Développement d'une infrastructure d'analyse multi-niveaux pour la découverte des relations entre génotype et phénotype dans les maladies génétiques humaines

Tien Dao Luu

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTE

IGBMC – CNRS UMR 7104 – Inserm U 964

THÈSE présentée par :

Tien Dao LUU

soutenue le : **24 octobre 2012**

pour obtenir le grade de : **Docteur de l'université de Strasbourg**

Discipline/ Spécialité : Bioinformatique

Développement d'une infrastructure d'analyse multi-niveaux pour la découverte des relations entre génotype et phénotype dans les maladies génétiques humaines

THÈSE dirigée par :

M POCH Olivier	IGBMC, Strasbourg

RAPPORTEURS :

Mme DEVIGNES Marie-Dominique	LORIA, Nancy
M DELEAGE Gilbert	IBCP, Lyon

AUTRES MEMBRES DU JURY :

M ZUCKER Jean-Daniel	IRD, Paris/Hanoi
M LACHICHE Nicolas	LSIIT, Illkirch
M NGUYEN Ngoc Hoan	IGBMC, Illkirch

REMERCIEMENTS

Avant tout, je voudrais adresser mes plus sincères remerciements à Gilbert Deléage, Marie-Dominique Devignes, Jean-Daniel Zucker et Nicolas Lachiche pour l'honneur qu'ils me font de juger cette thèse.

Il me tient à cœur de témoigner ici de ma sincère reconnaissance envers Olivier Poch, mon cher directeur de thèse. Merci à toi, Olivier, pour avoir accueilli dans ton laboratoire un étudiant qui ne savait rien sur la biologie et qui ne parle pas un français compréhensible. Merci pour ta confiance, ta patience ainsi que ta tolérance et ta générosité. J'espère que tu continueras à accueillir de nouveaux étudiants vietnamiens à bras ouverts. Pour notre pays, le Vietnam, nous avons besoin de docteurs biens formés dans les meilleurs laboratoires, surtout pour un domaine comme la bioinformatique, très nouveau chez nous.

Ce travail a été réalisé grâce au soutien inconditionnel, allant de la science à la vie, de Nguyen Ngoc Hoan, mon encadrement et mon « grand frère ». Je te remercie du fond du cœur !

Je tiens aussi à remercier le Ministère de l'Education et de la Formation du Vietnam, sponsor financier de « cette aventure ».

Je remercie vivement Anne Friedrich qui m'a présenté clairement SM2PH-db version 1.0, la suite PipeAlign, les banques de données biologiques et les outils bioinformatiques utilisés dans SM2PH-db. Pour une personne ayant un parcours 100% informatique, ces connaissances bioinformatiques étaient indispensables pour me permettre de commencer ma nouvelle aventure il y a 4 ans.

Je tiens à remercier toutes les personnes du laboratoire pour leurs encouragements, leurs conseils et la sympathie dont ils ont fait preuve jour après jour. Je remercie tout particulièrement :

- Julie pour les très précieuses corrections apportées à mes écrits en anglais. J'ai aussi appris beaucoup sur l'alignement et MACSIMS grâce à elle.
- Raymond pour son support technique et pour les corrections apportées à mon français pour ce manuscrit.
- Laetita pour sa disponibilité et son aide concernant STRING et GxDb. Elle est toujours présente quand on a besoin d'aide.
- Nicolas et Luc, avec qui j'ai eu l'occasion de partager le bureau ainsi que leur bonne humeur.
- Wolfgang qui a partagé notre quotidien en essayant de comprendre mon « franco-vietnamien ».
- Odile et sa gentillesse.
- Mr SNP (Jean) pour ses commentaires précieux sur MSV3d et KD4v.

Je vous remercie, Alan, Alexis, Alin, Tao, Vincent, Vinod, Xavier et tout particulièrement Ben, pour votre amitié, les déjeuners ensembles, les explications biologiques et pour les échanges sur tous les « trucs » de la vie! J'ai beaucoup appris sur la vie « internationale » à vos cotés.

Je remercie également Isabelle Audo et Christina Zeitz de l'Institut de la Vision de Paris pour m'avoir fourni les 2 gènes et leurs mutations faux-sens très intéressants sur lesquelles j'ai eu l'occasion de travailler et de constater les avantages et les limites de SM2PH Central.

Un grand merci à Véro pour son support sur PolyPhen-2.

Je n'oublie pas ton rire, Nicodème. Merci d'avoir partagé ton savoir sur les méthodes d'apprentissage automatique.

Merci à Serge aussi, pour la gestion des serveurs et autres aléas informatiques.

Permettez-moi d'écrire ici quelques lignes en vietnamien pour mes parents, ma femme et mes amis vietnamiens.

Con cám ơn ba mẹ về những điều tốt đẹp nhất ba mẹ luôn dành cho con ngay từ lúc con còn ở trong bụng mẹ.

Con cám ơn ba mẹ (vợ). Không giống những gia đình Việt Nam khác có sự khác biệt giữa con ruột và con rể, ba mẹ đã thương con như con ruột. Cuối mỗi cuối tuần gọi điện thoại về Việt Nam, ba lúc nào cũng động viên con cố gắng hoàn thành sứ mệnh học tập. Ba bảo : đừng lo cho ở nhà, con cứ yên tâm mà học tập. Còn mẹ thì dặn đừng gọi, sợ con tốn tiền. Những lần ngắn ngủi con về Việt Nam thăm nhà, mẹ cứ hỏi : con thích ăn gì mẹ nấu cho. Hay hôm rồi mẹ bảo mẹ đi chùa cầu xin cho con hoàn thành tốt đẹp việc học. Đôi khi những câu nói không cần có động từ thương, động từ yêu trong đó, nhưng người nghe vẫn cảm nhận được hoàn toàn sự yêu thương của người nói.

Anh cám ơn em, người vợ nhỏ nhỏ xinh xinh về tình yêu và sự chờ đợi. Nếu không chat, điện thoại với em mỗi cuối tuần, chắc hẳn anh không đủ sức mạnh tinh thần để đi đến ngày hôm nay.

Con (em) cám ơn cô chú Hưng, cô chú Châu, anh Hoan, chị Bình, anh chị Sáu, anh Phú, chị Cương, chị Lan vì đã xem con (em) như con cháu (em út) trong nhà. Những tình cảm này là vô cùng quý giá đối với con (em) khi phải sống và học tập một mình trên đất khách quê người. Một chữ cám ơn bằng tiếng Pháp hay tiếng Việt cũng không đủ nói lên lòng biết ơn của con (em) dành cho các cô chú, anh chị.

Cám ơn anh Khắc, anh Nguyên, anh Lai về những khoảng thời gian cùng nhau chia sẻ Desperados, Pinot blanc, Riesling, Gewurztraminer và nhất là eau de vie và hors d'age !!!

Cám ơn tất cả những người bạn sinh viên đã và đang học tập ở Strasbourg, như Kiên, Khải, vợ chồng Quang, Huy, Linh, Nhung, Toàn, Hà lớn, Hà bé, Thiện, Hiền, vợ chồng Nghĩa Dược, anh Trì, vợ chồng Minh Anh, vợ chồng em Xuân Thủy, Hưng Annecy, Tuấn, Nam, Danh, … cám ơn tất cả về sự thân thiện và tình bằng hữu.

LISTE DES ABREVIATIONS

Å	Angström
AUC	Area Under Curve
BIPS	BioInformatics Platform of Strasbourg
BIRD	Biological Integration and Retrieval Data
BIRD-QL	BIRD Query Language
BMRB	Biological Magnetic Resonance Data Bank
BNL	Brookhaven National Laboratory
CDD	Centre de Données Décrypthon
CRIHAN	Centre de Ressources Informatiques de HAute-Normandie
DMLA	Dégénérescence Maculaire Liée à l'Âge
DSSP	Define Secondary Structure of Proteins
EBI	European Bioinformatics Institute
ECD	Extraction de Connaissances à partir de Données
EMBL	European Molecular Biology Laboratory
GO	Gene Ontology
GWAS	Genome Wide Association Studies
HPO	Human Phenotype Ontology
http	Hypertext Transfer Protocol
IC	Ingénierie des connaissances
Icarus	Interpreter of commands and recursive syntax
IGBMC	Institut de Génétique et de Biologie Moléculaire et Cellulaire
ILP	Inductive Logic Programming
KD4v	Comprehensible Knowledge Discovery System For Missense Variants
KDD	Knowledge Discovery in Databases
KEGG	Kyoto Encyclopedia of Genes and Genomes
LBGI	Laboratoire de Bioinformatique et Génomique Intégratives
LMS	Local Maximum Segments

LEON	multiple aLignment Evaluation Of Neighbours
LGO	Gene Ontology log-odds score
LORIA	Laboratoire Lorrain de Recherche en Informatique
LOVD	Leiden Open source Variation Database
LSDB	Locus-Specific DataBase
NCBI	National Center for Biotechnology Information
NHGRI	National Human Genome Research Institute
NorMD	Normalized Mean Distance
MACS	Multiple Alignment of Complete Sequences
MACSIMS	Multiple Alignment of Complete Sequences Information Management System
MAO	Multiple Alignment Ontology
MSF	Multiple Sequence Format
MSV3d	Database of human missense variants mapped to 3D protein structures
OMIM	Online Mendelian Inheritance in Man
PDB	Protein Data Bank
PDBe	PDB in Europe
PDBj	PDB of Japan
PIR	Protein Information Resource
PLI	Programmation Logique Inductive
RASCAL	Rapid Scanning and Correction of Alignment errors
RCSB	Research Collaboratory for Structural Bioinformatics
RefSeq	Reference Sequence database
RMSD	Root Mean Square Distance
ROC	Receiver Operating Characteristics
SCOP	Structural Classification of Proteins
SIB	Swiss Institute of Bioinformatics
SIFT	Sorting Intolerant From Tolerant
SM2PH	de la Mutation Structurale au Phénotype des Pathologies Humaines
SNP	Single Nucleotide Polymorphism
SOAP	Simple Object Access Protocol

SQL	Structured Query Language
SRS	Sequence Retrieval System
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
SVILP	Support Vector Inductive Logic Programming
SVM	Support Vector Machine
Tcl	Tool Command Language
Tk	ToolKit
UniMES	UniProt Metagenomic and Environmental Sequences
UniParc	UniProt Archive
UniProt	Universal Protein resource
UniProtKB	UniProt Knowledgebase
UniRef	UniProt Reference clusters
UMD	Universal Mutation Database
URI	Uniform Resource Identifier
URL	Uniform Resource Locator
wwPDB	Worldwide PDB
XML	eXtensible Markup Language
XGMML	eXtensible Graph Markup and Modeling Language

TABLE DES MATIERES

TABLE DES FIGURES

TABLE DES TABLEAUX

INTRODUCTION GENERALE

Le début du 21ième siècle a été marqué par la mise à disposition de la séquence complète du génome humain, après une dizaine d'années d'efforts de la communauté internationale. Ce déchiffrage complet du génome humain et l'introduction des biotechnologies à haut débit ont montré, entre autres, que les liens biologiques entre modifications du génome et maladies humaines qui en découlent sont extrêmement complexes.

Suivant en cela la révolution technologique qui traverse la biologie dans son ensemble, la médecine moderne a vu l'étude des liens génotype-phénotype s'inscrire également dans le flux croissant de quantités énormes de données. Au delà des aspects de volumétrie, ces données, qui doublent chaque année, sont très hétérogènes et incluent aussi bien des données comportementales ou environnementales du patient, des données génétiques (contexte génomique, statut d'expression des gènes, activité des produits d'un gène, leurs interactions, leurs modifications...) sans omettre des informations concernant les processus, protocoles ou traitements utilisés lors de la création des données. Ces nouvelles conditions ont abouti à des taux de production et d'hétérogénéité des données qui dépassent largement les capacités d'analyse et d'expertise humaines ainsi que les possibilités de traitement des plus puissants ordinateurs. Dès lors, de nouveaux concepts et développements sont nécessaires pour, d'une part, assurer le déploiement d'un système capable d'intégrer et d'analyser de gros volumes de données hétérogènes et d'autre part, inférer de nouvelles hypothèses et théories en présentant ces connaissances aux biologistes/médecins de manière fluide et intuitive. Cette association entre méthodes biotechnologiques à haut débit, stratégies d'analyse des informations et algorithmes de découverte de connaissances a contribué à l'émergence d'un nouveau domaine : la biologie intégrative.

En termes de recherche biomédicale, pour atteindre une meilleure compréhension des maladies génétiques humaines et la mise en œuvre de diagnostics ou de solutions thérapeutiques efficaces, un enjeu primordial est la capacité à comprendre et prédire les effets des variations génétiques sur le phénotype de l'individu les portant. Idéalement, cela implique la prise en compte de multiples aspects provenant de la génétique, de la physiopathologie, de la progression ou de la réponse thérapeutique qui nécessitent tous des solutions bioinformatiques originales souvent regroupées sous le terme de bioinformatique translationnelle.

Dans ce cadre, 2 difficultés majeures ont été identifiées :

- « Il s'agit de déterminer par quels mécanismes un gène défectueux engendre une protéine qui fonctionne mal et comment cette dernière perturbe la vie normale de la cellule », a souligné Stéphane Roques, responsable du programme Décrypthon à l'AFM.
- il s'agit de distinguer les mutations délétères, à l'origine de modifications phénotypiques, des variations neutres qui seront sans conséquence directe.

Dans cette nouvelle orientation vers une bioinformatique translationnelle, j'ai été amené à m'impliquer fortement dans la mise en place d'une nouvelle infrastructure dans la continuité et l'évolution de développements antérieurs réalisés au laboratoire autour de SM2PH. Le projet SM2PH (de la Mutation Structurale au Phénotype des Pathologies Humaines), un projet pilote

du programme Décrypthon (http://www.decrypthon.fr), s'est inscrit dans ce contexte scientifique. L'objectif initial de ce projet était le déploiement et la mise à disposition d'un prototype d'infrastructure informatique à même de faciliter la compréhension de la relation entre le génotype et le phénotype pour l'ensemble des gènes codant pour les protéines impliquées dans des maladies génétiques humaines, notamment les maladies neuromusculaires. Dans ce cadre, SM2PH-db version 1.0 (Friedrich et al., 2010) a été développée principalement par Anne Friedrich (de l'IGBMC de Strasbourg) et Nicolas Garnier (de l'Institut de Biologie et de Chimie des Protéines de Lyon). Ils ont implémenté l'ensemble du protocole qui conduit de la protéine d'intérêt à l'obtention des 2 alignements multiples annotés, en passant par la sélection des empreintes structurales et l'extraction des alignements nécessaires à la modélisation par homologie. Cette version regroupait des données relatives à 2 249 protéines impliquées dans des maladies monogéniques humaines (au mois d'août 2009). Cette version a représenté le socle de l'infrastructure globale visant à mieux comprendre les relations génotype-phénotype.

Cependant, outre le manque de certaines données de génomique, protéomique, interactomique ou métabolomique, l'infrastructure développée ne disposait pas de modèles d'extraction de connaissance susceptibles d'être appliqués aux diverses données et de contribuer ainsi à la création de nouvelles découvertes biologiques. Ces constats nous ont amenés à développer une infrastructure plus en adéquation avec les problématiques modernes d'étude du lien génotype-phénotype. L'architecture générale de cette infrastructure est présentée schématiquement dans la Figure 1 qui met en exergue les 2 axes permettant d'aller des aspects gènes/protéines aux mutations.

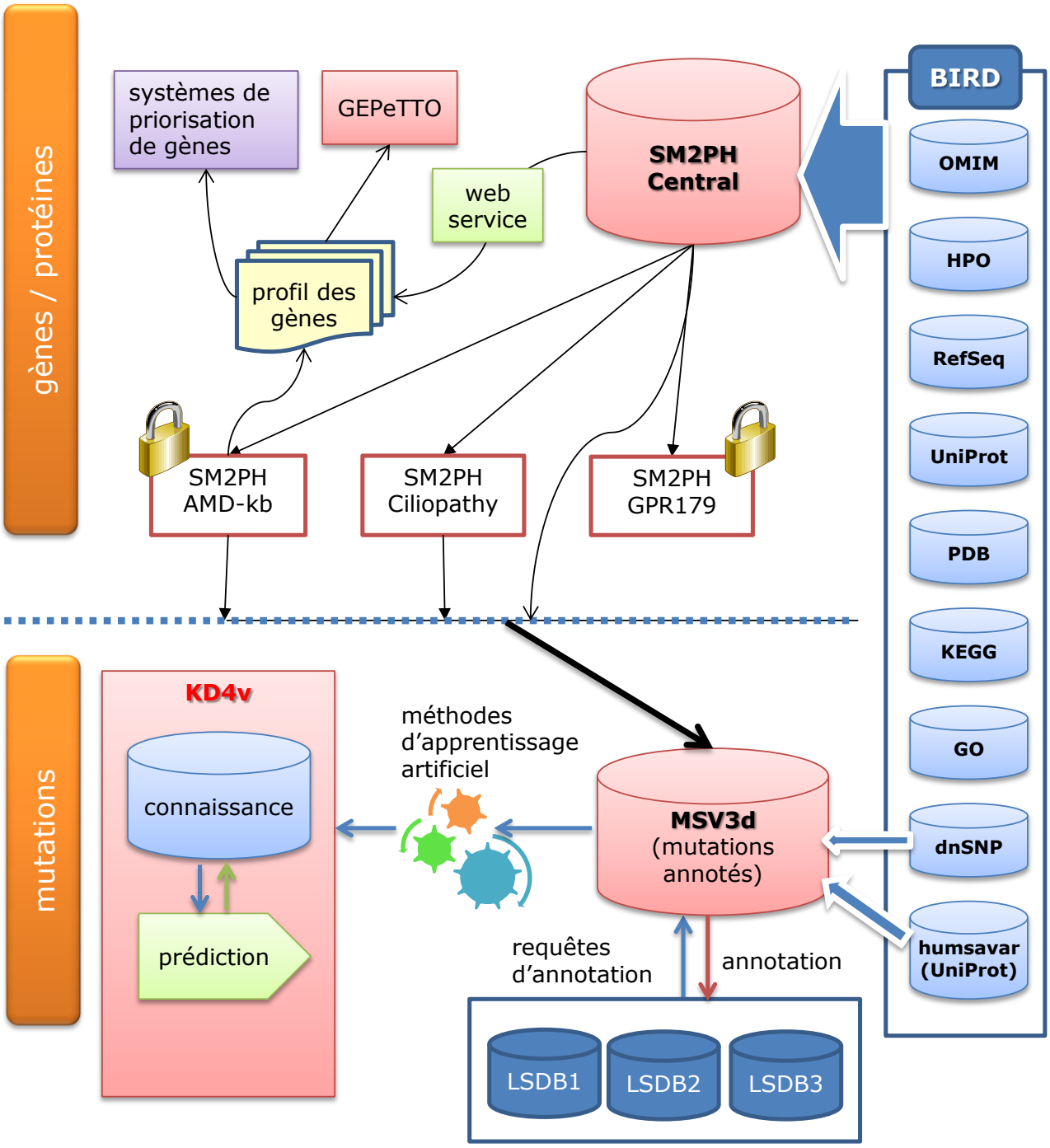


Figure 1. Architecture globale de notre infrastructure. Les cylindres symbolisent des bases de données, les cadenas des bases de données très confidentielles et les flèches indiquent les distributions de données.

La première partie de notre infrastructure est focalisée sur l'axe gène/protéine et concerne le développement d'un système à même de faciliter la compréhension des relations qui existent entre la séquence de la protéine, son évolution, sa structure 3D, sa localisation à l'intérieur des réseaux biologiques ou d'interaction, son expression différentielle dans divers tissus humains et les pathologies qui y sont associées. Pour cette première partie, nous avons développé et implémenté le système SM2PH Central. Un système unique englobant l'ensemble des données et informations précédemment décrites pour toutes les protéines humaines (~20 199 protéines) et intégrant tous les logiciels (~ 20 programmes) nécessaires à l'annotation intégrative automatique de séquences protéiques. SM2PH Central récupère des données à partir de BIRD (*Biological Integration and Retrieval of Data*), un système de gestion et

d'interrogation de données hétérogènes. Les informations collectées ou générées par SM2PH Central sont disponibles dans un format structuré permettant une exploitation automatique à haut débit par ordinateur et sont aussi accessibles aux biologistes pour une analyse visuelle à travers une interface web simple et conviviale. D'autre part, SM2PH Central fournit également un environnement logiciel permettant la création de bases de données spécialisées et consacrées à l'étude de maladies ou de gènes spécifiques *via* de nouvelles instances (sous-systèmes, telles que SM2PH-Ciliopathy, SM2PH-GPR179, SM2PH-AMD-kb). A titre d'exemple, on peut noter que SM2PH-AMD-kb (http://decrypthon.igbmc.fr/amdkb/) a été utilisée pour l'étude de la DMLA (Dégénérescence Maculaire Liée à l'Âge) afin de fournir un accès à l'ensemble des informations nécessaires à la priorisation de gènes candidats de la DMLA. Nous avons également développé un prototype afin de prioriser les gènes d'une liste de candidats potentiellement pathogènes en exploitant les informations disponibles dans SM2PH Central.

La seconde partie de notre infrastructure est centrée sur les mutations faux-sens. Nous avons développé MSV3d, une base de données et un site web dédiés à l'analyse à haut débit des mutations structurales impliquées dans les maladies génétiques humaines. Grâce à BIRD, MSV3d intègre l'ensemble des 445 574 mutations faux-sens connues des bases dbSNP et UniProt. Ces mutations sont séparées en 2 grandes catégories : les mutations délétères qui seront à l'origine de modifications phénotypiques et les mutations neutres qui seront sans conséquence. Une suite de programmes (~ 10 programmes) associés aux données provenant de SM2PH Central est lancée automatiquement pour annoter ces mutations. Afin d'étudier les liens entre mutations et phénotypes, mes efforts se sont concentrés sur le développement d'un environnement spécifique dédié à l'exploitation de MSV3d au moyen de méthodes d'extraction de connaissances. Cet environnement nommé KD4v a été développé en s'appuyant sur les données disponibles dans MSV3d et en utilisant des algorithmes de Programmation Logique Inductive (PLI). La base de connaissances de KD4v induite par la PLI contient des règles exploitables par un humain ou un ordinateur et des facteurs prédictifs caractérisant les mutations neutres ou délétères. KD4v permet aux biologistes d'exploiter cette base de règles et de prédire si une nouvelle mutation est neutre ou délétère. Si elle est délétère, KD4v fournit aussi l'explication qui réside derrière le jeu de règles.

Une application biologique a été réalisée dans la cadre de ma thèse. Nous avons étudié la cécité nocturne en utilisant SM2PH Central, en combinaison avec le service d'annotation de MSV3d et la méthode de prédiction KD4v pour analyser le gène GPR179 et valider l'impact phénotypique de ses 2 mutations nouvellement identifiées.

Ce manuscrit est le fruit de mon travail de thèse qui a été effectuée au Laboratoire de Bioinformatique et Génomique Intégratives de l'IGBMC, sous la direction d'Olivier Poch et la supervision de Nguyen Ngoc Hoan de Janvier 2009 à Octobre 2012. Durant cette période j'ai pu explorer de multiples aspects de la biologie intégrative et de l'ingénierie des connaissances. Ce manuscrit est organisé en 5 parties : Introduction, Données et Méthodes, Systèmes d'information, Découverte de connaissances et Application.

L'introduction présente succinctement les connaissances mises à contribution durant la thèse. Le Chapitre 1 décrit la relation liant le génotype au phénotype, le Chapitre 2 décrit, dans ses grandes lignes, le contexte bioinformatique en mettant l'accent sur les domaines de la biologie intégrative et de l'ingénierie des connaissances.

La deuxième partie présente rapidement le matériel, en l'occurrence essentiellement les données, et les méthodes utilisés durant ma thèse. Le Chapitre 3 concerne les bases de données et les méthodes générales, puis, dans le Chapitre 4, les principes de la méthode d'acquisition automatique de connaissances ou Programmation Logique Inductive (PLI), sont

décrits en englobant les notations et définitions de la logique des prédicats, l'algorithme générique de PLI, les différents systèmes PLI existant dans la littérature ainsi que les utilisations des PLI dans diverses applications en biologie surtout dans les problèmes de biologie structurelle.

La troisième partie comprend la présentation des 2 systèmes d'informations que j'ai dévéloppés lors de ce travail de thèse. Il s'agit du système SM2PH Central consacré à l'analyse intégrative des protéines humaines (Chapitre 5) et du système MSV3d concernant la caractérisation des mutations faux-sens (Chapitre 6). L'article décrivant MSV3d, publié dans le journal *Database (Oxford)*, est joint au Chapitre 6.

La quatrième partie détaille les résultats obtenus dans le cadre de l'extraction de connaissances avec 2 chapitres. Le Chapitre 7 regroupe la présentation du système KD4v et la publication réalisée dans le journal *Nucleic Acids Research*. Enfin, le Chapitre 8, qui peut être considéré comme une transition entre les développements réalisés et les perspectives de nos travaux, décrit notre prototype de la priorisation de gènes.

L'application de nos développements dans l'étude de l'impact structural et phénotypique de nouvelles mutations du gène GPR179 impliqué dans la cécité nocturne est décrite dans le chapitre 9 de la cinquième partie qui intègre le papier publié dans la revue *American Journal of Human Genetics*.

Enfin, dans la partie Conclusions et Perspectives, je présente la synthèse des principales contributions de mes travaux et conclus en proposant quelques pistes de recherche particulièrement intéressantes qui découlent des résultats présentés.

PREMIERE PARTIE : INTRODUCTION

La première partie recouvre une introduction qui présente succinctement les connaissances mises à contribution durant la thèse. Le Chapitre 1 décrit la relation liant le génotype au phénotype, le Chapitre 2 décrit, dans ses grandes lignes, le contexte bioinformatique global en mettant l'accent sur les domaines de la biologie intégrative et de l'ingénierie des connaissances.

CHAPITRE 1. RELATION GENOTYPE ET PHENOTYPE

Ce premier chapitre décrit la relation génotype-phénotype en mettant l'accent sur les différents niveaux d'information où la notion de mutation est susceptible d'intervenir depuis le gène et son organisation jusqu'aux éléments liés aux familles de patients. L'essentiel de ce chapitre très « biologique » provient de la thèse d'Anne Friedrich (Friedrich, 2007) qui a participé au développement de SM2PH-db version 1.0.

Le génotype est défini comme l'ensemble de l'information génétique d'un individu, présente, pour l'essentiel, dans l'ADN. En vis-à-vis du génotype, on place généralement le phénotype. Le phénotype est l'ensemble des caractéristiques observables ou détectables d'un individu, par exemple la couleur des yeux, de la peau, la forme d'un organe, les conséquences de maladies génétiques.... Il existe une relation complexe entre le génotype, l'environnement et les manifestations phénotypiques (Figure 2), ce qui explique qu'il est extrêmement difficile de mesurer l'influence de chacun.

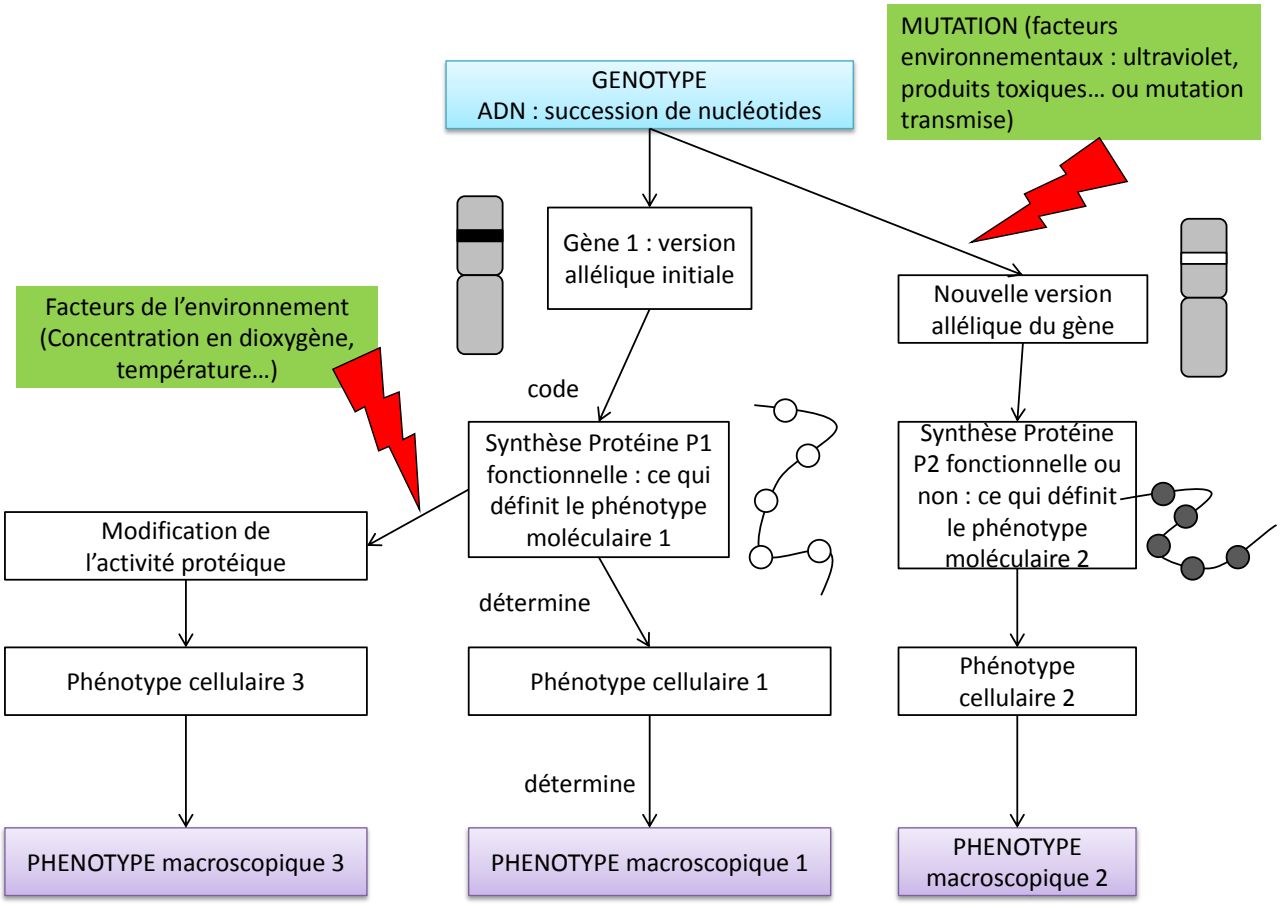


Figure 2. Relations entre génotype, phénotype et environnement. (Adaptée de http://www2.ac-lyon.fr/etab/lycees/lyc-01/bichat/IMG/pdf/Chap_2.pdf).

Dans ce chapitre, nous décrirons dans ses grandes lignes, le génome humain et sa variabilité, pour finalement, nous concentrer sur la description des éléments impliqués dans les maladies génétiques humaines.

1.1 Organisation du génome humain

Le génome est l'ensemble de l'information héréditaire chez un organisme. Chez l'homme, il est composé du génome nucléaire hérité des 2 parents et du génome mitochondrial maternel. Seul

le génome nucléaire suit les règles de la transmission mendélienne classique. Le support matériel du génome est l'ADN. L'ADN est une molécule en double hélice formée de 2 brins complémentaires et anti-parallèles. Chaque brin est un polymère constitué par l'enchaînement d'unités chimiques individuelles appelées nucléotides. Les nucléotides sont construits à partir d'un sucre (désoxyribose), d'un groupe phosphate et de quatre bases nucléotidiques : l'adénine (notée A), la cytosine (notée C), la guanine (notée G) et la thymine (notée T). Ces bases s'assemblent par paires dans la double hélice d'ADN, A:T (2 liaisons hydrogènes) et G:C (3 liaisons hydrogènes). Cette molécule d'ADN s'enroule sur elle même, avec l'aide de protéines d'histones pour se compacter et former les chromosomes (Figure 3).

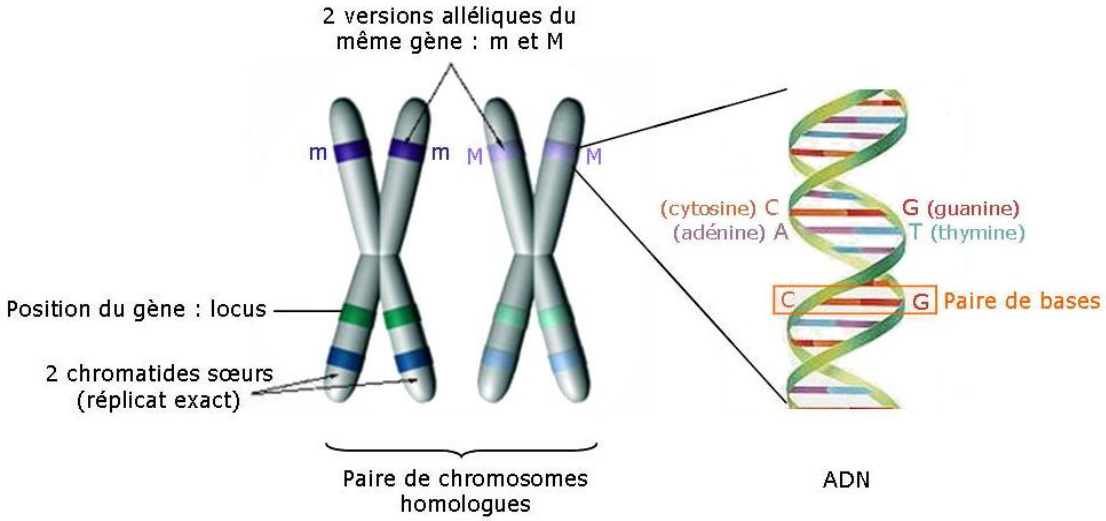


Figure 3. Représentation d'une paire de chromosomes homologues Adaptée de (Friedrich, 2007).

Le génome humain est constitué de 46 chromosomes répartis en 23 paires : 22 paires d'autosomes et une paire de gonosomes ou chromosomes sexuels (XX chez la femme, XY chez l'homme). Les 2 chromosomes d'une paire sont dits homologues, à l'exception de la paire XY.

La définition classique, et simpliste, d'un gène s'est longtemps référée à une séquence d'ADN qui contient les informations nécessaires à la détermination d'un caractère particulier. Chaque gène est donc un fragment de la molécule d'ADN dans un ordre précis constituant le chromosome.

1.1.1 Architecture des gènes

La structure interne d'un gène protéique humain est relativement complexe. Un gène est composé à la fois d'exons et d'introns (**Figure 4**). Le gène est transcrit de sa forme ADN à sa forme de transcrit primaire puis, pour aboutir au transcrit mature ou ARN messager (ARNm), l'étape d'épissage permet de supprimer les introns, d'ajouter en 5' du transcrit, une coiffe et en 3' du transcrit, une queue poly–adényl sur un site spécialisé (site de polyadénylation). L'ARNm, composé uniquement des exons, contient 2 zones non codantes (5' et 3' UTR (pour *UnTranslated Region*) et la région codante appelée CDS (pour *CoDing Sequence*) qui est une suite de codons (l'enchaînement des 3 nucléotides codant pour un acide aminé) et est bornée par le codon initiateur de la traduction (Ci) et le codon Stop ou d'arrêt de la traduction (Cs). La fraction codante du génome humain ne représente qu'une part très réduite de ce dernier, estimée à 1,2% du génome euchromatique (ensemble des régions peu condensées de la chromatine, actives pour la transcription) (IGHSC, 2004).

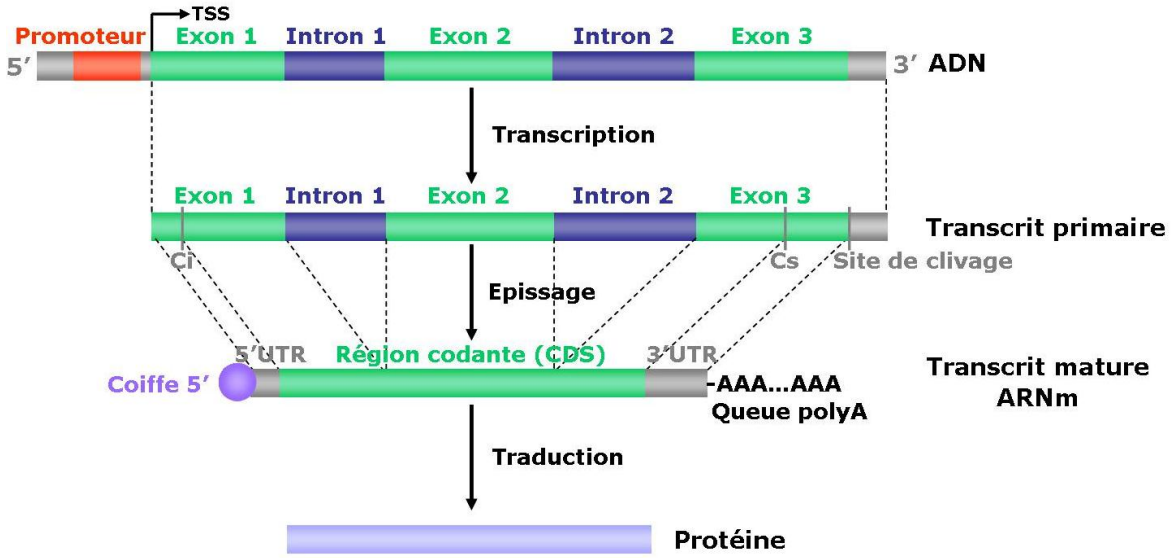


Figure 4. Structure du gène humain : de l'ADN génomique à la protéine (Friedrich, 2007).

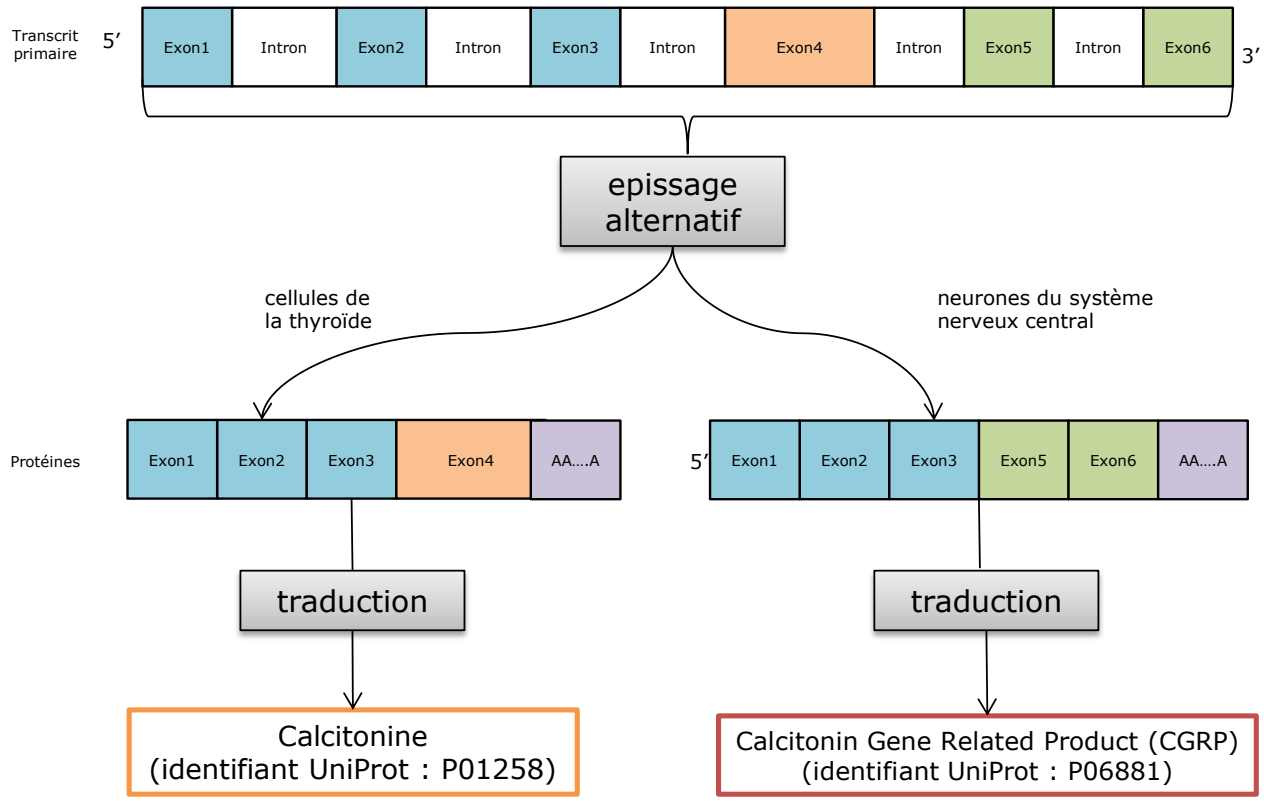


Figure 5. Epissage alternatif du gène CALCA. Le gène humain CALCA (Calcitonin) est situé sur le chromosome 11 et comprend 6 exons et 5 introns. Il code : i) pour l'hormone calcitonine qui s'exprime dans certaines cellules de la thyroïde et provient de l'union des exons 1 à 4, et ii) pour un neuromédiateur, le CGRP qui s'exprime dans de nombreux neurones du système nerveux central et périphérique et provient de l'union des exons 1, 2, 3, 5 et 6 (Smith et al., 1989).

Un degré supérieur de complexité est ajouté aux gènes de nombreux eucaryotes par l'épissage alternatif qui consiste à manipuler, pour un gène donné, son répertoire d'exons afin d'obtenir des ARNm différents. Ce processus permet d'augmenter le nombre de protéines codées par un même gène et donc, potentiellement, le nombre de fonctions d'un gène (Figure 5). Il peut avoir lieu dans le CDS et entraîner la production de protéines différentes, mais également dans

les régions 5' et 3' UTR avec, comme conséquence possible, une stabilité modifiée du transcrit mature et une modification du niveau de production d'une protéine par ailleurs identique.

Ce phénomène d'épissage alternatif est loin d'être une exception mais plutôt une règle générale puisque, selon diverses estimations, de 50% à 80% des gènes humains possèdent un, ou plusieurs, variant d'épissage (Kampa et al., 2004). Cet ensemble contribue ainsi à l'augmentation du répertoire des possibilités d'un organisme, sans avoir besoin de multiplier le nombre de gènes.

Chez l'humain, la structure des gènes est très variable tant par leur taille ou leur organisation, que par le nombre ou la longueur des introns et des exons qui les constituent (Tableau 1). Les introns ont une taille très variable de quelques paires de bases (pb) à environ 1 million, la moyenne se situant à 5746 pb. Les exons ont des tailles plus modestes comprises entre 2 pb et 22 kb avec pour moyenne 314 pb. Le nombre d'exons peut aller de 1 pour des gènes tels que ceux codant les histones jusqu'à 363 pour le gène de la titine.

	Total	**Minimum**	**Maximum**	**Moyenne**
Génome humain : 21 224 gènes codants identifiés				
Exons	214 934	2 pb	22 753 pb	312 pb
Introns	198 357	2 pb	1 160 411 pb	6,392 pb
Transcrit	39 932	33 pb	101 520 pb	2,946 pb
Nombre d'exons/gène	-	0	316	8
Nombre d'exons/transcrit	-	1	312	9

Tableau 1. Quelques statistiques du génome humain. Le nombre de gènes identifiés a été extrait du site d'Ensembl : http://www.ensembl.org/Homo_sapiens/Info/StatsTable?db=core. Les statistiques sont issues du NCBI (assemblage 37, version 3) : http://www.ncbi.nlm.nih.gov/mapview/stats/BuildStats.cgi?taxid=9606&build=37&ver=3

1.1.2 Expression des gènes humains

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

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

1.1.3 Architecture des protéines

Si les gènes représentent l'unité informationnelle génétique, les protéines représentent les unités fonctionnelles majeures. Ces dernières peuvent être classées selon leur fonction biologique et incluent :

- les enzymes, responsables de la catalyse des milliers de réactions chimiques au cœur des cellules;
- les protéines de structure comme la tubuline, la kératine ou le collagène;
- les protéines de transport à l'exemple de l'hémoglobine;
- les protéines de régulation telles que les facteurs de transcription;
- les molécules de signalement comme les hormones et leurs récepteurs;
- les protéines du système immunitaire;
- les protéines assurant un rôle mécanique telles que l'actine et la myosine.

Une protéine est un polymère d'acides aminés reliés entre eux par une liaison peptidique. A l'exception des protéines intrinsèquement non structurées qui ne possèdent pas de repliement particulier, pour être fonctionnelle, une protéine doit adopter une conformation 3D précise (Dyson and Wright, 2005).

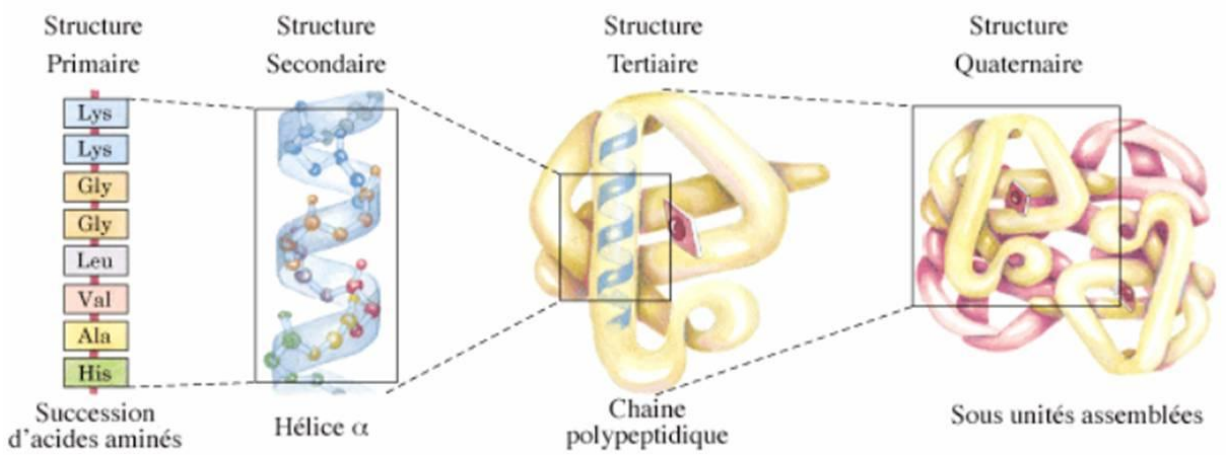


Figure 6. Repliement des protéines selon les 4 niveaux de structuration. Adaptée de (Friedrich, 2007)

On peut distinguer 4 niveaux d'organisation dans une protéine (**Figure 6**). La séquence des acides aminés liés dans la chaîne polypeptidique constitue la structure primaire de la protéine. La structure secondaire est un premier niveau de repliement, adopté par des portions de la protéine, résultant d'interactions entre des acides aminés voisins sur la chaîne. 2 motifs de repliement caractéristiques peuvent ainsi se former : les hélices α et les feuillets β, réunis par des boucles ou des demi-tours. La structure tertiaire correspond au repliement de la protéine dans l'espace, à l'agencement des éléments de structures secondaires entre eux et à l'organisation spatiale des chaînes latérales. Enfin, certaines protéines sont constituées de plusieurs chaînes polypeptidiques, ou sous-unités, qui s'arrangent spatialement en une structure quaternaire.

1.1.4 Réseau biologique

Chaque gène n'est pas isolé dans la cellule, mais au centre de complexes réseaux biologiques. En comprenant mieux ces réseaux, on peut mieux comprendre la fonction de ce gène et son influence sur le phénotype.

Un réseau biologique est une représentation de la circulation d'un certain type d'information dans la cellule. Il en existe 3 types (Figure 7) :

- Réseau génétique ou de régulation : un gène régule l'expression des autres gènes. En général, les gènes n'interagissent pas physiquement. Une relation entre 2 gènes A et B dans un réseau de ce type, signifie qu'un changement dans l'activité du gène A cause un changement dans l'activité du gène B, ce changement pouvant être le résultat d'une succession de modifications d'activité au niveau des produits associés aux 2 gènes. Par exemple, la relation entre les 2 gènes ESR1 et TFF1 est connue pour des tumeurs mammaires (Townson et al., 2006). Le gène ESR1 code pour un récepteur nucléaire des œstrogènes. TFF1 est un gène induit par les œstrogènes et impliqué dans divers processus biologiques. En l'absence d'expression du gène ESR1, des œstrogènes ne peuvent réguler le niveau d'ARN messager du gène TFF1.

- Réseau d'interaction protéine-protéine : une protéine interagit physiquement avec les autres protéines ou une protéine transmet un signal informatif aux autres protéines. L'identification des interactions entre protéines permet de mieux comprendre leur fonction et de découvrir de nouvelles cibles thérapeutiques pour le développement de médicaments. Par exemple, la protéine p53, codée par le gène TP53, est une cible thérapeutique attractive dans l'étude de cancers humains en raison de sa capacité d'induction de la mort cellulaire par apoptose, et donc de supprimer des cellules cancéreuses. Toutefois, p53 est régulée négativement par la protéine MDM2, codée par le gène MDM2, (Kussie et al., 1996), dont l'amplification dans de nombreux cancers favorise une prolifération incontrôlée (MDM2 : *Double Minute 2 Protein*). L'inhibition de l'interaction p53-MDM2 dans ces cancers représente donc une stratégie intéressante pour activer une apoptose p53-dépendante dans les tumeurs sur-exprimant MDM2 (Mukherjee et al., 2001).

- Réseau métabolique : l'ensemble des réactions chimiques dans une cellule. Ces réactions transforment des métabolites substrats en métabolites produits. Certaines réactions sont spontanées (i.e. les substrats se transforment en produits sans catalyseur), mais la plupart nécessite la présence d'une enzyme pour avoir lieu à une vitesse observable. Le rôle de l'enzyme est d'accélérer la réaction. La connaissance globale du réseau métabolique d'un organisme permet d'étudier les conséquences de la modification de la capacité d'une enzyme en un point du réseau, sur le fonctionnement global.

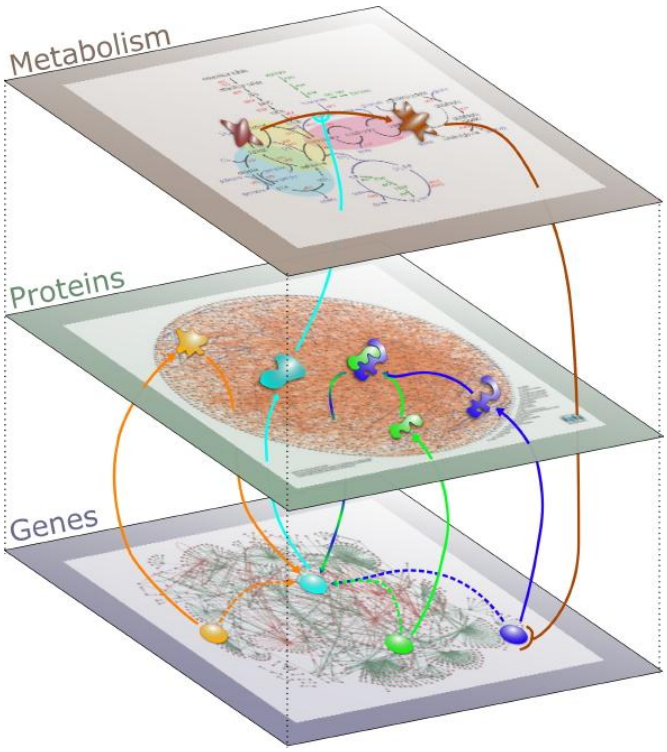


Figure 7. 3 réseaux biologiques. (Adaptée de http://biodev.extra.cea.fr/interoporc/).

1.2 Variabilité génétique

Chaque être humain est unique. En effet, à l'exception des « vrais » jumeaux, chaque individu dispose de son génome propre et l'on considère que le génome de 2 individus non apparentés varie en moyenne de 0,1%, toutes ethnies confondues.

La variabilité génétique est caractérisée par la variabilité allélique, chaque gène pouvant exister sous plusieurs formes. Une variation allélique est traditionnellement décrite comme un polymorphisme si l'on retrouve plus d'un allèle à un certain locus dans la population avec une fréquence supérieure à 1%. Les principaux mécanismes conduisant à la variabilité des génomes sont le brassage génétique.

Deux niveaux de variabilité génétique peuvent être distingués, d'une part, au niveau de réarrangements chromosomiques et donc touchant un grand nombre de gènes et, d'autre part, au niveau de nucléotides de la molécule d'ADN, donc plus local. Ces variations génomiques, qui peuvent apparaître spontanément ou être induites par des facteurs extérieurs, représentent le moteur de l'évolution, mais elles peuvent aussi être associées à l'apparition de maladies génétiques. On peut distinguer les modifications dites germinales, qui affectent les gamètes et sont donc potentiellement transmissibles à la descendance, des modifications somatiques, qui affectent les autres cellules d'un individu et ne sont pas transmissibles d'une génération à l'autre. Ces dernières n'interviennent pas dans le processus de l'évolution.

Nous allons voir dans le paragraphe suivant ces 2 types de variations ainsi que leurs conséquences sur l'expression du génome. Nous nous concentrerons plus particulièrement sur les variations locales qui nous intéressent dans ce travail de thèse.

1.2.1 Réarrangements chromosomiques

Les réarrangements chromosomiques entraînant une anomalie du nombre de chromosomes sont généralement liés à une mauvaise ségrégation des chromosomes au moment de la

première division méiotique à l'origine de la formation des gamètes. En effet, au cours de cette première division cellulaire, une cellule diploïde, contenant 2 copies de chaque chromosome, donne naissance à 2 gamètes haploïdes, comprenant chacun un chromosome de chaque paire. Lorsque les 2 chromosomes d'une même paire migrent vers la même cellule fille, un des gamètes formés aura 2 copies d'un chromosome et son complémentaire aucune : on parle d'aneuploïdie du fait que les gamètes ne contiennent pas le nombre attendu de chromosomes. En cas de fécondation, le zygote formé sera lui aussi aneuploïde, vis-à-vis de quoi la nature est très intolérante. La monosomie n'est pas compatible avec la vie, mis à part quand elle concerne les chromosomes sexuels. Certaines trisomies sont quant à elles viables, mais entraînent l'apparition de symptômes sévères, la plus connue étant la trisomie du chromosome 21, associée au syndrome de Down (Lejeune et al., 1959).

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

Les réarrangements chromosomiques entraînant une anomalie de structure des chromosomes peuvent concerner un ou plusieurs chromosomes et avoir lieu avant une division cellulaire. Les principaux types de remaniements (Figure 8) nécessitent une double cassure du brin d'ADN. Ils seront dits équilibrés s'ils n'entraînent pas la perte de matériel génétique et déséquilibrés dans le cas contraire. De manière générale, les remaniements équilibrés n'ont pas de conséquence sur le sujet porteur, alors que les remaniements déséquilibrés se traduisent par des manifestations cliniques d'autant plus sévères que la perte ou le gain de matériel est plus important.

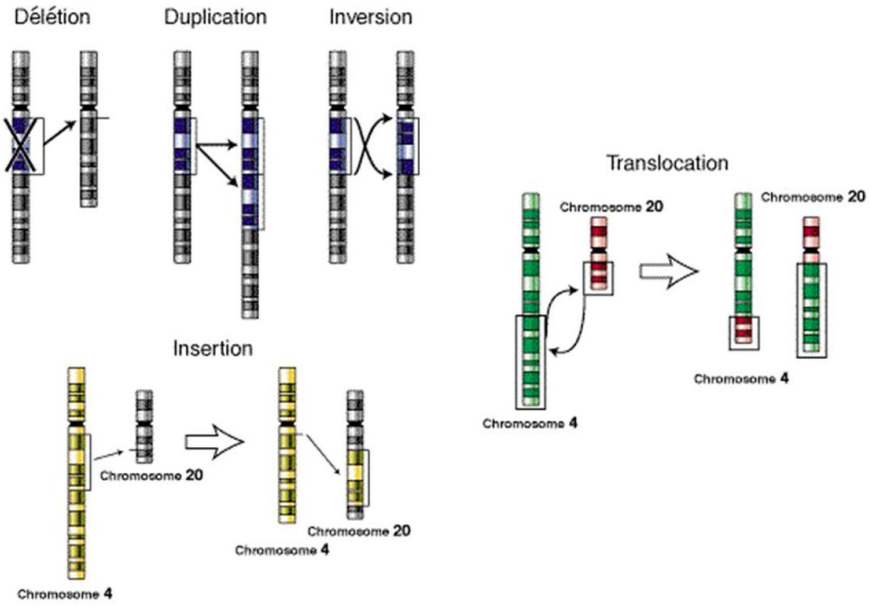


Figure 8. Remaniements chromosomiques entraînant une anomalie de structure. (Adaptée de http://en.wikipedia.org/wiki/Image:Types-of-mutation.png)

1.2.2 Modifications locales au niveau de l'ADN

Les modifications locales au niveau de l'ADN sont de loin les plus importantes quantitativement. On distingue les mutations spontanées issues d'erreurs non corrigées lors de la réplication et les mutations induites faisant intervenir l'action d'un agent dit mutagène.

Celui-ci peut être endogène, comme les radicaux libres, ou exogène, par exemple les rayons ultraviolets (UV).

1.2.2.1 Origine de l'apparition des mutations

Avant chaque division cellulaire, le génome humain doit être copié de manière précise dans son intégralité, ce qui représente un véritable défi. La réplication du génome est principalement assurée par l'ADN polymérase, un complexe enzymatique qui se trouve être d'une remarquable efficacité. Au cours de la réplication, chaque brin parental d'une molécule d'ADN sert de matrice pour la production d'un nouveau brin, lui-même déterminé par la complémentarité des bases. Chaque nouvelle molécule d'ADN double brin est constituée d'un brin parental et d'un brin néoformé (Figure 9) : on parle alors de réplication semi-conservative.

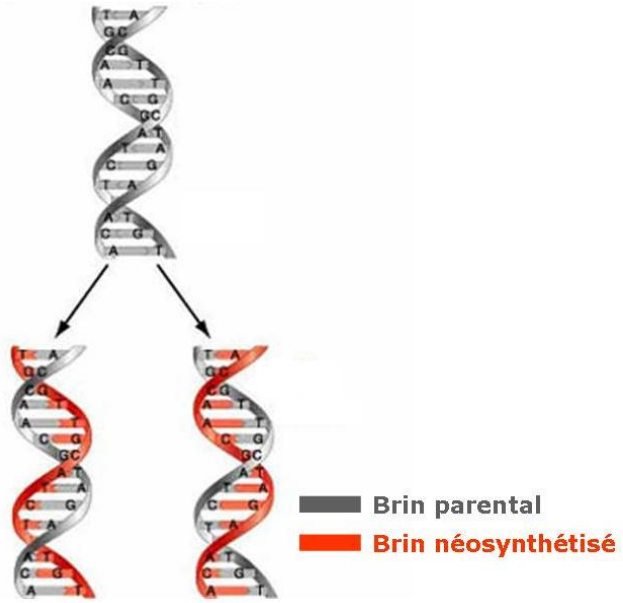


Figure 9. Réplication semi-conservative du génome. L'ADN polymérase se sert de chacun des brins de l'ADN parental comme matrice pour la synthèse de 2 molécules d'ADN double brin, constituée chacune d'un brin parental associé à un brin néosynthétisé. (Adaptée de http://fig.cox.miami.edu/~cmallery/150/gene/mol_gen.htm).

Bien que d'une incroyable fidélité, l'ADN polymérase introduit des erreurs lors de la copie du matériel génétique. Une grande partie de ces erreurs est corrigée immédiatement par des mécanismes complexes et efficaces de réparation de l'ADN (Kunkel, 2004), notamment grâce à l'activité exonucléolytique associée à la polymérase, encore appelée édition, ainsi qu'à un système post-réplicatif de réparation des mésappariements qui corrige les erreurs de réplication ayant échappé à l'édition (de Wind and Hays, 2001).

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

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

Dans certaines circonstances, le taux de mutations peut être considérablement augmenté par l'intervention d'agents environnementaux mutagènes. Ces derniers peuvent être des facteurs

physiques (rayons ultraviolets, radiations ionisantes, etc.) ou chimiques (agents désaminants, alkylants, etc.). Ils sont le plus souvent des agents cancérigènes et nocifs du fait de leur présence non contrôlée, et parfois non contrôlable, dans l'environnement. Ils agissent généralement sur un brin de l'ADN parental, produisant un changement de structure qui affecte la complémentarité du nucléotide altéré.

1.2.2.2 Effets des mutations sur le génome

Quelque soit l'origine de l'apparition d'une mutation, elle est susceptible de provoquer à elle seule une cascade de modifications pouvant entraîner des changements importants à plusieurs niveaux, de la séquence génomique aux traits physiques de l'individu. Nous allons à présent détailler les effets directs les plus connus d'une mutation sur la séquence du génome.

1.2.2.2.1 Substitution

Une substitution caractérise une mutation ponctuelle, qui se traduit par le changement d'un nucléotide par un autre. Parmi les substitutions, on peut distinguer les transitions et les transversions. La transition correspond au remplacement d'une purine (A ou G) par une purine ou d'une pyrimidine (T ou C) par une pyrimidine. La transversion correspond au remplacement d'une purine par une pyrimidine ou d'une pyrimidine par une purine. (**Figure 10**).

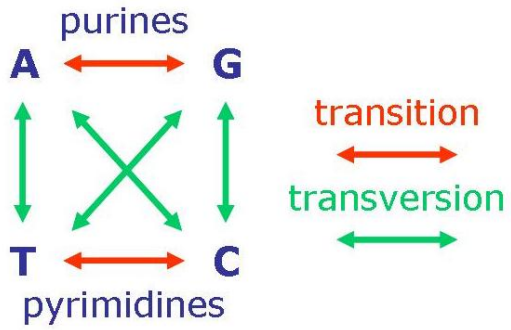


Figure 10. Possibilités de substitutions des 4 bases nucléotidiques. Adaptée de (Friedrich, 2007).

Les substitutions conduisent donc à la variation d'un seul nucléotide, on parle donc de SNP (*Single Nucleotide Polymorphism*) qui représentent près de 90% de la variabilité du génome humain (Collins et al., 1998) et auxquels on s'est intéressé au cours de cette thèse. Il a été estimé que 93% des gènes humains contiennent au moins un SNP et que 98% des gènes sont à proximité (+/- 5 kb) d'un SNP (Chakravarti, 2001). Le terme de polymorphisme est utilisé ici par abus de langage puisqu'il ne tient pas compte de la fréquence d'apparition dans la population.

1.2.2.2.2 Insertion et délétion

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

Les insertions/délétions, bien que susceptibles de se produire tout le long du génome, sont particulièrement fréquentes au niveau des séquences répétées. En effet, au moment de la duplication du génome, celles-ci peuvent être à l'origine d'un « glissement de réplication », au cours duquel le brin matrice, ou sa copie, peuvent dévier de leurs positions relatives en formant des boucles d'ADN simple brin par appariement de bases complémentaires de sorte qu'une partie de la matrice sera omise ou copiée 2 fois selon le cas (Figure 11). Ce phénomène explique, par exemple, la grande variabilité des séquences des microsatellites par la création

de nouveaux variants qui s'ajoutent à la collection des allèles déjà présents au sein de la population.

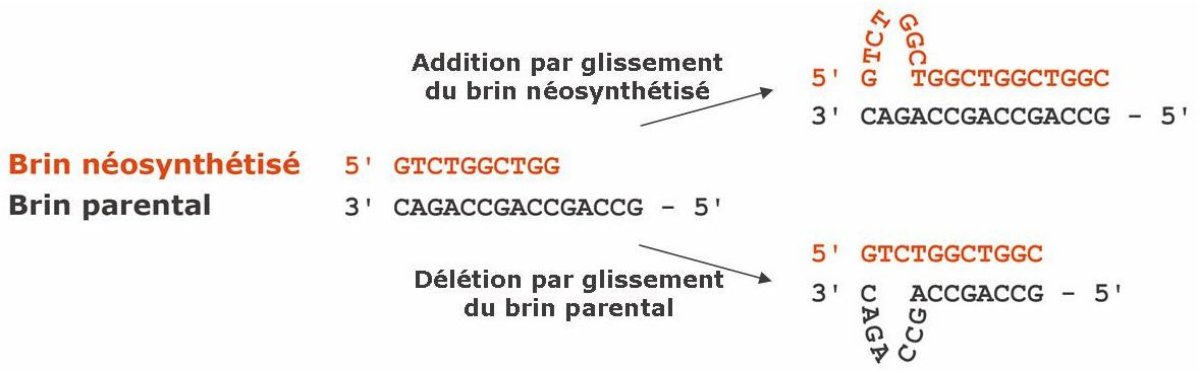


Figure 11. Glissement de réplication. Lors de la réplication, un glissement affectant le brin parental peut engendrer une délétion sur le brin néoformé, alors qu'un glissement de ce dernier peut être à l'origine d'une insertion. Adaptée de (Friedrich, 2007).

Il arrive également qu'un ou plusieurs nucléotides consécutifs soient remplacés par un ou plusieurs autres nucléotides (le même nombre ou un nombre différent). Dans ce dernier cas, il s'agit de l'insertion et la délétion simultanées de nucléotides.

On utilise plus simplement le terme d'indel (INsertion-DELétion) pour nommer tout événement impliquant une insertion/délétion. Bien que la plupart des indels ait été décrite comme étant associée à l'apparition de maladies génétiques chez l'homme, ils peuvent occasionnellement représenter des variants polymorphiques au sein de la population (Fernie and Hobart, 1997).

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

1.2.3 Conséquences des mutations

Les conséquences d'une mutation dépendent de la variation elle-même, mais également de l'endroit où elle s'est produite. Des mutations peuvent résider non seulement dans les régions codantes des gènes, mais également dans les régions non codantes, au niveau des promoteurs, au sein des introns, voire à distance du gène, et leurs conséquences sont variées. Dans ce travail de thèse, nous nous concentrerons plus particulièrement sur les mutations dans une région codante d'un gène.

Au moment de la traduction, la suite des acides aminés qui vont constituer la protéine est déduite de l'enchaînement des triplets de nucléotides, encore appelés codons, de l'ARNm par la relation unique qui existe entre un codon donné et son acide aminé correspondant, selon la règle édictée par le code génétique universel. Il existe 64 triplets possibles à partir de A, T, G, C qui codent pour vingt acides aminés et 3 codons stop (Figure 12). Comme on peut le voir dans la Figure 12, le code génétique est dégénéré, c'est-à-dire qu'un même acide aminé peut être codé par plusieurs codons différents. De plus, 3 codons (UAA, UAG, UGA) sont des codons qui ne peuvent pas être traduits en acides aminés : ce sont des codons appelés non sens ou STOP. Leur rôle est de marquer la fin de la traduction d'un gène en protéine.

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

Figure 12. Code génétique universel. Il a la particularité d'être dégénéré sur la 3ème base dite flottante.

Les conséquences sur le produit de l'expression d'un gène d'une mutation au niveau d'une partie codante de ce gène peuvent être de plusieurs natures (Figure 13).

1.2.3.1 Mutation silencieuse

A cause de la dégénérescence du code génétique, certaines mutations par substitution ne produisent aucun effet sur la protéine codée. En effet, la traduction du codon muté donnera le même acide aminé. On parle dans ce cas de mutation silencieuse.

1.2.3.2 Mutation exprimée

Si la mutation observée au niveau de l'ADN entraîne la modification d'un acide aminé, on parlera de mutation ponctuelle au niveau de la protéine. On peut classer les mutations ponctuelles en 3 catégories :

- mutations faux-sens : La substitution du nucléotide entraine une modification de l'acide aminé dans la protéine. La protéine mutée aura dans ce cas la même longueur que la protéine sauvage. Soit il n'y a pas de conséquences pour sa fonction, ce qui est souvent le cas lorsque l'acide aminé remplaçant possède les mêmes propriétés physicochimiques que l'original, soit la protéine subit une perte de fonction qui peut être due par exemple à la déstabilisation de la protéine. C'est à ce type de mutations que nous nous intéresserons par la suite (voir CHAPITRE 6).
- mutations non-sens : changement d'un codon sens en un codon non-sens (codon stop). La traduction s'arrête prématurément, il en résulte un polypeptide plus court et pour cette raison le plus souvent non fonctionnel. Plus la mutation sera proche du N-terminal, plus les effets seront délétères sur la protéine.
- mutation d'un codon stop : la mutation modifie un codon stop en un acide aminé et allonge la taille de la protéine. Les effets sur la structure et la fonction de la protéine dépendront généralement de l'emplacement du prochain codon stop.

