) also holds in *A. pompejana*, despite its challenging habitat. It particularly contrasts with the fast evolutionary rates observed in other Protostome model species with completely sequenced genomes or complete proteomes such as *C. elegans* or *S. japonicum*. This strong heterogeneity may partly explain the contradictory results obtained in whole-genome based phylogenetic studies that favoured either the classical "Coelomata hypothesis" positioning Nematodes and Plathyhelminthes as early branching clades (Wolf et al. 2004; Ciccarelli et al. 2006) or the new animal phylogeny dividing Protostomes in Ecdysozoa (including Arthropods and Nematodes) and Lophotrochozoa (including Annelids, Molluscs and Platyhelminthes) (Putnam et al. 2007). Our bayesian and maximum parsimony (data not shown) analyses of a concatenation of slowly evolving proteins from 17 complete genomes globally agrees with the new animal phylogeny. The rapidly evolving Platyhelminthe *S. japonicum* is however placed as a sister clade of the Nematodes within the Ecdysozoa. The split between Annelids and Platyhelminthes is also clear from the comparative genomics results, since only 2% of *A. pompejana* proteins have a Blast first hit in Platyhelminthes and only 5 proteins in our dataset are specific to Annelids and Platyhelminthes. As we only have access to the proteome of a parasitic Platyhelminthe, this apparent split may actually be the consequence of an adaptation to a parasitic lifestyle. This illustrates the urgent need for complete genomes from free-living species of Platyhelminthes and more generally from diverse representatives of Lophotrochozoa in order to unravel the relations and synapomorphies unifying the Spiralia (Annelids and Molluscs) and Platyhelminthes within Lophotrochozoa. Taking into account the high proportion of *A. pompejana* "specific" genes (20%) found in this study, these genomes would be especially valuable for the discrimination of genes truly specific to Alvinellidae (and possibly linked to environmental adaptation) and those that are in fact shared by other lineages of Lophotrochozoa.

The slow evolutionary rate observed in *A. pompejana* leads to strong sequence conservation with other slow evolving species, in particular Deuterostomes and Cnidaria. The proximity between annelid and Vertebrate sequences has previously been reported for a set of 442 proteins from the polychaete *Platynereis dumerilii* (Raible et al. 2005) and is now observable at a larger scale. This proximity is also apparent in the gene repertoire of *A. pompejana,* with an important pool of *A. pompejana* genes specifically found in Deuterostomes or in Deuterostomes and Cnidaria compared to the restricted set of Protostome-specific genes. The former pool is enriched in typically "animal" functions (intercellular communication and adhesiveness, hormonal regulation, immunity). These

genes may have been present in the cenancestor of Bilaterian and Eumetazoan respectively and subsequently lost or diverged beyond recognition in the representative species of Ecdysozoa and Platyhelminthes used in our study. Indeed, with the multiplication of genome and EST sequencing projects in Invertebrates, many "vertebrate novelties" turn out to be present in Cnidaria and/or Placozoa, but lost in the canonical model Protostomes, i.e. *D. melanogaster* and *C. elegans* (Hughes and Friedman 2004; Miller et al. 2007; Putnam et al. 2007; Zmasek et al. 2007; Srivastava et al. 2008). There is now an increasing body of evidence for the prominent role of lineage-specific losses in animal evolution (for a review, see (De Robertis 2008)), especially in Protostomes. The present analysis suggests that massive losses are not a shared trait of the Protostomes, since genes involved in major metazoan functions are retained in *A. pompejana*, a Lophotrochozoan representative. However, we cannot exclude that losses exist in Annelids and/or Molluscs. For instance, no enzyme involved in urea excretion has been identified in the *Alvinella* database, as expected from previous studies reporting an incomplete or non-functional urea cycle in a number of annelid species (loss of the citrulline-arginine segment, see (Natesan et al. 1992)).

Genes are not the sole genome features differentially lost in the course of Metazoan evolution, as suggested by a study of genomic regions of the Polychaete *Platynereis dumerilii* that revealed intron-rich genes in Annelids (Raible et al. 2005). According to the authors' estimates, two-thirds of human introns would have been present in the bilaterian ancestor and retained in Annelids, while lost in the insect and Nematode genomes. The hypothesis of an intron-rich Bilaterian ancestor (discussed in (Roy 2006)) has been extended to the ancestor of Metazoa through the examination of the exon-intron structure of *Nematostella* and *Trichoplax* genes (Putnam et al. 2007; Srivastava et al. 2008). Thus, the emerging picture of evolution is one of a complex ancestor of Metazoa, with a gene toolkit and a gene structure closer to those of extant Vertebrates and Annelids than to model Ecdysozoa. This contradicts the intuitive view of a linear evolution, from simple ancestral networks to complex ones in Vertebrates, although this is in line with several studies suggesting a reductive evolution from a complex community of ancestors as a general trend in life evolution (see (Glansdorff et al. 2008) and references therein).

Methods

cDNA libraries and EST sequencing

A. pompejana samples were collected during the Biospeedo 2004 oceanographic cruise on the south East Pacific Rise at latitudes ranging from 14°S to 21°33S (25 individuals of which 6 were dissected on board). All individuals and/or tissues were conserved in liquid nitrogen. Specific tissues from adults were dissected and preserved in RNAlater stabilization and storage solution prior to being stored in liquid nitrogen. Four non-normalized and full-length-enriched cDNA libraries were constructed at the CNS Genoscope (http://www.genoscope.cns.fr/). One was prepared from 6 whole animals while the others were constructed from specific tissues: gills (5 individuals), ventral tissue (5 individuals) and pygidium (3 individuals). The whole animal and pygidium libraries were constructed with the CloneMiner cDNA construction kit (Invitrogen), which is designed to construct cDNA libraries without the use of traditional restriction enzyme cloning methods. This technology combines the action of SuperscriptII reverse Transcriptase with the Gateway Technology. Single-stranded mRNA was converted into double stranded cDNA containing attB sequences on each end. Through site-specific recombination, attB-flanked cDNA was cloned directly into attP-containing donor vector by homologue recombination. The gill and ventral tissues libraries were prepared using the oligo-capping approach. Full length RNAs were enriched by the action of the bacterial alkaline phosphatase to digest 5'-uncapped mRNAs. A 30-mer 5' oligo was linked using T4 RNA ligase after removing the 5'cap using Tobacco acid pyrophosphatase. The first strand cDNA was primed with an oligo(dT)-SfI primer and double stranded using specific 5' and 3' primers and amplified by PCR. The PCR SfI-digested cDNA products were size selected to exclude fragments smaller than 1 kb and then linked into pME18S-FL3 DraIII-digested vector. A total of 100,177 different clones were sequenced.

EST filtering, assembly and clustering

We developed a semi-automatic pipeline (TCL scripts) to manage and process the data, from the chromatograms to the assembled sequences. The raw SCF chromatogram files were clipped with respect to quality, repeats, and vector content. Quality clipping was based on phred (Ewing and Green 1998; Ewing et al. 1998) quality values: a window of size 20 bp was slid through the quality files from both sides, and the clip positions (left/right) were determined by the first window position with a phred-value above a threshold of 13. Vector masking was performed by cross_match (http://www.phrap.org/phrap_documentation.html) against the UniVec database (http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html) and the pME18S and pDONR222 vector

sequences. Masked sequences were cleaned from empty vector sequences and short sequences (<100nt) were filtered out. To avoid misassembly, 5' poly(A) and 3' poly(T) sequence boundaries were masked using a 20/25nt sliding window (in-house TCL script).

The EST assembly was performed using cap3 (Huang and Madan 1999) with default parameters, with the exception of 'overlap percent identity cutoff' (-p) and 'clipping range' (-y) parameters set to 90% and 30nt, respectively. The assembly was performed independently on each library as well as on the full complement of sequences. According to the stringent parameters used throughout the assembly process, the resulting sets of contigs and singletons can be considered as a set of unique transcripts: paralogs (as well as potential homeologs), divergent allelic forms and splicing variants are separated. However, some specific analyses require a non-redundant set of unique genes, with highly similar alleles and splicing variants grouped together. Thus, we further clustered sequences resulting from the global assembly by an all-against-all BLASTN comparison (Altschul et al. 1997). Sequences sharing >95% identity on a minimal overlap of 200nt were pooled into the same cluster.

cDNA characterization

Complete and partial CoDing Sequences (CDS) were determined from assembled sequences using two independent approaches, similarity and *ab initio* prediction. Contig and singleton sequences were compared to protein sequences of the UniprotKB (consortium 2008) and PDB (Berman et al. 2000) databases using BLASTX (Altschul et al. 1997). Coding frames were deduced from BLASTX best hit alignments (E-value ≤ $1e^{-05}$). Then, the CDS were created by extending the matching region in both 5' and 3' directions to the end of the cDNA sequence or a stop codon. If a stop codon was encountered in the 5'end, the first ATG codon following this stop codon was chosen as the initiation codon. When a frameshift was detected in the cDNA sequence, the translation of the incriminated region was replaced by masking symbols.

In parallel, we used ESTScan (Lottaz et al. 2003) to detect CDS. Since no large set of coding and noncoding sequences of annelids or molluscs are available for training, we used the *H. sapiens* model. In order to optimise a threshold for the ESTScan score, we established the distribution of sequences with or without homologs according to ESTScan cut-off values (Supplemental Figure S7). By setting an optimal cut-off value ≥ 200, we obtained a specificity of 70% and a sensitivity of 66%. To be considered as complete, a CDS must start with an initiation codon and end with a stop codon. Additionally, for CDS sequences deduced from BLASTX, the protein sequence must cover at least 80% of the best BLASTX hit.

The GC content study on mRNA encoding ribosomal proteins was performed for 84 almost complete cDNA (including 15 mitochondrial cDNA). The GC content was plotted against the number of repeats and subsequently tested with several regression models (linear, exponential, logarithmic and power) using the software SigmaPlot. The model that best fitted the dataset was a power function ($y=a^x+b$).

Integrative functional annotation

MACS (Lecompte et al. 2001) protein alignments were generated with the PipeAlign (Plewniak et al. 2003) toolkit. Integrative annotation was based on the MACSIMS (Thompson et al. 2006) and GOAnno (Chalmel et al. 2005) software frameworks. MACSIMS divides the multiple alignments into subfamilies according to conservation patterns. It then validates or corrects functional and structural information mined from public databases before propagation to the query sequence. Pfam-A (Finn et al. 2006) annotations are extracted from MACSIMS. GOAnno provides Gene Ontology (Ashburner et al. 2000) annotations for the query, after analysis of the GO terms obtained for the query subfamily. In addition to these programs, we have developed new software (Gagniere et al., manuscript in preparation) to: (1) generate a text mining functional definition from close homologs, (2) generate a consensus Enzyme Commission number from close homologs, (3) map the annotated proteins to the KEGG (Kyoto Encyclopedia of Genes and Genomes) pathways (Okuda et al. 2008) and the EMBL STRING (Search Tool for the Retrieval of Interacting Genes/Proteins) database (von Mering et al. 2007).

The mapping of *A. pompejana* proteins to the STRING database was performed by retrieving data for the closest human Uniprot homolog. This homolog was then used to search the STRING database using different identifiers (Uniprot ID, Uniprot Ensembl, RefSeq and Genome Reviews accession numbers and gene names in this order). If no STRING homolog was found using this textual search, a BLASTP search was performed on STRING human protein sequences and the first best hit (E-value ≤ $1e^{-05}$) was chosen. Then, STRING networks were built (combined score cut-off ≥ 0.9) and sub-networks were extracted by retrieving level 1 neighbours for each protein. These small sub-networks were scored by the Ratio of consistency (Rc), defined as the ratio between the observed number of sub-network edges (protein-protein interactions) and the maximum theoretical number of edges. Rc will be high in sub-networks exhibiting a high level of intra sub-network interactions. In contrast, a low ratio indicates large sub-networks with few intra sub-network interactions. In order to reduce the set of sub-networks, two sub-networks *A* and *B* were fused if they matched the following criteria:

$$1 - (1 - Nc) \times (1 - Rc_A) \times (1 - Rc_B) \geq 0.7$$

Nc≥0.5

with *Nc* the ratio between the number of *A* and *B* shared nodes and the number of nodes in the smaller sub-network. These criteria help to preferentially fuse highly related sub-networks, while avoiding low consistency sub-networks that would otherwise agglomerate weakly related sub-networks. The final sub-networks were visualized using Cytoscape (Cline et al. 2007).

Functional enrichment

The functional annotation clustering tool in the DAVID (Dennis et al. 2003) software was used to study the pools of tissue-specific transcripts and the sets of differentially conserved genes. For each set, the closest homologs of the *A. pompejana* proteins from a given species (depending on the set under study) were processed against the background of this species. Enrichment with a P-value ≤ 0.01 was considered to be significant.

Phylogeny and molecular evolution

Phylogenetic reconstruction was performed on 17 model taxa covering the main eukaryotic lineages (choanoflagellate: 1, cnidaria: 1, platyhelminthe: 1, nematod: 2, arthropod: 2, lophotrochozoa: 3 including *A. pompejana*, chordate: 6). The tree was rooted with the yeast *Saccharomyces cerevisiae*. The phylogenetic tree and rates of amino acid substitution for each branch were inferred on a concatenated alignment of 76 ribosomal protein families using MrBayes 3 (Ronquist and Huelsenbeck 2003) under the WAG model.

Both observed and simulated amino-acid frequencies associated with the orthologous set of protein coding regions (65167 amino acids) were obtained using the codeML package of the software PaML v3.14 (Yang 1997) and the 'universal' genetic code. Amino-acid alignments were validated manually, concatenated and exported in a PHYLIP format using the software Se-AL v2.0 (http://evolve.zoo.ox.ac.uk/software/Se-Al/). Regions containing gaps, misalignments or uncertainties were excluded from the analysis. PAML analyses were performed using a reference tree previously obtained from the ProML package of PHYLIP 3.68 (Felsenstein 1989) for the 9 taxa, using the JTT model of amino-acid substitutions. Amino acid frequencies were calculated using the aaML package (aadist = 'equal', with the jones.dat matrix) and standard deviations of frequencies were obtained from 100 rearrangements (bootstrap) of the dataset. This allowed us to estimate the proportion of hydrophobic, positively-charged and negatively-charged amino acids associated with the translated sequences across taxa, and to calculate hydrophobicities using the hydrophobic index based on the OMH scale of Sweet & Eisenberg (Sweet and Eisenberg 1983). This index is known to

take into account the ability of an amino acid to be replaced by another during the course of evolution.

Comparative genomics

A massive best hit taxonomy assessment of the *A. pompejana* proteins was performed using BlastP (E-value threshold ≤ $1e^{-05)}$). Sequences were compared to the Uniprot database and classified according to the taxonomy of their best hit: Deuterostomia, Nematoda, Arthropoda, Platyhelminthes, Cnidaria, other Metazoa, Choanoflagellida, "protists" (Alveolata, Diplomonadida, Cryptophyta, Entamoebidae, Euglenozoa, Mycetozoa, Parabasalidea, Rhodophyta and Stramenopiles), Viridiplantae, Fungi, other Eukaryota, Prokaryotes and Viruses. Each group includes at least one species with a complete proteome represented in Uniprot. For proteins with a first hit in Annelids or Mollusca, subsequent hits were considered, since no complete proteome is available in these taxa. When all BlastP hits (E-value ≤ $1e^{-05)}$) of an *A. pompejana* protein were restricted to a unique taxon (with the exception of Annelids and Molluscs), this protein was considered to be specific to this taxon and Annelids (and potentially Molluscs).

Alvinella database and website

The database is managed by the PostgreSQL (http://www.postgresql.org/) relational database management system and is backed-up on a weekly basis. The *Alvinella* website uses an Apache HTTP server (http://httpd.apache.org/) and PHP5 (http://www.php.net/) and was built from scratch using the Smarty (http://www.smarty.net/) template engine, the ADOdb (http://adodb.sourceforge.net/) database abstraction library and the phpGACL (http://phpgacl.sourceforge.net/) generic access control list library.

Acknowledgments

We wish to acknowledge Julie Thompson for a critical reading of the manuscript. We would also like to thank the Bioinformatics Platform of Strasbourg and the Structural Biology and Genomics Platform for their assistance during this work. We thank the crew and pilots of the RV L'Atalante and the DSV Nautile for their assistance and technical support during the cruise BioSpeedo'04 for collecting animals and Florence Pradillon for sending preliminary samples helping in the present studies. This work was supported by institutional funds from INSERM, CNRS and UDS, by European Commission

funding through the SPINE2–COMPLEXES project LSHG–CT–2006–031220 and by ANR-05-BLAN-0407 grant.

Figure legends

Figure 1

The 30 most frequent PFAM-A domains identified in *A. pompejana* protein sequences. For each domain, the bar height indicates the total number of reads encoding proteins with this domain. We used a logarithmic scale for representation constraints. Bars are coloured according to functional role categories: respiration, oxidative stress and/or detoxification in blue, protein-protein or protein-DNA/RNA interactions in green, structural proteins in red, anti-bacterial defence in orange and others in grey.

Figure 2

Bayesian phylogeny of Metazoa. The analysis was performed on a concatenation of 76 ribosomal protein family alignments, with *Saccharomyces cerevisiae* as the outgroup. The scale bar indicates the expected number of amino acid substitutions per aligned position. All nodes were resolved in 100% of the sampled topologies from the Bayesian analysis, except the node indicated by a square (support value of 96%).

Figure 3

Amino acid composition across model taxa. The y-axis indicates the proportion of charged amino acids, the x-axis represents the hydrophobicity index.

Figure 4

Taxonomic distribution of protein sequence BLAST best hits. The eukaryotic taxa accounting for less than 2% of *A. pompejana* best hits were fused into the "Other eukaryota" category for clarity: Choanoflagellida (0.6%), Protists (1.9%), Fungi (1.5%), Viridiplantae (1.3%).

Figure 5

Screenshots of the *Alvinella* website illustrating some of the visualisation tools.

Figures

Figure 1

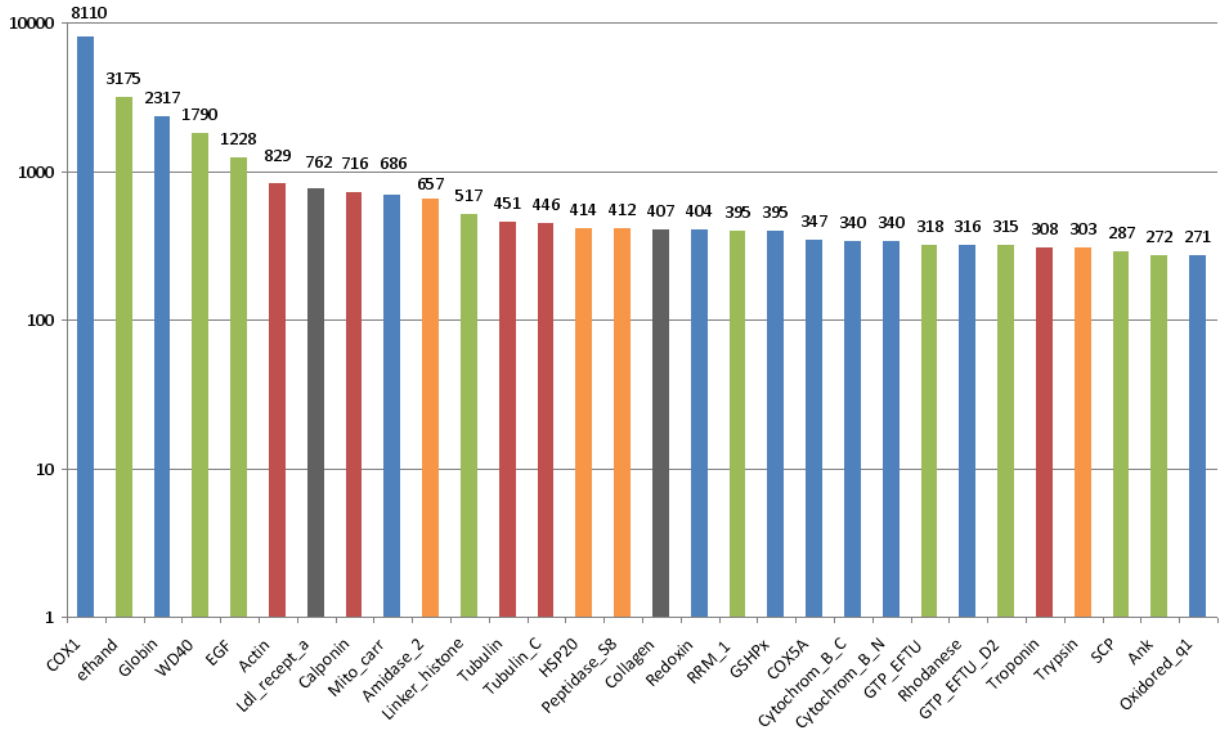


Figure 2

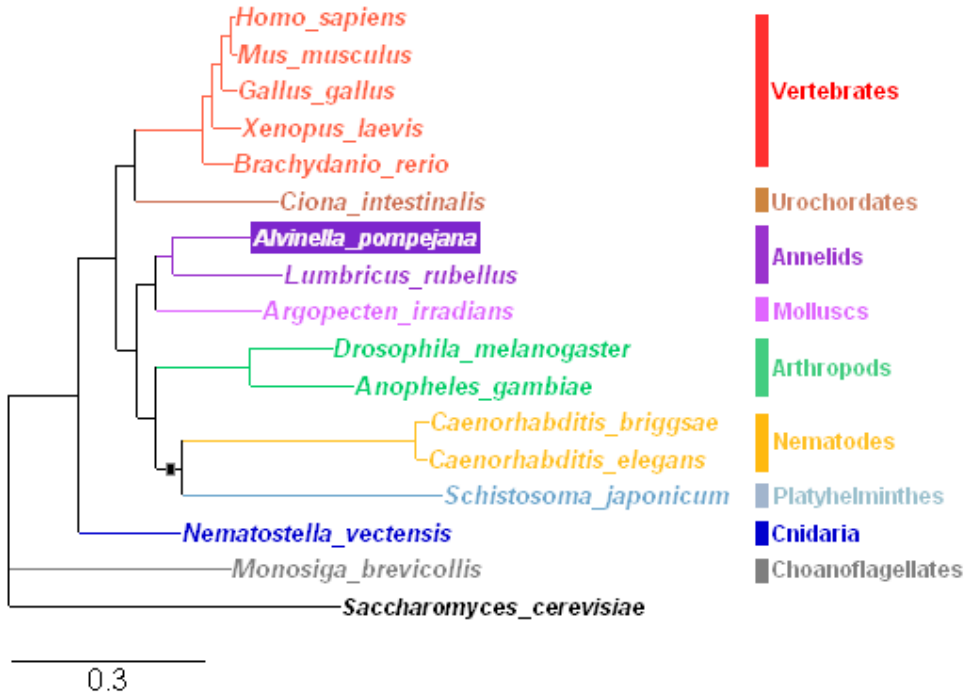


Figure 3

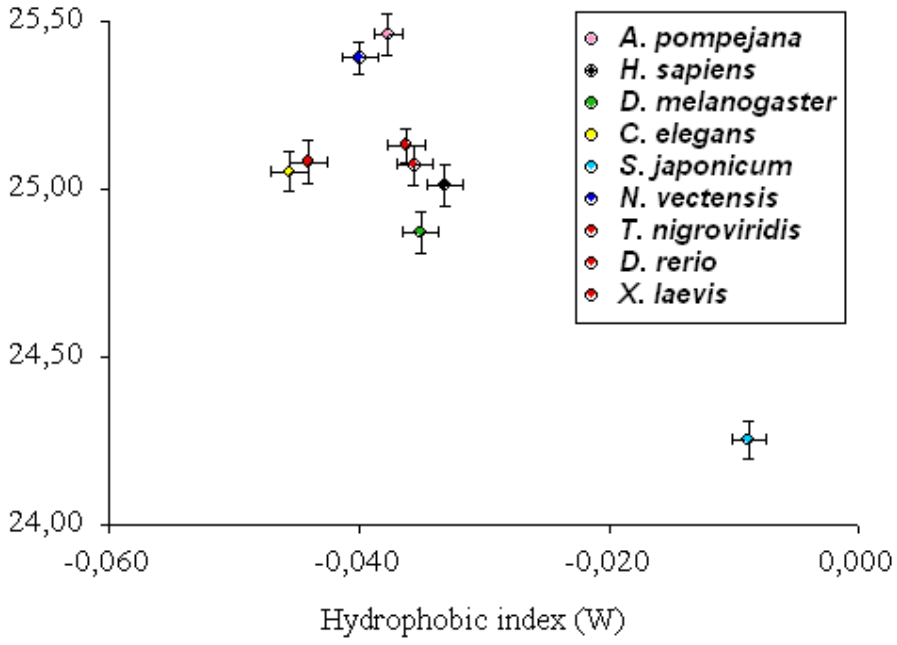


Figure 4

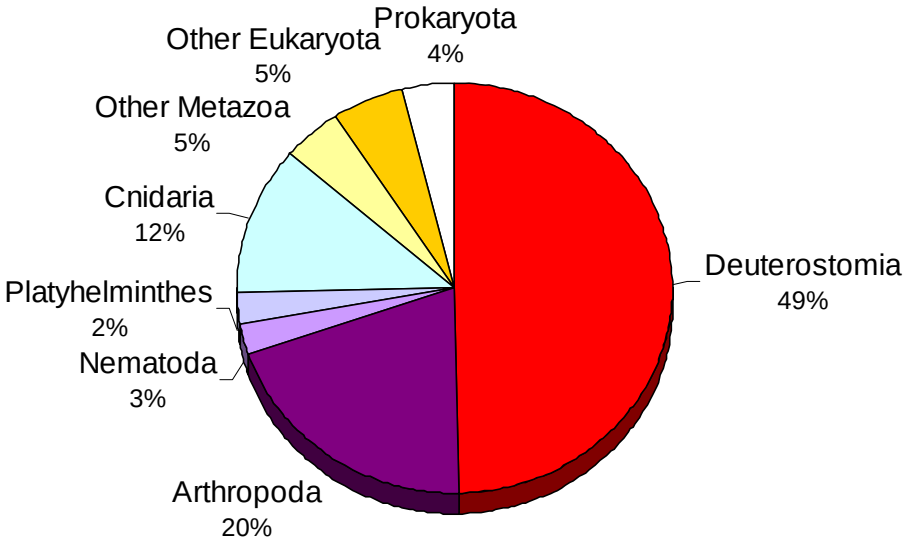
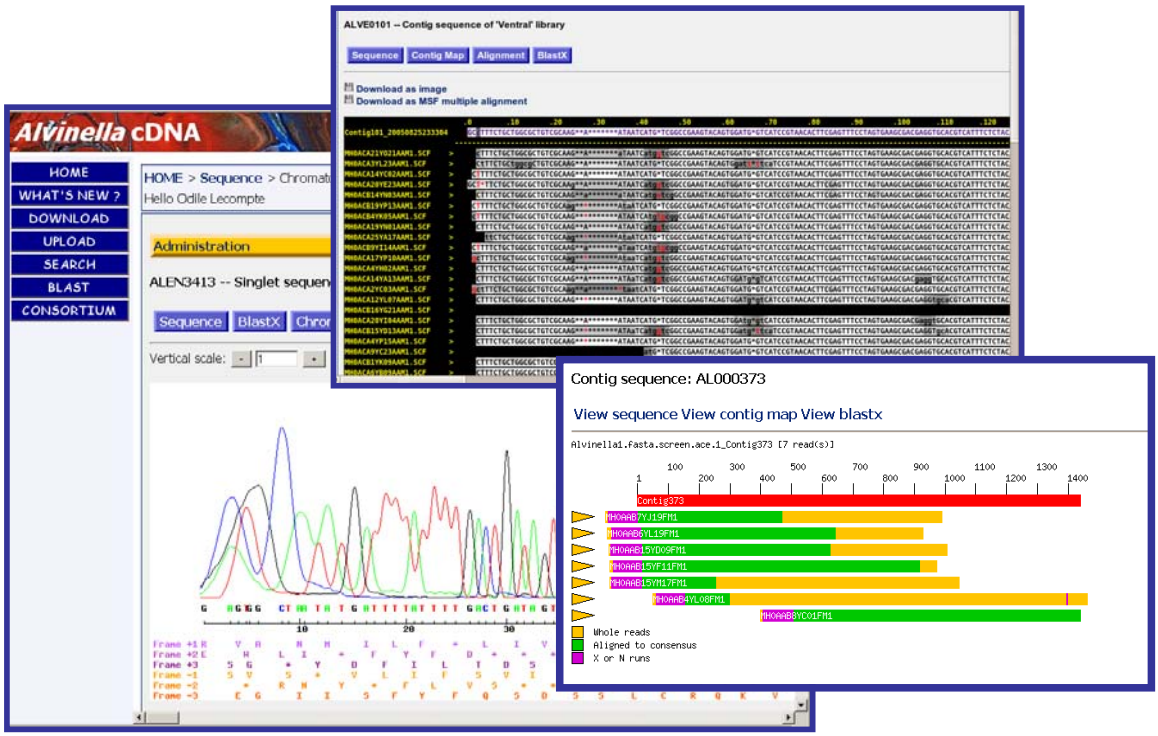


Figure 5



Tables

Table 1 Summary of *A. pompejana* cDNA libraries and assemblies

Libraries	CloneMiner		Oligo-capping		Global assembly
	Whole animal	Pygidium	Ventral tissue	Gills	
Initial chromatograms	20,549	36,648	16,411	26,569	100,177
Clean sequences	19,739 (96%)	25,419 (69%)	12,871 (78%)	18,105 (68%)	76,134 (76%)
Number of 3' poly(A) (%)	1,599 (8%)	3,156 (12%)	5,465 (43%)	1,467 (8%)	11,687 (15%)
Mean length (bp)	633	610	720	776	674
Assembly					
Contigs	1,365	2,327	917	1,193	4,993
Singletons	4,060	6,355	1,914	2,567	10,865
Contig mean length (bp)	993	951	852	931	1,017
Redundancy (%)	73	66	78	79	79

Table 2 Overview of the annotation results.

Annotation	Number of items
Level 1	**Proteins**
Proteins with homologs	7,353 (100%)
Pfam-A domains	4,767 (65%)
Gene Ontology	5,949 (81%)
Biological process	5,072 (69%)
Cellular component	4,530 (62%)
Molecular function	5,601 (76%)
Text mining definition	4,611 (63%)
Enzyme Classification	1,243 (17%)
Level 4 (X.X.X.X)	1,180 (16%)
Annotated proteins	6,252 (85%)
Level 2	**Networks**
KEGG pathways	345
Mapped	202
Coverage >50%	82
STRING subnetworks	385
Mapped	264
Coverage >50%	63

Table 3 Highly expressed genes in *A. pompejana* libraries

Access	Fonction	Reads*
TERA04282	Cytochrome c oxidase subunit 1 (EC 1.9.3.1)	7845
TERA02741	Hypothetical protein	2029
TERA02189	Hypothetical protein	917
TERA02142	Hypothetical protein	879
TERA03177	Actin	533
TERA00344	Extracellular globin (Haemoglobin A2 chain precursor)	524
TERA02067	Hypothetical protein	424
TERA00650	Hypothetical protein	422
TERA03305	Intracellular haemoglobin	386
TERA00833	Extracellular haemoglobin (Haemoglobin B2 chain precursor)	370
TERA00205	Extracellular haemoglobin linker L1	349
TERA03100	Cytochrome c oxidase subunit 5A (EC 1.9.3.1)	344
TERA00354	Cytochrome b	338
TERA04769	Hypothetical protein	335
TERA00845	Extracellular haemoglobin (Haemoglobin B1 chain precursor)	322
TERA02090	Hypothetical protein	304
TERA01907	Heat shock protein	295
TERA02261	Myosin essential light chain	285
TERA01929	Hypothetical protein	275
TERA00421	Extracellular haemoglobin linker L3	267
TERA03231	Glutathione peroxidase	264
TERA01189	Hypothetical protein	263
TERA02903	Hypothetical protein	258
TERA01828	Tropomyosin	232
TERA00984	Hypothetical protein	230
TERA02160	Elongation factor 1-alpha	218
TERA01465	Peptidoglycan recognition protein	200

*The last column indicates the number of reads in the global assembly.

Table 4 Percent identities between orthologous protein sequences of Metazoa.

	Ap	Hs	Dm	Ce	Sj	Nv
Alvinella pompejana (Ap)	100,0					
Homo sapiens (Hs)	65,9	100,0				
Drosophila melanogaster (Dm)	62,7	61,7	100,0			
Caenorhabditis elegans (Ce)	56,3	55,3	55,1	100,0		
Schistosoma japonicum (Sj)	57,7	56,0	54,8	51,0	100,0	
Nematostella vectensis (Nv)	65,1	64,5	60,8	54,9	55,3	100,0

References

Aguinaldo, A.M., Turbeville, J.M., Linford, L.S., Rivera, M.C., Garey, J.R., Raff, R.A., and Lake, J.A. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* **387:** 489-493.

Ahearn, G.A., Mandal, P.K., and Mandal, A. 2004. Mechanisms of heavy-metal sequestration and detoxification in crustaceans: a review. *J Comp Physiol B* **174:** 439-452.

Alcivar, A., Hu, S., Tang, J., and Yang, X. 2003. DEDD and DEDD2 associate with caspase-8/10 and signal cell death. *Oncogene* **22:** 291-297.

Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D.J. 1997. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res* **25:** 3389-3402.

Ao, J.Q., Ling, E., Rao, X.J., and Yu, X.Q. 2008. A novel ML protein from Manduca sexta may function as a key accessory protein for lipopolysaccharide signaling. *Mol Immunol* **45:** 2772-2781.

Arendt, D. and Nubler-Jung, K. 1994. Inversion of dorsoventral axis? *Nature* **371:** 26.

Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T. et al. 2000. Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. *Nat Genet* **25:** 25-29.

Bechtel, J.M., Wittenschlaeger, T., Dwyer, T., Song, J., Arunachalam, S., Ramakrishnan, S.K., Shepard, S., and Fedorov, A. 2008. Genomic mid-range inhomogeneity correlates with an abundance of RNA secondary structures. *BMC Genomics* **9:** 284.

Berezovsky, I.N. and Shakhnovich, E.I. 2005. Physics and evolution of thermophilic adaptation. *Proc Natl Acad Sci U S A* **102:** 12742-12747.

Berman, H.M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T.N., Weissig, H., Shindyalov, I.N., and Bourne, P.E. 2000. The Protein Data Bank. *Nucleic Acids Res* **28:** 235-242.

Bingle, C.D. and Vyakarnam, A. 2008. Novel innate immune functions of the whey acidic protein family. *Trends Immunol* **29:** 444-453.

Bonnivard, E., Catrice, O., Ravaux, J., Brown, S.C., and Higuet, D. 2009. Survey of genome size in 28 hydrothermal vent species covering 10 families. *Genome* **52:** 524-536.

Burjanadze, T.V. 2000. New analysis of the phylogenetic change of collagen thermostability. *Biopolymers* **53:** 523-528.

Cary, S.C., Shank, T., and Stein, J. 1998. Worms bask in extreme temperatures. *Nature* **391:** 545-546.

Chalmel, F., Lardenois, A., Thompson, J.D., Muller, J., Sahel, J.A., Leveillard, T., and Poch, O. 2005. GOAnno: GO annotation based on multiple alignment. *Bioinformatics* **21:** 2095-2096.

Chevaldonné, P., Desbruyeres, D., and Childress, J.J. 1992. Some like it hot... and some even hotter. *Nature* **359:** 593-594.

Ciccarelli, F.D., Doerks, T., von Mering, C., Creevey, C.J., Snel, B., and Bork, P. 2006. Toward automatic reconstruction of a highly resolved tree of life. *Science* **311:** 1283-1287.

Cipollone, R., Ascenzi, P., and Visca, P. 2007. Common themes and variations in the rhodanese superfamily. *IUBMB Life* **59:** 51-59.

Cline, M.S., Smoot, M., Cerami, E., Kuchinsky, A., Landys, N., Workman, C., Christmas, R., Avila-Campilo, I., Creech, M., Gross, B. et al. 2007. Integration of biological networks and gene expression data using Cytoscape. *Nat Protoc* **2:** 2366-2382.

consortium, U. 2008. The universal protein resource (UniProt). *Nucleic Acids Res* **36:** D190-195.

Dahlhoff, E. and Somero, G.N. 1991. Pressure and temperature adaptation of cytosolic malate dehydrogenases of shallowand deep-living marine invertebrates: evidence for high body temperatures in hydrothermal vent animals *Journal of Experimental Biology* **159:** 473-487.

De Robertis, E.M. 2008. Evo-devo: variations on ancestral themes. *Cell* **132:** 185-195.

Deak, F., Wagener, R., Kiss, I., and Paulsson, M. 1999. The matrilins: a novel family of oligomeric extracellular matrix proteins. *Matrix Biol* **18:** 55-64.

Dennis, G., Jr., Sherman, B.T., Hosack, D.A., Yang, J., Gao, W., Lane, H.C., and Lempicki, R.A. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4:** P3.

Desbruyeres, D., Chevaldonné, P., Alayse, A.M., Jollivet, D., Lallier, F.H., Jouin-Toulmond, C., Zal, F., Sarradin, P.M., Cosson, R., Caprais, J.C. et al. 1998. Biology and ecology of the Pompeii worm (Alvinella pompejana Desbruyères and Laubier), a normal dweller of an extreme deep-sea environment : A synthesis of current knowledge and recent developments. *Deep-sea research* **45:** 383-422.

Desbruyeres, D. and Laubier, L. 1980. Alvinella pompejana gen. sp. now., aberrant Ampharetidae from East Pacific Rise hydrothermal vents. . *Oceanol. Acta* **3:** 267-274.

Dixon, D., Dixon, L., Shillito, B., and Gwynn, J. 2002. Background and induced levels of DNA damage in PaciWc deep-sea vents polychaetes: the case for avoidance. *Cahier de Biologie Marine* **43:** 333-336.

Dixon, D.R., Jolly, M.T., Vevers, W.F., and Dixon, L.R.J. 2009. Chromosomes of Pacific hydrothermal vent invertebrates: towards a greater understanding of the relationship between chromosome and molecular evolution. *Journal of the Marine Biological Association of the United Kingdom*.

Ewing, B. and Green, P. 1998. Base-calling of automated sequencer traces using phred. II. Error probabilities. *Genome Res* **8:** 186-194.

Ewing, B., Hillier, L., Wendl, M.C., and Green, P. 1998. Base-calling of automated sequencer traces using phred. I. Accuracy assessment. *Genome Res* **8:** 175-185.

Felsenstein, J. 1989. PHYLIP - Phylogeny Inference Package (Version 3.2). *Cladistics* **5:** 164-166.

Fields, A.P., Frederick, L.A., and Regala, R.P. 2007. Targeting the oncogenic protein kinase Ciota signalling pathway for the treatment of cancer. *Biochem Soc Trans* **35:** 996-1000.

Finn, R.D., Mistry, J., Schuster-Bockler, B., Griffiths-Jones, S., Hollich, V., Lassmann, T., Moxon, S., Marshall, M., Khanna, A., Durbin, R. et al. 2006. Pfam: clans, web tools and services. *Nucleic Acids Res* **34:** D247-251.

Finn, R.D., Tate, J., Mistry, J., Coggill, P.C., Sammut, S.J., Hotz, H.R., Ceric, G., Forslund, K., Eddy, S.R., Sonnhammer, E.L. et al. 2008. The Pfam protein families database. *Nucleic Acids Res* **36:** D281-288.

Gaill, F., Halpern, S., Quintana, C., and Desbruyeres, D. 1984. Presence intracellulaire d'arsenic et de zinc assocides au soufre chez une Polychete des sources hydrothermales. *C R Acad Sci III* **298:** 331-335.

Gaill, F. and Hunt, S. 1991. The biology of annelid worms from high temperature hydrothermal vent regions. *Rev Aquat Sci* **4**: 107-137.

Girguis, P.R. and Lee, R.W. 2006. Thermal preference and tolerance of alvinellids. *Science* **312:** 231.

Glansdorff, N., Xu, Y., and Labedan, B. 2008. The last universal common ancestor: emergence, constitution and genetic legacy of an elusive forerunner. *Biol Direct* **3:** 29.

Gordon, D. 2003. Viewing and editing assembled sequences using Consed. *Curr Protoc Bioinformatics* **Chapter 11:** Unit11 12.

Grzymski, J.J., Murray, A.E., Campbell, B.J., Kaplarevic, M., Gao, G.R., Lee, C., Daniel, R., Ghadiri, A., Feldman, R.A., and Cary, S.C. 2008. Metagenome analysis of an extreme microbial symbiosis reveals eurythermal adaptation and metabolic flexibility. *Proc Natl Acad Sci U S A* **105:** 17516-17521.

Haney, P.J., Stees, M., and Konisky, J. 1999. Analysis of thermal stabilizing interactions in mesophilic and thermophilic adenylate kinases from the genus Methanococcus. *J Biol Chem* **274:** 28453-28458.

Hauser, F., Gertzen, E.M., and Hoffmann, W. 1990. Expression of spasmolysin (FIM-A.1): an integumentary mucin from Xenopus laevis. *Exp Cell Res* **189:** 157-162.

Henscheid, K.L., Shin, D.S., Cary, S.C., and Berglund, J.A. 2005. The splicing factor U2AF65 is functionally conserved in the thermotolerant deep-sea worm Alvinella pompejana. *Biochim Biophys Acta* **1727:** 197-207.

Hourdez, S., Lallier, F.H., De Cian, M.C., Green, B.N., Weber, R.E., and Toulmond, A. 2000. Gas transfer system in Alvinella pompejana (Annelida polychaeta, Terebellida): functional properties of intracellular and extracellular hemoglobins. *Physiol Biochem Zool* **73:** 365-373.

Hourdez, S. and Weber, R.E. 2005. Molecular and functional adaptations in deep-sea hemoglobins. *J Inorg Biochem* **99:** 130-141.

Huang, X. and Madan, A. 1999. CAP3: A DNA sequence assembly program. *Genome Res* **9:** 868-877.

Hughes, A.L. and Friedman, R. 2004. Differential loss of ancestral gene families as a source of genomic divergence in animals. *Proc Biol Sci* **271 Suppl 3:** S107-109.

Ichikawa, S., Hatanaka, H., Yuuki, T., Iwamoto, N., Kojima, S., Nishiyama, C., Ogura, K., Okumura, Y., and Inagaki, F. 1998. Solution structure of Der f 2, the major mite allergen for atopic diseases. *J Biol Chem* **273:** 356-360.

Irvine, S.M. and Martindale, M.Q. 1996. Cellular and molecular mechanisms of segmentation in annelids. *Seminars in Cell & Developmental Biology* **7:** 593-604.

Jollivet, D., Desbruyeres, D., Ladrat, C., and Laubier, L. 1995. Evidence for differences in the allozyme thermostability of deep-sea hydrothermal vent polychaetes (Alvinellidae): a possible selection by habitat. *Marine Ecology Progress Series* **123:** 125-136.

Le Bris, N. and Gaill, F. 2007. How does the annelid Alvinella pompejana deal with an extreme hydrothermal environment? *ReViews in Environmental Science and BioTechnology* **6:** 102-119.

Le Bris, N., Zbinden, M., and Gaill, F. 2005. Processes controlling the physico-chemical micro-environments associated with Pompeii worms. *Deep-Sea Research Part I-Oceanographic Research Papers* **52:** 1071-1083.

Lecompte, O., Thompson, J.D., Plewniak, F., Thierry, J., and Poch, O. 2001. Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene* **270:** 17-30.

Lescure, A., Rederstorff, M., Krol, A., Guicheney, P., and Allamand, V. 2009. Selenoprotein function and muscle disease. *Biochim Biophys Acta*.

Lottaz, C., Iseli, C., Jongeneel, C.V., and Bucher, P. 2003. Modeling sequencing errors by combining Hidden Markov models. *Bioinformatics* **19 Suppl 2:** ii103-112.

Mann, H.H., Sengle, G., Gebauer, J.M., Eble, J.A., Paulsson, M., and Wagener, R. 2007. Matrilins mediate weak cell attachment without promoting focal adhesion formation. *Matrix Biol* **26:** 167-174.

Marie, B., Genard, B., Rees, J., and Zal, F. 2006. Effect of ambient oxygen concentration on activities of enzymatic antioxidant defences and aerobic metabolism in the hydrothermal vent worm, Paralvinella grasslei. *Marine Biology* **150:** 273-284.

McDougall, C., Hui, J.H., Monteiro, A., Takahashi, T., and Ferrier, D.E. 2008. Annelids in evolutionary developmental biology and comparative genomics. *Parasite* **15:** 321-328.

Mignone, F., Grillo, G., Licciulli, F., Iacono, M., Liuni, S., Kersey, P.J., Duarte, J., Saccone, C., and Pesole, G. 2005. UTRdb and UTRsite: a collection of sequences and regulatory motifs of the untranslated regions of eukaryotic mRNAs. *Nucleic Acids Res* **33:** D141-146.

Miller, D.J., Hemmrich, G., Ball, E.E., Hayward, D.C., Khalturin, K., Funayama, N., Agata, K., and Bosch, T.C. 2007. The innate immune repertoire in cnidaria--ancestral complexity and stochastic gene loss. *Genome Biol* **8:** R59.

Morris, S.C. 1998. *The Crucible of Creation : The Burgess Shale and the Rise of Animals*. Oxford University Press.

Naik, M.U. and Naik, U.P. 2003. Calcium-and integrin-binding protein regulates focal adhesion kinase activity during platelet spreading on immobilized fibrinogen. *Blood* **102:** 3629-3636.

Natesan, S., Jayasundaramma, B., Ramamurthi, R., and Reddy, S.R. 1992. Presence of a partial urea cycle in the leech, Poecilobdella granulosa. *Experientia* **48:** 729-731.

Nishio, Y., Nakamura, Y., Kawarabayasi, Y., Usuda, Y., Kimura, E., Sugimoto, S., Matsui, K., Yamagishi, A., Kikuchi, H., Ikeo, K. et al. 2003. Comparative complete genome sequence analysis of the amino acid replacements responsible for the thermostability of Corynebacterium efficiens. *Genome Res* **13:** 1572-1579.

Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. 2008. KEGG Atlas mapping for global analysis of metabolic pathways. *Nucleic Acids Res* **36:** W423-426.

Pesole, G., Grillo, G., Larizza, A., and Liuni, S. 2000. The untranslated regions of eukaryotic mRNAs: structure, function, evolution and bioinformatic tools for their analysis. *Brief Bioinform* **1:** 236-249.

Piccino, P., Viard, F., Sarradin, P.M., Le Bris, N., Le Guen, D., and Jollivet, D. 2004. Thermal selection of PGM allozymes in newly founded populations of the thermotolerant vent polychaete Alvinella pompejana. *Proc Biol Sci* **271:** 2351-2359.

Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J. et al. 2003. PipeAlign: A new toolkit for protein family analysis. *Nucleic Acids Res* **31:** 3829-3832.

Pradillon, F., Zbinden, M., Mullineaux, L.S., and Gaill, F. 2005. Colonisation of newly-opened habitat by a pioneer species, Alvinella pompejana (Polychaeta : Alvinellidae), at East Pacific Rise vent sites. *Marine Ecology-Progress Series* **302:** 147-157.

Putnam, N.H., Srivastava, M., Hellsten, U., Dirks, B., Chapman, J., Salamov, A., Terry, A., Shapiro, H., Lindquist, E., Kapitonov, V.V. et al. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science* **317:** 86-94.

Raible, F., Tessmar-Raible, K., Osoegawa, K., Wincker, P., Jubin, C., Balavoine, G., Ferrier, D., Benes, V., de Jong, P., Weissenbach, J. et al. 2005. Vertebrate-type intron-rich genes in the marine annelid Platynereis dumerilii. *Science* **310:** 1325-1326.

Robinson-Rechavi, M., Alibes, A., and Godzik, A. 2006. Contribution of electrostatic interactions, compactness and quaternary structure to protein thermostability: lessons from structural genomics of Thermotoga maritima. *J Mol Biol* **356:** 547-557.

Ronquist, F. and Huelsenbeck, J.P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics* **19:** 1572-1574.

Roy, S.W. 2006. Intron-rich ancestors. *Trends Genet* **22:** 468-471.

Shin, D.S., Didonato, M., Barondeau, D.P., Hura, G.L., Hitomi, C., Berglund, J.A., Getzoff, E.D., Cary, S.C., and Tainer, J.A. 2009. Superoxide dismutase from the eukaryotic thermophile Alvinella pompejana: structures, stability, mechanism, and insights into amyotrophic lateral sclerosis. *J Mol Biol* **385:** 1534-1555.

Sicot, F.X., Mesnage, M., Masselot, M., Exposito, J.Y., Garrone, R., Deutsch, J., and Gaill, F. 2000. Molecular adaptation to an extreme environment: origin of the thermal stability of the pompeii worm collagen. *J Mol Biol* **302:** 811-820.

Srivastava, M., Begovic, E., Chapman, J., Putnam, N.H., Hellsten, U., Kawashima, T., Kuo, A., Mitros, T., Salamov, A., Carpenter, M.L. et al. 2008. The Trichoplax genome and the nature of placozoans. *Nature* **454:** 955-960.

Sweet, R.M. and Eisenberg, D. 1983. Correlation of sequence hydrophobicities measures similarity in three-dimensional protein structure. *J Mol Biol* **171:** 479-488.

Szilagyi, A. and Zavodszky, P. 2000. Structural differences between mesophilic, moderately thermophilic and extremely thermophilic protein subunits: results of a comprehensive survey. *Structure* **8:** 493-504.

Thompson, J.D., Muller, A., Waterhouse, A., Procter, J., Barton, G.J., Plewniak, F., and Poch, O. 2006. MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics* **7:** 318.

Tinel, A. and Tschopp, J. 2004. The PIDDosome, a protein complex implicated in activation of caspase-2 in response to genotoxic stress. *Science* **304:** 843-846.

Vilmos, P., Gaudenz, K., Hegedus, Z., and Marsh, J.L. 2001. The Twisted gastrulation family of proteins, together with the IGFBP and CCN families, comprise the TIC superfamily of cysteine rich secreted factors. *Mol Pathol* **54:** 317-323.

Vogt, G., Woell, S., and Argos, P. 1997. Protein thermal stability, hydrogen bonds, and ion pairs. *J Mol Biol* **269:** 631-643.

von Mering, C., Jensen, L.J., Kuhn, M., Chaffron, S., Doerks, T., Kruger, B., Snel, B., and Bork, P. 2007. STRING 7--recent developments in the integration and prediction of protein interactions. *Nucleic Acids Res* **35:** D358-362.

Vovelle, J. and Gaill, F. 1986. Données morphologiques, histochimiques et microanalytiques sur l'élaboration du tube organominéral d'Alvinella pompejana, Polychète des sources hydrothermales, et leurs implications phylogénétiques. *Zool. Scripta.* **15:** 33-43.

Wolf, Y.I., Rogozin, I.B., and Koonin, E.V. 2004. Coelomata and not Ecdysozoa: evidence from genome-wide phylogenetic analysis. *Genome Res* **14:** 29-36.

Yang, Z. 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. *Comput Appl Biosci* **13:** 555-556.

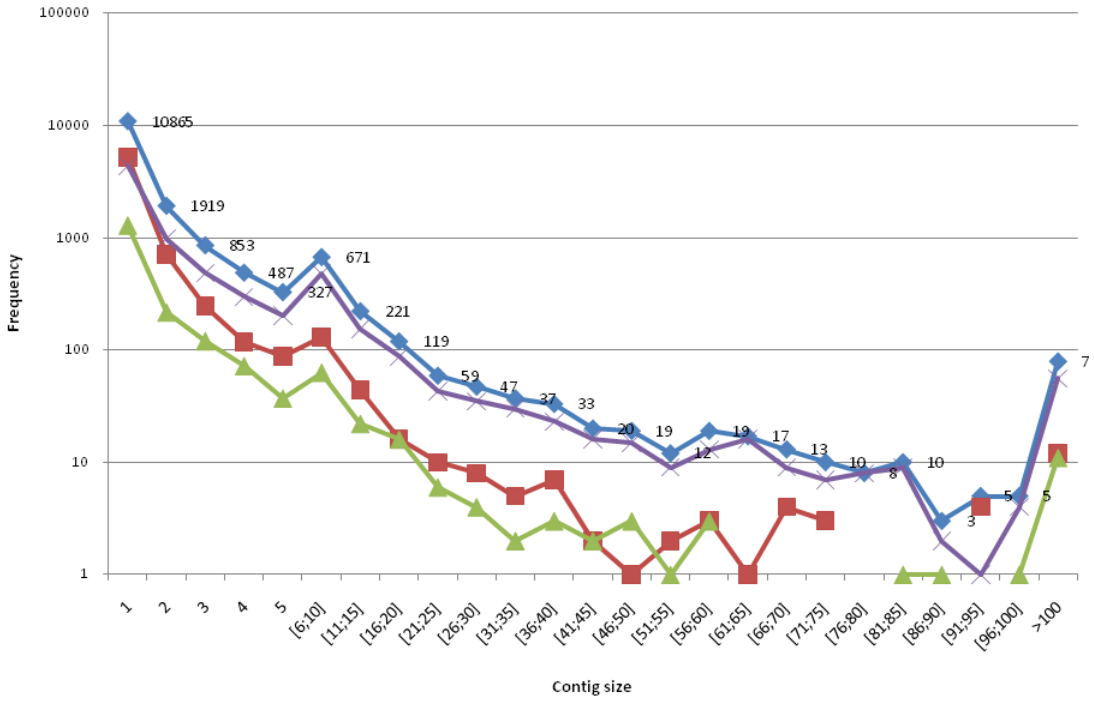
Zhang, L., Kasif, S., Cantor, C.R., and Broude, N.E. 2004. GC/AT-content spikes as genomic punctuation marks. *Proc Natl Acad Sci U S A* **101:** 16855-16860.

Zmasek, C.M., Zhang, Q., Ye, Y., and Godzik, A. 2007. Surprising complexity of the ancestral apoptosis network. *Genome Biol* **8:** R226.

Supplementary data

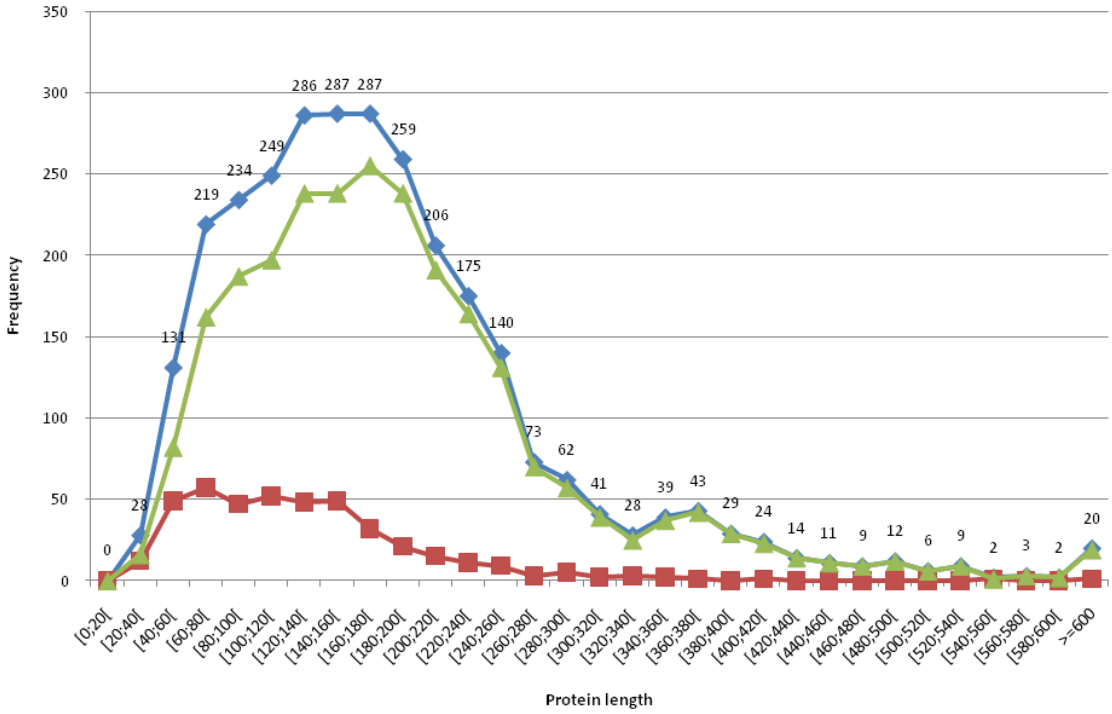
Supplemental Fig. S1.

Size distribution of the 4,993 contigs and 10,865 singletons resulting from the global assembly of 76,134 clones. Frequencies of singletons and contigs are indicated in blue for the full set of sequences, in red for sequences without CDS, in green for CDS with no detected similarity and in purple for CDS with homologs. We used a logarithmic scale for representation convenience.



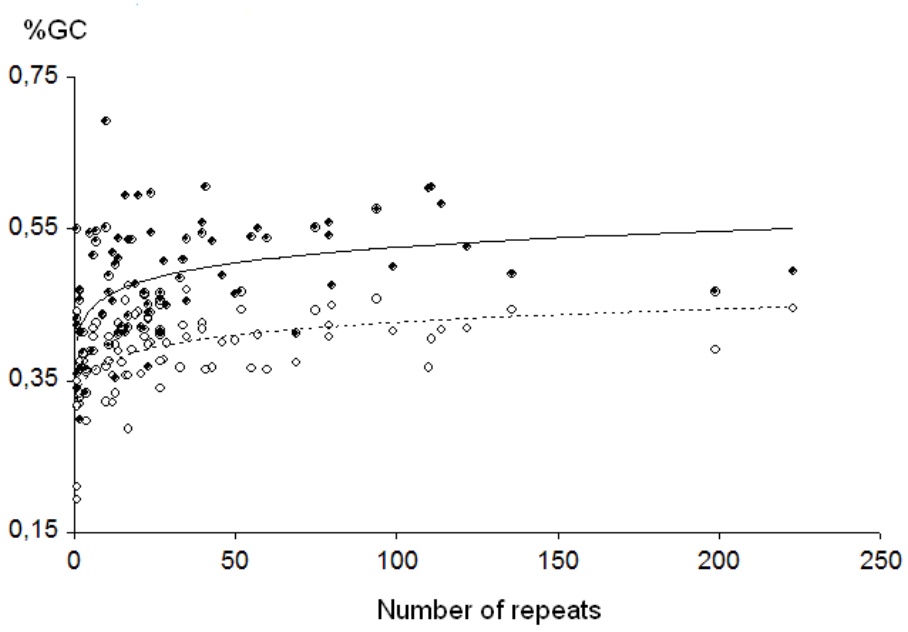
Supplemental Fig. S2.

Distribution of complete protein lengths. Frequencies of the whole set of complete proteins are indicated in blue. Frequencies of complete proteins with and without homologs are indicated in green and red respectively.



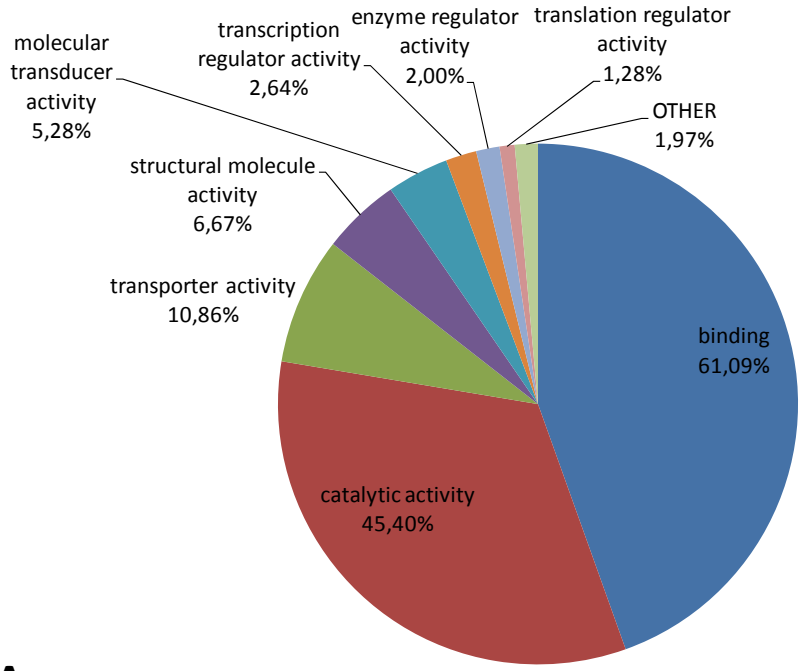
Supplemental Fig. S3

GC content and expression level of ribosomal genes. GC3 (CDS) and GC (UTRs) contents are indicated in black and white respectively.

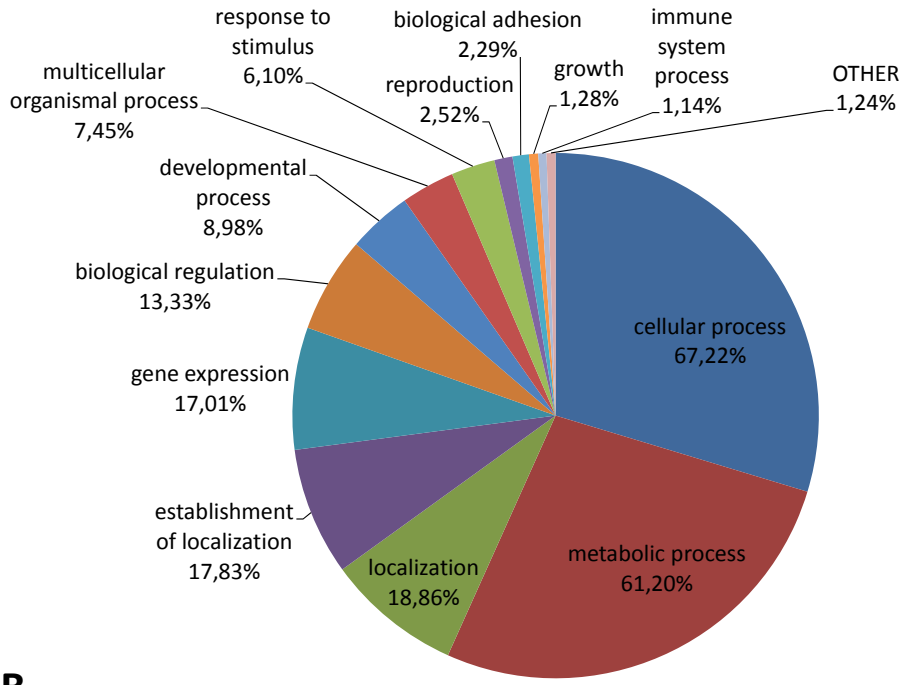


Supplemental Fig. S4

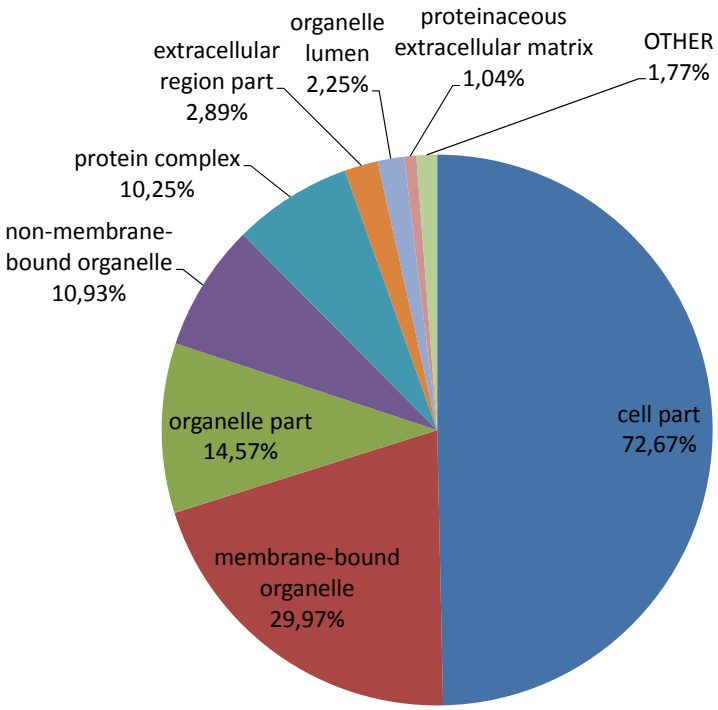
Repartition of *Alvinella* annotated proteins in GO categories. (A) Molecular function, (B) Biological process, (C) Cellular component.



A



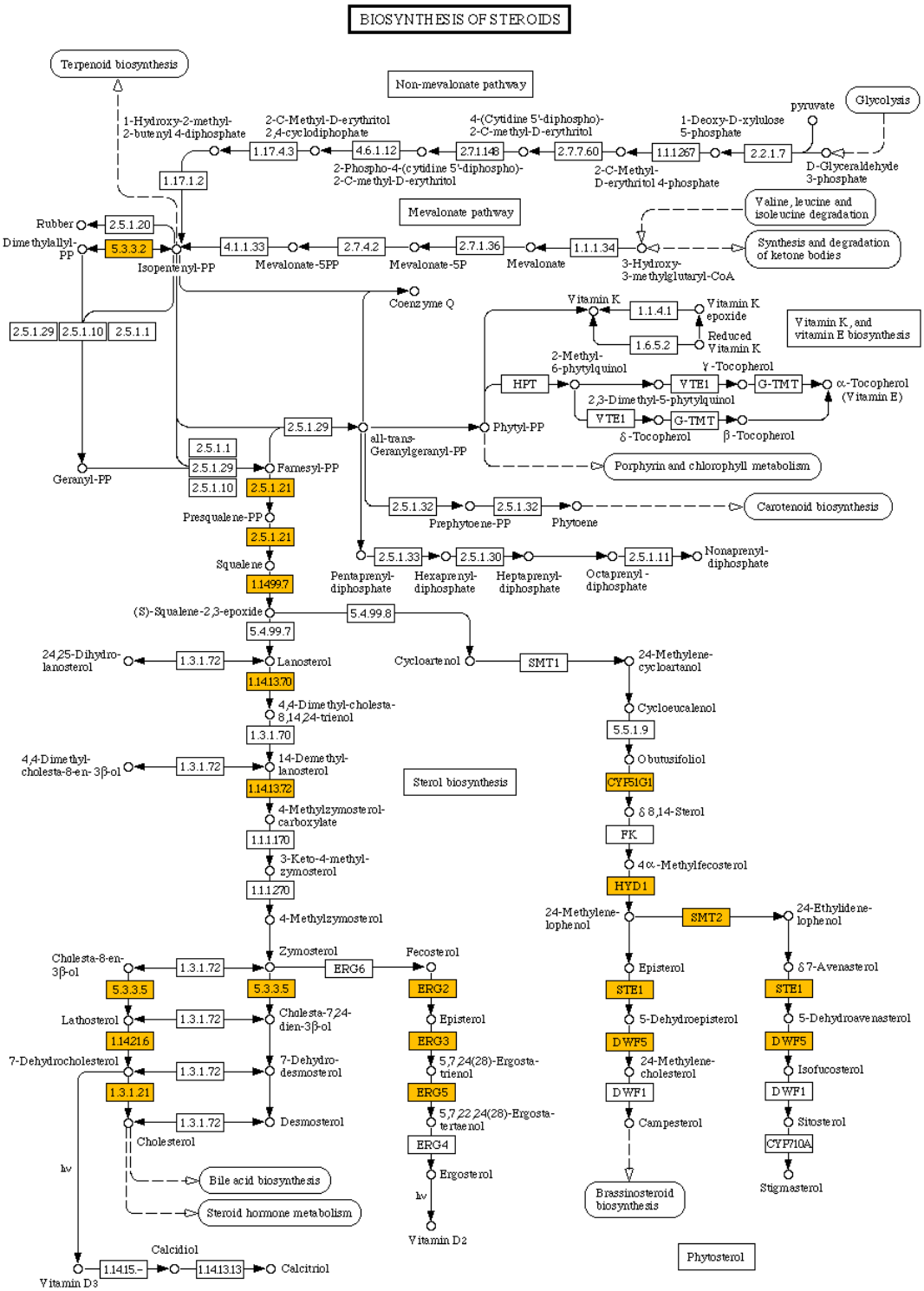
B


organelle lumen 2,25%
proteinaceous extracellular matrix 1,04%
OTHER 1,77%
extracellular region part 2,89%
protein complex 10,25%
non-membrane-bound organelle 10,93%
organelle part 14,57%
cell part 72,67%
membrane-bound organelle 29,97%

C

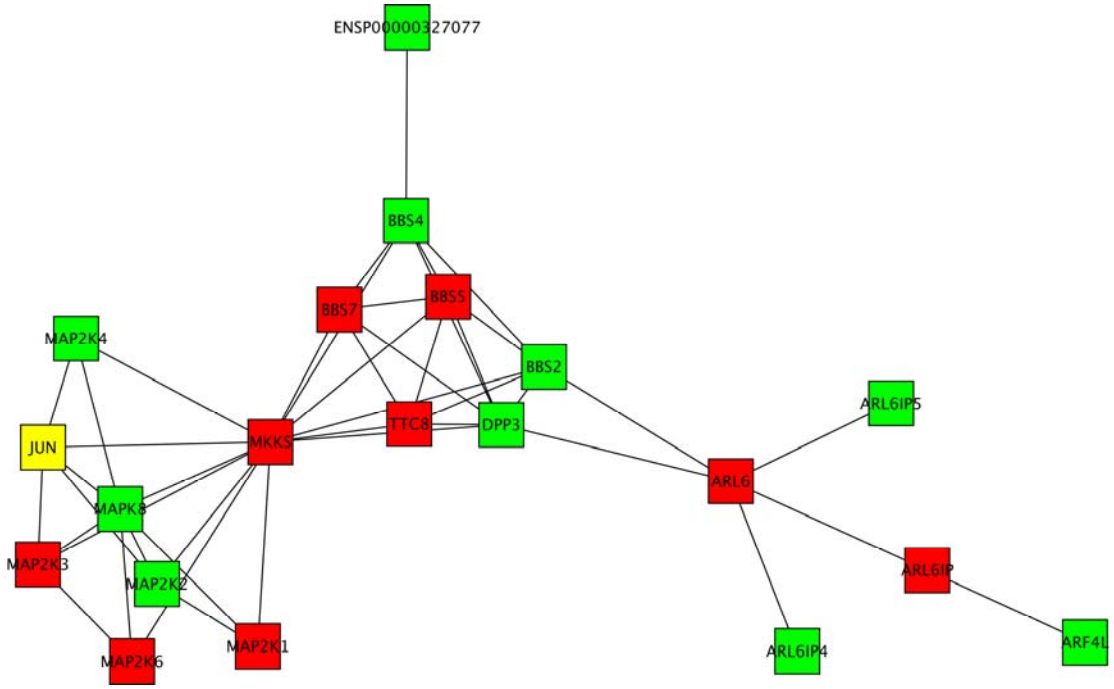
Supplemental Fig. S5.

KEGG map of the steroid biosynthesis pathway (http://www.genome.jp/kegg/pathway.html). Enzyme Classification numbers represented in our *Alvinella* libraries are indicated in orange.



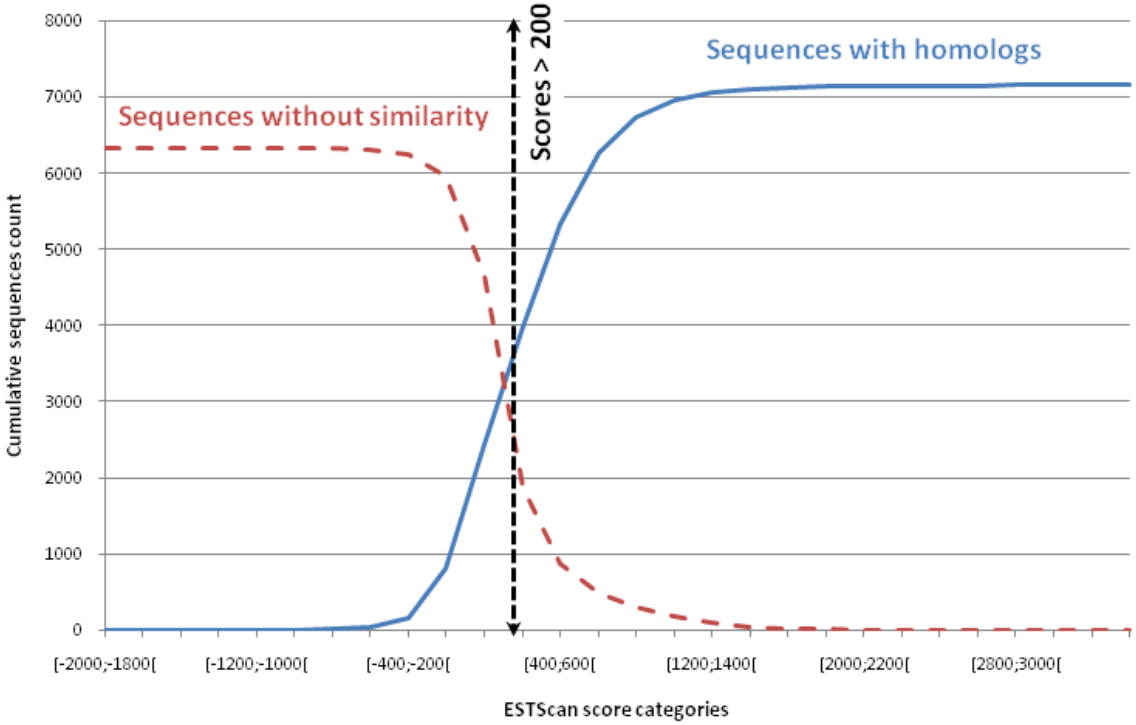
Supplemental Fig. S6

Example of a human sub-network from STRING containing genes involved in the Bardet-Biedl syndrome. The proteins are colored according to their presence or absence in *Alvinella*: in green, proteins not found in our cDNA libraries, in yellow, proteins with an ortholog in *Alvinella* and in red, proteins with a close ortholog in *Alvinella*.



Supplemental Fig. S7

Optimisation of ESTScan score cutoff. Cumulative counts of sequences with and without homologs are represented by blue and dashed red lines respectively.



Supplemental Table S1. Annotated tissue-specific transcripts in each library.

Library / Access	Definition	Reads*
Gills		
Cluster164	SCP domain containing protein	36
TERA01424	Zinc metalloproteinase	30
*TERA02097***	*Pterin-4-alpha-carbinolamine dehydratase 2 (EC 4.2.1.96)*	11
TERA01979	*Dehaloperoxidase A.*	9
TERA02011	*Mitochondrial ribosomal protein L18*	9
TERA01756	*Peroxiredoxin 1 (EC 1.11.1.15)*	8
TERA00973	Aquaporin	7
TERA01631	Fibrinogen-like protein	7
TERA01893	Cathepsin D (Aspartic protease)	7
TERA00948	OCIA domain containing protein	6
Cluster150	*Alpha tubulin.*	6
Cluster157	Polyamine oxidase	6
TERA01246	*EGF-like containing protein*	6
TERA02070	*Inositol oxygenase (EC 1.13.99.1)*	6
Ventral tissue		
TERA02620	Trefoil factor	19
TERA02630	Niemann-Pick Type C-2 protein homolog	16
TERA02543	Probable small nuclear ribonucleoprotein Sm D1	13
TERA02489	*Cytochrome c oxidase-assembly factor COX16*	9
TERA02465	WAP domain containing protein	8
TERA02592	Signal peptidase complex subunit 1	8
TERA03170	WAP domain containing protein	8
TERA02431	Histone H4	7
Pygidium		
TERA03518	*Extracellular hemoglobin subunit A2 precursor*	9
Cluster362	Alpha-aminoadipic semialdehyde dehydrogenase (EC 1.2.1.31)	8
Cluster338	*Ras-like GTP-binding protein rhoA*	7
TERA03618	Matrilin precursor	7
TERA04170	DnaJ-like protein	6
TERA04487	Calcium and integrin-binding protein	6
TERA04554	Protein kinase C iota type (EC 2.7.11.13)	6

*Last column indicates the number of reads. Only clusters exhibiting at least 6 reads are indicated.

** Clusters probably corresponding to splicing variants of transcripts found in other libraries are written in italics.

13 VISUALISATION DES DONNÉES

Afin de partager les résultats d'un projet d'annotation de génome, ou autre, entre les différentes personnes de la communauté scientifique concernées, une application Web apparaît comme l'outil de choix. Les biologistes ont alors accès à un environnement de travail homogène et dont les données sont automatiquement à jour.

Plusieurs navigateurs/visionneuses de génomes sont disponibles pour la création de sites Web, les plus connus et les plus répandus étant Ensembl (Hubbard *et al.*, 2009), UCSC Genome Browser (Kuhn *et al.*, 2009) et Generic genome browser (Stein *et al.*, 2002). Les deux premiers centralisent les données des génomes de plusieurs organismes modèles et sont très bien connus par les biologistes. Le dernier est utilisé par plusieurs sites dédiés à un organisme ou une classe d'organismes en particulier, FlyBase (Tweedie *et al.*, 2009), WormBase (Bieri *et al.*, 2007), SGD (Hong *et al.*, 2008) ou MGD (Bult *et al.*, 2008).

Cependant, ces systèmes sont conçus à des fins génériques de visualisation et ils peuvent ne pas combler certains besoins liés à un projet particulier. Au début du projet *Alvinella*, aucun de ces systèmes ne permettait de remplir certains des critères jugés indispensables pour que les biologistes puissent estimer la qualité de nos travaux et pour qu'ils puissent disposer de l'ensemble des informations possibles, notamment :

- La visualisation des données d'assemblage et des chromatogrammes,
- La visualisation des alignements multiples annotés par MACSIMS.