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

- si la modification concerne un nombre de nucléotides multiple de 3, l'effet sur la protéine sera le gain ou la perte ou le remplacement d'un ou plusieurs acides aminés dans sa séquence et ses conséquences dépendront principalement de l'emplacement des résidus dans la structure et de leur(s) fonction(s). Schématiquement, si une telle mutation a lieu au cœur d'un élément de structure secondaire de la protéine, cette dernière risque d'être fortement déstabilisée par la modification.
- si la modification concerne un nombre de nucléotides non multiple de 3, on assistera à un changement du cadre de lecture de la protéine, plus couramment appelé *frameshift*. La séquence en acides aminés de la protéine est totalement modifiée après l'endroit où s'effectue la mutation et l'apparition d'un codon stop peut se faire à n'importe quel moment : la protéine peut être tronquée ou rallongée. Ces mutations ont souvent des conséquences très lourdes sur la protéine et ses fonctions, à l'exception de celles affectant l'extrémité C-terminale de la protéine qui entraînent le plus souvent, l'apparition rapide d'un codon stop.

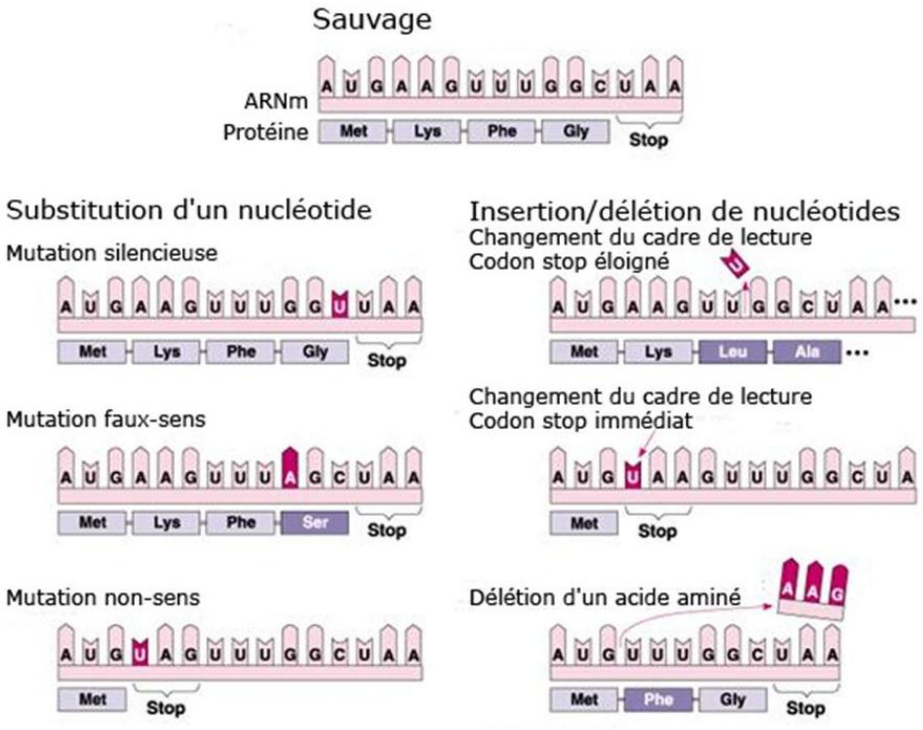


Figure 13. Conséquences des mutations sur la synthèse de la protéine. (Adaptée de (Friedrich, 2007)).

1.2.4 Impact des mutations sur les protéines

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

En revanche, les impacts des mutations faux-sens ou des indels (y compris les anomalies d'épissage) sont moins flagrantes. Ces mutations sont responsables d'un changement de la séquence protéique et sont susceptibles d'affecter par exemple la stabilité de la protéine, sa

maturation, son assemblage dans une structure multimérique, les sites essentiels à son activité enzymatique ou à l'interaction avec des ligands, etc.

Le changement de la séquence de la protéine peut également être sans conséquence sur sa fonction. Ainsi, lorsqu'un changement nucléotidique, de type faux-sens ou indel, survient au sein de la séquence codante d'un gène, un faisceau d'arguments est nécessaire pour déterminer les impacts de cette anomalie dans la pathologie.

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

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

L'enjeu de la discrimination des mutations délétères par rapport aux mutations neutres peut être schématisé par la capacité à prédire les substitutions ou modifications qui seront tolérées par la protéine. Le diagramme de Venn ayant trait aux propriétés des acides aminés (Figure 14) donne une idée de la compatibilité des résidus en fonction des caractéristiques physico-chimiques à maintenir.

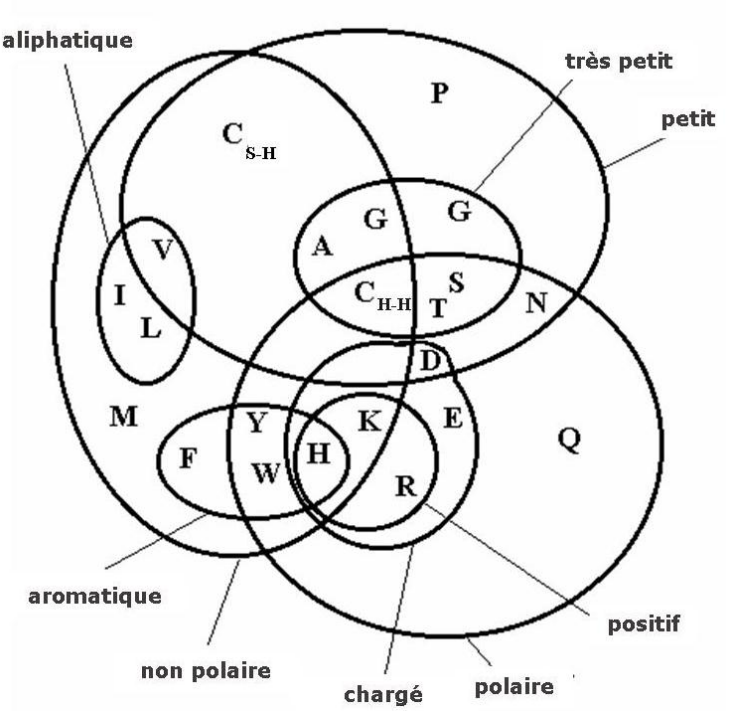


Figure 14. Classification des acides aminés d'après leurs propriétés physico-chimiques (Taylor, 1986) : diagramme de Venn.

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

Schématiquement, les résidus fonctionnels critiques au sein d'une protéine sont les suivants :

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

Les caractéristiques structurales critiques en rapport à la position relative de la variation dans la protéine sont les suivantes :

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

1.3 Maladies génétiques humaines

Comme nous venons de le voir, les mutations sont à la base des fondements de l'évolution, mais elles peuvent également avoir des effets négatifs et ainsi être à l'origine de maladies génétiques.

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

1.3.1 Définition d'une maladie génétique

Dans un premier temps, la définition d'une maladie génétique correspond à une affection qui est le résultat d'anomalies dans les gènes ou les chromosomes d'un individu. Une maladie génétique peut être héréditaire, c'est-à-dire transmissible de parent à enfant, ou somatique, c'est-à-dire survenir sporadiquement dans la population.

Schématiquement, on peut caractériser une maladie génétique par :

- son génotype qui est l'ensemble des allèles des gènes d'une cellule ou d'un individu,
- son phénotype qui correspond aux manifestations « observables » du génotype,
- la relation qui lie le génotype au phénotype.

Le génotype est caractéristique d'un individu, il représente son patrimoine génétique. Le phénotype est quant à lui beaucoup plus complexe. C'est le résultat de l'expression du génotype sous les contraintes de son environnement. Par exemple, les gènes exprimés dans une cellule musculaire ne seront pas les mêmes que dans une cellule nerveuse. Bien qu'ayant le même génotype, ces cellules exprimeront des phénotypes totalement différents. Au niveau de l'individu, le phénotype qui nous intéresse va être le phénotype global autrement dit l'ensemble des symptômes qui vont caractériser la maladie génétique.

Il n'est pas aisé de caractériser une maladie génétique par son phénotype car de nombreux symptômes peuvent apparaître en fonction de la position de la mutation dans le gène. Inversement, des symptômes similaires peuvent provenir de différentes maladies génétiques ou exprimer l'altération de plusieurs gènes différents chez un même individu. Il faut également considérer que le phénotype s'exprime à tous les niveaux de l'organisme, des niveaux moléculaire, cellulaire, tissulaire à l'organisme dans son entier. Il peut également s'exprimer différentiellement dans le temps. On peut distinguer :

- les phénotypes congénitaux, exprimés dès la naissance comme la drépanocytose qui affectent les globules rouges ;
- les phénotypes développementaux qui apparaissent plus tard, au cours de la croissance de l'individu comme la chorée de Huntington qui débute après 30 ans ;
- les phénotypes inductibles qui surviennent en réponse à un facteur de l'environnement comme l'hypolactasie (indigestion au lactose) qui apparaît lors de l'absorption de lait.

La relation qui existe entre le génotype et le phénotype est très complexe, même si le phénotype permet d'orienter le diagnostic d'une maladie, la détermination du génotype demeure une étape obligatoire. Schématiquement, on peut distinguer dans cette relation : (i) les facteurs génotypiques, représentés par l'ensemble des relations existant entre les 2 allèles de chacun des gènes d'un individu ; (ii) les facteurs épigénétiques , c'est-à-dire des modifications transmissibles et réversibles de l'expression des gènes ne s'accompagnant pas de changements des séquences nucléotidiques (Jirtle and Skinner, 2007) et (iii) les facteurs extérieurs, comme l'exposition environnementale ou le mode de vie de l'individu. Ces facteurs environnementaux peuvent notamment être à l'origine d'une variabilité interindividuelle des manifestations cliniques pour une même mutation.

Comme nous l'avons vu précédemment, certaines mutations entraînent une variation du phénotype sans pour autant provoquer de maladie chez l'individu. Il n'est pas évident de déterminer à partir de quel moment telle mutation sera considérée comme responsable d'une maladie génétique ou non. A l'origine d'une maladie génétique, on considère toute mutation induisant l'apparition de troubles médicaux chez un individu et transmissible de manière héréditaire, ce qui exclut les cancers de cette définition.

Les maladies génétiques peuvent être classées en différents groupes selon les gènes impliqués et leur mode de transmission.

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

1.3.2 Mode de transmission des maladies génétiques

L'homme possède 2 allèles de chaque gène, identiques (homozygotie) ou différents (hétérozygotie). Une maladie génétique sera transmise sur le mode dominant si les conséquences de la mutation de l'une des 2 copies du gène ne sont pas compensées par la copie normale, sinon on parlera de transmission en mode récessif. La transmission d'allèles, même délétères, peut se faire de manière transparente si elle est récessive.

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

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

Il existe globalement 3 modes de transmission qui suivent les lois mendéliennes.

- La transmission autosomique dominante. L'allèle muté se situe sur un chromosome autosome et est dominant ; la présence d'un seul allèle muté est suffisante pour que la maladie se déclare. C'est le cas de la Chorée de Huntington. On retrouvera ici autant d'hommes que de femmes atteints au cours des générations successives, au moins un parent atteint pour chaque individu atteint, mais aussi 2 parents atteints ayant des enfants sains (Figure 15A).

- La transmission autosomique récessive. L'allèle muté se trouve sur un chromosome autosome et est récessif, il est donc nécessaire que la mutation soit présente en double pour qu'un individu soit malade. On peut citer la thalassémie et la drépanocytose ayant ce mode de transmission. On trouvera dans un arbre généalogique caractéristique d'une maladie autosomique récessive des enfants atteints nés de 2 parents non-atteints, les garçons et les filles étant touchés à une fréquence équivalente. Tous les descendants de 2 parents atteints seront eux-mêmes atteints (Figure 15B).

- La transmission liée à un chromosome sexuel. Le cas le plus rare est celui lié au chromosome Y, la transmission se fait uniquement de père en fils. Le chromosome Y porte de nombreux gènes de différenciation sexuelle et par conséquent les mutations sur ce chromosome conduisent souvent à des individus stériles. Dans le cas général, il s'agit donc de mutations *de novo* qui ne donnent pas lieu à transmission. Pour le chromosome X, on détermine le caractère dominant ou récessif d'un gène lié à l'X grâce aux phénotypes des femmes qui possèdent 2 chromosomes X.

    - Dans le cas d'un état lié à l'X et dominant, un homme malade transmet sa maladie à toutes ses filles et aucun de ses fils, alors qu'une femme malade transmet sa maladie à 50% de ses enfants (Figure 15C). Globalement, la maladie atteint autant les hommes que les femmes. Parmi ces maladies, on peut citer : le syndrome de Rett, le syndrome de l'X fragile.

- Dans le cas d'une maladie liée à l'X et récessive, tous les fils d'une mère atteinte seront atteints, les pères atteints ne transmettront jamais le caractère à leur fils, des parents non atteints peuvent donner naissance à des fils atteints (Figure 15D). Globalement, la maladie sera plus fréquente chez les garçons que chez les filles, les filles sont conductrices de l'allèle malade. Des exemples de maladies récessives liées à l'X sont l'hémophilie, la myopathie de Duchenne.

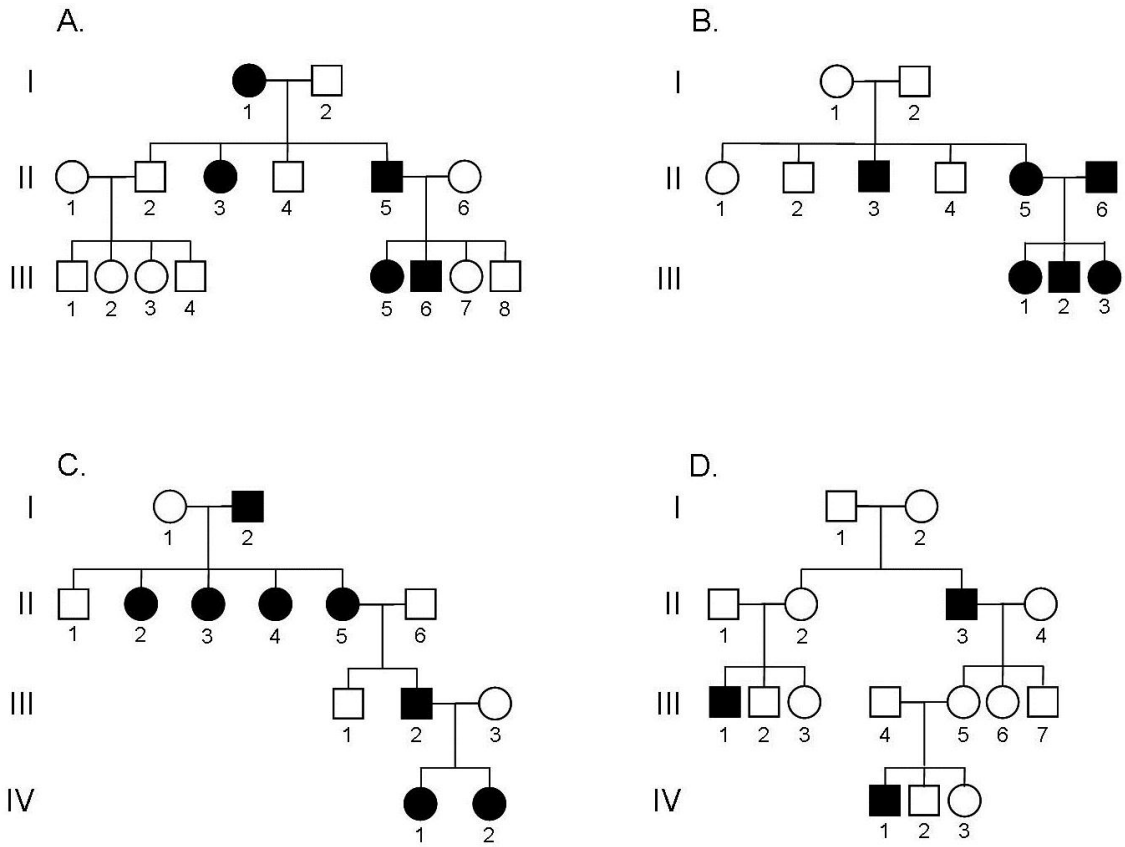


Figure 15. Arbres généalogiques : schémas de transmission des maladies monogéniques (Friedrich, 2007). Les ronds représentent les femmes et les carrés les hommes. Les ronds et carrés noirs représentent les individus malades. Les chiffres romains désignent les générations et les chiffres arabes précisent les individus de chaque génération. (A) arbre caractéristique d'une transmission autosomique dominante ; (B) arbre caractéristique d'une transmission autosomique récessive ; (C) arbre caractéristique d'une transmission liée à l'X et dominante ; (D) arbre caractéristique d'une transmission liée à l'X et récessive.

CHAPITRE 2. BIOLOGIE INTEGRATIVE DANS L'ETUDE DES LIENS COMPLEXE ENTRE PHENOTYPE ET GENOTYPE

« If you want to understand life,

don't think about vibrant, throbbing gels and oozes,

think about information technology. »

Richard Dawkins

2.1 Biologie intégrative

Depuis la mise en évidence de l'ADN comme source première de l'information génétique et la détermination, en 1953, de la structure de la double hélice d'ADN, la bioinformatique est devenue une discipline à part entière dans la recherche et les développements des sciences du vivant. Initialement conçues autour de méthodes informatiques dédiées à l'organisation et à l'analyse des données déposées dans les premières bases de données biologiques, les analyses bioinformatiques classiques étaient réalisées par des experts qui validaient visuellement ou expérimentalement les résultats obtenus *in silico*.

Le programme de séquençage du génome humain (*Human Genome Project*) a permis de déchiffrer la séquence des quelque trois milliards de bases présentes dans notre ADN. La bioinformatique a été traversée par cette révolution liée à la disponibilité de nombreuses séquences de génomes complets coïncidant avec la production d'une vaste quantité de données liées à l'émergence des biotechnologies à haut débit. Dès lors, les nouveaux systèmes intégrés sont développés pour analyser et exploiter des multitudes données provenant de la génomique (génome, gène, annotation…) et de la génomique fonctionnelle (transcriptome, protéome, interactome, métabolome…). Ces nouveaux systèmes ouvrent la voie vers l'étude des liens complexes entre génotype et phénotype (Blagosklonny and Pardee, 2002).

Cependant, les nouvelles biotechnologies ont abouti à des taux de production et d'hétérogénéité des données qui dépassent largement les capacités d'analyse et d'expertise humaines ainsi que les possibilités de traitement des plus puissants ordinateurs. Des développements sont nécessaires pour, d'une part, assurer le déploiement d'un système capable de gérer de gros volumes de données hétérogènes et de traiter rapidement des requêtes croisées entre différentes sources de données et d'autre part, créer des systèmes d'extraction de connaissances efficaces et pertinents capables de traiter les données fortement bruitées de la génomique fonctionnelle. Un tel système d'extraction de connaissances recouvre un processus itératif piloté par les connaissances elles mêmes incluant de nombreuses étapes de génération, épuration, validation, comparaison, analyse et représentation des données aboutissant à une nouvelle connaissance susceptible de relancer l'ensemble du processus. Pour cela, il faut un mariage entre la bioinformatique et l'ingénierie des connaissances.

2.2 Ingénierie des connaissances

On peut trouver une définition de l'ingénierie des connaissances (IC) récente et intéressante dans les comptes rendus des Journées Francophones d'Ingénierie des Connaissances 2009 qui se sont tenues à Hammamet en Tunisie : « L'ingénierie des connaissances permet de modéliser et d'acquérir des connaissances dans un but d'opérationnalisation et de gestion. Elle

propose des outils et méthodes pour les systèmes d'inscription, d'organisation, de diffusion et d'exploitation des connaissances au sens large. Les modèles de l'ingénierie des connaissances capturent des connaissances métiers, les processus cognitifs, les processus de coopération et les savoir-faire ».

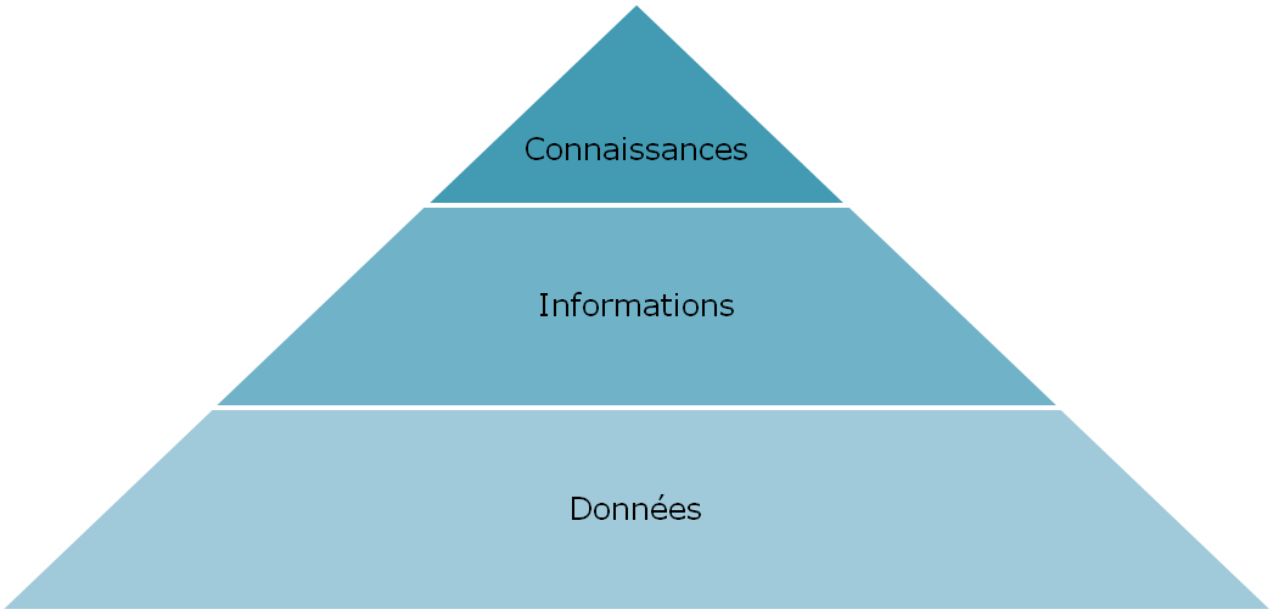


Figure 16. La pyramide des connaissances.

Dans l'IC, la distinction entre les données, les informations et les connaissances a été décrite explicitement (Figure 16) (Milan, 1987). Les données peuvent être définies comme une liste de faits, des éléments bruts ou des observations sans aucun contexte ou sens. Le contexte et les associations (ou les relations) entre les données sont nécessaires afin que les données puissent être transformées en informations utiles. Ainsi, l'information peut être considérée comme une collection de données organisées. La façon d'organiser les données est de faire des relations entre les données et les mettre en contexte. Les informations répondent aux questions du type : Qui ? Quoi ? Quand ? Où ? Par exemple, des entrées uniques dans une base de données sont des données, alors que le rapport créé à partir d'un ensemble de requêtes intelligentes sur la base de données résulte de l'information. Lorsque la combinaison de données et de sens à créer de l'information est extrêmement utile, la détection des modèles, des tendances et des exceptions étend la valeur de l'information vers les connaissances. Les connaissances appartiennent au sommet de la pyramide (Figure 16). Les connaissances sont différentes des données ou des informations en ce sens qu'elles peuvent être créées par l'accumulation de suffisamment d'informations ou par l'aide d'inférences logiques. Elles répondent aux questions du type : Pourquoi ? Comment ?

Prenons un exemple relevant du domaine de la biologie et considérons la séquence d'ADN constitutive d'un gène au sein d'une cellule. La séquence de nucléotides peut être considérée comme des données brutes. Le fait que l'on sache que cette séquence est reconnue par la machinerie cellulaire comme un gène particulier est une information. Enfin, les règles de fonctionnement de la machinerie cellulaire, et particulièrement le code génétique de la cellule, constituent les connaissances qui permettent d'interpréter ce gène comme une protéine, utilisée ensuite dans la mise en œuvre de fonctions biologiques.

L'Extraction de Connaissances à partir de Données (ECD) ou en anglais, *Knowledge Discovery in Databases* (*KDD*) est un des processus de l'IC. Il s'agit de l'utilisation de méthodes de fouille

de données, associées à une préparation des données préalables, et à une interprétation des résultats de fouille, afin d'extraire des connaissances pertinentes au regard des objectifs visés par l'expert du domaine étudié : l'analyste. Le processus d'ECD peut être appliqué à la fois de façon itérative et interactive. Itérative car les connaissances produites peuvent être réutilisées lors d'itérations suivantes du processus. Interactive car le processus d'ECD est réalisé sous le contrôle de l'analyste. C'est lui qui guide le processus en fonction de ses objectifs, de ses propres connaissances du domaine, et des résultats obtenus lors des précédentes itérations de l'extraction. Fréquemment, les connaissances obtenues sont exprimées sous forme de modèles numériques ou logiques : une série de coefficients pour un modèle de prévision numérique ou, des règles logiques du type « si Condition alors Conclusion ».

Le processus d'ECD est décrit dans (Fayyad et al., 1996) (**Figure 17**). Dans ce processus, la phase de préparation de données consiste à sélectionner la partie des données de la base globale qui nous intéresse, à nettoyer, valider et intégrer les données. Les données obtenues sont éventuellement transformées sous un format demandé par les méthodes de fouille de données. Les données seront entrées de la phase de fouille de données. La phase de fouille de données recherche et retourne les connaissances découvertes. Ces connaissances seront ensuite évaluées, interprétées et présentées à l'utilisateur pour une exploitation aisée : c'est la phase d'interprétation et d'évaluation.

Le choix de la méthode de fouille de données est déterminant et se fait essentiellement en fonction de l'objectif visé par l'analyste. Les différents objectifs de la fouille sont (Han and Kamber, 2001) :

- La recherche d'associations entre des attributs qui prennent des valeurs particulières de façon concomitante. Par exemple, l'un des objectifs de l'analyse de données du transcriptome est d'essayer de déterminer comment l'expression d'un gène particulier peut affecter l'expression d'autres gènes. Ces associations peuvent être révélatrices de mécanismes de co-régulation génique.

- La classification et la prédiction basées sur la définition d'un modèle à partir d'un jeu de données d'apprentissage. Par exemple, les biologistes sont intéressés par la découverte de facteurs permettant la discrimination des mutations neutres/délétères à partir des mutations connus.

- La construction de clusters qui regroupent les données ou les informations en différents groupes selon des mesures de similarité. Par exemple, on peut réaliser des clusters de protéines basés sur la proximité de leurs structures 3D pour inférer leurs fonctions.

Les méthodes de fouille de données sont souvent classifiées en fonction des divers objectifs exposés ci-dessus. En fonction du type de données qu'elles manipulent, il est également possible de classifier les méthodes de fouille de données :

- la fouille de données attribut-valeur pour de simples tables. Chaque ligne du tableau est un individu. Chaque colonne de la table a un nom qui est un attribut de l'individu. Pour un individu, la colonne a une valeur particulière. Ces approches de fouille de données se limitent donc au traitement d'une simple table et ne peuvent opérer sur plusieurs tables issues par exemple d'une base de données. On peut citer les machines à vecteurs de support (en anglais *Support Vector Machine*, SVM), les arbres de décision classiques, les régressions linéaires.

- la fouille de données relationnelles pour des tables de données relationnelles. Ces approches tirent profit de l'intérêt premier du modèle relationnel. Ce dernier permet en

effet de définir des relations entre les différentes tables de la base, des dépendances fonctionnelles entre les différents attributs. On peut citer la Programmation Logique Inductive qui sera introduite dans le CHAPITRE 4.

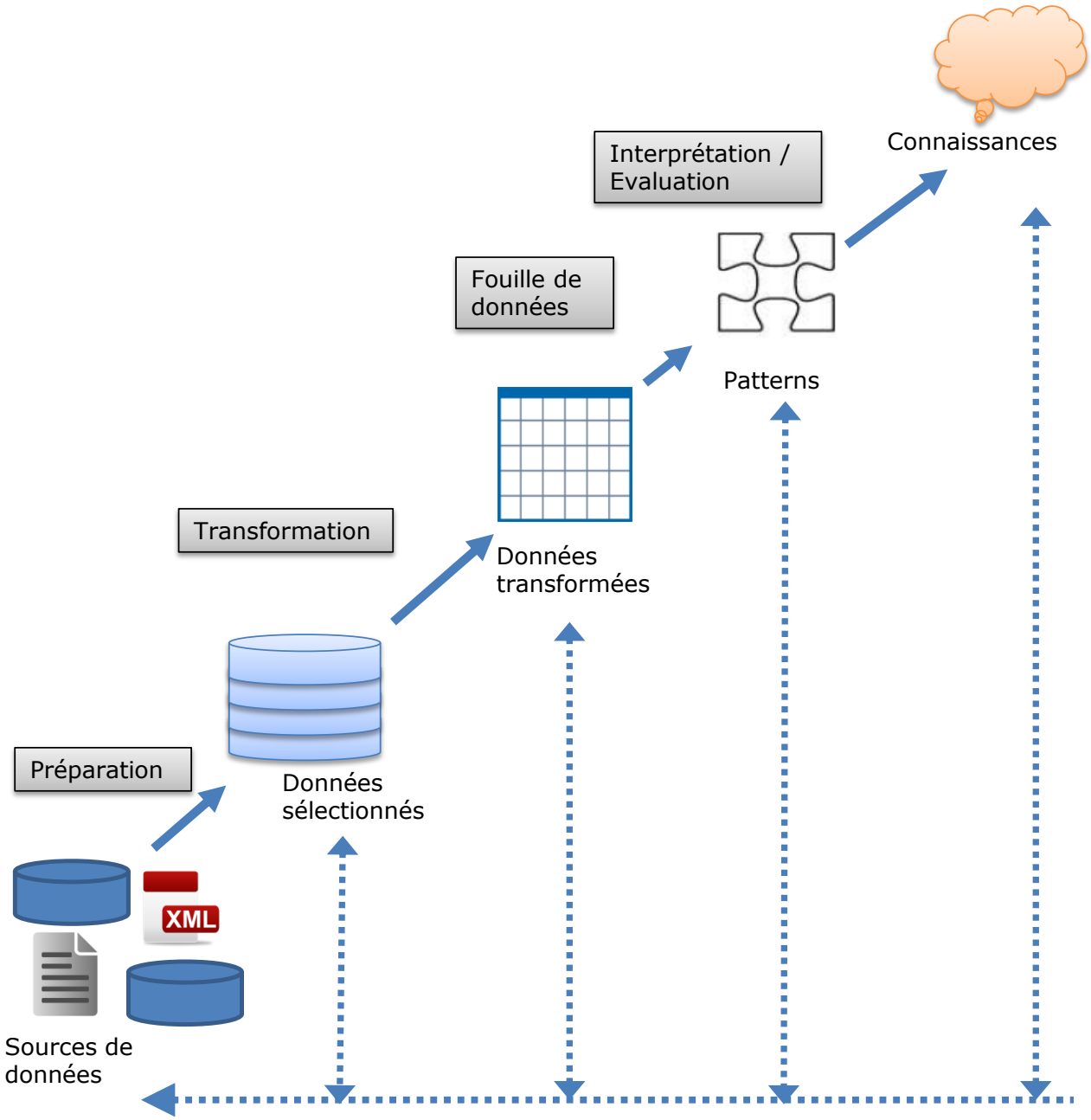


Figure 17. La représentation classique du processus d'extraction de connaissances à partir de données. Adaptée de (Fayyad et al., 1996).

2.3 Intégration de données biomédicales hétérogènes

Les analyses bioinformatiques actuelles en combinaison avec les processus ECD nécessitent l'accès à un grand nombre de vastes banques de données dont les tailles interdisent le transfert sur le réseau. Dès lors, il a fallu développer une infrastructure permettant d'intégrer les banques de données, accessibles tant par les utilisateurs que par les autres systèmes. De plus, la plupart des applications de calculs biologiques sont amenées à traiter des données (séquences nucléotidiques ou protéiques, transcriptomes, etc.) dont les formats sont de type « fichier plat ». Cette manière d'exploiter les données fait appel à de nombreuses opérations

répétées de transferts des fichiers originaux qui consomment énormément de temps sur le réseau.

Plusieurs approches et systèmes d'intégration (Hernandez and Kambhampati, 2004; Olund et al., 2007) sont disponibles pour créer et déployer facilement des bases de données hétérogènes. Ces approches sont brièvement classifiées en 3 types : fédération, entrepôt de données et indexation.

Dans l'approche fédérative, BioBank (Muilu et al., 2007) est un projet emblématique pour fédérer des bases de données sous une nouvelle architecture et offrir des services assurant la gestion des données biomédicales en utilisant la plate-forme Websphere Information Integrator (WII) de IBM. La plateforme WII a été mise au point pour relever le défi de l'intégration et de l'analyse des données scientifiques de divers domaines des Sciences du vivant. WII permet la création aisée d'une base de données virtuelle (schéma virtuel) à partir de Web Services et de bases de données relationnelles à distance. Cette technologie est totalement adaptée à la fusion de bases de données cliniques à distance avec des bases de données relationnelles locales.

Dans l'approche dite d'entrepôt de données, BioWarehouse (Lee et al., 2006) est un outil permettant la création d'un entrepôt de données biologiques. Cet outil est capable d'intégrer de multiples sources d'informations provenant des génomes, des ressources métaboliques ou d'enzymes dans une base relationnelle gérée par Oracle/MySQL. Cependant, la complexité et l'inter-dépendance du modèle de données rendent délicate l'intégration de nouvelles sources de données.

Finalement, dans le cadre de l'approche dite d'indexation, le *Sequence Retrieval System* (SRS) est largement utilisé pour l'indexation des banques biologiques. SRS est un système d'indexation en texte entier des fichiers en « format plat ». Toutefois, SRS n'est pas un système relationnel permettant l'opération de mise à jour fondée sur les transactions atomiques. Il ne convient pas bien à la gestion du contenu scientifique et de l'extraction direct des connaissances dans une base de données relationnelle (Bertone and Gerstein, 2001).

Dans ce contexte, le LBGI de l'IGBMC de Strasbourg a travaillé à la conception d'un nouveau système d'information biomédicale, nommé BIRD (pour *Biological Integration and Retrieval of Data*), fédérant des données généralistes (séquences, structures, fonction, évolution, etc.) et des données spécialisées de la biologie à haut débit (protéomique, transcriptomique, interactomique, phénomique, SNP, littérature, etc.). BIRD intègre efficacement les données hétérogènes dans une base de données relationnelle en utilisant des modèles de données configurables et des règles de configuration. Il s'est équipé d'un moteur de requête original afin de répondre aux requêtes de haut niveau. Nous reviendrons en détails sur ce système dans la section 3.1.

La tâche des bioinformaticiens est non seulement d'organiser efficacement les masses de données résultant des expériences à haut débit et d'en extraire une information fiable mais aussi d'élaborer des modèles qui vont permettre de répondre à des questions biologiques complexes. Traquer les molécules à l'origine de maladies est un exemple parmi d'autres. Pour cela, il faut d'abord trouver les gènes responsables de la maladie et ensuite « plonger » dans ces gènes pour détecter des mutations responsables et surtout, comprendre comment la simple modification d'une base peut être responsable de profondes altérations de la protéine, des processus dans lesquels cette dernière est impliquée et finalement, des différents phénotypes de la maladie chez des patients d'origine et d'histoire différentes. La section suivante va aborder ces démarches.

2.4 Outils bioinformatiques de prédiction des impacts des mutations faux-sens

Plusieurs groupes de recherche se sont attelés au développement d'outils de prédiction de l'impact des mutations faux-sens sur la fonction d'une protéine, avec plus ou moins de succès (Thusberg et al., 2011). Les méthodes actuelles de prédiction peuvent être schématiquement divisées en 2 catégories : celles exclusivement basées sur des données de séquences et celles qui combinent des données de séquences et de structures et qui dépendent par conséquent généralement de la disponibilité de données structurales.

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

Les facteurs prédictifs les plus couramment utilisés sont :

- la conservation des résidus au cours de l'évolution (Adzhubei et al., 2010; Bromberg and Rost, 2007; Chasman and Adams, 2001; Ng and Henikoff, 2003; Saunders and Baker, 2002),
- la compatibilité des résidus sauvages et mutés (Bromberg and Rost, 2007; Chasman and Adams, 2001; Ng and Henikoff, 2003; Ramensky et al., 2002; Saunders and Baker, 2002),
- la proximité d'un site fonctionnel (actif, liaison au ligand) (Chasman and Adams, 2001) (Ramensky et al., 2002; Saunders and Baker, 2002),
- le changement de charge ou d'hydrophobicité engendré par la substitution d'un résidu enfoui (Chasman and Adams, 2001; Ramensky et al., 2002; Saunders and Baker, 2002),
- l'impact d'un changement dans le cœur hydrophobe sur la solubilité de la protéine,
- la destruction d'un pont disulfure ou l'insertion d'une proline dans une hélice (Bromberg and Rost, 2007; Chasman and Adams, 2001; Ramensky et al., 2002; Saunders and Baker, 2002),
- la rigidité moléculaire (Chasman and Adams, 2001; Saunders and Baker, 2002),
- nombre de termes GO associés à la protéine (Calabrese et al., 2009; Capriotti and Altman, 2011).

Les approches considérées pour la prise de décision sont très diverses et recouvrent de multiples jeux de mutations, de protéines et des stratégies. De façon synthétique, on peut regrouper les approches utilisées autour :

- de fonctions de score (Ng and Henikoff, 2003),
- de règles de décision (Ramensky et al., 2002),
- de *Support Vector Machine* (Bao and Cui, 2005),
- de réseaux de neurones (Bromberg and Rost, 2007),
- de *Bayesien Naïf* (Adzhubei et al., 2010).

A l'heure actuelle, les 2 méthodes les plus couramment utilisées sont SIFT (Ng and Henikoff, 2003) et PolyPhen-2 (Adzhubei et al., 2010).

SIFT (*Sorting Intolerant From Tolerant*) s'appuie exclusivement sur l'homologie de séquence. Il prédit si une substitution sera tolérée ou non à chaque position de la séquence. Une substitution d'acide aminé dite tolérante est considérée comme n'ayant pas d'effet délétère sur la fonction de la protéine. A l'inverse, une substitution dite intolérante semble avoir un impact entrainant la perte partielle ou totale de la fonction protéique. Le postulat utilisé par SIFT est que toute évolution d'une protéine est corrélée à sa fonction protéique. Cela implique que les parties fonctionnelles qui sont sous forte pression de sélection sont donc conservées. La forte probabilité d'avoir un effet délétère provient d'une variation d'acide aminé survenant à une position fortement conservée d'une famille de protéine ; et inversement, on a une faible probabilité, si cette variation est à une position peu conservée. La probabilité, pondérée en fonction du nombre de séquences alignées, est calculée en comparant la matrice de mutation issue de l'alignement multiple avec celles prédites par BLOSUM62. SIFT élimine les séquences 100% identiques à la séquence cible. La valeur seuil permettant la prédiction est 0,05. C'est-à-dire, les positions avec des probabilités normales $< 0,05$ sont prédites délétères ; et celles $\geq 0,05$ sont prédites tolérantes.

Le principal avantage d'une méthode basée exclusivement sur des données de séquences comme SIFT est qu'elle ne nécessite pas l'accès à des données structurales et peut par conséquent être appliquée à la quasi totalité des protéines d'un organisme. L'inconvénient majeur réside dans la réduction des capacités de prédiction lorsque le nombre d'homologues est limité.

PolyPhen-2 (*Polymorphism Phenotyping*) utilise des données de séquences et de structures, quand la structure expérimentale de la protéine d'intérêt ou d'un homologue partageant plus de 50% d'identité est disponible. La position de la substitution est caractérisée par :

- la conservation de séquences en se servant du score de l'alignement de séquence calculé par PSIC (*Position-Specific Independent Counts*) (Sunyaev et al., 1999),
- le domaine PFAM extrait de Swiss-Prot,
- l'accessibilité au solvant et les structures secondaires assignées sur la structure par le logiciel DSSP,
- le facteur B, facteur cristallographique décrivant l'agitation thermique renseignant sur l'incertitude sur la position d'un atome, d'un acide aminé et/ou d'une structure secondaire.

Les substitutions sont classées en 3 niveaux : bénignes, potentiellement délétères et probablement délétères. La prédiction dite « bénigne » signifie que la fonction de cette protéine ne semble pas être affectée par cette variation. A l'inverse, la prédiction dite « probablement délétère » signifie que la protéine a une forte probabilité d'être altérée par cette variation.

Il a été montré que les méthodes utilisant l'information structurale donnent dans l'ensemble de meilleures prédictions que les autres (Bao and Cui, 2005; Calabrese et al., 2009; Capriotti and Altman, 2011; Masso and Vaisman, 2008; Ramensky et al., 2002; Yue et al., 2006). En principe, les données structurales devraient apporter une aide à la prédiction des effets des mutations. Cependant, en pratique, traduire une connaissance structurale sur le phénotype est loin d'être trivial et les méthodes basées exclusivement sur la séquence restent très puissantes. Ceci démontre une nouvelle fois la robustesse de l'étude de la conservation des résidus dans des protéines homologues et reflète l'importance structurale et fonctionnelle des contraintes évolutives subies par la protéine. Tous s'accordent pour dire que l'information

évolutive est le facteur prédictif le plus puissant, mais que les informations structurales sont indispensables lorsque peu de séquences homologues sont disponibles (Thusberg et al., 2011).

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

DEUXIEME PARTIE : DONNEES ET METHODES

La deuxième partie présente rapidement le matériel, en l'occurrence essentiellement les données, et les méthodes utilisés durant ma thèse. Le Chapitre 3 concerne les bases de données et les méthodes bioinformatiques, puis, dans le Chapitre 4, les principes de la méthode d'acquisition automatique de connaissance dite, de Programmation Logique Inductive (PLI) sont décrits en détail.

CHAPITRE 3. DONNEES BIOLOGIQUES ET OUTILS BIOINFORMATIQUES

Ce chapitre présente les données et outils bioinformatiques utilisés pour aboutir à la mise en place d'une infrastructure dédiée à l'analyse globale des relations qui lient le phénotype d'un individu à son génotype dans le cadre des maladies humaines.

3.1 Fédération des données biologiques par le système BIRD

Le Centre de Données Décrypthon (CDD) est capable d'héberger des informations biologiques et biomédicales hétérogènes, issues de diverses banques de données publiques (Tableau 2). Le CDD fait partie intégrante de l'infrastructure mise en place dans le cadre du programme Décrypthon initialement développé par Anne Friedrich. Il est hébergé par le serveur de stockage situé au CRIHAN (Centre de Ressources Informatiques de HAute-Normandie) et a, en particulier, été développé pour être exploité dans le cadre de la grille de calcul universitaire Décrypthon (http://www.decrypthon.fr/). Le CDD permet ainsi la mise à disposition directe d'un ensemble de données nécessaires au bon fonctionnement des applications portées sur cette grille, sans avoir besoin de les transférer sur chaque nœud de calcul.

Nom	Format	Source d'origine
UniProt	Flat	Universal Protein Resource
RefSeq	Flat	National Center for Biotechnology Information
InterPro	XML	EMBL European Molecular Biology Laboratory
PDB	PDB	RCSB Protein Data Bank
Gene Ontology, Human Phenotype Ontology	OBO/OWL	Open Biological and Biomedical Ontologies Foundry
OMIM	Flat	National Center for Biotechnology Information
MEDLINE	XML	National Library of Medicine
Taxonomy	Table relationnelle	National Center for Biotechnology Information
dbSNP	XML	National Center for Biotechnology Information
STRING	Table relationnelle	EMBL European Molecular Biology Laboratory
SCOP	Fichiers tableurs CSV	Laboratory of Molecular Biology in Cambridge

Tableau 2. Banques de données intégrées au Centre de Données Décrypthon.

Le CDD tire profit du système BIRD (Nguyen et al., 2008), pour *Biological Integration and Retrieval Data*. BIRD est un système générique qui permet la création de systèmes de bases de données à façon, à partir de schémas relationnels préexistants, en intégrant des liens croisés entre les données. Un modèle de données configurable est d'ores et déjà disponible pour plusieurs banques de données, notamment celles incorporées au CDD. BIRD permet ainsi l'intégration efficace de données hétérogènes.

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

Enfin, le centre Décrypthon partage ses données intégrées avec des applications/clients et des utilisateurs de plusieurs manières : par l'intermédiaire d'une API Java, d'une interface web et du langage BIRD–QL *via* des services web (HTTP) (la section 3.8.2).

Dans le cadre de la conception du nouveau système d'information SM2PH Central, présenté dans le CHAPITRE 5, BIRD a été utilisé comme le noyau de fusion de données hétérogènes. Il permet la mutualisation des données nécessaires à la réalisation des différents projets ainsi qu'un accès simplifié aux données pour la communauté scientifique.

3.2 Données génomiques / protéomiques

3.2.1 Banques de séquences protéiques

3.2.1.1 UniProt

UniProt (*Universal Protein resource*) (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 4 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 d'identité (100%, 90% et 50%) les séquences d'un même organisme provenant d'UniProtKB. Cette banque permet d'accélérer et d'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.

Au cours de ma thèse, j'ai utilisé les banques de données UniProtKB et UniRef dont je vais faire une description sommaire dans les sections suivantes.

3.2.1.1.1 UniProtKB

UniProtKB (Wu et al., 2006) est considérée comme la référence en matière de banque de protéines annotées. Elle est accessible à l'adresse suivante : http://www.uniprot.org/. Ses données sont distribuées publiquement sous forme de fichiers plats. UniProtKB est issue de la réunion de 2 banques de séquences protéiques : Swiss-Prot et TrEMBL.

La banque Swiss-Prot contient des séquences protéiques peu redondantes provenant de 4 origines : la traduction des gènes annotés dans la banque EMBL, banque européenne de référence de séquences nucléotidiques ; certaines protéines issues d'autres banques protéiques ; la consultation de publication scientifique et enfin plus rarement, de soumissions directes par les auteurs. Une des caractéristiques de cette banque est la qualité et la richesse des annotations des séquences collectées. Les différentes catégories d'informations figurant dans la banque comprennent notamment, pour chaque entrée, la, ou les, fonctions des protéines, les modifications post-traductionnelles connues, les sites et domaines structuraux ou fonctionnels identifiés, des références bibliographiques, les structures secondaires et

quaternaires, les similarités avec d'autres protéines, les positions conflictuelles ainsi que les mutations connues associées à d'éventuelles pathologies, etc. Un des atouts supplémentaires de cette banque repose sur l'introduction d'un nombre très important de références croisées avec d'autres banques de données (nucléiques, de structure, de motifs, etc.). Cette spécificité se révèle particulièrement utile avec le développement des systèmes d'interrogation de banques. Le flot grandissant de nouvelles séquences (Figure 18) issues de projets de séquençage de génomes complets pénalise la banque Swiss-Prot, dont la richesse des annotations ne peut suivre une telle explosion de données.

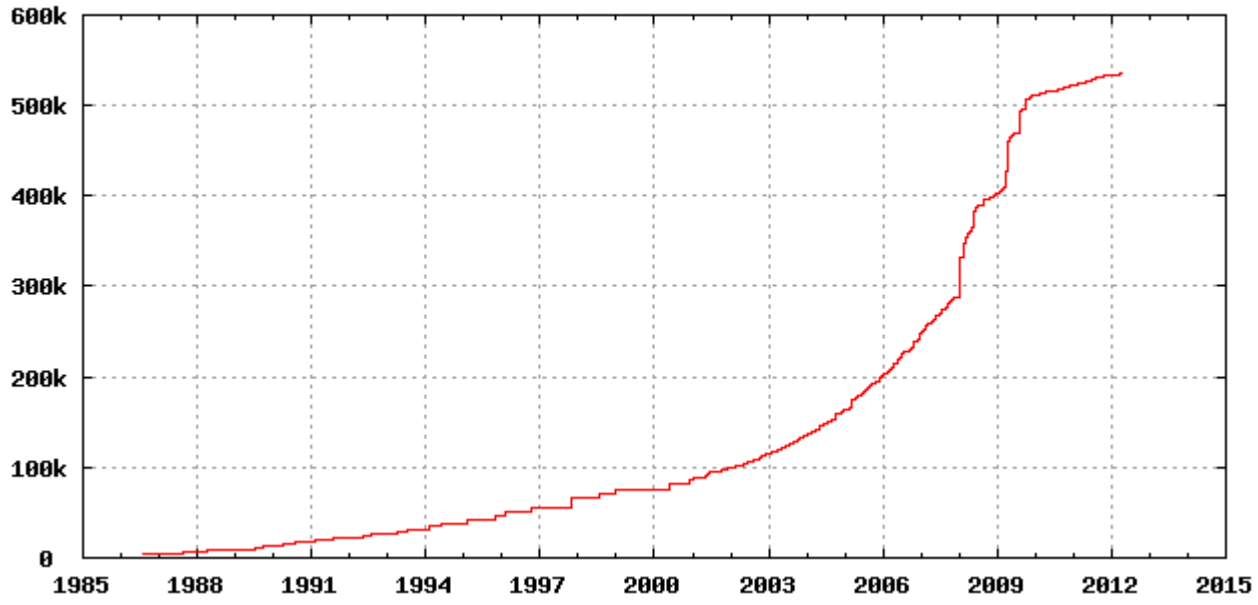


Figure 18. Evolution du nombre d'entrées de la banque Swiss-Prot depuis sa création en 1986. Adaptée de http://web.expasy.org/docs/relnotes/relstat.html.

Introduite en 1996, la banque TrEMBL, pour *Translation from EMBL*, est un supplément de Swiss-Prot. En effet, TrEMBL, distribuée par l'EBI, contient la traduction de toutes les parties codantes annotées figurant à l'EMBL, la banque européenne de référence de séquences nucléotidiques. Elle contient donc des séquences non validées, souvent incomplètes, faiblement annotées. Les séquences présentes dans TrEMBL constituent un complément de la banque Swiss-Prot qui seront éventuellement intégrées dans Swiss-Prot après vérification minutieuse par les annotateurs de Swiss-Prot.

3.2.1.1.2 UniRef

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

UniRef100 regroupe, dans un seul enregistrement, les séquences et fragments de séquences identiques provenant d'un même organisme.

Les banques UniRef90 et UniRef50 sont construites à partir des clusters d'UniRef100. Elles regroupent les séquences de tous les organismes sur la base d'un seuil d'identité de 90% et 50% respectivement.

[2]Pour chaque cluster des banques UniRef, une séquence est choisie comme représentant et des références croisées permettent d'avoir accès aux autres séquences du cluster.

L'utilisation des banques UniRef permet d'accélérer les recherches de similarité, mais souvent aussi d'augmenter leur significativité. Ces banques répondent en effet à un besoin de « simplification » de l'information, qui fait écho aux volumes de données actuellement disponibles et à la forte redondance présente dans les banques. Cette redondance peut être illustrée non seulement par la présence de séquences strictement identiques, mais aussi par une « redondance fonctionnelle » entre les séquences proches.

3.2.1.2 RefSeq

RefSeq, pour *Reference Sequence database*, est une collection de séquences distribuées et générées par le NCBI (*National Center for Biotechnology Information*) depuis 2003, regroupant à la fois des séquences protéiques et des séquences nucléiques (Pruitt et al., 2005). Cette banque est hébergée par le site http://www.ncbi.nlm.nih.gov/RefSeq/ et disponible gratuitement à la communauté scientifique sous la forme de fichiers plats.

L'idée est de distribuer des séquences de référence stables et utilisables pour différentes études fonctionnelles et médicales. La banque est disponible par organisme ou par taxon et permet donc de disposer d'une collection plus ou moins complète des séquences présentes dans un organisme donné (l'homme par exemple). Au niveau nucléique, RefSeq dérive de la banque GenBank, mais se distingue de cette dernière par plusieurs points. En effet, à la différence de GenBank (Benson et al., 2006) qui se veut un dépôt de toutes les séquences nucléiques existantes et contient de ce fait une redondance importante et nécessaire pour garder le caractère original des séquences soumises, RefSeq contient des séquences validées, annotées, non redondantes qui reflètent la synthèse des connaissances actuelles pour chacune de ses entrées. Au niveau protéique, la même politique de validation est appliquée.

3.2.2 Banques de mutations

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

Comme nous l'avons vu, UniProt (Wu et al., 2006), la banque de référence des protéines contient également des données sur les variants faux-sens. Les données sont disponibles dans un fichier plat à l'adresse www.uniprot.org/docs/humsavar.txt. Les variants référencés ne sont pas exhaustifs mais touchent un grand nombre de protéines; le phénotype est décrit par le nom de la maladie et fait parfois état d'une notion de sévérité. Si le variant a pu être expérimentalement relié à une maladie, la mention *disease* sera présente. Si le variant n'est pas associé à une maladie, alors il sera mentionné *polymorphism*, si le lien entre une maladie et le variant n'est pas clairement établi, il sera mentionné *unclassified*. À la date du 11 juillet 2012, la banque UniProt comporte un total de 66 399 variants faux-sens, dont 22 893 (34.4%) variants sont reliés à une maladie, 37 159 (56%) sont liés à la notion de *polymorphism* et 6 347 (9.6%) variants sont mentionnés comme *unclassified*.

Cependant, le catalogue le plus complet de variations génétiques connues au sein de divers génomes est disponible dans la banque de données dbSNP (Sherry et al., 2001). Elle a été créée au NCBI en collaboration avec le NHGRI (*National Human Genome Research Institute*,

États-Unis) pour stocker et rendre accessibles les polymorphismes, qu'il s'agisse de substitutions de bases nucléotidiques simples ou de courtes indels. Le terme polymorphisme est ici employé par abus de langage dans le sens où, de manière générale, la fréquence de la variation dans la population n'est pas prise en compte. Il est important de noter ici que les polymorphismes déposés dans dbSNP sont validés biologiquement ou non. De plus, le lien entre le génotype et le phénotype n'est souvent pas établi. La banque dbSNP est mise à jour régulièrement et répertorie actuellement (build 135) environ 47.8 millions de polymorphismes humains, dont 503 847 sont mutations faux-sens. La plupart des polymorphismes a été identifiée chez les principales espèces modèles telles que l'homme, la souris (*Mus musculus*) ou le riz (*Oriza sativa*). Cette banque est hébergée par le site (http://www.ncbi.nlm.nih.gov/projects/SNP/) et disponible gratuitement à la communauté scientifique sous la forme de fichiers XML. Les données contenues dans dbSNP sont intégrées avec d'autres données génomiques du NCBI.

Les banques spécifiques d'un locus, couramment appelées LSDB, pour *Locus-Specific DataBase*, contiennent, comme leur nom l'indique, des informations relatives à un gène ou à une famille de gènes spécifiques impliqués dans le même type de maladie. Ces banques sont gérées par des chercheurs ayant une expertise dans le gène ou les maladies qui y sont associées. Le principal intérêt de ces banques est la qualité des données et des annotations qui y sont collectées. Selon les banques, les descriptions génotypiques et phénotypiques peuvent atteindre des degrés de détails très élevés (description des 2 allèles du patient, plusieurs niveaux de description phénotypique, etc.). Près de la moitié des données contenues dans les LSDB sont des données non publiées (Cotton, 2000), saisies directement par les expérimentateurs. A ce jour (Aout 2012), on compte près de 4 116 LSDB référencées et accessibles par le site de la *Human Genome Variation Society* (http://www.hgvs.org/dblist/glsdb.html). Le nombre de LSBD disponibles ne cesse de croître, mais toutes ne sont pas maintenues de manière régulière. De plus, les formats de ces banques sont très variables, mais des efforts sont déployés pour leur homogénéisation et des outils génériques sont mis à la disposition des cliniciens, notamment par l'intermédiaire des systèmes UMD, pour *Universal Mutation Database* (Beroud et al., 2005) et LOVD, pour *Leiden Open (source) Variation Database* (Fokkema et al., 2005).

Comme nous le verrons par la suite, toutes les mutations faux-sens d'UniProt et de dbSNP sont intégrées dans MSV3d. La situation actuelle ne nous permet pas d'envisager, dans l'immédiat, l'alimentation automatique de MSV3d avec des données issues de LSDB pour chaque protéine impliquée dans une maladie génétique humaine. Nous avons par conséquent choisi, afin de couvrir le spectre le plus large possible d'entrées, de collecter manuellement des mutations faux-sens dans quelques LSDBs tel que UMD-MTM1 [Biancalana et al., in preparation] et *Tissue Nonspecific Alkaline Phosphatase Gene Mutations Database* (http://www.sesep.uvsq.fr/database_hypo/Mutation.htm).

3.2.3 PDB

La PDB (*Protein Data Bank*) (Berman et al., 2007; Henrick et al., 2008; Kouranov et al., 2006; Rose et al., 2011) est la principale banque internationale de structures 3D de macromolécules biologiques. Cette banque a été établie dès 1971 au *Brookhaven National Laboratory* (BNL) à partir de sept structures. Depuis 1998, elle a été transférée au *Research Collaboratory for Structural Bioinformatics* (RCSB). En 2003, l'organisation wwPDB (*Worldwide PDB*), regroupant le RCSB, le groupe EBI-PDBe (*PDB in Europe*), la PDBj (*PDB of Japan*) et la BMRB (*Biological Magnetic Resonance Data Bank*), a été fondée pour superviser la distribution de la PDB.

À la date du 21 août 2012, la PDB comporte un total de 83 938 structures très majoritairement résolues par cristallographie aux rayons X (87.8%), mais on y trouve 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 (92.6%), le reste se partageant entre les structures d'acides nucléiques et les complexes protéine-acide nucléique.

Chaque entrée comporte les structures primaires et secondaires des molécules considérées, les coordonnées cartésiennes des atomes, souvent des détails des expériences (conditions de cristallisation, collecte des données, résolution de structure…) ainsi que des références croisées de banques de données et des références bibliographiques. Chaque entrée est un fichier plat au format PDB et disponible à l'adresse http://www.rcsb.org/pdb/.

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 19), notamment grâce aux projets massifs de génomique structurale de la *Protein Structure Initiative* (Berman et al., 2009) (Nair et al., 2009)

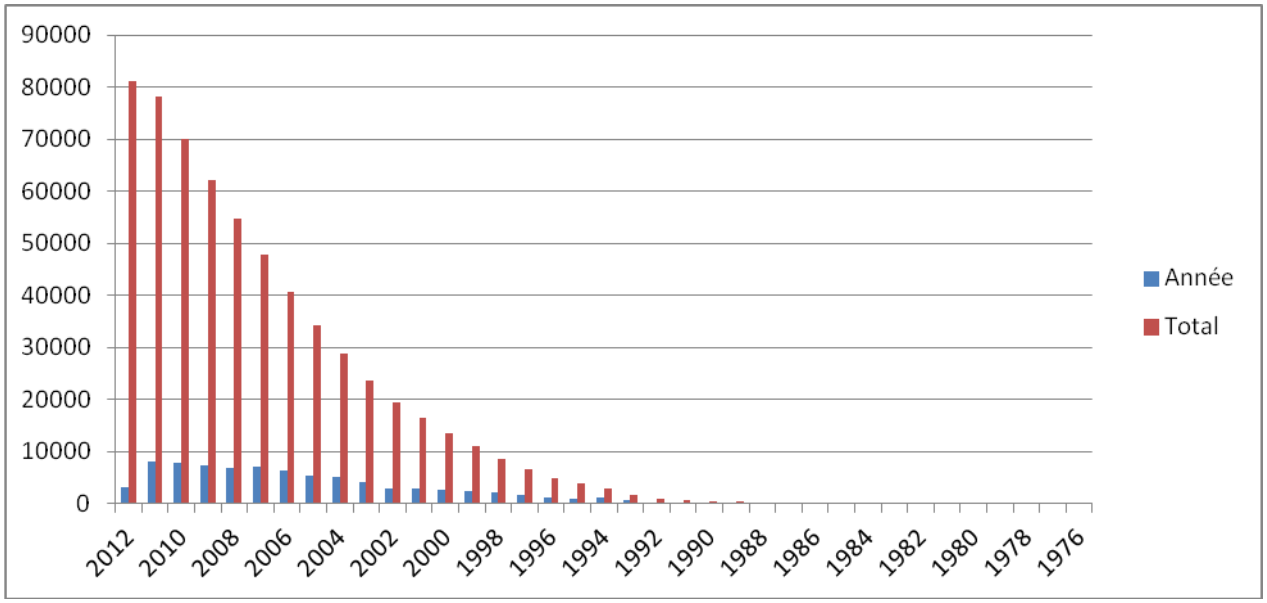


Figure 19. Évolution du nombre d'entrées de la PDB de 1976 à juin 2012. Les barres bleues illustrent le taux de croissance annuelle et les barres rouges la croissance cumulée. Données adaptées de http://www.rcsb.org/pdb/statistics/contentGrowthChart.do?content=total&seqid=100

On peut cependant noter une redondance importante dans la PDB, puisque plusieurs structures 3D peuvent correspondre à la même séquence, selon les conditions d'obtention de la structure, la présence ou non de ligand, l'existence de mutations... Quand on regroupe les séquences sur la base d'un seuil de similarité de 100%, on obtient 47 453 (56.5%) structures 3D différentes.

3.2.4 SCOP

La comparaison par paire des structures tertiaires des protéines contenues dans la banque PDB a permis leur classification hiérarchique dès les années 1995, avec la création de la base de données SCOP (*Structural Classification of Proteins*) (Murzin et al., 1995). SCOP fournit une description détaillée des structures et des relations évolutives des protéines. SCOP regroupe les séquences de structures connues en familles, en se basant sur les liens évolutifs et les principes gouvernant le repliement 3D des protéines (informations issues de la littérature sur

les repliements 3D bien documentés). La classification de SCOP s'appuie essentiellement sur des méthodes de comparaison automatique pour les séquences les plus proches et sur l'inspection manuelle à l'aide d'outils de visualisation pour les séquences les plus distantes. Les structures sont tout d'abord découpées en domaines et sont ensuite regroupées selon leur similarité de séquence, puis leur similarité de structure, puis leur similarité en composition et organisation des structures secondaires. Les 4 niveaux de classification (Figure 20).sont, du plus général au plus fin : (i) la classe, (ii) le repliement, (iii) la superfamille, (iv) la famille. SCOP est accessible à l'adresse suivante : http://scop.mrc-lmb.cam.ac.uk/scop. Ses données sont distribuées publiquement sous forme de fichiers tableurs CSV (*Comma-Separated Values*).

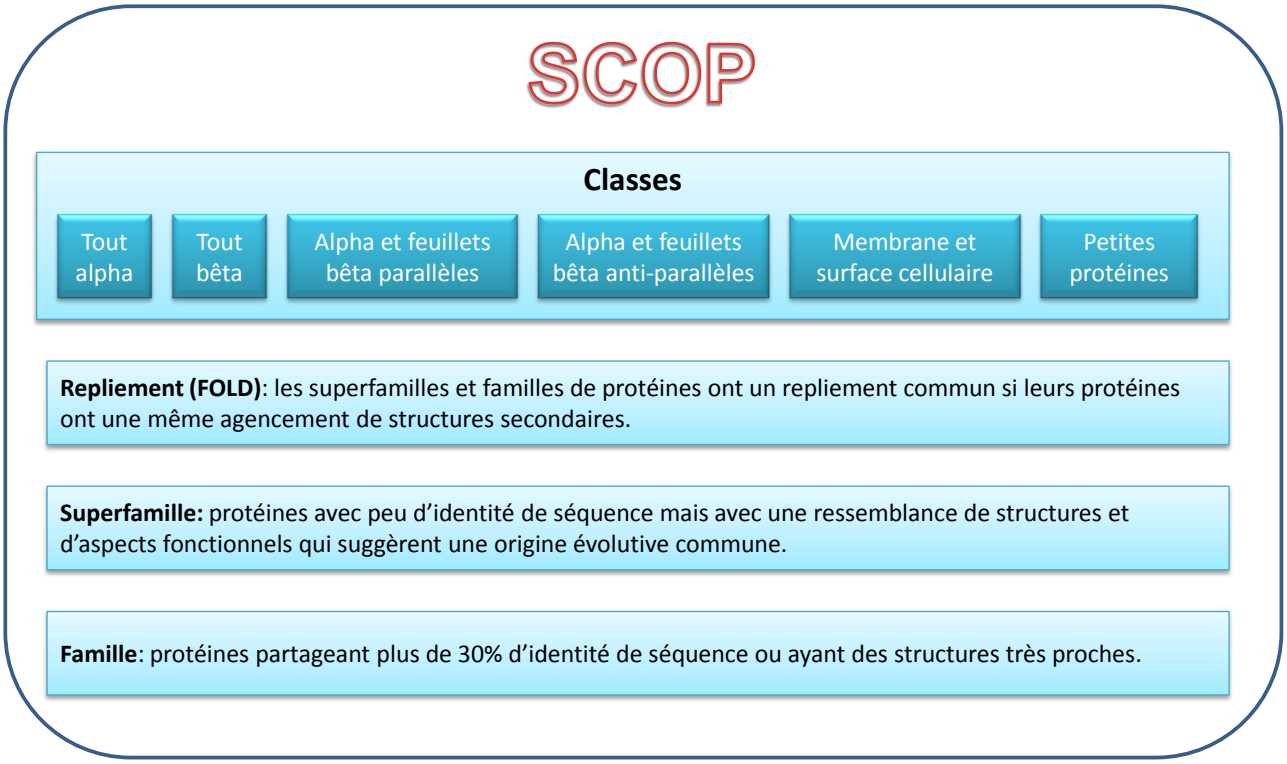


Figure 20. Classification hiérarchique des structures protéiques dans SCOP. La base de données SCOP regroupe les structures en fonction de leurs identités de séquences, et de leur ressemblance au niveau des structures secondaires et tertiaires.

3.3 Données transcriptomique : GxDB

Au cours de ces dernières années, la transcriptomique, à savoir la connaissance du niveau d'expression de tous les gènes d'un organisme dans des conditions définies, est devenue un outil essentiel dans l'étude du vivant. Un transcriptome offre en effet l'opportunité de mieux comprendre le fonctionnement globale d'une cellule, d'un tissu ou d'un organisme et est à même d'informer sur la réponse du génome à un environnement donné, sur l'état physiologique d'un tissu, sur l'état de santé d'un patient, sur l'état d'un milieu dans son ensemble…. Dans le cadre de l'étude des maladies génétiques humaines, le transcriptome est amené à jouer un rôle de plus en plus important. Le transcriptome offre souvent les premiers éléments de compréhension des mécanismes reliant une maladie à de nouveaux loci, détectés, par exemple, lors d'études d'associations pangénomiques (*GWAS : Genome Wide Association Study*). Enfin, à condition de savoir caractériser l'impact des divers facteurs environnementaux, il est possible d'utiliser le transcriptome soit comme une variable d'ajustement, soit comme un trait phénotypique à part entière. Cependant, de telles utilisations impliquent que l'on puisse comparer de façon objective, des transcriptomes de

différentes origines (organismes, tissulaires, cellulaires…) ou provenant de diverses conditions liées au développement, au cycle cellulaire, à l'état physiologique, à l'existence de traitement…. L'analyse et la méta-analyse (la comparaison transversale entre différentes expériences) de ces données restent assez laborieuses et sont souvent difficilement accessibles aux biologistes.

Pour répondre à ces besoins, notre laboratoire a développé une plateforme innovante, GxDB, permettant une grande automatisation des tâches, une visualisation des données et fournissant au biologiste des outils d'annotations et d'analyses approfondies. De plus cette plateforme doit permettre au bioinformaticien de développer et d'intégrer de nouveaux outils.

Dans le cadre de ma thèse, j'ai été amené à utiliser GxDB qui combine une base de données relationnelle, le traitement entièrement automatisé des données brutes de transcriptomique, des outils d'analyse et un site web sécurisé. Le traitement automatique de GxDB excluant l'intervention humaine nous permet d'analyser et de comparer de façon assez objective des données de natures très différentes. Après le téléchargement des données grâce une interface web conviviale, une série d'analyses est lancée en automatique et peut être affinée par un expert humain. GxDB permet de combiner différents outils tout au long de l'analyse d'une expérience et donc de comparer plusieurs méthodes, mais aussi des expériences entre elles.

GxDB fournit des services pour distribuer automatiquement ses données. SM2PH Central utilise en particulier le service nommé ExpressTiss qui fournit les expressions tissulaires de protéines humaines dans 79 tissus calculées par 6 méthodes de normalisation (RMA, gcRMA, dChip, MAS5.0, VSN et Plier). Il requiert en entrée une liste de gènes ou de protéines et rend en sortie un tableau de niveaux d'expression où chaque ligne correspond à une *probeset* (caractéristique d'un gène) et chaque colonne à un tissu. Les données en sortie sont stockées dans un fichier CSV.

3.4 Données métaboliques et réseaux fonctionnels : 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 du *Human Genome Program* lancé par le Ministère japonais de l'éducation, de la culture, des sports, des sciences et des technologies. KEGG se définit comme une série de bases de données orientée vers la biologie des systèmes permettant la compréhension de mécanismes et de concepts fonctionnels complexes de la cellule ou d'un organisme à partir des gènes et des relations ou produits liés à ces gènes.

KEGG est constituée de 4 banques principales connectées entre elles et jouant le rôle de points d'entrée pouvant aboutir aux autres sous-banques qui les composent :

- PATHWAY est la base de connaissances des voies métaboliques dessinées manuellement et des 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 sont les arêtes (Figure 21).

PATHWAY est sans doute le point central de la banque KEGG. Ses entrées sont organisées en une hiérarchie à 2 niveaux reflétant la résolution de chaque voie. 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 2 sous-graphes : un graphe d'interactions protéine-protéine et un graphe de réactions enzymatiques.

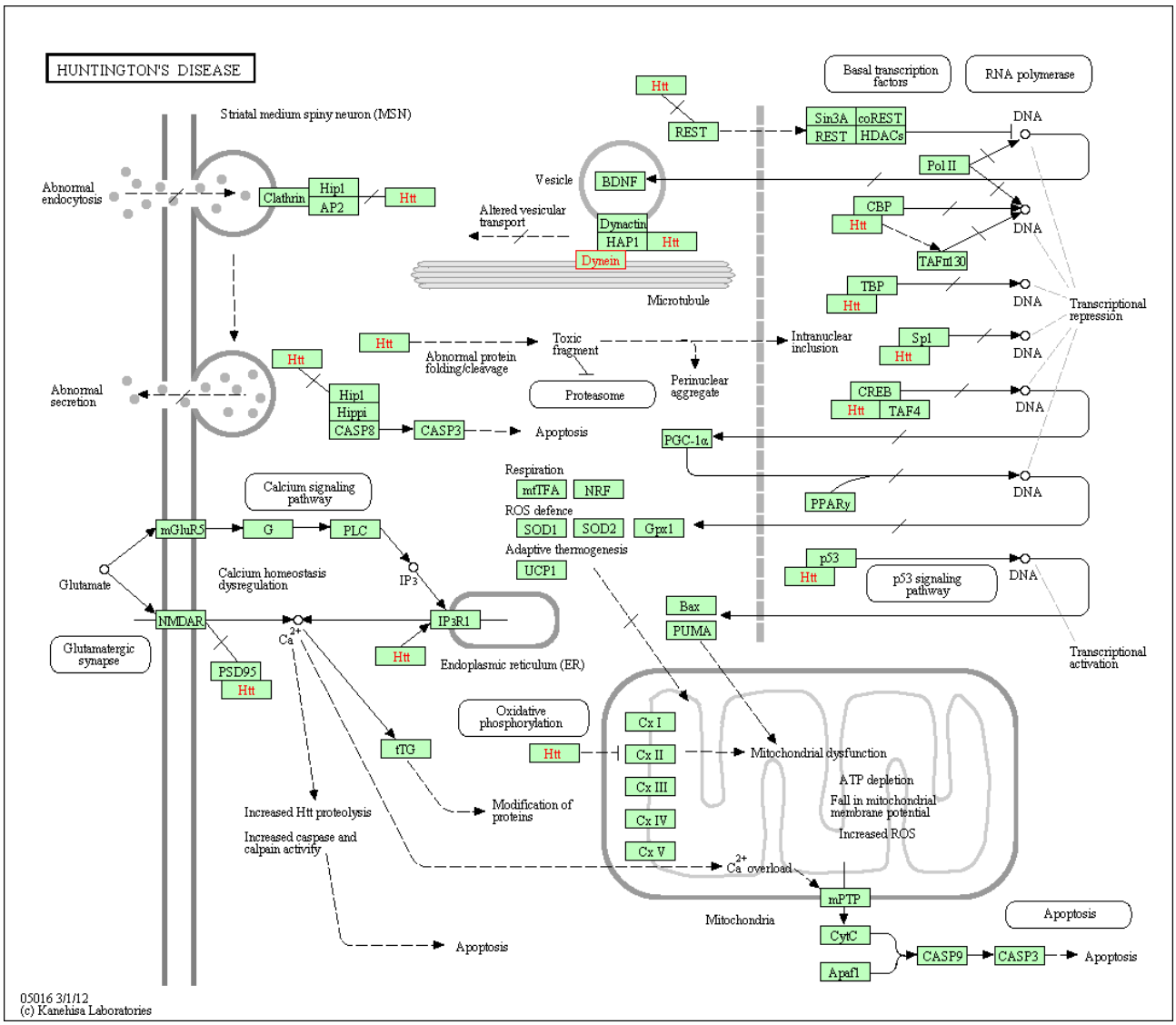


Figure 21. Voie métabolique de Huntington. Les nœuds rouges caractérisent les gènes impliqués dans des maladies du type Ciliopathie. Image générée à partir de la sous-banque de données spécialisée dans l'analyse des ciliopathies (instance SM2PH–Ciliopathy).

3.5 Données interactomiques

3.5.1 STRING

STRING (*Search Tool for the Retrieval of Interacting Genes/Proteins*) (Szklarczyk et al., 2011) 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. Cette banque est hébergée par le site (http://string-db.org) et disponible gratuitement à la communauté scientifique sous la forme de tables relationnelles.

Les données d'interactions sont entièrement pré-calculées 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., 2009), 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 d'expression :

- Conservation de gènes dans un voisinage proche : une suite de gènes conservée entre plusieurs génomes procaryotes ou à l'intérieur d'une distance chromosomique compatible avec un mécanisme de régulation commun peut indiquer une relation fonctionnelle.
- Fusion de gènes : il a été démontré que la présence de gènes fusionnés chez certains organismes indique potentiellement une étroite relation fonctionnelle entre ces gènes. Cette observation est utilisée pour prédire l'existence d'une interaction physique entre les protéines codées par les gènes homologues non fusionnés présents chez d'autres organismes.
- 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 directe ou indirecte.
- Coexpression : des associations fonctionnelles reflétées par des profils d'expression proches sont recherchées à l'intérieur des données d'expression de puces à ADN à l'aide du logiciel ArrayProspector (Jensen et al., 2004).

A 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 bien documentées et 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 22).

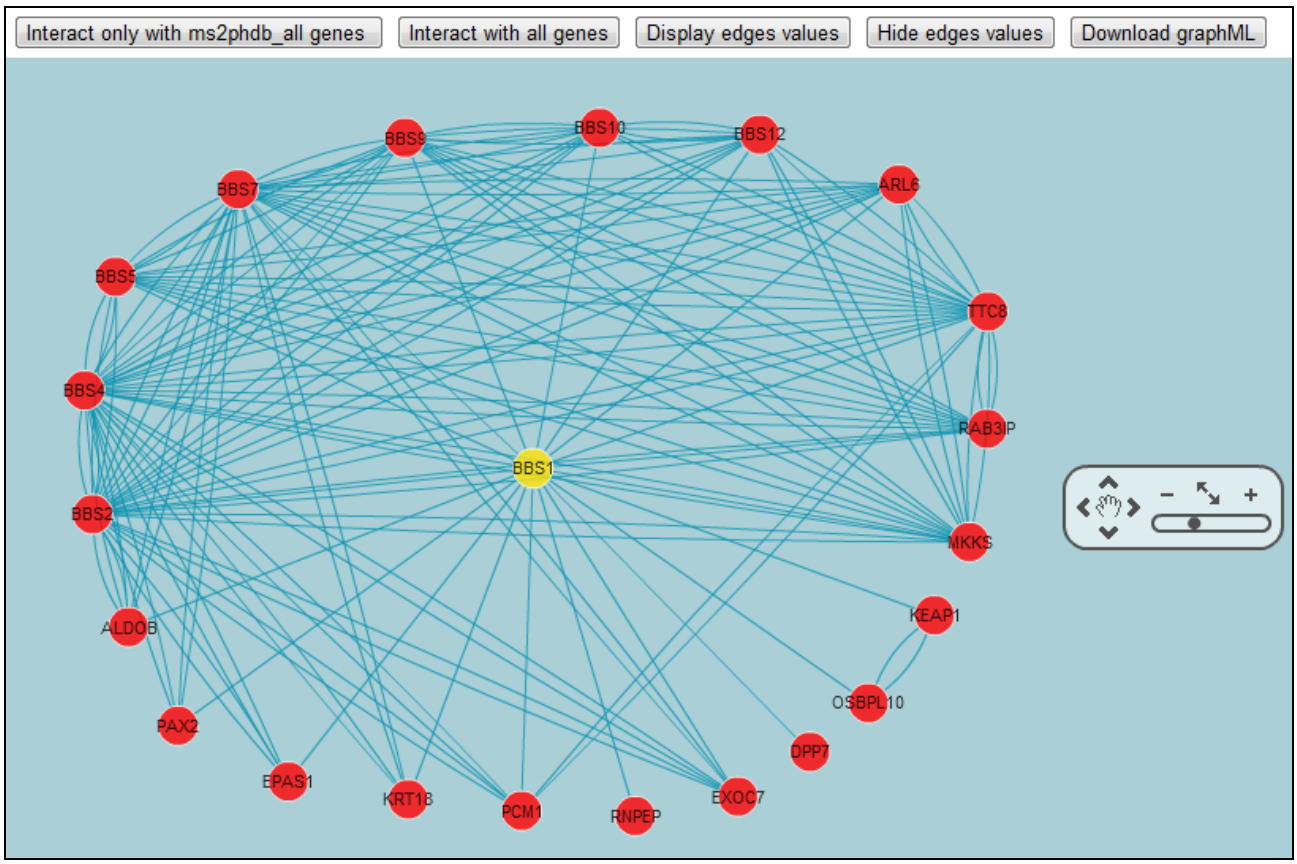


Figure 22. Visualisation d'un sous-graphe STRING rassemblant les interactants du gène BBS1 (Bardet-Biedl Syndrome 1). Seuls les interactants au premier degré de BBS1 sont représentés (seuil de confiance ≥ à 0,700). Chaque nœud du graphe représente une protéine et chaque arête une interaction. Image générée à partir du site SM2PH Central.

Enfin, les interactions connues ou prédites de STRING sont transférées entre organismes selon le principe des intérologues qui s'appuie sur la notion d'orthologie et suppose que si 2 protéines soient connues pour interagir dans un organisme, leurs homologues proches dans un autre organisme (orthologues) ont de fortes chances d'interagir également. 2 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 organisme 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.

SM2PH Central utilise des interactions de 20 199 protéines dans la version 9.0 actuelle de STRING.

3.5.2 Visualisation des interactions

Cytoscape (Shannon et al., 2003) (http://www.cytoscape.org/) est un logiciel *open-source* pour visualiser et manipuler des graphes. Ce logiciel est enrichi par de nombreux modules

d'analyse bioinformatique développés par la communauté internationale permettant d'exploiter les réseaux biologiques. Cytoscape permet notamment d'explorer les interactions d'un groupe de protéines ou de modéliser des régulations génétiques ou métaboliques.

A l'origine, Cytoscape est un logiciel natif à installer sur une machine. Une version simplifiée sous forme de librairie web est disponible. La version web est écrite en JavaScript, le rendu et l'interface sont gérés par le programme Flash. Cette version a été installée dans SM2PH Central pour montrer les interactions entre protéines humaines.

3.6 Données phénotypes

3.6.1 OMIM

OMIM (Amberger et al., 2009), pour *Online Mendelian Inheritance in Man*, est une banque de données qui présente le catalogue des gènes humains et des maladies associées. La force de cette banque réside dans la qualité et la diversité de son contenu. L'accès à des informations tant cliniques, génétiques que moléculaires y est possible. Chaque maladie génétique est répertoriée, décrite et annotée. Chaque entrée OMIM donne une description du syndrome, des informations sur son mode de transmission, le nom et la séquence du gène, un certain nombre de mutations ainsi que de nombreuses références vers des articles spécifiques. Cette banque est hébergée par le site (http://www.ncbi.nlm.nih.gov/omim) et disponible gratuitement à la communauté scientifique sous la forme de fichiers plats.

Au 28 août 2012, OMIM contenait 2 792 gènes dont la séquence est connue et auxquels au moins une mutation est associée (http://omim.org/statistics/geneMap). Plusieurs phénotypes distincts pouvant être associés à un même gène, selon la mutation qui l'affecte, le nombre de phénotypes, dont les bases moléculaires sont connues, est légèrement supérieur (3 485) dans le tableau des statistiques d'OMIM (Tableau 3).

Classification des entrées	**Autosome**	**Lié à l'X**	**Lié à l'Y**	**Mitochondrial**	**Total**
Gène avec séquence connue (>= 1 mutation)	13 177	641	48	35	13 901 **(2 792)**
Gène avec séquence connue et phénotype	145	5	0	2	152
Description phénotypique, bases moléculaires connues	3 191	262	4	28	**3 485**
Phénotype ou locus mendélien, bases moléculaires inconnues	1 632	135	5	0	1 772
Autres, principalement phénotypes avec bases mendéliennes soupçonnées	1 781	127	2	0	1 910
Total	19 926	1 170	59	65	21 220

Tableau 3. Statistiques d'OMIM.

3.6.2 HPO

La base de données OMIM gère la plus vaste collection des maladies humaines, mais sa description phénotypique (l'annotation des données phénotypiques) est peu structurée et ne convient pas à une analyse informatique automatisée. Cette absence de langage normalisé et contrôlé pour décrire les anomalies développementales est une des causes du retard de

l'informatisation de l'étude des malformations humaines (Oti et al., 2008) La construction de ce langage passe par le recours aux ontologies. Une ontologie décrit un domaine d'intérêt à l'aide de concepts et de relations clairement définis. Les concepts sont organisés de manière structurée (souvent une structure hiérarchique). Le sens d'un terme est utilisé de façon univoque. Les termes utilisés doivent être lisibles par des machines.

HPO (Robinson and Mundlos, 2010), pour *Human Phenotype Ontology*, est développé à partir des synopsis cliniques de OMIM dans un effort pour fournir une structure sémantique de l'ontologie des phénotypes humains. Il est constitué d'un ensemble d'environ 10 000 termes décrivant les caractéristiques des phénotypes humains, qui sont disponibles pour le téléchargement ou peuvent être consultés *via* le site Web (http://www.human-phenotype-ontology.org/). HPO est organisé comme un graphe orienté acyclique (en anglais *directed acyclic graph* ou DAG) dans lequel les termes représentent les sous-classes (cas plus spécifiques) de leur terme parent. Les termes qui sont situés près de la racine du graphe sont moins spécifiques que des termes qui sont plus loin d'elle. Dans l'extrait de l'HPO représenté dans la Figure 23, le terme « Syndactylie des 2ème-3ème doigts » est une sous-classe du terme « Syndactylie » dans le sens où la Syndactylie des 2ème-3ème doigts est une sorte spécifique de syndactylie. La racine de ce graphe est le terme « Anomalies des extrémités des membres ». Les données hiérarchisées de HPO sont très utiles dans l'aide au diagnostic des syndromes malformatifs. Elle permettent également de mesurer les distances entre les syndromes et de définir des familles de syndrome (*syndrome families*) ou agrégats phénotypiques (*phenotype clusters*) (Oti et al., 2009).

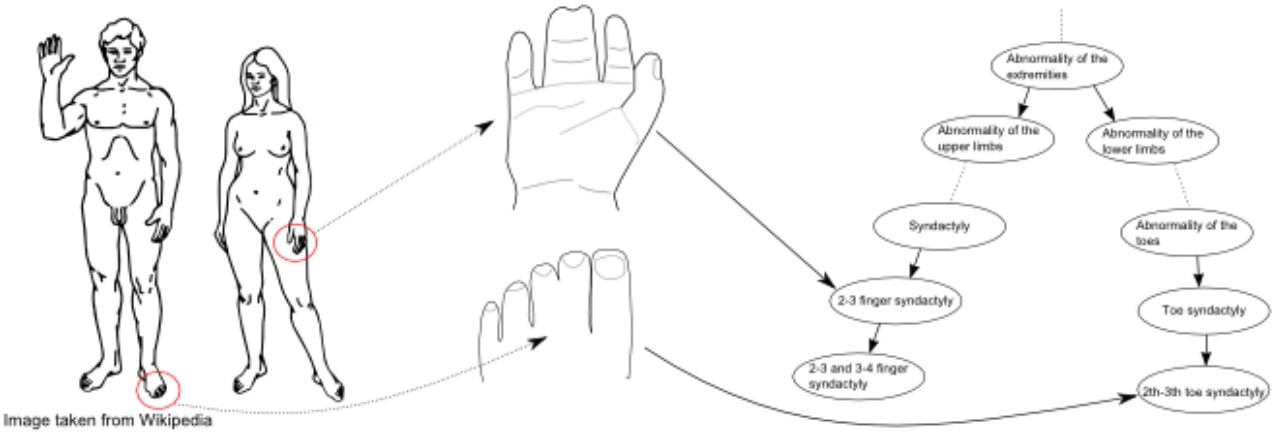


Figure 23. Une ontologie des phénotypes humains. Issue du site web HPO.

3.7 EvoluCode : Code-barres évolutionnaires

EvoluCode (Linard et al., 2012), développé par notre laboratoire, est un nouveau formalisme intégrant différents paramètres liés à l'évolution dans un format unique : les code-barres évolutionnaires. Cette nouvelle approche permet de faciliter l'analyse et la visualisation de l'histoire évolutive de complexes protéiques à l'échelle du génome. La visualisation (Figure 24) permet de faire ressortir le caractère atypique des protéines au cours de l'évolution. Une base de données contenant toutes les données EvoluCode est disponible à l'adresse : http://lbgi.igbmc.fr/barcodes.

EvoluCode prend actuellement en compte dix paramètres calculés et estimés automatiquement pour toutes les protéines en confrontant les protéines de l'homme (qui sert de référence) et l'ensemble des protéines orthologues de seize organismes vertébrés. Parmi ces paramètres, sept sont déduits des alignements multiples des séquences complètes ou MACS (*Multiple*

Alignment of Complete Sequences) : i) la longueur (*length*) de la séquence, ii) la différence entre la longueur (*length_difference*) de la séquence de référence et la séquence du vertébré considéré, qui permet de souligner un événement génétique tel que la perte/gain d'exons/domaines ou une erreur de prédiction, iii) le nombre de régions (*no_of_regions*) conservées et partagées entre la séquence de référence et la séquence du vertébré, iv) le pourcentage d'identifié (*sequence_identity*) entre la séquence de référence et la séquence de vertébré, v) le nombre de domaines (*no_of_domains*) connus dans la séquence du vertébré, basé sur les annotations PFAM (Finn et al., 2010), vi) la conservation des domaines (*domain_conservation*) sous forme d'un paramètre qualitatif indiquant la présence éventuelle de changements entre la structure de la protéine orthologue du vertébré considéré et la protéine de référence, qui permet de mettre en évidence la réorganisation/perte/gain de domaines et vii) le caractère hydrophile moyen (*hydrophilicity*) de la séquence du vertébré.

2 autres paramètres sont extraits de la base de données OrthoInspector (Linard et al., 2011) : i) le nombre de paralogues (*inparalog*) de l'organisme de référence par rapport à l'organisme vertébré, qui donne une indication sur les duplications récentes du gène de référence et ii) le nombre de co-orthologues (*co-ortholog*) de l'organisme vertébré par rapport à l'organisme de référence, qui donne une indication sur les duplications dans la lignée non-humaine. Et enfin, le dernier paramètre (*synteny*) qui indique si la synténie, c'est-à-dire l'ordre des gènes observés sur le génome de référence, est conservée (d'un ou des 2 côtés du gène considéré) ou non.

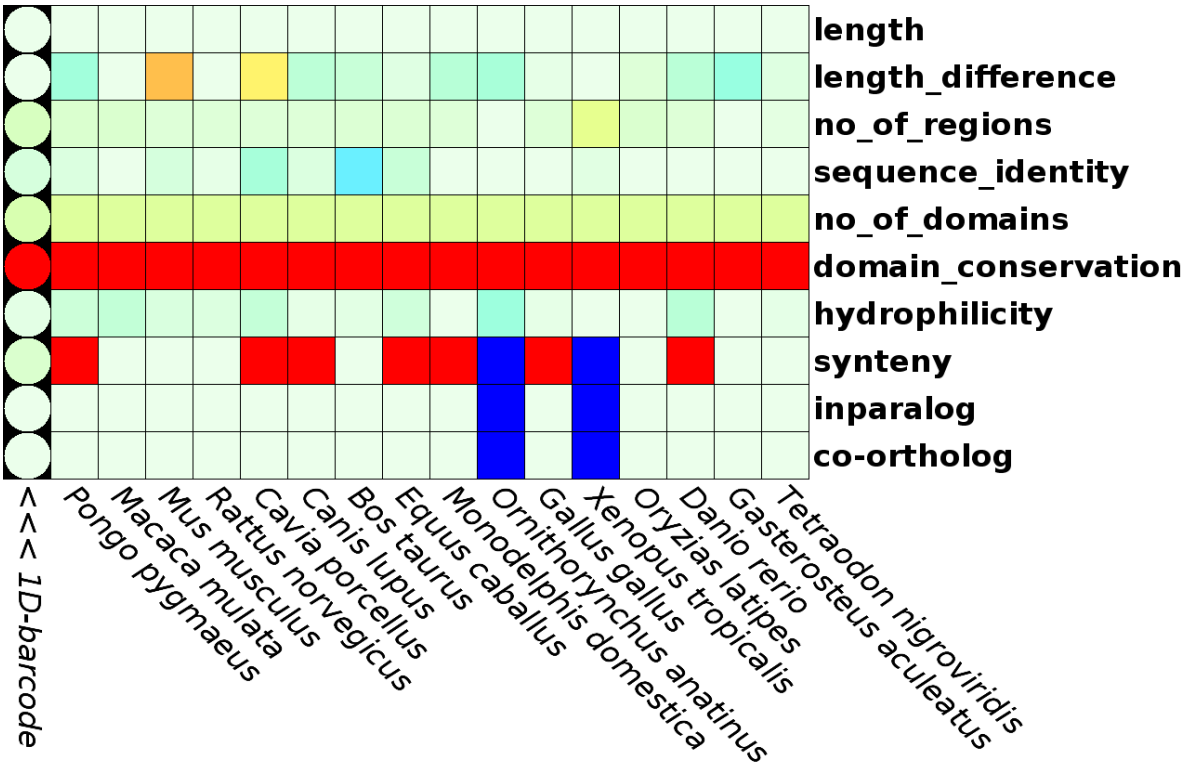


Figure 24. Visualisation d'un code-barre évolutionnaire (EvoluCode) sous sa forme 2D pour le gène LIPC impliqué dans la Dégénérescence Maculaire Liée à l'Âge. Le code-barre 1D est sur la gauche. Les couleurs bleues et rouges correspondent aux valeurs atypiques pour les différentes paramètres attribués selon une courbe gaussiennes (rouge = paramètre atypiquement fréquent, bleu = paramètre atypiquement rare).

Grâce à un traitement statistique et un code couleur dédié, la visualisation (Figure 24) permet de faire ressortir le caractère atypique des protéines au cours de l'évolution. L'ensemble de ces paramètres est ensuite décrit statistiquement pour estimer quelles valeurs peuvent être

considérées comme atypique par rapport à ce qui est généralement observé dans une espèce particulière par rapport à l'homme.

L'ensemble de ces paramètres est représenté dans une matrice (code-barre 2D) (Figure 24). Le code-barre 1D correspond à la valeur moyennée de chaque paramètre par rapport aux organismes présents pour le code-barre 2D. Dans un premier temps, ces code-barres 1D sont intégrés dans SM2PH Central.

3.8 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 2 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.

3.8.1 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.

3.8.2 BIRD-QL

Les données hétérogènes intégrées dans BIRD (voir section 3.1) sont représentées sous plusieurs tables relationnelles. L'exploitation de ces données par des requêtes SQL (*Structured Query Language*) n'est pas évidente sauf pour les développeurs ou utilisateurs experts du domaine. Pour cette raison, BIRD s'est équipé d'un langage et d'un moteur de requête de haut niveau : BIRD-QL, qui permet aux biologistes de fouiller facilement des données de BIRD sans avoir besoin de connaissances en base de données. L'utilisateur peut développer et envoyer des requêtes BIRD-QL (voir l'exemple dans la Figure 25) *via* des commandes http ou des scripts pour rechercher des séquences intéressantes et recevoir des données sous différents formats. Ce moteur est également capable de générer rapidement et à la volée des sous-banques de données selon les besoins des applications.

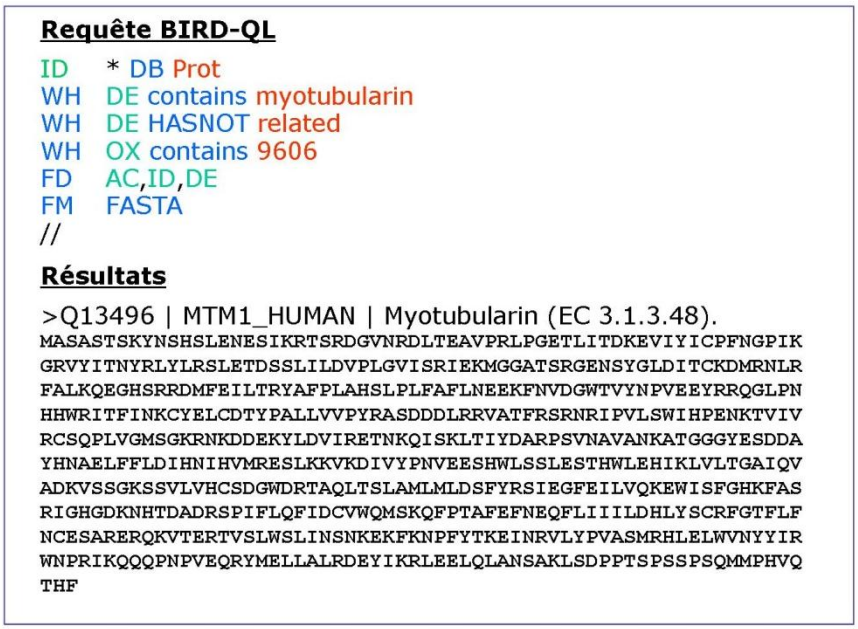


Figure 25. Exemple de requête BIRD-QL. En bleu, le « dialecte » propre à BIRD-QL, en vert les champs de la banque interrogée et en rouge les conditions spécifiées par l'utilisateur. La requête présentée interroge les banques de séquences protéiques (Prot) et demande de retourner, au format FASTA (dont la ligne de description contient le numéro d'accession de la séquence AC, l'identifiant ID et la définition DE), les protéines humaines (dont l'identifiant de taxonomie est 9606) dont la définition DE contient le mot «myotubularin», mais pas le mot « *related* ».

La structure de ce langage permet aux utilisateurs de manipuler facilement ses requêtes et de construire des conditions de recherche selon la logique booléenne. L'utilisateur peut ainsi filtrer les données sélectionnées en ajoutant des conditions au fur et à mesure. Ces conditions sont entièrement indépendantes les unes des autres et sont reliées par les opérateurs logiques AND. Le Tableau 4 donne le nombre possible d'utilisations de chaque clé et les contraintes dans une requête BIRD-QL.

Clé	Nom	Cardinalité	Exemple
ID	Séquence id, protéine id, query id	1	ID Q92PK5
WH	Où	0..N	WH OS contains «Homo sapiens»
WH BLAST	Contrainte blast	0..N	WH Blast score >1
WH PATTERN	Motifs/Fonctions scientifiques prédéfinies dans BIRD	0..N	StructureDistance(PDB_id,x,y,z) >10
FD	Paramètres de sortie ou une des valeurs d'une fonction scientifique	0..1	FD AC,ID,SQ FD Get_Count(ID) FD GetStatitic(Taxid,OS,)
OFFSET	OFFSET indique de passer ce nombre de lignes avant de renvoyer les lignes restantes	0..1	OFFSET 20
LM	Limite	0..1	LM 10
FM	Format	0..1	FM fasta/embl/genbank/xml

Tableau 4. Liste des clés de BIRD-QL et exemple de requête.

Un des avantages de BIRD-QL est lié au fait que l'on peut envoyer cette requête à n'importe quelle banque de données hébergée ou à plusieurs instances SM2PH à la fois par des Web Services ou par une API (*Application Programme Interface*). Cette requête ne dépend pas des données intégrées (noms des banques de données, tables relationnelles) dans chaque base BIRD.

Suivant une logique purement objet, on imaginerait le traitement des requêtes se faisant sur des collections d'objets persistants et retournant des listes d'instance venant de différentes classes (format hétérogène : UniProt, PDB, etc.). Cette approche serait pénalisante en termes de mémoire et de vitesse. Pour être plus performant, le moteur de requêtes BIRD-QL contourne la couche objet. Il accède directement à la base de données par SQL et ne retourne que des OIDs (*Object Identifier*). Cela permet de ne pas faire transiter sur le réseau ou en mémoire des données inutiles. On peut ainsi facilement faire des requêtes retournant un grand nombre de données sachant que ne seront copiés en mémoire que les objets réellement lus. Ce mécanisme nous permet d'implémenter facilement les d'interfaces graphiques lisibles dans SM2PH Central et MSV3d et aide les utilisateurs à dialoguer plus facilement avec les bases générées.

3.9 PipeAlign : un outil d'analyse des protéines

Comme nous l'avons vu précédemment (par exemple EvoluCode), une grande partie des informations évolutives utilisées dans nos analyses est basée sur l'exploitation d'alignements multiples de séquences complètes, que nous appelons couramment MACS, pour *Multiple Alignment of Complete Sequences*. Nous décrirons dans cette section la méthodologie et les outils adoptés pour la construction d'alignements multiples de bonne qualité et leur analyse (Friedrich, 2007).

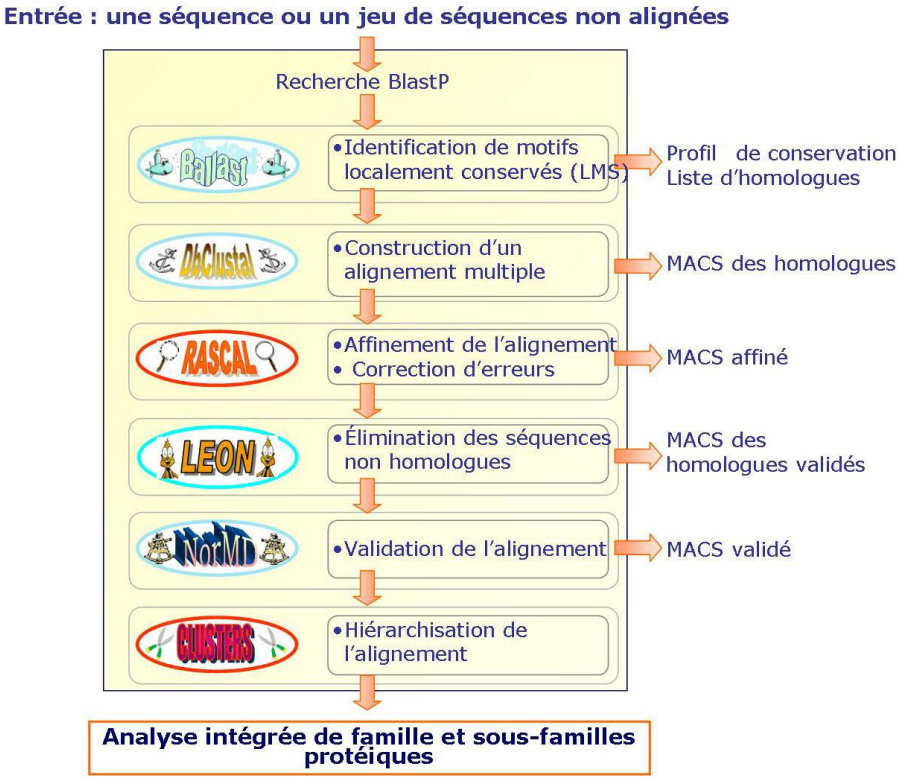


Figure 26. Aperçu de la cascade de programmes constituant PipeAlign. Adapté de (Plewniak et al., 2003) par (Friedrich, 2007).

Nous avons construit les MACS en utilisant la cascade logicielle PipeAlign (Plewniak et al., 2003), un outil d'analyse de famille de protéines, développé au sein du laboratoire. PipeAlign est composé d'une suite de programmes d'analyse de séquences (Figure 26) et permet, la construction automatique d'un MACS de qualité, à partir d'une protéine d'intérêt ou d'une série de séquences. La première étape du PipeAlign consiste à rechercher les protéines similaires à la protéine requête à l'aide du programme BLASTP. Les différents programmes de PipeAlign sont décrits brièvement dans les paragraphes suivants.

3.9.1 Ballast : traitement des résultats des recherches BLASTP

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

3.9.2 DbClustal : construction du MACS

DbClustal (Thompson et al., 2000) est un programme d'alignement multiple de séquences complètes qui conjugue les avantages des algorithmes d'alignement global et d'alignement local. En effet, le programme d'alignement multiple ClustalW (Thompson et al., 1994), basé sur l'algorithme d'alignement global développé par Needleman et Wunsch (Needleman and Wunsch, 1970), a longtemps été utilisé au laboratoire pour la construction des MACS mais il présente les inconvénients des méthodes basées uniquement sur l'alignement global à savoir, des performances médiocres pour aligner des séquences contenant de longues insertions ou des extensions N-terminales ou C-terminales.

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

3.9.3 RASCAL : correction des alignements

A ce jour, aucun algorithme d'alignement de séquences complètes n'est en mesure de réaliser un alignement optimal de tous types de séquences. Aussi est-il fréquent que des erreurs d'alignements entre divers segments soient introduites. Le programme RASCAL (Thompson et al., 2003), pour *RApid Scanning and Correction of ALignment errors*, a été développé pour détecter ces erreurs d'alignement et les corriger. L'alignement multiple obtenu en sortie du programme DbClustal est divisé horizontalement et verticalement pour former un « quadrillage » au sein duquel les régions bien alignées, et donc fiables, peuvent être différenciées. Les erreurs potentielles d'alignement sont détectées en comparant les profils des régions fiables. RASCAL réaligne les régions les moins fiables ou mal alignées en utilisant un algorithme proche de celui utilisé dans ClustalW. La correction de l'alignement est restreinte aux régions les moins fiables, permettant ainsi une stratégie de ré-affinement plus performante.

3.9.4 LEON : extraction des séquences non homologues

Un alignement multiple n'ayant de sens que si les séquences protéiques alignées sont homologues, le programme LEON (Thompson et al., 2004), pour *multiple aLignment Evaluation Of Neighbours*, a été mis en place pour détecter, au sein d'un MACS, les séquences n'appartenant pas à la famille d'intérêt. L'extraction des séquences non homologues 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 ainsi d'un MACS de bonne qualité qui ne contient que des séquences partageant au moins une région homologue à la séquence d'intérêt.

3.9.5 NorMD : évaluation de la qualité d'un MACS

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

Le score assigné à l'alignement est normalisé par rapport au nombre de séquences que cet alignement contient, leur pourcentage d'identité, leur longueur, etc. Ceci permet de comparer les scores NorMD d'alignements indépendants. Le score assigné par NorMD est généralement compris entre 0 et 1. Plus le score tend vers 1, plus la qualité de l'alignement est considérée comme satisfaisante. Un score inférieur à 0,3 est considéré non satisfaisant.

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

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

La Figure 26 récapitule les différentes étapes réalisées par PipeAlign. La sortie finale de la suite de programme PipeAlign est un MACS validé de haute qualité, dans lequel les séquences sont classées en sous-familles potentiellement fonctionnelles.

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

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

MACSIMS (*MACS Information Management System*) (Thompson et al., 2006) est un système d'annotation des alignements multiples basé sur l'ontologie MAO (*Multiple Alignment Ontology*) (Thompson et al., 2005). L'annotation se fait de manière automatique en intégrant différents types de données provenant des bases de données biologiques ou de méthodes de prédiction *ab initio*. Parmi les banques de données consultées on peut citer : UniProt, PDB et InterPro. Les prédictions faites à partir de la séquence primaire donnent, entre autres, les régions transmembranaires, les peptides signaux, les régions de faible complexité, etc. Les annotations de l'ontologie GO sont également incluses. Une étape de validation croisée des données permet de conserver les informations avérées qui sont ensuite propagées aux séquences non annotées. Un résumé des étapes intégrées à MACSIMS est fourni dans la Figure 27.

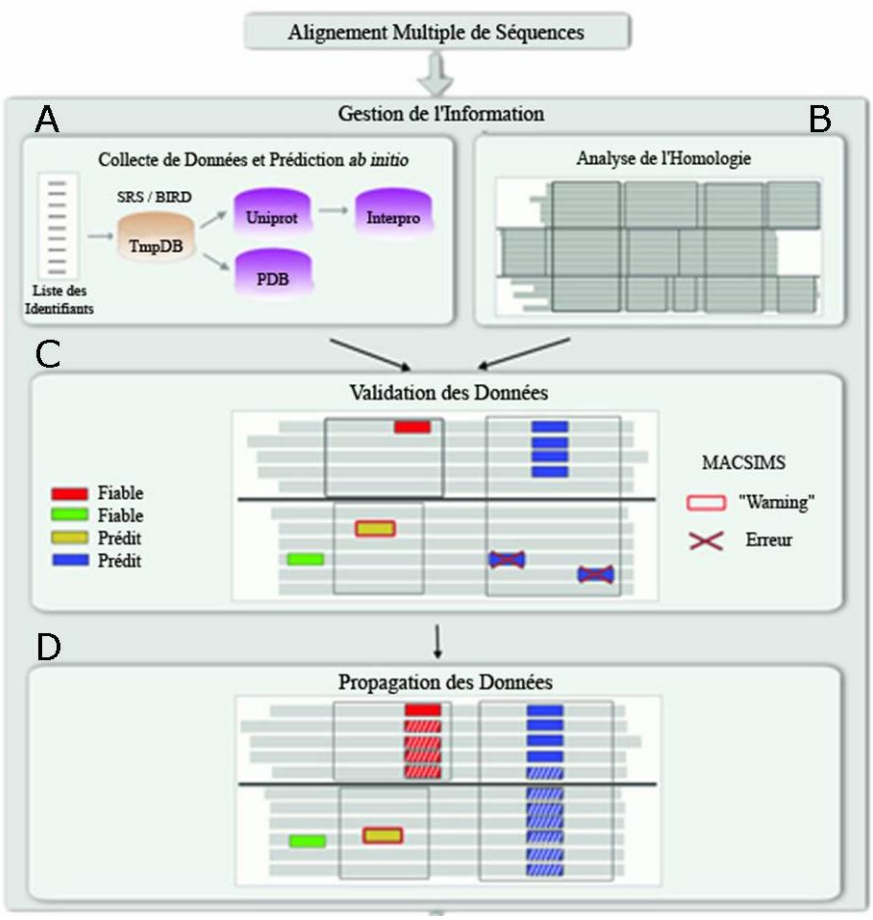


Figure 27. Etapes successives de MACSIMS. Pour chaque séquence de l'alignement, des informations sont collectées au sein de banques de données publiques et certaines caractéristiques sont prédites *ab initio* (A). Les blocs de conservation de l'alignement sont déterminés en parallèle (B). Les données recueillies fiables sont validées, les autres sont éliminées (C). Les données fiables sont ensuite transférées des séquences annotées aux séquences inconnues (D). Adaptée de (Thompson et al., 2006) par (Friedrich, 2007).

L'alignement annoté est donné en sortie sous la forme d'un fichier XML (*eXtensible Markup Language*), un format structuré permettant une exploitation informatique automatique à haut débit. L'éditeur d'alignements multiples Jalview (Waterhouse et al., 2009) a été adapté pour permettre la visualisation, de manière simple et conviviale, des alignements annotés.

3.11 Analyse structurale des protéines

3.11.1 Modeller : construction de modèles par homologie

Modeller (Eswar et al., 2008) permet de construire des modèles 3D de protéines grâce à une méthode de minimisation des contraintes spatiales, qu'il définit en terme de « fonction de probabilités de densité ». Modeller est souvent utilisé pour la modélisation comparative. L'utilisateur peut fournir un alignement comprenant la séquence qu'il veut modéliser (séquence cible) et une ou plusieurs structures de référence (empreinte). Le processus de calcul automatique du modèle 3D comporte 3 étapes. Dans la première étape, Modeller extrait les contraintes spatiales (contraintes de distance des liaisons, des angles, des angles dièdres et des contacts entre les atomes non liés) pour les résidus identiques entre la séquence cible et la (ou les) empreinte choisie, à partir des informations trouvées dans l'alignement structural. Dans la seconde étape, il estime ces contraintes, soit à partir de la structure de référence (pour les résidus conservés dans l'alignement), soit à partir de celles calculées d'après une base de données constituée de 105 alignements contenant 416 protéines de structures connues (pour les résidus non conservés). Ainsi, les « relations » extraites, exprimées en terme de « densité de probabilités conditionnelles », sont utilisées comme des contraintes spatiales. Par exemple, les probabilités pour les différentes valeurs des angles dièdres de la chaîne principale sont calculées en fonction du type de résidu considéré, de la conformation du résidu équivalent et de l'identité de séquence entre les 2 protéines. Dans la dernière étape, Modeller combine ces contraintes spatiales avec les termes du champ de force CHARMm dans une fonction qu'il optimise pour construire le « modèle 3D tout atome ». Pour chaque étude de modélisation, nous avons généré 5 modèles et conservé celui présentant l'énergie la plus basse.

3.11.2 Visualisation et mise en forme des structures 3D

Nous avons utilisé le logiciel PyMOL (http://pymol.sourceforge.net) et l'application Jmol (http://jmol.sourceforge.net) pour visualiser nos structures et modèles 3D et les analyser. Jmol présente l'avantage d'être disponible sous forme d'applet et peut ainsi être intégré au sein de pages Web, sans que l'utilisateur n'ait à réaliser d'installation préalable.

Des fonctions permettent de visualiser et d'explorer les différentes parties ou propriétés d'une molécule (comme les charges, les résidus hydrophobes et polaires). Ces 2 outils sont capables d'importer tous les formats de modèles moléculaires disponibles et peuvent être contrôlés au moyen de scripts. La qualité et la diversité des représentations disponibles ainsi que les possibilités offertes à un utilisateur averti rendent ces outils indispensables à l'analyse structurale.

CHAPITRE 4. PROGRAMMATION LOGIQUE INDUCTIVE

« Knowledge is power. »

Sir Francis Bacon (1561 - 1626)

La programmation logique inductive (PLI) (Muggleton, 1991) se fonde sur les principes de la programmation logique pour induire des hypothèses à partir d'exemples et de connaissances du domaine. Dans ce chapitre, nous allons tout d'abord présenter quelques principes et notions de programmation logique (section 4.1) puis, la PLI (section 4.2). Dans la section 4.3, nous allons aborder la problématique de la structuration de l'espace des hypothèses. Différents biais d'apprentissage (section 4.4) et différentes stratégies de recherche (section 4.5) sont introduits pour orienter et limiter les recherches de l'espace des hypothèses. Un système de PLI qui s'appelle Aleph est présenté dans la section 4.6. Nous terminerons ce chapitre en présentant divers travaux appliqués à la biologique.

4.1 Rappels sur la Programmation Logique

Nous commençons cette section par quelques rappels de logique et de programmation logique (Lloyd, 1987) qui vont nous permettre de préciser les notations utilisées dans ce document.

En programmation logique inductive, le langage de description des exemples, connaissances du domaine et des hypothèses est la logique du premier ordre. Cette logique est un langage formel particulièrement efficace pour exprimer les relations complexes entre individus et pour déduire de nouvelles relations à partir des relations connues. Ces notions sont détaillées dans le paragraphe suivant.

4.1.1 La syntaxe de la logique du premier ordre

La syntaxe de la logique du premier ordre est l'ensemble de toutes les formules construites à partir des symboles primitifs. Nous disposons de 6 classes de symboles primitifs :

- Constantes. Une constance est le nom d'un objet précis, par convention le nom d'une constante commence par une lettre minuscule ou un chiffre. Exemples : 1, 2, vrai, faux, lysine, lys, gpr179.
- Variables. Une variable prend ses valeurs parmi les éléments d'un domaine. Des exemples de domaines sont : l'ensemble des nombres entiers, l'ensemble des acides aminés. Par convention, le nom d'une variable commence par une lettre majuscule. Exemples : X, Y, Z, Acide, Gene.
- Fonctions. Une fonction possède une arité (le nombre d'arguments) qui doit valoir au moins 1. Une fonction d'arité 0 n'est autre qu'une constante.
- Prédicats. Un prédicat est une fonction dans {vrai, faux}f. Il est représenté par un symbole muni d'une arité, par exemple acide_amine/1.
- Connecteurs : $\neg$ (non), $\wedge$ (et), $\vee$ (ou), $\Rightarrow$ (implique), $\Leftrightarrow$ (équivalent).
- Quantificateurs : $\forall$ (le quantificateur universel) et $\exists$ (le quantificateur existentiel).

Définition (Terme). Un *terme* se définit récursivement comme suit:

- Des constantes et des variables sont des termes.
- f(t1,...,tn) est un terme si f est une fonction à n arguments et t1,...,tn sont des termes.

Définition (Atome). Un *atome* est un prédicat appliqué à des termes. Exemple : acide_amine(lysine) est un atome du prédicat acide_amine/1.

Définition (Littéral). Un *littéral* est un prédicat appliqué à des termes, éventuellement précédé du symbole de négation $\neg$. Exemples : acide_amine(cystéine) est un littéral, $\neg$acide_amine(cytosine) aussi.

Définition (Formule). Une *formule* se définit récursivement comme suit :

- Un littéral est une formule.
- Si f est une formule alors $\neg$f est une formule.
- Si f et g sont des formules alors ($f \wedge g$), ($f \vee g$), ($f \Rightarrow g$) et ($f \Leftrightarrow g$) sont des formules.
- Si f est une formule et x est une variable, alors (($\forall x$)f) et (($\exists x$)f) sont des formules.

Définition (Clause). Une *clause* est une formule de type particulier : c'est une disjonction finie de littéraux dans laquelle toutes les variables sont quantifiées universellement.

On simplifie généralement l'écriture de ces clauses en supprimant les quantificateurs universels. Ainsi, $\forall X \forall Y(f(X) \vee g(Y))$ s'écrira simplement $f(X) \vee g(Y)$.

Par exemple : $\forall X$(molecule(X) $\vee \neg$acide_amine(X)) est une clause et elle s'écrira simplement molecule(X) $\vee \neg$acide_amine(X)

Définition (Clause de Horn). On appelle *clause de Horn* toute clause comportant au plus un littéral positif.

Par exemple :

$$\text{molecule}(X) \vee \neg\text{acide\_amine}(X)$$

$$\text{acide\_amine}(\text{lysine})$$

Définition (Clause définie). Une *clause définie* est une clause de Horn qui a exactement un littéral positif. Donc, une clause définie s'écrit :

$$A_0 \vee \neg A_1 \vee \neg A_2 \vee \neg A_3 \vee \ldots \neg A_n$$

Une clause définie peut se récrire sous forme de règle :

$$A_0 \leftarrow A_1 \wedge A_2 \wedge A_3 \wedge \ldots \wedge A_n$$

Par exemple :

$$\text{molecule}(X) \leftarrow \text{acide\_amine}(X)$$

Le connecteur $\leftarrow$ signifie l'implication du membre de gauche par le membre droit. A_0 est appelé tête de la clause, les autres atomes formant son corps.

En remplaçant la conjonction par une virgule et le connecteur $\leftarrow$ par le symbole :-, on obtient la syntaxe utilisée en Prolog :

$$A_0 :- A_1 , A_2 , A_3 , \dots , A_n.$$

Par exemple :

molecule(X) :- acide_amine(X)

Définition (Programme logique défini). Un *programme logique défini*, ou pour simplifier un *programme logique*, est un ensemble de clauses définies.

4.1.2 Raisonnement en logique du premier ordre

Etant donné un programme logique P, le but est de faire des raisonnements à partir de ce programme afin de savoir, par exemple, quels faits sont vrais étant donné P. Considérons l'exemple suivant où chacune des propositions est accompagnée de sa représentation Prolog :

(A) Les acides aminés sont des molécules (règle).

molecule(X) :- acide_amine(X)

(B) La lysine est un acide aminé (observation).

acide_amine(lysine)

(C) La lysine est une molécule (conséquence).

molecule(lysine)

La PLI utilise 2 types de raisonnement : la déduction et l'induction. Ils sont décrits comme suit :

- La déduction consiste à tirer une conséquence (C) à partir d'une règle générale (A) et d'une observation empirique (B).
- L'induction consiste à trouver une règle générale (A) qui pourrait rendre compte de la conséquence (C) si l'observation empirique (B) était vraie.

Nous allons maintenant décrire les éléments logiques impliqués dans le raisonnement que nous allons aborder dans les paragraphes suivants.

4.2 Cadre général de la Programmation Logique Inductive

Alors que la programmation logique concerne l'inférence déductive, comme son nom l'indique, la PLI concerne préférentiellement l'inférence inductive. Le but de la PLI est de créer des outils et des techniques permettant de trouver des hypothèses à partir d'un ensemble d'observations. Il s'agit de synthétiser des règles à partir d'observations et d'une base de connaissances.

Dans (Lavrac and Dzeroski, 1994), le problème de la PLI est formulé comme suit :

Entrées :

+ Une base de connaissances *B* (théorie du domaine), c'est-à-dire un ensemble de clauses qui reflètent la connaissance *a priori*.

+ Un ensemble *E* composé d'exemples positifs E^+ et d'exemples négatifs E^-.

Sorties:

Trouver une hypothèse *H* telle que les conditions de cohérence suivantes soient

satisfaites :

+ Complétude : $\forall e^+ \in E^+$, $H \wedge B \models e^+$

+ Consistance : $\forall e^- \in E^-$, $H \wedge B \not\models e^-$

B, *E* et *H* sont ici des ensembles de clauses de Horn.

On cherche donc une hypothèse *H* qui couvre tous les e^+ (complétude) et ne couvre aucun e^- (consistance).

Par exemple, on se pose le problème d'apprendre la relation *is_deleterious*(X), qui se lit : X est une mutation délétère, à partir des relations *modification_size*, *conservation* et *gain_contact*. Il y a 2 exemples positifs et 2 exemples négatifs. Tous sont donnés dans le Tableau 5.

exemples positifs	**base de connaissances**
is_deleterious(o15305.p.Glu139Lys). is_deleterious(q9NRR6.p.Arg512Trp).	modification_size(o15305.p.Glu139Lys, increase) conservation(o15305.p.Glu139Lys, rank1) gain_contact(o15305.p.Glu139Lys, phob) gain_contact(o15305.p.Glu139Lys, arom) modification_size(q9NRR6.p.Arg512Trp, unchanged) modification_size(p27930.p.Glu292Lys, increase) conservation(p27930.p.Glu292Lys, no_conservation) …
exemples négatifs	
is_deleterious(p27930.p.Glu292Lys). is_deleterious(q96JF6.p.Ile171Thr).	

Tableau 5. Un exemple du problème PLI dans la classification des mutations délétères/neutres.

Un programme PLI est capable de trouver une hypothèse H, par exemple :

H = is_deleterious(X) :- conservation(X, rank1)

Cette nouvelle relation se lit que la mutation X est délétère si la position mutée est bien conservée.

Muggleton et Raedt ont proposé un squelette pour les algorithmes de PLI (Muggleton and Raedt, 1994). Le pseudo code de ce squelette est illustré ci-après (Figure 28) :

Début

Initialiser H

Tant que la condition d'arrêt de H n'est pas satisfaite **faire**

Retirer h de H

Choisir des règles d'inférence r_1, …, r_k à appliquer à h

Appliquer les règles r_1, …, $r_k \in R$ à h pour obtenir h_1, …, h_n. Ajouter h_1, …, h_n à H.

Elaguer H

Fin tant que

Fin

Figure 28. Algorithme générique de PLI.

L'algorithme fonctionne comme suit. On prend un ensemble d'hypothèses candidates H. De façon répétitive, on supprime une hypothèse h de H. On applique à h des règles d'inférences (inductives ou déductives) r_1, …, r_k qui produisent de nouvelles hypothèses h_1, …, h_n que l'on rajoute à H. On élague H et on recommence jusqu'à ce que l'on remplisse la condition d'arrêt. Cette condition d'arrêt est souvent le simple fait d'avoir trouvé une solution i.e. une hypothèse qui couvre tous les exemples positifs et aucun exemple négatif.

Dans la suite, nous présentons le détail des algorithmes de PLI. Dans la section 4.3, la structuration de l'espace des hypothèses est établie à l'aide de la définition d'une relation de généralité entre clauses. Différents types de biais d'apprentissages qui permettent de limiter la taille de l'espace de recherche sont introduits dans la section 4.4. Enfin, les différentes stratégies d'exploration de cet espace sont abordées dans la section 4.5.

4.3 Structuration de l'espace des hypothèses

La problématique de la PLI se ramène à une recherche dans un espace des hypothèses, nommé aussi espace de recherche, satisfaisant un certain nombre de propriétés. La recherche des hypothèses dans cet espace est guidée par une notion de généralité (notée $\succeq$) existant entre les clauses. Intuitivement, une hypothèse h1 est plus générale qu'une hypothèse h2 si l'ensemble des exemples couverts par h1 inclut l'ensemble des exemples couverts par h2. En PLI, la relation de généralité la plus communément utilisée est la θ-subsomption de Plotkin (Plotkin, 1970).

Définition (Substitution). Une *substitution* est un ensemble de couples variable/terme. Une *substitution* θ s'applique à une formule I ou une clause C en remplaçant chaque occurrence des variables par le terme correspondant. Par exemple, l'application de la substitution θ = {X/ o15305.p.Glu139Lys ; Y/rank1} à la clause C = conservation (X, Y) donne la clause Cθ = conservation(o15305.p.Glu139Lys, rank1).

Définition (θ-subsomption de 2 clauses). On dit qu'une clause C_1 θ-subsume une clause C_2, noté $C_1 \succeq_\theta C_2$ si et seulement s'il existe une substitution θ telle que $C_1\theta \subseteq C_2$. C_1 est alors une généralisation de C_2 (ou que C_2 est une spécialisation de C_1) par θ-subsomption. La θ-subsomption est une relation transitive entre clauses et permet la structuration de l'espace des hypothèses en un treillis.

4.4 Les biais de recherche dans l'espace des hypothèses

En pratique, l'espace des hypothèses en PLI est souvent très grand, voire infini. Il est donc important d'imposer des contraintes ou des restrictions sur les hypothèses pour en réduire la taille et garantir ainsi une certaine qualité de l'apprentissage. Ces restrictions sont appelées biais déclaratif puisque elles sont définies de façon déclarative par l'utilisateur. On distingue principalement 2 grandes familles au sein des biais déclaratifs : les biais syntaxiques et les biais sémantiques.

Les biais syntaxiques permettent de définir l'ensemble des hypothèses envisageables en spécifiant explicitement leur syntaxe. Il est possible de limiter le nombre de clauses induites ou de limiter le nombre de variables dans une clause ou encore de limiter la profondeur de la récursion dans les termes.

À l'opposé des biais syntaxiques, les biais sémantiques imposent des restrictions sur le sens des hypothèses. La déclaration de types et de modes permet de restreindre la taille de l'espace des hypothèses. Par exemple, nous pourrions limiter le prédicat habite_dans(Personne,Ville) à seulement nous donner la ville pour une personne déterminée et pas nous donner toutes les personnes qui vivent dans la ville donnée. Cela peut être fait avec la déclaration de mode habite_dans (+Personne, -Ville) où '+' signifie argument en entrée, et '-' argument en sortie.

Le système Aleph que nous utilisons dans nos travaux, permet de déclarer différents modes sur les variables des littéraux grâce aux déclaratives *modeh* et *modeb*. Le premier spécifie les prédicats pouvant apparaître en tête de clause et le second, ceux pouvant apparaître dans le corps de clause. Par exemple, dans la classification des mutations délétères et neutres, nous avons déclaré la tête des clauses avec le modeh comme suivant : modeh(1, is_deleterious(+mutation))

4.5 Exploration de l'espace des hypothèses

Les systèmes de PLI utilisent différents algorithmes pour parcourir l'espace des hypothèses. Il existe plusieurs stratégies de recherche dont les recherches descendante et ascendante qui sont présentées ci-dessous.

4.5.1 Recherche descendante

La recherche descendante (ou *top-down* en anglais) repose sur une exploration de l'espace des hypothèses de la plus générale à la plus spécifique en appliquant successivement des règles d'inférence déductive.

Définition (Règle d'inférence déductive). Une règle d'inférence déductive r fait correspondre une conjonction de clauses S à une conjonction de clauses G telle que G⊨S. r est une règle de spécialisation. Les règles de spécialisation utilisées dans la PLI sont :

- l'ajout de littéral à la clause,
- l'application d'une substitution, afin de transformer des variables de la clause en constantes ou d'unifier plusieurs variables.

La recherche descendante explore l'espace des hypothèses comme suit : à chaque étape de l'exploration, on recherche une clause *h* couvrant un maximum d'exemples de E^+ et pas ou peu d'exemples de E^-. Pour effectuer cette recherche, on part de l'hypothèse la plus générale du concept (la clause vraie) puis, on la spécialise à l'aide des règles de spécialisation qui proposent toutes les clauses qui lui sont plus spécifiques selon la notion de généralité retenue. Le choix des spécialisations suivies peut se faire à l'aide d'heuristiques, souvent calculées à l'aide des taux de couverture des exemples. A la fin, la clause *h* retenue couvre un maximum d'exemples de E^+ et aucun/peu exemple E^-. Il suffit, ensuite, d'enlever à E^+ les exemples couverts par *h* et de recommencer la recherche d'une nouvelle clause. L'algorithme s'arrête quand E^+ est vide, on obtient ainsi l'hypothèse H qui est un ensemble de clauses et qui couvre tous les exemples positifs et aucun/peu exemple négatif.

On peut citer des logiciels qui utilisent cette approche comme le précurseur MIS de Shapiro (Shapiro, 1981), FOIL (Quinlan and Cameron-Jones, 1993), Progol (Muggleton, 1995) et TILDE (Blockeel and Raedt, 1998).

4.5.2 Recherche ascendante

A l'inverse de la méthode précédente, la recherche ascendante (ou *bottom-up* en anglais) correspond à une exploration des hypothèses des plus spécifiques aux plus générales. Ils considèrent les exemples et généralisent itérativement l'hypothèse recherchée en appliquant des règles d'inférence inductive et la notion des moindres généralisés. Ils sont dit *guidés par les exemples*.

Définition (Règle d'inférence inductive). Une règle d'inférence inductive r fait correspondre une conjonction de clauses S à une conjonction de clauses G telle que G⊨S. r est une règle de généralisation.

Définition (moindre généralisé). Un *moindre généralisé* G de H_1, …, H_n doit vérifier :

- $\forall H_i : G \geq H_i$
- $[\exists G' : (\forall H_i : G' \geq H_i) \wedge (G \geq G')] \Rightarrow (G \sim G')$

Lorsque les moindres généralisés sont associés à la relation de θ-subsomption, on parle de la généralisation la moins générale (en anglais *least general generalization* ou *lgg*) (Plotkin, 1970). Les détails sur la construction de la lgg de 2 clauses sont données dans (Lavrac and Dzeroski, 1994). Nous nous contenterons ici de donner un exemple de *lgg* (Figure 29).

Soit C1 et C2 deux clauses :

C1 = is_deleterious(o15305.p.Glu139Lys) :- modification_size(o15305.p.Glu139Lys, increase), conservation(o15305.p.Glu139Lys, rank1)

C2 = is_deleterious(q9NRR6.p.Arg512Trp) :- modification_size(q9NRR6.p.Arg512Trp, unchanged), conservation(q9NRR6.p.Arg512Trp, rank1)

La *lgg* de C1 et C2 est alors :

lgg(C1, C2) = is_deleterious(X) :- conservation(X, rank1)

Figure 29. Exemple d'une généralisation la moins générale.

Cette approche est utilisée dans un logiciel pionnier de la PLI, GOLEM (Muggleton and Feng, 1990).

4.6 Aleph : un système de PLI multiforme

Aleph (*A Learning Engine for Proposing Hypotheses*) (Srinivasan, 2004) est un successeur de Progol. Il a été amélioré de manière à pouvoir s'identifier à un grand nombre de systèmes existants ayant des stratégies et des algorithmes de recherche variés tels que FOIL, TILDE ou GOLEM. Dans Aleph, les programmeurs peuvent définir un très grand nombre de paramètres permettant de modifier le choix de la meilleure clause, la construction de la clause la plus spécifique de l'espace de recherche (appelé la *bottom clause*), les stratégies de recherche, les fonctions d'évaluation, etc. On peut dire qu'Aleph est hautement paramétrable. C'est pourquoi nous avons utilisé Aleph dans nos travaux.

L'algorithme de base d'Aleph peut être défini en 4 étapes comme dans la Figure 30.

Entrées : un ensemble d'exemples $E = E^+ + E^-$ et une base de connaissances B.

Répéter tant que $E^+ \neq \varnothing$

1. Choisir aléatoirement e^+ dans E^+;
2. Etape de saturation : construire la clause $\perp$ la plus spécifique de l'espace de l'hypothèse telle que $T \wedge e^+ \models \perp$ ($\perp$ est appelé la *bottom clause*). Cette clause est généralement une clause définie contenant le plus grand nombre de littéraux.
3. Etape de réduction : parcourir l'espace des hypothèses pour trouver une hypothèse correcte plus générale que $\perp$.
4. La clause ayant obtenu le meilleur score est ajoutée à la théorie courante et tous les exemples déjà couverts par cette théorie sont retirés de l'ensemble des exemples.

Fin répéter

Figure 30. Algorithme de base d'Aleph.

Par défaut, dans Aleph, la recherche bornée par la *bottom clause* $\perp$ s'effectue par couverture séquentielle de la clause la plus générale vers les plus spécifiques. Notons que dans l'étape 3, réduire l'espace de recherche aux sous ensembles de la *bottom clause* ne permet pas de prendre en compte toutes les clauses plus générales. En pratique, Aleph utilise un algorithme *branch-and-bound* qui permet une énumération intelligente des clauses candidates.

4.7 Applications dans le domaine de la biologie

La PLI a déjà prouvé son intérêt dans quelques applications dans le domaine de la biologie (Kelley et al., 2009) car :

- Il peut extraire des connaissances à partir d'un modèle de données complexe constituée de plusieurs tables ou relations. Naturellement, les modèles de données biologiques sont souvent complexes. Par exemple, un gène peut coder une ou plusieurs protéines, chaque protéine a des « vues » différentes comme une vue informationnelle, une vue structurale, une vue évolutive, etc.
- Les connaissances sont générées sous la forme de règles qui peuvent être interprétées facilement par les biologistes ou les cliniciens.
- Les biologistes ou les cliniciens peuvent introduire leurs propres connaissances au moment du processus d'apprentissage.

On peut citer des applications dans le domaine de la biologie.

En 1992, (Muggleton et al., 1992) a utilisé la PLI pour prédire la structure secondaire des protéines. Les auteurs ont utilisés le système GOLEM pour l'apprentissage de règles de prédiction des structures secondaires. Un ensemble de 12 protéines non homologues de structure connue a été donné à l'algorithme avec une connaissance *a priori* décrivant les propriétés physico-chimiques des résidus. GOLEM a appris un petit ensemble de règles qui prédit quels résidus appartiennent à une hélice α en fonction des relations et des propriétés physico-chimiques. Les règles ont ensuite été testées sur 4 protéines indépendantes non homologues donnant un taux de prédiction de 81%. L'expérimentation reste très limitée, mais les résultats montrent que l'approche est potentiellement intéressante.

Des travaux ont été publiés sur l'apprentissage de signatures sur le repliement 3D des protéines (Cootes et al., 2001). Pour 20 repliements issus de la base de données de classifications structurelles de protéines SCOP (Murzin et al., 1995), 59 règles ont été générées automatiquement. Le taux de prédiction correcte est de 74%. Les auteurs ont montré que leurs règles peuvent être exprimées en termes de concepts structurels et/ou fonctionnels, tels que la localisation de sites actifs. Des améliorations ont été apportées dans (Cootes et al., 2003). Les auteurs ont travaillé sur un ensemble de domaines protéiques extraits de la base de données SCOP sur lesquels ils ont utilisé des alignements multiples de structure et les éléments de structure secondaire pour décrire des repliements. Ils ont généré 66 règles pour apprendre 45 repliements communs tels que *Immunoglobulin*, *Prealbumin-like*, *TIM barrel*, etc.

(Nguyen and Ho, 2008) a envisagé le problème de la prédiction des interactions protéine-protéine. Pour cela, ils ont collecté 22 prédicats correspondant aux données génomiques et protéomiques extraites à partir de plusieurs bases de données (UniProt, InterPro, Gene Ontology, et les banques des données d'expression des gènes). Les règles obtenues permettent de prédire les interactions protéine-protéine avec une précision de 82%.

La PLI a également été utilisée pour automatiser un processus scientifique d'expérimentation visant à établir la fonction de gènes (King, 2004) ou pour prédire l'affinité de la liaison protéine-ligand (Amini et al., 2007).

Cependant, on peut noter que la PLI n'est pas utilisée à la hauteur de la qualité des résultats obtenus dans les quelques applications biologiques et ceci, sans doute, car c'est un langage de programmation. L'utilisateur doit programmer pour extraire des règles. De plus, à l'heure actuelle, pour la PLI, il n'y pas de logiciels couplés à une interface graphique tel que Weka (http://weka.wikispaces.com/).

Dans le CHAPITRE 7 de cette thèse, nous avons utilisé la PLI pour fouiller les 8 000 mutations faux-sens structurales. Les règles obtenues permettent de caractériser et prédire les mutations neutres ou délétères avec une très bonne efficacité.

TROISIEME PARTIE : SYSTEMES D'INFORMATION DEDIES A L'ANALYSE GLOBALE PROTEINES-MUTATIONS FAUX-SENS

La troisième partie comprend la présentation des 2 systèmes d'informations que j'ai construits lors de ce travail de thèse. Il s'agit du système SM2PH Central consacré à l'analyse intégrative des protéines humaines (Chapitre 5) et du système MSV3d concernant la caractérisation des mutations faux-sens (Chapitre 6). L'article décrivant MSV3d, publié dans le journal Database (Oxford), est joint au Chapitre 6.

CHAPITRE 5. SM2PH CENTRAL : SYSTEME D'INFORMATION POUR PERCER LE SECRET DES PROTEINES HUMAINES

« As a general rule the most successful man in life is the man who has the best information. »

Benjamin Disraeli (1804 - 1881)

SM2PH Central a été conçu pour « percer le secret des protéines humaines » dans le cadre de l'étude des maladies génétiques humaines. SM2PH Central est à fois un centre de références pour décrire la relation phénotype-génotype et un générateur automatique de bases de données dédiées (instance) permettant de répondre aux problèmes spécifiques d'un projet biologique particulier. SM2PH Central vise à automatiser la procédure d'intégration de données multi-sources et à faciliter l'enrichissement des contenues scientifiques en créant des interconnections entre des données hétérogènes dans un schéma unique.

Au cours de ce chapitre, nous allons détailler les atouts de SM2PH Central en présentant son architecture globale et les éléments qui ont présidé à sa conception, sa base de données, le processus de mise à jour et d'annotation des données ainsi que les différentes interfaces d'interrogation ou d'analyse et ses services web. Dans la dernière partie de ce chapitre, les SM2PH-Instances consacrées à l'étude de maladies ou de gènes spécifiques seront rapidement décrites.

5.1 Conception de SM2PH Central

5.1.1 Stratégie architecturale

La conception architecturale du système générique de SM2PH Central s'articule autour des principes suivants :

i. Modèle de données flexible : Il permet de mettre en valeur le contenu scientifique des données hébergées en gérant des relations qualifiées entre enregistrements. Par exemple : les relations entre protéines sont qualifiées par le score d'interaction provenant de la banque d'interaction STRING ou les relations entre une protéine et sa structure 3D sont qualifiées par le score d'alignement entre la séquence de la protéine et celle de son empreinte structurale. Le modèle de données permet également la gestion simultanée de plusieurs sources de données hétérogènes venant du système BIRD ou d'une banque publique. De plus, ce modèle est extensible afin de pouvoir s'adapter aux évolutions des banques de données et aux besoins de l'utilisateur. Pour faciliter l'implémentation, nous avons utilisé un mécanisme configurable du modèle de données en utilisant des fichiers XML prédéfinis.

ii. Architecture orienté service : Pour faciliter la construction, l'exploitation et la maintenance d'un système qui associe des données à de nombreux logiciels

d'annotation ou de traitement (MACSIMS, Modeller, ..), SM2PH Central a été bâti sur une Architecture Orienté Service ou architecture modulable contenant de multiples services (Figure 31). On peut citer comme services notables de SM2PH Central : la construction d'alignements multiples de séquences complètes (ou MACS pour *Multiple Alignment of Complete Sequences*), l'annotation des MACS par MACSIMS, la sélection d'empreinte structurale pour la construction d'un modèle 3D, la construction des modèles 3D, la collecte et l'intégration de données transcriptomiques, la construction de graphes d'interactions protéine-protéine ainsi que le générateur des SM2PH-Instances.