C'est pourquoi nous avons décidé de développer une interface Web capable d'intégrer l'ensemble de ces données tout en fournissant les services de bases que l'on peut attendre d'un tel système. Cette décision est confortée par le fait qu'un système hautement similaire a été développé dans le cadre de GarlicESTdb (Kim *et al.*, 2009), une base de séquences d'ADNc de l'aïl commun, mais qui n'est hélas pas disponible publiquement.

Dans ce chapitre, nous allons présenter l'architecture et les fonctionnalités du site Web qui ont été mises en œuvre pour le projet d'ADNc d'*Alvinella*. Enfin, sera décrit, *Sortez*, un prototype de portail de recherche original qui permet de fédérer les résultats de recherche du site Web *Alvinella* et des autres sites du laboratoire.

13.1 Interface d'accès Web

Le site Web du projet *Alvinella pompejana* est accessible à l'adresse http://alvinella.igbmc.fr (Figure 63). L'accès est à l'heure actuelle réservé aux membres du consortium Alvinella mais deviendra public après publication de l'article présenté page 153. Ce site a été développé en PHP et est hébergé sur le serveur alnitak.

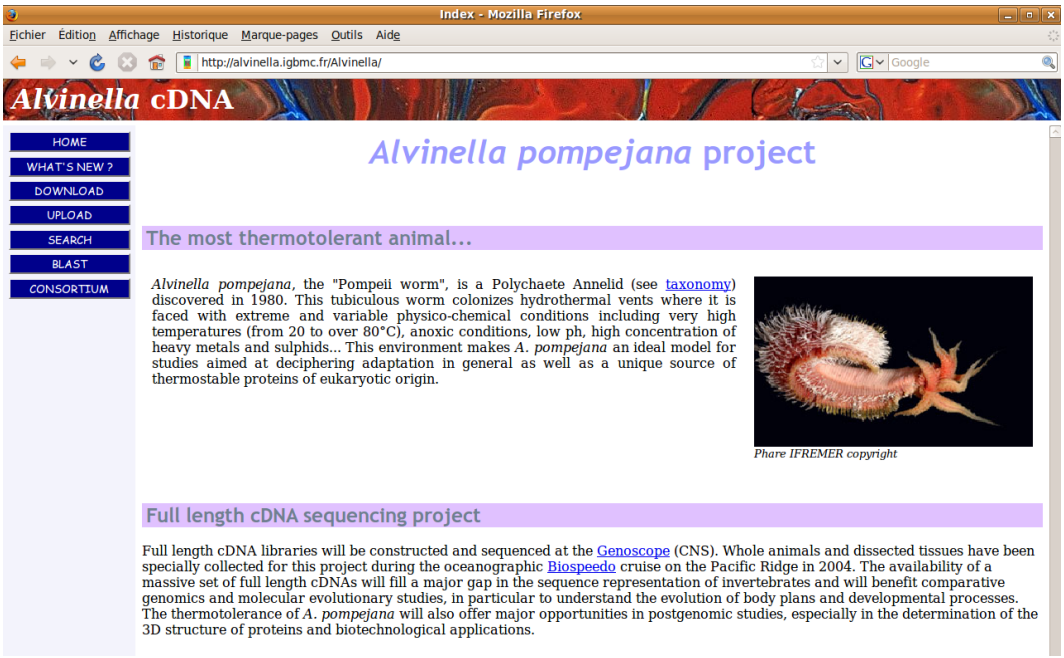


Figure 63 – Page d'accueil du site Web du projet *Alvinella pompejana*

13.1.1 Architecture du site

Le site est constitué d'une part, par plusieurs pages statiques décrivant le projet et le consortium *Alvinella* et d'autre part, par une partie entièrement dynamique dédiée à l'interrogation et à l'affichage des données générées par les deux pipelines de traitement et d'annotation. L'accès à ces informations passe en grande partie par la base de données d'assemblage, à l'exception des fichiers de chromatogrammes, des résultats de BLAST et de MACSIMS, auxquels le site accède par l'intermédiaire des disques réseau.

Toute la partie dynamique a été développée afin d'être totalement modulable et chaque module peut être utilisé indépendamment. Pour y parvenir, nous avons implémenté une architecture Modèle-Vue-Contrôleur (MVC), puis nous avons décomposé l'interface en composants réutilisables.

13.1.1.1 Architecture MVC

Cette architecture impose une séparation d'une application en trois parties (Figure 64) :

- Le modèle est la partie qui encapsule les données de l'application. C'est par son intermédiaire que l'application ajoute, récupère, met à jour ou supprime (CRUD – *Create, Retrieve, Update, Delete*) des données. Lorsque cette partie est reliée à une base de données, ou à tout autre moyen rendant les données persistantes, comme c'est le cas ici avec la base de données d'assemblage, on parle alors de DAO (*Data Access Object*),

- La vue est la partie qui prend en charge l'affichage de l'interface. Dans le cas d'un site Web, c'est elle qui génère le code HTML (*Hypertext Markup Language*) et CSS (*Cascading Style Sheets*) de la page qui va être affichée par le navigateur de l'utilisateur. En règle générale, la vue n'interroge jamais directement le modèle.
- Le contrôleur est la partie qui fait le lien entre le modèle et la vue. C'est lui qui récupère les données et réalise éventuellement des calculs ou différentes opérations sur ces données, puis les fournit à la vue. À l'inverse, lorsque l'utilisateur interagit avec l'interface, la vue transmet les informations au contrôleur qui décide de l'action à réaliser sur les données en fonction des informations qui lui sont transmises.

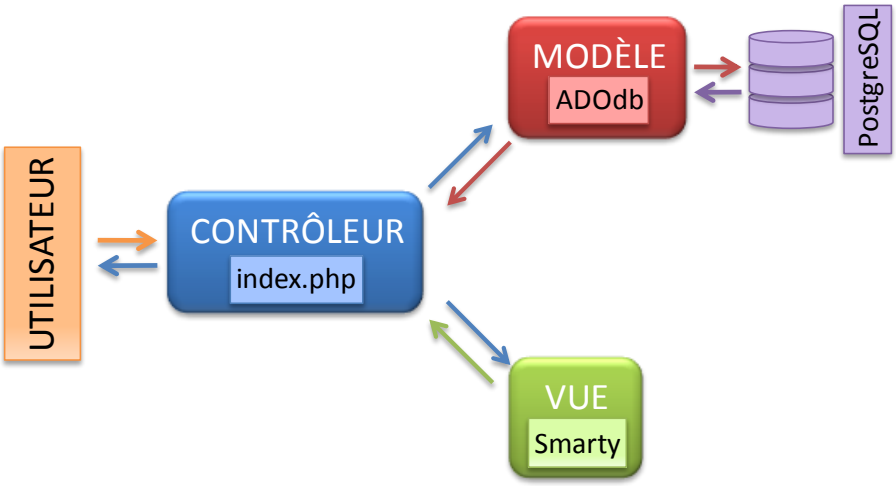


Figure 64 – Architecture MVC du site Web *Alvinella*

Puisque le développement du site a débuté juste avant la sortie de PHP5, l'accès à la base de données par le modèle n'est pas réalisé avec PDO (*PHP Data Objects*), livré avec PHP5, mais est réalisé à l'aide de la librairie ADOdb (http://adodb.sourceforge.net/). Ces deux librairies sont des couches qui permettent de s'abstraire du type de SGBDR. Ainsi, si on désire par la suite passer de PosgreSQL à MySQL ou à tout autre SGBDR, il suffira simplement de modifier l'URI (*Uniform Resource Identifier*) de connexion à la base de données dans la configuration du site.

Enfin, afin de contrôler l'accès des utilisateurs à certaines données jugées confidentielles, nous avons intégré au contrôleur la librairie de listes de contrôle d'accès phpGACL (http://phpgacl.sourceforge.net/). Ce contrôle peut alors être appliqué sur un module entier du site, la page du module va être remplacée par un message d'erreur (Figure 65), ou sur une partie de page qui sera masquée aux yeux de l'utilisateur.

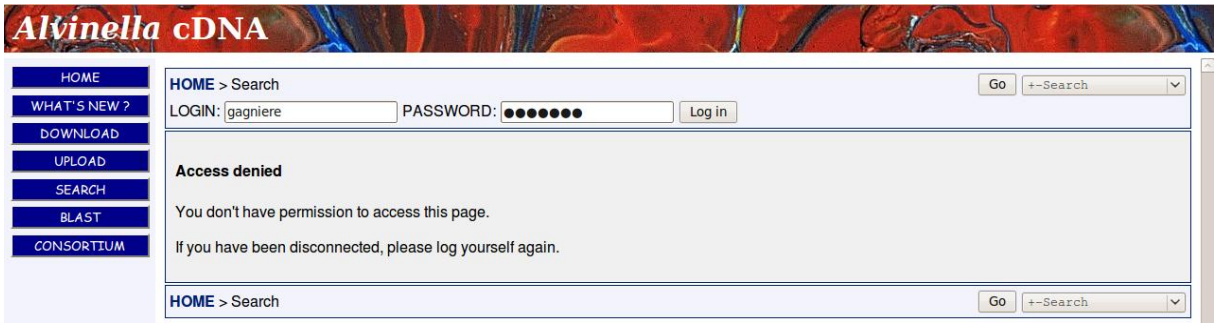


Figure 65 – Page d'erreur affichée lors de droits utilisateur insuffisants

13.1.1.2 Architecture modulaire

Le site est construit autour de composants modulables dont chacun gère une partie des pages dynamiques. Chaque composant peut être intégré au site en le déclarant dans la configuration du contrôleur.

Un composant peut gérer une ou plusieurs pages dynamiques et est composé au minimum de deux fichiers :

- un fichier contenant la logique métier, c'est-à-dire le code spécifique de la ou des pages gérées, et qui sera exécuté au niveau du contrôleur
- un fichier « modèle » (*template*) contenant du code HTML et des balises spéciales qui vont être reconnues par le moteur de template Smarty (http://www.smarty.net/). Un moteur de template permet non seulement de séparer clairement la logique métier de l'interface graphique, symbolisée par la page HTML, mais permet aussi de *designer* plus naturellement le code HTML. À l'époque du début de la conception du site, Smarty était tout simplement le moteur de template le plus flexible et le plus puissant en PHP. Smarty va traiter toutes les balises qu'il reconnait et va les substituer par un code différent, issu du traitement (Figure 66).

```
{* Smarty *}
{if isset($seqFasta)}
<pre>
{$seqFasta}
</pre>
{else}
Unknown sequence
{/if}
```

```
<pre>
&gt;TERA01043
GTGGTATATAGTACCGA
AAGGATGCTATACCTCA
CGTTTTCGAGAGTAACA
TACATTCAATGTCAGGT
…
</pre>
```

Figure 66 – Template Smarty et son résultat en HTML
À gauche le template qui affiche le contenu de la variable *$seqFasta* si elle existe dans le contexte de Smarty, sinon le message '*Unknown sequence*' est affiché. À droite le résultat en code HTML.

En fonction de paramètres fournis au contrôleur, ce dernier va vérifier la présence du composant correspondant à la page demandée, exécuter son code le cas échéant et retourner la page HTML générée par Smarty. Lorsqu'un composant demandé n'existe pas ou lorsque le composant retourne une erreur quelconque (mauvais paramètres, erreur inattendue…), la page d'accueil et les messages d'erreurs sont affichés à la place de la page demandée.

13.1.2 Description de l'interface

Outre le fait de pouvoir télécharger les fichiers de séquences des banques complètes d'*Alvinella*, et de pouvoir visualiser différentes informations liées au projet et au Consortium *Alvinella*, l'utilisateur a accès à un certain nombre de modules. Ces modules peuvent être divisés en deux catégories : les modules de recherche et les modules de visualisation.

13.1.2.1 Modules de recherche

13.1.2.1.1 Module de recherche textuelle

Le module de recherche textuelle permet de rechercher des séquences par l'intermédiaire de leur annotation, ou permet d'accéder directement à des séquences en fournissant une liste de numéro d'accès. Il est possible de sélectionner dans quelles annotations rechercher et de limiter la recherche à plusieurs banques et types de séquences (Figure 67).

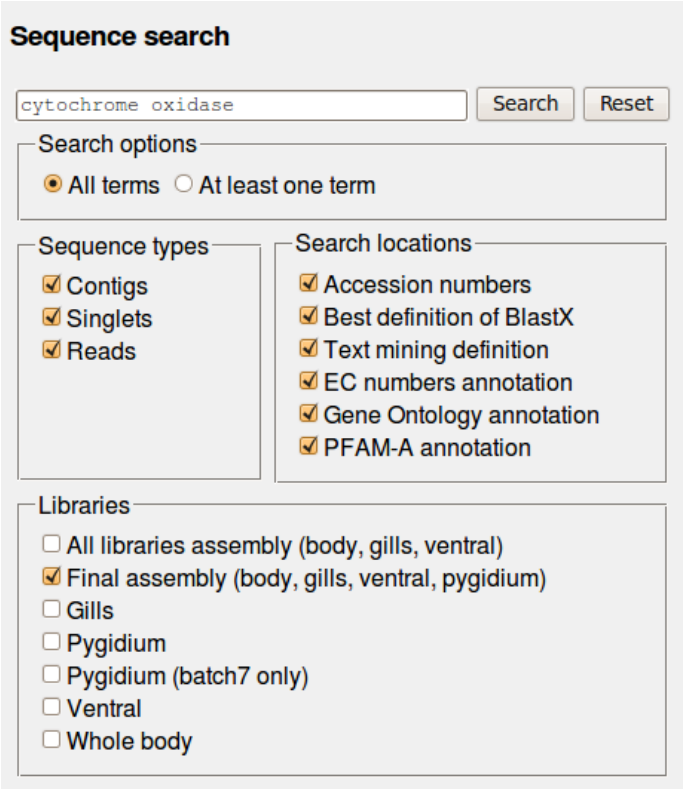


Figure 67 – Formulaire de recherche textuelle

Les résultats de ce type de recherche sont disposés sous forme de liste de séquences nucléotidiques auxquelles sont rattachées des informations sur les ESTs qui les composent et les annotations des séquences protéiques correspondantes (Figure 68). Les annotations ne sont pas affichées par défaut, offrant ainsi une vision synthétique des résultats. Cependant, un ou plusieurs types d'annotations peuvent être sélectionnés et affichés à la demande.

Des indicateurs apparaissent dans les colonnes de gauche, afin de renseigner l'utilisateur sur quels types annotations sa requête a été retrouvée, et en combien d'exemplaires.

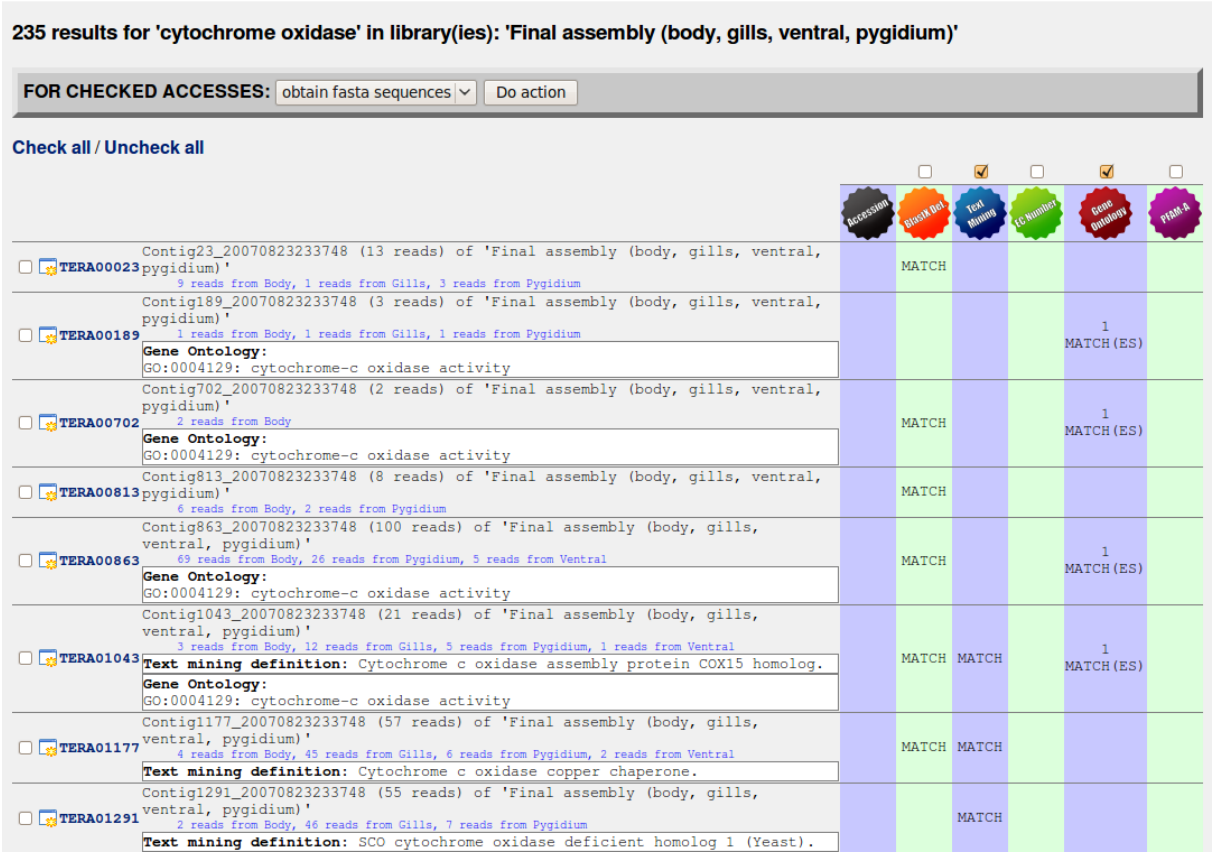


Figure 68 – Résultats de recherche textuelle des termes 'cytochrome oxidase'
Ici, les annotations de définition fonctionnelle et de Gene Ontology apparaissent dans les cadres blancs du volet de gauche.

13.1.2.1.2 Module de recherche par similarité

Ce module permet de réaliser des recherches BLAST (BLASTN, TBLASTN et TBLASTX) à l'intérieur des différentes banques d'ADNc d'*Alvinella* (Figure 69). Lorsque ce module a la possibilité d'exécuter ses recherches dans une banque protéique, il active les recherches de type BLASTP et BLASTX.

La page de résultats de recherches transforme les résultats bruts en ajoutant des liens hypertexte afin de naviguer facilement au travers de cette page et d'accéder directement à une page de séquence en cliquant sur un numéro d'accès. Si les numéros d'accès sont

reconnus comme appartenant à une banque de séquence publique, l'utilisateur est redirigé sur le site SRS du BIPS (http://bips.u-strasbg.fr/srs83/frontpage.do). L'utilisateur a aussi la possibilité de télécharger le fichier de résultats au format plat pour pouvoir le conserver.

Ce module a été réutilisé dans la conception de plusieurs sites Web du laboratoire, dont RETINOBASE (Annexe 2 – Publication 3 : RETINOBASE page 191).

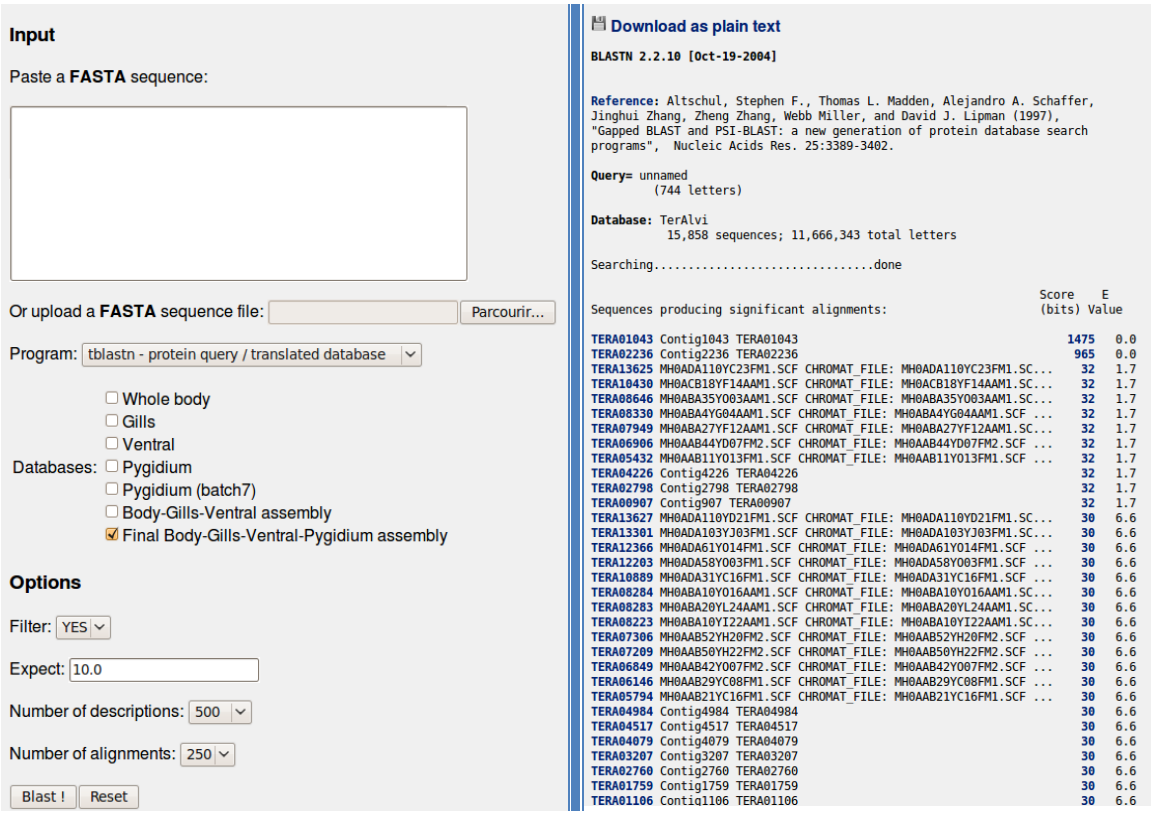


Figure 69 – Formulaire du module de recherche par similarité et un résultat de recherche

13.1.2.2 Modules de visualisation

Les différents modules de visualisation sont intégrés au sein d'une même page où apparait une barre de navigation symbolisée par des boutons violets (Cf. Figure 70, page 164). Chaque bouton représente un module de visualisation et l'assortiment de modules disponibles varie selon le type de séquence visualisée. Par exemple, on ne trouvera pas de module de visualisation d'alignement pour une séquence de singleton.

13.1.2.2.1 Module de visualisation de séquence

Ce module est le plus simple et permet la visualisation de séquences converties au format FASTA (Figure 70).

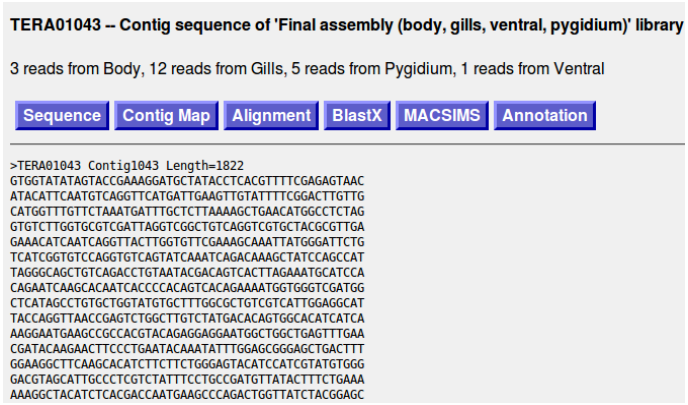


Figure 70 – Module de visualisation de séquence

13.1.2.2.2 Module de visualisation de carte de contig

Ce module construit à la volée une image schématique d'un contig et de ses séquences d'EST, à la manière d'un navigateur de génome (Figure 71). Chaque élément de la carte est un lien qui permet d'accéder à une séquence particulière ainsi qu'à ses modules de visualisation. Le sens de l'alignement d'une séquence est symbolisé par une pointe de flèche. Des indications colorées permet de marquer des zones de séquences non alignées (en jaune, limité à 30 bp selon les paramètres d'assemblage Cap3 décrits page 117) ou des zones masquées par des caractères 'X' (en violet).

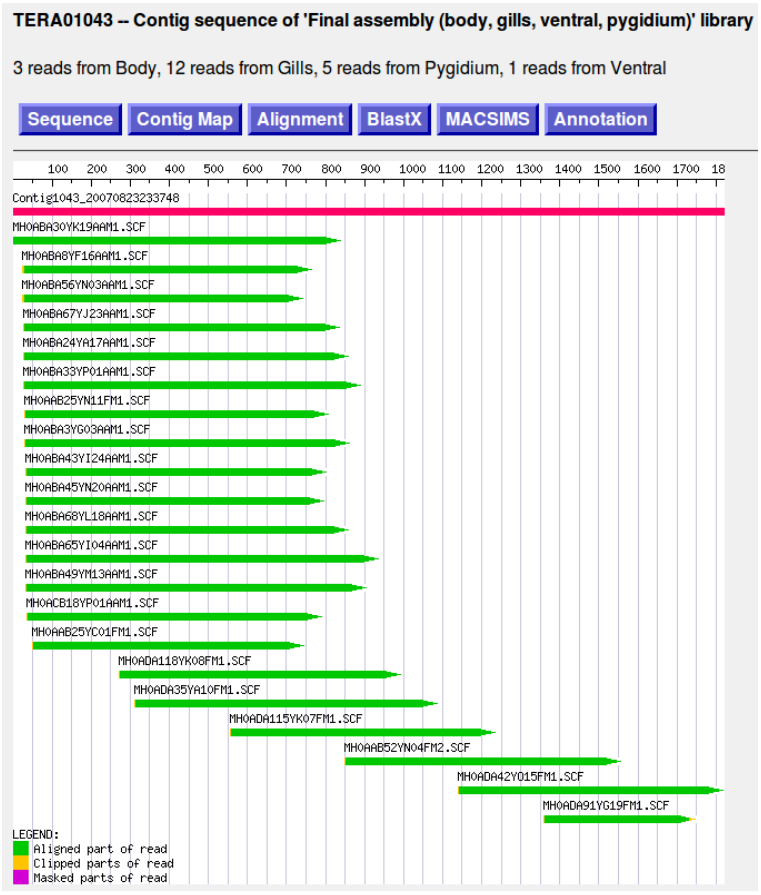


Figure 71 – Visualisation d'un contig sous forme de carte

13.1.2.2.3 Module de visualisation d'alignement de contig

Ce module représente l'alignement d'un contig en mimant l'affichage du logiciel Consed (Figure 72). Consed oriente son affichage sur la qualité des séquences visualisées. En conséquence, chaque résidu est « surligné » d'un dégradé de gris, la couleur blanche représentant de hautes valeurs de qualité (≥ 40), et les couleurs grises des qualités moindres. Les résidus qui présentent un désaccord avec la séquence consensus apparaissent en rouge.

Il est possible de télécharger l'alignement sous forme d'image (pour publication) ou dans un format de fichier MSF. Tout comme la carte de contig, chaque séquence est un lien hypertexte qui mène directement au module de visualisation de chromatogramme de la séquence considérée.

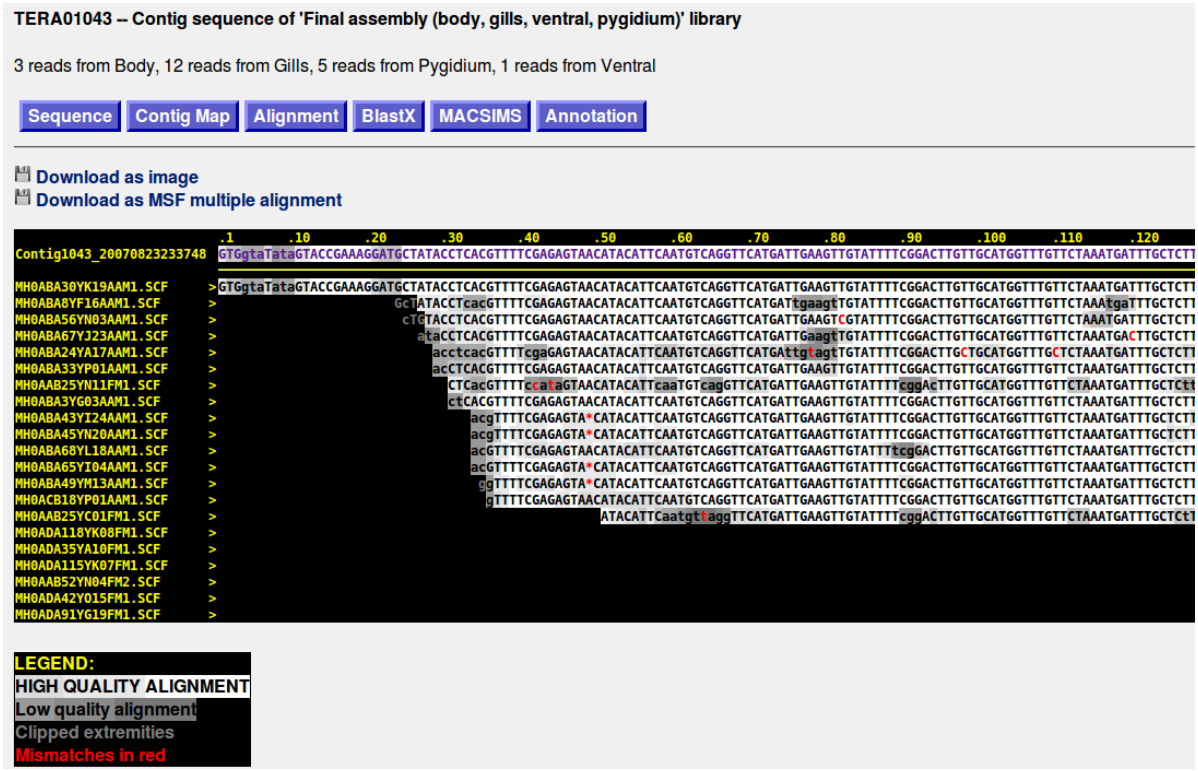


Figure 72 – Visualisation d'un contig sous forme d'alignement

13.1.2.2.4 Visualisation de chromatogramme

Ce module permet l'affichage de chromatogrammes au format SCF. Une traduction dans les 6 cadres de lecture de la séquence nucléotidique contenue dans chaque chromatogramme apparait en bas (Figure 73).

Cet affichage permet de vérifier que le travail de nettoyage en amont de toute la cascade d'assemblage et d'annotation a bien été réalisé. Dans la Figure 73, il apparait clairement que

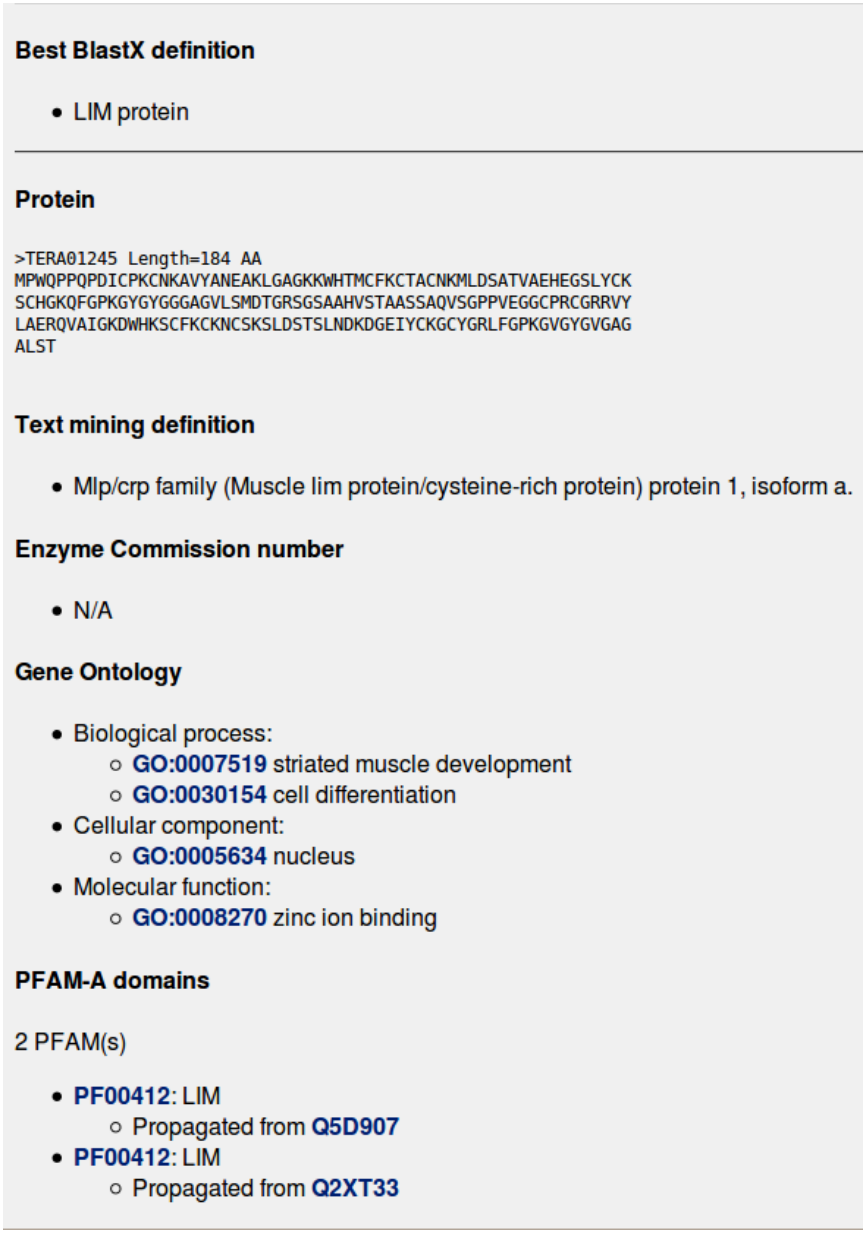
la première partie de mauvaise qualité du chromatogramme n'a pas été prise en compte dans la création de la séquence nucléotidique. De plus, la visualisation des cadres de lecture permet de vérifier si l'apparition de codons stops ou de *frameshifts* est bien réelle ou s'il s'agit d'un problème de *base-calling* (pics chevauchants par exemple).

Figure 73 – Visualisation d'un chromatogramme

13.1.2.2.5 Module de visualisation d'annotation intégrative

Ce module regroupe en une seule page les résultats du pipeline d'annotation intégrative (Figure 74). Les résultats de cartographie KEGG PATHWAY et STRING ne sont pas encore intégrés à ce module. Plusieurs liens hypertexte apparaissent selon les annotations attribuées :

- Les termes GO, sont liés au navigateur Gene Ontology AmiGO (http://amigo.geneontology.org/)
- Les numéros EC sont liés au site ENZYME (http://www.expasy.ch/enzyme/)
- Les domaines Pfam sont liés au site Pfam (http://pfam.sanger.ac.uk/)
- Les numéros d'accès de séquences publiques sont liés au site SRS du BIPS



Best BlastX definition

- LIM protein

Protein

```
>TERA01245 Length=184 AA
MPWQPPQPDICPKCNKAVYANEAKLGAGKKWHTMCFKCTACNKMLDSATVAEHEGSLYCK
SCHGKQFGPKGYGYGGGAGVLSMDTGRSGSAAHVSTAASSAQVSGPPVEGGCPRCGRRVY
LAERQVAIGKDWHKSCFKCKNCSKSLDSTSLNDKDGEIYCKGCYGRLFGPKGVGYGVGAG
ALST
```

Text mining definition

- Mlp/crp family (Muscle lim protein/cysteine-rich protein) protein 1, isoform a.

Enzyme Commission number

- N/A

Gene Ontology

- Biological process:
  - GO:0007519 striated muscle development
  - GO:0030154 cell differentiation
- Cellular component:
  - GO:0005634 nucleus
- Molecular function:
  - GO:0008270 zinc ion binding

PFAM-A domains

2 PFAM(s)

- PF00412: LIM
  - Propagated from Q5D907
- PF00412: LIM
  - Propagated from Q2XT33

Figure 74 – Visualisation de l'annotation intégrative d'une séquence protéique
Exemple de protéine liée à la biologie du muscle qui comporte 2 domaines doigt de zinc de type LIM

13.1.2.2.6 Module de visualisation de résultats MACSIMS

Ce module permet la visualisation d'alignements multiples annotés par MACSIMS. Un menu déroulant permet à tout moment de choisir le type d'annotations à visualiser (Figure 75). Ces annotations englobent les blocs de conservation (BLOCK), les régions homologues englobant plusieurs blocs de conservation (REGION), les régions de faible complexité (LOWCOMP), les domaines Pfam-A ou PROSITE, les sites actifs (SITE), les domaines transmembranaires (TRANSMEMB), les peptides signaux (SIGNAL)…

En plus des annotations réalisées par MACSIMS, un mode de coloration basé sur les propriétés physico-chimiques des résidus, inspiré du logiciel Seqlab (package GCG) a été rajouté.

Le changement d'affichage du type d'annotations s'effectue par l'intermédiaire de requêtes AJAX (*Asynchronous JavaScript and XML*) à l'aide de la librairie JavaScript MooTools (http://mootools.net/), ce qui évite de recharger continuellement la page et permet d'alléger la charge du serveur tout en rendant plus confortable l'expérience utilisateur.

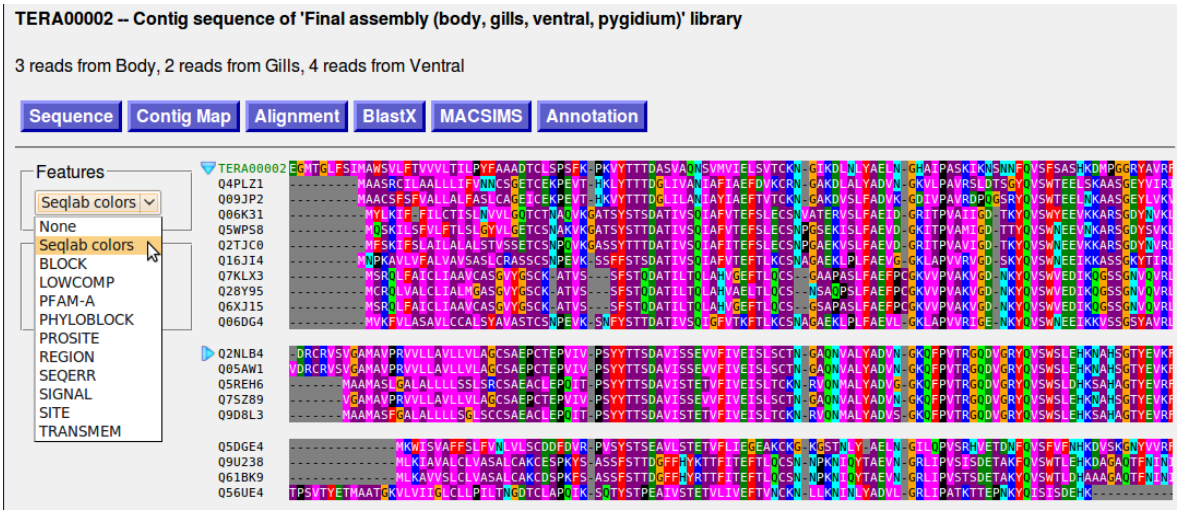


Figure 75 – Affichage d'un alignement annoté par MACSIMS avec la liste des types d'annotations

Lors de l'affichage d'annotations, le passage de la souris au dessus d'une annotation provoque l'affichage d'une bulle d'aide contenant les détails de cette annotation (Figure 76).