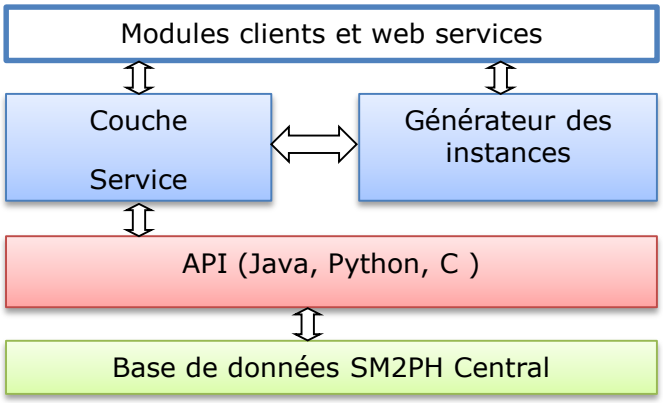


Figure 31. Architecture Orientée Service de SM2PH Central.

iii. Système configurable : Il permet de configurer les composantes du système. Par exemple, on peut aisément supprimer le module des réseaux d'interactions protéine-protéine, ou on peut ajouter le module permettant d'étudier le contexte génomique d'un gène.

iv. Echange de données par le langage de requêtes BIRD-QL : Pour éviter les duplications ou les transferts des gros fichiers des banques de données publiques sur le réseau informatique, SM2PH utilise les requêtes BIRD-QL pour récupérer uniquement les données nécessaires à l'annotation ou au traitement désiré. Ce langage manipule aisément des requêtes incluant des contraintes complexes sur le modèle de données généré sans qu'il soit nécessaire de connaître en détail la structure cachée du système relationnel ou de maîtriser la logique des requêtes SQL. BIRD-QL permet l'édition de requêtes à partir d'une interface Web et échange de données facilement entre les SM2PH Instances et les outils associés (MSV3d, KD4v, Gene Priorisation).

v. Standardisation et interopérabilité : Le but de la standardisation est de faciliter la diffusion et l'échange d'informations entre les sources, c'est-à-dire de les rendre interopérables. Le service web est une composante d'application, accessible sur le web via une interface standard, qui peut interagir dynamiquement avec d'autres applications en utilisant des protocoles de communication basés sur le format XML du fichier structuré et cela indépendamment du système d'exploitation et des langages de programmation utilisés. Ces services permettent à des clients distants (humains ou logiciels) d'accéder à SM2PH Central et à ses instances via Internet.

5.1.2 Stratégies fonctionnelles et intégratives

Les stratégies qui ont guidé l'élaboration et le développement de SM2PH Central concernent :

- L'ajout de données pertinentes provenant des technologies à haut-débit (génomes, transcriptomes, protéomes, métabolomes…),
- l'extension de notre système à toutes les protéines humaines,
- L'offre d'outils compréhensifs d'analyse décisionnelle permettant une meilleure interprétation des résultats en termes de relations entre les différents jeux de données,
- la mise à disposition de moyens pour exploiter de façon automatique les données intégrées afin d'extraire de nouvelles connaissances, par exemple un outil pour la priorisation des gènes impliqués dans une maladie,
- la création d'une interface d'utilisation la plus simple possible,
- le maintien d'une interopérabilité avec les autres systèmes bioinformatiques.

5.1.3 Conception « *use case* »

SM2PH Central a été développé pour atteindre les objectifs suivants :

- Un centre de ressources qui permette d'étudier la relation phénotype-génotype dans les maladies génétiques humaines.
- La mise en valeur du contenu scientifique des données biomédicales : Les données intégrées ou annotées par les banques publiques sont généralement disponibles sur des sites web consultables à travers un navigateur par l'utilisateur lui-même. Elles peuvent aussi, parfois, être interrogées par des programmes par requêtes http ou par service web. Certaines banques peuvent être téléchargées partiellement ou en totalité en tant que simples fichiers texte et même en image des bases de données relationnelles que l'on peut ainsi installer localement. Ces banques publiques concernent chacune une thématique donnée (génome, protéome, mutations, maladies, etc.) et, même si elles sont quelques fois interconnectées, elles ne répondent pas pleinement à notre besoin qui est de fournir à l'utilisateur le maximum de données, de relations, d'analyses croisées concernant « tout » de la séquence au phénotype. SM2PH Central collecte, rassemble et relie cet ensemble de données en une base unique, dédiée, qui doit permettre d'interroger, d'analyser et surtout fournir des outils pour créer de nouvelles connaissances et aider à comprendre les liens entre mutation et maladie.
- L'adaptation aux projets à court terme : classiquement, le développement d'une base de données contenant des fonctionnalités de haut niveau nécessite de mobiliser de nombreuses ressources humaines et techniques sur de longues périodes. Ceci ne convient pas aux projets scientifiques actuels souvent spécialisés et à court terme. SM2PH Central apporte une solution globale en permettant : d'une part, l'utilisation de la ressource intégrée SM2PH Central et d'autre part, la possibilité de développer directement, à partir de SM2PH Central, une SM2PH-Instance comme un nouveau système adapté à des besoins spécifiques.

  L'intégration du processus itératif de création d'une instance dans un pipeline de traitement de données ou simulation (Figure 32) : ces instances générées peuvent servir de bases réduites ou spécifiques pour être intégrées dans des chaînes de calculs ou de simulation, par exemple de priorisation des gènes. De plus, les processus de traitement des données devant s'appliquer à un jeu plus spécialisé de gènes peuvent s'automatiser plus facilement afin de définir des boucles itératives remplissant les conditions et des critères de choix prédéfinis permettant par exemple, l'élimination de

redondance, la validation, la mise à jour automatique et/ou le suivi d'un domaine scientifique particulier…

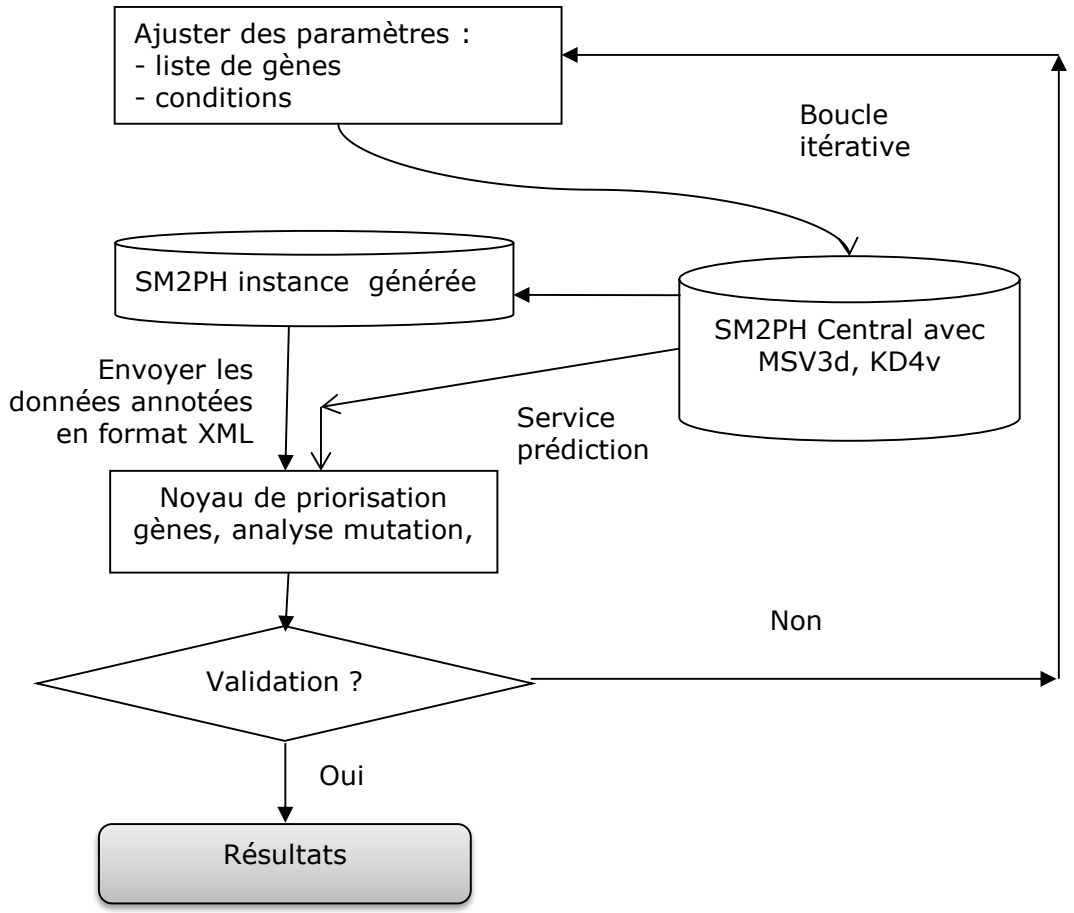


Figure 32. Exemple d'intégration SM2PH-Instance dans une boucle de priorisation de gènes.

5.1.4 Cycle de développement

SM2PH Central est un projet de recherche et de développement. Il fait converger des technologies informatiques au service des applications biologiques. Il s'inscrit dans un cycle de développement itératif (Figure 33), partant de l'analyse des besoins et des problèmes biologiques, il conduit à la définition et à l'implémentation de modules (recherche textuelle, collecte des données, etc.). Ces nouvelles fonctionnalités peuvent être proposées par les biologistes qui effectuent eux-mêmes les tests du système.

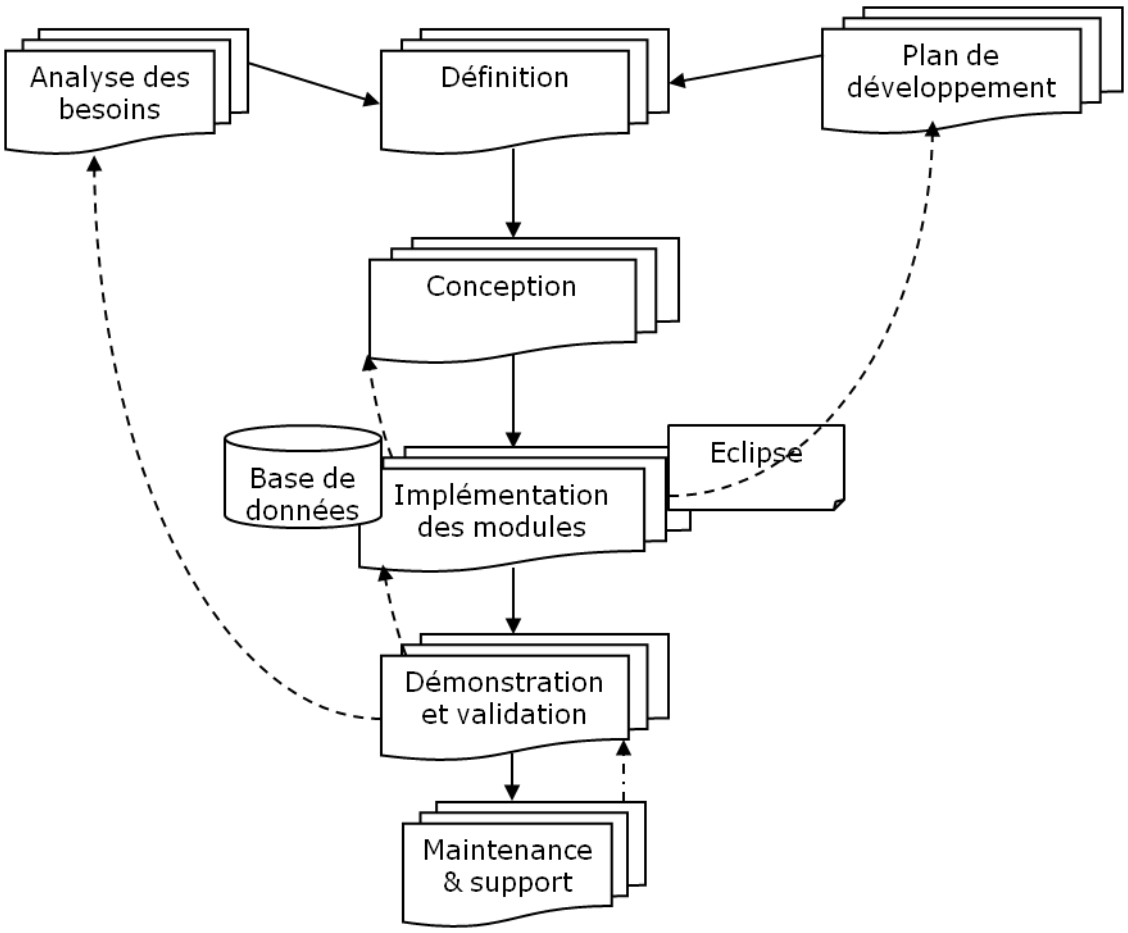


Figure 33. Cycle itératif de développement de SM2PH Central et des outils associés. Eclipse est un environnement de développement intégré spécialement conçu pour le multi-langage de programmation : Java, Python, TCL, etc.

5.2 Implémentation d'architecture

Selon la conception présentée ci-dessus, SM2PH Central est implémenté avec une base de données relationnelle et un serveur web (Figure 34).

La structure de la base de données est principalement centrée sur la table contenant des informations générales relatives à la protéine. Les autres tables regroupent différents niveaux d'informations, relatifs aux modèles 3D construits et à leurs caractéristiques (structures secondaires, coordonnées dans l'espace), aux réseaux des interactions et phénotypes associés, etc.

Le serveur web dispose d'un « explorateur » permettant d'accéder facilement à différentes sources et donc, en utilisant différentes stratégies. Par exemple, l'utilisateur peut (i) chercher les protéines qui se trouvent dans une voie ou réseau biologique, (ii) chercher les protéines impliquées dans une maladie génétique humaine d'intérêt, ou (iii) chercher les gènes humains à partie d'un nom de gène chez la souris.

Les données de SM2PH Central sont aussi accessibles par les modules de recherche. Des requêtes peuvent être formulées soit, à l'aide de mots clés et d'opérateurs de type AND, OR et NOT, soit, par le biais de formulaires permettant de guider l'utilisateur dans sa formulation. De cette manière, l'utilisateur peut affiner sa requête.

La récupération des données SM2PH Central peut aussi s'effectuer par les services web, ce qui permet une interrogation automatisée à partir de n'importe quel langage de programmation

supportant le protocole de transfert http. Un des buts de ces services est de distribuer facilement des profils de caractéristiques de gènes à divers systèmes d'analyse complexe, tel que celui de priorisation de gènes.

De nombreux outils d'analyse puissants ont été implémentés dans le serveur web. L'emploi de ces outils permet aux utilisateurs d'exploiter au mieux les données récupérées. Il est possible par exemple, de visualiser simultanément la séquence protéique, l'alignement multiple, et le modèle 3D en y affichant en parallèle les multiples annotations et informations disponibles, mises en lumière résidu par résidu en fonction du déplacement de la souris sur l'une des trois fenêtres.

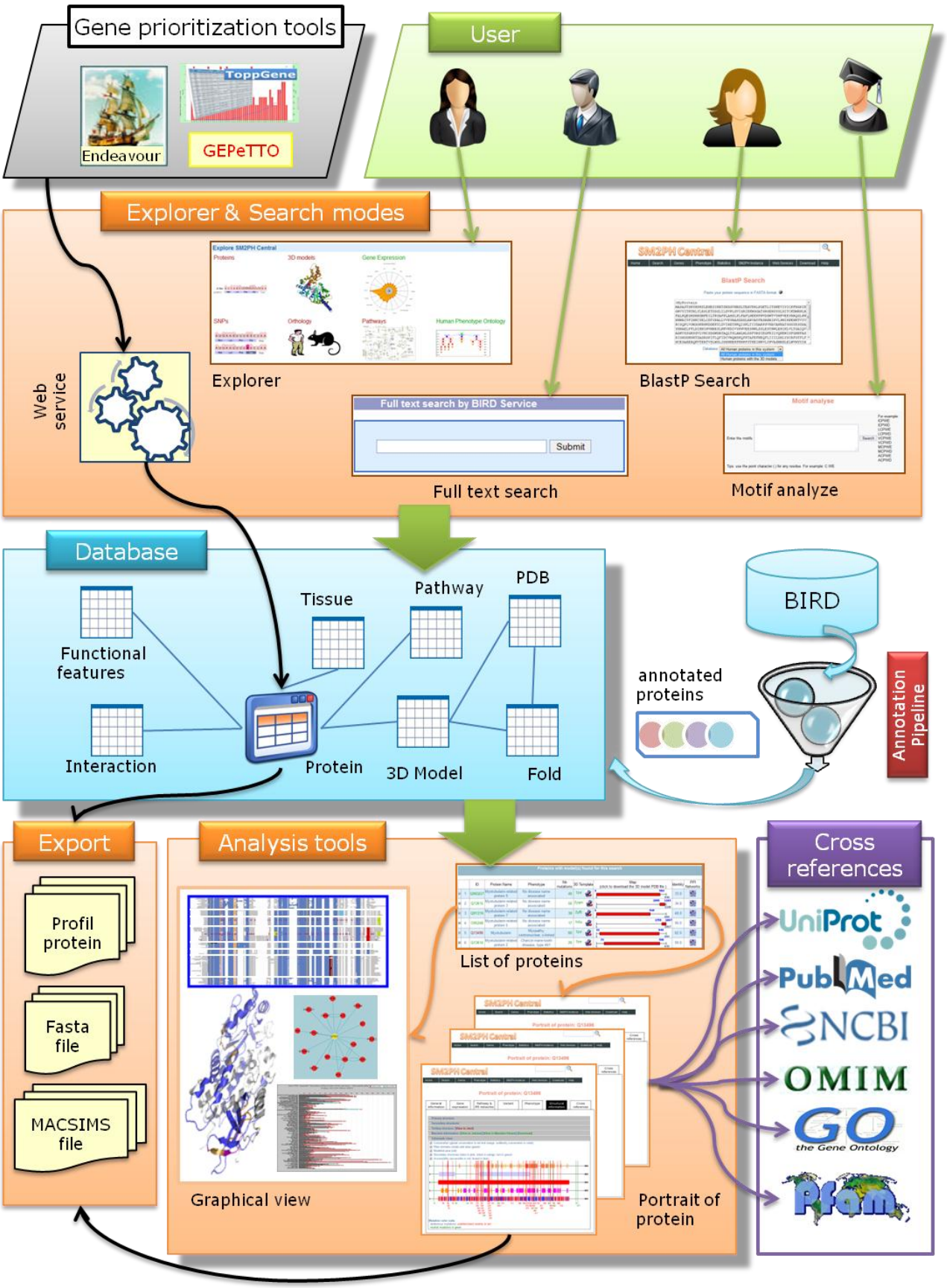


Figure 34. Architecture globale du système SM2PH Central.

5.3 Contenu de la base de données

La base de données SM2PH Central est constituée de 37 tables distinctes reliées entre elles par des clés étrangères (Annexe 2). Elle est modélisée par un schéma en étoile. C'est le modèle de conception le plus utilisée pour l'implantation des entrepôts de données parce que il est le plus approprié aux requêtes et analyses des utilisateurs. Dans ce modèle, le centre de l'étoile est la table principale qui contient le sujet que l'on veut étudier, tandis que les branches qui rayonnent à partir de cette table correspondent aux éléments descriptifs du sujet, appelés « dimensions ». Dans le cas de la base de données de SM2PH Central, la table protéine contenant des informations générales relatives à la protéine est au centre de l'étoile et les dimensions correspondent aux différentes perspectives de l'analyse d'une protéine : une vue évolutive, une vue structurale, une vue informationnelle, des réseaux biologiques et le profil d'expression génique (Figure 35).

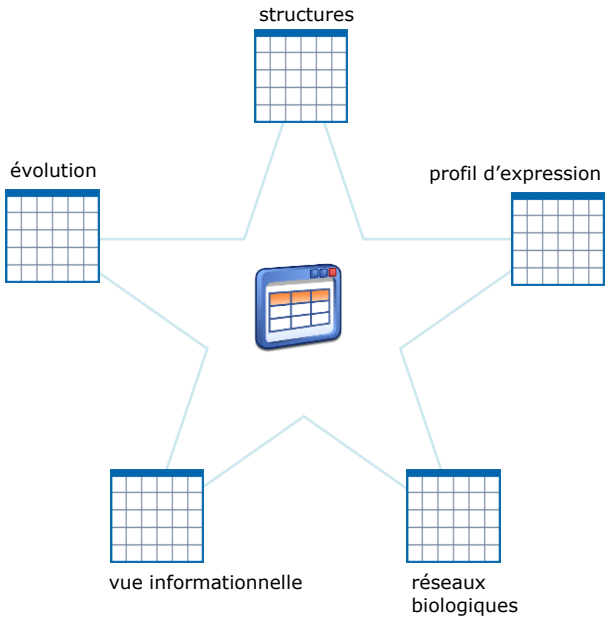


Figure 35. Schéma en étoile de la base de données de SM2PH Central.

- La vue évolutive est liée à 2 alignements multiples de séquences complètes. Nous avons choisi de présenter un alignement regroupant les séquences eucaryotes homologues, ainsi qu'un alignement dédié à l'analyse structurale.
- La vue structurale est définie par l'intermédiaire d'un ou plusieurs modèles 3D construits par homologie. Les protocoles mis en place pour la construction de cette vue seront détaillés dans la section suivante.
- La vue informationnelle est, quant à elle, caractérisée par plusieurs types de données : (i) des données que l'on peut qualifier de « généralistes », comme le nom de la protéine, le nom du gène, etc., complétées par des liens croisés vers des banques de données comme UniProt, OMIM ou encore GO ; (ii) les annotations fonctionnelles et structurales des alignements ; (iii) des données de mutations, associées à des données phénotypiques.
- Les réseaux biologiques intégrés dans SM2PH Central contiennent une collection de cartes représentant des voies métaboliques de la banque de données KEGG ainsi que des interactions entre protéines connues et prédites provenant dans la banque de données STRING.

- Enfin, les profils d'expression proviennent de l'analyse des données transcriptomiques par le programme GxDb provenant de l'expérience *Human Gene Atlas* réalisée par puce à ADN sur 79 tissus humains. Les niveaux d'expression ont été calculés selon 6 méthodes de normalisation (RMA, gcRMA, dChip, MAS5.0, PLIER, VSN). De par la complexité des données transcriptomiques, ces différentes méthodes ne donnent pas systématiquement les mêmes résultats, l'utilisateur ou d'éventuelles analyses automatisées peuvent tirer parti de ces variations qui traduisent des faits biologiques.

5.4 Chargement et mise à jour des données

La banque de données est mise à jour tous les 3 mois de manière automatique. La liste des protéines humaines, qui est censée ne pas trop changer, est mise à jour en premier. Cette étape est effectuée sur le serveur BIRD. Le fichier, qui contient l'ensemble des protéines entrées de la banque au format FASTA, est ensuite fourni à la grille de calcul pour commencer l'annotation.

La section suivante (section 5.5) va décrire dans le détail l'annotation des protéines dans SM2PH Central, mais on peut déjà préciser que le processus complet de mise à jour prend environ 2 semaines pour l'ensemble des étapes : construction et annotation des alignements multiples; recherche d'empreinte et modélisation des structures 3D; localisation à l'intérieur des réseaux biologiques et d'interaction; intégration des données transcriptomiques et finalement, chargement de la base de données complète.

A l'heure actuelle, SM2PH Central contient 20 199 protéines humaines. 10 713 modèles 3D ont pu être construits. Le bilan global de l'annotation intégrative est fourni dans le Tableau 6.

Type	**Nombre d'éléments**
Protéines	20 199
Gènes	19 579
Gènes codent une protéine	19 551
Gènes possédant plusieurs ID UniProt	28
Gènes de souris (dans les relations d'orthologie homme-souris)	17 043
Modèles 3D	10 713
Protéines avec un ou plusieurs modèles 3D	9 347
SCOP	7 810
repliements	1 393
superfamilles	2 223
familles	4 194
Gene Ontology	9 805
Processus biologique	5 968
Composant cellulaire	938

Fonction moléculaire	2 899
Domaines Pfam	12 345
SITE	5 740
KEGG Pathway	226 (réseaux)
Protéines localisées dans les réseaux	5 724
Sous-graphes STRING (un sous-graphe par protéine)	14 179
Gènes ayant un profil d'expression	13 087
Gènes avec des annotations OMIM	16 023
HPO (*Human Phenotype Ontology*)	1 612
Liens vers des LSDB	4 116

Tableau 6. SM2PH Central en quelques chiffres.

5.5 Annotation intégrative automatique de chaque protéine

Nous avons développé, dans SM2PH Central, un pipeline d'annotation automatique (Figure 36) pour le calcul des données relatives à chaque protéine humaine. Ce pipeline se compose de 2 niveaux d'annotation. Le premier niveau permet de réaliser un certain nombre d'annotations structurales et fonctionnelles. Le second niveau vise à localiser les protéines à l'intérieur de réseaux, tels que : i) ceux définies dans la banque KEGG PATHWAY incluant des voies métaboliques, des grands processus biologiques, des maladies, des complexes de signalisation ou ii) ceux disponibles dans la banque STRING et regroupant, pour l'essentiel, les réseaux d'interaction protéine-protéine. Ce niveau localise également chaque protéine humaine au regard de l'expression différentielle de son gène dans différents tissus.

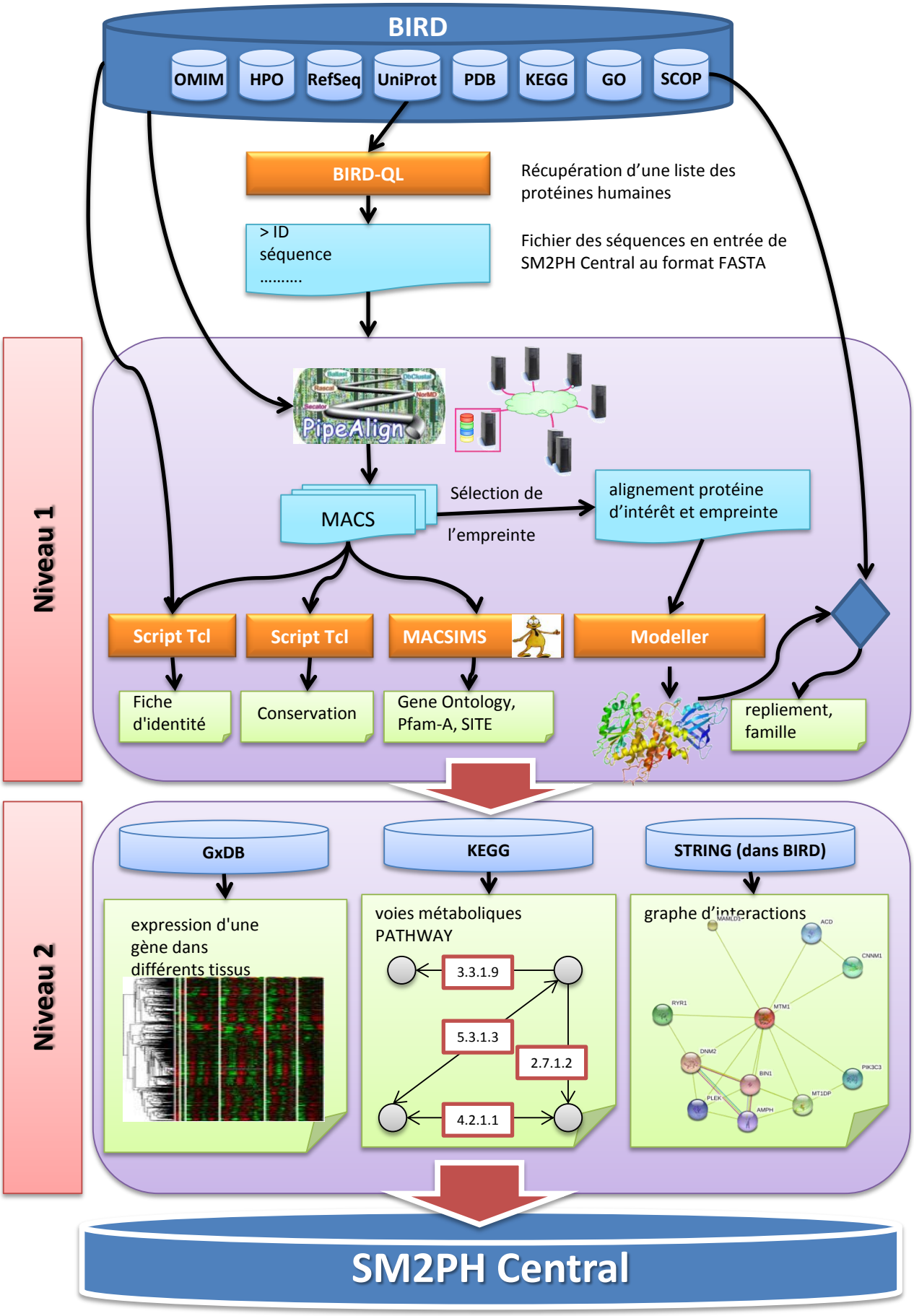


Figure 36. Schéma général du pipeline d'annotation intégrative de séquences protéiques.

5.5.1 Premier niveau d'annotation

A partir d'une séquence protéique, le premier niveau d'annotation permet de construire, de manière couplée, des alignements multiples de séquences complètes annotés avec des données fonctionnelles et structurales et un modèle 3D de la protéine d'intérêt. Au cours de cette annotation, on attribue à la protéine :

- un alignement multiple annoté des protéines eucaryotes homologues,
- un alignement multiple annoté dédié à l'analyse structurale,
- un ou plusieurs modèles 3D si plusieurs empreintes 3D existent,
- une fiche d'identité de la protéine décrite par des annotations issues d'UniProt, GO, OMIM, HPO ainsi que des liens croisés vers d'autres banques. Ces annotations sont obtenues à partir de BIRD.

5.5.1.1 Construction et annotation des alignements multiples

La construction automatique de l'alignement multiple est réalisée à partir de la séquence soumise à une version ajustée de PipeAlign (voir section 3.9 pour sa version originale). Les 2 MACS sont calculés et annotés de manière indépendante. Les différences entre les 2 alignements résident, d'une part, dans le choix des banques utilisées pour effectuer les recherches de similarité Blast et, d'autre part, dans les méthodes de sélection des séquences à aligner.

Dans le cadre de l'alignement « structural », la recherche de similarité se fait dans les banques PDB et UniRef90 (regroupement des séquences avec plus de 90% d'identité) et la méthode d'échantillonnage des bandelettes (Friedrich et al., 2007) est appliquée pour la sélection des séquences à aligner. Le nombre maximal de séquences à aligner est fixé à 100.

Dans le cadre du second alignement, la recherche de similarité se fait simultanément dans les banques UniProt et PDB. Le filtre Eucaryote est appliqué sur les séquences d'UniProt pour ne conserver que séquences issues d'eucaryotes. Le nombre total de séquences sélectionnées est limité à 200.

Les 2 alignements retenus sont ensuite annotés par l'intermédiaire de MACSIMS.

5.5.1.2 Sélection de l'empreinte et création de l'alignement protéine d'intérêt / empreinte structurale

Si l'administrateur n'a pas précisé dans le fichier de configuration quelle structure doit servir d'empreinte pour la construction du modèle 3D par homologie, celle-ci sera déterminée automatiquement. Pour cela, des empreintes sont sélectionnées à partir de l'alignement « structural ». La méthodologie développée permet la sélection d'empreintes partageant au moins 30% d'identité avec la séquence d'intérêt, sachant que le pourcentage d'identité est calculé par rapport à la séquence la plus courte, sur la base d'un alignement global par paires.

Si une ou plusieurs empreintes ont été identifiées, il faut générer un alignement entre la séquence d'intérêt et l'empreinte. La construction d'un modèle 3D par homologie est basée sur cet alignement. Il est primordial que l'alignement utilisé soit de très bonne qualité pour que le modèle généré soit pertinent.

L'alignement de ces 2 séquences est extrait de l'alignement multiple considéré en sortie du PipeAlign. Il sera potentiellement plus fiable que s'il avait été calculé de manière indépendante. En effet, la construction d'un alignement multiple permet de prendre en compte plus d'informations relatives aux membres de la famille protéique et donc d'améliorer l'ajustement de l'alignement, particulièrement au niveau des domaines de moindre conservation.

5.5.1.3 Construction du modèle 3D

La protéine d'intérêt est modélisée en utilisant le programme Modeller, qui accepte en entrée l'alignement de la protéine d'intérêt avec son empreinte structurale pour générer les modèles 3D. Parmi les modèles ayant un RMSD (*Root Mean Square Deviation*) inférieur ou égal à 2Å, le modèle présentant l'énergie minimale est finalement retenu (voir section 3.11.1). Les structures secondaires déterminées à partir du modèle par le programme DSSP (Kabsch and Sander, 1983) sont ensuite ajoutées au fichier PDB du modèle.

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

5.5.1.4 Identification des familles protéiques par structure 3D

A partir des empreintes PDB utilisées lors de la construction du modèle 3D, nous identifions les familles structurales en nous appuyant sur la banque SCOP qui classifie l'ensemble des structures 3D connues (voir section 3.2.4). Dans notre modélisation de structure 3D, une protéine peut avoir plusieurs empreintes. Nous avons développé un script Python permettant de localiser les différentes parties modélisées et d'effectuer la correspondance avec les familles, superfamilles et des types de repliements (*folds*) de SCOP (Figure 37).

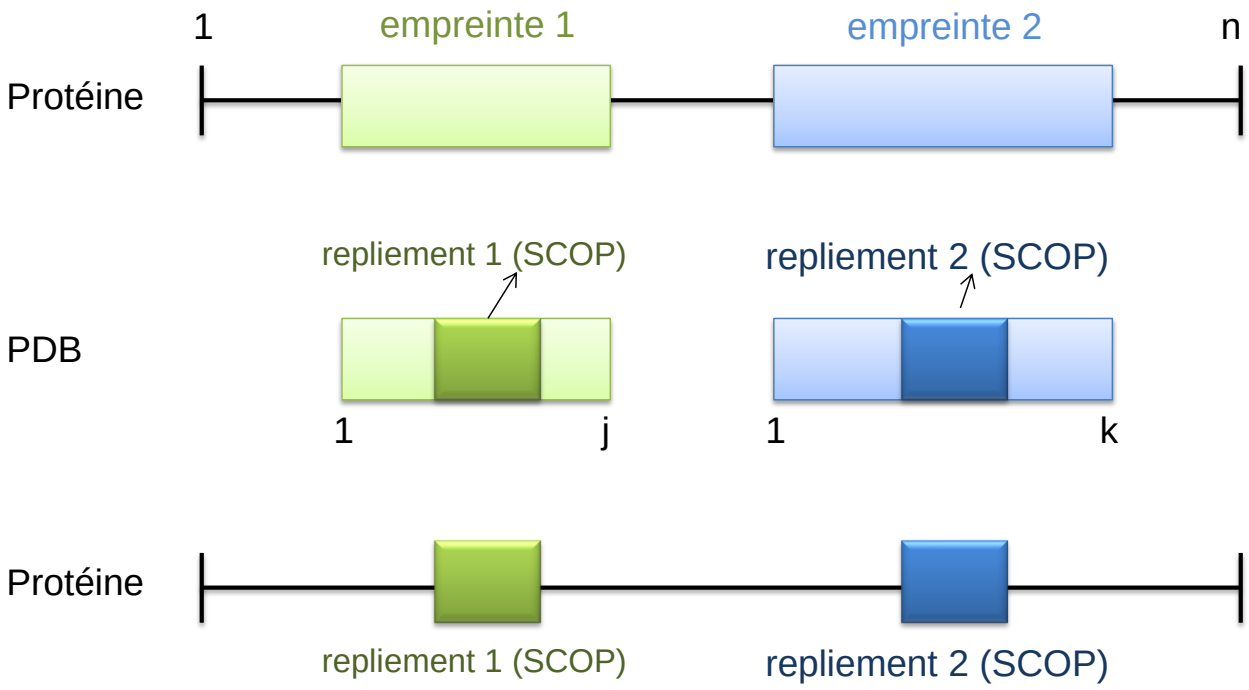


Figure 37. Schéma de la localisation des régions d'une protéine avec les repliements de SCOP.

5.5.1.5 Fiche d'identité des protéines

Les informations concernant la protéine sont regroupées dans un fichier au format XML qui sert au chargement de la base de données. Les informations généralistes concernant les

identifiants, les noms de la protéine et du gène, les synonymes, les GO, les maladies associées, sont stockées dans ce fichier.

5.5.2 Second niveau d'annotation

Au second niveau d'annotation, nous avons développé des modules permettant de replacer les protéines annotées à l'intérieur de réseaux de relations complexes.

5.5.2.1 Construction du graphe d'interactions fiables

Les interactions protéine-protéine sont extraites de la banque STRING, et seules les informations nécessaires sont conservées dans SM2PH Central (Figure 38).

Les 20 199 protéines du réseau humain sont extraites de la banque STRING (version 9.0) et vont représenter les nœuds du graphe. Les interactions dont le score est supérieur ou égal à 0.7 (score de grande fiabilité) sont ensuite sélectionnées. 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 décrit dans SM2PH Central comporte 14 179 nœuds.

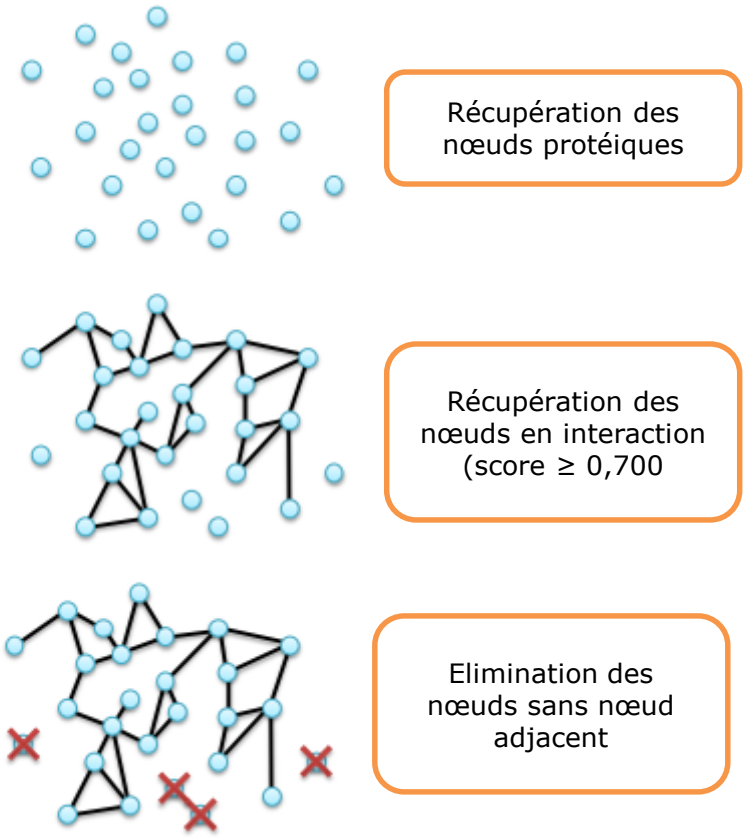


Figure 38. Construction du graphe d'interactions fiables.

5.5.2.2 Intégration des données d'expression des gènes

Les expressions tissulaires des gènes présents dans SM2PH Central ont été intégrées en se basant sur la banque de donnée transcriptomique : GxDB (voir section 3.3). Schématiquement, l'intégration s'effectue de la manière suivante (Figure 39). Pendant la mise

à jour des données, SM2PH Central exporte un fichier texte dont chaque ligne est une identifiant UniProt. Cette liste de protéines est transférée au serveur GxDB. Après vérification du format de la liste soumise, le serveur GxDB lance le service ExpressTiss (voir section 3.3). Le service ExpressTiss calcule les données d'expression et les renvoie sur le serveur SM2PH Central qui après analyse lexicale les importe directement dans sa base de données.

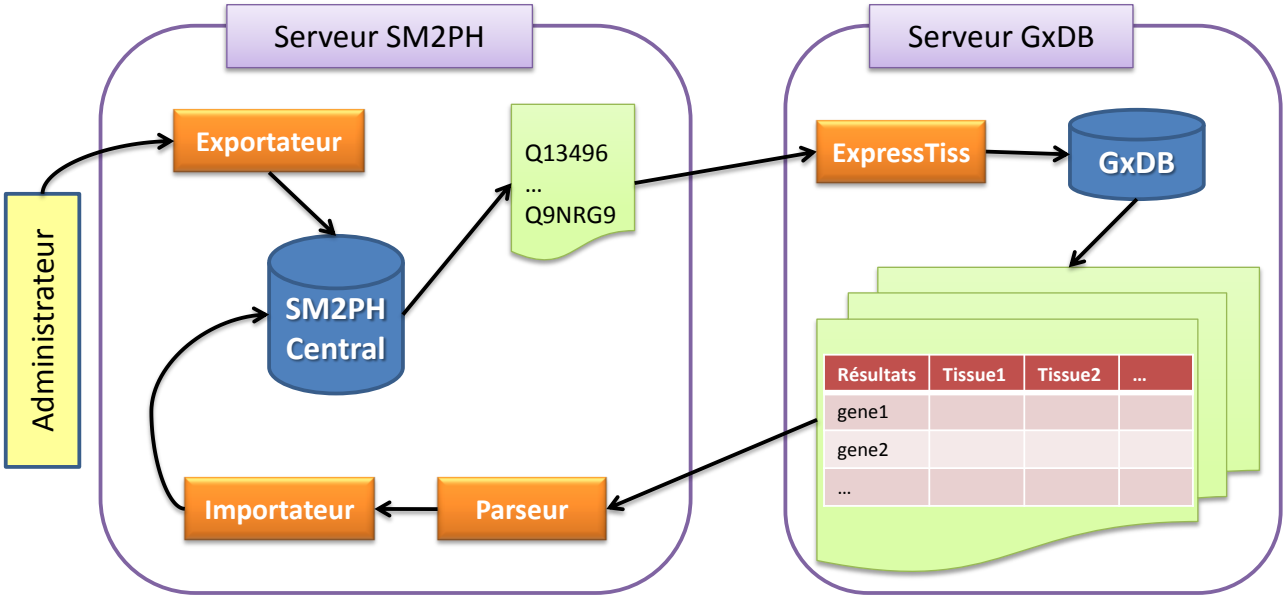


Figure 39. Processus de l'intégration d'expression des gènes.

5.6 Description de l'interface de SM2PH Central

Le site web de SM2PH Central est accessible à l'adresse http://decrypthon.igbmc.fr/sm2ph/ (Figure 40). Ce site a été développé en Python et est hébergé sur un serveur web de l'IGBMC.

5.6.1 SM2PH Explorateur

L'explorateur SM2PH, qui a été implémenté dans la page d'accueil de SM2PH Central, est la véritable pierre angulaire de notre système. Il est conçu afin de faciliter l'utilisation pour les différents types d'utilisateurs non-experts : biologistes, pharmaciens, médecins, cliniciens…. L'utilisateur dispose de toute une variété d'options pour accéder aux données de SM2PH en fournissant : un identifiant UniProt, le nom d'un gène, le nom d'un pathway, le code PDB ou une mutation faux-sens. Plusieurs points d'accès avec SM2PH Explorateur contribuent à exploiter des données le plus rapidement possible (Figure 40). En un clic, l'utilisateur peut récupérer les données. Par exemple, l'utilisateur peut cliquer sur le point d'accès Pathways, une liste de tous les pathways intégrés dans SM2PH Central est alors récupérée. Pour chaque entrée, on dispose de son identifiant avec un lien vers la base de données KEGG, du nom de la voie, du nombre et de la liste des protéines de SM2PH localisées dans cette voie. Chaque protéine a un lien vers sa page. Ces informations sont chargées très rapidement parce qu'elles sont distribuées sur des pages différentes créées à la volée. L'utilisateur peut naviguer entres ces pages grâce à un outil de navigation implémenté au dessus de chaque page ou utiliser le filtre pour chercher la voie d'intérêt.

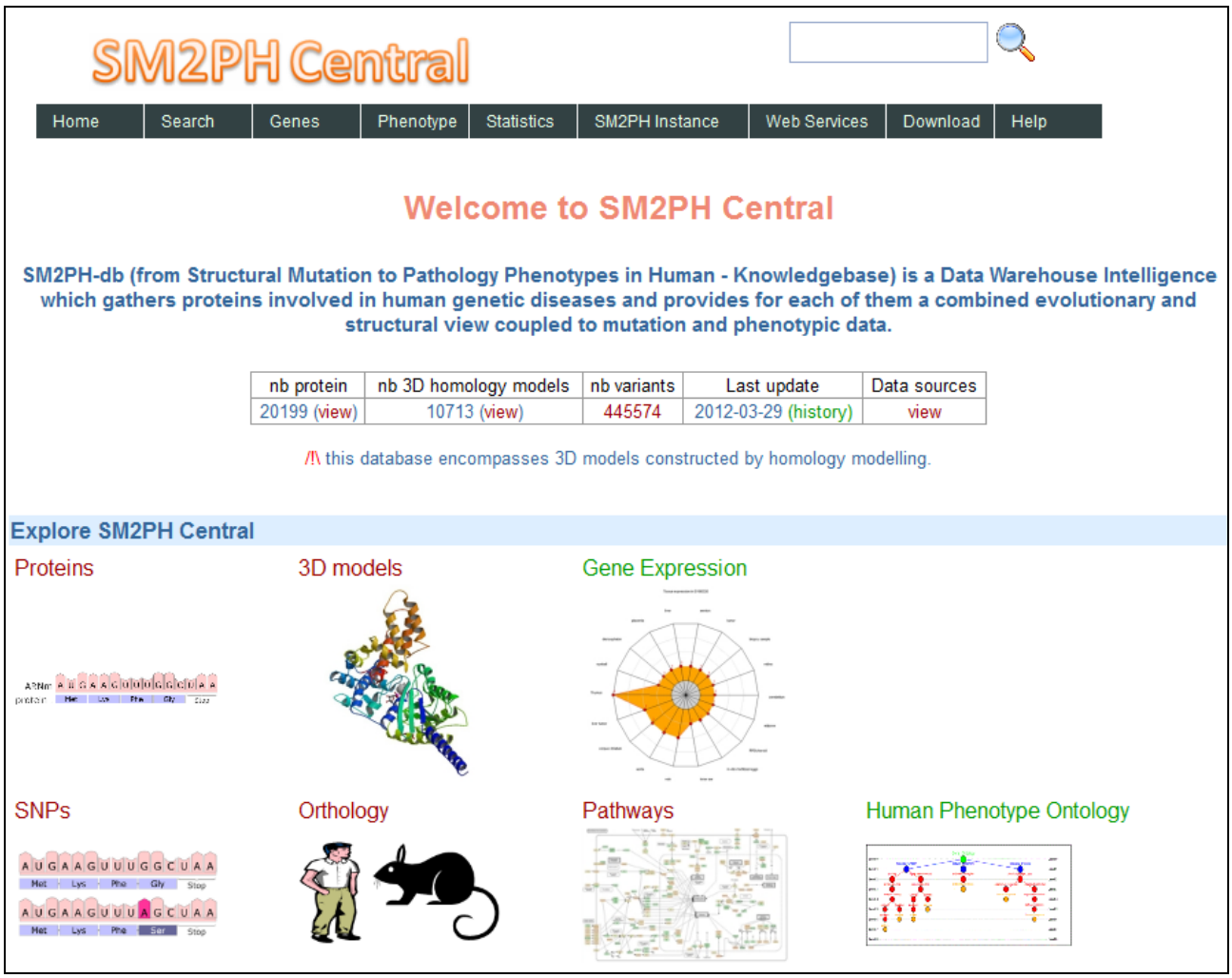


Figure 40. Capture d'écran de la page d'accueil de SM2PH Central avec SM2PH Explorateur.

5.6.2 Modules de recherche

Pour faciliter l'accès et le questionnement par des utilisateurs non-expert, SM2PH Central fournit pas moins de 7 modes d'interrogation distincts accessibles par l'intermédiaire du menu *Search* (Figure 41).

Figure 41. Modules de recherche sur le site SM2PH Central.

SM2PH Central propose une recherche de chaînes de caractères très rapide dans les données via le module *Full Text Search*. Ce module est lié au moteur de recherche et d'indexation de BIRD Server qui utilise les fonctions de recherche de texte puissant et rapide du système IBM DB2. Quand SM2PH envoie un mot ou un ensemble de mots à BIRD celui-ci lui retourne une liste d'identifieurs (OIDs ou *Object Identifier*). Cela permet de gagner du temps en ne faisant pas transiter des données inutiles sur le réseau ou en mémoire. Ce module est également intéressant par son efficacité et sa simplicité d'utilisation. De plus, l'utilisateur peut accéder à ce service depuis toutes les pages de SM2PH Central grâce à la fenêtre *textbox* présente en haut à droite de tous les écrans SM2PH (Figure 41). L'intégration de la recherche en texte entier est une des grandes améliorations de SM2PH Central.

Le module *Simple Search* est destiné à la réalisation d'une recherche simple. Il comporte 2 champs et un nombre limité de mots clés (numéro d'accession de la protéine, nom de la protéine, nom d'une maladie, etc.). Une aide contextuelle très utile a été mise en place pour ce formulaire : au moment de la frappe des premières lettres, les premiers résultats sont dynamiquement affichés afin d'aider l'utilisateur si sa recherche est peu précise.

Pour mieux cibler vos recherches et obtenir une liste de résultats à la fois plus courte et plus pertinente, il suffit d'utiliser le module *Advanced Search* qui permet de faire une recherche avancée, en combinant jusqu'à 4 mots clés différents avec les opérateurs logiques ET, OU et NON. En plus des mots clés disponibles dans la recherche simple, cette recherche peut s'effectuer sur la proportion de la protéine modélisée, sur le nombre d'hélices du modèle généré, etc.

Le module *3D Accessible Pattern Search* permet de rechercher les protéines présentant un motif de 2 à 10 résidus situés dans un environnement 3D restreint dont le diamètre est ajustable par l'utilisateur.

L'interrogation de la banque, par l'intermédiaire d'une recherche de similarité, réalisée par BlastP, est possible grâce à la génération automatique, lors de chaque mise à jour, d'une banque de séquences au format Blast, à partir des séquences protéiques en entrée de SM2PH Central. Cette fonction se trouve dans le module *Blast Search*.

Le module *Motif Search* permet de rechercher, dans l'ensemble des séquences des protéines humaines, la présence d'un ou plusieurs motifs. Ce module est très utile parce que certains motifs ou signatures donnent parfois un indice sur la fonction des protéines (site catalytique, site de liaison à un ion, etc.). Nous avons intégré les expressions régulières pour permettre à l'utilisateur de manipuler le texte de façon très puissante et très concise. Par exemple le caractère '.' correspond à n'importe quel caractère.

Le critère de la recherche dans le module *SNP Search* est l'identifiant de la mutation faux-sens dans la banque dbSNP. Quand vous avez décrit une mutation dans la requête, SM2PH Central va essayer de trouver la protéine qui contient cette mutation. Ce module permet de passer de SM2PH Central à la banque de mutations MSV3d présentée dans le chapitre suivant.

Dans tous ces modules de recherche, une aide interactive a été intégrée pour faciliter la saisie et les recherches. Elle est accessible par l'intermédiaire d'une icône représentant un point d'interrogation.

5.6.3 Modules de visualisation et d'analyse des données

Toutes les modules d'interrogation et d'extraction de données appellent des scripts CGI qui interrogent la base de données et retournent une page HTML comme résultat. La Figure 42

présente l'interface de résultats obtenus suite à une requête et qui sont présentés sous la forme de 2 tableaux consécutifs où chaque ligne est associée à une protéine. Le premier tableau contient les protéines avec au moins un modèle associé. A l'inverse, le second tableau représente les protéines n'ayant pas de modèle 3D.

Le premier tableau fournit un récapitulatif des données associées aux protéines répondant aux critères de la recherche effectuée : il contient l'identifiant de la protéine, son nom ainsi que le nom d'une maladie associée à cette protéine, le nombre de mutations connues, le code PDB de l'empreinte structurale utilisée pour la génération du modèle, le diagramme situé dans la colonne *Map* représente la partie modélisée par rapport à la longueur totale de la protéine. Le diagramme est une zone interactive, par l'intermédiaire de laquelle il est possible de télécharger le fichier du modèle 3D au format PDB. Le pourcentage d'identité entre la protéine d'intérêt et l'empreinte structurale est précisé dans la colonne *Identity*. Enfin, la dernière colonne *PPI Networks* permet la visualisation des réseaux d'interactions protéine-protéine.

Proteins with model(s) found for this search

	ID	Protein Name	Phenotype	Nb mutations	3D Template	Map (click to download the 3D model PDB file)	Identity	PPI Networks
1	Q96QG7	Myotubularin-related protein 9	No disease name associated	46	1zvr	3 503 1 549	33.0	
2	Q13615	Myotubularin-related protein 3	No disease name associated	56	2yqm	1095 1183 1 1198	34.0	
3	Q9Y216	Myotubularin-related protein 7	No disease name associated	38	2yf0	1 510 1 660	48.0	
4	O95248	Myotubularin-related protein 5	No disease name associated	17	1v5u	1755 1867 1 1867	90.0	
5	Q13496	Myotubularin	Myopathy, centronuclear, x-linked	66	1zvr	31 544 1 603	62.0	
6	Q13614	Myotubularin-related protein 2	Charcot-marie-tooth disease, type 4b1	26	1zvr	73 586 1 643	99.0	
7	Q9NYA4	Myotubularin-related protein 4	No disease name associated	39	1zvr	16 846 1 1195	36.0	
8	Q13613	Myotubularin-related protein 1	No disease name associated	9	1zvr	95 607 1 665	74.0	
9	Q86WG5	Myotubularin-related protein 13	Charcot-marie-tooth disease, type 4b2	69	1v5u	1737 1849 1 1849	64.0	
10	Q9Y217	Myotubularin-related protein 6	No disease name associated	25	2yf0	4 505 1 621	92.0	
11	Q96EF0	Myotubularin-related protein 8	No disease name associated	41	2yf0	33 506 1 704	62.0	

Proteins without models found for this search

	ID	Protein Name	Phenotype	Nb Mutations	MACSIMS Euca	MACSIMS Sample	PPI Networks
1	A4FU01	Myotubularin-related protein 11	No disease name associated	33	Jalview	Jalview	
2	Q9C0I1	Myotubularin-related protein 12	No disease name associated	19	Jalview	Jalview	
3	Q8NCE2	Myotubularin-related protein 14	Myopathy, centronuclear, 1	14	Jalview	Jalview	
4	Q9NXD2	Myotubularin-related protein 10	No disease name associated	13	Jalview	Jalview	
5	Q9Y2M0	Fanconi-associated nuclease 1	No disease name associated	64	Jalview	Jalview	

Figure 42. Capture d'écran du résultat d'une recherche en texte entier du terme « myotubularin ». Les résultats sont séparés entre les protéines associées ou non à un modèle structural.

Dans le second tableau, les colonnes relatives aux données structurales sont remplacées par l'accès aux alignements multiples annotés par MACSIMS.

Plusieurs liens sont accessibles par l'intermédiaire de cette page résultat. Par exemple, cliquer sur le numéro d'accession d'une protéine permet d'accéder à la page d'accueil de cette dernière (Figure 43), qui peut être considérée comme un portrait de la protéine d'intérêt. Le portrait d'une protéine regroupe :

- des informations généralistes comme l'identifiant UniProt, son nom, le nom du gène qui la code, les relations d'orthologie identifiées entre l'homme et la souris. Les liens vers le téléchargement des fichiers générés sont accessibles via l'onglet *General information* (Figure 43A).
- les réseaux de voies métaboliques ou d'interactions protéine-protéine sont accessibles grâce à l'onglet *Pathway & PPI Networks* (Figure 43B). Quand l'utilisateur interroge une protéine d'intérêt, le SM2PH Central va générer à la volée un fichier au format XGMML (*eXtensible Graph Markup and Modeling Language*) et va l'envoyer au client pour afficher le graph. Le format XGMML permet de décrire des graphes pour les visualiser à l'aide de différents logiciels, dont Cytoscape.
- le profil transcriptomique, qui permet de déterminer dans quels tissus un gène d'intérêt est exprimé, est accessible via l'onglet *Gene expression* (Figure 43C).
- des informations relatives aux mutations faux-sens connues et aux maladies associées sont accessibles par l'onglet *Variant* qui donne des informations sur la mutation contenue dans la banque MSV3d et, lorsque les données sont disponibles, des informations sur l'impact structural (voir CHAPITRE 6).
- les maladies humaines et les phénotypes associés à la protéine d'intérêt sont aisément accessibles en utilisant le lien vers la banque OMIM et HPO dans l'onglet *Phenotype* (Figure 43D). Nous avons développé une visualisation de l'ontologie HPO qui permet d'associer les gènes provenant de SM2PH Central aux phénotypes de l'ontologie HPO (Figure 44). Grâce à cette visualisation, en parcourant les phénotypes, on observe immédiatement les gènes SM2PH Central impliqués dans un phénotype donné. Le fichier d'ontologie HPO a une taille de 15,6 Mo, le code JavaScript étant exécuté côté client, le traitement de celui-ci peut être lent selon la puissance de la machine client.
- des détails sur les structures protéiques sont accessibles par l'intermédiaire de l'onglet *Structural information*. On y retrouve des données relatives au modèle structural construit telles que l'identifiant PDB de la structure utilisée, le pourcentage d'identité que partagent la protéine et l'empreinte (Figure 45C). L'alignement utilisé pour la modélisation par homologie peut également être visualisé (accessible par le lien *View in Jmol* ou le lien *View in Jalview*). Les structures secondaires assignées par le logiciel DSSP sont également disponibles (Figure 45B) de manière visuelle, sous la forme d'un diagramme schématisant la succession d'hélices et de feuillets, ainsi que sous la forme d'un tableau qui reprend les positions de début et de fin de chaque élément de structure secondaire. Des annotations relatives aux domaines et sites fonctionnels associés, par MACSIMS, à la séquence de la protéine sont disponibles. MACSIMS récupère des informations relatives aux domaines et aux sites fonctionnels dans plusieurs banques de données (Figure 45A).
- des liens croisés vers d'autres banques publiques, comme GO, UniProt, LSDB et PubMed sont proposés au niveau de l'onglet *Cross references*.

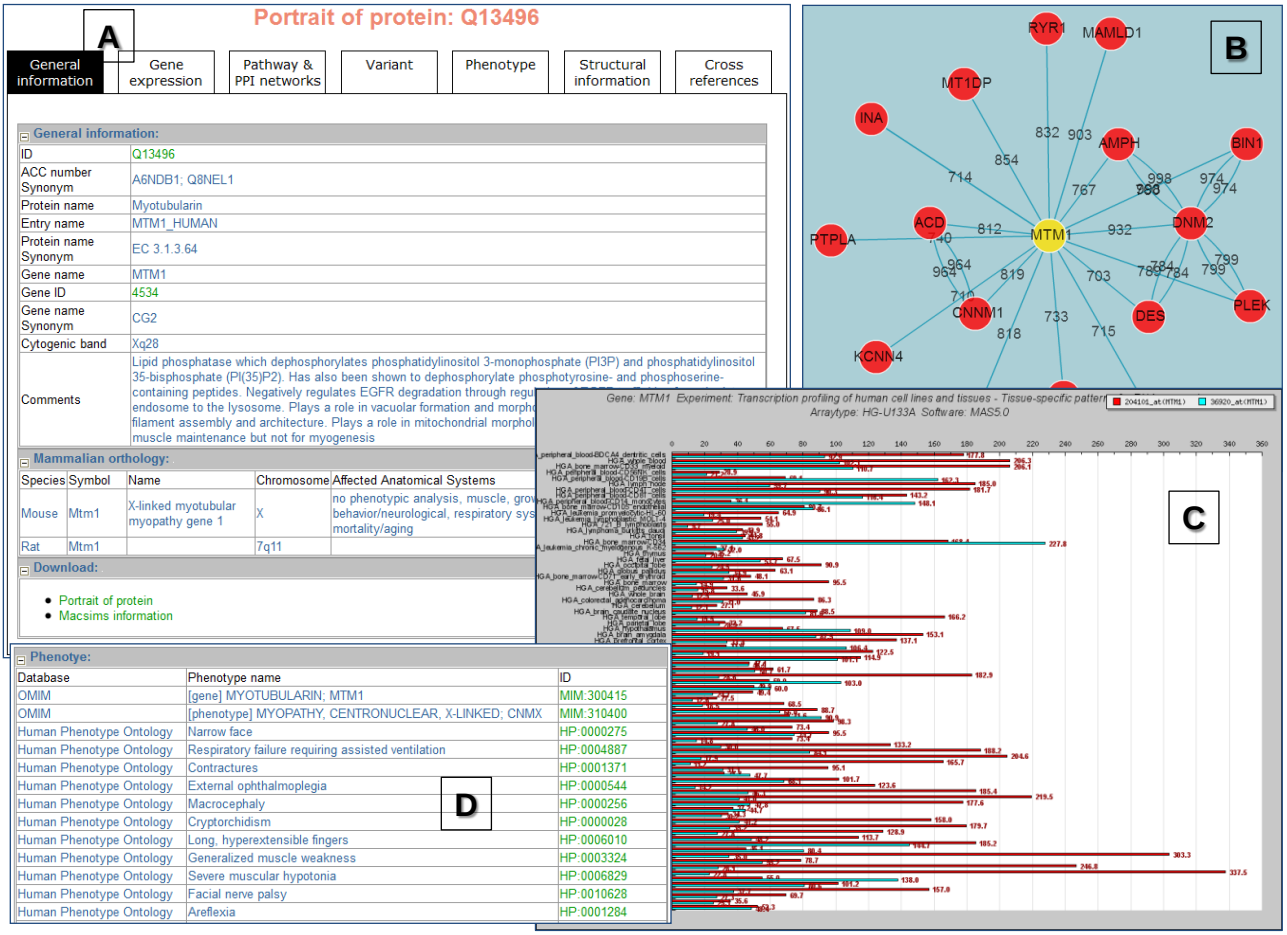


Figure 43. Portrait d'une protéine de SM2PH Central.

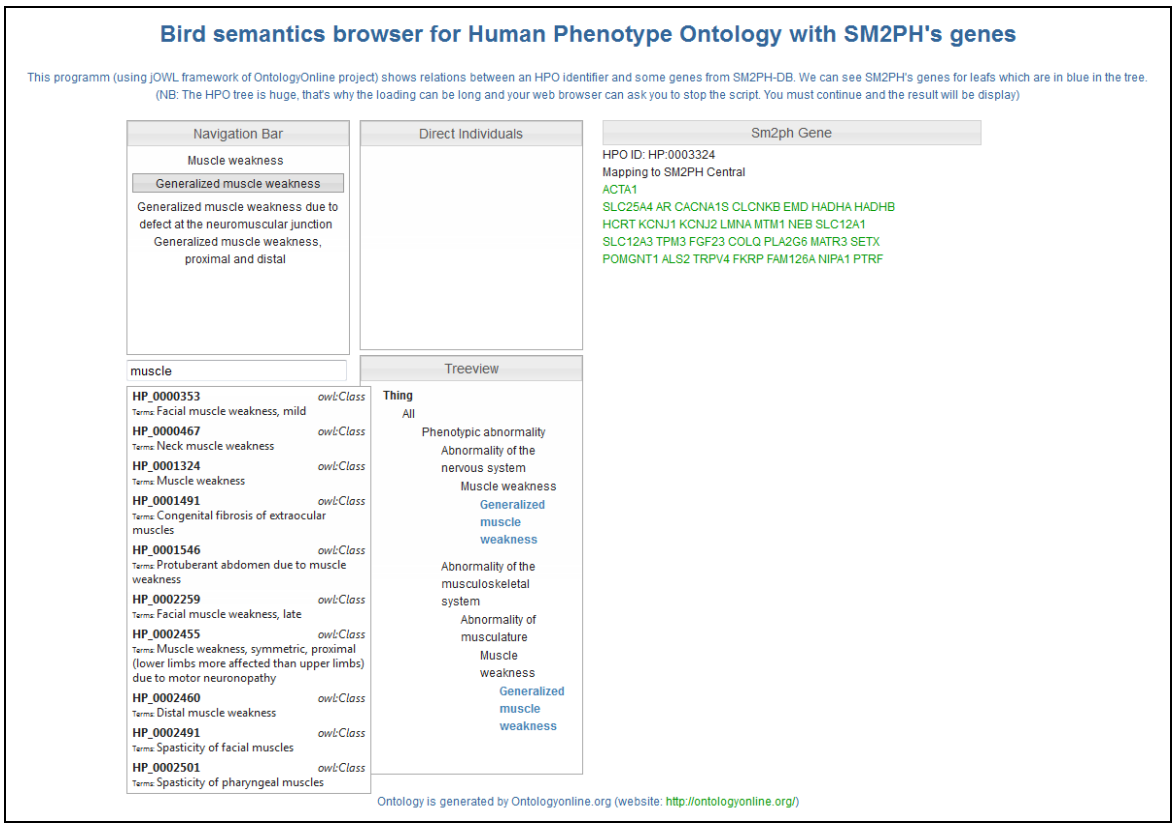


Figure 44. Visualisation de l'ontologie HPO associée aux gènes SM2PH Central.

Figure 45. Données structurales de la myotubularine dans SM2PH Central.

Afin de connecter l'ensemble des données relatives à la séquence, la structure, la fonction et l'évolution d'une protéine, nous avons réutilisé une interface web de visualisation développé par Nicolas Garnier dans SM2PHdb version 1.0 (Garnier, 2008). Elle est basée sur l'applet Java Jmol. L'accès à cette interface Jmol peut se faire par le lien *View in Jmol* dans l'onglet *Structural information*. La génération de cette page est programmée en Python et utilise l'AJAX (*Asynchronous Javascript And XML*) pour la mise à jour dynamique de la page. Cette interface (Figure 46) permet l'interconnexion de la séquence protéique avec la structure 3D dans le contexte évolutif de l'alignement multiple annoté. Les annotations fonctionnelles issues de l'alignement sont visualisées à la fois sur la séquence et dans le contexte structural permettant de mieux appréhender les relations qui les unissent.

L'interface de résultat est découpée en 2 parties interconnectées (Figure 46A).

La partie de droite présente l'alignement multiple. Le score NorMD de l'alignement est affiché au bas de la page. La séquence de la protéine d'intérêt se situe toujours dans la première ligne de l'alignement et la fraction modélisée est soulignée. Les annotations fonctionnelles et structurales générées par MACSIMS sont accessibles par l'intermédiaire d'une liste de sélection *Features*. Lors de la sélection d'une annotation, l'alignement multiple est coloré en fonction de celle-ci ainsi que le modèle 3D. Les séquences des sous-familles de protéines sont colorées de façon différentielle. Les boutons « collapse » et « uncollapse » permettent respectivement d'afficher toutes les séquences des sous-familles ou bien de les cacher.

Le modèle attaché à la protéine d'intérêt est présenté dans le cadre gauche de l'interface au cœur d'une fenêtre *Jmol*. L'alignement multiple annoté et le modèle sont connectés afin de visualiser simultanément les annotations sur la partie modélisée et sur l'alignement, en

respectant toujours les mêmes codes couleurs. La sélection d'un résidu au niveau de la structure mettra en évidence ce résidu au sein de la séquence et vice versa. Un certain nombre d'options de visualisation sont disponibles sous la fenêtre de visualisation du modèle, allant des options de bases (zoom, arrière-plan, couleurs, etc.) accessibles à tout un chacun, à l'utilisation de lignes de commande permettant des manipulations poussées du modèle par des utilisateurs aguerris.

De par la disponibilité des données de mutation associées à des données phénotypiques, l'interface offre la possibilité de localiser la position des mutations faux-sens engendrant l'émergence d'une maladie sur le modèle structural et sur la séquence de la protéine au sein de l'alignement multiple.

La visualisation peut être basculée pour coupler le modèle 3D à l'alignement de la séquence d'intérêt avec la séquence de l'empreinte (Figure 46B). Les résidus identiques entre les 2 séquences sont indiqués par une étoile rouge. La coloration correspond aux structures secondaires, les hélices sont en rose et les feuillets en jaune.

Enfin, l'option *Around residues* permet de visualiser les résidus situés dans un environnement structural proche d'un résidu d'intérêt (Figure 46C). 2 images de la représentation 2D des résidus proches et accessibles dans un rayon de 12Å par défaut. La coloration est fonction de l'hydrophobie ou de la polarité pour la première image et des propriétés physico-chimiques des résidus pour la seconde. Le diamètre des cercles est proportionnel à l'accessibilité relative des résidus. La coloration dans l'applet Jmol est la même que la première image et est choisie par l'utilisateur, tout comme le diamètre de recherche autour du résidu. La recherche de ces résidus est facilitée par la base de données SM2PH Central qui contient la valeur d'accessibilité pour chaque acide aminé. La construction de l'image est effectuée pour réaliser une projection 2D des coordonnées des résidus proches et accessibles de l'acide aminé sélectionné.

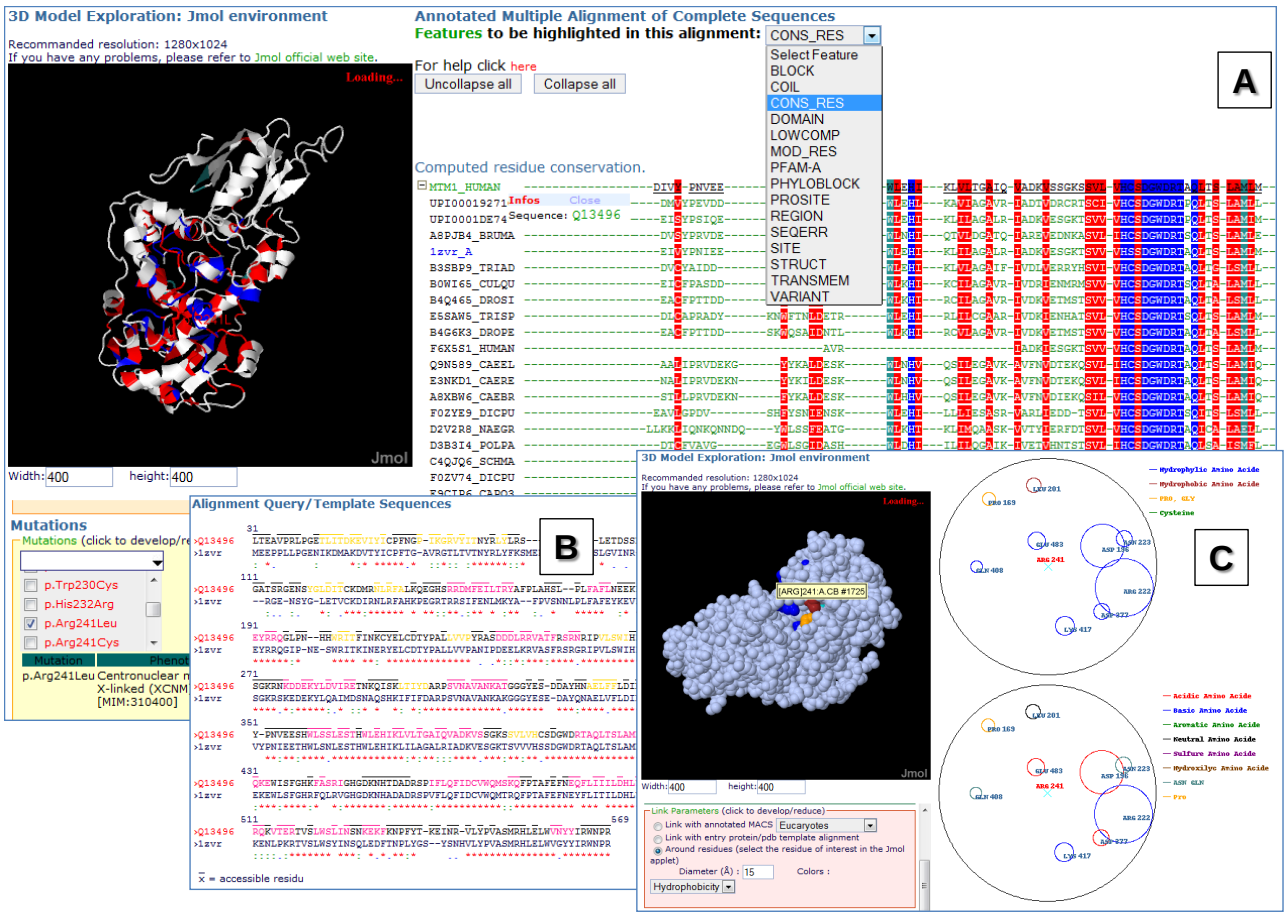


Figure 46. Interface Jmol d'interconnexion des différentes vues afférentes à la protéine. A : interface de visualisation d'un alignement multiple annoté couplé de manière interactive au modèle 3D. **B** : alignement par paire (séquence protéique cible et empreinte structurale) utilisé pour la construction du modèle 3D. Les étoiles rouges symbolisent les résidus identiques entre la séquence d'intérêt et l'empreinte structurale. **C** : visualisation des résidus structuralement proches d'un résidu d'intérêt. La taille des cercles est fonction de l'accessibilité et la couleur de la nature du résidu.

5.7 Web services de SM2PH Central

Pour faciliter l'accès aux ressources de la relation phénotype-génotype dans les maladies génétiques humaines, nous avons implémenté les services web suivantes (Figure 47) :

- *Prediction Service* : permet de prédire l'impact phénotype d'une mutation faux-sens. En fait, ce service est installé sur KD4v (voire CHAPITRE 7).
- *Data Fetch Service* : sur la base de l'identifiant UniProt d'une protéine d'intérêt, ce service renvoie toutes les informations concernant cette protéine dans un fichier XML.
- *Full Text Search Service* : renvoie tous les identifiants UniProt pour un mot clé de recherche (par exemple : dmd, myopathies, retina, etc.).

Figure 47. Page web (http://decrypthon.igbmc.fr/sm2ph/cgi-bin/webservices) qui liste tous les services web implémentés dans SM2PH Central. L'utilisateur peut trouver ici des exemples et le code pour utiliser ces services web.

5.8 SM2PH-Instances

Une des caractéristiques de SM2PH Central réside dans la possibilité de générer aisément, et à façon, des sous-banques ou « instances » dédiées à l'étude d'un sous-groupe particulier de protéines humaines. Ceci est rendu possible d'une part, par un environnement logiciel puissant et configurable et d'autre part, par le fait que toutes les protéines humaines sont présentes dans SM2PH Central. La modularité élevée de l'architecture de SM2PH Central permet d'assurer une mise à jour coordonnée et concomitante de SM2PH Central et des Instances associées qui sont généralement développées par des chercheurs experts d'un domaine et contiennent des informations additionnelles spécialisées publiques ou privées. Le temps de calcul nécessaires pour la construction d'une nouvelle instance est court, environ une heure. En créant une SM2PH-Instance, le biologiste ou médecin bénéficie directement des avantages que propose cet outil :

- L'utilisateur peut réaliser son système en très peu de temps tout en donnant une liste de gènes. Il le fait tout seul, sans avoir forcément besoin de connaître le langage de programmation.

- Avec des données très confidentielles, l'utilisateur peut administrer lui-même son système.

Les instances peuvent être hébergées sur les serveurs de notre laboratoire ou l'administrateur peut télécharger le site web et la base de données pour préserver la confidentialité.

D' ores et déjà, plusieurs instances consacrées à l'étude de maladies (tels que les ciliopathies (http://decrypthon.igbmc.fr/kbmc)) ou des gènes spécifiques (tels que GPR179 (Audo et al., 2012)) ont été développées. Enfin, SM2PH-AMD-kb (Figure 48) a été spécifiquement élaborée en collaboration avec T. Léveillard de l'Institut de la Vision (Paris) pour faciliter l'étude de la Dégénérescence Maculaire Liée à l'Âge (DMLA). Cette instance permet d'accéder notamment à l'ensemble des informations nécessaires à la priorisation de gènes candidats de la DMLA (voir CHAPITRE 8).

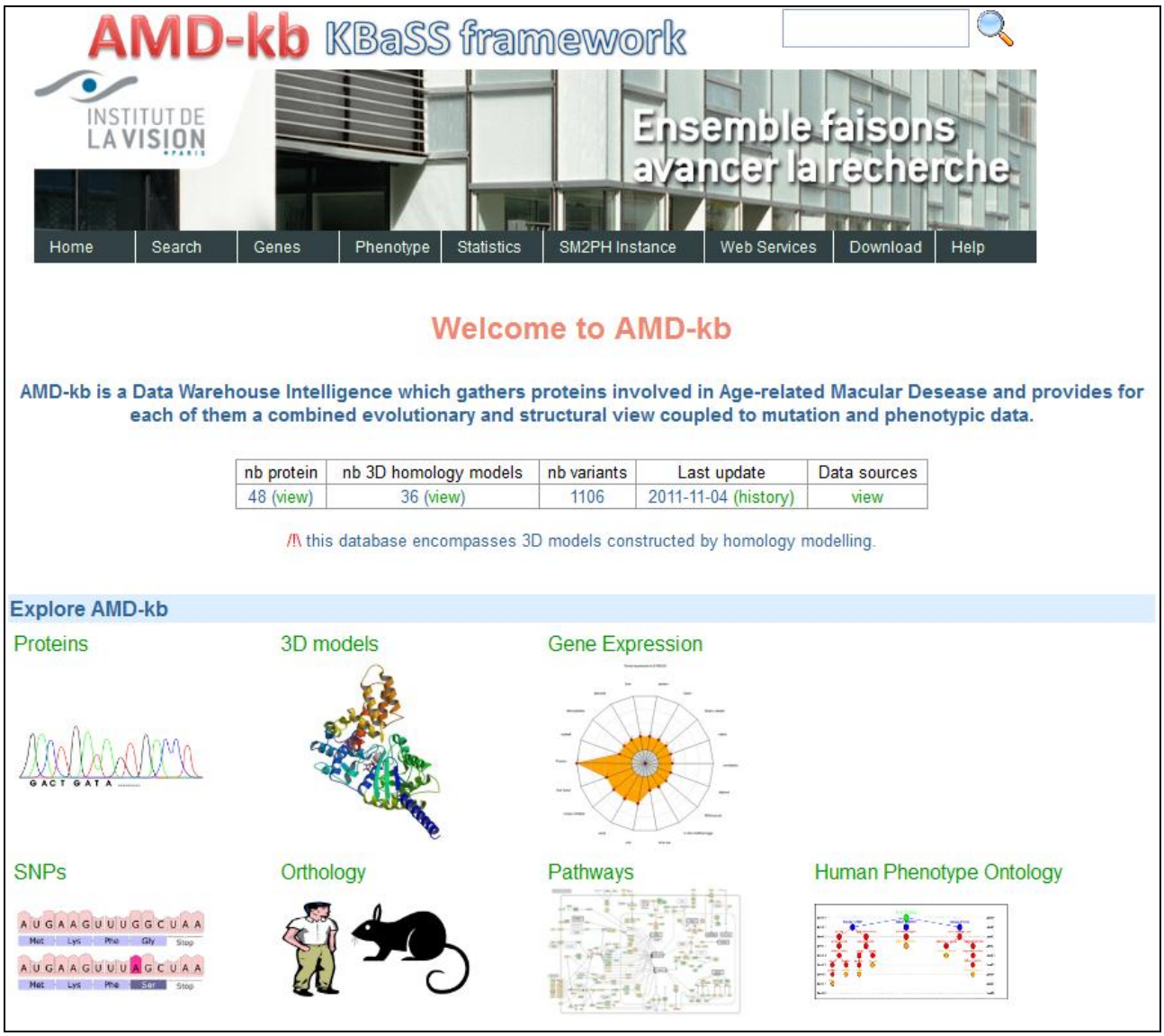


Figure 48. SM2PH-AMD-kb, une SM2PH-Instance consacrée à l'étude de la Dégénérescence Maculaire Liée à l'Âge.

CHAPITRE 6. MSV3D : UN SYSTEME DEDIE A L'ANALYSE GLOBALE DES MUTATIONS FAUX-SENS

6.1 Introduction

L'objectif principal de ma thèse réside dans la mise à disposition d'une infrastructure bioinformatique visant à faciliter la compréhension des relations génotype–phénotype à différents échelles : macroscopique, cellulaire et moléculaire. Grâce au développement de SM2PH Central, présenté dans le chapitre précédent, nous disposons d'un environnement de travail permettant d'analyser et d'explorer l'ensemble des gènes/protéines humains intégrant les niveaux d'informations, allant, des aspects structuraux, évolutifs et génomiques en passant par la transcriptomique, l'interactomique, la protéomique et la métabolomique. La volonté d'associer à SM2PH Central des mutations faux-sens et leurs annotations nous a amenés à développer MSV3d. MSV3d est un système d'information dédié à l'analyse globale des mutations faux-sens et de leurs conséquences structurales et phénotypiques dans le cadre des maladies génétiques humaines.

Près de 90% des variations du génome concernent la modification d'une seule base de l'ADN appelée *Single Nucleotide Polymorphism* (*SNP*) (Collins et al., 1998). Parmi ceux-ci, on distingue les mutations faux-sens (nsSNPs), qui induisent le changement d'un acide aminé au niveau de la protéine (Stenson et al., 2008). MSV3d (http://decrypthon.igbmc.fr/msv3d) offre à la communauté un accès rapide à l'information annotée des mutations faux-sens impliqués dans toutes les protéines humaines. Pour cela, MSV3d intègre toutes les mutations faux-sens connues disponibles dans les banques de données publiques : dbSNP, UniProt et quelques banques de données spécifiques d'un locus (ou *LSDB*). Ces mutations sont séparées en 2 grandes classes : celles qui n'auront pas d'impact et qui seront fonctionnellement neutres pour la protéine, et celles qui auront un effet délétère sur la protéine et qui seront potentiellement associées à une maladie génétique. Une suite de programmes (~ 10 programmes) est lancée parallèlement pour annoter des mutations faux-sens avec 33 paramètres. Ainsi, pour chaque mutation faux-sens, nous avons des éléments qui décrivent la modification introduite dans la taille, la charge, la polarité par exemple. Ces informations ont été codées par un système à 3 états correspondants à l'augmentation, la diminution ou la conservation d'une propriété. De même, différentes données concernant le degré de conservation du résidu muté (disponible dans SM2PH Central) ont été codées sous la forme d'entier ou de pourcentage. Enfin, des informations concernant l'environnement 3D du résidu muté, telles que sa position dans une structure secondaire, le nombre de contacts gagnés, perdus ou maintenus, ont été codées dans la base de données. MSV3d est aussi équipé d'un service d'annotation. L'utilisateur peut faire analyser ses propres mutations soit par une interface graphique soit par les APIs (Java, C#).

MSV3d est facilement accessible grâce à une simple interface web couplée à un moteur de recherche puissant supporté par BIRD et par des services web. Son contenu est totalement ou partiellement téléchargeable en fichiers aux formats plats ou XML. Cette caractéristique, qui permet d'intégrer des données traitées complexes dans d'autres traitements ou outils d'analyse, devient incontournable dans le domaine des bases de données biologiques. Pour l'anecdote, on peut noter que ce point a été particulièrement apprécié par les rapporteurs de notre article accepté dans le journal Database : « *I am particularly impressed with the SOAP API and the easy links to download the data in text formats. This kind of interface polish is*

very much appreciated, since many existing tools have really unpleasant user interfaces and are difficult to script. ».

Les données de MSV3d peuvent également être utilisées comme un jeu de données de référence pour le chercheur qui veut développer et tester une méthode d'apprentissage automatique pour la classification ou la prédiction de mutations délétères/neutres.

Le contenu principal de MSV3d est décrit dans l'article « MSV3d: database of human MisSense variants mapped to 3D protein structure » publié dans le journal Database (Oxford).

6.2 Publication

Luu TD, Rusu AM, Walter V, Ripp R, Moulinier L, Muller J, Toursel T, Thompson JD, Poch O, Nguyen NH. *MSV3d: database of human MisSense variants mapped to 3D protein structure.* Database (Oxford). 2012.

Database, Vol. 2012, Article ID bas018, doi:10.1093/database/bas018

Database tool

MSV3d: database of human MisSense variants mapped to 3D protein structure

Tien-Dao Luu[1], Alin-Mihai Rusu[1], Vincent Walter[1], Raymond Ripp[1], Luc Moulinier[1], Jean Muller[1,2], Thierry Toursel[3], Julie D. Thompson[1], Olivier Poch[1] and Hoan Nguyen[1,*]

[1]Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire (UMR7104), 67404 Illkirch, [2]Laboratoire de Diagnostic Génétique, CHU Strasbourg Nouvel Hôpital Civil, 67000 Strasbourg and [3]Association Française contre les Myopathies, 91002, EVRY cedex, France

***Corresponding author:** Tel: +33 (0)3 88 65 32 65; Fax : +33 (0)3 88 65 32 01; Email: nguyen@igbmc.fr

Submitted 5 December 2011; Revised 6 March 2012; Accepted 8 March 2012

The elucidation of the complex relationships linking genotypic and phenotypic variations to protein structure is a major challenge in the post-genomic era. We present MSV3d (Database of human MisSense Variants mapped to 3D protein structure), a new database that contains detailed annotation of missense variants of all human proteins (20 199 proteins). The multi-level characterization includes details of the physico-chemical changes induced by amino acid modification, as well as information related to the conservation of the mutated residue and its position relative to functional features in the available or predicted 3D model. Major releases of the database are automatically generated and updated regularly in line with the dbSNP (database of Single Nucleotide Polymorphism) and SwissVar releases, by exploiting the extensive Décrypthon computational grid resources. The database (http://decrypthon.igbmc.fr/msv3d) is easily accessible through a simple web interface coupled to a powerful query engine and a standard web service. The content is completely or partially downloadable in XML or flat file formats.

Database URL: http://decrypthon.igbmc.fr/msv3d

Introduction

Single nucleotide polymorphisms (SNPs) refer to a genetic change in which one nucleotide is replaced by another one and represent one of the most common forms of human genomic variation. Although SNPs are primarily associated with population diversity and individuality, they can also be linked to the emergence or the predisposition to disease, influencing its severity, its progression or its drug sensitivity. Several public repositories of SNPs exist, including GWAS Central (1), SwissVar (2) and dbSNP (3). Among these, dbSNP is probably the most extensive, with release 135 hosting more than 50 million human SNPs including 535 660 synonymous and 873 308 non-synonymous SNPs. The non-synonymous SNPs (nsSNPs), also called missense variants, are particularly important since they result in an alteration of the amino acid sequence of the encoded protein. Missense variants have been linked to a wide variety of diseases, for example by affecting protein function, by reducing protein solubility or by destabilizing protein structure (4). With the huge amount of protein information now available in various biological databases, including sequences, structures, functions, interactions, pathways, together with the development of *in silico* analysis tools, it is now possible to better understand the correlation between a missense mutation and the associated molecular phenotypes. Research groups have addressed this topic and have developed tools aimed at predicting the effects of missense variants on the function of a protein and its 3D structure, with varying degrees of success [for recent reviews, see refs (5–7)].

Over the last decade, numerous web servers have been developed to explore the effects of missense variants on

Database, Vol. 2012, Article ID bas018, doi:10.1093/database/bas018

characterized protein 3D structures and functions, including ModSNP (8), PolyPhen-2 (9), SNPs3D (10), StSNP (11), TopoSNP (12), LS-SNP (13), SNPeffect (14), MutDB (15), etc., which are publically available on the internet. These bioinformatics resources have different strengths and weaknesses (16). Although many of the web servers provide predictive analyses, according to a recent review (17), 'most of the tested servers use NCBI's dbSNP database as a primary source of SNP data, but are not up-to-date, increasing the chances that annotations for SNPs of interest will not be available to users'.

We have previously developed the SM2PH (from Structural Mutation to Pathology Phenotypes in Human) infrastructure (18), with the goal of contributing new useful bioinformatics resources dedicated to monogenic disorders for the research community. SM2PH-db was developed in the framework of the Décrypthon grid project (19), resulting from a collaboration between the AFM (French Muscular Dystrophy Association), IBM, and the CNRS (French National Research Centre) and involved the creation of a suite of an online analysis and visualization tools for the analysis of the correlation between genetic variations in disease-causing genes and the associated human phenotype. The initial version of SM2PH-db (18) contained about 2300 genes involved in human monogenetic diseases and has been regularly and automatically updated since September 2009. An integrative 'genotype–phenotype relationship' analysis was also performed, involving the characterization of how genetic alterations affect gene products (proteins) at the molecular level. Thus, the phenotypes associated with human pathologies were represented by their structural, functional and evolutionary context of all genes/proteins known to be involved in human diseases.

In this context, we designed SM2PH Central (Figure 1): a gene centric knowledgebase dedicated to the integration of and unified access to the information associated with any human protein (pathway, tissue expression, interactions, evolution, etc.). SM2PH Central provides access to a wide range of interconnected information that provides a global view from the gene to the phenotype. The system can be used to automatically generate any SM2PH database instance (i.e. a specialized database centred on a thematic use case, for example genes involved in a specific human disease, gene lists resulting from high-throughput analysis, etc.).

Here we present MSV3d, a new database of previously identified missense variants involved in all human proteins mapped to 3D structure. MSV3d provides a unified access for SM2PH Central and its instances, for integration in any specialized database requiring interoperable and interconnected missense-related data and information, as well as for the biological community. The database is dedicated to automatic annotation of all human missense variants involved in 20 199 human proteins, thus covering all genes and diseases included in the Online Mendelian Inheritance in Man (OMIM) database (20). It facilities user exploration of the relationships between genetic variations and 3D structure via a unified access to databases, including SOAP web services, a Java API, simple queries and full or partial database download services. Statistical plots dynamically coupled with a powerful query engine allow the user to filter and analyse the data. In addition, the database also represents a useful benchmark set for any researcher who wants to develop and evaluate a machine learning method for classification or prediction of deleterious/neutral mutations. The database is automatically updated every 3 months and a major release is performed every year.

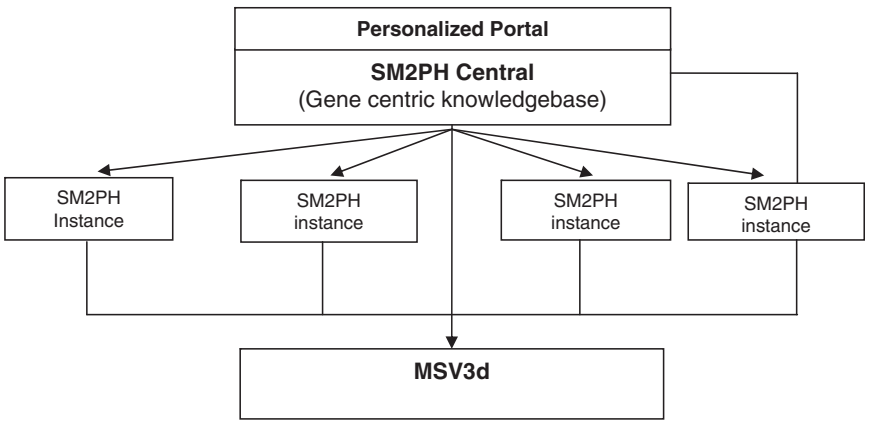


Figure 1. General architecture of SM2PH central. SM2PH Central allows the generation of SM2PH instances (focusing on specific sets of target genes) which can access variant information through the new MSV3d, devoted to human variant data and information.

Description of database

Database content

The human missense variants in MSV3d are mainly retrieved from the dbSNP and SwissVar databases, but also from several locus-specific databases (LSDBs), e.g. the ALPL gene mutations database (http://www.sesep.uvsq.fr/database_hypo/Mutation.html). We classified all these variants in two main categories: disease-causing variant associated with OMIM diseases and Variant(s) of Uncertain Significance (VUS). In MSV3d, each missense variant is then characterized using four main levels of information:

Mutant information. This level involves data related to the gene and its associated protein, the chromosome position, the OMIM disease and genotype population reference. Pathogenicity prediction scores from external tools are provided by locally running the latest version of SIFT (21) and Polyphen-2 (9) to predict damaging effects of all missense variants in MSV3d. The SCOP fold classification (22) is also identified.

Conservation and physico-chemical changes. This level covers the information related to the mutated position in the context of its protein family. A multiple sequence alignment of the protein and up to 500 homologues [*UniRef90* (23)] is constructed using PipeAlign (24) and annotated by MACSIMS (25). The MACSIMS annotation provides several descriptions of conservation, such as the conservation score of the substituted position, the percentage of mutated residues at the same position and the number of known mutations at this position. The physico-chemical changes induced by the amino acid substitution such as modifications in size, charge, polarity and hydrophobicity have been described previously (26). Modification of glycine or proline in the mutation is also identified. A global score reflecting the degree of modification induced by the substitution is also assigned. This score corresponds to the distance between the substituted residues based on a vector representation of the amino acids (see the MSV3d website for more details), where larger distances imply less conservative substitutions.

Structural features and modifications. These features include the structural annotations provided by MACSIMS, as well as detailed descriptions of the 3D context (e.g. residue relative accessibility, contact with an annotated site, etc.). Structural modifications induced by the amino acid substitution are predicted based on the mutant 3D models. These are automatically constructed using MODELLER (27) for missense variants that can be mapped onto a wild-type 3D model sharing $>50\%$ identity with the template used for the model construction. Secondary structures are deduced from the PDB (28) entry using the DSSP program (29). The effect in the protein relative stability upon single-site mutation is predicted with I-Mutant2.0 (30).

Spatial contacts. Four types of spatial contact have been defined: (i) the contacts between a residue and its direct 3D neighbours, based on the wild-type 3D model, (ii) the contacts between a mutant residue and its direct 3D neighbours based on the mutant 3D model, (iii) the contacts between residues in contact with the wild-type residue and their direct 3D neighbours, based on the wild-type 3D model and (iv) the contacts between residues in contact with the mutant residue and their direct 3D neighbours, based on the mutant 3D model.

Database statistics

MSV3d currently contains more than 445 574 missense variants mapped to 20 199 human proteins. Of these missense variants, 58 159 were found in SwissVar, 424 541 in dbSNP (build 135) and 37 209 in both SwissVar and dbSNP. A total of 24 379 the missense variants are considered as disease-causing variants and 421 195 as VUS.

Concerning the structural data, 10 713 structural templates from the PDB database have been identified allowing the mapping of 63 528 variants to a 3D structure. Among those mapped variants, 13 421 are identified in 265 SCOP fold classifications and 8023 variants are associated with 1479 OMIM diseases. Concerning gene conservation and function, 49 164 variants are mapped to one of the 2342 functional domains identified in the database (extracted from the Pfam protein family database (31), validated and propagated by MACSIMS) and 1799 HPO ontology terms from the HPO (Human Phenotype Ontology) database (32).

Up-to-date statistics concerning the physico-chemical changes induced by the amino acid substitutions, the conservation patterns, the localization in a secondary structure and/or functional domain are available on the 'Statistics' page of the website. Distributions of missense variants in SCOP folds or Pfam domains are also provided. As an example, Figure 2 illustrates the top 20 SCOP folds enriched in missense variants. By default, these statistics take into account the missense variants of all genes in the database. However, the user can also submit his own gene list in order to personalize the statistics analysis.

Web interface and search engine

The MSV3d web interface (Figure 3) is designed to allow the user to rapidly query the complete database, for example by entering a protein name, gene name, SNP ID, OMIM ID, PDB template ID, chromosome position, protein fold or Pfam domain and to retrieve and export a list of missense variant data. MSV3d also provides a powerful full-text search service, allowing the user to search for any

Database, Vol. 2012, Article ID bas018, doi:10.1093/database/bas018

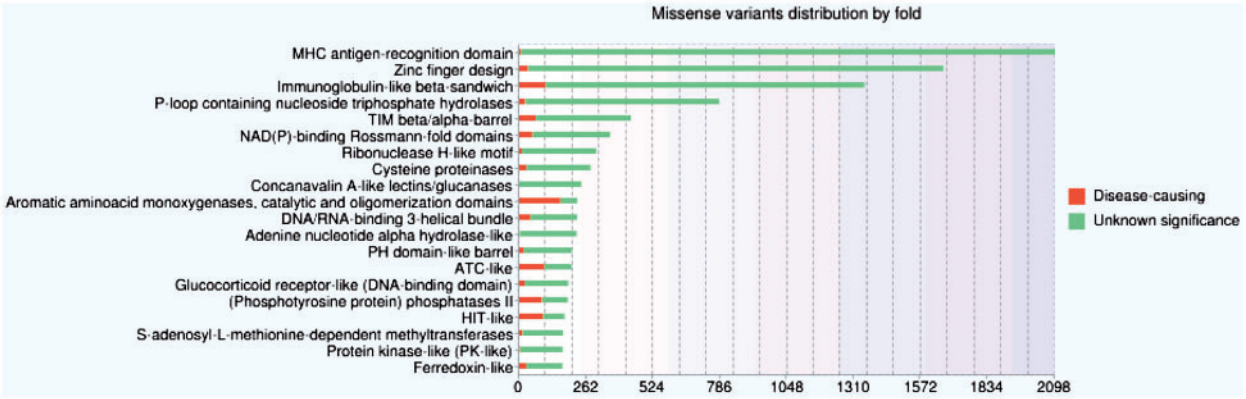


Figure 2. Histogram showing distributions of missense variants by SCOP fold. Each bar contains two parts: the red part represents deleterious substitutions and the green part represents tolerated substitutions.

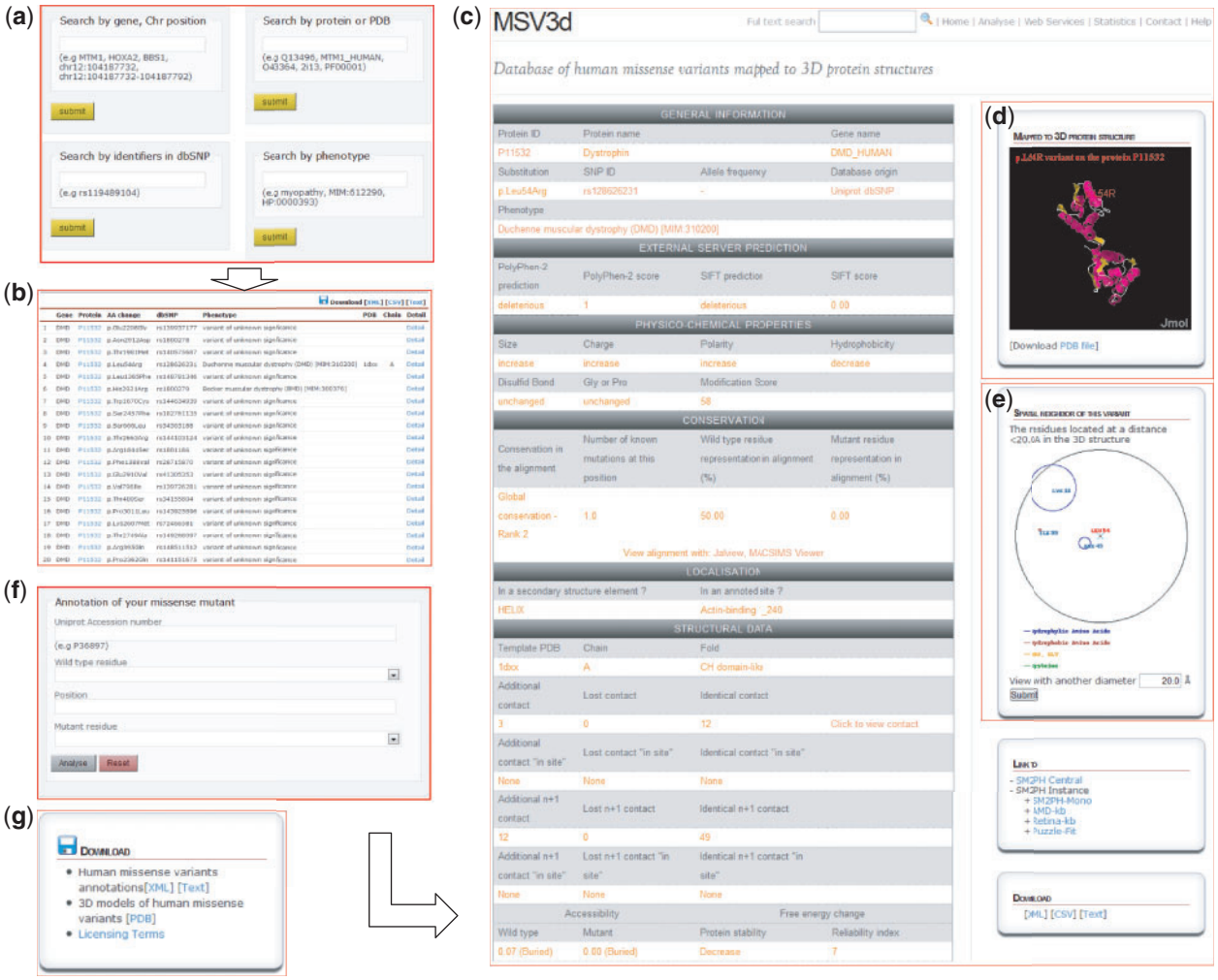


Figure 3. MSV3d web interface contains numerous functionalities including: (**a**) field search, (**b**) free text search, (**c**) detailed information, (**d**) 3D structure visualization using Jmol, (**e**) spatial neighbouring residue visualization, (**f**) missense annotation service and (**g**) download service.

keyword stored in MSV3d without restriction to the index and field names. The results of a search can be visualized on the web or downloaded (Figure 3g) in a variety of formats such as XML or flat files. The user can also download the full database release in different formats.

To facilitate the structural analysis of missense mutation, we have incorporated the Jmol software (33) in the MSV3d interface. The Jmol applet is loaded automatically with an available structure model when a variant is selected on the web interface. Figure 3d shows the Jmol-based visualization for variations mapped onto the 3D structure. Mutations are automatically highlighted and neighbouring variants can also be identified.

Finally, the environment of neighbouring amino acid residues around a missense mutation is defined as follows: residues are considered to be neighbours of a mutation if they occur within a limited sphere in 3D space. Figure 3e shows the neighbouring residues of the p.Leu54Arg mutation in protein P11532 (with radius 20 Å).

Missense variant annotation with standard web service

The user can annotate a new missense variant using the web interface (Figure 3f) or using a programming interface via a SOAP web service. SOAP provides standard interoperability functions to communicate between applications running on different operating systems, with different programming languages. The SOAP WSDL protocol and API client of MSV3d (Java and C#) can be downloaded from our website (http://decrypthon.igbmc.fr/msv3d/cgi-bin/webservices).

Database construction

MSV3d pipeline

Taking advantage of the previous developments (18, 19), we have designed the MSV3d pipeline, involving more than 20 programs, firstly, to facilitate the investigation of the structural impacts of known or unknown missense mutations from all 20 199 human proteins, thus covering all known human genetic diseases and secondly, to guarantee a rapid update of the complete database content. The software pipeline has been deployed with high interoperability between all programs and their parallel application.

The schema in Figure 4 shows the main steps in the MSV3d pipeline. In general, the MSV3d pipeline takes a protein sequence as input and extracts associated missense variants from public databases. For each sequence, similarity searches are performed in public databases stored in the BIRD System (19), which is a local data warehouse supported by IBM DB2. BIRD provides a common architecture and relational schema for the integration of both local and public databases, as well as a unifying query system

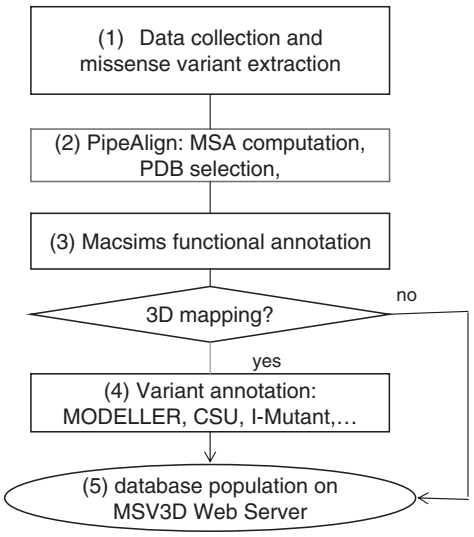


Figure 4. Schema of software pipeline.

(BIRD-QL) for non-integrated data. The identification of background conservation and reconstruction of the evolutionary history of each reference sequence is based on Multiple Alignments of Complete Sequences (MACS) (34), thanks to a modified version of the PipeAlign program (24) and to the MACSIMS (MACS Information Management System) software. After PDB template selection and modelling where necessary, variant annotation is performed in the context of the 3D structure. The main steps in the process of MSV3d database creation are as follows:

Step 1: data input and missense variant extraction. This involved the implementation of automated protocols for the creation of a comprehensive collection of data related to (i) missense variants obtained from dbSNP, SwissVar and LSDBs, (ii) phenotypes obtained from the OMIM database. Most applications in the MSV3d pipeline use specialized BIRD-QL queries via the http protocol in order to automate retrieval, integration and mining of information in dbSNP and associated data such as proteins, genes and phenotype information.

Step 2: PipeAlign. The sequence analysis process (PDB selection, conservation, evolutionary information, etc.) has been automated using our in-house software cascade that has been shown to be robust and efficient (9). The PipeAlign cascade integrates eight programs to process: (i) protein sequence and structure database searches (Blast, Ballast), (ii) multiple sequence alignment creation [DbClustal (35)], correction [Rascal (36)] and quality estimation [NorMD (37), Leon (38)] and (iii) hierarchical classification into subfamilies [DPC (39), Secator (40)]. To address the challenges of the current sequence data deluge, a modified version of PipeAlign integrating a sequence sampling step

Database, Vol. 2012, Article ID bas018, doi:10.1093/database/bas018

(Sampler) after the Blast searches (41) has been implemented in the Décrypthon grid and recently in our local cluster at the Institute of Genetics and Molecular and Cellular Biology (IGBMC). More than 20 000 MSA are computed and indexed in the local data warehouse. The availability of a 3D structure or model of the protein is essential to gain insight into the structural impact of a missense variant. The best source of protein structural information is the PDB (28), which stores almost all the experimentally resolved crystallographic structures. 3D models of the wild-type proteins are automatically constructed by homology, using MODELLER (27). The models are built by inferring the structure of a protein (the target) from the structure of another putatively homologous protein solved by experimental methods (the template). Five homology models are constructed and the one with the best normalized DOPE score (27) is integrated in MSV3d.

Step 3: MACSIMS functional annotation. To characterize the background conservation and exploit different types of evolutionary data, we used MACSIMS to annotate the MACS with information such as: (i) taxonomic data, (ii) functional descriptions, (iii) known domains or domains similar to a known 3D structure, (iv) potential disordered regions, (v) blocks that do not correspond to disordered regions or known domains but that are conserved at the family or subfamily level and thus may constitute uncharacterized domains and (vi) conservation pattern of domains and residues. All the information associated with the MACS is collected and stored in XML format files.

Step 4: missense variant annotation. If the variant position is mapped to a 3D structure identified in Step 3, the structural context of each individual mutation is modelled based on 33 descriptors combining sequence/structure-related data using several software tools such as MODELLER, CSU (42), I-Mutant (30) (detail of the descriptors and computational software are provided on the MSV3d website).

Step 5. Finally, the full database is populated on the web server thanks to the BIRD-QL query engine, which is capable of managing the large volumes of heterogeneous data and provides up-to-date biological data for MSV3d.

Computer resource specification

To rapidly generate and update the very large database content, we use the Décrypthon grid and the IGBMC local cluster. The Décrypthon grid represents a total of 58 machines and 475 processors distributed on six nodes. The servers include multiprocessor machines (4–16 physical processors) under the AIX operating system and a cluster of single processor machines under the Linux system. To guarantee a permanent powerful CPU resource, we also deployed the complex software pipeline on a local cluster, representing a total of 16 machines and 240 processors. In order to facilitate the deployment of the MSV3d pipeline in the grid environment and local cluster, we developed interoperability protocols for the various programs and automatic procedures to compute, transfer and integrate the information from heterogeneous sources (software and biological databases) as well as to perform regular updates. Today, the complete update and annotation process for all human proteins (20 199 proteins and more than 400 000 missense variants) in MSV3d, takes up to 1 week.

Conclusions and future work

The large missense variant database mapped to 3D structures with regular updates is available for our scientific partners and for the wider community. Several access standards were developed to allow users to rapidly identify and retrieve the variant annotations. Facilitated access to such databases is an essential step to better understand how human genetic alterations affect the gene products at the structural level and subsequently to elucidate the relationships between genotypic and phenotypic variations. We have improved our original infrastructure and architecture in order to rapidly generate and manage the new data concerning all human proteins, thus facilitating an integrated approach to study any human genetic disease. The main advantages of MSV3d are (i) the full missense variant annotation for proteins without PDB structures, based on automated 3D modelling and (ii) the ergonomic and comprehensive database interface complemented with a SOAP-based remote API.

In the future, we plan to enhance the data integration by including structural surface topology descriptions using the M-ORBIS (for Mapping of mOleculaR Binding sItes and Surfaces) approach (43). This method, based on α-shape analysis, allows the precise mapping of different 'functional' regions such as the protein core and the non-interacting or interacting surfaces. The latter can then be further characterized as participating in homodimeric, heterodimeric, protein–peptide, protein–small peptide or protein–ligand interactions. With richer and more relevant knowledge, we hope to discover and extract pertinent relationships between missense variants and structural information using Inductive Logic Programming (44) or Support Vector Machine (45) approaches. We will also incorporate a novel formalism for the representation of protein evolutionary histories in the form of Evolutionary Barcodes (46). This new formalism allows the integration of different evolutionary parameters in a unifying format and facilitates the multilevel analysis and visualization of complex evolutionary histories. In the next major release of MSV3d, annotations for every possible amino acid replacement in the proteins will be integrated in the database. Finally, concerning data standards

and semantic interoperability of biological data, we plan to implement a BioMart (47) interface to facilitate the exploitation and diffusion of our database in the biological community. More specifically, the standards proposed by the BioDBcore consortium (48) will be incorporated in MSV3d.

Acknowledgements

The authors are grateful to Serge Uge for his assistance in implementing the local cluster. The IGBMC common services and platforms are acknowledged for assistance.

Funding

The work was performed within the framework of the Decrypthon program, co-funded by Association Française contre les Myopathies (AFM, 14390-15392), IBM and Centre National de la Recherche Scientifique (CNRS). We acknowledge financial support from the ANR (EvolHHuPro: BLAN07-1-198915) and Institute funds from the CNRS, INSERM, and the Université de Strasbourg. Funding for open access charge: ANR-10-BINF-03-02.

Conflict of interest. None declared.

References

1. Thorisson,G.A., Lancaster,O., Free,R.C. *et al*. (2009) HGVbaseG2P: a central genetic association database. *Nucleic Acids Res.*, **37**, D797–D802.
2. Mottaz,A., David,F.P., Veuthey,A.L. and Yip,Y.L. (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, **26**, 851–852.
3. Sherry,S.T., Ward,M.H., Kholodov,M. *et al*. (2001) dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res.*, **29**, 308–311.
4. Chasman,D. and Adams,R.M. (2001) Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.*, **307**, 683–706.
5. Thusberg,J. and Vihinen,M. (2009) Pathogenic or not? And if so, then how? Studying the effects of missense mutations using bioinformatics methods. *Hum. Mutat.*, **30**, 703–714.
6. Jordan,D.M., Ramensky,V.E. and Sunyaev,S.R. (2010) Human allelic variation: perspective from protein function, structure, and evolution. *Curr. Opin. Struct. Biol.*, **20**, 342–350.
7. Cline,M.S. and Karchin,R. (2011) Using bioinformatics to predict the functional impact of SNVs. *Bioinformatics*, **27**, 441–448.
8. Yip,Y.L., Scheib,H., Diemand,A.V. *et al*. (2004) The Swiss-Prot variant page and the ModSNP database: a resource for sequence and structure information on human protein variants. *Hum. Mutat.*, **23**, 464–470.
9. Adzhubei,I.A., Schmidt,S., Peshkin,L. *et al*. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
10. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
11. Uzun,A., Leslin,C.M., Abyzov,A. and Ilyin,V. (2007) Structure SNP (StSNP): a web server for mapping and modeling nsSNPs on protein structures with linkage to metabolic pathways. *Nucleic Acids Res.*, **35**, W384–W392.
12. Stitziel,N.O., Binkowski,T.A., Tseng,Y.Y. *et al*. (2004) topoSNP: a topographic database of non-synonymous single nucleotide polymorphisms with and without known disease association. *Nucleic Acids Res.*, **32**, D520–D522.
13. Karchin,R., Diekhans,M., Kelly,L. *et al*. (2005) LS-SNP: large-scale annotation of coding non-synonymous SNPs based on multiple information sources. *Bioinformatics*, **21**, 2814–2820.
14. Reumers,J., Schymkowitz,J., Ferkinghoff-Borg,J. *et al*. (2005) SNPeffect: a database mapping molecular phenotypic effects of human non-synonymous coding SNPs. *Nucleic Acids Res.*, **33**, D527–D532.
15. Singh,A., Olowoyeye,A., Baenziger,P.H. *et al*. (2008) MutDB: update on development of tools for the biochemical analysis of genetic variation. *Nucleic Acids Res.*, **36**, D815–D819.
16. Tavtigian,S.V., Greenblatt,M.S., Lesueur,F. and Byrnes,G.B. (2008) In silico analysis of missense substitutions using sequence-alignment based methods. *Hum. Mutat.*, **29**, 1327–1336.
17. Karchin,R. (2009) Next generation tools for the annotation of human SNPs. *Brief. Bioinform.*, **10**, 35–52.
18. Friedrich,A., Garnier,N., Gagnière,N. *et al*. (2010) SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases. *Hum. Mutat.*, **31**, 127–135.
19. Bard,N., Bolze,R., Caron,E. *et al*. (2010) Decrypthon grid - grid resources dedicated to neuromuscular disorders. *Studies Health Technol. Informatics*, **159**, 124–133.
20. McKusick,V.A. (2007) Mendelian inheritance in man and its online version, OMIM. *Am. J. Hum. Genet.*, **80**, 588–604.
21. Kumar,P., Henikoff,S. and Ng,P.C. (2009) Predicting the effects of coding non-synonymous variants on protein function using the SIFT algorithm. *Nat. Protoc.*, **4**, 1073–1081.
22. Andreeva,A., Howorth,D., Chandonia,J.M. *et al*. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
23. Suzek,B.E., Huang,H., McGarvey,P. *et al*. (2007) UniRef: comprehensive and non-redundant UniProt reference clusters. *Bioinformatics*, **23**, 1282–1288.
24. Plewniak,F., Bianchetti,L., Brelivet,Y. *et al*. (2003) PipeAlign: a new toolkit for protein family analysis. *Nucleic Acids Res.*, **31**, 3829–3832.
25. Thompson,J.D., Muller,A., Waterhouse,A. *et al*. (2006) MACSIMS: multiple alignment of complete sequences information management system. *BMC Bioinformatics*, **7**, 318.
26. Taylor,W.R. (1986) The classification of amino acid conservation. *J. Theor. Biol.*, **119**, 205–218.
27. Eswar,N., Eramian,D., Webb,B. *et al*. (2008) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **426**, 145–159.
28. Berman,H.M., Westbrook,J., Feng,Z. *et al*. (2000) The protein data bank. *Nucleic Acids Res.*, **28**, 235–242.
29. Joosten,R.P., te Beek,T.A., Krieger,E. *et al*. (2011) A series of PDB related databases for everyday needs. *Nucleic Acids Res.*, **39**, D411–D419.
30. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
31. Finn,R.D., Mistry,J., Tate,J. *et al*. (2010) The Pfam protein families database. *Nucleic Acids Res.*, **38**, D211–D222.

Database, Vol. 2012, Article ID bas018, doi:10.1093/database/bas018

32. Robinson,P.N. and Mundlos,S. (2010) The human phenotype ontology. *Clin. Genet.*, **77**, 525–534.
33. Hanson,R. (2010) Jmol - a paradigm shift in crystallographic visualization. *J. Appl. Crystallogr.*, **43**, 1250–1260.
34. Lecompte,O., Thompson,J.D., Plewniak,F. *et al*. (2001) Multiple alignment of complete sequences (MACS) in the post-genomic era. *Gene*, **270**, 17–30.
35. Thompson,J.D., Plewniak,F. *et al*. (2000) DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. *Nucleic Acids Res.*, **28**, 2919–2926.
36. Thompson,J.D., Thierry,J.C. and Poch,O. (2003) RASCAL: rapid scanning and correction of multiple sequence alignments. *Bioinformatics*, **19**, 1155–1161.
37. Thompson,J.D., Plewniak,F., Ripp,R. *et al*. (2001) Towards a reliable objective function for multiple sequence alignments. *J. Mol. Biol.*, **314**, 937–951.
38. Thompson,J.D., Prigent,V. and Poch,O. (2004) LEON: multiple aLignment Evaluation Of Neighbours. *Nucleic Acids Res.*, **32**, 1298–1307.
39. Wicker,N., Dembele,D., Raffelsberger,W. and Poch,O. (2002) Density of points clustering, application to transcriptomic data analysis. *Nucleic Acids Res.*, **30**, 3992–4000.
40. Wicker,N., Perrin,G.R., Thierry,J.C. and Poch,O. (2001) Secator: a program for inferring protein subfamilies from phylogenetic trees. *Mol. Biol. Evol.*, **18**, 1435–1441.
41. Friedrich,A., Ripp,R., Garnier,N. *et al*. (2007) Blast sampling for structural and functional analyses. *BMC Bioinformatics*, **8**, 62.
42. Sobolev,V., Sorokine,A., Prilusky,J. *et al*. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
43. Albou,L.P., Poch,O. and Moras,D. (2011) M-ORBIS: mapping of molecular binding sites and surfaces. *Nucleic Acids Res.*, **39**, 30–43.
44. Muggleton,S. (1991) Inductive logic programming. *New Generation Comput.*, **8**, 295–318.
45. Vapnik,V. (1998) *Statistical Learning Theory*. John Wiley & Sons Inc, New York.
46. Linard,B., Nguyen,H., Prosdocimi,F. *et al*. (2012) EvoluCode: Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data. *Evol. Bioinform.*, **8**, 61–77.
47. Kasprzyk,A. (2011) BioMart: driving a paradigm change in biological data management. *Database*, **2011**, bar049.
48. Gaudet,P., Bairoch,A., Field,D. *et al*. (2011) Towards BioDBcore: a community-defined information specification for biological databases. *Nucleic Acids Res.*, **39**, D7–D10.

6.3 Contenu de la base de données

Dans la cadre de l'article présenté, nous n'avons pu décrire en détails le contenu de la base de données MSV3d ainsi que le pipeline d'annotation des mutations. Ces éléments sont présentés dans cette section.

La Figure 49 présente le schéma logique de la base de données de MSV3d.

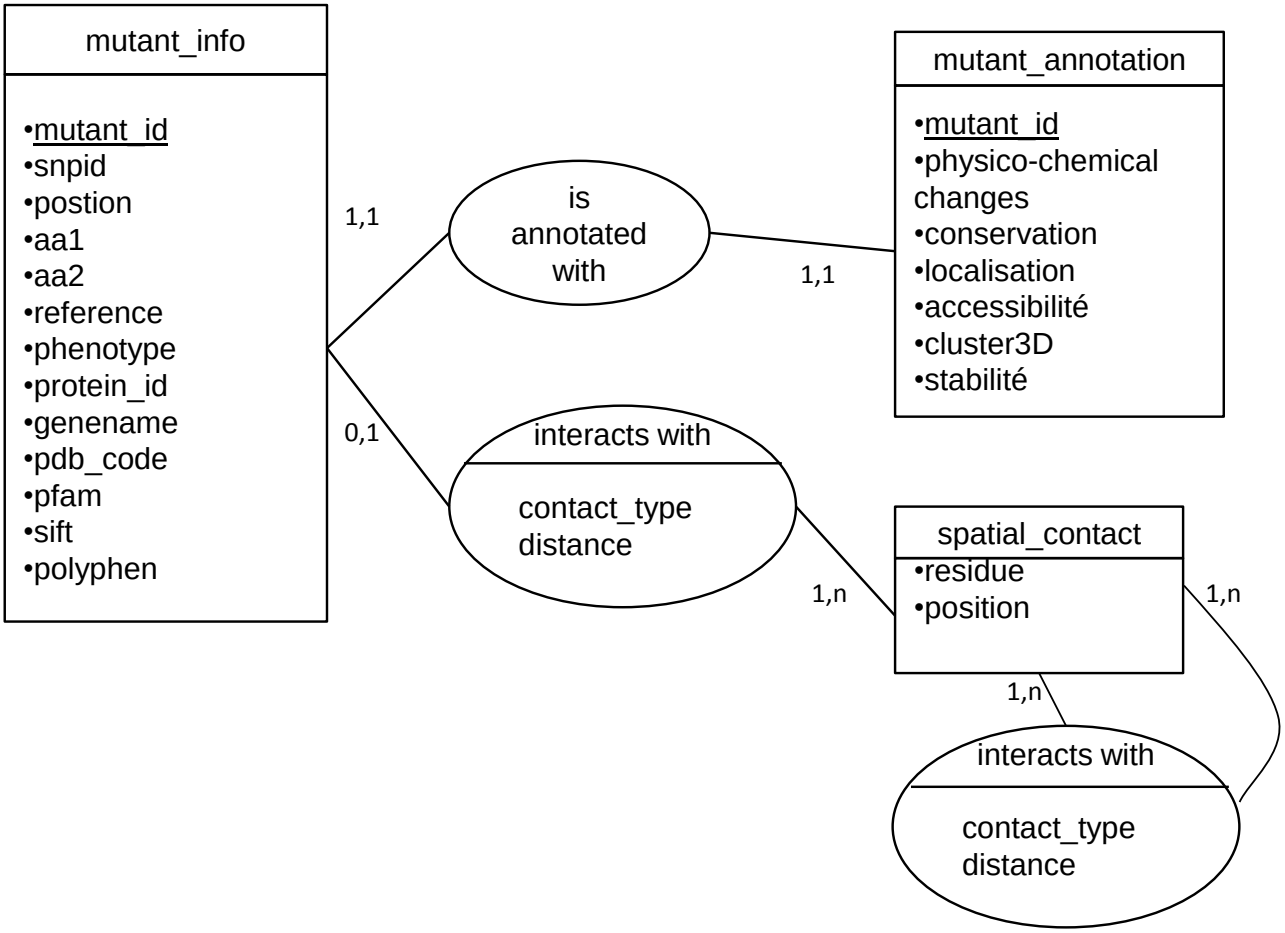


Figure 49. Schéma logique de la base de données de MSV3d.

Chaque entrée MSV3d est une mutation faux-sens. A partir des données de SM2PH Central (des alignements multiples de séquences complètes et des modèles 3D des protéines), nous avons développé un pipeline permettant d'annoter des mutations (Figure 50). Il intègre des programmes externes et des outils développés par Nicolas Garnier et Anne Friedrich dans le projet SM2PH. Pour chacune de ses mutations, MSV3d fournit 2 types de données : des données « connues » et des données « calculées ». Les données « connues » concernent aussi bien les données génotypiques et phénotypiques que les données récupérées automatiquement par MACSIMS (Thompson et al., 2006), notamment celles relatives au découpage en domaines fonctionnels des protéines. Les données « calculées » correspondent, quant à elles, à la conservation des résidus au cours de l'évolution et aux modèles 3D générés pour chacune de nos protéines d'intérêt. Les MACSIMS et les modèles 3D sont disponibles au sein de SM2PH Central.

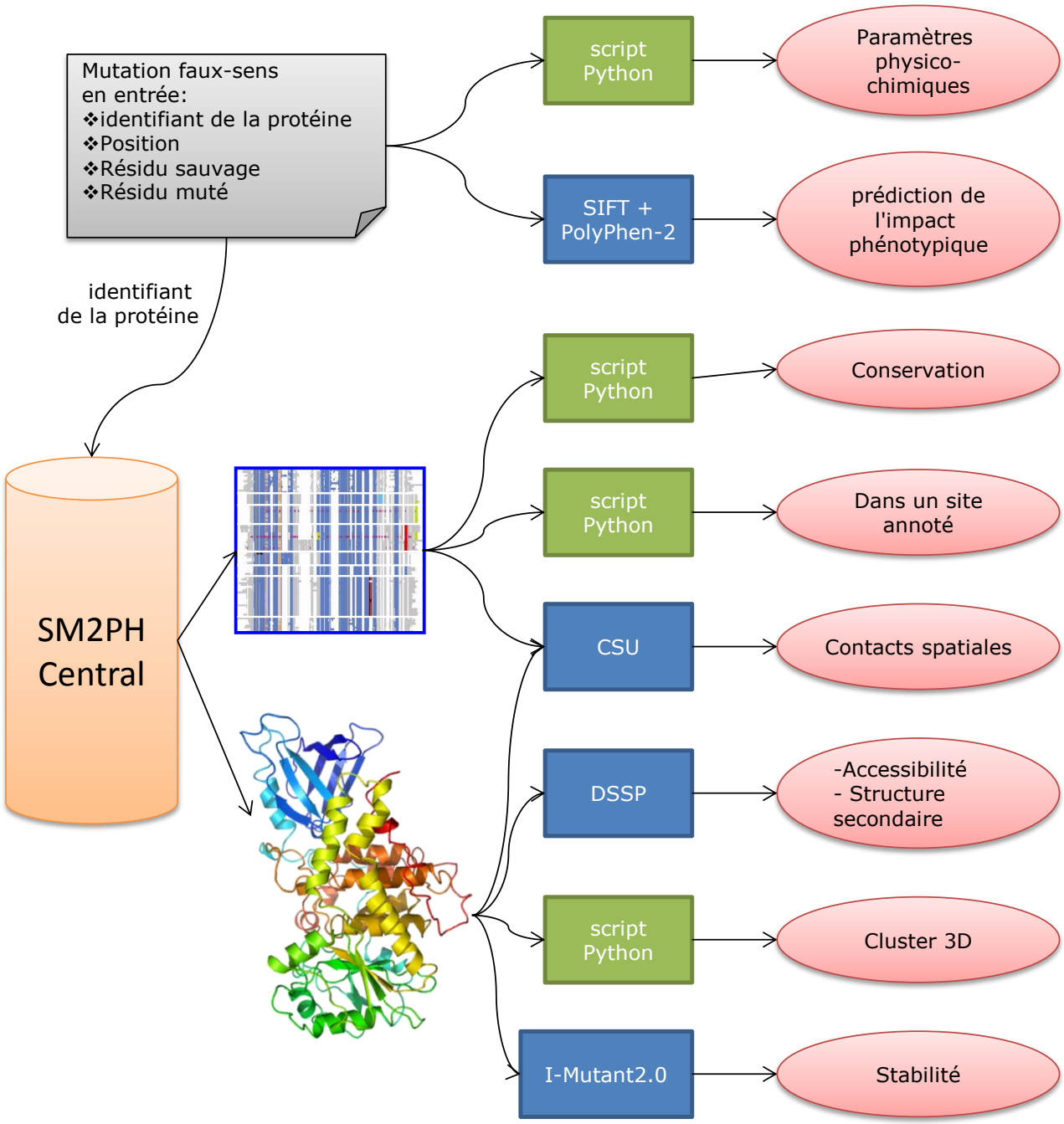


Figure 50. Pipeline d'annotation des mutations de MSV3d. En gris sont représentées les données initiales. Les rectangles bleus symbolisent les programmes externes, les rectangles verts les outils développés dans le projet SM2PH. Les ellipses roses sont les paramètres associés aux mutations.

Le Tableau 7 énumère ces données et leur méthode de calcul :

Type	Paramètres	Value	Méthode de calcul
Entité : mutant_annotation			
Propriétés physico-chimiques	Charge, taille, polarité, hydrophobie	+,-,=, !=	Matrices de substitution générées à partir de la classification des acides aminés de Taylor (Taylor, 1986)
	Rupture/création de pont disulfure	+,-,=	Logiciel DSSP et notre script Python
	Perte/apparition de Glycine ou Proline	+,-,=	Notre script Python

Conservation	Nombre de mutations dans un cluster	entier	Script qui regroupe les mutations dans un cluster si la prochaine mutation est à moins de 5 résidus en séquence.
	Fréquence de mutations à la position	entier	Dénombrement
	Conservation sauvage et mutant	pourcentage	Calcul sur l'alignement multiple
Localisation	Dans une structure secondaire	Helix, sheet, no	Logiciel DSSP
	Dans un site annoté	yes, no	Annotation MACSIMS
Accessibilité	Sauvage et mutant	pourcentage	DSSP
Cluster 3D	Regroupement des mutations en cluster 3D de 10, 20 et 30Å	entier	Script python de calcul de distances géométrique.
Stabilité	Calcul du ΔΔG du sauvage et mutant	augmentation/ diminution	I-Mutant2.0
Entité : Contacts spatiales			
Contacts gagnés	Contacts gagnés	entier	Logiciel CSU et notre script Python de calcul de la perte/gain des contacts spatiaux
	Contacts gagnés un site annoté	texte	
	Contacts n+1 gagnés	entier	
	Contacts n+1 gagnés dans un site annoté	texte	
Contacts perdus	Contacts perdus	entier	
	Contacts perdus un site annoté	texte	
	Contacts n+1 perdus	entier	
	Contacts n+1 perdus dans un site annoté	texte	
Contacts identiques	Contacts identiques	entier	
	Contacts identiques un site annoté	texte	
	Contacts n+1 identiques	entier	
	Contacts n+1 identiques dans un site annoté	texte	

Tableau 7. Ensemble des paramètres caractérisant une mutation. Les symboles utilisés décrivent : '+' : l'augmentation de la propriété suite à la mutation, '-' : la diminution, '=' : la conservation enfin le symbole ' !=' est utilisé lorsque la mutation introduit un changement de charge (positif/négatif).

6.3.1 Entité : mutant_annotation

a. Propriétés physico-chimiques

Chaque résidu ayant certaines caractéristiques physico-chimiques comme la taille, la charge, l'hydrophobie, etc., les mutations conservant ces propriétés seront mieux tolérées que les autres (Bowie et al., 1990). Par exemple, un résidu hydrophobe localisé dans le cœur comme la valine (V) remplacée par une leucine (L) également hydrophobe aura probablement moins d'impact que s'il est remplacé par une arginine (R) hydrophile et chargée.

Les variations de propriétés physico-chimiques sont établies à partir de matrices de substitution générées à partir du diagramme de Venn proposé par Taylor en 1986 (voir Figure

14). Les 4 matrices de substitutions utilisées pour les variations de taille, charge, polarité et hydrophile sont disponibles en Annexe 1.

b. Conservation

L'analyse de mutants passe également par l'intermédiaire des alignements multiples qui fournissent des informations essentielles sur la conservation des résidus dans la famille protéique concernée. Au regard de l'importance de cette information, la détermination des classes de conservation des colonnes d'un alignement de N séquences se fait en 3 étapes (Figure 53) :

- calcul de la conservation de chaque colonne de l'alignement. Ce calcul est réalisé de la manière suivante : les vecteurs des N résidus présents dans une colonne de l'alignement (les vecteurs de chaque acide aminé sont extraits de la « rosace des acides aminés » (Figure 51) (French and Robson, 1983)) sont additionnés pour obtenir le vecteur représentatif de la colonne, ou vecteur moyen Vm. La norme de Vm est divisée par la somme des normes des vecteurs des résidus de la colonne (Figure 52). Le score S de conservation de la colonne est ensuite normalisé, en le multipliant par le nombre de résidus dans la colonne et en le divisant par N. Le score de conservation S vaut 1 si les vecteurs représentant les résidus de la colonne sont tous identiques et colinéaires. On peut noter qu'au sein de la rosace des acides aminés, seuls les résidus identiques ont des vecteurs colinéaires de même longueur. Si les résidus ne sont pas tous identiques au sein de la colonne, le score S sera donc inférieur à 1.

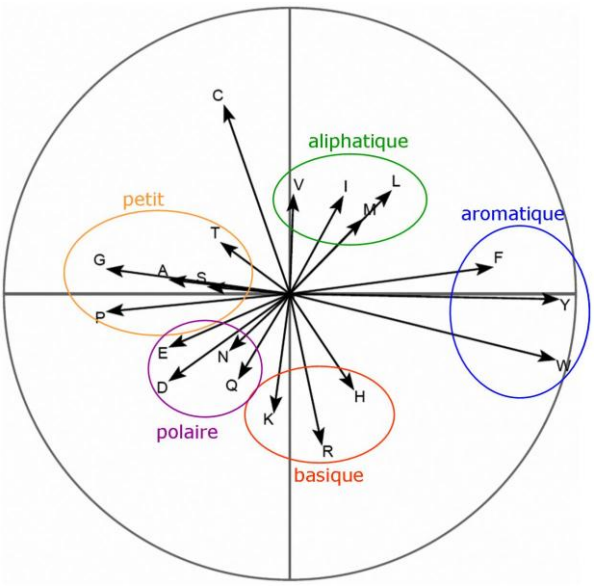


Figure 51. « Rosace des acides aminés ». Cette présentation, en 2D, des caractéristiques évolutives les plus importantes des résidus permet de dépeindre de manière visuelle les relations de similarité entre résidus.

- classification des scores de conservation à l'aide de la méthode DPC (*Density of Points Clustering*) (Wicker et al., 2002) pour la création automatique de groupes de scores ;
- tri des groupes de scores. Ceux qui présentent les valeurs les plus élevées sont représentatifs des colonnes les mieux conservées au sein de l'alignement.

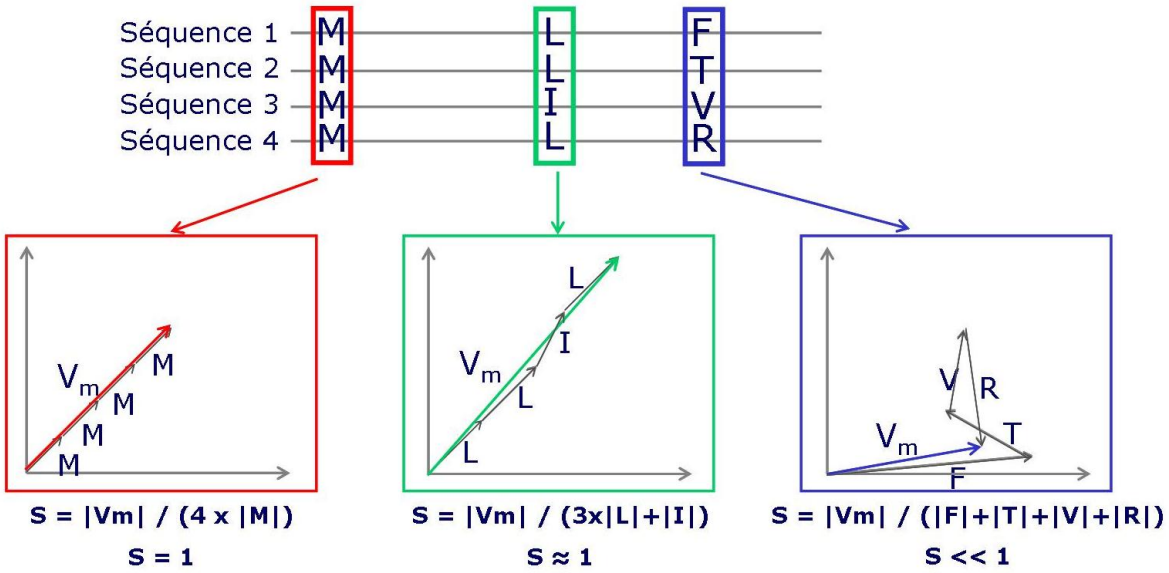


Figure 52. Scores de conservation dans les colonnes de l'alignement, par la méthode de la norme des vecteurs moyens.

Ce processus est répété pour chaque sous-famille du MACS. Ainsi, nous disposons de 4 valeurs pour décrire la conservation en chaque position : 2 valeurs liés à l'ensemble des séquences alignées (rang 1 : strictement conservé, rang 2 : moyennement conservé), une valeur de conservation pour la sous-famille à laquelle appartient la protéine d'intérêt et une valeur pour une position variable.