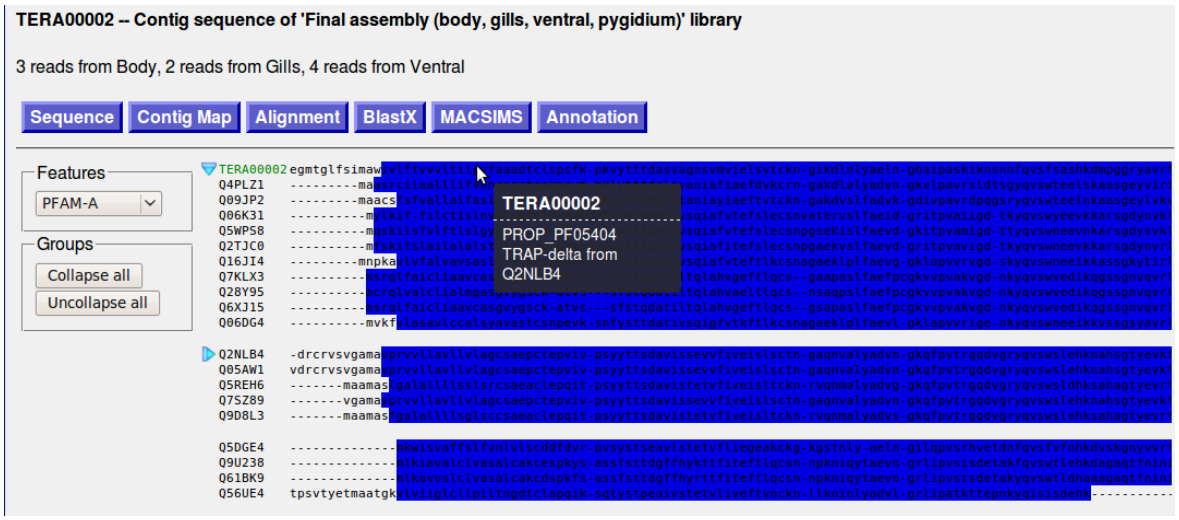


Figure 76 – Bulle d'aide lors du passage de la souris sur une annotation
Ici, le domaine Pfam-A TRAP-delta (numéro d'accès PF05404) a été propagé sur la séquence d'*Alvinella* TERA00002 à partir de la séquence Q2NLB4.

13.2 Sortez

Sortez est un prototype de portail de recherche développé à l'initiative de Guillaume Berthommier (LBGI) et moi-même afin de faciliter les recherches à l'intérieur des différentes bases de données du laboratoire. L'idée de créer un portail de recherche a germé lorsque l'on s'est aperçu que les données du LBGI étant distribuées au sein de sites Web différents, cela obligait un utilisateur à passer de l'un à l'autre pour effectuer ses recherches. Afin de fédérer les recherches sur de multiples bases de données par l'intermédiaire d'un seul site Web, nous nous sommes inspirés des portails *Entrez* du NCBI (http://www.ncbi.nlm.nih.gov/) et *EB-eye Search* de l'EBI (http://www.ebi.ac.uk/).

Le nom de « Sortez » a bien évidemment été choisi en référence au portail de recherche « Entrez » du NCBI. Le logo est quant à lui un hommage à la sublimissime série *Ace Attorney* publiée par CAPCOM (http://www.ace-attorney.com/). Sortez est disponible publiquement à l'adresse http://sortez.igbmc.fr/ (Figure 77).

Figure 77 – Page d'accueil du portail de recherche Sortez
La barre de recherche apparait en permanence sous le logo un peu volumineux de Sortez. Cette barre est facilement intégrable à d'autres sites pour lancer des recherches sur Sortez.

13.2.1 Enregistrement des sites Web auprès de Sortez

Sortez ne requiert qu'un seul fichier de configuration XML dans lequel sont uniquement indiqués les URI des fichiers de description XML des sites que l'on souhaite enregistrer auprès de Sortez. Ce descripteur comporte obligatoirement 4 informations (Cf. Annexe 6, page 199) :

- Le nom du site,
- La description du site,
- L'URL (*Uniform Resource Locator*) de la page d'accueil,
- Les URL du logo et d'une icône identifiant le site.

Ces informations permettent au site d'apparaître dans la vitrine de Sortez (Figure 78), qui affiche de manière homogène la liste de tous les sites enregistrés.

Figure 78 – Extrait de la vitrine de Sortez
Chacun des sites enregistrés apparait de manière standardisée avec son titre, sa description, son logo et un lien vers sa page d'acceuil.

Le fait d'apparaître dans la vitrine ne permet pas de réaliser des recherches à travers les sites enregistrés. Pour cela il faut déclarer un *Tracker* dans le descripteur associé au site Web.

13.2.2 Trackers Sortez

Dans la terminologie de Sortez, un *tracker* désigne le programme qui effectue réellement un travail de recherche sur un site Web, en accédant directement à sa base de données. Le langage dans lequel est implémenté le *tracker* importe peu, il suffit juste qu'il puisse dialoguer avec le noyau de Sortez, par l'intermédiaire d'un langage standardisé très simple au travers d'Internet (protocole HTTP). Un *tracker* Sortez est donc un service Web.

13.2.2.1 Architecture orientée service

Pour que Sortez soit le plus indépendant possible vis-à-vis des sites Web qu'il interroge, il ne doit pas accéder directement aux bases de données des autres sites. Ce travail est délégué

aux différents *trackers* qui sont interrogés par Sortez en tant que service Web. Ce type d'architecture basée sur plusieurs services Web est dénommée SOA (*Service Oriented Architecture*) (Figure 79).

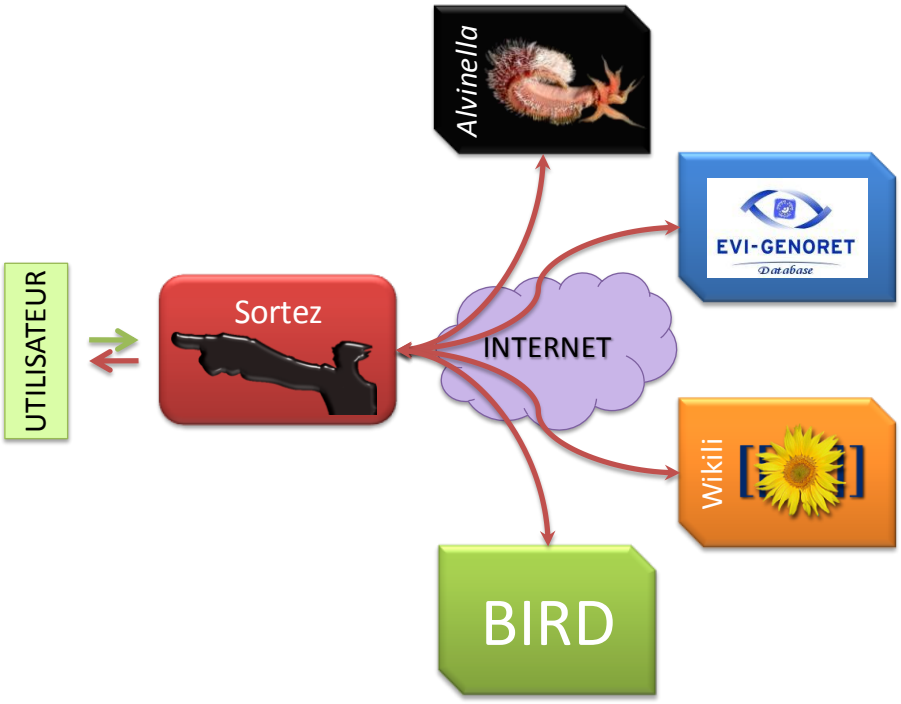


Figure 79 – Architecture SOA de Sortez, interrogeant quatre services Web
L'utilisateur lance une recherche sur le portail Sortez, qui la retransmet aux différents *trackers* enregistrés. Les *trackers* retournent alors leur réponse qui est affichée à l'utilisateur par Sortez.

13.2.2.2 Recherches Sortez

Lorsqu'un utilisateur lance une recherche, Sortez questionne en parallèle chaque *tracker* enregistré à l'aide de requêtes AJAX. La possibilité est laissée à l'utilisateur de ne lancer la recherche que dans un seul site.

Nous avons souhaité que l'architecture de Sortez reste relativement simple et légère. Pour cela, nous avons évité l'utilisation de standards lourds tels que WSDL (*Web Services Description Language*) et SOAP (*Simple Object Access Protocol*). S'ils présentent de nets avantages en termes de compatibilité et d'autonégociation de découverte de service, ces standards restent très verbeux et demanderaient un investissement de temps important à une personne qui voudrait implémenter son propre *tracker*.

Nous avons donc opté pour que les messages échangés entre Sortez et les services Web soient en JSON (*JavaScript Object Notation*) en suivant un « standard » facile à comprendre et à mettre en oeuvre.

Puisqu'un site Web peut vouloir hiérarchiser ses résultats de recherche, chaque *tracker* se doit de retourner ses résultats classifiés dans des sections et doit retourner le nombre total

de résultats (Figure 80). Ces sections doivent obligatoirement être déclarées dans le descripteur Sortez (Cf. Annexe 6, page 199).


```
{
    "query":"cytochrome oxidase",
    "database":{
        "count":125
    },
    "sections":{
        "accession":{
            "count":0,
            "hits":[]
        },
        "blastx": {
            "count":125,
            "hits":[
                {
                    "libraryName":"Whole body",
                    "type":"Contig",
                    "accessLink":"<a href=\"...\">ALBO0227<\/a>"
                },
                {
                    "libraryName":"Whole body",
                    "type":"Contig",
                    "accessLink":"<a href=\"...\">ALBO0540<\/a>"
                },
                // 123 autres résultats...
            ]
        }
    }
}
```


Figure 80 – Exemple de réponse JSON retournée par le *tracker* du site Web *Alvinella*

En tant que prototype, seules deux sections apparaissent : les numéros d'accès (‘*accession*’) et les descriptions des meilleurs hits BLASTX (‘*blastx*’). Ici, 125 résultats sont retournés suite à la recherche des termes ‘cytochrome oxidase’. Comme cela va être décrit plus loin, chaque résultat est composé d'un ensemble de variables spécifiques au site Web interrogé.

Lorsque Sortez reçoit une réponse de la part de l'un des *trackers*, le nombre de résultats retournés est mis à jour en regard du nom du site et de chaque section associées, dans la page de recherche (Figure 81).

Figure 81 – Résultats d'une recherche dans Sortez avec la liste dépliée des sections du site *Alvinella*

13.2.3 Résultats de recherche détaillés

À partir de la page de recherche, l'utilisateur peut accéder à une page affichant la liste détaillée des résultats de recherche pour un site distinct (Figure 82). Les résultats de chaque section de recherche sont rassemblés à l'aide d'onglets, les résultats de chaque onglet étant paginés. Les sections de recherche n'ayant retourné aucun résultat n'apparaissent pas dans la liste d'onglets.

Figure 82 – Détails d'une recherche Sortez

Comme on peut le remarquer dans la Figure 80 (page 172), chaque résultat de recherche est retourné par le *tracker* sous la forme d'une liste de variables JavaScript, le nombre et la nature de ces variables étant dépendants du *tracker* et de la section de recherche.

Afin de laisser un certain degré de liberté aux sites Web interrogés par Sortez, l'affichage de chaque élément de la liste détaillée est personnalisable par l'intermédiaire d'un mini moteur de *templates* intégré à Sortez. Les *templates* sont déclarés dans les descripteurs des sites (Cf. Annexe 6, page 199) et peuvent être différents pour chaque section de recherche du site.

De cette manière, chaque site peut maîtriser la quantité d'informations qu'il fournit ainsi que la manière de les afficher. Dans le cas d'un site où la sécurité des informations est importante, comme c'est le cas pour le site Web *Alvinella*, il devient possible d'afficher un minimum d'informations, accompagnées d'un lien hypertexte qui redirige l'utilisateur vers le site en question, qui demandera un mot de passe avant d'afficher tout autre d'information.

13.2.4 Trackers Sortez disponibles

Outre le site Web *Alvinella*, il existe actuellement 3 autres trackers dans lesquels Sortez permet de lancer des recherches :

- Le site EVI-GENORET database (http://genoret.igbmc.fr/), qui centralise des données phénotypiques, génomiques et protéomiques de maladies rétiniennes du projet européen EVI-GENORET (http://www.evi-genoret.org/),
- Le système BIRD, qui recherche à l'intérieur de toutes les banques de séquences publiques disponibles localement,
- Wikili (http://alnitak.igbmc.fr/wikili/), le wiki public de notre laboratoire, basé sur le moteur MediaWiki (http://www.mediawiki.org/).

13.3 Discussion et perspectives

Par le déploiement de ces deux sites, et plus particulièrement du site Web *Alvinella*, nous avons pris grand soin de mettre à disposition des membres du consortium *Alvinella*, par un vecteur d'information simple et convivial, un accès exhaustif aux travaux effectués sur les banques d'ADNc d'*Alvinella* (bien qu'il reste encore à intégrer l'annotation de second niveau des réseaux KEGG PATHWAY et STRING).

A l'avenir, lorsque les séquences seront rendues publiques et accessibles à partir des banques de séquences publiques, il est envisagé de déployer un système DAS (*Distributed Annotation System*) (Dowell *et al.*, 2001) afin de partager avec la communauté scientifique notre travail d'annotation.

De plus, la conception de ces sites m'a permis d'acquérir des compétences dans les technologies Web et l'administration système. Ces compétences ont été mises à profit lors

de la conception et/ou du déploiement des systèmes SM2PH-db et RETINOBASE, qui ont chacun fait l'objet d'une publication présenté en Annexe 1 et Annexe 2, respectivement.

13.3.1 SM2PH-db

SM2PH-db est une base de données mutationnelles des maladies monogéniques humaines élaborée dans le cadre du projet Decrypthon (http://www.decrypthon.fr/). SM2PH-db réalise l'intégration automatique et périodique de sources de données hétérogènes (OMIM, UniProtKB, Genecards, GO, plusieurs bases de données locus-spécifiques…).

Afin de procurer un environnement regroupant les données liées à la séquence, la structure et l'évolution, permettant d'interconnecter les informations génotypiques et phénotypiques, SM2PH-db établit automatiquement des MACS annotés par MACSIMS, ainsi que des prédictions de modèles structuraux des protéines mutantes à partir des modèles de protéines sauvages.

L'interface Web de SM2PH-db permet de représenter cet environnement sous la forme d'une interface interactive avec un affichage des données structurales et des données d'alignement multiple synchronisés. Outre l'affichage des modèles pré-calculés, cette interface permet à un utilisateur de générer à la volée des modèles de mutants arbitraires afin de mener à bien ses investigations.

À l'heure actuelle, SM2PH-db contient 28 000 mutations non-synonymes dont 20 000 sont considérées comme induisant une maladie. Un total de 1 600 modèles structuraux de protéines sauvages, dont 1 400 protéines différentes, est déjà pré-calculé et accessible.

SM2PH-db est accessible à l'adresse http://decrypthon.igbmc.fr/sm2ph/.

13.3.2 RETINOBASE

RETINOBASE est une base de données dédiée à l'intégration d'expériences de puces à ADN Affymetrix menées sur la rétine, et qui permet le stockage et l'analyse automatique des données d'expression. RETINOBASE est couplée à un site Web permettant la visualisation et l'interrogation de ces données.

L'analyse automatique effectuée par RETINOBASE englobe la normalisation des données à l'aide de 3 algorithmes (RMA, dChip et MAS5) et le partitionnement des profils d'expression à l'aide de 3 méthodes (2 méthodes dérivées de celle des k-means et une basée sur un modèle de mélanges).

Le site Web de RETINOBASE permet la visualisation et l'interrogation de ces résultats d'analyses à l'aide de formulaires questionnant les gènes des différentes puces Affymetrix, les intensités des signaux lumineux, et le profils d'expression. Les formulaires pour questionner les gènes sont complétés par le module de similarité BLAST décrit page 162.

Actuellement, RETINOBASE comprend 37 expériences différentes qui englobent 6 organismes modèles (humain, souris, rat, drosophile, poisson zèbre et poulet). RETINOBASE est accessible à l'adresse http://alnitak.u-strasbg.fr/RetinoBase/, sous réserve de demander un compte à l'adresse ripp@igbmc.fr.

13.3.3 BioG et Wolcanno… le futur ?

Lors de la conception du site Web *Alvinella*, il est apparu un manque évident de librairies PHP permettant de traiter les données biologiques.

L'*Open Bioinformatics Foundation* (OBF) (http://www.open-bio.org/) est une organisation supportant des projets *open source* dédiés à la bioinformatique, dont la famille de librairies Bio* (BioJava, BioPerl, BioPython, BioRuby, et BioSQL), qui sont des librairies permettant de manipuler de manière exhaustive des données biologiques dans un langage de programmation donné. Une initiative de librairie BioPHP existe également (http://biophp.org/), mais l'OBF ne la supporte pas du fait de sa qualité critiquable.

C'est pourquoi j'ai entrepris de créer ma propre librairie, appelée BioG, en m'inspirant fortement du modèle objet de BioJava. Au moment de l'écriture de ce manuscrit, BioG comporte 98 classes et interfaces, regroupées en 8 *packages* (Tableau 14) et permet déjà la lecture et l'écriture de 9 formats de fichiers biologiques (FASTA, MSF, RSF, EMBL, GENBANK…) ainsi que le rendu de séquences et d'alignements multiples sous forme de composants HTML ou sous forme d'images.

Tableau 14 – Les différents *packages* de la librairie BioG

Package	***Description***
core	Classes de base du modèle (sequence, annotation, feature, …)
collection	Réimplémentation de quelques structures de données du langage Java
exception	Hiérarchie d'exceptions utilisées par la librairie pour signaler des erreurs
utils	Diverses classes utilitaires
io	Classes de lecture et d'écriture de différents formats de fichiers
rendering	Classes permettant des rendus différents de séquences et des alignements en HTML ou sous forme d'image
html	Classes permettant de générer automatiquement du code HTML et CSS
image	Classes permettant de composer des images

L'idée principale est d'utiliser cette librairie pour le futur Wolcanno (*Web Viewer Of aLvinella cDNA ANNOtation*), une refonte totale du site Web *Alvinella* en utilisant un *framework* PHP5 moderne tel que Symfony (http://www.symfony-project.org/) ou Zend Framework

(http://framework.zend.com/), en lieu et place d'un ensemble de librairies hétérogènes (ADOdb, Smarty…).

Ces deux *frameworks* proposent tout un environnement de développement Web architecturé autour d'un modèle MVC complet et cohérent. L'accès aux bases de données s'effectue à l'aide d'un modèle ORM (Object Relational Mapping) qui permet de s'abstraire de la couche d'interrogation SQL des SGBDR en manipulant les données sous forme d'objets.

Les *frameworks* Symfony et Zend se différencient surtout au niveau du principe développement : Symfony est orienté développement par paramétrage des fichiers de configurations d'un squelette d'application préétabli, alors que Zend propose plutôt des blocs de construction à assembler afin de n'utiliser que les fonctionnalités requises pour l'application en cours de développement.

Cependant le squelette préétabli de Symfony a l'avantage d'être standardisé et facilite donc la distribution et l'échange de modules applicatifs. Il dispose d'ailleurs de fonctionnalités permettant de télécharger automatiquement des modules à partir de dépôts de type PEAR (http://pear.php.net/). Cette fonctionnalité semble idéale pour distribuer et réutiliser les modules du site *Alvinella*.

CONCLUSION ET PERSPECTIVES

14 CONCLUSIONS ET PERSPECTIVES

L'ère post-génomique est caractérisée par un flux de données s'accroissant au rythme des avancées technologiques réalisées dans les différents domaines gravitant autour de la biologie moléculaire et cellulaire et de l'automatisation de l'acquisition de données biologiques. Cette abondance de données nous offre dès à présent l'opportunité d'étudier non seulement l'expression de millions de gènes et de protéines d'un panel d'organismes couvrant une bonne partie de l'arbre de la vie, mais aussi leurs interactions en modélisant des système biologiques à des niveaux cellulaires, tissulaires ou même au niveau d'organismes entiers et ce aux différents stades de développement et de maturité.

Au cours de cette thèse, j'ai eu la chance d'être impliqué directement en amont de ce flux de données en analysant les données de séquençage des banques d'ADNc d'*Alvinella pompejana*. Cela m'a permis d'appréhender trois aspects essentiels à la compréhension de ces données ; la transformation des données brutes en séquences exploitables de haute qualité par l'élimination d'une grande partie du bruit résiduel associé aux techniques de clonage et séquençage, la caractérisation de ces séquences et des interactions potentielles de leurs équivalent protéiques par l'intermédiaire de l'annotation intégrative de protéines, et la présentation et la diffusion de ces données à différents niveaux de leurs traitements à l'aide d'interfaces Web.

Pour chacun de ces différents aspects, plusieurs développements ont été réalisés et ont abouti à trois programmes majeurs : un pipeline de traitement des données brutes de séquençage, un pipeline d'annotation intégrative, et le site Web *Alvinella* qui est la vitrine principale du projet.

Dans un premier temps, le pipeline de traitement des données brutes de séquençage nous a permis de mettre en valeur assez rapidement certaines spécificitées des différentes banques d'ADNc d'*Alvinella* encore en cours de clonage ou de séquençage. Nous avons alors pu contacter les personnes impliquées afin d'en comprendre les causes et de potentiellement y remédier. Outre la technique de clonage Oligo-capping qui produit des clones de longueur moyenne inférieure à la technique CloneMiner, nous nous sommes aperçus de la faible redondance de la banque de pygidium. Nous avons ainsi pu demander et obtenir un autre lot de séquençage afin de couvrir un plus grand répertoire de gènes. Le nombre de séquences a alors été doublé pour aboutir au final à une redondance de seulement 66 % (contre 60% auparavant). Ceci suggère fortement l'extrême diversité des transcrits dans cette banque et augure un nombre élevé de nouvelles protéines dont la fonction est encore inconnue. Dans un second temps, ce pipeline a été réutilisé dans d'autres projets collaboratifs. Il a ainsi permis le traitement de plusieurs banques d'ADNc de souris construites dans le cadre de l'étude de la dégénérescence rétinienne menée par Thierry

Léveillard (projet européen EVI-GENORET). Il a également été utilisé dans le cadre de notre collaboration avec l'équipe Évolution et Génétique des Populations Marines (EGPM) de la Station Biologique de Roscoff (Sb. Roscoff) afin de traiter les ADNc d'autres espèces marines hydrothermales (*Paralvinella grasslei*, *Branchipolynoe seepensis et Bathymodiolus azoricus*). Ces organismes étant phylogénétiquement proches d'*Alvinella* mais ne vivant pas au milieu des mêmes gammes de température autour des monts hydrothermaux, des études comparatives entre ces ADNc et ceux d'*Alvinella* devraient fournir des pistes pour l'étude des mécanismes liés à l'adaptation aux conditions extrêmes rencontrées par *A. pompejana* (Didier Jollivet, EGPM, Sb. Roscoff, manuscrit en préparation).

Bien que la robustesse du pipeline d'annotation intégrative n'ait pas encore été démontrée au niveau de l'homogénéité de l'annotation lors de l'intégration des différentes sources de données (Pfam-A, numéros EC, définition fonctionnelle, …), sa capacité à annoter jusqu'à 85 % des séquences protéiques d'*Alvinella* présentant des homologies nous a permis d'obtenir une bonne couverture de la plupart des voies métaboliques et des réseaux d'interactions protéine-protéine. Compte tenu du manque d'organismes proches d'*Alvinella* dans les banques de séquences, ce résultat est très encourageant et laisse présager de bonnes performances dans le cas de l'annotation d'autres organismes inconnus. L'annotation des séquences protéiques de *P. grasslei* (Arnaud Tanguy, Sb. Roscoff) et de *B. azoricus* (Isabelle Boutet, Sb. Roscoff) est d'ores et déjà envisagée. Ce pipeline offre en outre l'avantage d'être applicable aux séquences protéiques procaryotes comme l'a montré la ré-annotation des protéines de la bactérie *M. smegmatis*.

Le site Web d'*Alvinella* a, quant à lui, considérablement facilité la diffusion des informations au sein du consortium pendant notre travail d'analyse et deviendra, nous l'espérons, une vitrine attractive et un support efficace pour la communauté scientifique intéressée par les espèces hydrothermales, les Annélides et plus généralement par l'évolution des Métazoaires.

Un point important dans cette thèse, aura été l'effort constant pour mettre en œuvre des techniques robustes lors des traitements, afin de diminuer le bruit et d'éviter la propagation d'erreurs tout au long de la chaine de traitements, sans compromettre la qualité informative des données biologiques. Ceci est illustré par les paramètres d'assemblage relativement stricts afin de ne pas fusionner des gènes paralogues en un seul et même contig, ou par la double approche de création de séquences protéiques. Cela est aussi particulièrement évident lors de l'annotation intégrative, où les séquences sont annotées dans le cadre de leur famille protéique, *via* les MACS, au lieu de se baser uniquement sur des alignements deux à deux.

La mise à profit du contexte évolutif d'une protéine par l'intermédiaire du MACS est également envisagée dans le contexte du développement d'une plate-forme de clonage *in*

silico au LBGI. Cette plate-forme a pour objectifs (i) la caractérisation des protéines cibles (fiabilité des prédictions de fonction, hydrophobicité, usage des codons,…), (ii) la détermination des bornes des domaines pour l'expression et (iii) le design des primers. La troisième partie est d'ores et déjà assurée par GScopeClonage, un module de GScope développé par R. Ripp (LBGI). Nous avons pour notre part conçu un prototype baptisé OliDA (*Oligo Design Automatization*) qui s'acquitte de la deuxième étape. A partir d'un ADNc et/ou d'une protéine, le programme génère un MACS puis un MACSIMS. Les résultats fournis par ce dernier, en particulier, les domaines PFAM-A, les structures tridimensionnelles disponibles et les conservations observées au sein de la famille, sont analysées par notre programme afin de déterminer des bornes optimisées. L'ensemble est visualisable grâce à une interface Web interactive.

Les données générées au cours de ce travail de thèse ont permis de soulever de nouvelles questions biologiques et sont le point de nucléation de plusieurs études complémentaires.

En premier lieu, ces données ont servi de base à l'aboutissement de la première publication du Consortium *Alvinella* dont le résultat majeur est sans conteste la mise en évidence de la proximité des Annélides et des Deutérostomiens, aussi bien en termes de répertoire de gènes que de séquences protéiques. Ce résultat vient confirmer la complexité de l'ancêtre commun des Métazoaires que de récentes études (pour une revue, voir (De Robertis, 2008)) commencent à révéler, modifiant profondément notre vision de l'évolution.

Ces données ont également permis la mise évidence de la richesse en résidus chargés des protéines d'*A. pompejana*, ce qui pourrait participer à la thermostabilité des protéines. Des comparaisons de séquences et de structures complémentaires ont débuté. Une première étude menée par Didier Jollivet vise à élargir le spectre des séquences comparées en incluant davantage de représentants des Lophotrochozoaires, en particulier des Annélides, afin de vérifier si le biais observé est relié à la thermostabilité d'*Alvinella* ou à un *background* phylogénétique. Le second volet porte sur la recherche des bases structurales de la thermostabilité. Cette étude, réalisée par Laurent-Philippe Albou (LBGI), s'appuie sur l'identification de résidus spécifiques à *Alvinella* au sein des MACS regroupant différentes espèces mésophiles. Ces résidus seront classés selon leur localisation et leur environnement au niveau des modèles tridimensionnels réalisés à partir de séquences protéiques d'*Alvinella*. Les modèles obtenus chez *A. pompejana* seront également comparés aux structures humaines disponibles. À ce jour, près de 500 modèles ont déjà été réalisés.

Sur un plan expérimental, les données générées au cours de cette thèse ont constitué l'élément moteur de plusieurs études en cours ou sur le point d'aboutir. Une étude du stress

oxydatif à l'aide d'une puce à ADN Agilent réalisée à partir des clones d'*Alvinella* est prévue pour la fin de l'année 2009 (Arnaud Tanguy, Sb. Roscoff). Une autre étude du stress oxydatif s'appuyant cette fois sur des expériences de protéomique à l'aide de gels bidimensionnels et de spectrométrie de masse est en cours de réalisation (Jean Mary, Sb. Roscoff).

En parallèle, le LBGI s'est impliqué dans la valorisation des ADNc dans le contexte de la génomique structurale au travers de différentes actions. Tout d'abord, un test de clonage et d'expression automatisés d'un jeu de 53 clones d'*Alvinella* couvrant un panel varié de fonctions biologiques a été réalisé par E. Perrodou (LBGI) en collaboration avec la Plate-forme de Génomique Structurale de Strasbourg (Didier Busso) afin d'en estimer la faisabilité à plus grande échelle. Ce test a abouti à l'expression de 12 protéines solubles, ce qui est tout à fait encourageant dans le cadre de protéines eucaryotes (Gräslund *et al.*, 2008). Parmi ces protéines, divers cibles ont ensuite été retenues pour des expériences de cristallographie. Des cristaux ont déjà été obtenus pour un récepteur nucléaire (Yann Brelivet, LBGI) et une structure d'une protéine impliquée dans la fixation de peptidoglycane a été résolue (Claudine Mayer, Institut Pasteur, Université Paris 7, manuscrit en préparation). Ces résultats sur des protéines isolées ont motivé le LBGI à se diriger vers la production de complexes protéiques. Sur la base des complexes identifiés lors de l'annotation des ADNc, le signalosome COP9, composé de 8 sous-unités, a été retenu. Les 8 sous-unités sont présentes et complètes dans nos banques. COP9 est principalement impliqué dans la régulation de la dégradation des protéines. Il interagit avec de nombreuses protéines notamment c-Jun (Seeger *et al.*, 1998) et p53 (Bech-Otschir *et al.*, 2001) et est impliqué dans certaines maladies dont plusieurs cancers (Adler *et al.*, 2008). Les premiers tests ont abouti à l'expression de quatre sous-unités ou partenaires de COP9 dont deux sont solubles. Des tests de co-expression vont maintenant être menés.

Par ailleurs, *A. pompejana* a été retenu comme eucaryote modèle dans le cadre du projet européen SPINE2 qui vise à déterminer la structure tridimensionnelle de complexes macromoléculaires, et tout particulièrement de complexes impliqués dans des maladies humaines. Plusieurs équipes partenaires nous ont déjà sollicités pour obtenir des clones.

Ainsi, les retombées de nos travaux s'annoncent nombreuses sur le plan expérimental et nous espérons que la publication de l'article Consortium renforcera l'intérêt de la communauté scientifique pour *A. pompejana*. Il n'en reste pas moins que nous ne disposons pas de l'ensemble des transcrits d'*Alvinella*, ce qui peut être limitant dans certaines études. Le LBGI envisage donc de « s'attaquer » au séquençage du génome de cet organisme si particulier, afin de disposer du répertoire complet des gènes et des régions régulatrices. Des études très récentes révèlent que la taille du génome d'*A. pompejana* s'élève à 675 Mb (Bonnivard *et al.*, 2009) pour 2N=32 chromosomes (Dixon *et al.*, 2006) et que la taille du

génome est stable d'un individu à l'autre, contrairement à ce qui est observé chez certains Polychètes tels que *Platynereis* (Bonnivard *et al.*, 2009). Le LBGI a obtenu la construction d'une banque BAC du génome d'*Alvinella* de 150 000 clones de 50 kB par le Genoscope. Un total de 27 000 extrémités a déjà été séquencé et doivent à présent être analysées afin de déterminer la faisabilité du séquençage génomique.

ANNEXES

15 ANNEXES

15.1 Annexe 1 – Publication 2 : SM2PH-db

Friedrich A, Garnier N, Gagnière N, Nguyen H, Albou LP, Bettler E, Deléage G, Lecompte O, Muller J, Moras D, Toursel T, Moulinier L, Poch O. SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense variations in proteins involved in human monogenic diseases. Soumis à *Human mutation* en mai 2009.

SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human monogenic diseases

Anne Friedrich[1¶], Nicolas Garnier[2¶†], Nicolas Gagnière[1], Hoan Nguyen[1], Laurent-Philippe Albou[1], Valérie Biancalana[3], Emmanuel Bettler[2], Gilbert Deléage[2], Odile Lecompte[1], Jean Muller[1†], Dino Moras[1], Jean-Louis Mandel[3,4], Thierry Toursel[5], Luc Moulinier[1], Olivier Poch[1]

[1] Département de Biologie et Génomique Structurales, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), Centre National de la Recherche Scientifique/Institut National de la Santé et de la Recherche Médicale/Université de Strasbourg, Illkirch, France

[2] Institut de Biologie et Chimie des Protéines (UMR 5086); Centre National de la Recherche Scientifique/Université de Lyon, Lyon, France

[3] Laboratoire de Diagnostic Génétique, CHRU – Faculté de Médecine et laboratoire de Génétique Médicale EA3949, Université de Strasbourg, Strasbourg, France

[4] Département de Neurobiologie et Génétique, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), Centre National de la Recherche Scientifique/Institut National de la Santé et de la Recherche Médicale/Université de Strasbourg, Illkirch, France

[5] Association Française contre les Myopathies, Evry, France

[¶] These two authors contributed equally

[†] Present address: Molecular Biophysics Unit, Indian Institute of Science, Bangalore, India (N. Garnier); Computational Biology Unit, European Molecular Biology Laboratory, Heidelberg, Germany (J. Muller)

ABSTRACT

Understanding how genetic alterations affect gene products at the molecular level represents a first step in the elucidation of the complex relationships between genotypic and phenotypic variations and is thus a major challenge in the post-genomic era.

Here, we present SM2PH-db (http://decrypthon.igbmc.fr/sm2ph), a new database designed to investigate structural and functional impacts of missense mutations and their phenotypic effects in the context of human monogenic diseases. A wealth of interconnected information related to the relationship between genotype and phenotype is provided for each of the 2 249 disease-related entry proteins (August 2009), including data retrieved from biological databases and data generated from a **S**equence-**St**ructure-**E**volution **I**nference in **Sy**stems based approach, such as multiple alignments, three-dimensional structural models and multi-dimensional (physico-chemical, functional, structural and evolutionary) characterisations of missense mutations. SM2PH-db provides a robust infrastructure associated with interactive analysis tools supporting in-depth study and interpretation of the molecular consequences of mutations, with the more long-term goal of elucidating the chain of events leading from a molecular defect to its pathology.

The entire content of SM2PH-db is regularly and automatically updated thanks to a computational grid data federation facilities provided in the context of the Decrypthon program, guaranteeing users to work with up-to-date information.

KEY WORDS

Human monogenic disease, database, mutation impact, bioinformatics, genotype-phenotype relationship, SM2PH, structural homology model, missense mutation.

INTRODUCTION

The completion of the human genome has provided a large volume of data that represents the basis for the characterization of all human genes (International Human Genome Sequencing Consortium, 2004). It has also paved the way to the systematic study of its variability (Ring, et al., 2006; The International HapMap Consortium, 2003) which is related to the evolution of the human species, required for its constant development and adaptation; but also to the emergence of human diseases. Human variability is naturally expressed through e.g. genetic shuffling during meiosis, but also by all sorts of accidental DNA changes, ranging from the substitution of a single nucleic acid to major chromosomal rearrangements.

A major source of inter-individual human variation results from the substitution of one residue by another one, called Single Nucleotide Polymorphism (SNP). SNPs are highly abundant and distributed throughout the genome (Stranger, et al., 2007) and in addition, SNPs are the mutations that are the most related to human diseases (Antonarakis, et al., 2000). SNPs can be linked to the emergence of or to the predisposition to disease and influence its severity, progression as well as its drug sensitivity. Deleterious SNPs occur in both coding and non-coding regions. In non-coding regions, the variation mostly affects gene expression by disrupting functional sites at the transcriptional level (e.g. transcription factor binding sites) (Kim, et al., 2008), or result in splicing defects (Krawczak, et al., 2007). In coding regions, and in particular protein-coding genes, deleterious SNPs are mainly non synonymous SNPs (nsSNPs), also called missense mutations, which result in the modification of the amino acid sequence of the encoded protein. nsSNPs have been linked to a wide variety of diseases, for example by affecting protein function, by reducing protein solubility or by destabilizing protein structure (Chasman and Adams, 2001). All these perturbations can be considered as the primary molecular phenotype associated with the missense mutation, having a cascade of consequences and finally leading to the emergence of a genetic disease.

The elucidation of the complex relationships between genotypic and phenotypic variations is a major challenge in the post-genomic era. Indeed, this step is crucial to a better understanding of gene functions and networks, aimed at revealing the mechanism of diseases and the development of specific therapeutic solutions.

With the current amount of information available in various biological databases, including sequences, structures, functions, pathways, interactions, variations, etc. (Galperin and Cochrane, 2009) and the subsequent development of *in silico* analysis tools, it is now possible to better understand and/or predict the correlation between a missense mutation and its associated molecular phenotypes. However, to gain further insight into the mechanisms of diseases, these molecular phenotypes have to be linked to several levels of phenotypic consequences, from the cell to the organism, ideally on the basis of ontologies. Unfortunately, the access to genotypic data with precise phenotypic descriptions represents a bottleneck in the elucidation of these complex relationships. Briefly, two main classes of databases incorporate mutations and the description of their consequences (Horaitis and Cotton, 1999): (i) central databases, such as OMIM (McKusick, 2007), HGMD (Stenson, et al., 2008) and UniProtKB/Swissprot (Yip, et al., 2008), that include mutations related to a wide range of genes with limited genotypic/phenotypic descriptions; (ii) Locus-Specific DataBases (LSDBs), which are specific to a gene (or gene family) and typically contain much more precise genotypic and phenotypic descriptions. Around 700 LSDBs are listed on the Human Genome Variation Society web site (Horaitis, et al., 2007) as being currently available on the web. However, the LSDBs have very different data formats, although some efforts are being undertaken towards their homogenization, in particular the development of generic tools to facilitate LSDB implementation, such as the Universal Mutation Database (UMD) (Beroud, et al., 2005), the Leiden Open (source) Variation Database (LOVD) (Fokkema, et al., 2005) and MUTbase (Riikonen and Vihinen, 1999).