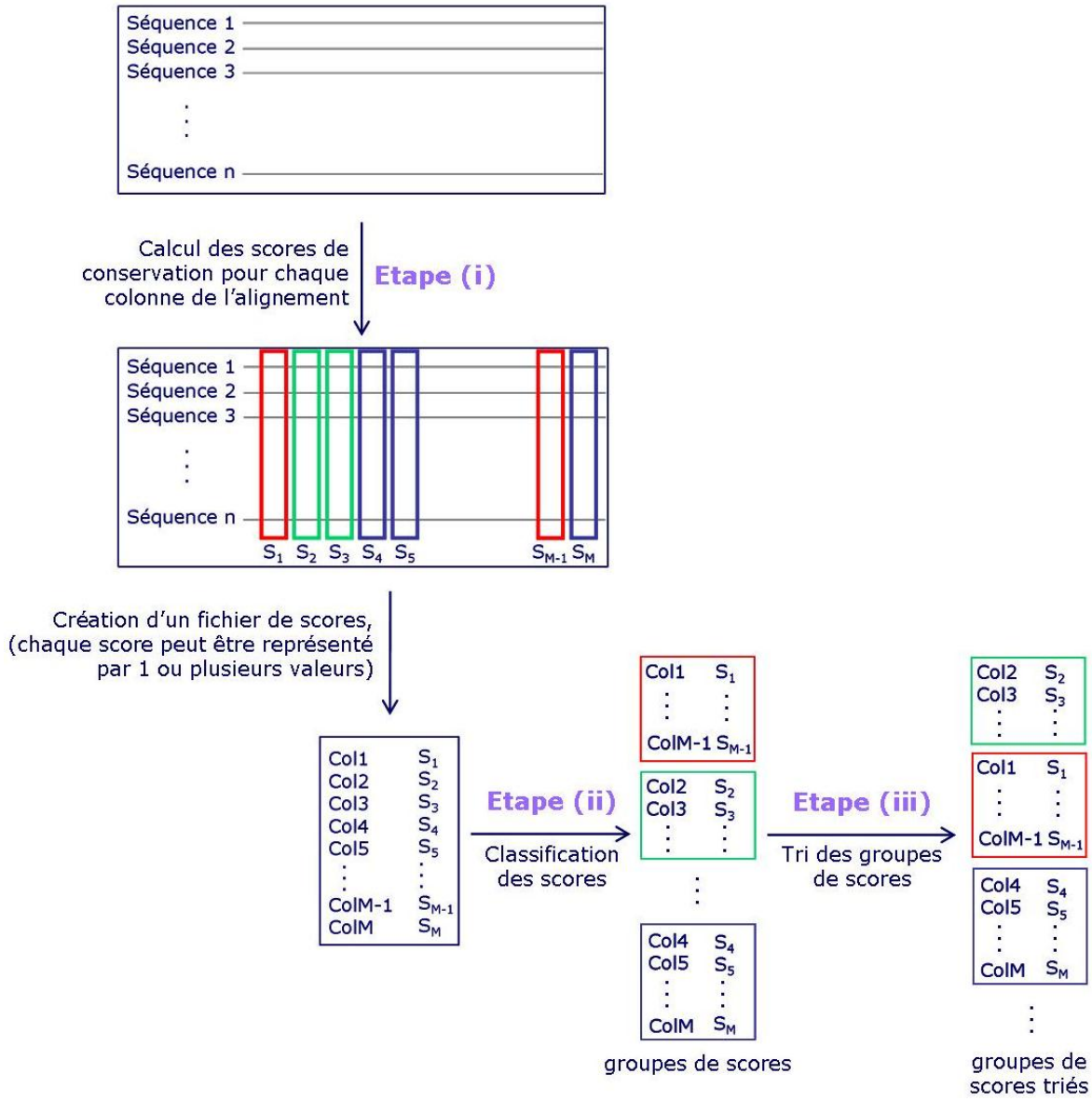


Figure 53. Principales étapes des méthodes de typification des colonnes de conservation. Sur cet exemple, les colonnes les mieux conservées sont les colonnes vertes, puis les rouges, puis les bleues etc.

c. *Localisation*

Les résidus impliqués dans une structure secondaire semblent fournir une bonne indication de la sévérité des mutations. (Adzhubei et al., 2010; Bao et al., 2005; Bromberg and Rost, 2007; Chasman and Adams, 2001; Li et al., 2009; Ramensky et al., 2002). Dans MSV3d, les structures secondaires sont déduites des fichiers PDB en utilisant le programme le plus utilisé, DSSP (*Define Secondary Structure of Proteins*) (Kabsch and Sander, 1983). Cette méthode est fondée sur l'identification des structures secondaires par les liaisons hydrogène formées entre les groupements de la chaîne principale. Elle assigne initialement huit conformations locales : hélice α, feuillet β (brin β), hélice 3-10, hélice π, coude, *bend* (région courbée), β-bridge isolé et *coil*. Nous avons regroupé ces diverses conformations dans les 3 catégories usuelles : hélice α, feuillet β et tout le reste.

Nous localisons aussi les mutations sur les sites « remarquables » annotés par MACSIMS.

d. *Accessibilité*

L'accessibilité au solvant des résidus permet de prédire si un résidu se situe à la surface de la protéine ou non. Les résidus enfouis au cœur de la protéine sont extrêmement importants pour son repliement et sa stabilité. L'accessibilité à la surface est une clé pour comprendre l'impact d'une mutation sur le phénotype (Adzhubei et al., 2010; Bao et al., 2005; Li et al., 2009; Ramensky et al., 2002).

Basée sur les modèles 3D, l'accessibilité au solvant des résidus est calculée par DSSP (Kabsch and Sander, 1983). Nous avons classé les résidus en 3 groupes :

- les résidus enfouis, pour lesquels moins de 5% de la surface de référence est accessible,
- les résidus « de surface » : accessibilité > 30%,
- les résidus « Intermédiaire » : 10% ≤ accessibilité ≤ 30%.

e. Stabilité

Il est intéressant d'analyser la stabilité de la protéine afin de voir si la structure n'est pas perturbée globalement. Jusqu'à présent, les études menées sur les mutations tendent à démontrer que la cause de nombreuses maladies est un changement au niveau de la stabilité de la structure 3D des protéines (Dobson et al., 2006).

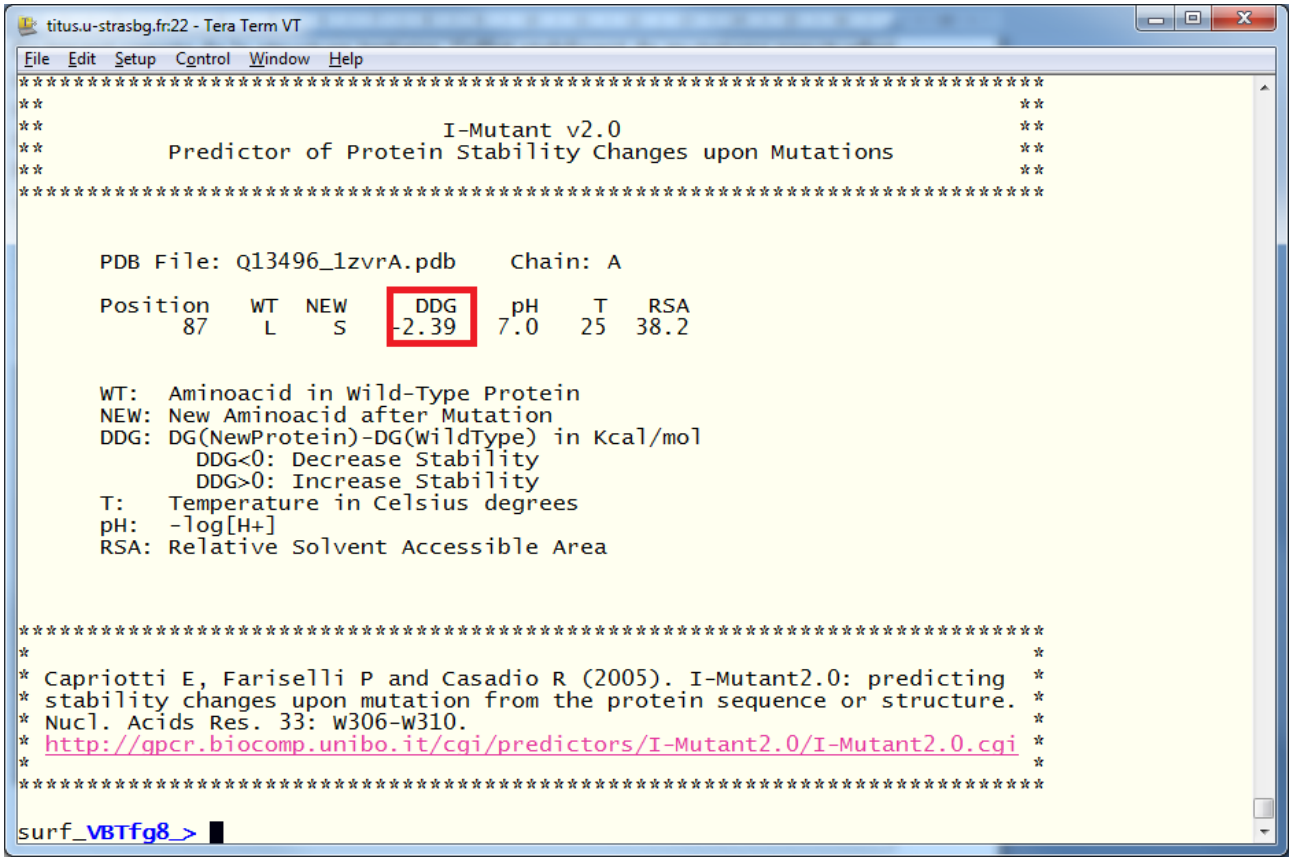


Figure 54. Résultat d'une prédiction de l'I-Mutant2.0 pour la mutation p.Leu87Ser affectant la myotubularine (Q13496).

En tenant compte de la structure tertiaire, l'effet stabilisant ou déstabilisant de mutations ponctuelles est évalué en utilisant le logiciel I-Mutant2.0 (Capriotti et al., 2005) basé sur une approche de type *Support Vector Machine* (SVM) qui estime une différence d'énergie libre

entre le sauvage et le mutant (ΔΔG) ainsi que le signe de cette différence. Une valeur positive ΔΔG implique une augmentation de la stabilité des protéines, tandis qu'une valeur négative ΔΔG suggère une mutation déstabilisante. La Figure 54 est un exemple de la sortie du logiciel I-Mutant2.0.

6.3.2 Entité : spatiale_contact

Les contacts sont calculés avec le logiciel CSU (Contacts of Structural Units) (Sobolev et al., 1999) pour le sauvage et le mutant, les données sont croisées, à l'aide de notre script d'analyse Python, pour connaître les contacts perdus, gagnés et inchangés. CSU a été créé pour calculer les contacts entre les unités structurales de protéines (hélice, feuillet ou résidu). La surface de contact entre 2 atomes A et B est déterminée comme étant la surface de la sphère centrée sur l'atome A et dont le rayon est égal à la somme du rayon de Van der Waals de l'atome A et du rayon d'une molécule de solvant. Cette aire prend en compte les points chevauchant la sphère de Van der Waals de l'atome B. Le logiciel CSU classe les contacts en 4 groupes :

- *hydrophilic - hydrophilic contact*,
- *aromatic - aromatic contact*,
- *hydrophobic - hydrophobic contact*,
- *hydrophobic - hydrophilic contact*.

6.4 Indexation du contenu du MSV3d dans Google

Pour que notre site MSV3D apparaisse lors d'une simple recherche Google, il est nécessaire que nos pages soient indexées par le moteur de recherche. Pour cela nous avons implémenté sur nos machines un robot Googlebot (également désigné par « robot » ou « robot d'indexation » ; « spider » en anglais). Googlebot a pour objectif de parcourir continuellement notre site pour y rechercher les pages nouvelles ou mises à jour de MSV3d et les faire connaître à Google pour qu'elles soient ajoutées dans l'index. Les moteurs de recherche tiennent compte de plusieurs paramètres pour déterminer la pertinence d'une page (*PageRanking*) : mots clés, nombre de citations, importance des sites, etc. Nous avons veillé à intégrer dans nos pages les éléments permettant d'améliorer au mieux notre score. Ainsi, par exemple, lorsqu'un utilisateur recherche l'identifiant rs119489104 d'une mutation, notre site MSV3d apparaît en 4ème et 5ème positions dans les résultats de recherche (Figure 55).

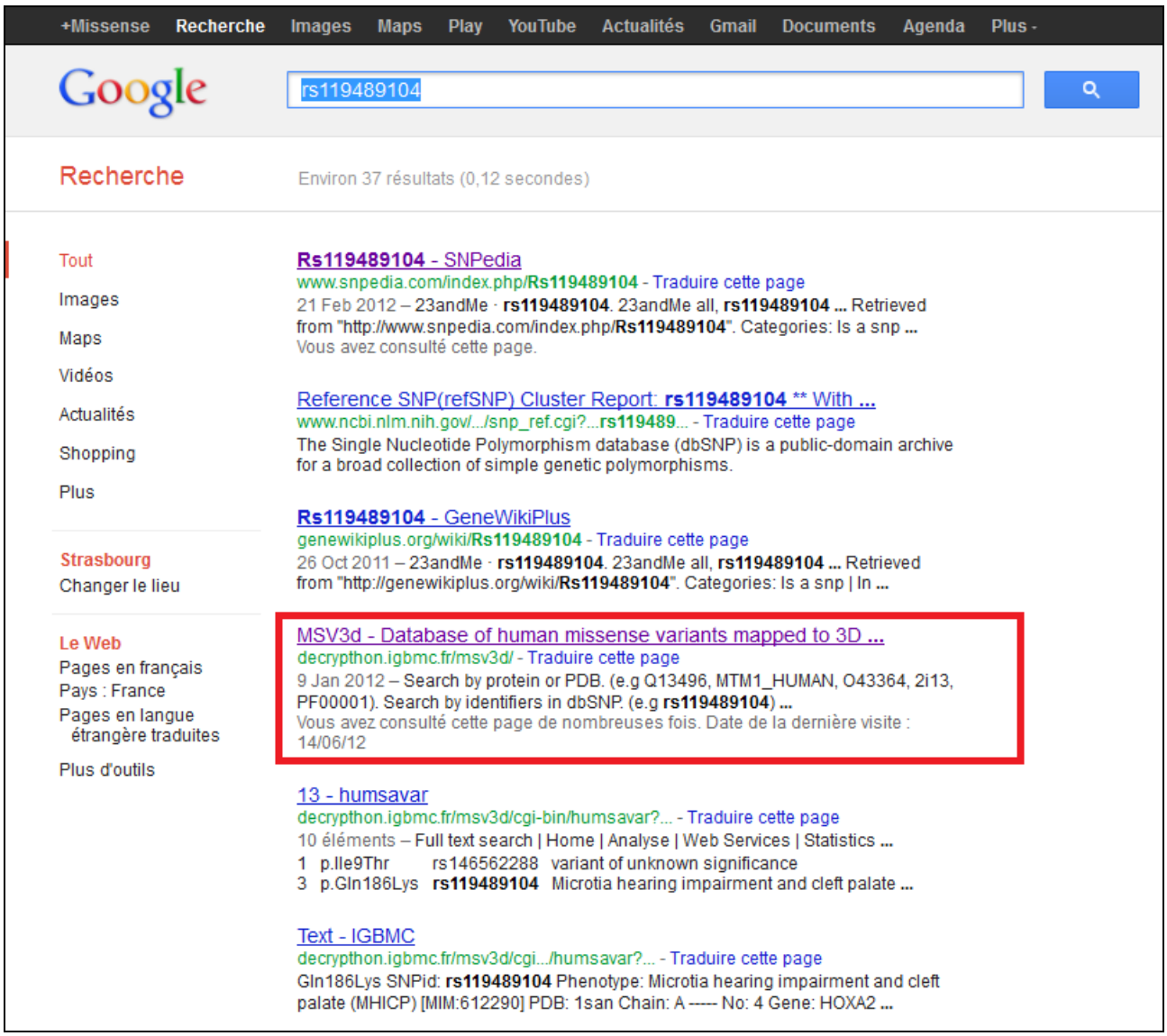


Figure 55. Capture d'écran d'une recherche 'rs119489104' sur Google. Le résultat dans le box rouge est notre site MSV3d.

6.5 Conclusions et perspectives

Par le développement de MSV3d, nous avons mis en place une base de données dédiée à l'analyse globale des conséquences d'une mutation, dans le cadre des maladies génétiques humaines. Le déploiement de cette base de données permet d'accéder, de manière aisée et conviviale, à diverses informations telles que les changements des propriétés physico-chimiques introduits par la mutation d'un résidu particulier, sa conservation, sa présence dans une structure secondaire ou encore son accessibilité relative. MSV3d a été mis en service en février 2012 et sa publication officielle date du mois de mars 2012. Depuis cette date, 2007 requêtes ont été soumises à MSV3d (au 01 juillet 2012), dont la plupart (19,61%) venaient des Etats Unis (Figure 56).

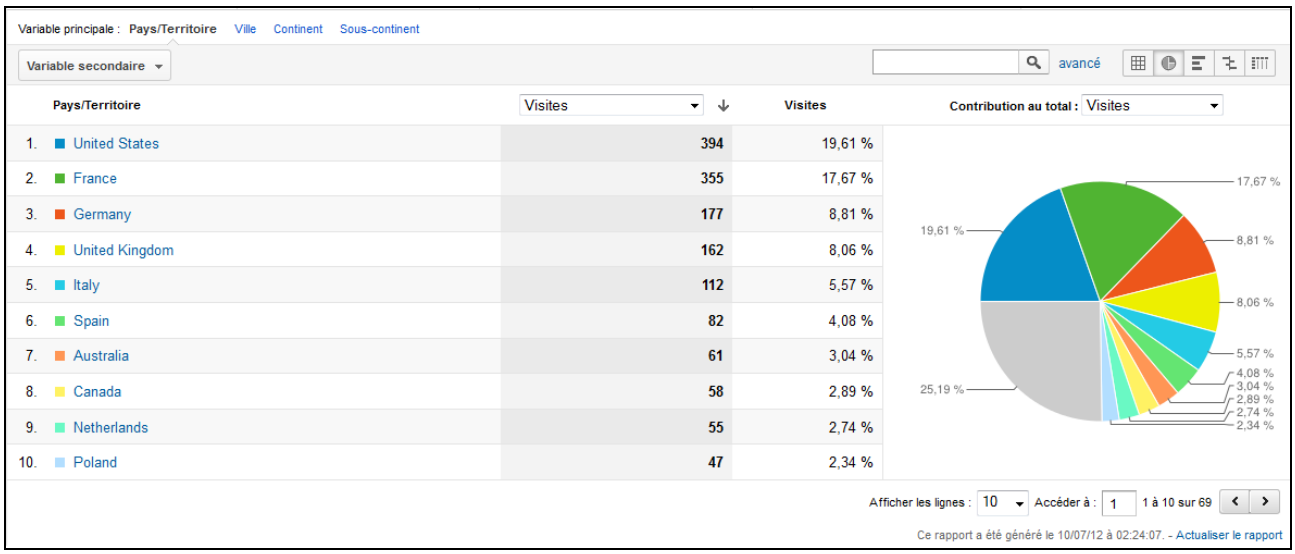


Figure 56. Répartition géographique des visiteurs de MSV3d. Ce rapport a été généré par le *Google Analytics*, un outil professionnel d'analyse d'audience Internet.

Les utilisations et potentialités de MSV3d sont nombreuses. Dès qu'il a été publié dans le journal Database, le consortium *Human Variant Project* nous a invités à participer à la 4ème conférence de ce projet. Après cette conférence, le lien vers MSV3d a été intégré dans le système LOVD de création de bases spécifiques d'un locus (*LSDB*) (Figure 57). Cela permet pour chaque mutation faux-sens décrites dans LOVD d'être annoté par MSV3d.

Figure 57. Lien croisé vers MSV3d intégré dans des systèmes LOVD (dans le rectangle rouge).

Une autre exploitation de MSV3d serait le pré-calcul pour l'ensemble des protéines humaines de toutes les mutations possibles (c'est-à-dire les 19 changements envisageables pour tous les résidus). Un tel pré-calcul, qui concernerait environ 214 242 746 de mutations, est tout à fait concevable dans MSV3d. Malheureusement, un tel projet est difficilement réalisable en l'état et cela, d'une part, à cause du temps d'annotation d'une mutation faux-sens localisée dans un modèle 3D (environ 3 minutes pour générer le modèle 3D muté à partir du modèle 3D sauvage) et d'autre part, au regard des structures disponibles et des moyens de calcul et de stockage à mobiliser. Cependant, au vu des progrès dans le domaine du séquençage personnalisé, on peut raisonnablement penser que l'utilisation de systèmes de type *cloud-computing* et/ou l'implication de grands centres de bioinformatique puisse permettre d'annoter automatiquement chaque nouveau variant identifié dans le génome complet d'un individu.

QUATRIEME PARTIE : DECOUVERTE DE CONNAISSANCES

La quatrième partie détaille les résultats obtenus dans le cadre de l'extraction de connaissances avec 2 chapitres. Le Chapitre 7 regroupe la présentation du système KD4v et la publication réalisée dans le journal *Nucleic Acids Research*. Enfin, le Chapitre 8, qui peut être considéré comme une transition entre les développements réalisés et les perspectives de nos travaux, décrit notre prototype de la priorisation de gènes.

CHAPITRE 7. KD4V : EXTRACTION DE CONNAISSANCES A PARTIR DES MUTATIONS

« You can know the name of a bird in all the languages of the world,

but when you're finished, you'll know absolutely nothing whatever about the bird...

So let's look at the bird and see what it's doing - that's what counts.

I learned very early the difference between knowing the name of something

and knowing something. »

Richard Feynman (1918 - 1988)

7.1 Introduction

En recherche biomédicale, l'un des objectifs est d'obtenir une meilleure compréhension des maladies génétiques humaines afin de pouvoir mettre en œuvre le développement de diagnostics ou de solutions thérapeutiques efficaces. Dans ce cadre, un enjeu primordial est de comprendre et prédire les effets des variations génétiques sur le phénotype d'un individu malade afin de distinguer les mutations délétères, responsables de diverses modifications phénotypiques, des variations neutres qui seront sans conséquence.

A l'heure actuelle, les méthodes de prédiction (voir section 2.4) ne fournissent qu'un score final et, en général, aucune information qui pourrait être utilisée pour évaluer la classification et estimer les relations entre la variation génotypique et la gravité du phénotype observé. Or, il est clair que, pour les biologistes, la compréhension des éléments qui ont abouti à une prédiction et, pour les bioinformaticiens, l'expertise que les biologistes pourraient apporter pour améliorer les prédictions peuvent s'avèrer d'une importance capitale. Pour dépasser ces limites, nous avons développé un système capable de générer sa propre base de connaissances pour caractériser et prédire les mutations neutres ou délétères en s'appuyant sur les données disponibles dans MSV3d et en utilisant des algorithmes de Programmation Logique Inductive (PLI). Le choix de l'approche PLI s'est fondé sur 2 avantages majeurs : i) il peut extraire des connaissances à partir d'un modèle de données complexe constituée de plusieurs tables ou relations; ii) les règles générées sont facilement interprétables par les biologistes ou les cliniciens et peuvent donc guider des experts humains pour la découverte de nouvelles corrélations entre mutation, séquence/structure et gravité du phénotype observé.

La Figure 58 présente les étapes principales établies pour extraire des connaissances à partir des données disponibles dans MSV3d en utilisant comme système PLI, le programme Aleph (voir section 4.6).

Dans un premier temps, des mutations structurales disponibles dans MSV3d et présentes dans le jeu de données de la base PolyPhen-2, sont sélectionnées et séparées en 2 jeux de données constitués de mutations choisies au hasard :

- un jeu d'entraînement contenant 6 000 mutations délétères et 2 000 mutations neutres,
- un jeu de test contenant 658 mutations délétères et 298 mutations neutres. Ce jeu de test a été utilisé pour comparer les performances de prédiction de KD4V à celles de PolyPhen-2 et SIFT, 2 systèmes populaires.

Ensuite, Aleph est utilisé pour induire des règles. Dans cette étape, nous avons utilisé la méthode « k cross validation » qui implique que le jeu d'entraînement est découpé en k parties égales. Aleph prend k-1 parties pour apprendre, le reste est utilisé pour valider. Cette étape est répétée k fois. On utilise « k cross validation » pour optimiser les paramètres d'apprentissage d'Aleph, notamment le paramètre *noise* qui permet de prendre en compte un certain nombre d'exemples négatifs dans la phase d'apprentissage. Les règles finales sont induites à partir du meilleur modèle d'apprentissage (voir Annexe 3). Dans une étape intermédiaire, les utilisateurs peuvent interpréter et sélectionner des règles afin d'intégrer leur expertise dans les itérations successives afin d'aboutir à la base de connaissances finale.

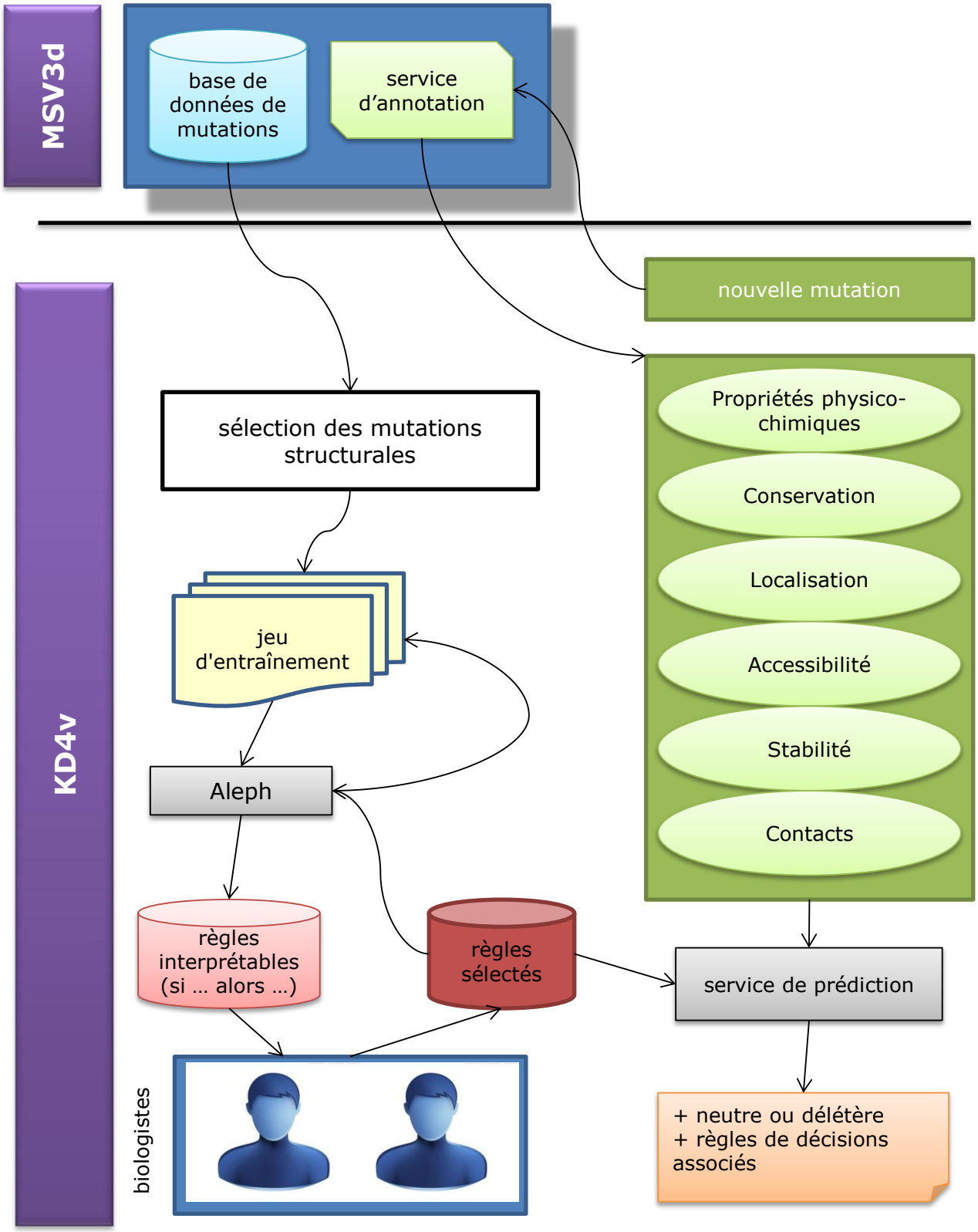


Figure 58. Organigramme pour la fouille de données de MSV3d avec Aleph.

Dans ce contexte, nous avons pu déduire un ensemble de 111 règles caractérisant les mutations neutres ou délétères (voir Annexe 4). Les règles confirment les conclusions antérieures concernant les caractéristiques physico-chimiques et évolutives spécifiques d'une mutation délétère (importance de la conservation, effets néfastes des modifications de charge, de volume ou d'hydrophobie des acides aminés mutés…). En outre, nous avons montré que la

méthode PLI peut être utilisée efficacement pour prédire les effets des mutations, avec des performances similaires aux méthodes les plus largement performantes : SIFT et PolyPhen. Tous ces travaux sont implémentés dans le serveur web « KD4v » (http://decrypthon.igbmc.fr/kd4v/) qui fournit une interface graphique intuitive afin, d'une part, d'accéder et d'interpréter l'ensemble des règles caractérisant des mutants délétères et d'autre part, de prédire si une nouvelle mutation est neutre ou délétère. Si elle est délétère, KD4v fournit les règles qui, au vu de leur format « si conditions x,y,z… alors mutation délétère », peuvent être considérées comme des « explications » interprétables par les utilisateurs humains. Les détails sur la stratégie de découverte des connaissances et sur le service de prédiction sont présentés dans l'article « *KD4v: comprehensible knowledge discovery system for missense variant* » publié dans le journal *Nucleic Acids Research*.

7.2 Publication du système KD4v

Luu TD, Rusu AM, Walter V, Linard B, Poidevin L, Ripp R, Moulinier L, Muller J, Raffelsberger W, Wicker N, Lecompte O, Thompson JD, Poch O, Nguyen NH (2012). *KD4v: comprehensible knowledge discovery system for missense variant.* Nucleic Acids Res. 40, W71-75.

Nucleic Acids Research Advance Access published May 27, 2012

Nucleic Acids Research, 2012, 1–5
doi:10.1093/nar/gks474

KD4v: comprehensible knowledge discovery system for missense variant

Tien-Dao Luu, Alin Rusu, Vincent Walter, Benjamin Linard, Laetitia Poidevin, Raymond Ripp, Luc Moulinier, Jean Muller, Wolfgang Raffelsberger, Nicolas Wicker, Odile Lecompte, Julie D. Thompson, Olivier Poch and Hoan Nguyen*

Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire, 67404 Illkirch, France

Received February 25, 2012; Revised May 4, 2012; Accepted May 6, 2012

ABSTRACT

A major challenge in the post-genomic era is a better understanding of how human genetic alterations involved in disease affect the gene products. The KD4v (Comprehensible Knowledge Discovery System for Missense Variant) server allows to characterize and predict the phenotypic effects (deleterious/neutral) of missense variants. The server provides a set of rules learned by Induction Logic Programming (ILP) on a set of missense variants described by conservation, physico-chemical, functional and 3D structure predicates. These rules are interpretable by non-expert humans and are used to accurately predict the deleterious/neutral status of an unknown mutation. The web server is available at http://decrypthon.igbmc.fr/kd4v.

INTRODUCTION

A wide variety of human diseases have been linked to non-synonymous SNPs (nsSNPs), also called Missense Variants, which result in an alteration of the amino acid sequence of the encoded protein and can affect the function, solubility or structure of the mutated protein. Today, with the huge amount of protein information available in various biomedical databases, it is now possible to better understand the correlation between a nsSNP and the associated phenotypes.

Several methods (1) have been developed to predict the effects of nsSNPs on the 3D structure of a protein and its function, based on the hypothesis that variants that modify the structure/function at the molecular level are more likely to be deleterious. The methods can be divided into two main categories: (i) sequence-based methods using multiple sequence alignments and incorporating different approaches to quantify residue conservation: SIFT (2), PANTHER (3), SNAP(4) and SNP/GO (5) and (ii) methods combining sequence and 3D structure features such as the widely used Polyphen-2 (6), nsSNPAnalyzer (7) and SNPs3D (8). Most of these methods can classify a nsSNP as either deleterious (strong functional effect) or neutral (weak functional effect) with high accuracy. However, they only provide a final score and in general, no information is provided that could be used to evaluate the classification and to estimate the relationships between genotypic and phenotypic variation.

To overcome these limitations, the KD4v (Comprehensible Knowledge Discovery System for Missense Variant) server aims to discover, exploit and provide the user with links between the computed impact of a mutation and the human disease phenotype. We applied the ILP method (9) to a set of nsSNPs involved in human diseases that are mapped to 3D structure and annotated by the MSV3d (MisSense Variant mapped to 3D structure) pipeline (10). KD4v provides two complementary services: (i) a knowledgebase consisting of ILP rules based on 16 sequence/structure/evolution predicates that characterize deleterious mutations in any human gene and that can be interpreted by biologists and (ii) a tool for mutation prediction based on the ILP rules with performances similar to the most widely used methods: PolyPhen-2 and SIFT. In addition, the KD4v server links the human genes to a rich set of up-to-date information encompassing tissue expression, protein-protein interactions or phenotypic descriptions hosted by SM2PH (11).

MATERIALS AND METHODS

Missense variant annotation

The nsSNPs observed in all human proteins were annotated by the MSV3d pipeline, which automatically performs a sequence/structure/evolution analysis and has

*To whom correspondence should be addressed. Tel: +33 3 88 65 32 65; Fax: +33 3 88 65 32 01; Email: nguyen@igbmc.fr

been shown to be robust and efficient (6,12). This includes various parameters which describe, among others, the physico-chemical changes induced by the amino acid substitution, the conservation pattern of the mutated residue, the status of mutated residues with respect to functional features. In KD4v, this multi-level sequence-based characterization of nsSNPs is complemented by parameters related to 3D models or the 3D Fold classification in SCOP (13). This results in pre-computed annotations for over 63 000 known nsSNPs in the 10 713 proteins with known or modelled 3D structures currently available. In addition, the user can also request a prediction for any new or unknown missense variant, if the protein can be mapped to a 3D structure.

The characterization of the background conservation and exploitation of the different types of evolutionary data has been described in detail previously (10). Briefly, we used MACSIMS (14) to annotate a multiple alignment, containing both Uniprot and PDB sequences, with information such as: (i) taxonomic data, (ii) functional descriptions, (iii) known domains or domains similar to a known 3D structure, (iv) potential disordered regions, (v) blocks that do not correspond to disordered regions or known domains but that are conserved at the family or subfamily level and thus may constitute uncharacterized domains and (vi) conservation pattern of domains and residues. If the variant position is mapped to an identified 3D structure, the structural context of each individual mutation is modelled based on several descriptors combining sequence/structure-related data using several software tools such as MODELLER (15), CSU (16), I-Mutant (17). Details of the predicates used in the KD4v server and computational methods/software are provided on the KD4v help page.

Dataset compilation and computer resource

We used the variant set from the Polyphen-2 training set (6) extracted from SwissVar (18) to train and test the KD4v server. Only nsSNPs that are mapped to 3D structures were retained and randomly split into a training set (6000 disease-causing mutations associated with distinct 881 OMIM phenotypes and 2000 neutral polymorphisms) and a first validation set (658 disease-causing variants associated with 311 distinct OMIM phenotypes and 298 neutral polymorphisms). We also created a second validation set (173 disease-causing mutations associated with distinct 39 OMIM phenotypes and 179 neutral polymorphisms), in which not only variants, but also protein sequences, were different in the training and validation sets. Our goal is to predict the deleterious nature of human variants, i.e. those variants associated with disease phenotypes, and it should be noted that these datasets do not specifically identify mutations that have a weaker effect on the function of the protein. The datasets are available for download from our website.

To guarantee a permanent powerful CPU resource for the KD4v server, we deployed the software on the Décrypthon grid (19) including a total of 58 machines and 475 processors under the AIX operating system distributed on six nodes.

Induction logic programming implementation

Induction Logic Programming (ILP) combines Machine Learning and Logic Programming (9). Briefly, given a formal encoding of the background knowledge and a set of examples, an ILP system will derive hypotheses explaining all positive examples and none, or almost none, negative examples. In this approach, logic is used as a language to induce hypotheses from the examples and background knowledge. The result of the learning step is a set of rules represented as logical formula, typically a Prolog program, that can be reused as a prediction service. The creation of the KD4v is based on distinct predicates deduced from the multi-level characterization provided by MSV3d (Supplementary Table S1) and involves various steps detailed in Supplementary Figure S1. We have limited our study to the task of discriminating the mutations linked to human diseases (deleterious) from those associated with the 'polymorphism' term (neutral). Thus, a positive example in Prolog syntax is defined as: 'is_deleterious(m_Q92947.p.Gly390Ala)' which indicates that, in protein Q92947, the replacement of the glycine at position 390 by an alanine is deleterious.

The implementation of the server also includes the optimization of the predicates using a 5-fold cross-validation on the training set with standard performance indicators including sensitivity, specificity, precision, recall, accuracy and F-measure (see legend of Supplementary Table S2 for a complete description). Thus, the final ILP model consists of 16 predicates (Supplementary Table S1) which can be separated into two major types: predicates describing the mutated residue or protein (functional and structural features) and predicates describing the physical, chemical or structural changes introduced by the substitution.

KD4v RULE SERVICE

Currently, the server hosts 111 rules that are comprehensible by humans. These ILP rules can be used, for example, to uncover the relationships between the deleterious effect of a mutation and the multi-class conservation pattern or the type of the physico-chemical alterations (e.g. size, charge and hydrophobicity) introduced by the substitution. Figure 1 shows some induced rules on the web page. To illustrate how to interpret ILP rules, we can consider the humvar398_44 rule:

```
deleterious(A) :-
    modif_charge(A, charge_increase) and
    modif_hydrophobicity(A, hydrophobicity_decrease) and
    secondary_struc(A, helix) and wt_accessibility(A, buried) and
    mut_accessibility(A, buried).
```

This rule states that a mutation A is deleterious if: (i) the charge of the residue is increased by the mutation; (ii) its hydrophobicity is decreased; (iii) the residue is found in a helix; (iv) the wild-type residue is buried; and (v) the mutant residue is also buried. This rule correctly identified 191 (3.18% of the 6000 studied) deleterious mutations, while misclassifying five neutral mutations as

There are total 111 rules.

How to interpret the rules

	Id	If Statement	Then	Coverage Positive	Coverage Negative	Rank
		Enter a key word: Submit				
⊞	humvar398_8	conservation_class(A, global_conservation_rank_1) and freq_at_pos(A, B) and B>=2.	deleterious(A)	475 (7.92%)	2 (0.1%)	1
⊞	humvar398_42	freq_at_pos(A, B) and B>=2 and secondary_struc(A, other) and wt_accessibility(A, buried).	deleterious(A)	397 (6.62%)	2 (0.1%)	2
⊞	humvar398_35	freq_at_pos(A, B) and B>=3 and secondary_struc(A, other).	deleterious(A)	249 (4.15%)	5 (0.25%)	3
⊞	humvar398_12	g_or_p(A, g_or_p_unchanged) and conservation_class(A, global_conservation_rank_1) and secondary_struc(A, other) and wt_accessibility(A, buried).	deleterious(A)	214 (3.57%)	3 (0.15%)	4
⊞	humvar398_37	modif_charge(A, charge_unchanged) and modif_polarity(A, polarity_increase) and conservation_class(A, global_conservation_rank_1).	deleterious(A)	211 (3.52%)	3 (0.15%)	5
⊞	humvar398_78	modif_hydrophobicity(A, hydrophobicity_decrease) and is_in_site(A, yes) and freq_at_pos(A, B) and B>=2 and secondary_struc(A, other).	deleterious(A)	211 (3.52%)	4 (0.2%)	6
⊞	humvar398_50	g_or_p(A, g_or_p_disparition) and freq_at_pos(A, B) and B>=2 and secondary_struc(A, other).	deleterious(A)	208 (3.47%)	5 (0.25%)	7
⊞	humvar398_11	modif_polarity(A, polarity_increase) and conservation_class(A, global_conservation_rank_2) and mut_accessibility(A, buried) and stability(A, decrease).	deleterious(A)	200 (3.33%)	5 (0.25%)	8
⊞	humvar398_55	modif_charge(A, charge_increase) and modif_polarity(A, polarity_increase) and conservation_class(A, global_conservation_rank_1).	deleterious(A)	196 (3.27%)	5 (0.25%)	9
⊞	humvar398_9	conservation_class(A, global_conservation_rank_1) and gain_contact(A, dc) and wt_accessibility(A, buried).	deleterious(A)	194 (3.23%)	4 (0.2%)	10
⊞	humvar398_44	modif_charge(A, charge_increase) and modif_hydrophobicity(A, hydrophobicity_decrease) and secondary_struc(A, helix) and wt_accessibility(A, buried) and mut_accessibility(A, buried).	deleterious(A)	191 (3.18%)	5 (0.25%)	11

Figure 1. ILP rules. The first column provides a link to the positive (deleterious mutations) and negative (neutral mutations) examples covered by a given rule and that can be seen by clicking on the + icon. The second column provides the rule identifier (Id). The next two columns provide the 'if' and 'then' clauses of the induced rules. The two right most columns indicate the number of positive and negative examples covered by the rule in each row.

deleterious (0.25% of the 2000 neutral mutations in the training set).

KD4v PREDICTION SERVICE

Input and output

KD4v provides a service aimed at estimating nsSNP effects based on the ILP rules. It can be accessed via the web interface or via the SOAP Web Service, which can be downloaded from the website. The input form of the web interface (Figure 2a) is supported by Ajax to facilitate the identification of the protein accession number and the location of a mutation on the protein sequence or on the schematic 3D map provided. Given the input data, the MSV3d pipeline generates a multi-level characterization of the variant to be predicted. If a 3D model is available, these values are translated into prolog facts, which then become the input for the prediction service. Thanks to the Prolog engine, the deductive reasoning process immediately derives a conclusion (deleterious or neutral nsSNP) with identified rules. Figure 2b shows the KD4v output for the substitution Gly138Phe in the human peroxisomal biogenesis factor 3, predicted to be deleterious. In the 3D model of this protein, which is involved in the Zellweger syndrome, this residue is buried and located in one of the central helices shaping the protein fold. Analyzing the rule associated with this deleterious prediction, it can be seen that, although this residue is not highly conserved (67% identity which corresponds to the rank2 in our conservation pattern classification), the gain in hydrophobic contact and the decrease in the overall stability might be responsible for the deleterious effect.

Prediction evaluation

We compared the performance of our ILP-based prediction service with two widely used methods: SIFT and PolyPhen-2. The different measures of predictive performance are reported for two independent nsSNP validation sets (Tables 1 and 2). The accuracy (72.28% in Table 1, 75.57% in Table 2) and F-measure (78.61% in Table 1, 71.52% in Table 2) indicate that the KD4v prediction service based on ILP is comparable to SIFT and PolyPhen-2 (although PolyPhen-2 is more accurate on one of the validation sets) and thus represents a competitive alternative solution. Moreover, the KD4v provides ILP rules associated with deleterious predictions that are more interpretable than the previous prediction methods. These rules should help to improve the understanding of

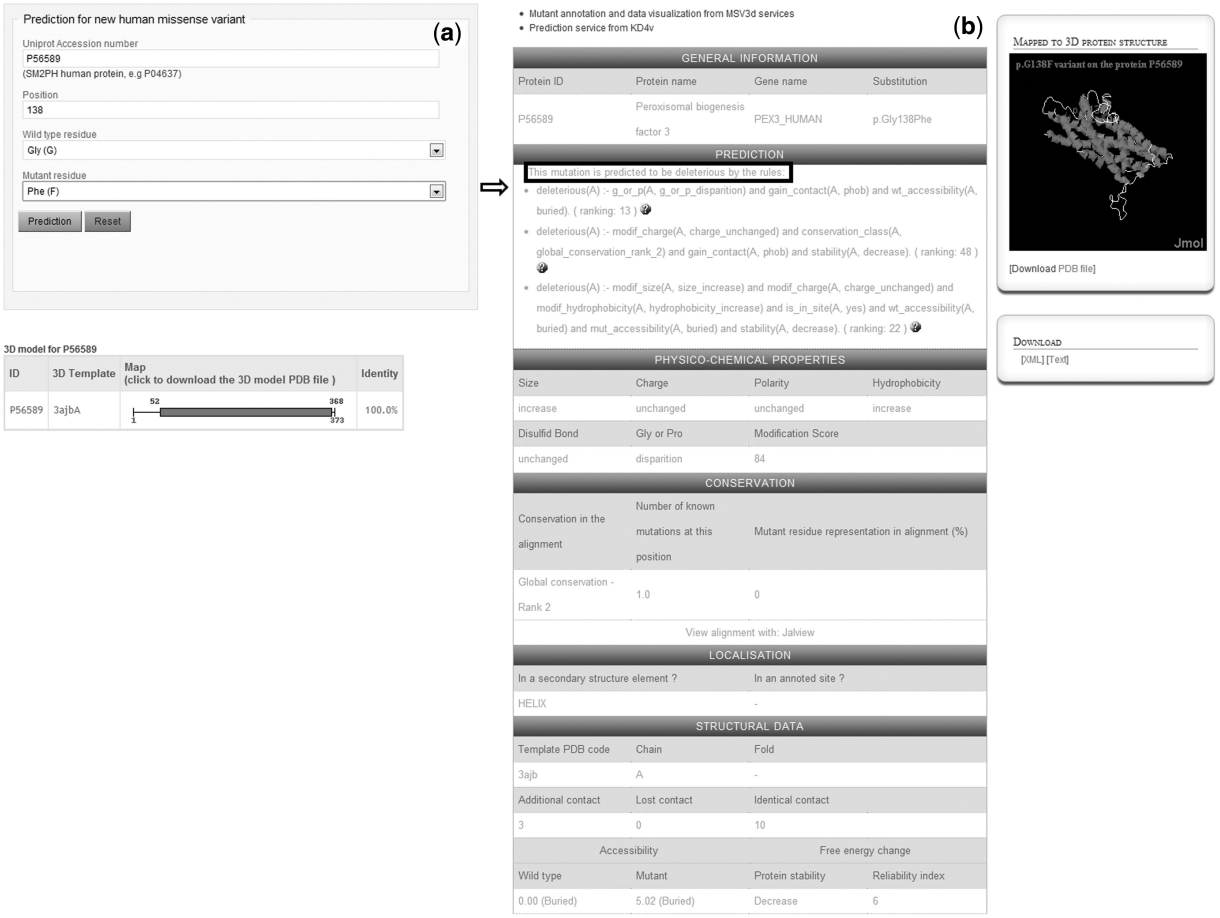


Figure 2. (**a**) Screenshot of the input form of the prediction service. (**b**) Screenshot of the output page providing the prediction results as well as the multi-level characterizations of the mutation. The rules are described if the variant is 'deleterious'. The annotated information related to the mutated position can be visualized in the MSV3d interface on the right.

Table 1. Comparison of prediction methods based on the PolyPhen-2 validation set [658 disease-causing (OMIM phenotype) mutations and 298 neutral polymorphisms]

	TP	FP	FN	TN	Sensitivity	Specificity	Precision	Recall	Accuracy	F-measure
SIFT	398	38	260	260	0.6049	0.8725	0.9128	0.6049	0.6883	0.7276
PolyPhen-2	576	111	77	184	0.8821	0.6237	0.8384	0.8821	0.8017	0.8597
KD4v	487	94	171	204	0.7401	0.6846	0.8382	0.7401	0.7228	0.7861

Table 2. Comparison of prediction methods based on the validation set that excludes proteins present in the training set (173 disease-causing mutations (OMIM phenotype) and 179 neutral polymorphisms)

	TP	FP	FN	TN	Sensitivity	Specificity	Precision	Recall	Accuracy	*F*-measure
SIFT	106	23	67	156	0.6127	0.8715	0.8217	0.6127	0.7443	0.702
PolyPhen-2	139	70	34	109	0.8035	0.6089	0.6651	0.8035	0.7045	0.7278
KD4v	108	21	65	158	0.6243	0.8827	0.8372	0.6243	0.7557	0.7152

the relationships between physico-chemical and structural features and deleterious mutations.

CONCLUSION

The KD4v server uses the available or modelled 3D structures and information provided by the MSV3d pipeline to characterize and predict the phenotypic effect of a mutation. The main advantages of KD4v are (i) valuable predicates and ILP rules associated with the predictions, allowing biologists to identify deleterious mutations and interpret the results, (ii) an ergonomic web interface, incorporating the comprehensive annotation of missense variants, complemented with a SOAP-based remote API

for multiple predictions. Furthermore, the effects of any unknown missense variant (1 of approximately 32 000 000 variants corresponding to all positions of mapped 3D structures and all possible amino acid replacements) can be predicted upon request by the user. In the future, we will extend the background knowledge, first by adding structural surface topology descriptions (20) of the proteins, allowing the precise mapping of different functional regions such as the protein core and the non-interacting or interacting surfaces, and second, by integrating useful knowledge about the functional impact of missense variants from the SNPdbe database (21). Finally, we intend to enhance the prediction performance by combining ILP with other machine learning methods.

SUPPLEMENTARY DATA

Supplementary Data are available at NAR Online: Supplementary Tables 1 and 2, Supplementary Figure 1.

ACKNOWLEDGEMENTS

The IGBMC common services and BIPS platforms are acknowledged for their assistance.

FUNDING

The work was performed within the framework of the Decrypthon program, co-funded by Association Française contre les Myopathies [AFM, 14390-15392]; IBM and Centre National de la Recherche Scientifique (CNRS); ANR [EvolHHuPro: BLAN07-1-198915 and Puzzle-Fit: 09-PIRI-0018-02]; Institute funds from the CNRS, INSERM, the Université de Strasbourg and the Vietnam Ministry of Education and Training (CT 322). Funding for Open access charge: ANR-10-BINF-03-02.

Conflict of interest statement. None declared.

REFERENCES

1. Thusberg,J., Olatubosun,A. and Vihinen,M. (2011) Performance of mutation pathogenicity prediction methods on missense variants. *Hum. Mutat.*, **32**, 358–368.
2. Ng,P.C. and Henikoff,S. (2003) SIFT: predicting amino acid changes that affect protein function. *Nucleic Acids Res.*, **31**, 3812–3814.
3. Thomas,P.D., Campbell,M.J., Kejariwal,A., Mi,H., Karlak,B., Daverman,R., Diemer,K., Muruganujan,A. and Narechania,A. (2003) PANTHER: a library of protein families and subfamilies indexed by function. *Genome Res.*, **13**, 2129–2141.
4. Bromberg,Y., Yachdav,G. and Rost,B. (2008) SNAP predicts effect of mutations on protein function. *Bioinformatics*, **24**, 2397–2398.
5. Calabrese,R., Capriotti,E., Fariselli,P., Martelli,P.L. and Casadio,R. (2009) Functional annotations improve the predictive score of human disease-related mutations in proteins. *Hum. Mutat.*, **30**, 1237–1244.
6. Adzhubei,I.A., Schmidt,S., Peshkin,L., Ramensky,V.E., Gerasimova,A., Bork,P., Kondrashov,A.S. and Sunyaev,S.R. (2010) A method and server for predicting damaging missense mutations. *Nat. Methods*, **7**, 248–249.
7. Bao,L., Zhou,M. and Cui,Y. (2005) nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. *Nucleic Acids Res.*, **33**, W480–W482.
8. Yue,P., Melamud,E. and Moult,J. (2006) SNPs3D: candidate gene and SNP selection for association studies. *BMC Bioinformatics*, **7**, 166.
9. Muggleton,S. (1991) Inductive logic programming. *N. Gen Comput*, **8**, 295–318.
10. Luu,T.D., Rusu,A.M., Walter,V., Ripp,R., Moulinier,L., Muller,J., Toursel,T., Thompson,J.D., Poch,O. and Nguyen,H. (2012) MSV3d: database of human MisSense variants mapped to 3D protein structure. *Database J. Biol. Databases Curation*, **2012**, bas018.
11. Friedrich,A., Garnier,N., Gagniere,N., Nguyen,H., Albou,L.P., Biancalana,V., Bettler,E., Deleage,G., Lecompte,O., Muller,J. *et al.* (2010) SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases. *Hum. Mutat.*, **31**, 127–135.
12. 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.
13. Andreeva,A., Howorth,D., Chandonia,J.M., Brenner,S.E., Hubbard,T.J., Chothia,C. and Murzin,A.G. (2008) Data growth and its impact on the SCOP database: new developments. *Nucleic Acids Res.*, **36**, D419–D425.
14. 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.
15. Eswar,N., Eramian,D., Webb,B., Shen,M.Y. and Sali,A. (2008) Protein structure modeling with MODELLER. *Methods Mol. Biol.*, **426**, 145–159.
16. Sobolev,V., Sorokine,A., Prilusky,J., Abola,E.E. and Edelman,M. (1999) Automated analysis of interatomic contacts in proteins. *Bioinformatics*, **15**, 327–332.
17. Capriotti,E., Fariselli,P. and Casadio,R. (2005) I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. *Nucleic Acids Res.*, **33**, W306–W310.
18. Mottaz,A., David,F.P., Veuthey,A.L. and Yip,Y.L. (2010) Easy retrieval of single amino-acid polymorphisms and phenotype information using SwissVar. *Bioinformatics*, **26**, 851–852.
19. Bard,N., Bolze,R., Caron,E., Desprez,F., Heymann,M., Friedrich,A., Moulinier,L., Nguyen,N.H., Poch,O. and Toursel,T. (2010) Decrypthon grid - grid resources dedicated to neuromuscular disorders. *Stud. Health Technol. Informat.*, **159**, 124–133.
20. Albou,L.P., Poch,O. and Moras,D. (2011) M-ORBIS: mapping of molecular binding sites and surfaces. *Nucleic Acids Res.*, **39**, 30–43.
21. Schaefer,C., Meier,A., Rost,B. and Bromberg,Y. (2012) SNPdbe: constructing an nsSNP functional impacts database. *Bioinformatics*, **28**, 601–602.

7.3 Evolution de KD4v

KD4v est toujours en cours d'évolution. 2 axes de recherche sont explorés pour :

- Introduire de nouveaux paramètres dans le modèle qui permettent de mieux discriminer, interpréter et prédire le caractère délétère/neutre d'une mutation.
- Combiner l'information sémantique de PLI avec les avantages des performances des approches SVM également pour augmenter les performances de prédiction.

7.3.1 Nouveaux paramètres plus discriminants

Le choix des nouveaux paramètres s'inscrit dans notre volonté d'introduire des informations supplémentaires qui ne soient pas directement liées à la position mutée et à son environnement spatial proche. Dans cette optique, nous avons déjà introduit dans KD4V, la notion de repliement (*fold*), telle qu'elle est définie dans la banque SCOP. Cette information qui prend en compte le mode d'organisation structurale de l'ensemble d'un domaine ou d'une protéine a réduit le nombre de règles (de 123 à 111) et a permis de vérifier la notion que tous les repliements ne présentaient pas la même susceptibilité aux mutations (Luu et al., 2012).

Actuellement, nous abordons la relation entre le type de mutation (délétère/neutre) et les fonctions des protéines en introduisant la fonction LGO. Cette fonction correspond au score log-odds calculé à partir d'une classification des termes de Gene Ontology (GO). Elle a été précédemment utilisée en combiné avec des données de séquences pour prédire les mutations faux-sens associées aux cancers (Kaminker et al., 2007). Récemment, ce score a été étendu et utilisé avec des données de structure pour distinguer les mutations délétères et neutres (Capriotti and Altman, 2011). Pour chaque terme de GO, la fréquence des mutations délétères ($f_{GO}(D)$) a été comparée à celle de mutations neutres ($f_{GO}(N)$) et la valeur LGO pour chaque protéine a été calculée comme suit :

$$LGO=\Sigma_{GO} \log2(f_{GO}(D)/f_{GO}(N))$$

Pour tester l'impact du score LGO dans notre système de prédiction basé sur la méthode PLI, 2 séries de tests ont été effectuées en ajoutant ou n'ajoutant pas le score LGO aux paramètres de mutations utilisés dans KD4v. LGO a réduit le nombre de règles (de 111 à 98).

	Sensitivity	**Specificity**	**Precision**	**Recall**	**Accuracy**	**F-measure**
KD4v	0.7401	0.6846	0.8382	0.7401	0.7228	0.7861
KD4v+LGO	0.8587	0.8209	0.9142	0.8587	0.847	0.8597

Tableau 8. Statistiques de prédiction avec ou sans LGO sur le jeu de test de KD4v (658 mutations délétères et 298 mutations neutres).

En première approche, au regard des résultats présentés dans le Tableau 8, on constate que LGO est un paramètre remarquable pour discriminer mutations délétères et neutres. Une analyse plus approfondie de l'importance du paramètre LGO, prenant en compte toutes les mutations (avec ou sans structure présentes dans MSV3d, est à effectuer pour valider ces observations.

7.3.2 Prédiction par la méthode hybride SVILP

Initialement, la méthode *Support Vector Inductive Logic Programming* (SVILP) a été introduite pour combiner la capacité de description des règles de PLI avec la puissance statistique de l'approche SVM (Stephen et al., 2005) et l'appliquer pour étudier la toxicité d'une molécule.

Dans notre problématique de prédiction de l'impact d'une mutation sur le phénotype des pathologies humaines, nous avons abordé l'approche SVILP en utilisant, comme précédemment dans KD4V, Aleph pour induire des règles à partir de 6 000 mutations délétères et 2 000 mutations neutres. Puis, l'ensemble des 111 règles (R1 à R111 dans la Figure 59) a été utilisé pour caractériser chaque mutation (M1 à M8000) attachée à son étiquette Y (+1 pour délétère et 2 pour neutre). Cela est comparable à décrire chaque mutation selon un vecteur binaire où la position ième vaut 1 si la règle i s'applique à l'exemple et 0 dans le cas contraire (Figure 59). Nous avons ensuite utilisé cette matrice comme entrée du modèle SVM en utilisant le *package* MSVMPack (Lauer and Guermeur, 2011) développé au LORIA (Laboratoire Lorrain de Recherche en Informatique). Les résultats ont été rassemblés dans le Tableau 9.

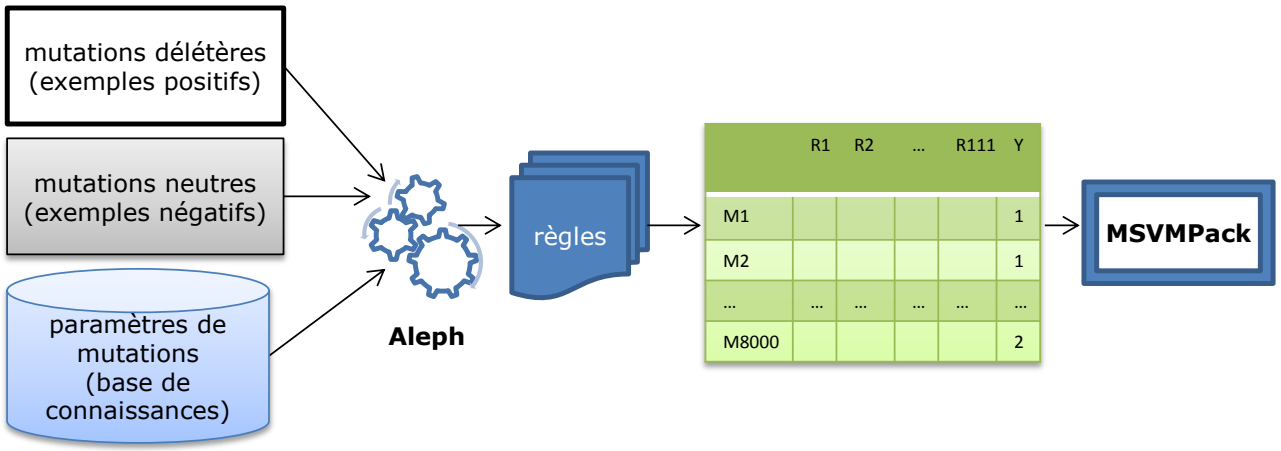


Figure 59. Méthode SVILP mise en œuvre pour l'étude du lien génotype/phénotype.

Les premiers résultats obtenus indiquent que la performance de prédiction n'est pas significativement améliorée par l'approche SVILP. En analysant plus en détail notre matrice, on se rend compte qu'elle est très asymétrique avec des lignes ne présentant que des zéro (vecteurs nuls caractéristiques des mutations neutres) et des lignes présentant des distributions très hétérogènes de 1 et de 0 (mutations délétères).

	Sensitivity	Specificity	Precision	Recall	Accuracy	F-measure
KD4v	0.7401	0.6846	0.8382	0.7401	0.7228	0.7861
SVILP	0.7249	0.6959	0.8413	0.7249	0.7159	0.7788

Tableau 9. Comparaison des méthodes de prédiction basée sur jeu de test de KD4v (658 mutations délétères et 298 mutations neutres).

7.4 Conclusions et perspectives

KD4V est, à notre connaissance, le premier système qui fournit des règles interprétables pour caractériser et prédire les mutations neutres ou délétères. Grâce à ces règles, le service de prédiction est capable de procurer non seulement une prédiction de qualité comparable aux meilleurs systèmes disponibles, mais aussi des corrélations entre l'impact sur le phénotype

d'une mutation et les paramètres attachés. Ces éléments ont permis de publier notre travail dans une revue (*Nucleic Acids Research*) ayant un bon *Impact Factor* et cela, malgré une pléthore d'algorithmes et de systèmes dédiés à la prédiction du caractère neutre ou délétère d'une mutation.

A l'avenir, nous comptons mettre à profit le fait que KD4v est une plateforme ouverte et modulable afin d'intégrer de nouvelles techniques efficaces de fouille ou d'analyse de données proposées par la communauté scientifique. L'objectif étant d'orienter KD4v vers un outil d'aide à la décision pour toutes maladies humaines impliquant une ou des mutations génétiques. Un tel objectif implique de nombreuses évolutions et notamment, d'étendre le champ d'analyse de KD4v à l'ensemble des mutations connues dans toutes les protéines humaines.

En effet, la version actuelle de KD4v prédit seulement les mutations qui se trouvent dans un modèle 3D, ce qui implique de disposer d'une structure de la protéine d'intérêt ou d'un de ses domaines ou, *a minima*, que l'on calcule un modèle structural sur la base de la structure d'une protéine proche. Ceci limite grandement KD4v car, à ce jour encore, peu de structures sont disponibles au regard des séquences protéiques disponibles. Dans ce cadre, nous sommes actuellement en train de développer et valider un système de prédiction des mutations présentes dans des régions sans donnée de structure disponible. Les résultats préliminaires obtenus par le stagiaire ingénieur, Alin-Mihai Rusu, sont très encourageants.

Comme nous l'avons vu dans le CHAPITRE 7, une autre approche envisageable pour étendre le champ d'action de KD4v pourrait consister à pré-calculer, dans MSV3d, les 33 paramètres de conservation et structuraux pour l'ensemble de 19 mutations possibles de chaque acide aminé présent dans une protéine ou un domaine d'intérêt. Ce pré-calcul permettrait à KD4V de prédire automatiquement toutes les mutations possibles des protéines humaines, et notamment les nouveaux et nombreux variants qui sont mis en évidence dans les projets de séquençage du génome complet d'un individu (Projet « 1000 *Genomes* » (http://www.1000genomes.org/). Cela permettrait d'aborder l'étude des mutations délétères d'un gène donné et les phénotypes observés chez un patient dans un contexte prenant en compte les SNPs présents dans l'ensemble des gènes interagissant (ou impliqués dans les voies) avec le gène responsable de la maladie.

Enfin, une évolution essentielle de KD4v va nécessiter d'échapper à la dichotomie : sévérité neutre/délétère d'une mutation, pour aborder la caractérisation et la prédiction d'effets phénotypiques plus complexes reflétant mieux la diversité des phénotypes des maladies génétiques humaines. Pour ce faire, en collaboration avec des chercheurs, médecins et cliniciens experts, nous comptons mettre à profit notre infrastructure SM2PH Central pour développer des Instances dédiées à des types particuliers de maladies humaines, telles que SM2PH-Ciliopathy ou SM2PH-AMD-kb. Dans le cadre plus restreint de maladies particulières, nous pourrons accéder, via MSV3d, à des banques *locus-specific* (LSDB) et exploiter les informations détaillées et de qualité disponibles. Ces descriptions plus complexes des maladies nous permettront non seulement, de construire des modèles de caractérisation et de prédiction pour plus de 2 degrés de sévérité mais également, à plus long terme, de comparer les modèles obtenus pour différentes maladies afin d'aborder l'étude des processus et réseaux biologiques dont la perturbation entraine des phénotypes similaires ou distincts entre les divers types de pathologies humaines.

CHAPITRE 8. VERS UNE PRIORISATION DES GENES

8.1 Introduction

La richesse des données intégrées dans SM2PH Central nous a amené à réfléchir à un nouveau prototype de priorisation de gènes basé sur la fusion multi-source de données incluant la séquence, l'évolution, la structure protéine-interaction, réseaux biologiques, etc. L'objectif principal de la priorisation des gènes est d'identifier les meilleurs candidats susceptibles d'être impliqués, par exemple, dans un processus biologique ou une pathologie. La priorisation « manuelle » est souvent basée sur l'intuition et l'expérience ce qui peut introduire des biais. De plus, c'est une tâche qui devient impossible avec le volume actuel de données disponibles, d'où la nécessité de l'intégrer dans un processus automatisé pour une approche à haut-débit. De façon synthétique, la priorisation des gènes permet de générer une liste de gènes « triée » selon leurs probabilités d'implication dans un processus en se basant sur la similarité avec des gènes déjà connus et validés pour leur implication dans le processus étudié.

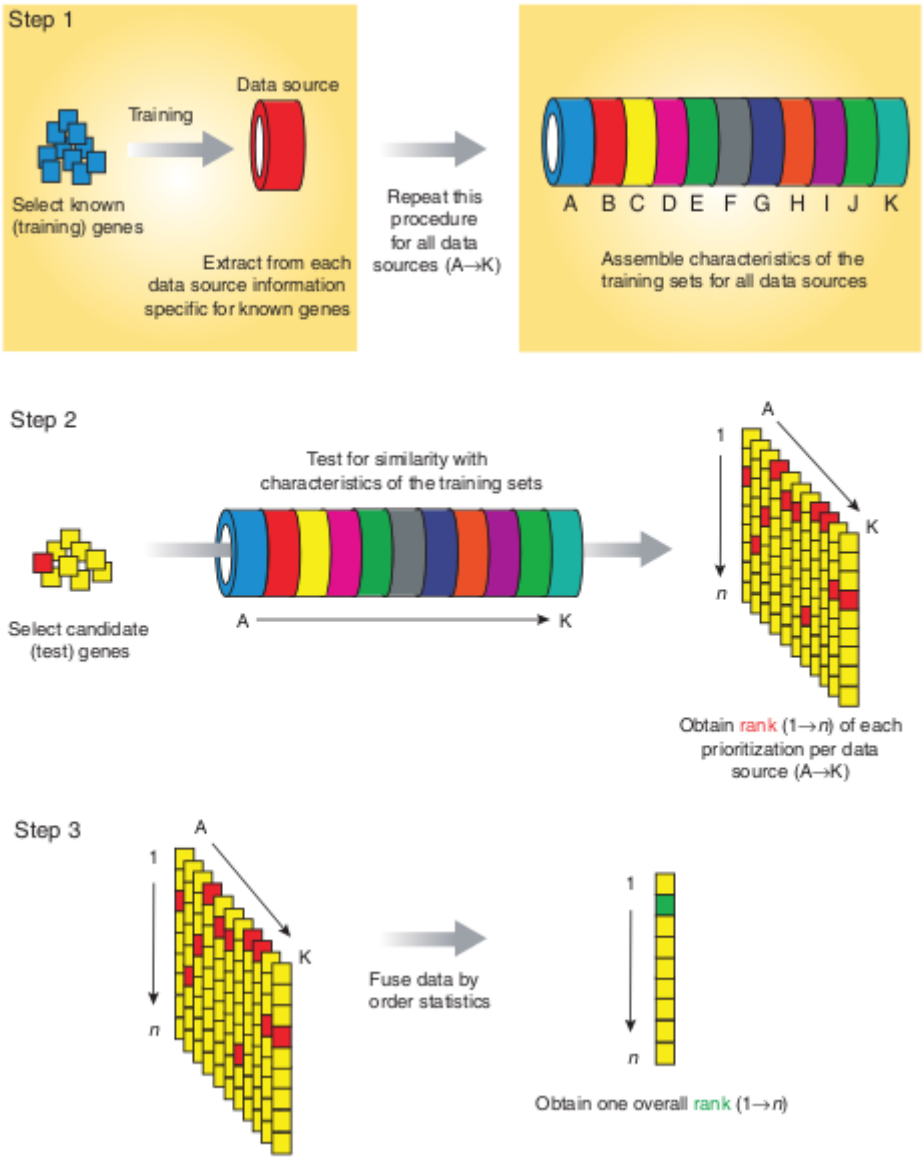


Figure 60. Description du principe global de la priorisation de gènes. Issue de: (Aerts et al., 2006b).