In this context, significant efforts have been devoted to providing links between human SNPs and their molecular effects and have led to the development of novel systems combining data storage and analysis tools, both of which are indispensable for the characterisation of the consequences of SNPs (Tavtigian, et al., 2008). Systems such as coliSNP (Kono, et al., 2008), LS-SNP (Karchin, et al., 2005), MutDB (Singh, et al., 2008), SAAPdb (Hurst, et al., 2009), SNPs3D (Yue, et al., 2006) and topoSNP (Stitziel, et al., 2004) all present structural information related to the mutated protein, which has been shown to be essential (Chasman and Adams, 2001; Wang and Moult, 2001). Among these however, LS-SNP is the only one that provides computationally generated comparative protein structure models when no experimentally determined structure is available. Concerning the phenotypic effects associated to the SNPs, only MutDB provides a direct access to the observed phenotype, although this is limited to the disease name. SNPs3D and LS-SNP give access to pathogenicity prediction such as the SIFT score (Ng and Henikoff, 2003), as well as in-house developed tools, but the absence of phenotypic descriptions clearly represents an obstacle in the interpretation of the molecular consequences of a nsSNP and the related disease mechanisms. Furthermore, the post-genomic era is characterised by a torrent of biological information flooding the databases: a major requirement for any newly developed database is hence to be based on a solid computing infrastructure and to ensure its regular and automated update (Philippi and Kohler, 2006). ColiSNP and SAAPdb are the only systems that are regularly updated and consequently respond to these requirements.

To address these limitations, we have developed SM2PH-db, which stands for "from Structural Mutation to Pathology Phenotypes in Human database". SM2PH-db provides access to a wide range of up-to-date interconnected information related to the relationship between genotype and phenotype, with particular attention being focused on structural information *via* 3D structure or comparative models. A detailed multi-dimensional (physico-

chemical, functional, structural and evolutionary) characterisation of the mutations based on a **S**equence-**St**ructure-**E**volution **I**nference in **Sy**stems (SStEISy), as well as an interactive analysis platform, ensure the investigation of structural and functional impacts of missense mutations toward a better understanding of their potential molecular effects, with regard to their individual phenotypic effects.

SM2PH-db deals with proteins involved in monogenic human diseases, also known as Mendelian diseases. These disorders are particularly studied since the emergence of a pathological phenotype is linked to the alteration of a single gene. Phenotypic diversity in monogenic diseases primarily reflects mutation heterogeneity although the action of gene modifiers and epigenetic and environmental factors must also be considered (Jirtle and Skinner, 2007; Weatherall, 1998). While these so-called genetic, epigenetic and environmental backgrounds are not integrated in our system, the accurate multi-dimensional characterisation of missense mutations will provide an initial insight into the molecular basis of monogenic diseases which could open the way in the future, to a significant improvement in our understanding of the molecular and genetic basis of common, complex diseases (Antonarakis and Beckmann, 2006).

The establishment of a suitable infrastructure for SM2PH-db required the development of automated procedures to allow the integration of information from heterogeneous sources as well as regular updates. This prompted us to perform our developments in the context of the Decrypthon program (http://www.decrypthon.fr/english/), which provides access to a computational grid as well as data storage and federation facilities, *via* the Decrypthon Data Center (Nguyen, et al., 2008).

Currently (August 2009), SM2PH-db holds a total of 2,249 human proteins related to monogenic diseases and is publicly accessible online at http://decrypthon.igbmc.fr/sm2ph.

SM2PH-db CONTENT

The information associated with each disease-related protein can be classified into two main categories: data retrieved from existing databases and information produced for SStEISy (**S**equence-**St**ructure-**E**volution **I**nference in **Sy**stems) purposes.

Data retrieved from existing databases

The retrieved information is organized into two classes:

1. General information such as the gene and protein names and synonyms, as well as the known splicing variants and sequences are extracted from the UniProtKB database (UniProt Consortium, 2008). Cytogenetic band information is downloaded from the Genecards database (Safran, et al., 2003), gene ontology annotations are extracted from the GO database (Harris, et al., 2004) and the associated disease names are obtained from the Online Mendelian Inheritance in Man (OMIM) database (McKusick, 2007).

2. Missense mutations related to the disease proteins linked to their associated phenotypes are extracted from the index of "Human polymorphisms and disease mutations" from UniProtKB/Swissprot (http://www.uniprot.org/docs/humsavar.txt) and from two LSDBs (UMD-MTM1 (Biancalana *et al.*, in preparation) and the Tissue Nonspecific Alkaline Phosphatase Gene Mutations Database, http://www.sesep.uvsq.fr/database_hypo/Mutation.html). Each disease-causing missense mutation is linked to the name of its related disease, supplemented by a severity description when available. Non-pathogenic missense mutations are associated with the "polymorphism" term, in accordance with the nomenclature used in the UniProtKB database.

At the time of writing (August 2009), 27,884 missense mutations are recorded in SM2PH-db, among which 20,252 are considered as disease-causing and 7,632 as non-pathogenic.

Information produced through SStEISy approaches

In order to establish an appropriate SStEISy workbench, data related to sequence, structure and evolution are processed for each disease-related protein and missense mutations are then

characterized in this context. The information resulting from the processing of these data is listed below:

1. Evolutionary background information network: structural and functional annotations are linked to each disease-related protein thanks to MACSIMS (Multiple Alignment of Complete Sequences Information Management System, (Thompson, et al., 2006)), an information management system which combines knowledge-based methods with complementary *ab initio* sequence based predictions. MACSIMS takes advantage of the multiple alignment ontology MAO (Thompson, et al., 2005) to integrate several types of data in the framework of Multiple Alignments of Complete Sequences (MACS). Indeed, subfamily characterization based on MACS allows to highlight some discriminative aspects of sequence information, such as distinct conservation/variability patterns or domain organization between different phylogenetic levels (Lecompte, et al., 2001). For each disease-related protein, two MACS are computed with a modified version of the PipeAlign suite of programs (Plewniak, et al., 2003), which integrates several steps ranging from homolog searches in protein sequence and structure databases to the definition of the hierarchical relationships between subfamilies. The first MACS is composed of the closest eukaryotic sequences and is used to identify evolutionary constraints at particular sequence positions that are characteristic of the protein family or subfamily. The second MACS is constructed with a sampling strategy to significantly reduce the number of aligned sequences, while at the same time maintaining the potential structural and functional information in the alignment (Friedrich, et al., 2007).

2. Structural information network: the availability of a 3D structure or model of the protein is essential to gain insight into the structural impact of a mutation. The best source of protein structural information is the PDB (Berman, et al., 2000), which stores almost all the experimentally resolved crystallographic structures. However, only 574 proteins out of our 2,249 human disease-related proteins are currently represented in the PDB. To enhance the

available experimental data, 3D models of the wild type proteins are automatically constructed by homology, using Modeller (Eswar, et al., 2008). The models are built by inferring the structure of a protein (the target) from the structure of another putatively homologous protein (i.e. a sequence sharing at least 30% identity) solved by experimental methods (the template). The selection of a suitable template is based on BLAST similarity searches in the PDB: templates covering the full protein are preferred, but shorter domains can be modelled when a full template is lacking, in which case several 3D models may be associated with a single protein. The pairwise alignment of the target and template proteins is extracted from the sampled MACS and is used as input to Modeller. Five homology models are constructed and the one with the best normalized DOPE score (Eramian, et al., 2008) is integrated in SM2PH-db.

Currently, 1,551 structures and 3D models related to wild type proteins are stored in the database, concerning 1,370 different proteins.

3. Missense mutation information network: missense mutations are characterized according to 31 parameters (Supp. Table S1), which can be classified into three main levels of information:

- physico-chemical changes induced by the amino acid substitution: modifications in size, charge, polarity and hydrophobicity are independently described (Taylor, 1986). A global score reflecting the degree of modification induced by the substitution is also assigned. This score corresponds to the residues inter-distance (Supp. Figure S1) based on a vector representation of the amino acids (French and Robson, 1983) and normalized on a scale from 0 to 100, with larger distances implying less conservative substitutions;

- functional and structural features related to the substituted position: these features include MACSIMS annotations, descriptions of the 3D context (e.g. residue relative accessibility, in contact with an annotated site, …) and a conservation ranking. This

ranking is calculated using an in-house developed method based on a 3-step process: (i) two independent conservation scores are computed for each column of the sampled MACS, namely the free energy score (Lockless and Ranganathan, 1999) and a score based on the two-dimensional vector representation of the amino acids (Supp Figure S2); (ii) these scores are classified with a Dirichlet mixture algorithm (Sjolander, et al., 1996) to define "groups of conservation" (iii) the groups are then ranked. This process is reiterated for each MACS sub-family. Thus, two global conservation classes (rank 1 and rank 2) and one sub-family conservation class per sub-family are finally defined.

- structural modifications induced by the amino acid substitution, based on the mutant 3D models. These are automatically constructed with Modeller for missense mutations that can be mapped onto a wild type 3D model sharing more than 50% identity with its PDB template: at the time of writing (August 2009), 9,435 3D mutant models are available in SM2PH-db (7,863 for disease-causing mutants and 1,572 for mutants considered as non-pathogenic). The change in protein relative stability upon single-site mutation ($\Delta\Delta G$ value, (Casadio, et al., 1995)) is predicted with I-Mutant2.0 (Capriotti, et al., 2005): a positive $\Delta\Delta G$ value implies a protein stability increase, whereas a negative $\Delta\Delta G$ value suggests a destabilizing mutation. The wild type residue contacts, computed with the CSU software (Sobolev, et al., 1999) are compared to the mutated residue ones in the 3D model and any change induced by the substitution is stored.

SM2PH-db DATA GENERATION AND UPDATE

A fully automated workflow has been developed for the generation and regular update of the entire database contents (Figure 1), which can be divided into five main processes:

1. Protein entry list update, based on the OMIM database. The list of gene entries with disease-causing mutations is obtained as described in (Amberger, et al., 2009). Based on this list, a file containing all the selected human entry sequences in FASTA format is created.

2. Mining of heterogeneous databases (OMIM, UniProtKB, Genecards, GO, several LSDBs) through the Decrypthon Data Center, based on the protein entry list. The entire set of retrieved data associated with each disease-related protein is constituted during this process.

3. Construction of the SStEISy workbench. First, structural templates are determined based on similarity searches in the PDB and two MACS are constructed and annotated, including the template sequences. Second, a new wild type 3D model is generated if the structural template differs from the one used in the previous version of SM2PH-db or if a new 3D template is available.

4. Multi-dimensional characterization of the mutants. This process initiates with the generation of mutant 3D models not available in the previous version of SM2PH-db (e.g. for novel proteins, newly generated wild type 3D models or new mutations). Physico-chemical changes and structural modifications induced by the substitution as well as functional and structural features related to the mutated position are then listed.

5. Finally, the entire content of the SM2PH-db is upgraded, by integrating all this information.

Particular attention has been paid to the automation and optimisation of this workflow. We have optimized communication and synchronization between the processes by analysing the data dependencies. In order to speed-up the entire workflow, processes 2 and 3 are launched in parallel. Moreover, as SM2PH-db includes about 2,250 entry proteins, the update of all the alignments and their annotation represented a potential bottleneck in terms of computing time and memory. To overcome these limitations, we have implemented PipeAlign and MACSIMS on the Decrypthon computation grid and the required databases have been integrated in the Decrypthon Data Center. The construction and annotation of all SM2PH-db alignments takes around two days using the Decrypthon infrastructure which is constituted of

120 computational nodes, while the same tasks took about one month to run on our internal server (four processor SUN Enterprise V40z).

Currently, the complete SM2PH-db update procedure is launched every two months, guaranteeing that the user is working with up-to-date information.

SOFTWARE IMPLEMENTATION

SM2PH-db is implemented in a relational database management system (PostgreSQL), has a schema with 26 tables and runs on a Linux server. It takes advantage of a molecular model database management system called Modeome3D (Garnier et al., in preparation). The generation of the web interface is programmed in Python and uses the AJAX (Asynchronous JavaScript and XML) methodology for dynamic updating of the pages.

SM2PH-db WEB INTERFACE

Database search

SM2PH-db can be queried using textual search forms (Figure 2A), based on a combination of keywords (e.g. protein name, disease name) or *via* a BLAST search (Altschul, et al., 1997) of the database entries.

The search results consist of a summary of the data related to each protein that matches the required criteria (Figure 2B). *Protein details* pages, that assemble all the retrieved data, can be accessed by clicking on the protein identifiers. When a 3D model is available, a diagram that schematizes the modelled region(s) of the sequence is presented and an interactive analysis interface can be accessed that allows the visualisation and manipulation of the processed information.

Protein details page

The *protein details* page can be considered to be a protein "identity card". It displays data retrieved from existing databases, links to external databases are provided and a visual summary of the processed information. The retrieved data are divided into 3 main sections:

General Information, *Cross-references* and *Mutations* (Figure 3A). The latter includes a list of missense mutations associated with a short description of their phenotypic consequences.

The processed information summary mainly concerns the functional domain annotations assigned to the protein by MACSIMS, details of the 3D model if one has been built and its secondary structure composition. The schematic view reveals all these features in a simple, linear representation (Figure 3B).

Interactive graphical interface

The SM2PH-db interactive interface (Figure 4A) is based on the MAGOS web server (Garnier, et al., 2006). It has been upgraded to meet the SStEISy requirements and enhanced to allow integrative missense mutation analyses. This graphical interface is divided into three frames, each related to a produced information network:

- Frame 1: the right-hand part of the interface shows the annotated MACS. The sequence of the human disease-related protein is always the first in the alignment and the modelled region is underlined. By default, one sequence of each sub-family of the MACS is displayed; sub-family sequences can be expanded with the "*Uncollapse all*" button. Functional and structural annotations such as active sites, PFAM domains, residues conservation during evolution, transmembrane regions, etc. are available through the *Features* menu: these are mapped on the MACS when selected.

- Frame 2: the upper part of the left-hand side of the interface displays the 3D model in a Jmol environment (http://jmol.sourceforge.net/) and allows its in-depth exploration. The *Display options* section is dedicated to a basic manipulation of the 3D model by the intermediary of predefined rendering types (e.g. cartoon, spacefill, wireframe) and residue coloring (e.g according to polarity, accessibility) while the *Console* section allows a high-level 3D model manipulation using the Jmol/Rasmol command line. The visualization of the residues positioned in a structural environment close to a residue of interest (i.e. within a distance

defined by the user) can be performed with the *Around residues* option found in the *Link Parameters* section (Figure 4B).

- Frame 3: the *Mutations* section, in the lower part of the left-hand side of the page, is devoted to the missense mutation informational network. The missense mutations mapped on the 3D model can be accessed *via* the top menu of this section. A table is shown under mutation selection, in which the associated phenotypes and 2 links are furnished. The first link provides access to the mutation characterisation page (Figure 4C), that includes the 31 parameters describing mutations (i.e. physico-chemical changes induced by the substitution, information related to the substituted position such as conservation, functional and structural features). The second link allows the replacement of the wild type 3D model by the mutant model in the Jmol window.

While these frames represent suitable analysis tools on their own, the main advantage of SM2PH-db graphical interface relies on their interconnection which allows high-level integrative missense mutation analysis to be performed. Firstly, the disease-related protein sequence in the MACS is connected to its 3D model: the annotations assigned to the underlined part of the sequence (i.e. the modelled part) are simultaneously displayed on the 3D model using the same colour code. The selection of a residue in the structure will also highlight this residue within the sequence and *vice versa*. Furthermore, the mutation frame is connected to the other two frames: the selection of a missense mutation leads to the simultaneously mapping of the wild type residue within the wild type 3D model and in the MACS. This allows the study of the mutated position in a SStEISy environment.

Thanks to this graphical interface, in addition to providing insights into the disease mechanisms, SM2PH-db represents an ideal workbench for an in-depth analysis of novel user's mutations through an interactive approach. Indeed, the user can study the substituted position in terms of structural/functional features, ask for the construction of a 3D model for

his mutant of interest and explore this model through the Jmol interface (*Generate* button in the *Mutations* section). Moreover, in order to support specific user investigations, the complete mutation characterization page, summarizing the main information related to this substitution and its position can be generated "on the fly".

SM2PH-db statistics page

General statistics related to the mutations stored in SM2PH-db can be accessed *via* the *Statistics* link, in the main menu. These statistics concern 10 parameters (out of the 31 computed) that characterize either the mutation substitutions or their substituted positions. For each of these parameters, 2 graphs representing the distribution of the associated values are provided: one for disease-causing mutations, the other for non-pathogenic mutations.

Each graph represents one single parameter and cannot be used in an independent manner to define a threshold to discriminate between disease-causing and non-pathogenic mutations. However, coupled with human expertise, these statistics could facilitate the interpretation of the molecular consequences of a given mutation. Therefore, these graphs can be viewed directly in a pop-up window from the mutation characterisation pages, by clicking on the statistics icon located in the lower right corner of the concerned parameter cell.

DISCUSSION

SM2PH-db has been organized to give the user an easy access to a wealth of information relevant to the study of genotype/phenotype relationships in the context of human monogenic diseases. Our database regroups up-to-date heterogeneous interconnected information, ranging from sequence to structure and including evolutionary and functional features, thus providing a high-level SStEISy workbench. This workbench, in combination with the interactive analysis platform, allows the investigation of known missense mutation molecular impacts with regard to their phenotypic effects as well as the in-depth exploration of novel missense mutations to infer their potential molecular and phenotypic effects.

At time of writing (August 2009), a total of 27,884 missense mutations are recorded in the database, among which 20,252 (73%) are disease-causing and 7,632 (27%) are considered as non-pathogenic. 9,435 3D mutant models have been created and are available *via* the *Interactive graphical interface*, of which 7,863 are disease-causing and 1,572 non-pathogenic.

SM2PH-db: insight into mutation effects

The molecular consequences of missense mutations are related to the functional and structural contexts of the affected position, as well as to the physico-chemical characteristics of the substitution (Saunders and Baker, 2002; Terp, et al., 2002). All these types of information are represented in SM2PH-db for the stored missense mutations.

The integration capabilities of SM2PH-db can be illustrated by the analysis of the molecular consequences of a selected missense mutation. Here, we consider the p.Arg421Gln missense mutation that affects Myotubularin which is associated with a severe deleterious phenotype.

Myotubularin can be searched by querying the *Protein name* field in the textual search form. This search indicates that 4 proteins whose names contain the term myotubularin are stored in the database, including 3 myotubularin-related proteins (Figure 2). It should be noted that a 3D model of Myotubularin has been constructed with a template that shares 69% of identity, suggesting that this model is of good quality (Figure 2B).

Before any further analysis, the protein details page should be visualized (Figure 3). Here, the domain and active site localisations are of particular interest when trying to interpret mutation consequences. These can be viewed in the *Macsims infos* section, as well as in the linear schematic view of this entry, where structural and evolutionary information are also provided.

The interactive graphical interface can then be accessed via the *View in Jmol* link. The Arginine in position 421 can be mapped onto the 3D structure under selection in the *Mutations* section (Figure 4A). A quick visual inspection shows that this residue is part of an alpha-helix and seems to be buried. The mutation characterisation page (Figure 4C) can then

be accessed *via* the link provided in the table in the *Mutations* section. The physico-chemical modifications linked to this substitution are not drastic: the modified score is 23 and the main change to be noted is the decrease of the residue size. The substituted position is located in the *Myotubularin phosphatase* domain and is well conserved during evolution: arginine represents more than 95% of this column in the alignment and has been classified as a rank 1 conserved position. In this example, an important feature can be observed, concerning the wild type and mutated residue contacts: a residue that is in contact with our substituted residue is in direct contact with the Myotubularin active site, at position 375. This contact, although predicted as maintained, can obviously be modified by the substitution by a smaller residue. As a consequence, one might hypothesize that the p.Arg421Gln substitution destabilizes the active site pocket, which could explain the severe phenotype associated.

In conclusion, this example shows that the SM2PH-db infrastructure is suitable for in-depth investigations of mutations and can support the formulation of hypotheses related to the molecular consequences of known or newly discovered mutations.

Phenotypic information

The availability of phenotypic information is central to SM2PH-db, since this information should help the scientist in the understanding of the disease pathogenesis. The retrieval of missense mutations from the UniProtKB/Swissprot database ensures the integration of phenotypic data for almost all the disease-related proteins, but our goal is to provide access to more precise information. Our system has consequently been organized to store information from LSDBs, excluding any patient data, although their integration has to be envisaged in a step-by-step manner because of the highly variable data formats. In the future, this task should become simpler, thanks to current standardization efforts in the context of the Gen2Phen European initiative (http://www.gen2phen.org) or the Human Variome Project (Ring, et al., 2006), devised to collect and curate all genetic variation, its phenotypes and associated

diseases, that will probably speed up the homogenization process. However, apart from the format aspects, another limiting factor for the integration of LSDB data is the reluctance of some clinicians to provide access to allelic variations and related information until after publication. To address this problem, we provide restricted access to private data on demand, based on the use of personal logins and passwords.

The Decrypthon Data Center is already capable of managing the UMD and LOVD formats and SM2PH-db stores data mined for example, from the UMD-MTM1 database. This database is dedicated to the Myotubularin protein (UniProt/Swiss-Prot:Q13496), involved in Myotubular Myopathy (MIM:310400) and includes data related to 68 different missense variants, of which 61 have a corresponding three-level degree of severity.

We are currently contacting LSDB curators, seeking their consent to access and possibly diffuse their non-confidential data, in order to provide users with precise genotypic/phenotypic information for a large number of disease-related proteins.

Conclusions and perspectives

SM2PH-db represents an initial effort towards an effective and automated integration of mutation and phenotypic data involved in human monogenic diseases. The sustainability of the database is guaranteed by the robust Decrypthon infrastructure on which it is based. Moreover, this latter also ensures that the biological data in SM2PH-db is always up-to-date.

In its current state, SM2PH-db and the numerous features offered by this novel system facilitates efficient study and interpretation of the molecular consequences of a given mutation.

In the future, a tool dedicated to the automated discrimination of disease-causing from non-pathogenic mutations will be integrated in our server. These predictions will be presented as propositions and complemented with all individual characterizing parameters, in order to give the user access to all the information required to judge the prediction accuracy. This complete

system will hopefully contribute to the elucidation of the chain of events leading from a molecular defect to the pathology.

To achieve this, we intend to further enhance the available data by including, not only more detailed genotypic and phenotypic information, but also interactomic data such as functional and physical interactions mined from the STRING (Jensen, et al., 2009) and IntAct (Kerrien, et al., 2007) databases, as well as structural surface topology descriptions and interacting interface predictions(Albou, et al., 2009). We also plan to create *variant description* pages "on the fly" for missense variants of interest to the user. This will allow us to provide a simple access to customized information, which will aid the interpretation of the potential molecular effects of the mutation on the user's protein.

ACKNOWLEDGEMENTS

The authors are grateful to Raymond Ripp for his assistance during this work and Julie Thompson for a critical reading of the manuscript. This work was funded by the French Association against Myopathy *via* the Decrypthon program (AFM12727-13836), EVI-GENORET LSHG-CT-2005-512036, EvolHHuPro ANR-07-BLAN-0054-02 as well as institute funds from the Institut National de la Santé et de la Recherche Médicale, the Centre National de la Recherche Scientifique and the Université de Strasbourg.

References

Albou LP, Schwarz B, Poch O, Wurtz JM, Moras D. 2009. Defining and characterizing protein surface using alpha shapes. Proteins 76(1):1-12.

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.

Amberger J, Bocchini CA, Scott AF, Hamosh A. 2009. McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res 37(Database issue):D793-6.

Antonarakis SE, Beckmann JS. 2006. Mendelian disorders deserve more attention. Nat Rev Genet 7(4):277-82.

Antonarakis SE, Krawczak M, Cooper DN. 2000. Disease-causing mutations in the human genome. Eur J Pediatr 159 Suppl 3:S173-8.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE. 2000. The Protein Data Bank. Nucleic Acids Res 28(1):235-42.

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.

Capriotti E, Fariselli P, Casadio R. 2005. I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res 33(Web Server issue):W306-10.

Casadio R, Compiani M, Fariselli P, Vivarelli F. 1995. Predicting free energy contributions to the conformational stability of folded proteins from the residue sequence with radial basis function networks. Proc Int Conf Intell Syst Mol Biol 3:81-8.

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.

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.

Eramian D, Eswar N, Shen MY, Sali A. 2008. How well can the accuracy of comparative protein structure models be predicted? Protein Sci 17(11):1881-93.

Eswar N, Eramian D, Webb B, Shen MY, Sali A. 2008. Protein structure modeling with MODELLER. Methods Mol Biol 426:145-59.

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.

French S, Robson B. 1983. What is a conservative substitution? Journal of molecular evolution 19(2):171-5.

Friedrich A, Ripp R, Garnier N, Bettler E, Deleage G, Poch O, Moulinier L. 2007. Blast sampling for structural and functional analyses. BMC Bioinformatics 8:62.

Galperin MY, Cochrane GR. 2009. Nucleic Acids Research annual Database Issue and the NAR online Molecular Biology Database Collection in 2009. Nucleic Acids Res 37(Database issue):D1-4.

Garnier N, Friedrich A, Bolze R, Bettler E, Moulinier L, Geourjon C, Thompson JD, Deleage G, Poch O. 2006. MAGOS: multiple alignment and modelling server. Bioinformatics 22(17):2164-5.

Harris MA, Clark J, Ireland A, Lomax J, Ashburner M, Foulger R, Eilbeck K, Lewis S, Marshall B, Mungall C and others. 2004. The Gene Ontology (GO) database and informatics resource. Nucleic Acids Res 32(Database issue):D258-61.

Horaitis O, Cotton RG. 1999. 6th International HUGO Mutation Database Meeting, March 27, 1999, Brisbane, Australia. Hum Mutat 14(3):183-5.

Horaitis O, Talbot CC, Jr., Phommarinh M, Phillips KM, Cotton RG. 2007. A database of locus-specific databases. Nat Genet 39(4):425.

Hurst JM, McMillan LE, Porter CT, Allen J, Fakorede A, Martin AC. 2009. The SAAPdb web resource: A large-scale structural analysis of mutant proteins. Hum Mutat.

International Human Genome Sequencing Consortium. 2004. Finishing the euchromatic sequence of the human genome. Nature 431(7011):931-45.

Jensen LJ, Kuhn M, Stark M, Chaffron S, Creevey C, Muller J, Doerks T, Julien P, Roth A, Simonovic M and others. 2009. STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res 37(Database issue):D412-6.

Jirtle RL, Skinner MK. 2007. Environmental epigenomics and disease susceptibility. Nat Rev Genet 8(4):253-62.

Karchin R, Diekhans M, Kelly L, Thomas DJ, Pieper U, Eswar N, Haussler D, Sali A. 2005. LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. Bioinformatics 21(12):2814-20.

Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, Dimmer E, Feuermann M, Friedrichsen A, Huntley R and others. 2007. IntAct--open source resource for molecular interaction data. Nucleic Acids Res 35(Database issue):D561-5.

Kim BC, Kim WY, Park D, Chung WH, Shin KS, Bhak J. 2008. SNP@Promoter: a database of human SNPs (single nucleotide polymorphisms) within the putative promoter regions. BMC Bioinformatics 9 Suppl 1:S2.

Kono H, Yuasa T, Nishiue S, Yura K. 2008. coliSNP database server mapping nsSNPs on protein structures. Nucleic Acids Res 36(Database issue):D409-13.

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.

Lecompte O, Thompson JD, Plewniak F, Thierry J, Poch O. 2001. Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene 270(1-2):17-30.

McKusick VA. 2007. Mendelian Inheritance in Man and its online version, OMIM. Am J Hum Genet 80(4):588-604.

Ng PC, Henikoff S. 2003. SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res 31(13):3812-4.

Nguyen H, Friedrich A, Berthommier G, Poidevin L, Moulinier L, Ripp R, Poch O. Introduction du nouveau Centre de Données Biomedicales Décrypthon; 2008; CORIA, Tregastel.

Philippi S, Kohler J. 2006. Addressing the problems with life-science databases for traditional uses and systems biology. Nat Rev Genet 7(6):482-8.

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.

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.

Safran M, Chalifa-Caspi V, Shmueli O, Olender T, Lapidot M, Rosen N, Shmoish M, Peter Y, Glusman G, Feldmesser E and others. 2003. Human Gene-Centric Databases at the Weizmann Institute of Science: GeneCards, UDB, CroW 21 and HORDE. Nucleic Acids Res 31(1):142-6.

Saunders CT, Baker D. 2002. Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol 322(4):891-901.

Singh A, Olowoyeye A, Baenziger PH, Dantzer J, Kann MG, Radivojac P, Heiland R, Mooney SD. 2008. MutDB: update on development of tools for the biochemical analysis of genetic variation. Nucleic Acids Res 36(Database issue):D815-9.

Sobolev V, Sorokine A, Prilusky J, Abola EE, Edelman M. 1999. Automated analysis of interatomic contacts in proteins. Bioinformatics 15(4):327-32.

Stenson PD, Ball E, Howells K, Phillips A, Mort M, Cooper DN. 2008. Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet 45(2):124-6.

Stitziel NO, Binkowski TA, Tseng YY, Kasif S, Liang J. 2004. topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. Nucleic Acids Res 32(Database issue):D520-2.

Stranger BE, Forrest MS, Dunning M, Ingle CE, Beazley C, Thorne N, Redon R, Bird CP, de Grassi A, Lee C and others. 2007. Relative impact of nucleotide and copy number variation on gene expression phenotypes. Science 315(5813):848-53.

Tavtigian SV, Greenblatt MS, Lesueur F, Byrnes GB. 2008. In silico analysis of missense substitutions using sequence-alignment based methods. Hum Mutat 29(11):1327-36.

Taylor WR. 1986. The classification of amino acid conservation. J Theor Biol 119(2):205-18.

Terp BN, Cooper DN, Christensen IT, Jorgensen FS, Bross P, Gregersen N, Krawczak M. 2002. Assessing the relative importance of the biophysical properties of amino acid substitutions associated with human genetic disease. Hum Mutat 20(2):98-109.

The International HapMap Consortium. 2003. The International HapMap Project. Nature 426(6968):789-96.

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.

UniProt Consortium. 2008. The universal protein resource (UniProt). Nucleic Acids Res 36(Database issue):D190-5.

Wang Z, Moult J. 2001. SNPs, protein structure, and disease. Hum Mutat 17(4):263-70.

Weatherall DJ. 1998. The phenotypic diversity of monogenic disease: lessons from the thalassemias. Harvey Lect 94:1-20.

Yip YL, Famiglietti M, Gos A, Duek PD, David FP, Gateau A, Bairoch A. 2008. Annotating single amino acid polymorphisms in the UniProt/Swiss-Prot knowledgebase. Hum Mutat 29(3):361-6.

Yue P, Melamud E, Moult J. 2006. SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics 7:166.

Figure legends

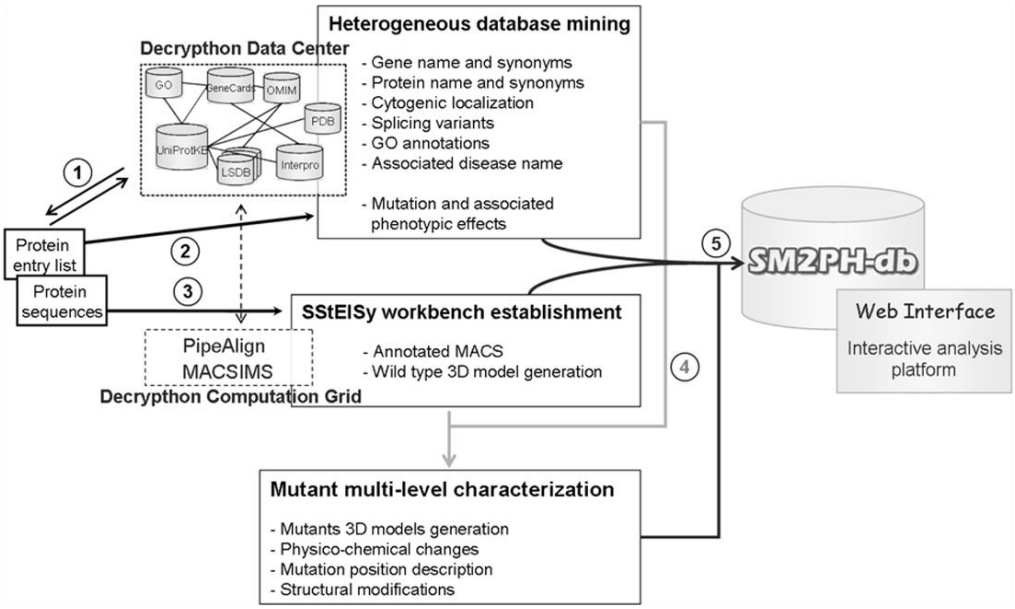


Figure 1: Schematic representation of SM2PH-db automated workflow for data generation and integration: (1) protein entry list update; (2) heterogeneous data mining *via* the Decrypthon Data Center; (3) SStEISy workbench construction; (4) mutant multi-dimensional characterization; (5) SM2PH-db data integration.

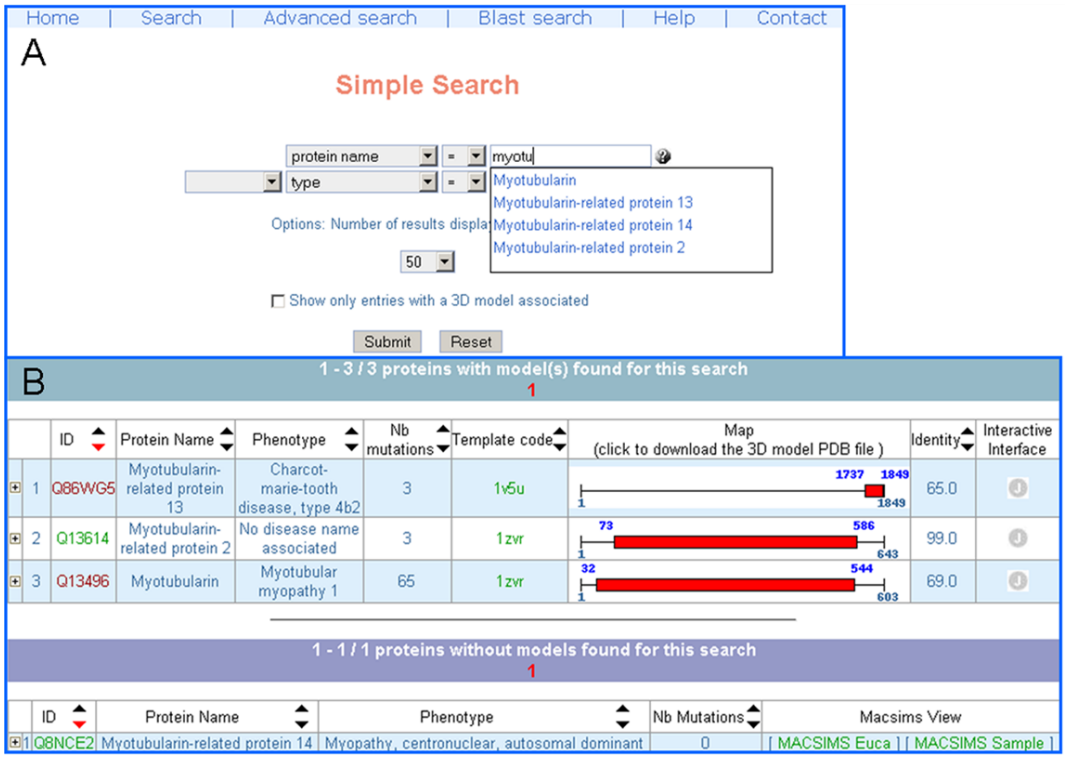


Figure 2: SM2PH-db search interface. (A) The textual search form allows querying of the database with a combination of keywords. A list of keywords matching the first letters entered by the user is automatically proposed to facilitate the search. (B) The result page displays a table containing the protein identifier, its name, the associated disease name and information related to the 3D model if one has been constructed. Links to the entry protein details and to the interactive analysis interface are provided.

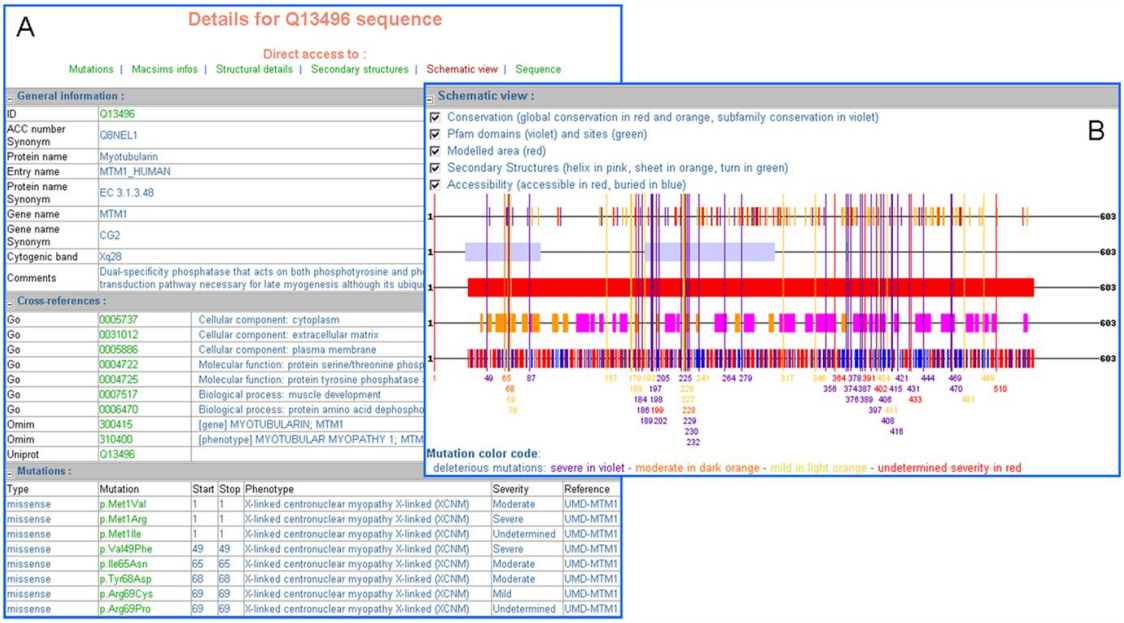


Figure 3: Screenshots of the Myotubularin details page. (A) The retrieved data are displayed, divided into several sections. (B) The schematic view associated with Myotubularin summarizes, in a linear representation, the main protein functional and structural features as well as the mutant positions.

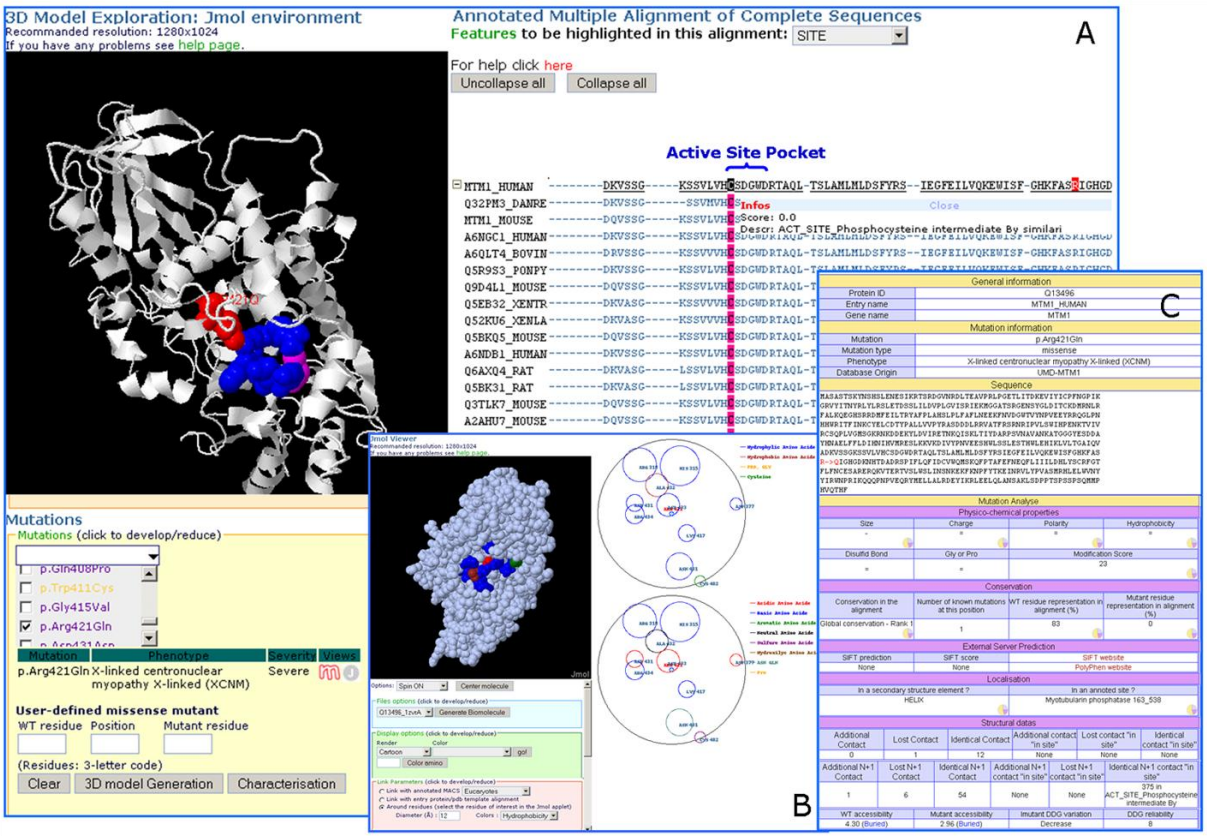


Figure 4: SM2PH-db interactive graphical interface for Myotubularin. (A) The dynamic web page is divided into three interconnected frames: the right-hand frame gives access to the annotated MACS interconnected with the 3D model/structure displayed in the upper left frame. All the features mined and predicted by MACSIMS can be displayed both on the query sequence in the context of the MACS and on its structural representation. The lower left frame shows the missense mutations that can be positioned and analysed within the structural and functional context of the protein. (B) Residues surrounding a position of interest can be highlighted with the *Around residues* option. The diameter of investigation can be modified by the user. (C) The p.Arg421Gln mutation characterisation page, which provides descriptions concerning the modifications induced by the substitution as well as information related to the conservation of the mutated residue, its position relative to functional features and within the 3D model, etc.

15.2 Annexe 2 – Publication 3 : RETINOBASE

Kalathur RK, Gagnière N, Berthommier G, Poidevin L, Raffelsberger W, Ripp R, Léveillard T, Poch O. RETINOBASE: a web database, data mining and analysis platform for gene expression data on retina. *BMC Genomics*, 2008 May 5;9:208

BMC Genomics

BioMed Central

Database

RETINOBASE: a web database, data mining and analysis platform for gene expression data on retina

Ravi Kiran Reddy Kalathur[1], Nicolas Gagniere[1], Guillaume Berthommier[1], Laetitia Poidevin[1], Wolfgang Raffelsberger[1], Raymond Ripp[1], Thierry Léveillard[2] and Olivier Poch*[1]

Address: [1]Laboratoire de Bioiformatique et de Genomique Integratives, Institut de Génétique et de Biologie Moléculaire et Céllulaire, CNRS/INSERM/ULP, BP 163, 67404 Illkirch Cedex, France and [2]Inserm U592 Universite Pierre et Marie Curie, Laboratoire de Physiopathologie Céllulaire et Moléculaire de la Retine, Hopital Saint-Antoine, Paris, France

Email: Ravi Kiran Reddy Kalathur - ravi@igbmc.u-strasbg.fr; Nicolas Gagniere - gagniere@igbmc.u-strasbg.fr; Guillaume Berthommier - berthomg@igbmc.u-strasbg.fr; Laetitia Poidevin - Laetitia.Poidevin@igbmc.u-strasbg.fr; Wolfgang Raffelsberger - wraff@igbmc.u-strasbg.fr; Raymond Ripp - ripp@igbmc.u-strasbg.fr; Thierry Léveillard - Thierry.Leveillard@st-antoine.inserm.fr; Olivier Poch* - poch@titus.u-strasbg.fr

* Corresponding author

Published: 5 May 2008

BMC Genomics 2008, **9**:208 doi:10.1186/1471-2164-9-208

Received: 30 October 2007
Accepted: 5 May 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/208

Abstract

Background: The retina is a multi-layered sensory tissue that lines the back of the eye and acts at the interface of input light and visual perception. Its main function is to capture photons and convert them into electrical impulses that travel along the optic nerve to the brain where they are turned into images. It consists of neurons, nourishing blood vessels and different cell types, of which neural cells predominate. Defects in any of these cells can lead to a variety of retinal diseases, including age-related macular degeneration, retinitis pigmentosa, Leber congenital amaurosis and glaucoma. Recent progress in genomics and microarray technology provides extensive opportunities to examine alterations in retinal gene expression profiles during development and diseases. However, there is no specific database that deals with retinal gene expression profiling. In this context we have built RETINOBASE, a dedicated microarray database for retina.

Description: RETINOBASE is a microarray relational database, analysis and visualization system that allows simple yet powerful queries to retrieve information about gene expression in retina. It provides access to gene expression meta-data and offers significant insights into gene networks in retina, resulting in better hypothesis framing for biological problems that can subsequently be tested in the laboratory. Public and proprietary data are automatically analyzed with 3 distinct methods, RMA, dChip and MAS5, then clustered using 2 different K-means and 1 mixture models method. Thus, RETINOBASE provides a framework to compare these methods and to optimize the retinal data analysis. RETINOBASE has three different modules, "Gene Information", "Raw Data System Analysis" and "Fold change system Analysis" that are interconnected in a relational schema, allowing efficient retrieval and cross comparison of data. Currently, RETINOBASE contains datasets from 28 different microarray experiments performed in 5 different model systems: drosophila, zebrafish, rat, mouse and human. The database is supported by a platform that is designed to easily integrate new functionalities and is also frequently updated.

Conclusion: The results obtained from various biological scenarios can be visualized, compared and downloaded. The results of a case study are presented that highlight the utility of RETINOBASE. Overall, RETINOBASE provides efficient access to the global expression profiling of retinal genes from different organisms under various conditions.

Background

The retina is a thin and highly structured layer of neuronal cells that lines the back of eye. Its main function is to convert light energy into an interpretable signal for cortical cells in the brain. The retina has two components – an inner neurosensory retina and an outer retinal pigment epithelium (RPE), which together form the structural and functional basis for visual perception.

The retina consists of several cell types, of which neural cells predominate. Photoreceptors, bipolar and ganglion cells are three principal neuron cell types whose activity is modulated by other groups of cells, such as horizontal and amacrine cells [1]. Defects in any of the above-mentioned cell types can lead to a variety of retinal diseases, including age-related macular degeneration (AMD), retinitis pigmentosa (RP), Leber congenital amaurosis (LCA) and glaucoma. These diseases may cause partial visual loss or complete blindness, depending on the severity.

The recent progress in genomic approaches has now led to an increase in the number of transgenic and knockout animal models that can be used to investigate the role of specific genes in retinal function and related disorders in humans, e.g., *rd1* is a mouse model for RP [2], *Nr2e3* for the Human Enhanced S-cone syndrome (ESCS) [3], *Rds* for macular dystrophy and *RPE65*$^{-/-}$ for LCA [4]. Experimental information from the above mentioned models, combined with high-throughput technologies, has led to an increase in the number of experiments related to retinal gene expression.

The recent development of high-throughput technologies has resulted in an enormous volume of gene expression data. General repositories such as GEO [5] and ArrayExpress [6] operate as central data distribution centres encompassing gene expression data from different organisms and from various conditions. In contrast, resources like CGED [7], SIEGE [8] and GeneAtlas [9] are specialized databases that address specific problems; CGED concentrates on gene expression in various human cancer tissues, SIEGE focuses on epithelial gene expression changes induced by smoking in humans and Gene Atlas provides the expression profiles of genes in various mouse and human tissues.

In order to address specific issues related to retina and to meet the needs of retinal biologists in their analysis of gene expression data, we have developed RETINOBASE, a microarray gene expression database for retina. RETINOBASE combines simplified querying, analysis and data visualization options, plus specifically developed meta analysis tools. The integration of gene expression data from various development stages of wild type retina and from diverse conditions and genetic backgrounds will hopefully, not only increase our understanding of the physiological mechanisms involved in normal retinal tissue, but also facilitate studies of gene expression patterns under diverse conditions. Furthermore, RETINOBASE provides a platform for the comparison of different analysis scenarios based on various normalization methods, such as RMA [10], dChip [11], MAS5 [12], and clustering methods, such as the K-means [13] and mixture models methods [14].

Construction and content

RETINOBASE uses open-source tools. The website is powered by an Apache web server, PHP and Javascript for dynamic web pages and a PostgreSQL object-relational open source database management system (DBMS) as the back end to store data. The RETINOBASE database schema has been developed using the same philosophy as that used to design BASE [15], with enhancements to accommodate data from different platforms and also complies to the Minimum Information About Microarray Experiment (MIAME) standard [16]. It is based on a well-designed relational schema where "realexp" acts as a central table linking expression data with an experiment, sample and array type. This kind of schema helps the system to manage data efficiently, and increases retrieval speed.

RETINOBASE is designed to store gene expression profiles from microarray experiments. We downloaded all publicly available retina-related expression profiles from Gene Expression Omnibus (GEO) yielding 21 experiments [17-32], GEO datasets (GSE 1816, 4756, 1835, 3791, 2868). In addition, 8 proprietary experiments have been incorporated that can be accessed with permission from the owner of the experiment. These experiments were performed under different conditions, including knockout models, treatments and time series experiments performed on different organisms such as drosophila, zebra fish, rat, mice and human. All experiments have complete data, except for one experiment [19] that has partial data at the level of fold change, due to the unavailability of raw data (.CEL) or signal intensity data. Currently, RETINOBASE contains approximately 27 million gene expression values resulting from 509 hybridizations. In future releases of the database, we plan to include data from other studies associated with retina, including the SAGE [33], datasets from Diehn and coworkers [34] who used cDNA array to study human eye tissues, and/or datasets from Blackshaw and coworkers [35] who used SAGE to study mouse retinal development.

Gene information

In RETINOBASE, the gene annotation information obtained from Affymetrix [36] is linked to information about genes and loci causing inherited retinal diseases,

obtained from the Retinal information network (RETNET) [37]. RETINOBASE also provides information obtained from literature about expression of approximately 200 retinal genes specific to certain types of cell, such as photoreceptors, Muller cells or retinal sphere cells.

Data information

Raw data was obtained in two different formats, either as .CEL files (20 experiments) or at the level of signal intensities (8 experiments). Data obtained at the level of .CEL files are first analysed with three different normalization programs – RMA [10], dChip [11] and MAS5 [12] and then processed using the R statistical package [38] and Bioconductor [39]; after preprocessing, the resulting background-corrected and normalized signal intensities are automatically uploaded to RETINOBASE using SQL scripts via pgAdminIII.

Identification of control samples in an experiment facilitated incorporation of data at the level of fold change in RETINOBASE. The fold-changes in gene expression were calculated as the ratio between the signal intensities of a given gene in the treated (or knockout) model and the control. In the case of experiments performed in replicate, signal intensities were averaged before calculation of the ratios. All the experiments in RETINOBASE were clustered using 3 independent methods: (i) the density of points clustering (DPC) method [40] which is implemented in the in-house FASABI (Functional And Statistical Analysis of Biological Data) software, (ii) the dot product K-means method [41] used in TM4 Multiexperiment Viewer (MeV) a free, open-source system for microarray data management and analysis [42], (iii) the mixture model method implemented in FASABI. Although cluster analyses often provide useful insights into the data, biological interpretation of the results is recommended, since alternative algorithms generally produce different cluster outputs and no single clustering algorithm is best suited for clustering genes into functional groups for all data sets [43]. We chose the DPC, K-means and mixture models methods because of their robustness in clustering large datasets. Although the K-means method generally requires the user to choose the number of clusters to be calculated, the TMEV system uses figure of merit (FOM) graphs [44] to make an appropriate suggestion. Other clustering algorithms, such as a graph-theoretic approach [45], and a neural network based method SOM [46], as well as different parameter options, will be incorporated in future releases of the database. Storing both the normalized and analyzed data in our relational model allows flexible comparisons across different chips at the level of individual genes.

Quality control

Quality control reports are generated using affyQCReport – an R package that generates quality control reports for Affymetrix array data [47] and RReportGenerator [48] for all experiments, where .CEL files are available. In addition, we also calculate a coefficient of variation for individual Probe Sets between the replicates, which provides a direct estimate of the quality between replicates.

Experiment and sample details

The RETINOBASE home page presents a list of all experiments available to the user and also provides access to experimental details such as title, short description etc. The "Sample details" option (Figure 1) gives details about sample description, organism, tissue, treatment, strain specific information and the array used for hybridisation for a given experiment.

Querying the database

RETINOBASE has three different querying modules: "Gene Information", "Raw Data System Analysis" and "Fold change system Analysis".

Gene information module

The "Gene Information" module offers three different query options – "Gene Query", "Ortholog Query" and "Blast Query". Using these, one can access information such as chromosomal location, linked retinal diseases, cellular localization, and gene ontologies for a given gene. Furthermore, gene details returned from these queries are linked to external databases such as GeneCards [49], NCBI [50], specifically to UniGene [51], ADAPT mapping viewer [52] and also to UCSC genome browser [53] that would yield more information (Figure 2).

"Gene Query" and "Ortholog Query" accept as input the gene name, symbol, Affymetrix Probe Set ID, Refseq or Unigene IDs, whereas "Blast Query" accepts sequences in FASTA format. "Ortholog Query" is useful in cross-referencing probe sets between different Affymetrix GeneChip arrays. The data based on reference sequence similarity is taken from HomoloGene and cross-referenced. In addition, the raw data and cluster information for a given gene (cluster number, software used for clustering and information about other genes present in the same cluster) for all experiments can be obtained through the "Gene Query" (Figure 2).

Raw data system analysis module

This module has "Data and Cluster Query" options and "Data visualization" which is both a query and visualization option. "Data Query" (Figure 3) provides gene expression information at the level of signal intensities for single or multiple genes in all experiments. "Cluster Query" (Figure 3) – unique to RETINOBASE, provides

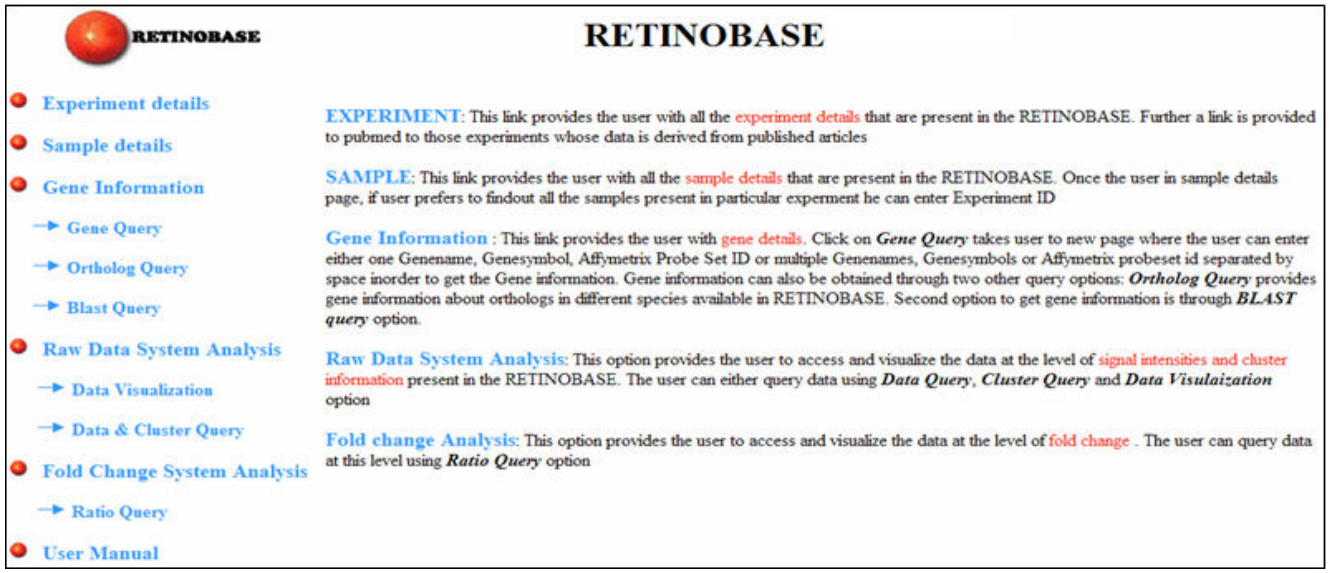


Figure 1
RETINOBASE home page. The home page of RETINOBASE [57] which has general information such as experiment and sample details. Specific query options are shown as in the database.

information about expression patterns of related genes across various conditions and genetic backgrounds. It also identifies any two given genes in the same cluster in one or more experiments. Apart from the above mentioned query options, RETINOBASE also provides a user-friendly transcriptomic data visualization tool that was developed to allow retinal biologists to graphically analyse gene expression profiles across all the experiments. A user can choose the experiment, chip, gene and analysis software to be used in a step-by-step process, following which the related samples can be labelled and organized for an easy comparison through histograms or radar-graph representations (Figure 4). This step-by-step process effectively increases querying speed, which in turn allows faster retrieval of specific data from large volumes of gene expression information. Additional information concerning the number of Probe Sets for a gene on a given chip, the normalization software used to obtain the signal intensities and the quality control report of the experiment are also provided.

Fold change system analysis module

Gene expression information at the level of fold change is provided for single or multiple genes in one or more experiments. In addition, "Ratio Query" supports a specialized query that permits retrieval of all genes from one or more experiments having a fold change greater and/or less than a given criteria.

Downloading results and user manual

In order to allow users to further compare and interpret data, the results from all querying modules available in RETINOBASE can be downloaded in the comma separated value (.CSV) file format using the "Download results" option.

A user manual is also available on the home page of RETINOBASE and it would provide a detailed description of the utilities.

Case study: Use of meta-analysis tools in RETINOBASE

In order to demonstrate the utility of RETINOBASE, we undertook a case study to identify novel genes that may have a potential role in retinal function. In the experiment "Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors" (experiment 7 in RETINOBASE) it was elegantly demonstrated that *Nrl* (neural retina leucine zipper) is a key regulator of photoreceptor differentiation in mammals [17]. We first performed cluster analysis using the "Signal intensity or Cluster query" tool in RETINOBASE by providing *Nrl* as the gene symbol and then retrieved the resulting clusters. In agreement with the original study by Akimoto *et al.*,, our "cluster query" found *Rho* (rhodopsin), *Nr2e3* (nuclear receptor subfamily 2, group E, member 3) and *Pde6b* (phosphodiesterase 6B, cGMP-specific, rod, beta) in the same cluster as *Nrl* in 4 out of 5 possible combinations (1. RMA normalized data and K-

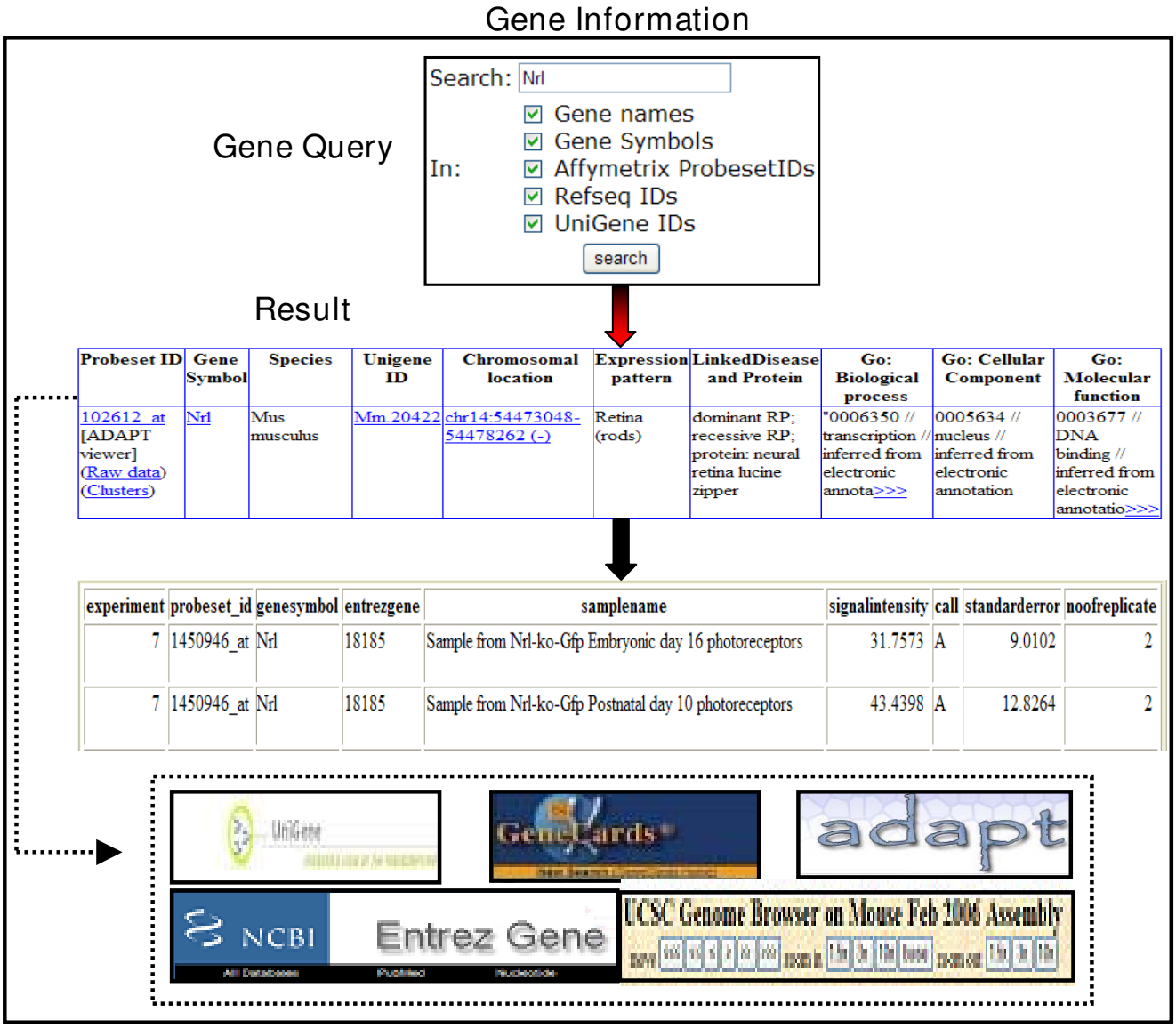
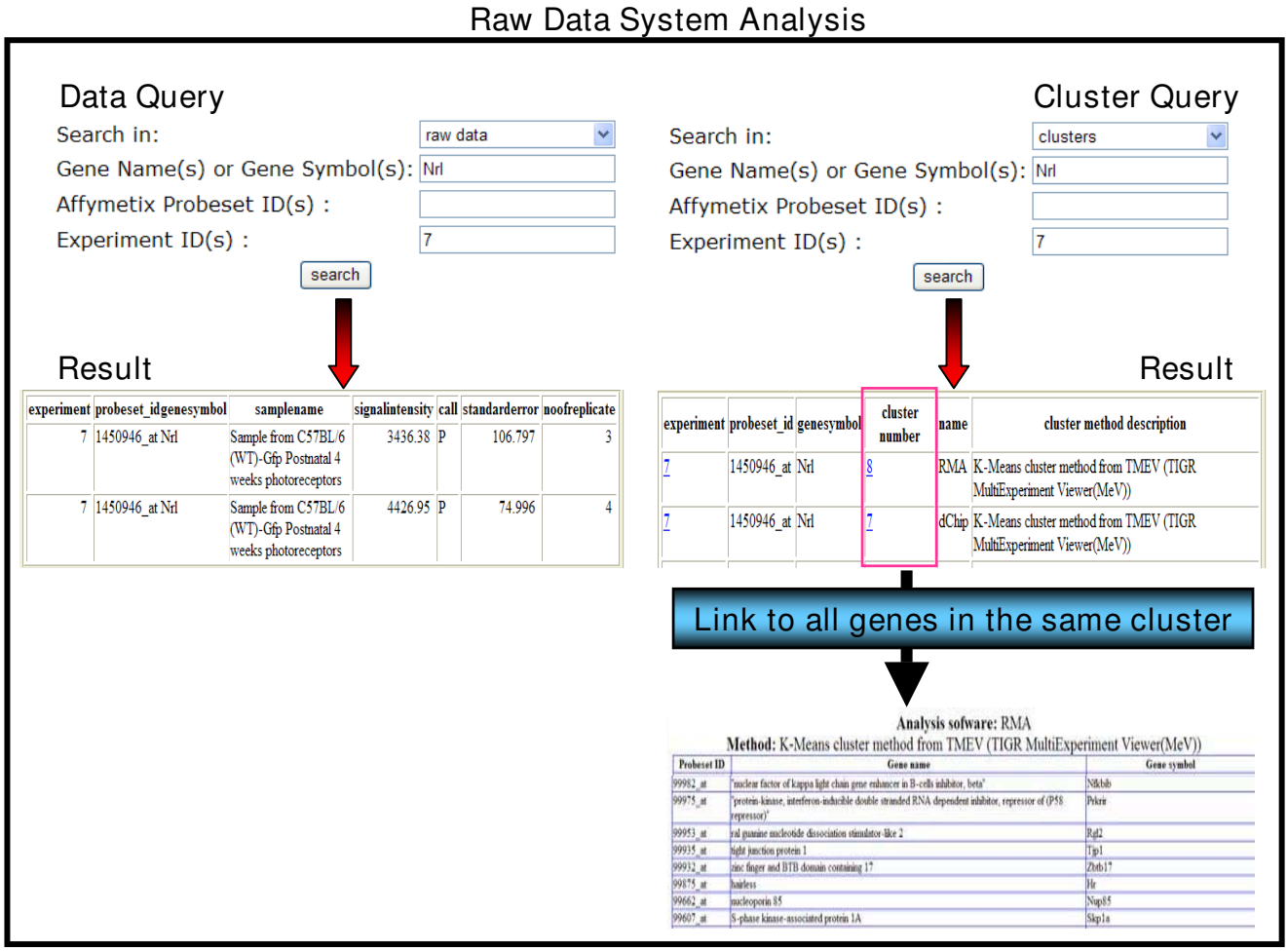


Figure 2
RETINOBASE Queries. A "Gene Query" yields information such as Unigene ID, chromosomal location, Entrez gene, expression pattern, linked diseases and gene ontology. The thick black arrow indicates that raw data and cluster information can be accessed directly from a "Gene Query" output, and the dotted line indicates links to external databases.

means clustering with TMEV, 2. RMA normalized data, K-means clustering with FASABI, 3. dChip normalized data, K-means clustering with TMEV, 4. dChip normalized data, K-means clustering with FASABI and 5. dChip normalized data, clustering with mixture model), confirming that genes specific for rods are coregulated with *Nrl*. In addition, *Gnat1* (guanine nucleotide binding protein (G protein), alpha transducing activity polypeptide 1), a gene implicated in congenital stationary night blindness [54], was also found in the same cluster in all 5 cluster combinations mentioned above, confirming its role in retinal function. This suggests that *Gnat1* is also coregulated with *Nrl* in retina. Based on the similar coexpression profiles in wild type mouse retina at time points corresponding to embryonic day 16, post natal day 2, 6, 10 and 28 (Figure 5), we further identified a novel gene that is likely to be implicated in regulating retinal differentiation, namely *D6Wsu176e* (DNA segment, Chr 6, Wayne State University 176, expressed), described as being expressed in the outer nuclear layer of neural retina [55]. The RETINOBASE "Ortholog query" for *D6Wsu176e* points to the human ortholog, *FAM3C*, that is involved in cell differen-



experiment	probeset_idgenesymbol	samplename	signalintensity	call	standarderror	noofreplicate
7	1450946_at Nrl	Sample from C57BL/6 (WT)-Gfp Postnatal 4 weeks photoreceptors	3436.38	P	106.797	3
7	1450946_at Nrl	Sample from C57BL/6 (WT)-Gfp Postnatal 4 weeks photoreceptors	4426.95	P	74.996	4

experiment	probeset_id	genesymbol	cluster number	name	cluster method description
7	1450946_at	Nrl	8	RMA	K-Means cluster method from TMEV (TIGR MultiExperiment Viewer(MeV))
7	1450946_at	Nrl	7	dChip	K-Means cluster method from TMEV (TIGR MultiExperiment Viewer(MeV))

Probeset ID	Gene name	Gene symbol
99982_at	"nuclear factor of kappa light chain gene enhancer in B-cells inhibitor, beta"	Nfkbib
99975_at	"protein-kinase, interferon-inducible double stranded RNA dependent inhibitor, repressor of (P58 repressor)"	Prkrir
99953_at	ral guanine nucleotide dissociation stimulator-like 2	Rgl2
99935_at	tight junction protein 1	Tjp1
99932_at	zinc finger and BTB domain containing 17	Zbtb17
99875_at	hairless	Hr
99662_at	nucleoporin 85	Nup85
99607_at	S-phase kinase-associated protein 1A	Skp1a

Figure 3
Data and Cluster Query options. Data and cluster query results for the *NRL* gene in experiment 7 [17]: "Targeting GFP to new born by NRL promoter and temporal expression profiling of flow-sorted photoreceptors". The user can subsequently obtain all genes present in the given cluster.

tiation and proliferation during inner ear embryogenesis [56]. With its known function in cell differentiation and its presence in the same cluster as *Nrl* in 3 out of 5 of the above mentioned clustering combinations, *D6Wsu176e* may be an interesting candidate for studying rod differentiation. We further went on to check whether *Nrl*, *Rho*, *Gnat1* and *D6Wsu176e* genes are coexpressed (present in the same cluster) in other experiments present in RETINOBASE, in particular checking experiment 12 (Gene expression patterns in the retina of rds mice treated with CNTF/rAAV virus and non-treated after 60 days of injection) (GEO: GSE4756) and experiment 14 (Biological characterization of gene response in Rpe65-/- mouse model of Leber's congenital amaurosis during progression of the disease) [21]. In these two experiments the four genes mentioned above were present in the same cluster indicating that they might be coregulated. This case study illustrates how RETINOBASE facilitates hypothesis testing for the biologist, and demonstrates how to generate novel hypotheses regarding retinal function and finally, how to identify potential novel targets for human retinopathies.

Future directions

RETINOBASE is under constant development, including addition of new experiments when available. In addition, data from proprietary experiments can be accessed on approval by individual researchers and will be made generally available after publication. Several functional enhancements are also planned for the future. We will continue to refine and update RETINOBASE with respect to data retrieval, mining and visualization options. Direct upload and meta-analysis options will also be provided.

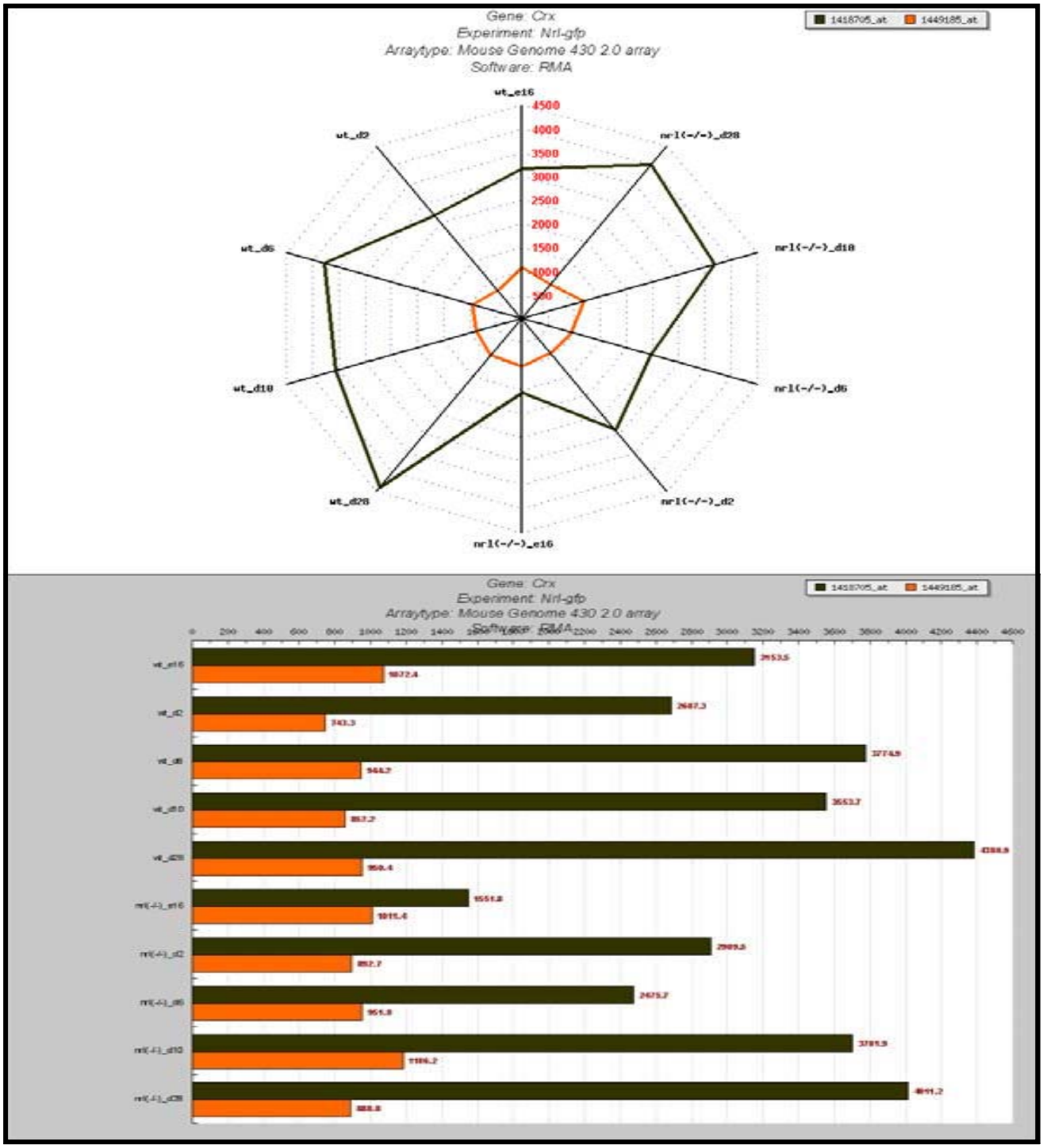


Figure 4

Data visualization. Expression profile of two Probe Sets of cone-rod homeobox containing gene (CRX) in the experiment 7 [17]: "Targeting GFP to new born by NRL promoter and temporal expression profiling of flow-sorted photoreceptors". Data is represented as radar plots on the top panel and as histograms in the bottom panel.

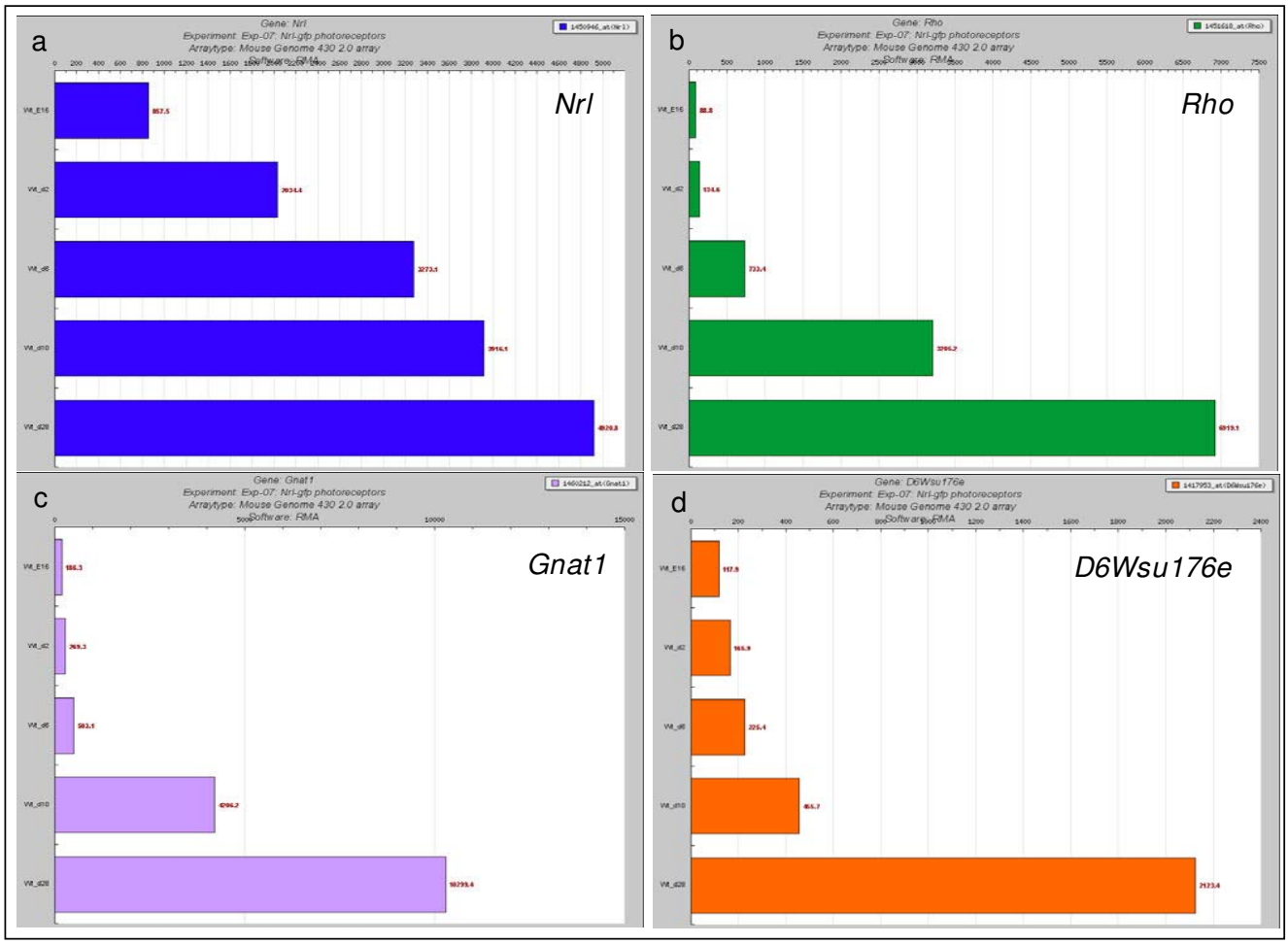


Figure 5
Expression levels (normalised signal intensities) of *Nrl*. (a) Rho, (b) Gnat1, (c) D6Wsu176e, and (d) at embryonic day 16, post natal day 2, 6, 10 and 28 in experiment 7 [17].