Schématiquement, la priorisation se décompose en 3 étapes (Figure 60) : i) la caractérisation des gènes du set d'entrainement qui consiste à établir le profil de chacun des gènes et éventuellement un profil « type » ii) la caractérisation de gènes candidats qui consiste à établir un profil pour chacun des gènes et à attribuer un score qui évalue la similarité entre le gène candidat et les gènes du set d'entraînement ou le profil « type », selon la méthode d'évaluation utilisée et iii) le classement des gènes candidats en fonction du score obtenu. Une fois que les différents classements locaux sont disponibles, une étape de fusion est effectuée pour obtenir un classement global.

Le Tableau 10 permet d'avoir un aperçu des différents outils de priorisation existants et des critères d'évaluation qu'ils utilisent. Certains de ces critères sont récurrents et ont également été intégrés dans l'outil de priorisation que nous avons développé (annotation fonctionnelle, interaction protéine-protéine, etc.).

Considérer un maximum de source de données est essentiel car cela permet d'accroître la robustesse de l'outil (Aerts et al., 2006b). En effet, il a été montré qu'observer un nombre élevé de paramètres pour la comparaison des gènes connus et des gènes candidats permet de diminuer le bruit produit pour certains gènes par certaines sources de données (par exemple, le bruit produit par une « surreprésentation » dans la littérature des gènes très étudiés ou impliqués dans un processus très étudié et/ou très référencé, l'exemple type étant les cancers). Les outils fondateurs de la priorisation de gènes figurant parmi les plus aboutis, à savoir Endeavour (Aerts et al., 2006b; Tranchevent et al., 2008) et ToppGene (Chen et al., 2009) utilisent plus de source de données que la moyenne (Tableau 10).

Le Tableau 11 permet de mettre en évidence un certain nombre d'avantages et d'inconvénients des différents outils (format, pertinence et nombre de résultats, etc.) qui ont été comparés en utilisant le jeu des 41 gènes définis dans le cadre de l'étude d'association pangénomique (GWAS) réalisée sur la Dégénérescence Maculaire Liée à l'Âge (voir section 8.2).

Tools	Functional annotations	Expression	Regulatory information	Text (co-citation)	Text (functional)	Interactions	Pathways	Sequence	Phenotype	Conservation/homology	Disease probabilities	Chemicals	
aGeneApart				X									1
Bitola				X									1
Caesar	X	X		X	X	X			X	X			7
Candid		X		X	X		X		X	X			6
CGI		X			X								2
DGP							X		X				2
DIR		X			X	X							3
DomainRBF							X						1
Endeavour	X	X	X		X	X	X	X		X			8
G2D	X			X		X		X					4
GeneDistiller	X	X		X		X	X		X				6
GeneProspector					X								1
GeneRank	X	X											2
GeneRanker	X			X	X	X			X				5
GeneSeeker		X			X				X	X			4
GeneWanderer	X	X		X		X	X						5
Génie					X					X			2
GenTrepid					X	X	X	X					4
MetaRanker						X					X		2
MimMiner	X				X	X		X	X				5
PGMapper					X								1
PhenoPred	X					X		X	X				4
Pinta	X	X		X		X	X						5
Pocus	X	X											2
PolySearch	X				X	X	X					X	5
PosMed	X			X	X	X			X			X	6
Prioritizer	X	X				X	X						4
S2G	X		X			X	X	X	X	X			7
SNPs3D	X				X	X	X	X	X				6
Suspects	X	X						X					3
TargetMine	X		X				X		X			X	5
Tom	X	X											2
ToppGene	X	X		X		X	X	X	X				7
Total	20	15	3	7	15	20	13	12	10	6	4	3	

Tableau 10. Présentation de différentes sources de données pris en compte dans la priorisation des gènes

OUTIL	FONCTIONNEL	VITESSE	ENTREE						SORTIE			INCONVENIENTS	AVANTAGES	REFERENCE
			Gènes connus	Mots-clés	Données d'expression	Région	Génome	DEG	Liste priorisée	Sélection de candidats	Test statistiques			
CANDID	X	+++			X		X		X		X	DMLA : gènes connus mal classés	Sortie facilement exploitable Poids des sources de données	Hutz et al., 2008
Endeavour	X	+++	X	X	X	X	X	X	X		X		Nombreux critères d'évaluation Nombreuses sources de données	Aerts et al. 2006 Tranchevent et al. 2008
G2D			X	X		X			X		X	Plus maintenu		Perez-Iratxeta et al. 2005, 2007
GeneDistiller	X	+	X	X		X			X			Un chromosome à la fois (chr) Information noyées dans tous les gènes du chr. Poids des sources de données	Code du programme Open Source	Seelow et al. 2008
GeneRanker			X	X			X		X	X	X	Plus maintenu		Gonzalez et al., 2008
Genie	X		X	X		X	X		X	X	X	DMLA : que 4 gènes candidats reconnus Ne retourne des résultats que pour 4 gènes candidats	Dictionnaire des termes discriminants	Fontaine et al. 2011
IDGP	X	+++	X	X			X		X		X	Un gène à la fois		Lopez-Bigas et al., 2007
MetaRanker	X	+	X	X		X	X		X		X	Phénotype parmi une liste prédéfinie		Pers et al. 2011
PhenoPred	X	+++	X				X		X			DMLA : que 2 gènes retenus		Redivojac et al. 2008
PosMed	X	+++	X			X	X		X	X	X	DMLA : gènes connus mal classés	Principalement basé sur la littérature	Yoshida et al. 2009
SNPs3D	X	+		X			X		X			Pas de filtrage des mots-clés : >50% sont des synonymes du mot-clé (nécessite un thésaurus)	Dictionnaire des termes discriminants DMLA, nom complet : que 3 gènes retenus AMD : nombreux gènes reconnus, pas tous les gènes connus	Yue et al. 2006
TargetMine	X	+++	X					X		X		Pas de ranking	ND	Chen et al. 2011
ToppGene	X	+++	X					X	X		X	Plusieurs sources retournent un score ND ou nul pour la plupart des gènes.	Nombreux critères d'évaluation Nombreuses sources de données	Chen et al. 2009
GPSy	X	++	X	X				X	X		X	Phénotype parmi une liste prédéfinie	Plus de 30 organismes disponibles	Britto et al. 2012

Tableau 11. Tableau comparatif de différents outils de priorisation de gènes dans le cas de l'étude de la Dégénérescence Maculaire Liée à l'Âge (DMLA).

Dans le cadre du développement de notre prototype, notre premier objectif a été d'intégrer de nouveaux concepts dans les procédures de priorisation de gènes (contexte génomique, données évolutives étendues, données cliniques etc.). Notre approche recouvre un processus itératif piloté par les connaissances elles-mêmes incluant de nombreuses étapes de génération, épuration, validation, comparaison, analyse et représentation des données aboutissant à une nouvelle connaissance susceptible de relancer l'ensemble du processus. Ce travail préliminaire a été réalisé en combinant SM2PH Central avec des bases de données développées au sein de l'équipe LBGI telles que, EvoluCode (voir section 3.7) ou le prototype GecoDB qui fournit des informations relatives au contexte génomique d'un gène, d'un SNP, des exons codants ou non codants, ou d'une région génomique. Le fait que nos sources de données soient mises à jour régulièrement et qu'elles soient synchronisées confère un avantage certain par rapport à plusieurs des outils testés qui ne sont plus mis à jour ou pas maintenus, parfois depuis plusieurs années, tel que SNPs3D, GeneSeeker ou CANDID.

8.2 Conception de notre système de priorisation de gènes

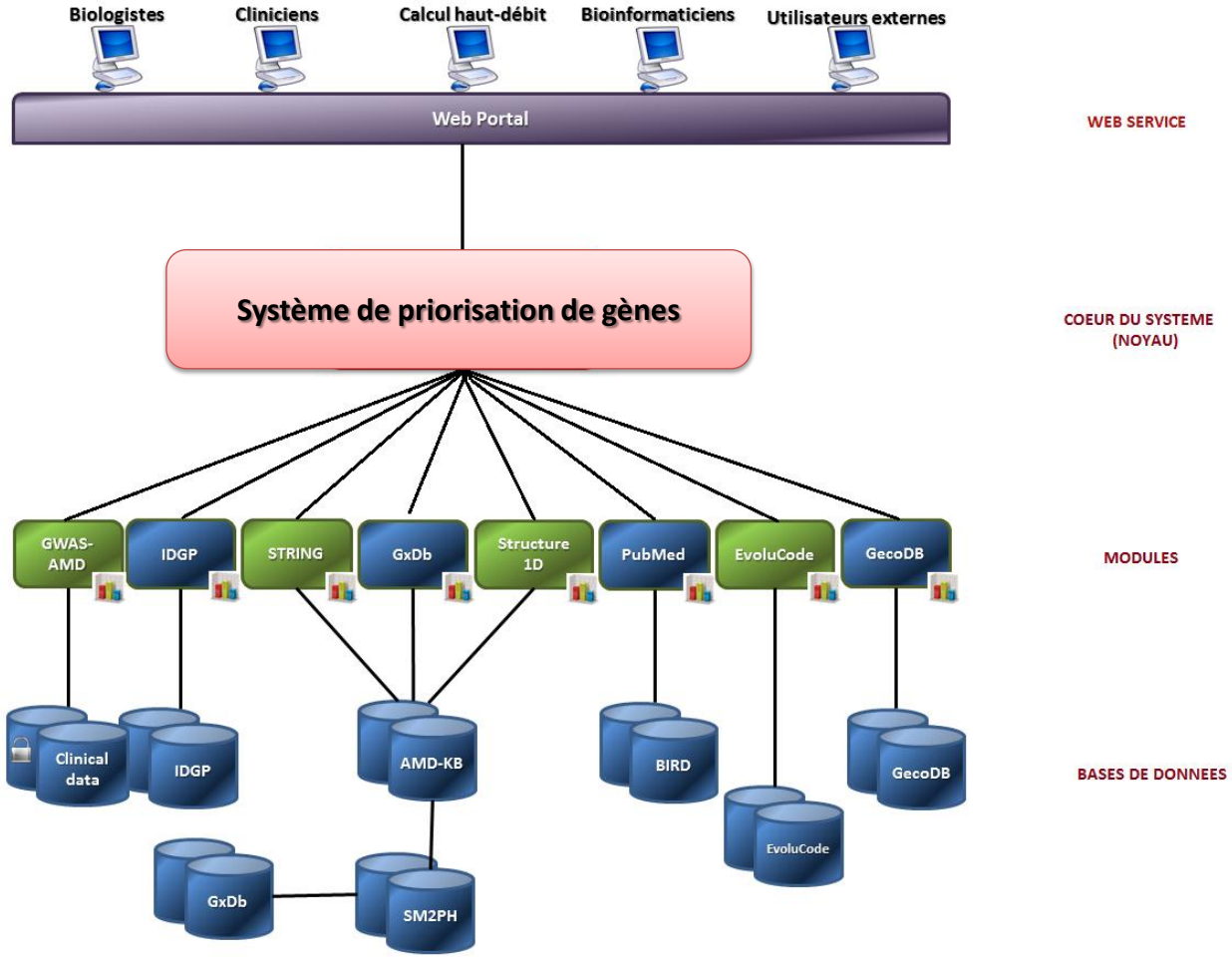


Figure 61. Architecture multicouches de notre système de priorisation de gènes. Le premier niveau correspond au stockage de l'information dans les différents entrepôts de données. Le second niveau correspond à l'ensemble des modules de priorisation qui traitent cette information. Le troisième niveau correspond au cœur de l'application qui administre les tâches et se charge d'interpréter les résultats locaux pour extraire un message global. Le quatrième niveau correspond à la couche d'interaction avec l'utilisateur.

Nous avons conçu un prototype du système de priorisation de gènes. La structure de ce système est entièrement modulaire permettant d'intégrer ou de supprimer un module très facilement. La structure du système est composée de 4 couches essentielles (Figure 61) : i) le

web service, ii) le cœur de l'application, iii) les différents modules de priorisation et iv) l'information stockées dans des bases de données ou dans des fichiers. Chaque module de priorisation est associé à une source de données différente.

Notre prototype utilise huit critères d'évaluation (Figure 61) afin de prendre en compte un maximum de critères d'évaluation et augmenter la robustesse de l'outil.

- Le module *IDGP* qui repose sur la prédiction de la probabilité d'implication des gènes dans les maladies génétiques humaines (Lopez-Bigas and Ouzounis, 2004). Cette prédiction propose une classification partiellement supervisée définissant une probabilité d'implication de chaque gène dans une maladie génétique, dominante ou récessive (Calvo et al., 2007).

- Le module *STRING* qui s'appuie sur les données d'interactions protéines-protéines de SM2PH Central. En effet, si une protéine codée par un gène candidat interagit avec une ou plusieurs protéines impliquées dans un processus, il y a plus de chances que ces 2 protéines soient impliquées dans un processus commun.

- Le module *GxDb* qui permet de comparer les profils d'expression des gènes connus avec les profils d'expression des gènes candidats. En effet, 2 gènes sont susceptibles d'être impliqués dans un même réseau s'ils sont fréquemment co-exprimés ou ont des profils d'expression très similaires dans différentes situations (tissus, maladie, traitement…) de façon cohérente. La coexpression est un indicateur très intéressant car elle peut renseigner sur une interaction potentielle (directe ou indirecte) entre 2 gènes.

- Le module *Structure1D* qui repose sur la similarité des séquences protéiques. Rappelons que dans certaines familles protéiques comme les globines, une homologie de séquence de seulement 15% suffit à avoir une forte similarité de structure et de fonctions. De même, des domaines avec une homologie supérieure à 30% suffisent souvent pour témoigner d'une similarité évidente de structure et de fonction.

- Le module *PubMed* qui fait appel aux capacités du système BIRD pour traiter et analyser l'ensemble des 20 494 848 résumés de PubMed (juillet 2012).

- Le module *EvoluCode* qui repose sur les codes-barres évolutionnaires 1D (voir section 3.7). Ce module donne une information multi-niveaux et la visualisation de l'histoire évolutive de complexes protéiques à l'échelle du génome.

- Le module *GecoDB* qui permet d'étudier la similarité du contexte génomique entre les gènes connus et les gènes candidats. Il permet d'une part, d'observer la similarité de l'environnement d'un gène en comparant la distance des plus proches éléments génétiques ou épigénétiques (disponibles à l'UCSC) en amont ou en aval et d'autre part, il permet d'observer la similarité des éléments présents dans la région transcrite. Une comparaison des cartes exoniques pourrait également être envisagée (mais n'est pas implémentée à l'heure actuelle).

- Le module *GWAS-AMD* qui repose sur des données confidentielles issues de l'étude d'association pangénomique (GWAS) fournies par l'équipe de T. Léveillard de l'Institut de la Vision (IdV) dans le cadre du consortium AMD mondial. Ce consortium a pour objectif de découvrir de nouveaux gènes potentiellement impliqués dans les Dégénérescences Maculaires liées à l'âge (DMLA ou AMD en anglais pour *Age-related Macular Degeneration*) via l'étude des associations SNPs-maladie observées chez plusieurs milliers de patients collectés sur toute la planète. Ces données correspondent

aux valeurs observées pour 40 polymorphismes d'intérêt qui ont permis de définir une liste de 52 gènes dont certains sont : i) déjà connus comme étant responsables de la DMLA, ii) révélés par l'étude pangénomique ou iii) potentiellement responsables de la DMLA au regard des valeurs limites obtenues dans l'étude pangénomique.

Le cœur de l'application (noyau), quant à lui, prend en charge les informations fournies par l'utilisateur (set d'entrainement, set de test, processus biologique étudié). Il établit alors les liens entre les gènes et les protéines résultant de leur transcription/traduction, puis transmet l'ensemble de ces informations aux modules de priorisation. Une fois l'information traitée par ces modules, le noyau récupère les scores et puis détermine le classement des gènes candidats.

8.3 Test

Notre prototype de priorisation de gènes cibles a été appliqué à l'étude de la Dégénérescence Maculaire Liée à l'Âge (DMLA) en mettant à profit la liste de 52 gènes et les données *GWAS* confidentielles fournies par notre collaborateur T. Léveillard de l'Institut de la Vision (Paris)

La DMLA est une maladie dégénérative de « vieillesse » très répandue chez l'homme qui est à l'origine de la cécité ou d'une baisse sévère de l'acuité visuelle d'un grand nombre d'occidentaux. Très difficile à diagnostiquer, elle est souvent décelée dix à quinze ans après l'apparition des premiers symptômes qui sont souvent mal pris en compte par le patient. De plus, la DMLA étant une maladie de la « vieillesse », le criblage des gènes incriminés est difficile car la pathologie intervient à un stade où de nombreux processus sont perturbés. L'origine de la pathologie peut donc être la conséquence d'un dérèglement quelconque qui n'est pas directement lié à cette pathologie. Cette manifestation tardive explique pourquoi il est très difficile de trouver des organismes modèles. A ce jour, il n'existe aucun modèle murin connu, car le cycle de vie très court de la souris (*Mus musculus*) ne permet pas l'apparition de la maladie. Dans le domaine de la « vision », où les données très abondantes provenant de la souris sont essentielles, l'incapacité à exploiter ce modèle animal pour l'étude de la DMLA est très problématique. Il est donc intéressant d'effectuer une première sélection des meilleurs gènes candidats en faisant appel à des techniques *in silico*, en l'occurrence la priorisation de gènes.

Suite à la version beta de notre système de priorisation, nommé GEPeTTO (pour *GEne PrioTization TOol*), implémenté par Vincent Walter (stagiaire), le set d'entrainement incluant 23 gènes connus et le set de test incluant les 29 gènes candidats, liés à la DMLA ont été soumis à différents outils pour effectuer une analyse comparative. Les outils utilisés sont Endeavour (Aerts et al., 2006a; Tranchevent et al., 2008), outil fondateur et ToppGene (Chen et al., 2009).

Pour mettre en évidence la justesse de l'outil, le set d'entrainement a été subdivisé en 2 sets puis l'un des 2 a été inclus dans le set de test. 3 jeux de données différents ont ainsi été construits :i) les 14 gènes connus de la littérature contre les autres gènes, ii) les 9 gènes validés confidentiels contre les autres gènes, iii) la moitié des gènes connus de la littérature et la moitié de gènes validés et confidentiels contre les autres gènes.

Le protocole de validation mis en place a visé à comparer la capacité des différents outils (GEPeTTO, ToppGene et Endeavour) à reconnaître les gènes connus inclus dans le set de test. Pour cela, nous avons utilisé la courbe ROC (*Receiver Operating Characteristics*) et son critère AUC. La courbe ROC constitue un outil adapté à la visualisation des performances d'un classifieur, elle représente le compromis entre les faux positifs (spécificité) et les vrais positifs

(sensibilité) (Figure 62). L'aire située sous la courbe (*Area Under Curve*, AUC) synthétise l'information de la courbe ROC. AUC correspond à la probabilité qu'un évènement positif soit considéré comme positif par le classifieur sur l'étendue des observations. Un modèle idéal aura une valeur AUC=1 et un modèle aléatoire une valeur AUC=0.5. Par convention, une valeur d'AUC>0.7 est caractéristique d'un bon modèle, un modèle très discriminant doit avoir une valeur AUC>0.85 et le modèle sera considéré comme excellent pour une valeur AUC>0.9.

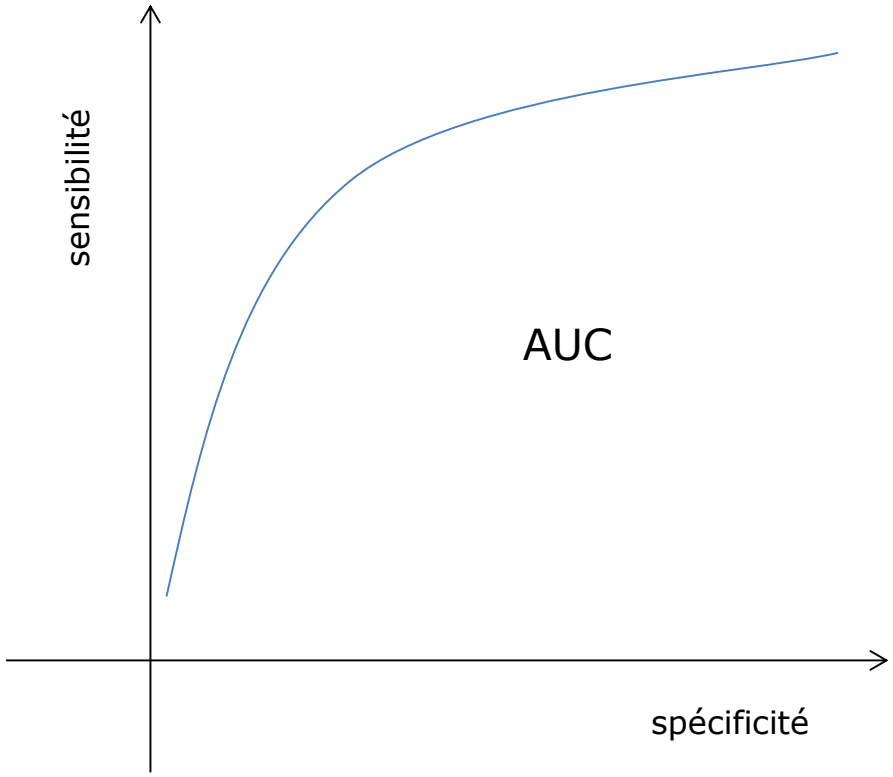
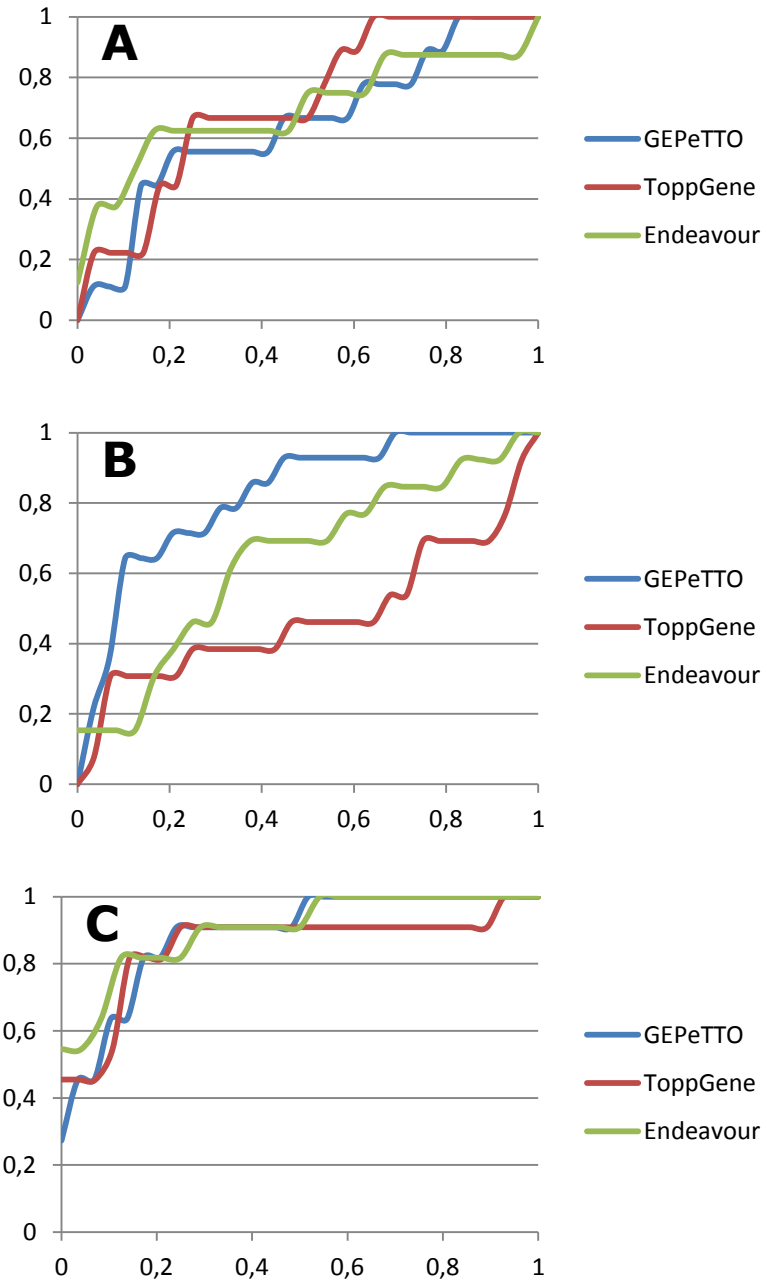


Figure 62. La courbe ROC et son critère AUC.

La courbe ROC obtenue avec les résultats fournies avec GEPeTTO est très proche des courbes ROC obtenues pour les 2 outils de priorisation Endeavour et ToppGene (Figure 63). Cela indique que les résultats obtenus sont d'une qualité à priori équivalente à ceux obtenus par ces 2 outils, considérés comme des les références actuelles. De plus, les valeurs d'AUC obtenues avec GEPeTTO montrent que GEPeTTO utilise un bon modèle de priorisation car les valeurs obtenues sont comprises dans l'intervalle [0.64;0.887] en étant globalement compris entre 0.8 et 0.9. GEPeTTO est donc un système de priorisation prometteur, du moins pour notre cas d'étude, à savoir la DMLA.



D

SET D'APPRENTISSAGE	VALEUR AUC		
	GEPETTO	ToppGene	Endeavour
Gènes connus (14)	0.649	0.720	0.701
Gènes validés (9)	0.825	0.479	0.643
Mélange (11)	0.887	0.854	0.903

Figure 63. Les courbes ROC de notre système de priorisation de gènes en comparaison par rapport à d'autres outils (Endeavour et ToppGene) suivant 3 sets d'entraînement A) les gènes connus dans la littérature pour être impliqué dans la DMLA B) les gènes validés mais pas encore publiés et C) un mélange entre des gènes connus et les gènes validés, D) Les valeurs AUC des courbes dans a,b,c.

8.4 Conclusions et perspectives

Nous avons conçu le prototype d'un système de priorisation GEPeTTO basé sur des sources de données multiples, ayant ainsi permis de prioriser les gènes candidats selon leurs probabilités d'implication dans la DMLA. Ce prototype, développé en mettant à profit SM2PH Central et divers outils ou bases de données, illustre bien l'interopérabilité des différents systèmes développés. Ceci confère un grand avantage, à savoir l'accès à des sources communes de données récupérées ou calculées, pérennes, maintenues et mises à jour constamment. De plus, de par sa conception, l'architecture entièrement modulaire de l'application permet d'ajouter facilement de nouveau modules de priorisation (c'est-à-dire de nouveaux critères d'évaluation). Enfin, les premiers résultats obtenus lors de la comparaison avec les outils les plus couramment utilisés ont montré que notre système de priorisation est un système acceptable.

Notre système de priorisation est actuellement en cours de validation. Néanmoins, en raison des divers modules non spécifiques à la DMLA, il est déjà possible d'affirmer que le système pourra être appliqué facilement à l'étude de nouveaux cas (processus biologiques, autres pathologies, etc.). Dans ce cadre, on envisage déjà de fournir une meilleure gestion de ces modules en permettant à l'utilisateur de donner un poids aux différents critères d'évaluation, comme le font déjà certains outils tel CANDID (Hutz et al., 2008) ou Endeavour.

CINQUIEME PARTIE : APPLICATIONS

La cinquième partie (Chapitre 9), présente un exemple d'application de nos développements réalisée dans le cadre de l'étude de l'impact structural et phénotypique de nouvelles mutations du gène GPR179 impliqué dans la cécité nocturne. Le manuscrit publié dans la revue *American Journal of Human Genetics* est joint à ce chapitre.

CHAPITRE 9. ILLUSTRATION DES CAPACITES DE NOS SYSTEMES

9.1 Introduction

Ce chapitre illustre comment la combinaison du système SM2PH Central, du service d'annotation MSV3d et de la méthode de prédiction KD4v a permis d'analyser et de valider l'impact phénotypique de mutations nouvellement identifiées. Cette application de notre infrastructure a été réalisée dans le cadre d'une collaboration avec l'équipe d'Isabelle Audo et Christina Zeitz (Institut de la Vision, Paris) centrée sur l'analyse de résultats de séquençage intensif d'exomes de patients. Brièvement, ces expériences ont abouti à la caractérisation de 3 mutations faux-sens au sein du gène GPR179. Ce gène code pour un récepteur couplé aux protéines G et les mutations identifiées ont été corrélées à l'émergence de la Cécité Nocturne Stationnaire Congénitale (en anglais *Congenital Stationary Night Blindness* ou *CSNB*), une maladie rare de la rétine qui se manifeste dès la naissance ou dans les premiers mois de la vie. Causé par une transmission défectueuse du signal lumineux dans la rétine, ce trouble de la vision est particulièrement handicapant lorsque que le patient est dans un environnement de faible luminosité. 3 modes d'hérédité de cette maladie existent : autosomique dominante, récessive et récessive liée à l'X. Actuellement, il n'existe pas de traitement de cette affection.

Les résultats de cette étude regroupant les aspects expérimentaux et médicaux, ainsi que nos prédictions sont décrits dans l'article présenté dans la suite de ce chapitre : «*Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness*» publié dans la revue *American Journal of Human Genetics*. Pour mener à bien ces travaux, nous avons du répondre à de nombreuses questions dont certaines sont récurrentes des analyses de mutations nouvelles et d'autres plus spécifiques du gène mis en évidence, on peut citer :

- Quelle est la fonction du gène GPR179 ? En effet, au delà du domaine correspondant à la fonction de GPCR, aucune information n'était disponible concernant les autres domaines de cette grande protéine de 2 367 acides aminés ?
- Est-ce que ce gène s'exprime ? dans quels tissus ? et avec quels gènes se co-exprime-t-il?
- La protéine GPR179 interagit-elle avec d'autres protéines ? Est-elle impliquée dans des réseaux biologiques ?
- Parmi les différentes structures 3D éloignées de GPCR disponibles, peut-on trouver une structure proche de GPR179 qui permette de construire un modèle 3D par homologie très informatif ?
- Les positions mutées sont-elles conservées ? Sont-elles situées à proximité de sites « remarquables » ?
- Basée sur les modèles 3D, les résidus mutées sont-ils enfouis ou « de surface » ?
- Et finalement, bien évidemment, peut-on prédire l'impact phénotypique de ces mutations ?

Cependant, on peut noter que pour répondre à ces questions, plusieurs interventions humaines ont été nécessaires et des améliorations ont été introduites dans SM2PH Central et dans les stratégies d'exploitation des informations de séquence/structure/fonction/évolution ou de génomique fonctionnelle. La recherche d'homologie a été réalisée pour augmenter la diversité des séquences au sein de l'alignement. Au regard de l'hétérogénéité de la qualité des séquences, nous avons dû i) intervenir sur cet alignement afin de ne retenir que les 100 premières séquences détectées par Blast et ii) améliorer manuellement l'alignement afin d'obtenir un score NorMD (1,74) de très bonne qualité. Les informations fournies par MACSIMS ont permis de clarifier les données taxonomiques, d'identifier des sites « remarquables », des domaines ou d'obtenir des annotations fonctionnelles via GO. La structure de la squid rhodopsine (PDB : 2ZIY) a été sélectionnée comme empreinte et le modèle 3D a été généré. Toutes ces informations ont été stockées dans une nouvelle SM2PH instance, SM2PH-GPR179, dédiée à l'étude du gène GPR179. Pour annoter 2 des 3 mutations faux-sens responsables des atteintes rétiniennes (les 2 mutations dont les positions 455 et 603 se situent dans le domaine GPCR), MSV3d a envoyé la requête à SM2PH-GPR179 pour accéder aux alignements et aux modèles 3D. Ces informations ont permis à MSV3d de calculer l'ensemble des 33 valeurs décrivant, la conservation, la localisation des mutations par rapport aux éléments de structures secondaires et tertiaires ainsi que les modifications de propriétés physico-chimiques, d'accessibilité au solvant, de contacts entre résidus, de stabilité et de variation d'énergie libre. Ces données ont ensuite été envoyées à KD4v pour prédire l'impact de ces mutations sur la fonction du GPR179 et leur rôle potentiel dans le phénotype pathologique rétinien. KD4v a prédit que les 2 mutations p.Gly455Asp et p.His603Tyr sont délétères. Les programmes SIFT et PolyPhen-2 ont donné le même résultat. La Figure 64 ci-dessous présente la sortie du système KD4v pour la mutation p.Gly455Asp qui illustre que 4 règles sont associées à cette prédiction délétère.

Ces 4 règles ont les rangs 13, 35, 56 et 103, ce qui, dans KD4v, indique indirectement l'importance de ces règles rangées selon le nombre de mutations dans le jeu d'entraînement qu'ils couvrent. Plus le rang est faible, plus le nombre de mutations est élevé.

Pour illustrer comment interpréter les règles PLI, on peut considérer l'exemple de la règle de rang 13, décrite ci-dessous. Cette règle est assez rigoureuse puisqu'elle a identifié correctement 160 mutations délétères (2.67% des 6 000 étudiées) et a mal classé 5 mutations neutres qu'elle a considérées délétère (0.25% de 2 000 mutations neutres dans le jeu d'entraînement).

```
deleterious(A) :-
        modif_size(A, size_increase) and
        modif_charge(A, charge_increase) and
        conservation_class(A, global_conservation_rank_2) and
        wt_accessibility(A, buried)
```

En français, cette règle peut se traduire traduite comme suit :

Cette mutation est prédite pour être délétère si :

- La taille du résidu muté est plus grand que celle du résidu sauvage et
- La substitution entraîne l'augmentation de charge (le résidu neutre est remplacé par le résidu ayant charge positive ou négative) et
- Cette substitution est très conservative et
- La position mutée est enfouie dans le modèle 3D.

En analysant cette règle, on peut aisément appréhender les éléments qui ont conduit à la prédiction délétère et notamment le fait que :

- La position mutée est très conservée. La fréquence du résidu sauvage (glycine) à la position muté est 97%, alors que le résidu mutant (acide aspartique) n'est jamais observé dans les séquences alignées.
- De même, dans le modèle 3D, la position mutée est enfouie (accessibilité <= 10) et le résidu sauvage (glycine) qui est neutre et petit est remplacé par un résidu (acide aspartique) qui est grand et chargé négativement.

Tous ces éléments suggèrent que la mutation pourrait engendrer une déstabilisation du cœur de la protéine et expliquer le phénotype observé.

Actuellement, une autre étude est réalisée avec la même méthodologie pour la découverte de mutations spécifiques du gène LRIT3.

GENERAL INFORMATION

Protein ID	Protein name	Gene name	Substitution
Q6PRD1	Probable G-protein coupled receptor 179 [Precursor]	GP179_HUMAN	p.Gly455Asp

PREDICTION

This mutation is predicted to be deleterious by the rules:

- deleterious(A) :- modif_polarity(A, polarity_increase) and conservation_class(A, global_conservation_rank_2) and mut_accessibility(A, buried) and stability(A, decrease). (ranking: 103)
- deleterious(A) :- g_or_p(A, g_or_p_disparition) and gain_contact(A, phob) and wt_accessibility(A, buried). (ranking: 56)
- deleterious(A) :- modif_charge(A, charge_increase) and modif_hydrophobicity(A, hydrophobicity_decrease) and secondary_struc(A, helix) and wt_accessibility(A, buried) and mut_accessibility(A, buried). (ranking: 35)
- deleterious(A) :- modif_size(A, size_increase) and modif_charge(A, charge_increase) and conservation_class(A, global_conservation_rank_2) and wt_accessibility(A, buried). (ranking: 13)

PHYSICO-CHEMICAL PROPERTIES

Size	Charge	Polarity	Hydrophobicity
increase	increase	increase	decrease
Disulfid Bond	**Gly or Pro**	**Modification Score**	
unchanged	disparition	28	

CONSERVATION

Conservation in the alignment	Number of known mutations at this position	Wild type residue representation in alignment (%)	Mutant residue representation in alignment (%)
Global conservation - Rank 2	1.0	97	0

View alignment with: Jalview

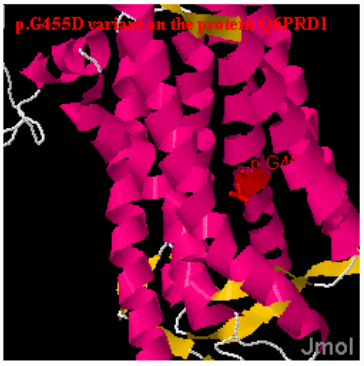
LOCALISATION

In a secondary structure element ?	In an annoted site ?
HELIX	PF00003 7tm_3 390_632

STRUCTURAL DATA

Template PDB code	Chain	Fold	
2ziy	A	-	
Additional contact	**Lost contact**	**Identical contact**	
3	0	9	
Accessibility		**Free energy change**	
Wild type	**Mutant**	**Protein stability**	**Reliability index**
1.72 (Buried)	0.69 (Buried)	Decrease	1

MAPPED TO 3D PROTEIN STRUCTURE



[Download PDB file]

DOWNLOAD

[XML] [Text]

Figure 64. Capture d'écran de la page fournissant le résultat de prédiction de KD4v ainsi que la caractérisation multi-niveaux de la mutation p.Gly455Asp du gène GPR179. L'utilisateur peut visualiser et/ou télécharger le modèle 3D dans la fenêtre Jmol à droite.

9.2 Publication

Audo I, Bujakowska K, Orhan E, Poloschek CM, Defoort-Dhellemmes S, Drumare I, Kohl S, Luu TD, Lecompte O, Zrenner E, Lancelot ME, Antonio A, Germain A, Michiels C, Audier C, Letexier M, Saraiva JP, Leroy BP, Munier FL, Mohand-Saïd S, Lorenz B, Friedburg C, Preising M, Kellner U, Renner AB, Moskova-Doumanova V, Berger W, Wissinger B, Hamel CP, Schorderet DF, De Baere E, Sharon D, Banin E, Jacobson SG, Bonneau D, Zanlonghi X, Le Meur G, Casteels I, Koenekoop R, Long VW, Meire F, Prescott K, de Ravel T, Simmons I, Nguyen H, Dollfus H, Poch O, Léveillard T, Nguyen-Ba-Charvet K, Sahel JA, Bhattacharya SS, Zeitz C. *Whole exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness*. *Am J Hum Genet* 90, 321-330. 2012.

Whole-Exome Sequencing Identifies Mutations in *GPR179* Leading to Autosomal-Recessive Complete Congenital Stationary Night Blindness

Isabelle Audo,[1,2,3,4,5,39] Kinga Bujakowska,[1,2,3,39] Elise Orhan,[1,2,3] Charlotte M. Poloschek,[6] Sabine Defoort-Dhellemmes,[7] Isabelle Drumare,[7] Susanne Kohl,[8] Tien D. Luu,[9] Odile Lecompte,[9] Eberhart Zrenner,[10] Marie-Elise Lancelot,[1,2,3] Aline Antonio,[1,2,3,4] Aurore Germain,[1,2,3] Christelle Michiels,[1,2,3] Claire Audier,[1,2,3] Mélanie Letexier,[11] Jean-Paul Saraiva,[11] Bart P. Leroy,[12,13] Francis L. Munier,[14] Saddek Mohand-Saïd,[1,2,3,4] Birgit Lorenz,[15] Christoph Friedburg,[15] Markus Preising,[15] Ulrich Kellner,[16] Agnes B. Renner,[17] Veselina Moskova-Doumanova,[1,2,3] Wolfgang Berger,[18,19,20] Bernd Wissinger,[8] Christian P. Hamel,[21] Daniel F. Schorderet,[22] Elfride De Baere,[12] Dror Sharon,[23] Eyal Banin,[23] Samuel G. Jacobson,[24] Dominique Bonneau,[25] Xavier Zanlonghi,[26] Guylene Le Meur,[27] Ingele Casteels,[28] Robert Koenekoop,[29] Vernon W. Long,[30] Francoise Meire,[31] Katrina Prescott,[32] Thomy de Ravel,[33] Ian Simmons,[30] Hoan Nguyen,[9] Hélène Dollfus,[34,35] Olivier Poch,[9] Thierry Léveillard,[1,2,3] Kim Nguyen-Ba-Charvet,[1,2,3] José-Alain Sahel,[1,2,3,4,5,36,37] Shomi S. Bhattacharya,[1,2,3,5,38] and Christina Zeitz[1,2,3,*]

Congenital stationary night blindness (CSNB) is a heterogeneous retinal disorder characterized by visual impairment under low light conditions. This disorder is due to a signal transmission defect from rod photoreceptors to adjacent bipolar cells in the retina. Two forms can be distinguished clinically, complete CSNB (cCSNB) or incomplete CSNB; the two forms are distinguished on the basis of the affected signaling pathway. Mutations in *NYX*, *GRM6*, and *TRPM1*, expressed in the outer plexiform layer (OPL) lead to disruption of the ON-bipolar cell response and have been seen in patients with cCSNB. Whole-exome sequencing in cCSNB patients lacking mutations in the known genes led to the identification of a homozygous missense mutation (c.1807C>T [p.His603Tyr]) in one consanguineous autosomal-recessive cCSNB family and a homozygous frameshift mutation in *GPR179* (c.278delC [p.Pro93Glnfs*57]) in a simplex male cCSNB patient. Additional screening with Sanger sequencing of 40 patients identified three other cCSNB patients harboring additional allelic mutations in *GPR179*. Although, immunhistological studies revealed Gpr179 in the OPL in wild-type mouse retina, Gpr179 did not colocalize with specific ON-bipolar markers. Interestingly, Gpr179 was highly concentrated in horizontal cells and Müller cell endfeet. The involvement of these cells in cCSNB and the specific function of GPR179 remain to be elucidated.

[1]Institut National de la Santé et de la Recherche Médicale, U968, Paris 75012, France; [2]Université Pierre et Marie Curie (UPMC Paris 06), UMR_S 968, Institut de la Vision, Paris 75012, France; [3]Centre National de la Recherche Scientifique, UMR_7210, Paris 75012, France; [4]Centre Hospitalier National d'Ophtalmologie des Quinze-Vingts, INSERM-DHOS CIC 503, Paris 75012, France; [5]Institute of Ophthalmology, University College of London, London EC1V 9EL, UK; [6]Department of Ophthalmology, University of Freiburg, Freiburg 79106, Germany; [7]Laboratoire Neurosciences Fonctionnelles et Pathologies, CNRS FRE 2726, Hôpital Roger Salengro, Lille 59037 Cedex, France; [8]Molecular Genetics Laboratory, Institute for Ophthalmic Research, Department for Ophthalmology, University of Tuebingen, Tuebingen 72076, Germany; [9]Institut de Génétique et de Biologie Moléculaire et Cellulaire, Illkirch 67404 Cedex, France; [10]Centre for Ophthalmology, Department for Ophthalmolgy, University Tuebingen, Tuebingen 72076, Germany; [11]IntegraGen, Genopole CAMPUS 1 bat G8 FR-91030, Evry 91000, France; [12]Center for Medical Genetics, Ghent University, Ghent 9000, Belgium; [13]Department of Ophthalmology, Ghent University, Ghent 9000, Belgium; [14]Unit of Oculogenetics, Jules Gonin Eye Hospital, Lausanne 1004, Switzerland; [15]Department of Ophthalmology, Justus-Liebig-University Giessen, Universitaetsklinikum Giessen and Marburg GmbH Giessen Campus, Giessen 35385, Germany; [16]AugenZentrum Siegburg, Siegburg 53721, Germany; [17]Department of Ophthalmology, University Medical Center Regensburg, 93053, Regensburg, Germany; [18]Institute of Medical Molecular Genetics, University of Zurich, Zurich 8057, Switzerland; [19]Neuroscience Center Zurich, University and ETH Zurich, Zurich 8057, Switzerland; [20]Center for Integrative Human Physiology, University of Zurich, Zurich 8057, Switzerland; [21]National Centre for Genetic Sensory Diseases, Montpellier 34295 Cedex 05, France; [22]IRO–Institut de Recherche en Ophtalmologie and Faculté des Sciences du Vivant, Ecole Polytechnique Fédérale de Lausanne, University of Lausanne, Sion 1950, Switzerland; [23]Department of Ophthalmology, Hadassah-Hebrew University Medical Center, Jerusalem 91120, Israel; [24]University of Pennsylvania, Scheie Eye Institute, Philadelphia 19104, PA, USA; [25]UMR Institut National de la Santé et de la Recherche Médicale, U771-CNRS6214 et CHU, Angers 49000, France; [26]Service Exploration Fonctionnelle de la Vision et Centre basse vision de la Clinique Sourdille, Nantes 44000, France; [27]CHU-Hotel Dieu, Service d'Ophtalmologie, Nantes 44093, France; [28]Department of Ophthalmology, University Hospitals, Leuven 3000, Belgium; [29]McGill Ocular Genetics Laboratory, McGill University, Montreal, QC H3H 1P3, Canada; [30]St James's University Hospital, Leeds LS9 7TF, UK; [31]Hopital Des Enfants Reine Fabiola, Brussels 1020, Belgium; [32]Yorkshire Regional Genetics Service, Department of Clinical Genetics, Chapel Allerton Hospital, Leeds LS7 4SA, UK; [33]Centre for Human Genetics, Leuven University Hospitals, Leuven 3000, Belgium; [34]Centre de Référence pour les Affections Rares en Génétique Ophtalmologique, Hôpitaux Universitaires de Strasbourg, Strasbourg 67000, France; [35]Laboratoire de Physiopathologie des Syndromes Rares Héréditaires, équipe avenir INSERM, Faculté de Médecine, Université de Strasbourg, Strasbourg 67000, France; [36]Fondation Ophtalmologique Adolphe de Rothschild, Paris 75019, France; [37]Académie des Sciences–Institut de France, Paris 75006, France; [38]Department of Celular Therapy and Regenerative Medicine, Andalusian Molecular Biology and Regenerative Medicine Centre (CABIMER), Isla Cartuja, Seville 41902, Spain

[39]These authors contributed equally to this work

*Correspondence: christina.zeitz@inserm.fr

DOI 10.1016/j.ajhg.2011.12.007.

Congenital stationary night blindness (CSNB) comprises a group of genetically and clinically heterogeneous retinal disorders. The associated genes encode proteins that are confined to the phototransduction cascade or are important in retinal signaling from photoreceptors to adjacent bipolar cells.[1] Most of the patients with mutations in these genes show a typical electrophysiological phenotype characterized by an electronegative waveform of the dark-adapted bright flash electroretinogram (ERG), in which the amplitude of the b-wave is smaller than that of the a-wave.[2] This so-called Schubert-Bornschein-type ERG response can be divided in two subtypes, incomplete CSNB ([icCSNB] CSNB2A [MIM 300071], CSNB2B [MIM 610427]) and complete CSNB ([cCSNB] CSNB1A [MIM 310500], CSNB1B [MIM 257270] and CSNB1C [MIM 613216]).[3] icCSNB has been characterized by both a reduced rod b-wave and substantially reduced cone responses because of both ON- and OFF-bipolar cell dysfunction, whereas the complete type is associated with a drastically reduced rod b-wave response because of ON-bipolar cell dysfunction but largely normal cone b-wave amplitudes.[4] icCSNB has been associated with mutations in *CACNA1F* [MIM 300110], *CABP4* [MIM 608965], and *CACNA2D4* [MIM 608171], whereas cCSNB has been associated with mutations in *NYX* [MIM 300278], *GRM6* [MIM 604096], and *TRPM1* [MIM 603576]. So far more than 280 mutations have been identified in these genes by us and others via direct sequencing of candidate genes (unpublished data) or microarray analysis.[5] Prevalence studies determined that *CACNA1F*, *NYX*, and *TRPM1* mutations leading to incomplete and complete CSNB occur more frequently (unpublished data). Genotyping studies of our CSNB cohort, comprising 160 patients, reveal that in ~13% of cases mutations in known genes underlying CSNB were not identified. This is a strong indication that mutations in other genes remain to be discovered or that mutations in unscreened regions, that is regulatory elements and introns, might be involved. Mutations in many genes leading to CSNB have been identified through a candidate gene approach by comparing the human phenotype to similar phenotypes observed in knockout or naturally occurring animal models.[6–13] The bottleneck of this approach is the size of a cohort and the identification of the "right" patient harboring the mutation in such a candidate gene. Novel techniques that use massively parallel sequencing of all human exons have recently been successful in identifying mutations in novel genes in other heterogeneous diseases such as Leber congenital amaurosis.[14,15] To rapidly identify the missing mutations in our CSNB cohort after whole-exome enrichment (IntegraGen, Evry, France), we sequenced four exomes from a consanguineous autosomal-recessive cCSNB family (that included parents who were first cousins and two of three affected children) and from a sporadic male cCSNB patient of Portuguese origin (Figure S1A and S2, available online, shows the typical cCSNB ERG of patient CIC02756). One index patient from each family was previously excluded by Sanger sequencing for mutations in *GRM6* and *TRPM1*. In addition, the sporadic male patient was also excluded for mutations in *NYX*. Research procedures were conducted in accordance with institutional guidelines and the Declaration of Helsinki. Prior to genetic testing, informed consent was obtained from all patients and their family members. Ophthalmic examination included best corrected visual acuity, slit lamp examination, fundoscopy, perimetry, full-field (ERG) incorporating the International Society for Clinical Electrophysiology of Vision (ISCEV) standards,[16] fundus autofluorescence (FAF), and optical coherence tomography (OCT) (the extent of investigation depended on the referring center). Exons of DNA samples were captured with in-solution enrichment methodology (SureSelect Human All Exon Kits Version 3, Agilent, Massy, France) with the company's biotinylated oligonucleotide probe library (Human All Exon v3 50 Mb, Agilent). Each genomic DNA was then sequenced on a sequencer as paired-end 75 bases (Illumina HISEQ, Illumina, San Diego, USA). Image analysis and base calling were performed with Real Time Analysis (RTA) Pipeline version 1.9 with default parameters (Illumina). The bioinformatic analysis of sequencing data was based on a pipeline (Consensus Assessment of Sequence and Variation [CASAVA] 1.8, Illumina). CASAVA performs alignment, calls the SNPs based on the allele calls and read depth, and detects variants (SNPs and indels). Genetic variation annotation was performed by an in-house pipeline (IntegraGen) and results were provided per sample or family in tabulated text files. After excluding variants observed in dbSNP 132, data were further filtered to keep only variants in coding and splice regions that were present in a homozygous state in the affected children and in a heterozygous state in the parents from the consanguineous family. This allowed us to reduce the number of variants from 5,901 indels to 1 and from 66,621 SNPs to 7. The observed deletion represented a repeat deletion in the penultimate exon of *VSIG10* and was therefore unlikely to be a disease-causing variant. However, three missense mutations predicted to be probably or possibly damaging were identified in three different genes (*KIAA0753*, *CRHR1* [MIM 122561], and *GPR179* [G protein-coupled receptor 179]) on chromosome 17. The p.Arg518Cys variant found in *KIAA0753* was considered unlikely to be disease causing because this arginine residue is not evolutionarily conserved. On the other hand, both the p.Arg259Gln substitution in *CRHR1* and the p.His603Tyr in *GPR179* affected highly evolutionary conserved amino acid residues (Figure 1 and Figure S1B). Interestingly, the other cCSNB patient (CIC02756), also studied by whole-exome sequencing, carried a homozygous 1 bp deletion, resulting in a frameshift and premature termination (p.Pro96Glnfs*57) in exon 1 of *GPR179*. These data strongly support the finding that mutations in *GPR179* lead to CSNB found in both families (Table 1). For the c.1807C>T (p.His603Tyr) mutation, both parents were found to be heterozygous because the nucleotide A was read 11 times and 7 times in the father and mother,

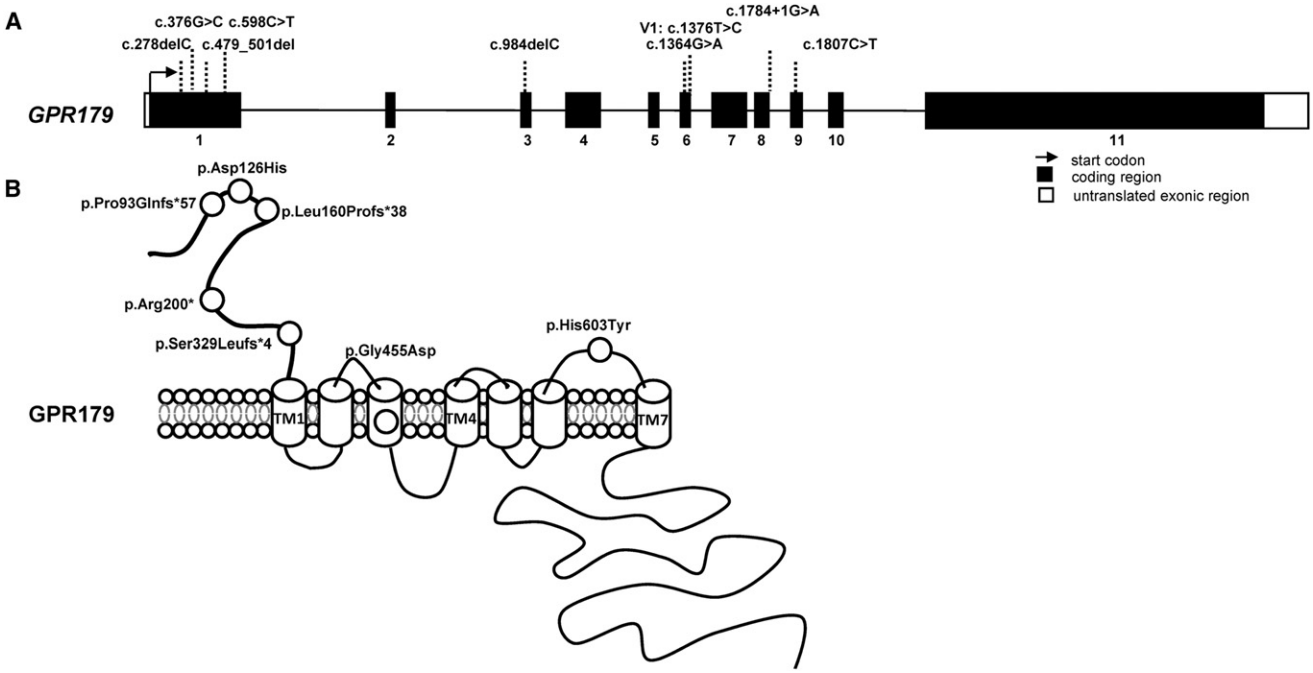


Figure 1. *GPR179* mutations in CSNB.
(A) *GPR179* structure containing 11 coding exons (NM_001004334.2). Different mutations identified in cCSNB patients are depicted. (B) The specific domains for GPR179 were estimated by a prediction program (UniProtKB/Swiss-Prot).

respectively, whereas the G was found 13 times and 11 times, respectively (reverse strand). The two affected children (patients CIC3308 and CIC04005) showed 26 times and 14 times the nucleotide A. The c.278delC deletion detected in the sporadic cCSNB patients was detected 22 times; 20 other reads of unknown type were also indicated. This might be due to the fact that at this position multiple Cs are present, and thus different reads might occur. Sanger sequencing confirmed the mutations in the index patients of each family. Both mutations cosegregated with the phenotype within the respective family (Figure S1A). In addition, next-generation sequencing data were used to analyze homozygous regions in the affected siblings (patients CIC03308 and CIC04005) of the consanguineous family. The analysis revealed seven major homozygous regions (>0.5 Mb), which were exclusively present on chromosome 17. *GPR179* was present in the second largest homozygous region (10.8 Mb), whereas *CRHR1* was present in a smaller region (1.3 Mb). In the other sporadic cCSNB patient, *GPR179* was not present in any major homozygous region; this can be explained by the fact that the parents were only distant cousins.

We screened 40 CSNB patients (cCSNB and unclassified CSNB) of various origins and from different clinical centers in Europe, the United States, Canada, and Israel by using Sanger sequencing for 27 fragments covering the 11 coding exons and flanking intronic regions of *GPR179* (NM_001004334.2). These were amplified by PCR in the presence of 1.5 mM $MgCl_2$ at an annealing temperature of 60°C. For one of the fragments a specific solution (solution S, 3×, fragment exon 11 m, Hot Fire Polymerase, Solis BioDyne, Tartu, Estonia, and primers; Table S1) was used. The PCR products were sequenced with a sequencing mix (BigDyeTerm v1.1 CycleSeq kit, Applied Biosystems, Courtabœuf, France), analyzed on an automated 48-capillary sequencer (ABI 3730 Genetic analyzer, Applied Biosystems), and the results interpreted by applying SeqScape software (Applied Biosystems). We detected three additional cCSNB patients who carried compound heterozygous disease-causing mutations (Table 1). The mutation spectrum identified herein comprises missense, splice-site, and nonsense mutations and deletions. None of these changes were present in control chromosomes (≥366 chromosomes). For patients whose family members could be investigated, the mutations cosegregated with the cCSNB phenotype, and the genotypes were indicative of an autosomal-recessive mode of inheritance (Table 1 and Figure S1A). Missense mutations were predicted to be pathogenic by PolyPhen and SIFT programs and were also found to affect evolutionarily conserved amino acid residues (Figure S1B). On the basis of all of the above evidence, we conclude that mutations in *GPR179* lead to cCSNB. Interestingly, we found four cCSNB patients with no mutations in *GRM6*, *TRPM1*, *NYX*, or *GPR179*, indicating that mutations in additional genes probably remain to be identified to explain these cases of cCSNB. In addition, a few rare variants (Table S2) in *GPR179* were identified in patients screened by Sanger sequencing and were classified as variants of unknown pathogenicity because only one mutation was observed or they did not affect conserved amino acid residues. The frequencies of *GPR179* polymorphisms found in our patients are provided in Table S3.

Table 1. Patients with Pathogenic *GPR179* Mutations

Patient Number	Relationship to Index Patient	Sex	Mutations Excluded in Following Genes	Ethnicity and Location	Exon	Nucleotide Exchange (RNA or Protein Effect)	Allele State	Control Alleles (Mutated or WT)	Phenotype Index
CIC02756[a]	–	male	*NYX, GRM6, TRPM1*	Portuguese-French; Paris, France	1	c.278delC (p.Pro93Glnfs*57)	homozygous	0/366	cCSNB, high myopia, nystagmus, moderate decreased visual acuity
CIC02757	unaffected father	male	–		1	c.278delC (p.Pro93Glnfs*57)	heterozygous		
CIC02758	unaffected mother	female	–		1	c.278delC (p.Pro93Glnfs*57)	heterozygous		
CIC03631	–	female	*GRM6, TRPM1*	French; Lille, France	1, 3	c.376G>C (p.Asp126His), c.984delC (p.Ser329Leufs*4)	compound heterozygous	0/366 and 0/372	cCSNB, high myopia, strabismus, micronystagmus
7699	–	female	*GRM6, TRPM1*	Tübingen, Germany	1, 6	c.479_501del (Leu160Profs*38), c.1364G>A (p.Gly455Asp)	compound heterozygous	0/366 and 0/384	cCSNB, strabismus, minimal rotational nystagmus, normal visual field
7692	unaffected father	male	–		1	c.479_501del (p.Leu160Profs*38)	heterozygous		
7697	unaffected mother	female	–		6	c.1364G>A (p.Gly455Asp)	heterozygous		
Y1049	–	female	*GRM6, TRPM1*	Lille, France	1, IVS8	c.598C>T (p.Arg200*), c.1784+1G>A (r.spl?)	compound heterozygous	0/378 0/378	cCSNB
Y1166	unaffected father	male	*GRM6, TRPM1*		IVS8	c.1784+1G>A (r.spl)	heterozygous		
Y1167	unaffected mother	female	*GRM6, TRPM1*		1	c.598C>T (p.Arg200*)	heterozygous		
Y1048	affected sister	female	*GRM6, TRPM1*		1, IVS8	c.598C>T (p.Arg200*), c.1784+1G>A (r.spl)	compound heterozygous		
26985[b,c]	–	male	*GRM6, TRPM1*	Lebanon; Freiburg, Germany	9	c.1807C>T (p.His603Tyr)	homozygous	0/366	cCSNB, left exotropia, until age of 2 nystagmus
CIC03306	father	male	–		9	c.1807C>T (p.His603Tyr)	heterozygous		ERG b-wave were slightly reduced for high flash strength
CIC03307	unaffected mother	female	–		9	c.1807C>T (p.His603Tyr)	heterozygous		-
CIC03308	affected sister	female	–		9	c.1807C>T (p.His603Tyr)	homozygous		cCSNB
CIC04005	affected sister	female	–		9	c.1807C>T (p.His603Tyr)	homozygous		cCSNB, visual acuity reduced

CSNB mutations are annotated according to the recommendation of the Human Genome Variation Society, with nucleotide position +1 corresponding to the A of the translation-initiation codon ATG in the cDNA nomenclature RefSeq NM_001004334.2.

[a] The parents of CIC02756 are far cousins (Figure S1A).

[b] The diagnostic for GRM6 for this patient was performed in Zurich, Switzerland.

[c] For this family consanguinity has been reported (Figure S1A).

To date no information is available on the functional characterization of *GPR179*. To predict the protein structure and the influence of the mutations identified herein, we created homology models. The human GPR179 sequence (UniProtKB identifier Q6PRD1) was used as a probe for similarity searches in the UniProtKB database with the use of the BlastP program.[17,18] In total, more than 100 metazoan sequences (excluding fragments) that were annotated or predicted as GPR179- or GPR158-like were highlighted and aligned with a customized version of the PipeAlign program.[19–21] *GPR179* codes for a protein with 2,367 amino acids that can be divided into four main regions corresponding to a small signal peptide (positions 1–25), the N-terminal extracellular region (position 26-381), the seven transmembrane (7TM)-spanning region (position 382-628), and the intracellular C-terminal region (position 629-2367) (Figure 1B). Sequence analysis predicted that the N-terminal extracellular region contains a calcium-binding EGF-like domain (position 278-324), whereas the C-terminal intracellular region is characterized by the presence of a short motif centered on the sequence CPWE, which is repeated at least 22 times in the GPR179-related proteins. GPR179 proteins are present in all vertebrates and are closely related to GPR158 and GPR158-like proteins. It is noteworthy that the major differences between GPR179 and the closely related GPR158 proteins rely on the absence of the calcium-binding EGF-like domain at the N-terminal part and a reduced number of CPWE motifs (up to three) in all GPR158 homologs. Interestingly, three other molecules, the regulator of G protein signaling 9 (RGS9 [MIM 604067]), the retinal rod rhodopsin-sensitive cGMP 3,5-cyclic phosphodiesterase subunit gamma (PDE6G [MIM 180073]) and the retinal cone rhodopsin-sensitive cGMP 3,5-cyclic phosphodiesterase subunit gamma (PDE6H [MIM 601190]) share the same protein motif CPWE. These molecules have been implicated in the inhibition of the G protein or amplification of the signal in the phototransduction cascade. Mutations in those genes lead to different retinal disorders, including bradyopsia [MIM 608415],[22] rod-cone dystrophy [MIM 613582],[23] and cone dystrophy [MIM 610024].[24]

Based on their seven transmembrane domain regions, both proteins (GPR179 and GPR158) belong to the glutamate receptor or class C GPCR proteins. This class includes, among others, metabotropic glutamate receptors (GRMs), two γ-aminobutyric acid B receptor (GABABR), the calcium-sensing receptor (CASR), the sweet and umami taste receptors and various orphan receptors.[25] The different deletions and the early termination mutation in *GPR179* identified in our patients are located in exons 1 and 3 and are predicted to lead to nonsense-mediated mRNA decay, which might result in the absence of a protein product. Alternatively, if a protein is formed, only the first extracellular part would be present but would lack all transmembrane domains of GPR179, resulting in truncated protein (Figure 1B). The missense alterations (p.Asp126His, p.Gly455Asp, and p.His603Tyr) affect evolutionarily conserved amino acid residues, which are predicted to be part of the first extracellular domain, within the third transmembrane domain, and in the last extracellular domain (Figure 1B). Multiple alignment analysis of more than 100 metazoan GPR179-related sequences shows strict conservation of the asparagine at position 126 (Asp126), the glycine at position 455 (Gly455), and the histidine at position 603 (His603) in vertebrate sequences. PolyPhen and SIFT programs annotated the three amino acid substitutions to be possibly pathogenic.[26] These programs use conservation among species and homologs to predict the pathogenic character of a mutation. In addition, an inductive logic programming prediction web server[27] predicted p.Gly455Asp and p.His603Tyr to be pathogenic. This program uses available 3D structures to predict the influence of a mutation. To date, no model of the 3D structure of the amino acid residues <300 is available, therefore the possible pathogenic effect of p.Asp126His could not be predicted with this program. To further gain insight into the deleterious effect of the missense mutations, we generated 3D models of the seven transmembrane (7TM)-spanning region of the human wild-type GPR179 and of the two altered proteins (p.Gly455Asp and p.His603Tyr) by homology modeling with MODELER software (Figure 2).[20] Two known 7TM templates, the bovine taste receptor (PDB 1F88) and the squid rhodopsine (PDB 2ZIY), were used to construct the homology models. For each 3D model construction, ten homology models were constructed, and the models with the best normalized discrete optimized potential energy (DOPE) score were selected.[28] The homology 3D models were visualized and analyzed by means of the SM2PH-db,[28] and figures were constructed with PyMOL software (version 0.99). Multilevel characterization of the mutants (physico-chemical changes and structural modifications induced by the substitution, as well as functional and structural features related to the mutated position) can be visualized and analyzed by the MSV3d web server. Structural analysis of the 3D homology models based on the squid RHO 3D model (2ZIY) localized the His603 in the external loop bridging the sixth and seventh transmembranes, whereas the Gly455 is localized within the third transmembrane helix, which is part of a binding pocket·(Figures 1B and 2). Our homology model predicts that the amino acid exchange p.Gly455Asp introduces a long negatively charged side chain that might point toward the cavity of the binding pocket. This suggests that the phenotypic consequences observed for this mutation might be related to some steric constraints hampering the normal functioning of the receptor. The steric constraints in respect to the p.His603Tyr mutation are less obvious. However, strong conservation across species and homologs are indicative of an important role for the histidine at position 603. Although, for the moment, the 3D structure of the amino acid residues <300 of GPR179 is not available, we know from other receptors that the N

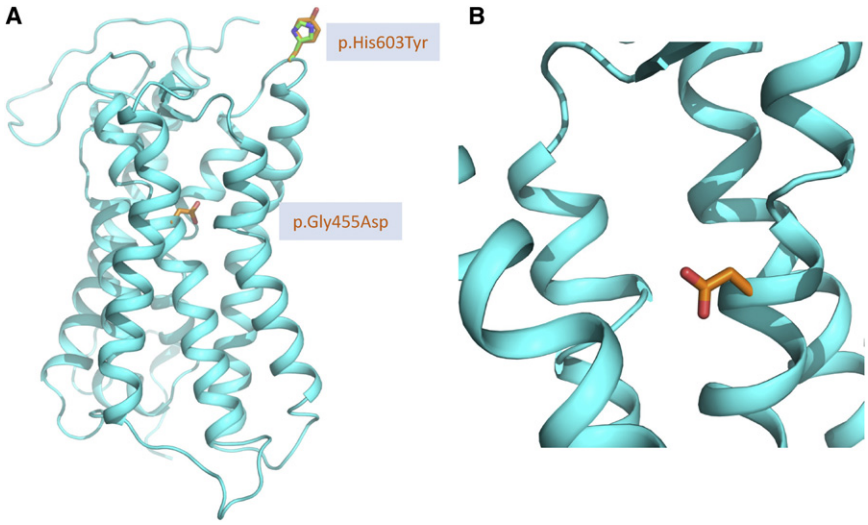


Figure 2. 3D Model of the Transmembrane Region of GPR179
(A) 3D homology model based on the 3D model of the wild-type squid rhodopsin (2ZIY). Wild-type Gly455 and His603 residues are indicated in green and the mutated Asp455 and Tyr603 in orange.
(B) Superimposition of 3D models of the wild-type residues and the Gly455Asp alteration (aspartate in orange).

terminus of such proteins is important for ligand binding, and thus the p.Asp126His mutation might be associated with loss of this binding. On the other hand, the amino acids that are mutated in our patients also might be important for structural properties of the protein in the endoplasmic reticulum. Thus a misfolded protein is likely to be excluded from the strictly regulated transport to the membrane. Similar findings were observed for mutations in *GRM6*, identified in cCSNB patients. Mutated metabotropic glutamate receptor 6 could not reach the membrane, leading to cCSNB.[29] Further functional analysis of the mutant variants is needed to determine whether these mutations lead to downregulation of the GPR179 transcript, trafficking problems, abolishment of ligand binding, or interactions with other proteins involved in signal transmission from photoreceptors to the adjacent bipolar cells.