Conclusion

RETINOBASE has been developed to store, analyse, visualize and compare retinal-related data in order to provide insights into retinal gene expression in various mouse models and other organisms under diverse conditions. Our database, with different types of query options and powerful visualization tools, allows comprehensive analysis of biological mechanisms/pathways of the retina in normal and diseased conditions. We demonstrated by means of a case study how novel genes such as *D6Wsu176e* (which potentially play an important role in retinal differentiation and development) can be identified using the meta analysis tools incorporated in RETINOBASE. With the addition of new experiments the variety of hypothesis testing options will continuously increase, providing biologists with a valuable tool to gain a better understanding of the retina.

Availability and requirements

The RETINOBASE can be accessed at [57]. All users must register (name and email address) to obtain a username and password.

Authors' contributions

RK is involved in database design and development, data analysis, design of the user interface and prepared the manuscript. NG, GB and RR developed the web services and database back end. LP is involved in testing various querying tools. WR is involved in data analysis and helped to draft the manuscript. TL participated in the design of the user interface. OP was involved in overall design of the project and in drafting the manuscript.

Acknowledgements

We would like to thank Naomi Berdugo for valuable suggestions, as well as "beta tester" users of the RETINOBASE, for their valuable suggestions. We thank Julie Thompson for proofreading the manuscript. This work was sup-

ported by the European Retinal Research Training Network (RETNET) MRTN-CT-2003-504003, EVI-GENORET LSHG-CT-2005-512036, CNRS, INSERM and University of Louis Pasteur (ULP), Strasbourg, France.

References

1. Masland RH: **The fundamental plan of the retina.** *Nat Neurosci* 2001, **4(9):**877-886.
2. Pittler SJ, Baehr W: **Identification of a nonsense mutation in the rod photoreceptor cGMP phosphodiesterase beta-subunit gene of the rd mouse.** *Proc Natl Acad Sci USA* 1991, **88(19):**8322-8326.
3. Akhmedov NB, Piriev NI, Chang B, Rapoport AL, Hawes NL, Nishina PM, Nusinowitz S, Heckenlively JR, Roderick TH, Kozak CA, *et al.*: **A deletion in a photoreceptor-specific nuclear receptor mRNA causes retinal degeneration in the rd7 mouse.** *Proc Natl Acad Sci USA* 2000, **97(10):**5551-5556.
4. Pang JJ, Chang B, Hawes NL, Hurd RE, Davisson MT, Li J, Noorwez SM, Malhotra R, McDowell JH, Kaushal S, *et al.*: **Retinal degeneration 12 (rd12): a new, spontaneously arising mouse model for human Leber congenital amaurosis (LCA).** *Mol Vis* 2005, **11:**152-162.
5. Barrett T, Troup DB, Wilhite SE, Ledoux P, Rudnev D, Evangelista C, Kim IF, Soboleva A, Tomashevsky M, Edgar R: **NCBI GEO: mining tens of millions of expression profiles – database and tools update.** *Nucleic Acids Res* 2007:D760-765.
6. Parkinson H, Kapushesky M, Shojatalab M, Abeygunawardena N, Coulson R, Farne A, Holloway E, Kolesnykov N, Lilja P, Lukk M, *et al.*: **ArrayExpress – a public database of microarray experiments and gene expression profiles.** *Nucleic Acids Res* 2007:D747-750.
7. Kato K, Yamashita R, Matoba R, Monden M, Noguchi S, Takagi T, Nakai K: **Cancer gene expression database (CGED): a database for gene expression profiling with accompanying clinical information of human cancer tissues.** *Nucleic Acids Res* 2005:D533-536.
8. Shah V, Sridhar S, Beane J, Brody JS, Spira A: **SIEGE: Smoking Induced Epithelial Gene Expression Database.** *Nucleic Acids Res* 2005:D573-579.
9. Su AI, Cooke MP, Ching KA, Hakak Y, Walker JR, Wiltshire T, Orth AP, Vega RG, Sapinoso LM, Moqrich A, *et al.*: **Large-scale analysis of the human and mouse transcriptomes.** *Proc Natl Acad Sci USA* 2002, **99(7):**4465-4470.
10. Irizarry RA, Hobbs B, Collin F, Beazer-Barclay YD, Antonellis KJ, Scherf U, Speed TP: **Exploration, normalization, and summaries of high density oligonucleotide array probe level data.** *Biostatistics* 2003, **4(2):**249-264.
11. Li C, Wong WH: **Model-based analysis of oligonucleotide arrays: expression index computation and outlier detection.** *Proc Natl Acad Sci USA* 2001, **98(1):**31-36.
12. Hubbell E, Liu WM, Mei R: **Robust estimators for expression analysis.** *Bioinformatics* 2002, **18(12):**1585-1592.
13. Hartigan JAWM: **A K-Means Clustering Algorithm.** *Applied Statistics* 1979, **28(1):**100-108.
14. Medvedovic M, Sivaganesan S: **Bayesian infinite mixture model based clustering of gene expression profiles.** *Bioinformatics* 2002, **18(9):**1194-1206.
15. Saal LH, Troein C, Vallon-Christersson J, Gruvberger S, Borg A, Peterson C: **BioArray Software Environment (BASE): a platform for comprehensive management and analysis of microarray data.** *Genome Biol* 2002, **3(8):**SOFTWARE0003.
16. Brazma A, Hingamp P, Quackenbush J, Sherlock G, Spellman P, Stoeckert C, Aach J, Ansorge W, Ball CA, Causton HC, *et al.*: **Minimum information about a microarray experiment (MIAME)-toward standards for microarray data.** *Nat Genet* 2001, **29(4):**365-371.
17. Akimoto M, Cheng H, Zhu D, Brzezinski JA, Khanna R, Filippova E, Oh EC, Jing Y, Linares JL, Brooks M, *et al.*: **Targeting of GFP to newborn rods by Nrl promoter and temporal expression profiling of flow-sorted photoreceptors.** *Proc Natl Acad Sci USA* 2006, **103(10):**3890-3895.
18. Yoshida S, Mears AJ, Friedman JS, Carter T, He S, Oh E, Jing Y, Farjo R, Fleury G, Barlow C, *et al.*: **Expression profiling of the developing and mature Nrl-/- mouse retina: identification of retinal disease candidates and transcriptional regulatory targets of Nrl.** *Hum Mol Genet* 2004, **13(14):**1487-1503.
19. Chen J, Rattner A, Nathans J: **The rod photoreceptor-specific nuclear receptor Nr2e3 represses transcription of multiple cone-specific genes.** *J Neurosci* 2005, **25(1):**118-129.
20. Liu J, Huang Q, Higdon J, Liu W, Xie T, Yamashita T, Cheon K, Cheng C, Zuo J: **Distinct gene expression profiles and reduced JNK signaling in retinitis pigmentosa caused by RP1 mutations.** *Hum Mol Genet* 2005, **14(19):**2945-2958.
21. Cottet S, Michaut L, Boisset G, Schlecht U, Gehring W, Schorderet DF: **Biological characterization of gene response in Rpe65-/- mouse model of Leber's congenital amaurosis during progression of the disease.** *Faseb J* 2006, **20(12):**2036-2049.
22. Vazquez-Chona F, Song BK, Geisert EE Jr: **Temporal changes in gene expression after injury in the rat retina.** *Invest Ophthalmol Vis Sci* 2004, **45(8):**2737-2746.
23. Cheng H, Aleman TS, Cideciyan AV, Khanna R, Jacobson SG, Swaroop A: **In vivo function of the orphan nuclear receptor NR2E3 in establishing photoreceptor identity during mammalian retinal development.** *Hum Mol Genet* 2006, **15(17):**2588-2602.
24. Gerhardinger C, Costa MB, Coulombe MC, Toth I, Hoehn T, Grosu P: **Expression of acute-phase response proteins in retinal Muller cells in diabetes.** *Invest Ophthalmol Vis Sci* 2005, **46(1):**349-357.
25. Steele MR, Inman DM, Calkins DJ, Horner PJ, Vetter ML: **Microarray analysis of retinal gene expression in the DBA/2J model of glaucoma.** *Invest Ophthalmol Vis Sci* 2006, **47(3):**977-985.
26. Cameron DA, Gentile KL, Middleton FA, Yurco P: **Gene expression profiles of intact and regenerating zebrafish retina.** *Mol Vis* 2005, **11:**775-791.
27. Abou-Sleymane G, Chalmel F, Helmlinger D, Lardenois A, Thibault C, Weber C, Merienne K, Mandel JL, Poch O, Devys D, *et al.*: **Polyglutamine expansion causes neurodegeneration by altering the neuronal differentiation program.** *Hum Mol Genet* 2006, **15(5):**691-703.
28. Kirwan RP, Leonard MO, Murphy M, Clark AF, O'Brien CJ: **Transforming growth factor-beta-regulated gene transcription and protein expression in human GFAP-negative lamina cribrosa cells.** *Glia* 2005, **52(4):**309-324.
29. Zhang J, Gray J, Wu L, Leone G, Rowan S, Cepko CL, Zhu X, Craft CM, Dyer MA: **Rb regulates proliferation and rod photoreceptor development in the mouse retina.** *Nat Genet* 2004, **36(4):**351-360.
30. Leung YF, Ma P, Dowling JE: **Gene expression profiling of zebrafish embryonic retinal pigment epithelium in vivo.** *Invest Ophthalmol Vis Sci* 2007, **48(2):**881-890.
31. Carter TA, Greenhall JA, Yoshida S, Fuchs S, Helton R, Swaroop A, Lockhart DJ, Barlow C: **Mechanisms of aging in senescence-accelerated mice.** *Genome Biol* 2005, **6(6):**R48.
32. Michaut L, Flister S, Neeb M, White KP, Certa U, Gehring WJ: **Analysis of the eye developmental pathway in Drosophila using DNA microarrays.** *Proc Natl Acad Sci USA* 2003, **100(7):**4024-4029.
33. Velculescu VE, Zhang L, Vogelstein B, Kinzler KW: **Serial analysis of gene expression.** *Science* 1995, **270(5235):**484-487.
34. Diehn JJ, Diehn M, Marmor MF, Brown PO: **Differential gene expression in anatomical compartments of the human eye.** *Genome Biol* 2005, **6(9):**R74.
35. Blackshaw S, Harpavat S, Trimarchi J, Cai L, Huang H, Kuo WP, Weber G, Lee K, Fraioli RE, Cho SH, *et al.*: **Genomic analysis of mouse retinal development.** *PLoS Biol* 2004, **2(9):**E247.
36. **The Affymetrix website** [http://www.affymetrix.com/]
37. **The Retinal information network** [http://www.sph.uth.tmc.edu/Retnet/]
38. **The R statistical package** [http://www.r-project.org]
39. Gentleman RC, Carey VJ, Bates DM, Bolstad B, Dettling M, Dudoit S, Ellis B, Gautier L, Ge Y, Gentry J, *et al.*: **Bioconductor: open software development for computational biology and bioinformatics.** *Genome Biol* 2004, **5(10):**R80.
40. Wicker N, Dembele D, Raffelsberger W, Poch O: **Density of points clustering, application to transcriptomic data analysis.** *Nucleic Acids Res* 2002, **30(18):**3992-4000.
41. Soukas A, Cohen P, Socci ND, Friedman JM: **Leptin-specific patterns of gene expression in white adipose tissue.** *Genes Dev* 2000, **14(8):**963-980.
42. Saeed AI, Bhagabati NK, Braisted JC, Liang W, Sharov V, Howe EA, Li J, Thiagarajan M, White JA, Quackenbush J: **TM4 microarray software suite.** *Methods Enzymol* 2006, **411:**134-193.

43. Datta S, Datta S: **Evaluation of clustering algorithms for gene expression data.** *BMC Bioinformatics* 2006, **7(Suppl 4):**S17.
44. Yeung KY, Haynor DR, Ruzzo WL: **Validating clustering for gene expression data.** *Bioinformatics* 2001, **17(4):**309-318.
45. Sharan R, Maron-Katz A, Shamir R: **CLICK and EXPANDER: a system for clustering and visualizing gene expression data.** *Bioinformatics* 2003, **19(14):**1787-1799.
46. Kohonen T: **Self-Organizing Maps.** 3rd edition. Springer-Verlag Berlin Hiedelberg New York; 2001.
47. Wilson CL, Miller CJ: **Simpleaffy: a BioConductor package for Affymetrix Quality Control and data analysis.** *Bioinformatics* 2005, **21(18):**3683-3685.
48. Raffelsberger W, Krause Y, Moulinier L, Kieffer D, Morand AL, Brino L, Poch O: **RReportGenerator: Automatic reports from routine statistical analysis using R.** *Bioinformatics* 2007.
49. **The GeneCards** [http://www.genecards.org]
50. **The NCBI** [http://www.ncbi.nlm.nih.gov]
51. **The UniGene** [http://www.ncbi.nlm.nih.gov/sites/entrez?db=unigene]
52. Leong HS, Yates T, Wilson C, Miller CJ: **ADAPT: a database of affymetrix probesets and transcripts.** *Bioinformatics* 2005, **21(10):**2552-2553.
53. **The UCSC genome browser** [http://genome.ucsc.edu/cgi-bin/hgGateway]
54. Szabo V, Kreienkamp HJ, Rosenberg T, Gal A: **p.Gln200Glu, a putative constitutively active mutant of rod alpha-transducin (GNAT1) in autosomal dominant congenital stationary night blindness.** *Hum Mutat* 2007, **28(7):**741-742.
55. Blackshaw S, Fraioli RE, Furukawa T, Cepko CL: **Comprehensive analysis of photoreceptor gene expression and the identification of candidate retinal disease genes.** *Cell* 2001, **107(5):**579-589.
56. Pilipenko VV, Reece A, Choo DI, Greinwald JH Jr: **Genomic organization and expression analysis of the murine Fam3c gene.** *Gene* 2004, **335:**159-168.
57. **RETINOBASE** [http://alnitak.u-strasbg.fr/RetinoBase/]

15.3 Annexe 3 – Exemple de format de fichier GFF3

```
##gff-version   3
##sequence-region   ctg123 1 1497228
ctg123 . gene            1000  9000  .  +  .  ID=gene00001;Name=EDEN

ctg123 . TF_binding_site 1000  1012  .  +  .  ID=tfbs00001;Parent=gene00001

ctg123 . mRNA            1050  9000  .  +  .
     ID=mRNA00001;Parent=gene00001;Name=EDEN.1
ctg123 . mRNA            1050  9000  .  +  .
     ID=mRNA00002;Parent=gene00001;Name=EDEN.2
ctg123 . mRNA            1300  9000  .  +  .
     ID=mRNA00003;Parent=gene00001;Name=EDEN.3

ctg123 . exon            1300  1500  .  +  .  ID=exon00001;Parent=mRNA00003
ctg123 . exon            1050  1500  .  +  .
     ID=exon00002;Parent=mRNA00001,mRNA00002
ctg123 . exon            3000  3902  .  +  .
     ID=exon00003;Parent=mRNA00001,mRNA00003
ctg123 . exon            5000  5500  .  +  .
     ID=exon00004;Parent=mRNA00001,mRNA00002,mRNA00003
ctg123 . exon            7000  9000  .  +  .
     ID=exon00005;Parent=mRNA00001,mRNA00002,mRNA00003

ctg123 . CDS             1201  1500  .  +  0
     ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             3000  3902  .  +  0
     ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             5000  5500  .  +  0
     ID=cds00001;Parent=mRNA00001;Name=edenprotein.1
ctg123 . CDS             7000  7600  .  +  0
     ID=cds00001;Parent=mRNA00001;Name=edenprotein.1

ctg123 . CDS             1201  1500  .  +  0
     ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS             5000  5500  .  +  0
     ID=cds00002;Parent=mRNA00002;Name=edenprotein.2
ctg123 . CDS           7000  7600    .  +  0
     ID=cds00002;Parent=mRNA00002;Name=edenprotein.2

ctg123 . CDS             3301  3902  .  +  0
     ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS           5000  5500    .  +  2
     ID=cds00003;Parent=mRNA00003;Name=edenprotein.3
ctg123 . CDS           7000  7600    .  +  2
     ID=cds00003;Parent=mRNA00003;Name=edenprotein.3

ctg123 . CDS             3391  3902  .  +  0
     ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS           5000  5500    .  +  2
     ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
ctg123 . CDS           7000  7600    .  +  2
     ID=cds00004;Parent=mRNA00003;Name=edenprotein.4
```

15.4 Annexe 4 – le standard MIGS/MIMS

Investigation	Report type					
	EU	BA	PL	VI	OR	ME
• Submit to trace archives and INSDC	M	M	M	M	M	M
• Investigation type (i.e., report type)	M	M	M	M	M	M
• Project name[2]	M	M	M	M	M	M
• Study						
• Environment						
• Geographic location (latitude and longitude[float (point, transect and region)], depth and altitude of sample)[(integer)]	M	M	M	M	M	M
• Time of sample collection[(UCT)]	M	M	M	M	M	M
• Habitat[EnvO]	M	M	M	M	M	M
MIMS extension: select to report a set of uniform measurements for a given habitat:						M
• Water body: (temperature, pH, salinity, pressure, chlorophyll, conductivity, light intensity, dissolved organic carbon (DOC), current, atmospheric data, density, alkalinity, dissolved oxygen, particulate organic carbon (POC), phosphate, nitrate, sulfates, sulfides, primary production)[(integer, unit)]						
• Nucleic acid sequence source						
• Subspecific genetic lineage (below lowest rank of NCBI taxonomy, which is subspecies) (e.g., serovar, biotype, ecotype)[(CABRI)]	M	M	M	M	M	–
• Ploidy (e.g., allopolyploid, polyploid)[(PATO)]	M					
• Number of replicons (EU, BA: chromosomes (haploid count); VI: segments)[(integer)]	M	M	–	M	–	–
• Extrachromosomal elements[(integer)]	X	M				
• Estimated size (before sequencing; to apply to all draft genomes)[(integer; base pairs)]	M	X	X	X	X	–
• Reference for biomaterial (primary publication if isolated before genome publication; otherwise, primary genome report)[(PMID or DOI)]	X	M	X	X	X	X
• Source material identifiers: (cultures of microorganisms: identifiers[(alphanumeric)] for two culture collections[(OBI)]; specimens (e.g., organelles and Eukarya): voucher condition and location[(CV)])	M	M	M	M	M	M
• Known pathogenicity		M		M		
• Biotic relationship (e.g., free-living, parasite, commensal, symbiont)[(OBI)]	X	M		X		
• Specific host (e.g., host taxid, unknown, environmental)[EnvO]	X	M	M	M		
• Host specificity or range[(taxid)]	X	X	X	M		
• Health or disease status of specific host at time of collection (e.g., alive, asymptomatic)[PATO]		M		M		
• Trophic level (e.g., autotroph, heterotroph)[PATO]	M	M	–	–	–	–
• Propagation (phage: lytic or lysogenic; plasmid: incompatibility group)[(CV)]	M		M	M	–	–
• Encoded traits (e.g., plasmid: antibiotic resistance; phage: converting genes)[(CV; see caption)]		X	M	M		X
• Relationship to oxygen (e.g., aerobic, anaerobic)[PATO]		M	–	–	–	–
• Isolation and growth conditions[(PMID or DOI)]	M	M	M	M	M	M
• Biomaterial treatment (e.g., filtering of sea water)[(OBI)]						M
• Volume of sample[(integer)]						M
• Sampling strategy (enriched, screened, normalized)[(CV)]						M
• Assay						
• Sequencing						
• Nucleic acid preparation (extraction method[(CV)]; amplification[(CV)]	M	M	M	M	M	M
• Library construction (library size[(integer)], number of reads sequenced[(integer)], vector[(CV)]						M
• Sequencing method (e.g., dideoxysequencing, pyrosequencing, polony)[(OBI)]	M	M	M	M	M	M
• Assembly (assembly method[(CV)], estimated error rate[(unit)] and method of calculation[(CV)])	M	M	M	M	M	M
• Finishing strategy (status—e.g., complete or draft[(CV)], coverage[(integer)], contigs[(integer)])	M	M	X	X	X	X
• Relevant Standard Operating Procedures (SOPs)	M	M	M	M	M	M
• Relevant electronic resources	M	M	M	M	M	M

La liste des critères à respecter se situe à gauche. Chaque colonne à droite représente un type d'expérience en fonction de l'organisme séquencé. Un « M » correspond à un critère obligatoire (*mandatory*), et « X » à un critère optionnel (*extra*). Colonnes de droite : EU (eucaryotes), BA (bactéries et archées), PL (plasmides), VI (virus), OR (organelles), ME (métagénomes). Source (Field, Garrity, Gray, *et al.*, 2008)

15.5 Annexe 5 – Détail des équipes du Consortium *Alvinella*

Équipes	Personnes impliquées
Laboratoire de Bioinformatique et Génomique Intégratives IGBMC, France	Olivier Poch Odile Lecompte
Département de Biologie et Génomique Structurales IGBMC, France	Dino Moras Jean-Claude Thierry
Équipe Écophysiologie : Adaptation et Évolutions Moléculaires Station Biologique de Roscoff, France	Franck Zal
Équipe Évolution et Génétique des Populations marines Station Biologique de Roscoff, France	Didier Jollivet Arnaud Tanguy
Systématique, Adaptation, Évolution - Adaptation en Milieux Extrêmes Université Pierre & Marie Curie, France	Françoise Gaill Bruce Shillito
Systématique, Adaptation, Évolution - Génétique et Évolution Université Pierre & Marie Curie, France	Dominique Higuet
Laboratoire Génome Populations Interactions IFREMER, France	Jacques Dietrich
Laboratoire de Microbiologie Industrielle Université de Reims Champagne-Ardenne, France	Francis Duchiron
Laboratoire de Spectrométrie de masse BioOrganique ECPM, France	Alain Van Dorsselaer Emmanuel Leize
Laboratoire de Biologie et Génétique Évolutive Université du Maine, France	Benoit Chénais Nathalie Casse
Laboratoire d'Études des Parasites Génétiques Université François Rabelais, France	Yves Bigot
Laboratoire de Biologie Cellulaire Institut des Sciences de la vie - Université catholique de Louvain, Belgique	Jean-François Rees Bernard Knoops

15.6 Annexe 6 – descripteur Sortez du site Alvinella

```
<?xml version="1.0" encoding="UTF-8"?>
<sortez_descriptor>
    <name>Alvinella cDNA project website</name>
    <description><![CDATA[
        Alvinella pompejana, the "Pompeii worm", is a Polychaete Annelid
        (see taxonomy) discovered in 1980. This tubiculous worm
        colonizes hydrothermal vents where it is faced with extreme and
        variable physico-chemical conditions including very high
        temperatures (from 20 to over 80°C), anoxic conditions, low ph,
        high concentration of heavy metals and sulphids...<br>This
        environment makes A. pompejana an ideal model for studies aimed
        at deciphering adaptation in general as well as a unique source
        of thermostable proteins of eukaryotic origin.
    ]]></description>
    <main_page url="http://www-alvinella.u-strasbg.fr/Alvinella/" />
    <logo
        url="http://www-alvinella.u-
strasbg.fr/Alvinella/newSDS/images/alvinelle.jpg" />
    <icon
        url="http://www-alvinella.u-
strasbg.fr/Alvinella/newSDS/images/Alvinella94x65.jpg" />
    <sortez_tracker>
        <webservice
            url="http://www-alvinella.u-
strasbg.fr/Alvinella/newSDS/tracker.php" />
        <templates>
            <tpl_definition id="searchResult">
                <css><![CDATA[
.bold {
    font-weight: bold;
}
                ]]></css>
                <tpl><![CDATA[
<span class="bold">Library:</span> ${libraryName}<br/>
<b>Type:</b> ${type}<br/>
<b>Accession:</b> ${accessLink}
                ]]></tpl>
            </tpl_definition>
        </templates>
        <sections>
            <section id="accession" templateId="searchResult">
                <name>Accession numbers</name>
                <description>Search in accession numbers</description>
            </section>
            <section id="blastx" templateId="searchResult">
                <name>Best definition of BlastX</name>
                <description>
                    Search in best definition of BlastX annotation
                </description>
            </section>
        </sections>
    </sortez_tracker>
</sortez_descriptor>
```

RÉFÉRENCES BIBLIOGRAPHIQUES

16 RÉFÉRENCES BIBLIOGRAPHIQUES

Aaronson, J. S., Eckman, B., Blevins, R. A., Borkowski, J. A., Myerson, J., Imran, S., et Elliston, K. O. (1996). Toward the development of a gene index to the human genome: an assessment of the nature of high-throughput EST sequence data. Genome Res *6*, 829-845.

Adams, M. D., Celniker, S. E., Holt, R. A., Evans, C. A., Gocayne, J. D., Amanatides, P. G., Scherer, S. E., Li, P. W., Hoskins, R. A., Galle, R. F., *et al.* (2000). The genome sequence of Drosophila melanogaster. Science *287*, 2185-2195.

Adams, M. D., Kelley, J. M., Gocayne, J. D., Dubnick, M., Polymeropoulos, M. H., Xiao, H., Merril, C. R., Wu, A., Olde, B., et Moreno, R. F. (1991). Complementary DNA sequencing: expressed sequence tags and human genome project. Science *252*, 1651-1656.

Adler, A. S., Littlepage, L. E., Lin, M., Kawahara, T. L. A., Wong, D. J., Werb, Z., et Chang, H. Y. (2008). CSN5 isopeptidase activity links COP9 signalosome activation to breast cancer progression. Cancer Res *68*, 506-515.

Adoutte, A., Balavoine, G., Lartillot, N., Lespinet, O., Prud'homme, B., et de Rosa, R. (2000). The new animal phylogeny: Reliability and implications. Proceedings of the National Academy of Sciences of the United States of America *97*, 4453-4456.

Aguinaldo, A. M., Turbeville, J. M., Linford, L. S., Rivera, M. C., Garey, J. R., Raff, R. A., et Lake, J. A. (1997). Evidence for a clade of nematodes, arthropods and other moulting animals. Nature *387*, 489-493.

Alfarano, C., Andrade, C. E., Anthony, K., Bahroos, N., Bajec, M., Bantoft, K., Betel, D., Bobechko, B., Boutilier, K., Burgess, E., *et al.* (2005). The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res. *33*, D418–D424.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W., et Lipman, D. J. (1990). Basic local alignment search tool. J. Mol. Biol *215*, 403-410.

Altschul, S. F., Madden, T. L., Schaffer, A. A., Zhang, J., Zhang, Z., Miller, W., et Lipman, D. J. (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Res *25*, 3389-3402.

Anderson, S., Bankier, A. T., Barrell, B. G., de Bruijn, M. H., Coulson, A. R., Drouin, J., Eperon, I. C., Nierlich, D. P., Roe, B. A., Sanger, F., *et al.* (1981). Sequence and organization of the human mitochondrial genome. Nature *290*, 457-465.

Angiuoli, S. V., Gussman, A., Klimke, W., Cochrane, G., Field, D., Garrity, G., Kodira, C. D., Kyrpides, N., Madupu, R., Markowitz, V., *et al.* (2008). Toward an online repository of Standard Operating Procedures (SOPs) for (meta)genomic annotation. OMICS *12*, 137-141.

Arakaki, A. K., Huang, Y., et Skolnick, J. (2009). EFICAz2: enzyme function inference by a combined approach enhanced by machine learning. BMC Bioinformatics *10*, 107.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., *et al.* (2000). Gene ontology: tool for the unification of biology. The Gene Ontology Consortium. Nat. Genet *25*, 25-29.

Attwood, T. K., Bradley, P., Flower, D. R., Gaulton, A., Maudling, N., Mitchell, A. L., Moulton, G., Nordle, A., Paine, K., Taylor, P., *et al.* (2003). PRINTS and its automatic supplement, prePRINTS. Nucleic Acids Res. *31*, 400–402.

Avery, O. T., MacLeod, C. M., et McCarty, M. (1944). Studies on the chemical nature of the substance inducing transformation of pneumococcal types. Induction of transformation by a desoxyribonucleic acid fraction isolated from Pneumococcus type III. 1944. Journal of Experimental Medicine *79*, 137-158.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., *et al.* (2008). The RAST Server: rapid annotations using subsystems technology. BMC Genomics *9*, 75.

Bairoch, A. (2000). The ENZYME database in 2000. Nucleic Acids Res *28*, 304-305.

Ball, C. A., et Brazma, A. (2006). MGED standards: work in progress. OMICS *10*, 138-144.

Barrell, D., Dimmer, E., Huntley, R. P., Binns, D., O'Donovan, C., et Apweiler, R. (2009). The GOA database in 2009—an integrated Gene Ontology Annotation resource. Nucleic Acids Res. *37*, D396–D403.

Bech-Otschir, D., Kraft, R., Huang, X., Henklein, P., Kapelari, B., Pollmann, C., et Dubiel, W. (2001). COP9 signalosome-specific phosphorylation targets p53 to degradation by the ubiquitin system. EMBO J *20*, 1630-1639.

Bendtsen, J. D., Nielsen, H., von Heijne, G., et Brunak, S. (2004). Improved prediction of signal peptides: SignalP 3.0. J. Mol. Biol *340*, 783-795.

Benson, D. A., Karsch-Mizrachi, I., Lipman, D. J., Ostell, J., et Sayers, E. W. (2009). GenBank. Nucleic Acids Res. *37*, D26–D31.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., et Bourne, P. E. (2000). The Protein Data Bank. Nucleic Acids Res *28*, 235-242.

Berman, H. M., Westbrook, J. D., Gabanyi, M. J., Tao, W., Shah, R., Kouranov, A., Schwede, T., Arnold, K., Kiefer, F., Bordoli, L., *et al.* (2009). The protein structure initiative structural genomics knowledgebase. Nucleic Acids Res. *37*, D365–D368.

Besemer, J., Lomsadze, A., et Borodovsky, M. (2001). GeneMarkS: a self-training method for prediction of gene starts in microbial genomes. Implications for finding sequence motifs in regulatory regions. Nucleic Acids Res *29*, 2607-2618.

Bieri, T., Blasiar, D., Ozersky, P., Antoshechkin, I., Bastiani, C., Canaran, P., Chan, J., Chen, N., Chen, W. J., Davis, P., *et al.* (2007). WormBase: new content and better access. Nucleic Acids Res *35*, D506-510.

Birney, E., Stamatoyannopoulos, J. A., Dutta, A., Guigó, R., Gingeras, T. R., Margulies, E. H., Weng, Z., Snyder, M., Dermitzakis, E. T., Thurman, R. E., *et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.

Blattner, F. R., Plunkett, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., *et al.* (1997). The complete genome sequence of Escherichia coli K-12. Science *277*, 1453-1462.

Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M., Estreicher, A., Gasteiger, E., Martin, M. J., Michoud, K., O'Donovan, C., Phan, I., *et al.* (2003). The SWISS-PROT protein knowledgebase and its supplement TrEMBL in 2003. Nucleic Acids Res. *31*, 365–370.

Bonnivard, E., Catrice, O., Ravaux, J., Brown, S. C., et Higuet, D. (2009). Survey of genome size in 28 hydrothermal vent species covering 10 families. Genome *52*, 524-536.

Breitkreutz, B., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D. H., Bähler, J., Wood, V., *et al.* (2008). The BioGRID Interaction Database: 2008 update. Nucleic Acids Res. *36*, D637–D640.

Bru, C., Courcelle, E., Carrere, S., Beausse, Y., Dalmar, S., et Kahn, D. (2005). The ProDom database of protein domain families: more emphasis on 3D. Nucleic Acids Res. *33*, D212–D215.

Bryson, K., Loux, V., Bossy, R., Nicolas, P., Chaillou, S., Guchte, M. V. D., Penaud, S., Maguin, E., Hoebeke, M., Bessières, P., *et al.* (2006). AGMIAL: implementing an annotation strategy for prokaryote genomes as a distributed system. Nucleic Acids Res. *34*, 3533–3545.

Bult, C. J., Eppig, J. T., Kadin, J. A., Richardson, J. E., Blake, J. A., et Group, T. M. G. D. (2008). The Mouse Genome Database (MGD): mouse biology and model systems. Nucleic Acids Res. *36*, D724–D728.

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

Burke, D. T., Carle, G. F., et Olson, M. V. (1987). Cloning of large segments of exogenous DNA into yeast by means of artificial chromosome vectors. Science *236*, 806-812.

Campbell, B. J., Stein, J. L., et Cary, S. C. (2003). Evidence of chemolithoautotrophy in the bacterial community associated with Alvinella pompejana, a hydrothermal vent polychaete. Appl. Environ. Microbiol *69*, 5070-5078.

Cannone, J. J., Subramanian, S., Schnare, M. N., Collett, J. R., D'Souza, L. M., Du, Y., Feng, B., Lin, N., Madabusi, L. V., Müller, K. M., *et al.* (2002). The Comparative RNA Web (CRW) Site: an online database of comparative sequence and structure information for ribosomal, intron, and other RNAs. BMC Bioinformatics. *3*, 2.

Cantarel, B. L., Korf, I., Robb, S. M. C., Parra, G., Ross, E., Moore, B., Holt, C., Sanchez Alvarado, A., et Yandell, M. (2008). MAKER: an easy-to-use annotation pipeline designed for emerging model organism genomes. Genome Res *18*, 188-196.

Cary, S. C., Shank, T., et Stein, J. (1998). Worms bask in extreme temperatures. Nature *391*, 545-546.

Caspi, R., Foerster, H., Fulcher, C. A., Kaipa, P., Krummenacker, M., Latendresse, M., Paley, S., Rhee, S. Y., Shearer, A. G., Tissier, C., *et al.* (2008). The MetaCyc Database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. Nucleic Acids Res *36*, D623-631.

CeSC (1998). Genome sequence of the nematode C. elegans: a platform for investigating biology. Science *282*, 2012-2018.

Chalmel, F., Lardenois, A., Thompson, J. D., Müller, J., Sahel, J., Léveillard, T., et Poch, O. (2005). GOAnno: GO annotation based on multiple alignment. Bioinformatics *21*, 2095-2096.

Chatr-aryamontri, A., Ceol, A., Palazzi, L. M., Nardelli, G., Schneider, M. V., Castagnoli, L., et Cesareni, G. (2007). MINT: the Molecular INTeraction database. Nucl. Acids Res. *35*, D572-574.

Chevaldonné, P., Fisher, C. R., Childress, J. J., Desbruyères, D., Jollivet, D., Zal, F., et Toulmond, A. (2000). Thermotolerance and the 'Pompeii worms'. Marine Ecology Progress Series *208*, 293-295.

Chou, P. Y., et Fasman, G. D. (1974). Conformational parameters for amino acids in helical, beta-sheet, and random coil regions calculated from proteins. Biochemistry *13*, 211-222.

Claudel-Renard, C., Chevalet, C., Faraut, T., et Kahn, D. (2003). Enzyme-specific profiles for genome annotation: PRIAM. Nucleic Acids Res *31*, 6633-6639.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., *et al.* (2009a). Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res *37*, D19-25.

Cochrane, G., Akhtar, R., Bonfield, J., Bower, L., Demiralp, F., Faruque, N., Gibson, R., Hoad, G., Hubbard, T., Hunter, C., *et al.* (2009b). Petabyte-scale innovations at the European Nucleotide Archive. Nucleic Acids Res. *37*, D19–D25.

Cokol, M., Nair, R., et Rost, B. (2000). Finding nuclear localization signals. EMBO Rep *1*, 411-415.

Corliss, J. B., et Ballard, R. D. (1977). Oases of Life in the Cold Abyss. National Geographic *152*, 441-453.

CRICK, F. H. (1958). On protein synthesis. Symp. Soc. Exp. Biol *12*, 138-163.

Dayhoff, M., et National Biomedical Research Foundation. (1965). Atlas of protein sequence and structure 1er éd. (Silver Spring Md.: National Biomedical Research Foundation).

Day-Richter, J., Harris, M. A., Haendel, M., The Gene Ontology OBO-Edit Working Group, et Lewis, S. (2007). OBO-Edit an ontology editor for biologists. Bioinformatics *23*, 2198-2200.

De Robertis, E. M. (2008). Evo-devo: variations on ancestral themes. Cell *132*, 185-195.

Dear, S., et Staden, R. (1992). A standard file format for data from DNA sequencing instruments. DNA Seq *3*, 107-110.

Delcher, A. L., Bratke, K. A., Powers, E. C., et Salzberg, S. L. (2007). Identifying bacterial genes and endosymbiont DNA with Glimmer. Bioinformatics, btm009.

Dennis, C., et Surridge, C. (2000). Arabidopsis thaliana genome. Introduction. Nature *408*, 791.

Dennis, G., Sherman, B. T., Hosack, D. A., Yang, J., Gao, W., Lane, H. C., et Lempicki, R. A. (2003). DAVID: Database for Annotation, Visualization, and Integrated Discovery. Genome Biol *4*, P3.

Desbruyères, D., Chevaldonné, P., Alayse, A. M., Jollivet, D., Lallier, F. H., Jouin-Toulmond, C., Zal, F., Sarradin, P. M., Cosson, R., Caprais, J. C., *et al.* (1998). Biology and ecology of the Pompei worm (Alvinella pompejana Desbruyères and Laubier), a normal dweller on an extreme deep-sea environment: a synthesis of current knowledge and recent developments. Deep-Sea Resaech Part II *45*, 383-422.

Desbruyères, D., et Laubier, L. (1986). Les Alvinellidae, une famille nouvelle d'annélides polychètes inféodées aux sources hydrothermales sous-marines : systématique, biologie et écologie. Canadian Journal of Zoology *64*, 2227-2245.

Desbruyères, D., et Laubier, L. (1980). Alvinella pompejana gen. sp. nov., Ampharetidae aberrant des sources hydrothermales de la ride Est-Pacifique. Oceanol Acta *3*, 267-274.

Deshayes, C., Perrodou, E., Gallien, S., Euphrasie, D., Schaeffer, C., Van-Dorsselaer, A., Poch, O., Lecompte, O., et Reyrat, J. (2007). Interrupted coding sequences in Mycobacterium smegmatis: authentic mutations or sequencing errors? Genome Biol *8*, R20.

Dickmeis, T., et Müller, F. (2005). The identification and functional characterisation of conserved regulatory elements in developmental genes. Brief Funct Genomic Proteomic *3*, 332-350.

Dixon, D. R., Jolly, M. T., Vevers, W. F., et Dixon, L. R. (2006). Chromosomes of Pacific hydrothermal vent invertebrates: towards a greater understanding of the relationship between chromosome and molecular evolution. Journal of the Marine Biological Association of the UK *Forthcoming*, 1-17.

Dowell, R. D., Jokerst, R. M., Day, A., Eddy, S. R., et Stein, L. (2001). The Distributed Annotation System. BMC Bioinformatics. *2*, 7.

Eilbeck, K., Lewis, S. E., Mungall, C. J., Yandell, M., Stein, L., Durbin, R., et Ashburner, M. (2005). The Sequence Ontology: a tool for the unification of genome annotations. Genome Biol. *6*, R44.

ENCODE Project Consortium (2004). The ENCODE (ENCyclopedia Of DNA Elements) Project. Science *306*, 636-640.

Etzold, T., et Argos, P. (1993). SRS--an indexing and retrieval tool for flat file data libraries. Comput. Appl. Biosci *9*, 49-57.

Ewing, B., et Green, P. (1998). Base-calling of automated sequencer traces using phred. II. Error probabilities. Genome Res *8*, 186-194.

Ewing, B., Hillier, L., Wendl, M. C., et Green, P. (1998). Base-calling of automated sequencer traces using phred. I. Accuracy assessment. Genome Res *8*, 175-185.

Field, D., Garrity, G., Gray, T., *et al.* (2008). The minimum information about a genome sequence (MIGS) specification. Nat Biotech *26*, 541-547.

Field, D., Garrity, G. M., Sansone, S., *et al.* (2008). Meeting report: the fifth Genomic Standards Consortium (GSC) workshop. OMICS *12*, 109-113.

Field, D., et Sansone, S. (2006). A Special Issue on Data Standards. OMICS: A Journal of Integrative Biology *10*, 84-93.

Finn, R. D., Tate, J., Mistry, J., Coggill, P. C., Sammut, S. J., Hotz, H., Ceric, G., Forslund, K., Eddy, S. R., Sonnhammer, E. L. L., *et al.* (2008). The Pfam protein families database. Nucl. Acids Res. *36*, D281-288.

Fitch, W. M., et Margoliash, E. (1967). Construction of Phylogenetic Trees. Science *155*, 279-284.

Fleischmann, R. D., Adams, M. D., White, O., Clayton, R. A., Kirkness, E. F., Kerlavage, A. R., Bult, C. J., Tomb, J. F., Dougherty, B. A., et Merrick, J. M. (1995). Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. Science *269*, 496-512.

Fraser, C. M., Gocayne, J. D., White, O., Adams, M. D., Clayton, R. A., Fleischmann, R. D., Bult, C. J., Kerlavage, A. R., Sutton, G., Kelley, J. M., *et al.* (1995). The minimal gene complement of Mycoplasma genitalium. Science *270*, 397-403.

FREIFELDER, D., et DAVISON, P. F. (1962). Studies on the sonic degradation of deoxyribonucleic acid. Biophys. J *2*, 235-247.

Freyhult, E. K., Bollback, J. P., et Gardner, P. P. (2007). Exploring genomic dark matter: A critical assessment of the performance of homology search methods on noncoding RNA. Genome Res. *17*, 117–125.

Friedberg, I. (2006). Automated protein function prediction--the genomic challenge. Brief. Bioinformatics *7*, 225-242.

Fukunishi, Y., et Hayashizaki, Y. (2001). Amino acid translation program for full-length cDNA sequences with frameshift errors. Physiol. Genomics *5*, 81-87.

Gallien, S., Perrodou, E., Carapito, C., Deshayes, C., Reyrat, J., Van Dorsselaer, A., Poch, O., Schaeffer, C., et Lecompte, O. (2009). Ortho-proteogenomics: multiple proteomes investigation through orthology and a new MS-based protocol. Genome Res *19*, 128-135.

Gardner, P. P., Daub, J., Tate, J. G., Nawrocki, E. P., Kolbe, D. L., Lindgreen, S., Wilkinson, A. C., Finn, R. D., Griffiths-Jones, S., Eddy, S. R., *et al.* (2009). Rfam: updates to the RNA families database. Nucl. Acids Res. *37*, D136-140.

Gardy, J. L., Laird, M. R., Chen, F., Rey, S., Walsh, C. J., Ester, M., et Brinkman, F. S. L. (2005). PSORTb v.2.0: Expanded prediction of bacterial protein subcellular localization and insights gained from comparative proteome analysis. Bioinformatics *21*, 617-623.

Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A. H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C. J. A., Lachaize, C., *et al.* (2003). Automated annotation of microbial proteomes in SWISS-PROT. Comput Biol Chem *27*, 49-58.

Gerstein, M. B., Bruce, C., Rozowsky, J. S., Zheng, D., Du, J., Korbel, J. O., Emanuelsson, O., Zhang, Z. D., Weissman, S., et Snyder, M. (2007). What is a gene, post-ENCODE? History and updated definition. Genome Res *17*, 669-681.

Gilchrist, D. A., Fargo, D. C., et Adelman, K. (2009). Using ChIP-chip and ChIP-seq to study the regulation of gene expression: genome-wide localization studies reveal widespread regulation of transcription elongation. Methods *48*, 398-408.

Godzik, A., Jambon, M., et Friedberg, I. (2007). Computational protein function prediction: are we making progress? Cell. Mol. Life Sci *64*, 2505-2511.

Gordon, D., Abajian, C., et Green, P. (1998). Consed: a graphical tool for sequence finishing. Genome Res *8*, 195-202.

Gordon, D., Desmarais, C., et Green, P. (2001). Automated finishing with autofinish. Genome Res *11*, 614-625.

Gordon, D. (2003). Viewing and editing assembled sequences using Consed. Curr Protoc Bioinformatics *Chapter 11*, Unit11.2.

Götz, S., García-Gómez, J. M., Terol, J., Williams, T. D., Nagaraj, S. H., Nueda, M. J., Robles, M., Talón, M., Dopazo, J., et Conesa, A. (2008). High-throughput functional annotation and data mining with the Blast2GO suite. Nucleic Acids Res *36*, 3420-3435.

Gouy, M., Gautier, C., Attimonelli, M., Lanave, C., et di Paola, G. (1985). ACNUC--a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage. Comput. Appl. Biosci *1*, 167-172.

Gouzy, J., Carrere, S., et Schiex, T. (2009). FrameDP: sensitive peptide detection on noisy matured sequences. Bioinformatics *25*, 670-671.

Gräslund, S., Nordlund, P., Weigelt, J., Hallberg, B. M., Bray, J., Gileadi, O., Knapp, S., Oppermann, U., Arrowsmith, C., Hui, R., *et al.* (2008). Protein production and purification. Nat. Methods *5*, 135-146.

Gross, S. S., Do, C. B., Sirota, M., et Batzoglou, S. (2007). CONTRAST: a discriminative, phylogeny-free approach to multiple informant de novo gene prediction. Genome Biol *8*, R269.

Haft, D. H., Selengut, J. D., et White, O. (2003). The TIGRFAMs database of protein families. Nucleic Acids Res. *31*, 371–373.

Hamosh, A., Scott, A. F., Amberger, J. S., Bocchini, C. A., et McKusick, V. A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res. *33*, D514–D517.

Hatzigeorgiou, A. G., Fiziev, P., et Reczko, M. (2001). DIANA-EST: a statistical analysis. Bioinformatics *17*, 913-919.

Henscheid, K. L., Shin, D. S., Cary, S. C., et Berglund, J. A. (2005). The splicing factor U2AF65 is functionally conserved in the thermotolerant deep-sea worm Alvinella pompejana. Biochim. Biophys. Acta *1727*, 197-207.

Higgins, D. G., et Sharp, P. M. (1988). CLUSTAL: a package for performing multiple sequence alignment on a microcomputer. Gene *73*, 237-244.

Hong, E. L., Balakrishnan, R., Dong, Q., Christie, K. R., Park, J., Binkley, G., Costanzo, M. C., Dwight, S. S., Engel, S. R., Fisk, D. G., *et al.* (2008). Gene Ontology annotations at SGD: new data sources and annotation methods. Nucleic Acids Res. *36*, D577–D581.

Huang, D. W., Sherman, B. T., et Lempicki, R. A. (2009). Systematic and integrative analysis of large gene lists using DAVID bioinformatics resources. Nat Protoc *4*, 44-57.

Huang, X., et Madan, A. (1999). CAP3: A DNA sequence assembly program. Genome Res *9*, 868-877.

Huang, Y., Gilna, P., et Li, W. (2009). Identification of ribosomal RNA genes in metagenomic fragments. Bioinformatics *25*, 1338-1340.

Hubbard, T. J. P., Aken, B. L., Ayling, S., Ballester, B., Beal, K., Bragin, E., Brent, S., Chen, Y., Clapham, P., Clarke, L., *et al.* (2009). Ensembl 2009. Nucleic Acids Res *37*, D690-697.

Hulo, N., Bairoch, A., Bulliard, V., Cerutti, L., Cuche, B. A., Castro, E. D., Lachaize, C., Langendijk-Genevaux, P. S., et Sigrist, C. J. A. (2008). The 20 years of PROSITE. Nucleic Acids Res. *36*, D245–D249.

Hunter, S., Apweiler, R., Attwood, T. K., Bairoch, A., Bateman, A., Binns, D., Bork, P., Das, U., Daugherty, L., Duquenne, L., *et al.* (2009). InterPro: the integrative protein signature database. Nucleic Acids Res. *37*, D211–D215.

Hurtado, L. A., Lutz, R. A., et Vrijenhoek, R. C. (2004). Distinct patterns of genetic differentiation among annelids of eastern Pacific hydrothermal vents. Mol. Ecol *13*, 2603-2615.

IHGSC (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Iseli, C., Jongeneel, C. V., et Bucher, P. (1999). ESTScan: a program for detecting, evaluating, and reconstructing potential coding regions in EST sequences. Proc Int Conf Intell Syst Mol Biol, 138-148.

Jenkinson, A., Albrecht, M., Birney, E., Blankenburg, H., Down, T., Finn, R., Hermjakob, H., Hubbard, T., Jimenez, R., Jones, P., *et al.* (2008). Integrating biological data - the Distributed Annotation System. BMC Bioinformatics *9 Suppl 8*, S3.

Jensen, L. J., Kuhn, M., Stark, M., Chaffron, S., Creevey, C., Muller, J., Doerks, T., Julien, P., Roth, A., Simonovic, M., *et al.* (2009). STRING 8--a global view on proteins and their functional interactions in 630 organisms. Nucleic Acids Res *37*, D412-416.

Jensen, L. J., Lagarde, J., Mering, C. V., et Bork, P. (2004). ArrayProspector: a web resource of functional associations inferred from microarray expression data. Nucleic Acids Res. *32*, W445–W448.

Jimenez, R. C., Quinn, A. F., Garcia, A., Labarga, A., O'Neill, K., Martinez, F., Salazar, G. A., et Hermjakob, H. (2008). Dasty2, an Ajax protein DAS client. Bioinformatics *24*, 2119-2121.

Jollivet, D. (1996). Specific and genetic diversity at deep-sea hydrothermal vents: an overview. Biodiversity and Conservation *5*, 1619-1653.

Kanehisa, M., Araki, M., Goto, S., Hattori, M., Hirakawa, M., Itoh, M., Katayama, T., Kawashima, S., Okuda, S., Tokimatsu, T., *et al.* (2008). KEGG for linking genomes to life and the environment. Nucleic Acids Res *36*, D480-484.

Kerrien, S., Alam-Faruque, Y., Aranda, B., Bancarz, I., Bridge, A., Derow, C., Dimmer, E., Feuermann, M., Friedrichsen, A., Huntley, R., *et al.* (2007). IntAct—open source resource for molecular interaction data. Nucleic Acids Res. *35*, D561–D565.

Keseler, I. M., Bonavides-Martinez, C., Collado-Vides, J., Gama-Castro, S., Gunsalus, R. P., Johnson, D. A., Krummenacker, M., Nolan, L. M., Paley, S., Paulsen, I. T., *et al.* (2009). EcoCyc: A comprehensive view of Escherichia coli biology. Nucl. Acids Res. *37*, D464-470.

Kim, D., Jung, T., Nam, S., Kwon, H., Kim, A., Chae, S., Choi, S., Kim, D., Kim, R. N., et Park, H. (2009). GarlicESTdb: an online database and mining tool for garlic EST sequences. BMC Plant Biol. *9*, 61.

Kitts, P. A., Madden, T. L., Sicotte, H., Black, L., et Ostell, J. A. UniVec. Available at: http://www.ncbi.nlm.nih.gov/VecScreen/UniVec.html.

Koonin, E. (2003). Sequence - evolution - function : computational approaches in comparative genomics (Boston: Kluwer Academic).

Krogh, A., Larsson, B., von Heijne, G., et Sonnhammer, E. L. (2001). Predicting transmembrane protein topology with a hidden Markov model: application to complete genomes. J. Mol. Biol *305*, 567-580.

Kuhn, R. M., Karolchik, D., Zweig, A. S., Wang, T., Smith, K. E., Rosenbloom, K. R., Rhead, B., Raney, B. J., Pohl, A., Pheasant, M., *et al.* (2009). The UCSC Genome Browser Database: update 2009. Nucleic Acids Res *37*, D755-761.

Kulesh, D. A., Clive, D. R., Zarlenga, D. S., et Greene, J. J. (1987). Identification of interferon-modulated proliferation-related cDNA sequences. Proc. Natl. Acad. Sci. U.S.A *84*, 8453-8457.

Kuska, B. (1998). Beer, Bethesda, and biology: how "genomics" came into being. J. Natl. Cancer Inst *90*, 93.

Lagesen, K., Hallin, P., Rødland, E. A., Stærfeldt, H., Rognes, T., et Ussery, D. W. (2007). RNAmmer: consistent and rapid annotation of ribosomal RNA genes. Nucleic Acids Res. *35*, 3100–3108.

Lander, E. S., Linton, L. M., Birren, B., Nusbaum, C., Zody, M. C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W., *et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.

Le Bris, N., et Gaill, F. (2007). How does the annelid Alvinella pompejana deal with an extreme hydrothermal environment? Reviews in Environmental Science and Biotechnology *6*, 197-221.

Lecompte, O., Ripp, R., Puzos-Barbe, V., Duprat, S., Heilig, R., Dietrich, J., Thierry, J. C., et Poch, O. (2001). Genome evolution at the genus level: comparison of three complete genomes of hyperthermophilic archaea. Genome Res *11*, 981-993.

Lecompte, O., Thompson, J. D., Plewniak, F., Thierry, J., et Poch, O. (2001). Multiple alignment of complete sequences (MACS) in the post-genomic era. Gene *270*, 17-30.

Lee, B., Hong, T., Byun, S. J., Woo, T., et Choi, Y. J. (2007). ESTpass: a web-based server for processing and annotating expressed sequence tag (EST) sequences. Nucleic Acids Res. *35*, W159–W162.

Letunic, I., Doerks, T., et Bork, P. (2009). SMART 6: recent updates and new developments. Nucleic Acids Res. *37*, D229–D232.

LEVINTHAL, C., et DAVISON, P. F. (1961). Degradation of deoxyribonucleic acid under hydrodynamic shearing forces. J. Mol. Biol *3*, 674-683.

Levy, S., Sutton, G., Ng, P. C., Feuk, L., Halpern, A. L., Walenz, B. P., Axelrod, N., Huang, J., Kirkness, E. F., Denisov, G., *et al.* (2007). The diploid genome sequence of an individual human. PLoS Biol *5*, e254.

Liang, F., Holt, I., Pertea, G., Karamycheva, S., Salzberg, S. L., et Quackenbush, J. (2000). An optimized protocol for analysis of EST sequences. Nucleic Acids Res. *28*, 3657–3665.

Liang, P., et Pardee, A. B. (1992). Differential display of eukaryotic messenger RNA by means of the polymerase chain reaction. Science *257*, 967-971.

Lipman, D. J., et Pearson, W. R. (1985). Rapid and sensitive protein similarity searches. Science *227*, 1435-1441.

Lomsadze, A., Ter-Hovhannisyan, V., Chernoff, Y. O., et Borodovsky, M. (2005). Gene identification in novel eukaryotic genomes by self-training algorithm. Nucleic Acids Res *33*, 6494-6506.

Lottaz, C., Iseli, C., Jongeneel, C. V., et Bucher, P. (2003). Modeling sequencing errors by combining Hidden Markov models. Bioinformatics *19 Suppl 2*, ii103-112.

Lowe, T. M., et Eddy, S. R. (1997). tRNAscan-SE: a program for improved detection of transfer RNA genes in genomic sequence. Nucleic Acids Res. *25*, 955–964.

Lukashin, A. V., et Borodovsky, M. (1998). GeneMark.hmm: new solutions for gene finding. Nucleic Acids Res *26*, 1107-1115.

Majoros, W. H., Pertea, M., et Salzberg, S. L. (2004). TigrScan and GlimmerHMM: two open source ab initio eukaryotic gene-finders. Bioinformatics *20*, 2878-2879.

Manatee Manatee. Available at: http://manatee.sourceforge.net/.

Mandel, M. J., Stabb, E. V., et Ruby, E. G. (2008). Comparative genomics-based investigation of resequencing targets in Vibrio fischeri: focus on point miscalls and artefactual expansions. BMC Genomics *9*, 138.

Mao, C., Cushman, J. C., May, G. D., et Weller, J. W. (2003). ESTAP--an automated system for the analysis of EST data. Bioinformatics *19*, 1720-1722.

Marcotte, C. J. V., et Marcotte, E. M. (2002). Predicting functional linkages from gene fusions with confidence. Appl. Bioinformatics *1*, 93-100.

Marcotte, E. M., Xenarios, I., et Eisenberg, D. (2001). Mining literature for protein-protein interactions. Bioinformatics *17*, 359-363.

Markowitz, V. M., Mavromatis, K., Ivanova, N. N., Chen, I. A., Chu, K., et Kyrpides, N. C. (2009). IMG ER: A System for Microbial Genome Annotation Expert Review and Curation. Bioinformatics. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19561336 [Accédé Juillet 20, 2009].

Maruyama, K., et Sugano, S. (1994). Oligo-capping: a simple method to replace the cap structure of eukaryotic mRNAs with oligoribonucleotides. Gene *138*, 171-174.

MATTHAEI, J. H., JONES, O. W., MARTIN, R. G., et NIRENBERG, M. W. (1962). Characteristics and composition of RNA coding units. Proc. Natl. Acad. Sci. U.S.A *48*, 666-677.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., *et al.* (2009a). Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res *37*, D619-622.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B., *et al.* (2009b). Reactome knowledgebase of human biological pathways and processes. Nucl. Acids Res. *37*, D619-622.

Maxam, A. M., et Gilbert, W. (1977). A new method for sequencing DNA. Proc Natl Acad Sci U S A. *74*, 560–564.

Mulder, N., et Apweiler, R. (2007). InterPro and InterProScan: tools for protein sequence classification and comparison. Methods Mol. Biol *396*, 59-70.

Mullis, K. (1994). The Polymerase chain reaction (Boston: Birkhäuser).

MySQL MySQL: The world's most popular open source database. Available at: http://www.mysql.com/.

Nagaraj, S. H., Deshpande, N., Gasser, R. B., et Ranganathan, S. (2007). ESTExplorer: an expressed sequence tag (EST) assembly and annotation platform. Nucleic Acids Res 3*5,* W143-147.

Nagaraj, S. H., Gasser, R. B., et Ranganathan, S. (2007). A hitchhiker's guide to expressed sequence tag (EST) analysis. Brief. Bioinformatics 8*,* 6-21.

Nair, R., Liu, J., Soong, T., Acton, T., Everett, J., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C., *et al.* (2009). Structural genomics is the largest contributor of novel structural leverage. Journal of Structural and Functional Genomics 1*0,* 181-191.

Nam, S., Kim, D., Jung, T., Choi, Y., Kim, D., Choi, H., Choi, S., et Park, H. (2009). PESTAS: a web server for EST analysis and sequence mining. Bioinformatics. Available at: http://www.ncbi.nlm.nih.gov/pubmed/19414531 [Accédé Juin 17, 2009].

Natale, D. A., Arighi, C. N., Barker, W. C., Blake, J., Chang, T., Hu, Z., Liu, H., Smith, B., et Wu, C. H. (2007). Framework for a protein ontology. BMC Bioinformatics 8 *Suppl 9,* S1.

National Research Council (U.S.). (1988). Mapping and sequencing the human genome Committee on Mapping and Sequencing the Human Genome, Board on Basic Biology, Commission on Life Sciences, National Research Council. (Washington D.C.: National Academy Press).

Needleman, S. B., et Wunsch, C. D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J. Mol. Biol 4*8,* 443-453.

Nguyen, H., Berthommier, G., Friedrich, A., Poidevin, L., Ripp, R., Moulinier, L., et Poch, O. (2008). Introduction du nouveau centre de données biomédicales Décrypthon. Dans CORIA, pp. 151-164.

O'Connor, M., Peifer, M., et Bender, W. (1989). Construction of large DNA segments in Escherichia coli. Science 2*44,* 1307-1312.

Ousterhout, J. (1994). Tcl and the Tk toolkit (Reading Mass.: Addison-Wesley).

Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G. D., et Maltsev, N. (1999). The use of gene clusters to infer functional coupling. Proc. Natl. Acad. Sci. U.S.A 9*6,* 2896-2901.

Parkinson, J., Anthony, A., Wasmuth, J., Schmid, R., Hedley, A., et Blaxter, M. (2004). PartiGene--constructing partial genomes. Bioinformatics 2*0,* 1398-1404.

Paulus, T., et Müller, M. C. M. (2006). Cell proliferation dynamics and morphological differentiation during regeneration in Dorvillea bermudensis (Polychaeta, Dorvilleidae). J. Morphol 2*67,* 393-403.

Pearson, H. (2006). Genetics: What is a gene? Nature 4*41,* 398-401.

Pearson, W. R., et Lipman, D. J. (1988). Improved tools for biological sequence comparison. Proc. Natl. Acad. Sci. U.S.A 8*5,* 2444-2448.

Perrodou, E., Deshayes, C., Muller, J., Schaeffer, C., Van Dorsselaer, A., Ripp, R., Poch, O., Reyrat, J., et Lecompte, O. (2006). ICDS database: interrupted CoDing sequences in prokaryotic genomes. Nucleic Acids Res 3*4,* D338-343.

Pertea, G., Huang, X., Liang, F., Antonescu, V., Sultana, R., Karamycheva, S., Lee, Y., White, J., Cheung, F., Parvizi, B., *et al.* (2003). TIGR Gene Indices clustering tools (TGICL): a software system for fast clustering of large EST datasets. Bioinformatics 1*9,* 651-652.

PHP PHP: Hypertext Preprocessor. Available at: http://www.php.net/.

Plewniak, F., Thompson, J. D., et Poch, O. (2000). Ballast: blast post-processing based on locally conserved segments. Bioinformatics 1*6,* 750-759.

Plewniak, F., Bianchetti, L., Brelivet, Y., Carles, A., Chalmel, F., Lecompte, O., Mochel, T., Moulinier, L., Muller, A., Muller, J., *et al.* (2003). PipeAlign: A new toolkit for protein family analysis. Nucleic Acids Res 3*1,* 3829-3832.

PostgreSQL PostgreSQL: The world's most advanced open source database. Available at: http://www.postgresql.org/.

Prasad, T. S. K., Goel, R., Kandasamy, K., Keerthikumar, S., Kumar, S., Mathivanan, S., Telikicherla, D., Raju, R., Shafreen, B., Venugopal, A., *et al.* (2009). Human Protein Reference Database—2009 update. Nucleic Acids Res. 3*7,* D767–D772.

Protein Data Bank (1973). Acta Crystallogr Sect B 2*9,* 1746-1746.

Pruesse, E., Quast, C., Knittel, K., Fuchs, B. M., Ludwig, W., Peplies, J., et Glockner, F. O. (2007). SILVA: a comprehensive online resource for quality checked and aligned ribosomal RNA sequence data compatible with ARB. Nucl. Acids Res. 3*5,* 7188-7196.

Pruitt, K. D., Tatusova, T., et Maglott, D. R. (2007). NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res. 3*5,* D61–D65.

Ranea, J. A. G., Yeats, C., Grant, A., et Orengo, C. A. (2007). Predicting protein function with hierarchical phylogenetic profiles: the Gene3D Phylo-Tuner method applied to eukaryotic genomes. PLoS Comput. Biol 3*,* e237.

Reeves, G. A., Eilbeck, K., Magrane, M., O'Donovan, C., Montecchi-Palazzi, L., Harris, M. A., Orchard, S., Jimenez, R. C., Prlic, A., Hubbard, T. J. P., *et al.* (2008). The Protein Feature Ontology: a tool for the unification of protein feature annotations. Bioinformatics 2*4,* 2767-2772.

Remm, M., Storm, C. E., et Sonnhammer, E. L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons1. Journal of Molecular Biology 3*14,* 1041-1052.

Reyrat, J. M., et Kahn, D. (2001). Mycobacterium smegmatis: an absurd model for tuberculosis? Trends Microbiol 9*,* 472-474.

Reysenbach, A. (2001). Thermophiles : biodiversity, ecology, and evolution (New York: Kluwer Academic).

Richard, G., Kerrest, A., et Dujon, B. (2008). Comparative genomics and molecular dynamics of DNA repeats in eukaryotes. Microbiol. Mol. Biol. Rev 7*2,* 686-727.

Richterich, P. (1998). Estimation of Errors in “Raw” DNA Sequences: A Validation Study. Genome Res. 8*,* 251–259.

Roberts, L. (1991). GRAIL seeks out genes buried in DNA sequence. Science 2*54,* 805.

Ronaghi, M., Uhln, M., et Nyrn, A. P. (1998). DNA SEQUENCING: A Sequencing Method Based on Real-Time Pyrophosphate. Science 2*81,* 363-365.

Rost, B. (2002). Enzyme function less conserved than anticipated. J. Mol. Biol 3*18,* 595-608.

Rubin, D. L., Noy, N. F., et Musen, M. A. (2007). Protégé: a tool for managing and using terminology in radiology applications. J Digit Imaging 2*0 Suppl 1,* 34-46.

Rusk, N. (2009). Cheap third-generation sequencing. Nat Meth 6*,* 244.

Salwinski, L., Miller, C. S., Smith, A. J., Pettit, F. K., Bowie, J. U., et Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res. 3*2,* D449–D451.

SANGER, F., et THOMPSON, E. O. P. (1953a). The amino-acid sequence in the glycyl chain of insulin. I. The identification of lower peptides from partial hydrolysates. Biochem. J 5*3,* 353-366.

SANGER, F., et THOMPSON, E. O. P. (1953b). The amino-acid sequence in the glycyl chain of insulin. II. The investigation of peptides from enzymic hydrolysates. Biochem. J 5*3,* 366-374.

Sanger, F., Air, G. M., Barrell, B. G., Brown, N. L., Coulson, A. R., Fiddes, C. A., Hutchison, C. A., Slocombe, P. M., et Smith, M. (1977). Nucleotide sequence of bacteriophage phi X174 DNA. Nature 2*65,* 687-695.

Sanger, F., Nicklen, S., et Coulson, A. R. (1977). DNA sequencing with chain-terminating inhibitors. Proc. Natl. Acad. Sci. U.S.A 7*4,* 5463-5467.

Schaefer, C. F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., et Buetow, K. H. (2009). PID: the Pathway Interaction Database. Nucleic Acids Res. 3*7,* D674–D679.

Schröder, J., Schröder, H., Puglisi, S. J., Sinha, R., et Schmidt, B. (2009). SHREC: a short-read error correction method. Bioinformatics 2*5,* 2157-2163.

Seeger, M., Kraft, R., Ferrell, K., Bech-Otschir, D., Dumdey, R., Schade, R., Gordon, C., Naumann, M., et Dubiel, W. (1998). A novel protein complex involved in signal transduction possessing similarities to 26S proteasome subunits. FASEB J 1*2,* 469-478.

Selengut, J. D., Haft, D. H., Davidsen, T., Ganapathy, A., Gwinn-Giglio, M., Nelson, W. C., Richter, A. R., et White, O. (2007). TIGRFAMs and Genome Properties: tools for the assignment of molecular function and biological process in prokaryotic genomes. Nucleic Acids Res. 3*5,* D260–D264.

Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., et Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res 1*3,* 2498-2504.

Shimokawa, K., Okamura-Oho, Y., Kurita, T., Frith, M. C., Kawai, J., Carninci, P., et Hayashizaki, Y. (2007). Large-scale clustering of CAGE tag expression data. BMC Bioinformatics 8*,* 161.

Shin, D. S., Didonato, M., Barondeau, D. P., Hura, G. L., Hitomi, C., Berglund, J. A., Getzoff, E. D., Cary, S. C., et Tainer, J. A. (2009). Superoxide dismutase from the eukaryotic thermophile Alvinella pompejana: structures, stability, mechanism, and insights into amyotrophic lateral sclerosis. J. Mol. Biol 3*85,* 1534-1555.

Sjölander, K. (2004). Phylogenomic inference of protein molecular function: advances and challenges. Bioinformatics 2*0,* 170-179.

Smith, A. F. A., Hubley, R., et Green, P. RepeatMasker Available at: http://repeatmasker.org.

Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L. J., Eilbeck, K., Ireland, A., Mungall, C. J., *et al.* (2007). The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat Biotech 2*5,* 1251-1255.

Smith, H. O., et Wilcox, K. W. (1970). A restriction enzyme from Hemophilus influenzae. I. Purification and general properties. J. Mol. Biol 5*1,* 379-391.

Smith, T. F., et Waterman, M. S. (1981). Identification of common molecular subsequences. J. Mol. Biol 1*47,* 195-197.

Staden, R., Beal, K. F., et Bonfield, J. K. (2000). The Staden package, 1998. Methods Mol. Biol 1*32,* 115-130.

Stein, L. D., Mungall, C., Shu, S., Caudy, M., Mangone, M., Day, A., Nickerson, E., Stajich, J. E., Harris, T. W., Arva, A., *et al.* (2002). The generic genome browser: a building block for a model organism system database. Genome Res 1*2,* 1599-1610.

Sterk, P., Kersey, P. J., et Apweiler, R. (2006). Genome Reviews: standardizing content and representation of information about complete genomes. OMICS 1*0,* 114-118.

Stewart, A. C., Osborne, B., et Read, T. D. (2009). DIYA: a bacterial annotation pipeline for any genomics lab. Bioinformatics. 2*5,* 962–963.

Tang, Z., Choi, J., Hemmerich, C., Sarangi, A., Colbourne, J. K., et Dong, Q. (2009). ESTPiper – a web-based analysis pipeline for expressed sequence tags. BMC Genomics. 1*0,* 174.

Tatusov, R. L., Fedorova, N. D., Jackson, J. D., Jacobs, A. R., Kiryutin, B., Koonin, E. V., Krylov, D. M., Mazumder, R., Mekhedov, S. L., Nikolskaya, A. N., *et al.* (2003). The COG database: an updated version includes eukaryotes. BMC Bioinformatics. 4*,* 41.

Taylor, C. F., Field, D., Sansone, S., Aerts, J., Apweiler, R., Ashburner, M., Ball, C. A., Binz, P., Bogue, M., Booth, T., *et al.* (2008). Promoting coherent minimum reporting guidelines for biological and biomedical investigations: the MIBBI project. Nat. Biotechnol 2*6,* 889-896.

Ter-Hovhannisyan, V., Lomsadze, A., Chernoff, Y. O., et Borodovsky, M. (2008). Gene prediction in novel fungal genomes using an ab initio algorithm with unsupervised training. Genome Res 1*8,* 1979-1990.

The UniProt Consortium (2009). The Universal Protein Resource (UniProt) 2009. Nucleic Acids Res. 3*7,* D169–D174.

Thompson, J. D., Higgins, D. G., et Gibson, T. J. (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 2*2,* 4673-4680.

Thompson, J. D., Plewniak, F., et Poch, O. (1999). A comprehensive comparison of multiple sequence alignment programs. Nucleic Acids Res 2*7,* 2682-2690.

Thompson, J. D., Plewniak, F., Ripp, R., Thierry, J. C., et Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. J. Mol. Biol 3*14,* 937-951.

Thompson, J. D., Plewniak, F., Thierry, J., et Poch, O. (2000). DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res 2*8,* 2919-2926.

Thompson, J. D., Thierry, J. C., et Poch, O. (2003). RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics 1*9,* 1155-1161.

Thompson, J. D., Muller, A., Waterhouse, A., Procter, J., Barton, G. J., Plewniak, F., et Poch, O. (2006). MACSIMS: multiple alignment of complete sequences information management system. BMC Bioinformatics 7*,* 318.

Thompson, J. D., Prigent, V., et Poch, O. (2004). LEON: multiple aLignment Evaluation Of Neighbours. Nucleic Acids Res 3*2,* 1298-1307.

Thompson, J. D., Holbrook, S. R., Katoh, K., Koehl, P., Moras, D., Westhof, E., et Poch, O. (2005). MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Res. 3*3,* 4164–4171.

Tian, W., et Skolnick, J. (2003). How well is enzyme function conserved as a function of pairwise sequence identity? J. Mol. Biol 3*33,* 863-882.

Tweedie, S., Ashburner, M., Falls, K., Leyland, P., McQuilton, P., Marygold, S., Millburn, G., Osumi-Sutherland, D., Schroeder, A., Seal, R., *et al.* (2009). FlyBase: enhancing Drosophila Gene Ontology annotations. Nucl. Acids Res. 3*7,* D555-559.

Vallenet, D., Labarre, L., Rouy, Z., Barbe, V., Bocs, S., Cruveiller, S., Lajus, A., Pascal, G., Scarpelli, C., et Medigue, C. (2006). MaGe: a microbial genome annotation system supported by synteny results. Nucl. Acids Res. 3*4,* 53-65.

Velculescu, V. E., Zhang, L., Vogelstein, B., et Kinzler, K. W. (1995). Serial analysis of gene expression. Science 2*70,* 484-487.

Venter, J. C., Adams, M. D., Myers, E. W., Li, P. W., Mural, R. J., Sutton, G. G., Smith, H. O., Yandell, M., Evans, C. A., Holt, R. A., *et al.* (2001). The sequence of the human genome. Science 2*91,* 1304-1351.

Walsh, S., et Barrell, B. (1996). The Saccharomyces cerevisiae genome on the World Wide Web. Trends Genet 1*2,* 276-277.

Warren, A. S., et Setubal, J. C. (2009). The Genome Reverse Compiler: an explorative annotation tool. BMC Bioinformatics. 1*0,* 35.

Wasmuth, J. D., et Blaxter, M. L. (2004). prot4EST: translating expressed sequence tags from neglected genomes. BMC Bioinformatics 5*,* 187.

Waterhouse, A. M., Procter, J. B., Martin, D. M. A., Clamp, M., et Barton, G. J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics 2*5,* 1189-1191.

Waterston, R. H., Lindblad-Toh, K., Birney, E., Rogers, J., Abril, J. F., Agarwal, P., Agarwala, R., Ainscough, R., Alexandersson, M., An, P., *et al.* (2002). Initial sequencing and comparative analysis of the mouse genome. Nature 4*20,* 520-562.

WATSON, J. D., et CRICK, F. H. (1953). Molecular structure of nucleic acids; a structure for deoxyribose nucleic acid. Nature 1*71,* 737-738.

Webb, E., et International Union of Biochemistry and Molecular Biology. (1992). Enzyme nomenclature 1992 : recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the nomenclature and classification of enzymes (San Diego: Published for the International Union of Biochemistry and Molecular Biology by Academic Press).

Wei, C., et Brent, M. R. (2006). Using ESTs to improve the accuracy of de novo gene prediction. BMC Bioinformatics 7*,* 327.

Wheeler, D. A., Srinivasan, M., Egholm, M., Shen, Y., Chen, L., McGuire, A., He, W., Chen, Y., Makhijani, V., Roth, G. T., *et al.* (2008). The complete genome of an individual by massively parallel DNA sequencing. Nature 4*52,* 872-876.

Wicker, N., Perrin, G. R., Thierry, J. C., et Poch, O. (2001). Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol. Biol. Evol 1*8,* 1435-1441.

Wicker, N., Dembele, D., Raffelsberger, W., et Poch, O. (2002). Density of points clustering, application to transcriptomic data analysis. Nucleic Acids Res 3*0,* 3992-4000.

Woese, C. R. (2000). Interpreting the universal phylogenetic tree. Proc Natl Acad Sci U S A. 9*7,* 8392–8396.

Yu, C., Zavaljevski, N., Desai, V., Johnson, S., Stevens, F. J., et Reifman, J. (2008). The development of PIPA: an integrated and automated pipeline for genome-wide protein function annotation. BMC Bioinformatics 9*,* 52.

Résumé

Les travaux présentés dans cette thèse s'inscrivent dans le cadre de la génomique à haut-débit, et tout particulièrement l'analyse de séquences d'*Expressed Sequence Tags* (EST).

Les efforts mis en œuvre ont abouti à la réalisation d'une suite logicielle permettant de gérer une cascade de traitements modulaires. Cette cascade inclut une première phase de prétraitements et d'assemblage visant à améliorer la qualité initiale des EST, qui sont ensuite traduits en séquences protéiques grâce à une combinaison d'approches. La dernière phase consiste en une annotation intégrative des protéines dont l'originalité repose sur l'exploitation du contexte évolutif grâce à l'alignement multiple. La protéine est ensuite replacée au sein de ses réseaux fonctionnels. Les résultats générés sont accessibles via plusieurs interfaces originales de recherche et de visualisation conçues au cours de cette thèse.

Les outils développés ont été utilisés pour analyser différentes collections d'EST et de protéines procaryotes et eucaryotes. Ils ont notamment permis l'exploitation de 100 000 séquences de transcrits d'*Alvinella pompejana*, un Annélide polychète thermotolérant, endémique des sources hydrothermales. Les études comparatives réalisées ont mis en évidence l'importance des gènes impliqués dans l'adaptation au stress oxydatif et à l'hypoxie chez *Alvinella* ainsi qu'un enrichissement des protéines en acides aminés chargés positivement qui pourrait participer à la thermotolérance de ce ver. Enfin, nos travaux ont révélé l'origine ancestrale de nombreux gènes jusqu'à présent considérés comme spécifiques des Deutérostomes, modifiant ainsi notre vision de l'évolution des Métazoaires.

Abstract

This thesis work concerns high-throughput genomics, and more particularly Expressed Sequence Tag (EST) analysis.

The project has led to the development of an EST analysis pipeline capable of managing a suite of analysis modules. The first phase of this pipeline includes pre-processing and assembly of the ESTs to improve their initial quality, and their subsequent translation into protein sequences using a combination of similarity and *ab initio* approaches. The last phase of the pipeline consists of an original integrative annotation of the proteins, based on their evolutionary context thanks to multiple alignments. The proteins are then mapped onto their functional networks. The generated results can be accessed by several dedicated Web querying and visualisation interfaces designed during this thesis.

These developments were used in several studies of prokaryotic and eukaryotic cDNA libraries and proteins. Notably, they enabled the exploitation of 100,000 *Alvinella pompejana* sequences, a thermotolerant polychaete Annelid, endemic to hydrothermal vents. These comparative studies highlighted crucial genes implicated in *Alvinella* oxidative stress and hypoxia adaptation, as well as an enrichment in positively charged amino acids of proteins that could be involved in this worm's thermotolerance. Finally, our work revealed the ancestral origin of several genes previously considered to be Deuterostome specific, thus modifying our vision of Metazoan evolution.