As mentioned above, little is known about the structure and function of GPR179. However, an EST profile is available (Unigene database) and shows a restricted pattern of expression in the human eye, heart, and brain. Furthermore, transcriptomic data of whole retina from *rd1* mice revealed increased expression of *GPR179* compared to the expression in wild-type mice starting from postnatal day 12 (Figure 3). The *rd1* mouse, carrying *Pde6b* mutations, is a naturally occurring model with progressive rod photoreceptor degeneration, leading to a complete loss of all rods by post natal day 36, and preserved inner retina.[30] This would suggest that *GPR179* is expressed in the inner nuclear layer of the retina. Interestingly, *Nyx*, another gene with mutations leading to cCSNB, shows a similar expression profile in the *rd1* mouse (Figure 3).

Real-time PCR experiments with two different primer sets (Table S4) confirmed the expression of *GPR179* in human retina (commercially available cDNA from Clontech, Saint-Germain-en-Laye, France), giving a signal of $\Delta C_T = 13.46$ ($C_{TGPR179} = 30.03$) in relation to beta-actin (*ACTB* [MIM 102630]) ($C_{T\ ACTB} = 16.57$) (primers Table S4). Sanger sequencing of the amplified RT-PCR products

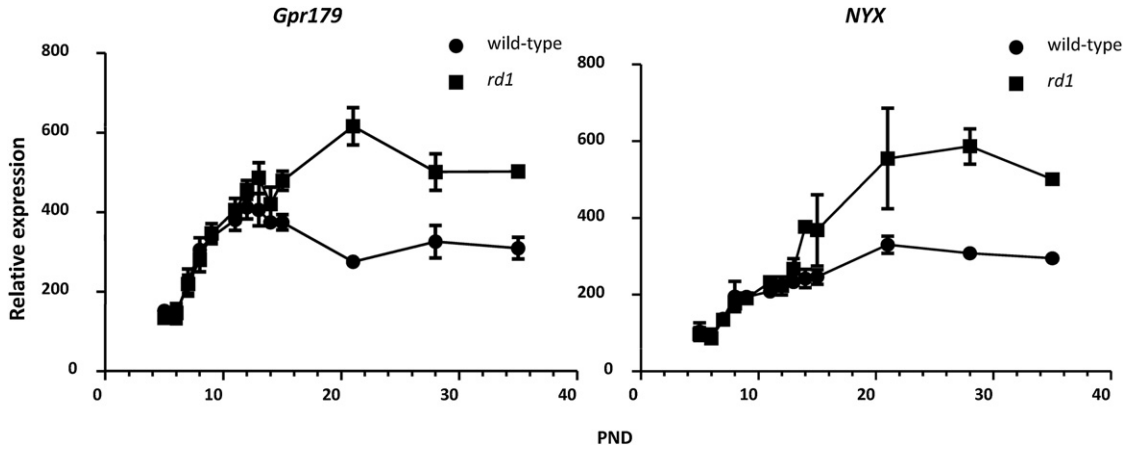


Figure 3. Indirect Expression Analysis of Gpr179 in rd1 and Wild-Type Mice
Expression of *Gpr179* (1459268_at) compared to the expression of *Nyx* (1446344_at, a known molecule expressed in the inner nuclear layer, also implicated in cCSNB) during rod degeneration in the *rd1* mouse. Neural retinas from *rd1* and wild-type mice on identical genetic backgrounds[47] were hybridized to the mouse genome 430 2.0 array (Affymetrix, High Wycombe, UK). The expression profiles are similar from postnatal day (PND) 5 to PND12. Thereafter, the expression of *Gpr179*, as well as *Nyx*, increases in the *rd1* retina. This phenomenon correlates temporally with the loss of rod photoreceptor cells and is likely due to the unaffected inner retinal cells in the *rd1* specimen at this age.

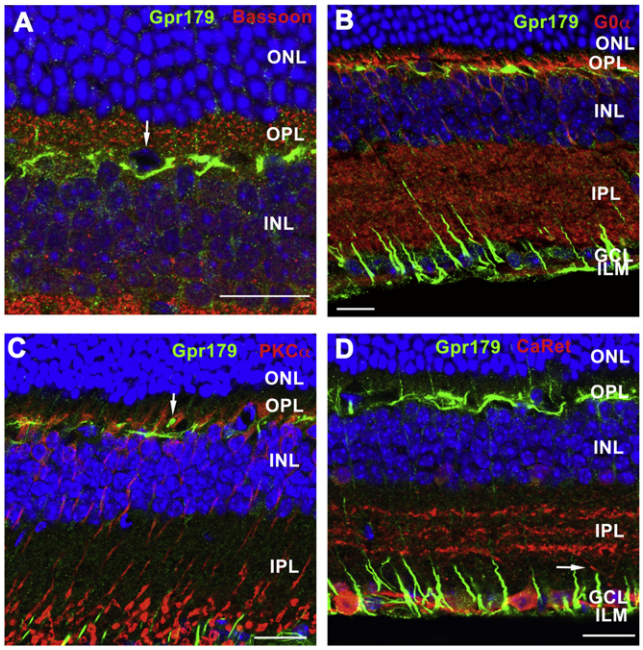


Figure 4. Gpr179 Immunohistochemistry on Retinal Sections of Wild-Type Mice

GPR179 signal (green) in the OPL and ILM double labeled with other retinal markers (red) were detected in wild-type mice by confocal microscopy. The scale bar represents 20 μm. The following abbreviations are used: ONL, outer nuclear layer; OPL, outer plexiform layer; INL, inner nuclear layer; GCL, ganglion cell layer; and ILM, inner limiting membrane. Gpr179 did not colocalize with the presynaptic marker Basoon (A), the ON-bipolar cell markers Goα (B) and PKCα (C), or with the ganglion cell labeled with calretinin (D). Some cells in the upper part of INL were surrounded by specific Gpr179 staining (arrow in A). The shape and the localization might indicate that these cells represent horizontal cells. Bipolar cell dendrites, stained with PKCα seem to surround Gpr179 (arrow in C). Calretinin antibody labeled specifically the ganglion cells, and their dendrites did not colocalize with Gpr179 (arrow in D). Instead, it seems that Gpr179 is highly concentrated in Müller cell endfeet (D); similar results have been previously shown for a potassium channel Kir4.1.[31]

from retinas confirmed the presence of the *GPR179* transcript. Using the same conditions, we could not detect the transcript in lymphocytes or HEK293 cells.

We investigated the localization of the Gpr179 protein in adult mouse retina by immunostaining coronal eye cryosections with a rabbit polyclonal antibody directed against GPR179 (Sigma-Aldrich, Saint-Quentin Fallavier, France). Bound primary antibody was detected with a secondary antibody (Alexa Fluor 488-conjugated, Invitrogen, Courtaboeuf, France), and the nuclei were counterstained (4′,6-diamidino-2-phenylindole [DAPI], Euromedex, Souffelweyersheim, France). Immunofluorescence was analyzed with a confocal microscope (FV1000 fluorescent, Olympus, Hamburg, Germany). Gpr179 expression could be detected in the outer plexiform layer (OPL) and in the inner limiting membrane (ILM), in close proximity to the ganglion cells (Figure 4, green). Colocalization studies with a mouse anti-Bassoon (Enzo Lifescienes, Lyon, France), a specific marker for ribbon synapse, excluded a close vicinity between Gpr179 and presynaptic terminals (Figure 4A). Furthermore, immunostaining with mouse antibodies against Goα (Millipore, Molsheim, France) and PKCα (Sigma-Aldrich), two specific ON-bipolar markers, demonstrated the absence of colocalization of Gpr179 with these proteins (Figures 4B and 4C). Instead, Gpr179 appears to be localized in a distinct compartment within bipolar cells or in other cells, such as horizontal cells (indicated by the arrow in Figure 4A). Interestingly, bipolar cell dendrites, stained with PKCα, seem to surround Gpr179 (indicated by the arrow in Figure 4C). Alternatively, the Gpr179 OPL staining could also be localized within Müller cell processes present within this layer. In addition, a mouse antibody against calretinin (Millipore), a specific marker for ganglion cells and their dendrites, was used and did not show colocalization with Gpr179 immunostaining (Figure 4D, an example of a ganglion cell dendrite is marked with an arrow). Instead, Gpr179 was highly expressed in Müller cell endfeet at the level of the ILM; similar results had previously been shown for the potassium channel Kir4 .1 macromolecular complex.[31–33] Therefore, immunolocalization of Gpr179 suggests its localization in the OPL either in bipolar cells in a cellular compartment distinct from the synaptic membrane and cell body, and/or in horizontal cells, and/or in Müller cell processes as well as within the Müller cell endfeet.

The OPL localization of Gpr179 and the same associated ON-bipolar dysfunction phenotype as for Grm6, Nyx, or Trpm1 alterations[34–38] would suggest that *GPR179* is part of the same transduction pathway and could directly interact with any of these proteins. However, immunolocalization studies are not in keeping with this hypothesis. Instead, immunostaining suggests Müller cell localization and could place the Gpr179 functional role within these cells, possibly through the Kir4.1 macromolecular complex. This complex was shown to involve at least the potassium channel Kir4.1, the water channel aquaporin-4 (AQP4), and the dystrophin isoform Dp71. Interestingly, although Kir4.1 and Aqp4 knockout mice do not show Schubert- Bornschein ERG abnormalities,[39,40] a subset of dystrophin mutations, responsible for Duchenne muscular dystrophy (DMD [MIM 310200]) are associated with such ERG abnormalities.[41–44] Therefore, one hypothesis would be that Gpr179 is part of the Kir4.1 macromolecular complex. Gpr179 might directly interact with dystrophin isoforms, and its dysfunction would lead to cCSNB in a similar mechanism as in DMD. In order to reconcile the Grm6/Nyx/Trpm1-signaling pathway within bipolar cells and Gpr179 within Muller cells, one might hypothesize that Gpr179 could be involved in an as-yet unknown interaction between ON-bipolar cells and Müller cells that would be essential for ON-bipolar cell depolarization resulting in b-wave formation. On the other hand, our immunostaining studies could also suggest specific localization of Gpr179 within horizontal cells. Therefore, another hypothesis could be that ON-bipolar cells directly interact with horizontal cells. Lack of this interaction due

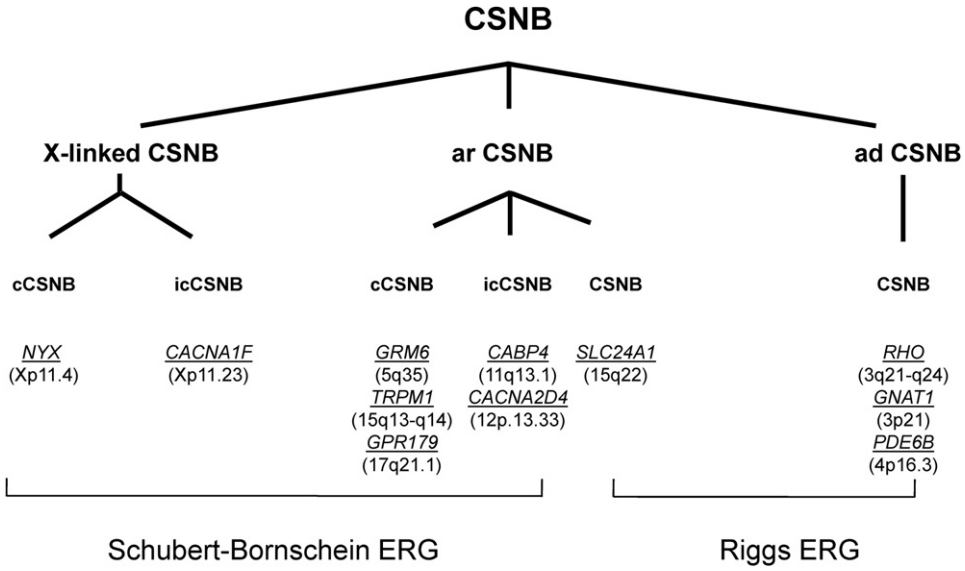


Figure 5. Genes Underlying CSNB
Different forms of CSNB in human are classified according to their electroretinographic feature, mode of inheritance, clinical phenotype, and mutated genes. Patients discussed herein show a complete Schubert-Bornschein type of ERG. The following abbreviations are used: cCSNB, complete CSNB; icCSNB, incomplete CSNB; ar, autosomal recessive; ad, autosomal dominant. Genes are indicated in italics and underlined. Chromosomal location is given between brackets. The phenotype of patients with mutations in icCSNB is more variable and can even lead to progressive cone or cone-rod dystrophy.[1]

to Gpr179 dysfunction could lead to the reduced b-wave observed in patients with cCSNB.

Including the current study, mutations in four genes (*NYX*, *GRM6, TRPM1*, and *GPR179*) have been implicated in cCSNB (Figure 5).[6,7,10–13,45,46] These genes code for nyctalopin, metabotropic glutamate receptor 6, transient receptor potential cation melastatin 1 channel, and G protein-coupled receptor 179, respectively. All but GPR179 localize postsynaptically to the photoreceptors in the retina in ON-bipolar cells.[35] Further functional studies will eventually clarify the exact role of this novel protein within the ON-bipolar cells pathway, which will also improve our understanding of the overall visual signal transduction through the retina.

Supplemental Data

Supplemental Data include two figures and four tables and can be found with this article online at http://www.cell.com/AJHG/.

Acknowledgments

The authors are grateful to the families described in this study; to Dominique Santiard-Baron and Christine Chaumeil for their help in DNA collection; to the clinical staff; and to Olivier Goreau, Serge Picaud, and Alvaro Rendon for antibodies used in the colocalization experiments. The project was supported by GIS-maladies rares (C.Z.), Retina France ([part of the 100-Exome Project] I.A., C.P.H., J.-A.S., H.D. and C.Z.), Foundation Voir et Entendre (C.Z.), Agence National de la Recherche (S.S.B), Foundation Fighting Blindness (FFB) grant CD-CL-0808-0466-CHNO (I.A. and the CIC503, recognized as an FFB center), FFB grant C-CMM-0907-0428-INSERM04, Ville de Paris and Région Ille de France, the French Association against Myopathy (AFM) grant KBM-14390 (O.P.), and National Institutes of Health grant 1R01EY020902-01A1 (K.B.).

Received: October 20, 2011
Revised: November 18, 2011
Accepted: December 8, 2011
Published online: February 9, 2012

Web Resources

The URLs for data presented herein are as follows:

ESEfinder, http://rulai.cshl.edu/cgi-bin/tools/ESE3/esefinder.cgi?process=home
GenCards, http://www.genecards.org/
PolyPhen, http://genetics.bwh.harvard.edu/pph/
Sm2ph Central, http://decrypthon.igbmc.fr/sm2ph/cgi-bin/home
National Center for Biotechnology Information, http://ncbi.nlm.nih.gov/
Online Mendelian Inheritance in Man (OMIM), http://www.omim.org
Pymol, http://pymol.org/
SIFT (Sorting Intolerant From Tolerant), http://blocks.fhcrc.org/sift/SIFT.html
UniProtKB/Swiss-Prot, http://www.uniprot.org
USCS Human Genome Browser, http://genome.ucsc.edu/

References

1. Zeitz, C. (2007). Molecular genetics and protein function involved in nocturnal vision. Expert Rev. Ophthalmol. *2*, 467–485.
2. Schubert, G., and Bornschein, H. (1952). Analysis of the human electroretinogram.Ophthalmologica *123*, 396–413.

3. Miyake, Y., Yagasaki, K., Horiguchi, M., Kawase, Y., and Kanda, T. (1986). Congenital stationary night blindness with negative electroretinogram. A new classification. Arch. Ophthalmol. *104*, 1013–1020.
4. Audo, I., Robson, A.G., Holder, G.E., and Moore, A.T. (2008). The negative ERG: Clinical phenotypes and disease mechanisms of inner retinal dysfunction. Surv. Ophthalmol. *53*, 16–40.
5. Zeitz, C., Labs, S., Lorenz, B., Forster, U., Uksti, J., Kroes, H.Y., De Baere, E., Leroy, B.P., Cremers, F.P., Wittmer, M., et al. (2009). Genotyping microarray for CSNB-associated genes. Invest. Ophthalmol. Vis. Sci. *50*, 5919–5926.
6. Dryja, T.P., McGee, T.L., Berson, E.L., Fishman, G.A., Sandberg, M.A., Alexander, K.R., Derlacki, D.J., and Rajagopalan, A.S. (2005). Night blindness and abnormal cone electroretinogram ON responses in patients with mutations in the GRM6 gene encoding mGluR6. Proc. Natl. Acad. Sci. USA *102*, 4884–4889.
7. Zeitz, C., van Genderen, M., Neidhardt, J., Luhmann, U.F., Hoeben, F., Forster, U., Wycisk, K., Mátyás, G., Hoyng, C.B., Riemslag, F., et al. (2005). Mutations in GRM6 cause autosomal recessive congenital stationary night blindness with a distinctive scotopic 15-Hz flicker electroretinogram. Invest. Ophthalmol. Vis. Sci. *46*, 4328–4335.
8. Zeitz, C., Kloeckener-Gruissem, B., Forster, U., Kohl, S., Magyar, I., Wissinger, B., Mátyás, G., Borruat, F.X., Schorderet, D.F., Zrenner, E., et al. (2006). Mutations in CABP4, the gene encoding the Ca2+-binding protein 4, cause autosomal recessive night blindness. Am. J. Hum. Genet. *79*, 657–667.
9. Wycisk, K.A., Zeitz, C., Feil, S., Wittmer, M., Forster, U., Neidhardt, J., Wissinger, B., Zrenner, E., Wilke, R., Kohl, S., and Berger, W. (2006). Mutation in the auxiliary calcium-channel subunit CACNA2D4 causes autosomal recessive cone dystrophy. Am. J. Hum. Genet. *79*, 973–977.
10. Audo, I., Kohl, S., Leroy, B.P., Munier, F.L., Guillonneau, X., Mohand-Saïd, S., Bujakowska, K., Nandrot, E.F., Lorenz, B., Preising, M., et al. (2009). TRPM1 is mutated in patients with autosomal-recessive complete congenital stationary night blindness. Am. J. Hum. Genet. *85*, 720–729.
11. Li, Z., Sergouniotis, P.I., Michaelides, M., Mackay, D.S., Wright, G.A., Devery, S., Moore, A.T., Holder, G.E., Robson, A.G., and Webster, A.R. (2009). Recessive mutations of the gene TRPM1 abrogate ON bipolar cell function and cause complete congenital stationary night blindness in humans. Am. J. Hum. Genet. *85*, 711–719.
12. van Genderen, M.M., Bijveld, M.M., Claassen, Y.B., Florijn, R.J., Pearring, J.N., Meire, F.M., McCall, M.A., Riemslag, F.C., Gregg, R.G., Bergen, A.A., and Kamermans, M. (2009). Mutations in TRPM1 are a common cause of complete congenital stationary night blindness. Am. J. Hum. Genet. *85*, 730–736.
13. Nakamura, M., Sanuki, R., Yasuma, T.R., Onishi, A., Nishiguchi, K.M., Koike, C., Kadowaki, M., Kondo, M., Miyake, Y., and Furukawa, T. (2010). TRPM1 mutations are associated with the complete form of congenital stationary night blindness. Mol. Vis. *16*, 425–437.
14. Sergouniotis, P.I., Davidson, A.E., Mackay, D.S., Li, Z., Yang, X., Plagnol, V., Moore, A.T., and Webster, A.R. (2011). Recessive mutations in KCNJ13, encoding an inwardly rectifying potassium channel subunit, cause leber congenital amaurosis. Am. J. Hum. Genet. *89*, 183–190.
15. Bamshad, M.J., Ng, S.B., Bigham, A.W., Tabor, H.K., Emond, M.J., Nickerson, D.A., and Shendure, J. (2011). Exome sequencing as a tool for Mendelian disease gene discovery. Nat. Rev. Genet. *12*, 745–755.
16. Marmor, M.F., Fulton, A.B., Holder, G.E., Miyake, Y., Brigell, M., and Bach, M.; International Society for Clinical Electrophysiology of Vision. (2009). ISCEV Standard for full-field clinical electroretinography (2008 update). Doc. Ophthalmol. *118*, 69–77.
17. UniProt Consortium. (2010). The Universal Protein Resource (UniProt) in 2010. Nucleic Acids Res. *38* (Database issue), D142–D148.
18. Altschul, S.F., Madden, T.L., Schäffer, 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.
19. 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.
20. Eswar, N., Eramian, D., Webb, B., Shen, M.Y., and Sali, A. (2008). Protein structure modeling with MODELLER. Methods Mol. Biol. *426*, 145–159.
21. Eramian, D., Eswar, N., Shen, M.Y., and Sali, A. (2008). How well can the accuracy of comparative protein structure models be predicted? Protein Sci. *17*, 1881–1893.
22. Nishiguchi, K.M., Sandberg, M.A., Kooijman, A.C., Martemyanov, K.A., Pott, J.W., Hagstrom, S.A., Arshavsky, V.Y., Berson, E.L., and Dryja, T.P. (2004). Defects in RGS9 or its anchor protein R9AP in patients with slow photoreceptor deactivation. Nature *427*, 75–78.
23. Dvir, L., Srour, G., Abu-Ras, R., Miller, B., Shalev, S.A., and Ben-Yosef, T. (2010). Autosomal-recessive early-onset retinitis pigmentosa caused by a mutation in PDE6G, the gene encoding the gamma subunit of rod cGMP phosphodiesterase. Am. J. Hum. Genet. *87*, 258–264.
24. Piri, N., Gao, Y.Q., Danciger, M., Mendoza, E., Fishman, G.A., and Farber, D.B. (2005). A substitution of G to C in the cone cGMP-phosphodiesterase gamma subunit gene found in a distinctive form of cone dystrophy. Ophthalmology *112*, 159–166.
25. Lagerström, M.C., and Schiöth, H.B. (2008). Structural diversity of G protein-coupled receptors and significance for drug discovery. Nat. Rev. Drug Discov. *7*, 339–357.
26. Ng, P.C., and Henikoff, S. (2001). Predicting deleterious amino acid substitutions. Genome Res. *11*, 863–874.
27. Luu, T., Nguyen, N., Friedrich, A., Muller, J., Moulinier, L., and Poch, O. (2011). Extracting knowledge from a mutation database related to human monogenic disease using inductive logic programming. International Conference on Bioscience, Biochemistry and Bioinformatics; Singapore. IACSIT *5*, 83–100.
28. Friedrich, A., Garnier, N., Gagnière, N., Nguyen, H., Albou, L.P., Biancalana, V., Bettler, E., Deléage, G., Lecompte, O., Muller, J., et al. (2010). SM2PH-db: An interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases. Hum. Mutat. *31*, 127–135.
29. Zeitz, C., Forster, U., Neidhardt, J., Feil, S., Kälin, S., Leifert, D., Flor, P.J., and Berger, W. (2007). Night blindness-associated mutations in the ligand-binding, cysteine-rich, and intracellular domains of the metabotropic glutamate receptor 6 abolish protein trafficking. Hum. Mutat. *28*, 771–780.
30. Carter-Dawson, L.D., LaVail, M.M., and Sidman, R.L. (1978). Differential effect of the rd mutation on rods and cones in the mouse retina. Invest. Ophthalmol. Vis. Sci. *17*, 489–498.

31. Connors, N.C., and Kofuji, P. (2006). Potassium channel Kir4.1 macromolecular complex in retinal glial cells. Glia *53*, 124–131.
32. Claudepierre, T., Rodius, F., Frasson, M., Fontaine, V., Picaud, S., Dreyfus, H., Mornet, D., and Rendon, A. (1999). Differential distribution of dystrophins in rat retina. Invest. Ophthalmol. Vis. Sci. *40*, 1520–1529.
33. Dalloz, C., Sarig, R., Fort, P., Yaffe, D., Bordais, A., Pannicke, T., Grosche, J., Mornet, D., Reichenbach, A., Sahel, J., et al. (2003). Targeted inactivation of dystrophin gene product Dp71: Phenotypic impact in mouse retina. Hum. Mol. Genet. *12*, 1543–1554.
34. Vardi, N., Duvoisin, R., Wu, G., and Sterling, P. (2000). Localization of mGluR6 to dendrites of ON bipolar cells in primate retina. J. Comp. Neurol. *423*, 402–412.
35. Morgans, C.W., Ren, G., and Akileswaran, L. (2006). Localization of nyctalopin in the mammalian retina. Eur. J. Neurosci. *23*, 1163–1171.
36. Gregg, R.G., Kamermans, M., Klooster, J., Lukasiewicz, P.D., Peachey, N.S., Vessey, K.A., and McCall, M.A. (2007). Nyctalopin expression in retinal bipolar cells restores visual function in a mouse model of complete X-linked congenital stationary night blindness. J. Neurophysiol. *98*, 3023–3033.
37. Morgans, C.W., Zhang, J., Jeffrey, B.G., Nelson, S.M., Burke, N.S., Duvoisin, R.M., and Brown, R.L. (2009). TRPM1 is required for the depolarizing light response in retinal ON-bipolar cells. Proc. Natl. Acad. Sci. USA *106*, 19174–19178.
38. Koike, C., Obara, T., Uriu, Y., Numata, T., Sanuki, R., Miyata, K., Koyasu, T., Ueno, S., Funabiki, K., Tani, A., et al. (2010). TRPM1 is a component of the retinal ON bipolar cell transduction channel in the mGluR6 cascade. Proc. Natl. Acad. Sci. USA *107*, 332–337.
39. Kofuji, P., Ceelen, P., Zahs, K.R., Surbeck, L.W., Lester, H.A., and Newman, E.A. (2000). Genetic inactivation of an inwardly rectifying potassium channel (Kir4.1 subunit) in mice: Phenotypic impact in retina. J. Neurosci. *20*, 5733–5740.
40. Li, J., Patil, R.V., and Verkman, A.S. (2002). Mildly abnormal retinal function in transgenic mice without Müller cell aquaporin-4 water channels. Invest. Ophthalmol. Vis. Sci. *43*, 573–579.
41. Cibis, G.W., Fitzgerald, K.M., Harris, D.J., Rothberg, P.G., and Rupani, M. (1993). The effects of dystrophin gene mutations on the ERG in mice and humans. Invest. Ophthalmol. Vis. Sci. *34*, 3646–3652.
42. De Becker, I., Riddell, D.C., Dooley, J.M., and Tremblay, F. (1994). Correlation between electroretinogram findings and molecular analysis in the Duchenne muscular dystrophy phenotype. Br. J. Ophthalmol. *78*, 719–722.
43. Pillers, D.A., Bulman, D.E., Weleber, R.G., Sigesmund, D.A., Musarella, M.A., Powell, B.R., Murphey, W.H., Westall, C., Panton, C., Becker, L.E., et al. (1993). Dystrophin expression in the human retina is required for normal function as defined by electroretinography. Nat. Genet. *4*, 82–86.
44. Pillers, D.A., Fitzgerald, K.M., Duncan, N.M., Rash, S.M., White, R.A., Dwinnell, S.J., Powell, B.R., Schnur, R.E., Ray, P.N., Cibis, G.W., and Weleber, R.G. (1999). Duchenne/Becker muscular dystrophy: Correlation of phenotype by electroretinography with sites of dystrophin mutations. Hum. Genet. *105*, 2–9.
45. Bech-Hansen, N.T., Naylor, M.J., Maybaum, T.A., Sparkes, R.L., Koop, B., Birch, D.G., Bergen, A.A., Prinsen, C.F., Polomeno, R.C., Gal, A., et al. (2000). Mutations in NYX, encoding the leucine-rich proteoglycan nyctalopin, cause X-linked complete congenital stationary night blindness. Nat. Genet. *26*, 319–323.
46. Pusch, C.M., Zeitz, C., Brandau, O., Pesch, K., Achatz, H., Feil, S., Scharfe, C., Maurer, J., Jacobi, F.K., Pinckers, A., et al. (2000). The complete form of X-linked congenital stationary night blindness is caused by mutations in a gene encoding a leucine-rich repeat protein. Nat. Genet. *26*, 324–327.
47. Viczian, A., Sanyal, S., Toffenetti, J., Chader, G.J., and Farber, D.B. (1992). Photoreceptor-specific mRNAs in mice carrying different allelic combinations at the rd and rds loci. Exp. Eye Res. *54*, 853–860.

9.3 Conclusions et perspectives

Comme cela a été noté dans notre manuscrit, cette étude, la première publiée sur le gène GPR179, ouvre la voie à diverses recherches sur sa fonction potentielle au sein des cellules rétiniennes. Ces travaux devraient conduire à une meilleure compréhension des causes de la Cécité nocturne stationnaire congénitale et plus globalement, des mécanismes de traitement de l'information visuelle par les différents types cellulaires.

D'un point de vue plus pratique, nous avons pu mettre à profit cette étude pour améliorer nos outils et observer comment des utilisateurs non-experts se les appropriaient. Ceci a ainsi permis de vérifier que les informations fournies par SM2PH Central (séquence, structure, fonction, évolution, réseaux biologiques…) sont rapidement et aisément questionnées par les biologistes, et qu'ils apprécient de ne pas avoir à s'adresser à chaque banque individuelle telle que UniProt, NCBI, STRING, KEGG, SCOP, etc. Comme cela a été souligné dans la section introductive de ce chapitre, l'alignement automatique n'était pas de très bonne qualité, ce qui a nécessité des ajustements manuels et une validation par un expert humain. A l'heure actuelle, cette édition de l'alignement est effectuée par un outil, SeqLab, externe à SM2PH Central. Ceci a révélé les limites de notre infrastructure, et notamment le fait qu'il n'avait pas été prévu de ne relancer que la partie pertinente du pipeline d'annotation et de modélisation. Ces aspects, partiellement traités durant ce projet, impliquent que la prochaine version de SM2PH Central devra disposer de son éditeur d'alignement afin de faciliter les corrections à la volée sans obligatoirement entrainer le calcul de l'ensemble des paramètres et modèles. De même, le devenir de données modifiées par un utilisateur dans une instance (en l'occurrence un alignement) reste également à préciser et à développer. En effet, ces données validées par un expert pourraient améliorer la qualité de SM2PH Central et profiter à l'ensemble de la communauté. Cependant, pour l'instant, aucun élément permettant de réaliser cette communication entre les 2 bases n'a été développé et les problèmes d'autorisation, de modalités, de définition des utilisateurs autorisés à modifier SM2PH Central… restent largement à définir.

Un autre élément est apparu, à savoir que SM2PH Central ne disposait pas de données d'expression sur tous les tissus humains, et notamment sur les tissus d'origine oculaire. Ceci est largement lié au fait que très peu de données sont disponibles pour l'œil humain car ce tissu, très symbolique, est rarement disponible que ce soit à partir d'un donneur sain, malade ou mort. Dès lors, nous n'avons pu réaliser d'analyse approfondie des données transcriptomiques. Cela nous a amené à prévoir d'intégrer, à l'avenir, des données transcriptomiques provenant d'autres organismes notamment de souris ou de rats et surtout, d'optimiser la création et la suppression rapides de liens entre SM2PH Central et la base de données transcriptomiques, GxDB. Par ailleurs, SM2PH Central n'a pas intégré d'informations relatives au contexte génomique du gène GPR179. Cet aspect devrait être réglé par l'intégration de GecoDB dans SM2PH Central. De plus, nous avons également constaté que la recherche rapide de petites séquences plus ou moins dévoyées (motifs) est importante. En effet, ces motifs donnent parfois un indice sur la fonction des protéines. La fonction *Motif Search* de SM2PH Central, introduite dans la section 5.6.2, a été implémentée pendant l'étude du gène GPR179 afin de vérifier, en temps réel et rapidement, si le motif CPW(E,D) était spécifique de la famille de protéines de GPR179 et d'analyser les fréquences des motifs équivalents au sein du protéome humain. A l'avenir, cette fonction très utile sera étendue aux protéomes d'autres organismes afin de faciliter des analyses croisées. Enfin, pendant la création du SM2PH-GPR179, il est apparu que les biologistes souhaitaient pouvoir intégrer

leurs propres données telles que des images par microscopie optique, des données phénotypiques ou cliniques. La prochaine version des instances SM2PH devra être en mesure d'implémenter cette fonction de façon conviviale et robuste.

Finalement, la combinaison entre MSV3d et KD4V a permis aux biologistes de caractériser et d'analyser les mutations dans le détail en exploitant les règles PLI. L'environnement convivial et intuitif de visualisation de ces résultats a abouti à une réelle participation des biologistes qui ont pu confirmer les résultats obtenus et compléter les informations de prédiction du caractère délétère/neutre des mutations.

CONCLUSIONS & PERSPECTIVES

Répondant au besoin de mieux comprendre les relations qui lient un génotype aux phénotypes moléculaires et cliniques associés, nous avons développé une nouvelle infrastructure bioinformatique qui unit, dans un même système, la collecte, la gestion, la maintenance et le traitement de multiples données ou informations ainsi que l'accès et l'exploitation de puissance de calculs locale ou délocalisée et d'algorithmes connus ou originaux.

La première contribution de cette thèse est SM2PH Central et sa capacité de générer des instances. SM2PH Central constitue notre centre de référence en ligne pour l'ensemble des protéines humaines intégrant des niveaux d'informations qui vont des aspects génomiques, structuraux, fonctionnels ou évolutifs aux aspects de transcriptomique, interactomique, protéomique ou métabolomique.

Dans ce contexte de système intégratif, la mise à jour des données est un point essentiel. Face à l'hétérogénéité, au volume et au débit des données générées par les biotechnologies à haut débit, la rapidité et la synchronisation sont des facteurs déterminants de toute infrastructure moderne. Le système BIRD, au cœur de SM2PH Central, nous a permis de relever ce défi pour un nombre imposant de données (voir les statistiques dans la section 5.4) et a abouti à ce que l'ensemble des données collectées et informations générées dans SM2PH Central soit automatiquement mis à jour en seulement 2 semaines. De plus, grâce à son architecture orientée service, SM2PH Central peut effectuer un service pour ne mettre à jour qu'une partie des données. Pour ses multiples avantages, cette architecture orientée service s'est imposée dans de nombreux domaines scientifiques, et notamment en biologie, où l'on observe une forte hétérogénéité des mises à jour entre banques de données (allant de quelques heures, jours ou semaines pour des données de séquences ou de structures à des mois, voire des années, pour des banques plus élaborées telles que STRING ou KEGG). Cette problématique de la synchronisation se retrouve au niveau des algorithmes et programmes implémentés dans SM2PH Central qui sont fréquemment modifiés ou supplantés par de nouvelles versions ou des programmes plus performants. Ceci aboutit souvent à des modifications des formats d'entrée ou de sortie qui peuvent s'avérer hautement pénalisant et entrainer de profonds remaniements au sein de banques intégratives telles que SM2PH Central. Là encore, notre architecture modulaire et le choix de BIRD comme système de gestion de données hautement configurable, nous a permis d'être réactifs et robustes et nous amène à penser que nous avons développé un système pérenne qui propose une solution prometteuse, applicable à d'autres champs de la biologie moderne.

La possibilité de générer à façon des sous-banques spécialisées ou SM2PH-Instances, participe de ce constat. Au-delà des aspects techniques et informatiques que la création d'instances implique, cette réalisation s'inscrit dans une ambition qui traverse toute la bioinformatique et la biologie à savoir, être en mesure de mobiliser, « d'agglomérer » et de mettre à disposition dynamiquement de vastes ressources de données ou de moyens informatiques, complexes et hétérogènes pour la résolution de problèmes précis. En l'état, nous avons fourni un début de solution qui permet de générer en 1 ou 2 heures, par simple courrier électronique (à l'heure actuel) ou par clic de souris (dans l'avenir proche), un système complet permettant à des experts humains d'analyser, par exemple, la relation génotype-phénotype d'un groupe restreint de gènes liés à une maladie définie. Cependant, il est clair que le problème est loin

d'être résolu et qu'il nous faudra encore enrichir notre système pour améliorer, entre autres, ses capacités d'extension à d'autres données ou programmes, sa connectivité, sa souplesse d'utilisation, etc… pour être en mesure, un jour, de pleinement profiter des compétences et de l'expertise couplées de bioinformaticiens et de biologistes.

La deuxième contribution directe de cette thèse est MSV3d comme une ressource d'annotation multi-niveau (propriétés physico-chimiques, fonction, évolution, structure) d'une mutation. Comme nous l'avons vu, MSV3d exploite SM2PH Central, et en ce sens MSV3d a été le premier « client » de SM2PH Central qui nous a permis de préciser comment lui faire jouer son rôle de distributeur universel d'informations complexes. En l'état, MSV3d englobe l'ensemble des 445 574 mutations humaines connues et son service d'annotation peut, à priori, caractériser automatiquement toute mutation inconnue. Fidèle à notre philosophie de création de systèmes autonomes interconnectés, MSV3d fournit l'ensemble des connaissances exploitées par la troisième contribution de cette thèse à savoir KD4v, notre base d'extraction de connaissances pour prédire l'impact phénotypique d'une mutation. Là encore, cette architecture nous permet de faire évoluer indépendamment chaque système. Ceci est essentiel dans le domaine de l'extraction de connaissances, où les nouveautés algorithmiques sont incessantes et traversent quasiment l'ensemble des champs de l'informatique depuis les statistiques et méthodes de classification jusqu'aux bases de données ou de connaissances.

KD4v a ouvert une nouvelle page dans l'étude de l'impact phénotypique des mutations, grâce à l'introduction de la méthode de Programmation Logique Induction et à la création de sa propre base de connaissances pour caractériser les mutations neutres ou délétères. Dans KD4v, les connaissances sont représentées sous forme de règles déduites automatiquement à partir des 8 000 mutations structurales connues et l'ensemble peut être aisément mis à jour quand le nombre des mutations disponibles augmente. Comme nous l'avons vu, les performances de prédiction sont très honorables et tout à fait comparables aux meilleurs outils du moment. Cependant, à nos yeux, le point le plus important est que ces règles soient exploitables par la machine et par l'homme. Ceci confère à notre système, la possibilité d'inclure activement les compétences et expertises des biologistes dans l'analyse et à terme, d'introduire dans les études des corrélations génotype-phénotype, des connaissances encore inconnues ou trop spécialisées.

Enfin, l'ultime contribution de cette thèse est liée au développement de GEPeTTO, un prototype de priorisation de gènes. Ce prototype est actuellement en cours de validation en utilisant le jeu de données confidentielles liés à la dégénérescence maculaire liée à l'âge (DMLA). Dans ce domaine très compétitif où les innovations algorithmiques et méthodologiques sont également très fréquentes, nous avons délibérément décidé de privilégier, en premier lieu, l'intégration de nouvelles données biologiques (EvoluCode ou GecoDB) en complément de celles déjà utilisées dans les meilleurs systèmes de priorisation. Cette approche nous a permis d'obtenir rapidement des performances très encourageantes mais surtout, cela nous a permis de vérifier que, là encore, l'une des clés réside dans la capacité d'agglomérer et d'exploiter simultanément de multiples niveaux de données, informations et connaissances biologiques.

A l'heure du constat final de nos réalisations, il est hélas clair que de nombreuses améliorations restent à réaliser pour atteindre l'objectif ambitieux d'un système adressant profondément l'étude du lien génotype-phénotype. Nous avons déjà décrit dans les chapitres de conclusion des différentes sections, certaines de ces améliorations en cours ou prévues. On peut schématiquement les réunir autour de plusieurs axes qui concernent :

i. les données ou informations à identifier et à ajouter dans SM2PH Central, incluant de nouvelles définitions de mutations (non-sens, indel, recombinaison…), de

nouvelles expériences (transcriptomes, métabolomes ou protéomes) ou de nouveaux organismes modèles. Un problème particulier concerne les informations disponibles dans les résumés bibliographiques qui sont susceptibles de renseigner sur de multiples champs mais qui impliqueront de choisir et d'implémenter des méthodes de fouille de textes efficaces et robustes. Cet aspect a été rapidement abordé dans le cadre de notre prototype de priorisation des gènes, ce qui nous a permis de vérifier que nous pouvions gérer et questionner de gros volumes de données dans BIRD (près de 20.494.848 résumés). Cependant, il semble que l'extraction efficace et non biaisée d'informations textuelles qui apporterait énormément à notre système est un problème très complexe qui s'inscrit dans un domaine de recherche en pleine effervescence. En effet, il est nécessaire de dépasser le simple décompte statistique des mots, citations ou références croisées pour prendre en compte de multiples aspects tels que i) le coté « médiatique » d'un domaine biologique (par exemple, le nombre, la richesse et les citations trans-domaines dans des maladies telle que le cancer ou l'Alzheimer ne sont pas comparables aux données liées à une maladie rare ou émergente) ; ii) l'importance des champs méthodologiques « modernes » par rapport aux outils spécifiques d'un domaine scientifique distinct (ainsi les aspects bibliographiques de la bactériologie, la virologie ou l'immunologie se retrouvent aisément autour de termes liés au séquençage, mais diffèrent dans la quasi-totalité des autres approches méthodologiques) ; iii) l'introduction du temps dans nos outils et d'un suivi de l'apparition de nouveaux termes dans les champs d'étude d'une maladie. Ce dernier point s'intègre dans la problématique générale du développement d'outils de veille des données et des technologies qui reste encore embryonnaire dans notre système.

ii. les modèles de gestion à rendre plus souples et plus flexibles pour l'administrateur et sa gestion en ligne des contenus des banques et des outils ou pour être en mesure de comparer les différentes versions et leur historique… Ces améliorations posent le problème du suivi de la qualité de SM2PH Central dans le cadre de données/méthodes très dynamiques. Cela implique que nous ayons à court ou moyen terme, une meilleure compréhension des éléments éventuellement inutiles, partiellement redondants ou essentiels. Cependant, notre réflexion actuelle se tourne surtout vers le rôle futur des instances de SM2PH et des aménagements indispensables pour faire participer au mieux nos partenaires utilisateurs et leurs données et protocoles propres. Ce point rejoint les besoins d'interopérabilité accrue de MSV3d avec des banques spécifiques d'un locus (LSDB) qui possèdent généralement des données phénotypiques plus détaillées sur les maladies humaines. La prise en compte des données de LSBD, est sans doute le seul moyen de sortir de la description « binaire » de l'impact phénotypique (neutre/délétère) d'une mutation faux sens afin d'aborder des descriptions plus complexes en accord avec la réalité phénotypique des maladies.

iii. la prise en compte de la médecine personnalisée et des notions de pluralités de données que cette nouvelle discipline introduit dans la médecine moderne. En effet, comme cela a été magnifiquement illustré dans l'étude réalisée par l'équipe du professeur Snyder (Chen et al., 2012), un même individu renferme plusieurs génomes, transcriptomes, protéomes… qui de plus, évoluent dans le temps en fonction des événements biologiques et environnementaux qui jalonnent sa vie. Dans une première analyse assez simpliste et très optimiste, on peut imaginer que

chacun pourrait posséder une « SM2PH-Instance » personnelle qui reste encore largement à inventer et à faire évoluer. Cependant, il est clair que ce nouveau regard sur la biologie humaine ouvre des perspectives prodigieuses dans l'étude des maladies en autorisant des analyses et des comparaisons de données jusqu'à présent inenvisageables. Les répercussions sur la compréhension des maladies ou sur la pratique médicale n'en sont encore qu'à leur balbutiements, mais on peut d'ores et déjà penser que, demain, ces données personnalisées représenteront la base descriptive fonctionnelle de tout un chacun et que des systèmes tels que ceux que nous avons mis en place participeront à cette révolution.

ANNEXES

ANNEXE 1 : MATRICES DE SUBSTITUTIONS

La classification des propriétés physico-chimiques est basée sur la diagramme de Venn proposé par Taylor en 1986 (Taylor, 1986) :

- Avec la taille, les acides aminés sont classés en 3 classes : très petit, petit et grand.
- 3 classes de charge : charge positive, charge négative et neutre.
- 2 classes de polarité : polaire et non polaire.
- 3 classes d'hydrophobicité : hydrophobe, neutre et hydrophile.

Basé sur cette classification, nous créons des matrices de scores associés aux modifications introduites dans la taille, la charge, la polarité et l' hydrophobicité comme suit :

- Matrices pour le changement de taille

Changement de taille		résidus muté		
		très petit	petit	grand
résidu sauvage	très petit	conservation	augmentation	augmentation
	petit	diminution	conservation	augmentation
	grand	diminution	diminution	conservation

- Matrices pour le changement de charge

Changement de charge		résidus muté		
		charge positive	charge négative	neutre
résidu sauvage	charge positive	conservation	opposite	augmentation
	charge négative	opposite	conservation	augmentation
	neutre	diminution	diminution	conservation

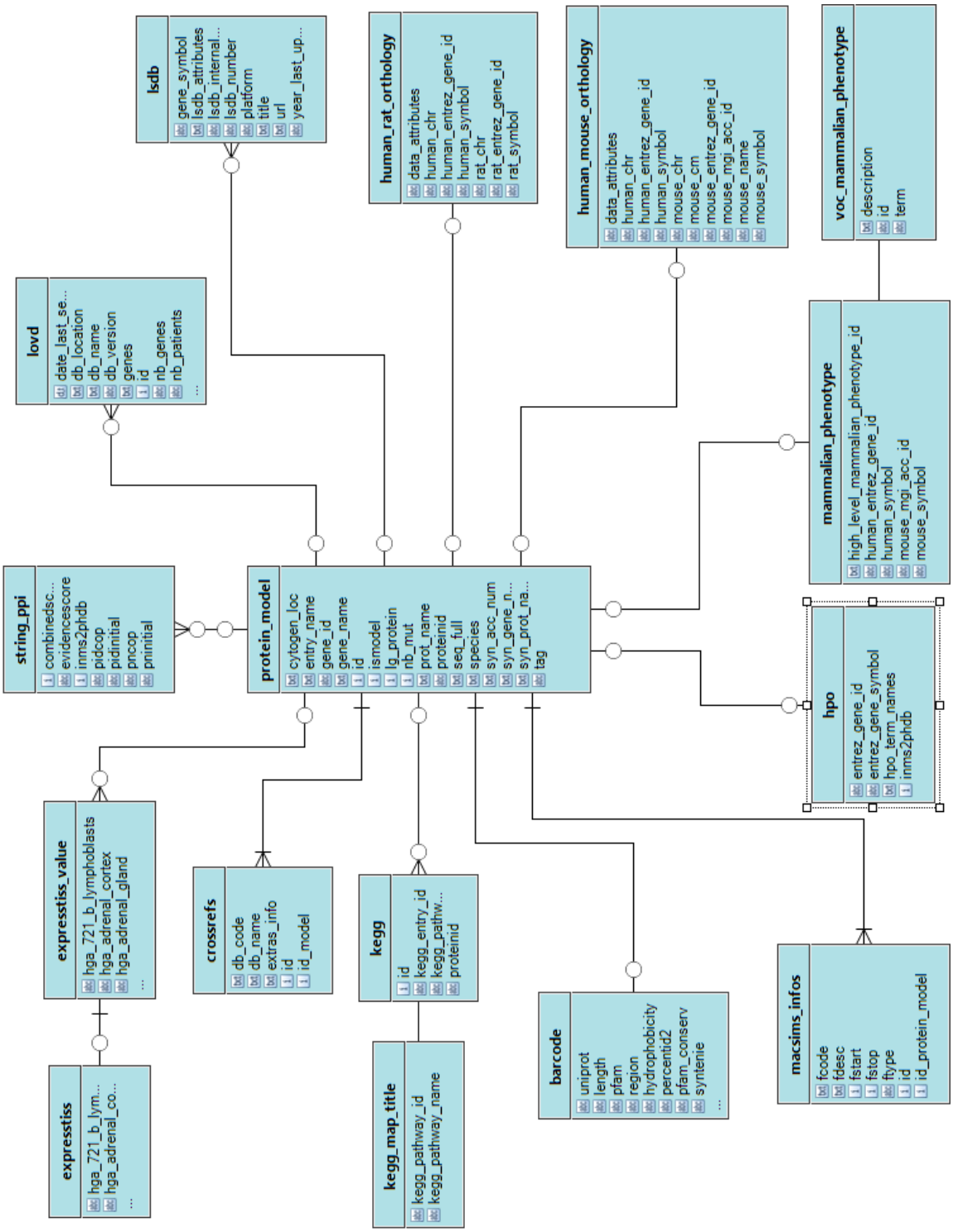
- Matrices pour le changement de polarité

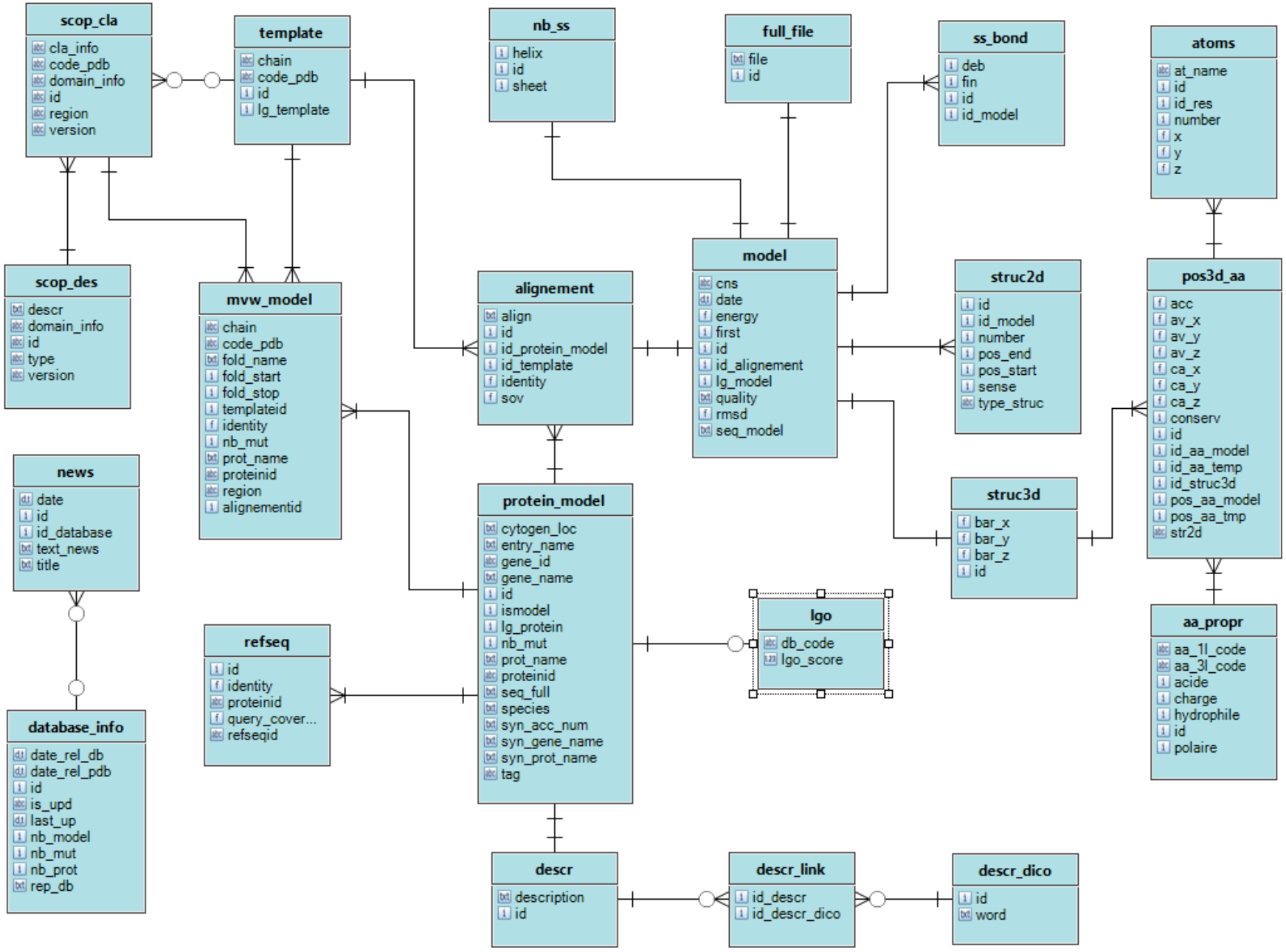
		résidus muté	
		non polaire	polaire
résidu sauvage	non polaire	conservation	augmentation
	polaire	diminution	conservation

- Matrices pour le changement de hydrophobicité

Changement de hydrophobicité		résidus muté		
		hydrophobe	neutre	hydrophile
résidu sauvage	hydrophobe	conservation	diminution	diminution
	neutre	augmentation	conservation	diminution
	hydrophile	augmentation	augmentation	conservation

ANNEXE 2 : SCHEMA LOGIQUE DE LA BASE DE DONNEES DE SM2PH CENTRAL




scop_cla
cla_info
code_pdb
domain_info
id
region
version
template
chain
code_pdb
id
lg_template
nb_ss
helix
id
sheet
full_file
file
id
ss_bond
deb
fin
id
id_model
atoms
at_name
id
id_res
number
x
y
z
scop_des
descr
domain_info
id
type
version
mvw_model
chain
code_pdb
fold_name
fold_start
fold_stop
templateid
identity
nb_mut
prot_name
proteinid
region
alignementid
alignement
align
id
id_protein_model
id_template
identity
sov
model
cns
date
energy
first
id
id_alignement
lg_model
quality
rmsd
seq_model
struc2d
id
id_model
number
pos_end
pos_start
sense
type_struc
pos3d_aa
acc
av_x
av_y
av_z
ca_x
ca_y
ca_z
conserv
id
id_aa_model
id_aa_temp
id_struc3d
pos_aa_model
pos_aa_tmp
str2d
news
date
id
id_database
text_news
title
protein_model
cytogen_loc
entry_name
gene_id
gene_name
id
ismodel
lg_protein
nb_mut
prot_name
proteinid
seq_full
species
syn_acc_num
syn_gene_name
syn_prot_name
tag
struc3d
bar_x
bar_y
bar_z
id
lgo
db_code
lgo_score
refseq
id
identity
proteinid
query_cover...
refseqid
aa_propr
aa_1l_code
aa_3l_code
acide
charge
hydrophile
id
polaire
database_info
date_rel_db
date_rel_pdb
id
is_upd
last_up
nb_model
nb_mut
nb_prot
rep_db
descr
description
id
descr_link
id_descr
id_descr_dico
descr_dico
id
word

ANNEXE 3 : CODE SOURCE POUR EXTRAIRE DES CONNAISSANCES A PARTIR DE MSV3D EN UTILISANT LE PROGRAMME ALEPH

```
% Setting parameters
:- set(i,4).
:- set(nodes, 200000).
:- set(clauselength, 60).
:- set(noise,5).
:- set(minpos,15).
:- set(record,true).
:- set(recordfile,'mutation.log').
:- set(verbosity,0).
:- set(show,1).

:- modeh(1, deleterious(+mutation)).

:- modeb(1, modif_size(+mutation, #modif_size_value)).
:- modeb(1, modif_charge(+mutation, #modif_charge_value)).
:- modeb(1, modif_hydrophobicity(+mutation, #modif_hydrophobicity_value)).
:- modeb(1, modif_polarity(+mutation, #modif_polarity_value)).
:- modeb(1, g_or_p(+mutation, #gp)).
:- modeb(1, is_in_site(+mutation,  #isinsite)).
:- modeb(1, conservation_class(+mutation,  #conservationclass)).
:- modeb(1, conservation_mut(+mutation, -conservationmut)).
:- modeb(1, freq_at_pos(+mutation, -freqatpos)).
:- modeb(1, secondary_struc(+mutation, #secondarystruc)).
:- modeb(*, gain_contact(+mutation, #contact)).
:- modeb(*, lost_contact(+mutation, #contact)).
:- modeb(1, wt_accessibility(+mutation, #wtaccessibility)).
:- modeb(1, mut_accessibility(+mutation, #mutaccessibility)).
:- modeb(1, stability(+mutation, #stabilityvalue)).
:- modeb(1, fold(+mutation, #foldname)).

:- modeb(1, (+freqatpos) @>= (#freqatpos)).
:- modeb(1, (+freqatpos) @=< (#freqatpos)).
:- modeb(1, (+conservationmut) @>= (#conservationmut)).
:- modeb(1, (+conservationmut) @=< (#conservationmut)).

:- determination(deleterious/1, modif_size/2).
:- determination(deleterious/1, modif_charge/2).
:- determination(deleterious/1, modif_hydrophobicity/2).
:- determination(deleterious/1, modif_polarity/2).
:- determination(deleterious/1, g_or_p/2).
:- determination(deleterious/1, is_in_site/2).
:- determination(deleterious/1, conservation_class/2).
:- determination(deleterious/1, conservation_mut/2).
:- determination(deleterious/1, freq_at_pos/2).
:- determination(deleterious/1, secondary_struc/2).
:- determination(deleterious/1, gain_contact/2).
:- determination(deleterious/1, lost_contact/2).
:- determination(deleterious/1, wt_accessibility/2).
:- determination(deleterious/1, mut_accessibility/2).
:- determination(deleterious/1, stability/2).
:- determination(deleterious/1, fold/2).
```

```
:- determination(deleterious/1, gteq/2).
:- determination(deleterious/1, lteq/2).
:- determination(deleterious/1, '='/2).
:- determination(deleterious/1, '@>='/2).
:- determination(deleterious/1, '@=<'/2).


:- discontiguous(modif_size/2, modif_charge/2, modif_hydrophobicity/2,
modif_polarity/2, modif_score/2, g_or_p/2,
      conservation_class/2, conservation_mut/2,
      freq_at_pos/2,
      is_in_site/2,
      secondary_struc/2,
      gain_contact/2, lost_contact/2,
      wt_accessibility/2, mut_accessibility/2,
      stability/2,
      fold/2).


% type definitions

mutation(D):-
      string_length(D, 18),
      sub_string(D, 0, 2, _, 'm_'), !.

modif_size_value('size_increase').
modif_size_value('size_decrease').
modif_size_value('size_unchanged').
modif_size_value('size_unknown').

modif_charge_value('charge_increase').
modif_charge_value('charge_decrease').
modif_charge_value('charge_unchanged').
modif_charge_value('charge_opposite').

modif_hydrophobicity_value('hydrophobicity_increase').
modif_hydrophobicity_value('hydrophobicity_decrease').
modif_hydrophobicity_value('hydrophobicity_unchanged').
modif_hydrophobicity_value('hydrophobicity_unknown').

modif_polarity_value('polarity_increase').
modif_polarity_value('polarity_decrease').
modif_polarity_value('polarity_unchanged').
modif_polarity_value('polarity_unknown').

gp('g_or_p_apparition').
gp('g_or_p_disparition').
gp('g_or_p_unchanged').

secondarystruc('sheet').
secondarystruc('helix').
secondarystruc('not_determined').

isinsite(yes).
isinsite(no).

stability('increase').
stability('decrease').

conservationclass(no_conservation_typification).
conservationclass(global_conservation_rank1).
conservationclass(global_conservation_rank2).
conservationclass(sub_family_conservation).
```

```
mutaccessibility(buried).
mutaccessibility(accessible).
mutaccessibility(intermediate).
wtaccessibility(buried).
wtaccessibility(accessible).
wtaccessibility(intermediate).

contact(phob).
contact(dc).
contact(arom).
contact(hb).

%score(X):-
%     integer(X),
%     X >= 0.

conservationmut(X) :-
      float(X),
      X >= 0.

freqatpos(X) :-
      integer(X),
      X >= 0.

% end type definitions

% Knowldege base

:- consult(modif_size).
:- consult(modif_charge).
:- consult(modif_hydrophobicity).
:- consult(modif_polarity).
:- consult(g_p).
:- consult(conservation_class).
:- consult(conservation_mut).
:- consult(is_in_site).
:- consult(freq_at_pos).
:- consult(secondary_struc).
:- consult(gain_contact).
:- consult(lost_contact).
:- consult(wt_accessibility).
:- consult(mut_accessibility).
:- consult(stability).
:- consult(fold).
```

ANNEXE 4 : LISTE DES REGLES

```
% humvar_1 [Pos cover = 140 (0.02), Neg cover = 5 (0.0025), Rank = 20/111]
deleterious(A) :-
      freq_at_pos(A, B), B@>=2, secondary_struc(A, other),
      lost_contact(A, dc), stability(A, decrease).

% humvar_2 [Pos cover = 61 (0.01), 2 (0.001), Rank = 60/111
deleterious(A) :-
      fold(A, 'HAD-like').

% humvar_3 [Pos cover = 57 (0.0095), Neg cover = 5 (0.0025), Rank = 70/111]
deleterious(A) :-
      modif_charge(A, charge_decrease), g_or_p(A, g_or_p_unchanged),
      is_in_site(A, yes), secondary_struc(A, helix),
      lost_contact(A, hb), stability(A, decrease).

% humvar_4 [Pos cover = 166 (0.028), Neg cover = 5 (0.0025), Rank = 15/111]
deleterious(A) :-
      modif_size(A, size_decrease), modif_charge(A, charge_decrease),
      g_or_p(A, g_or_p_unchanged),
      wt_accessibility(A, buried), mut_accessibility(A, buried),
      stability(A, decrease).


% humvar_5 [Pos cover = 122 (0.02), Neg cover = 5 (0.0025), Rank = 25/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      modif_polarity(A, polarity_unchanged),
      lost_contact(A, phob), stability(A, decrease).

% humvar_6 [Pos cover = 110 (0.018), Neg cover = 5 (0.0025), Rank = 30/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_increase),
      conservation_class(A, global_conservation_rank_1),
      mut_accessibility(A, intermediate), stability(A, decrease).

% humvar_7 [Pos cover = 81 (0.0135), Neg cover = 5 (0.0025), Rank = 51/111]
deleterious(A) :-
      modif_charge(A, charge_decrease),
      modif_hydrophobicity(A, hydrophobicity_increase),
      conservation_class(A, global_conservation_rank_2),
      mut_accessibility(A, intermediate).

% humvar_8 [Pos cover = 475 (0.079), Neg cover = 2 (0.001), Rank = 1/111]
deleterious(A) :-
      conservation_class(A, global_conservation_rank_1),
      freq_at_pos(A, B), B@>=2.

% humvar_9 [Pos cover = 194 (0.032), Neg cover = 4 (0.002), Rank = 10/111]
deleterious(A) :-
      conservation_class(A, global_conservation_rank_1),
      gain_contact(A, dc), wt_accessibility(A, buried).

% humvar_10 [Pos cover = 148 (0.025), Neg cover = 3 (0.0015), Rank = 18/111]
deleterious(A) :-
      modif_size(A, size_decrease), freq_at_pos(A, B), B@>=2,
```

```
      secondary_struc(A, sheet), mut_accessibility(A, buried).

% humvar_11 [Pos cover = 200 (0.033), Neg cover = 5 (0.0025), Rank = 8/111]
deleterious(A) :-
      modif_polarity(A, polarity_increase,
      conservation_class(A, global_conservation_rank_2),
      mut_accessibility(A, buried), stability(A, decrease).

% humvar_12 [Pos cover = 214 (0.036), Neg cover = 3 (0.0015), Rank = 4/111]
deleterious(A) :-
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_1),
      secondary_struc(A, other), wt_accessibility(A, buried).

% humvar_13 [Pos cover = 163 (0.027), Neg cover = 5 (0.0025), Rank = 16/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      secondary_struc(A, sheet), wt_accessibility(A, buried).

% humvar_14 [Pos cover = 173 (0.029), Neg cover = 5 (0.0025), Rank = 13/111]
deleterious(A) :-
      g_or_p(A, g_or_p_disparition),
      gain_contact(A, phob), wt_accessibility(A, buried).

% humvar_15 [Pos cover = 94 (0.016), Neg cover = 4 (0.002), Rank = 39/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_increase),
      modif_polarity(A, polarity_unchanged), g_or_p(A, g_or_p_unchanged),
      secondary_struc(A, other),
      wt_accessibility(A, buried), mut_accessibility(A, buried).

% humvar_16 [Pos cover = 42 (0.007), Neg cover = 5 (0.0025), Rank = 86/111]
deleterious(A) :-
      modif_size(A, size_increase), is_in_site(A, yes),
      gain_contact(A, phob), lost_contact(A, hb).

% humvar_17 [Pos cover = 138 (0.023), Neg cover = 5 (0.0025), Rank = 21/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_1).

% humvar_18 [Pos cover = 62 (0.0103), Neg cover = 5 (0.0025), Rank = 64/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_decrease),
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, sub_family_conservation),
      secondary_struc(A, helix), stability(A, decrease).

% humvar_19 [Pos cover = 40 (0.007), Neg cover = 1 (0.0005), Rank = 84/111]
deleterious(A) :-
      modif_size(A, size_unchanged), is_in_site(A, yes),
      gain_contact(A, hb).


% humvar_20 [Pos cover = 62 (0.0103), Neg cover = 3 (0.0015), Rank = 60/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_decrease),
      modif_polarity(A, polarity_increase),
      conservation_class(A, global_conservation_rank_2),
      mut_accessibility(A, intermediate).
```

```
% humvar_21 [Pos cover = 34 (0.006), Neg cover = 2 (0.001), Rank = 94/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_polarity(A, polarity_unchanged),
      conservation_class(A, sub_family_conservation),
      secondary_struc(A, other), wt_accessibility(A, buried).

% humvar_22 [Pos cover = 100 (0.017), Neg cover = 3 (0.0015), Rank = 35/111]
deleterious(A) :-
      modif_charge(A, charge_decrease),
      g_or_p(A, g_or_p_apparition), wt_accessibility(A, buried).

% humvar_23 [Pos cover = 144 (0.024), 0 (0), Rank = 19/111]
deleterious(A) :-
      fold(A, 'Aromatic aminoacid monoxygenases, catalytic and oligomerization
domains').

% humvar_24 [Pos cover = 121 (0.0201), Neg cover = 4 (0.002), Rank = 25/111]
deleterious(A) :-
      modif_charge(A, charge_decrease),
      modif_hydrophobicity(A, hydrophobicity_increase),
      secondary_struc(A, sheet), mut_accessibility(A, buried).

% humvar_25 [Pos cover = 46 (0.008), Neg cover = 3 (0.0015), Rank = 81/111]
deleterious(A) :-
      fold(A, 'Serum albumin-like').

% humvar_26 [Pos cover = 89 (0.0148), Neg cover = 2 (0.001), Rank = 44/111]
deleterious(A) :-
      freq_at_pos(A, B), B@>=3, wt_accessibility(A, accessible).

% humvar_27 [Pos cover = 51 (0.0085), Neg cover = 5 (0.0025), Rank = 74/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged), g_or_p(A, g_or_p_unchanged),
      is_in_site(A, yes), secondary_struc(A, helix),
      mut_accessibility(A, buried), stability(A, increase).

% humvar_28 [Pos cover = 81 (0.0135), Neg cover = 5 (0.0025), Rank = 51/111]
deleterious(A) :-
      modif_size(A, size_unchanged), modif_charge(A, charge_unchanged),
      is_in_site(A, yes), secondary_struc(A, other),
      wt_accessibility(A, buried).

% humvar_29 [Pos cover = 96 (0.016), Neg cover = 5 (0.0025), Rank = 38/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_decrease),
      conservation_class(A, no_conservation_typification),
      gain_contact(A, phob), wt_accessibility(A, buried).

% humvar_30 [Pos cover = 59 (0.01), Neg cover = 5 (0.0025), Rank = 68/111]
deleterious(A) :-
      modif_polarity(A, polarity_decrease),
      conservation_class(A, global_conservation_rank_2),
      stability(A, increase).

% humvar_31 [Pos cover = 36 (0.006), Neg cover = 5 (0.0025), Rank = 96/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_charge(A, charge_opposite),
      conservation_class(A, global_conservation_rank_2).

% humvar_32 [Pos cover = 38 (0.006), Neg cover = 3 (0.0015), Rank = 90/111]
deleterious(A) :-
      modif_size(A, size_decrease), modif_charge(A, charge_unchanged),
```

```
      conservation_class(A, sub_family_conservation),
      secondary_struc(A, helix), mut_accessibility(A, buried).

% humvar_33 [Pos cover = 125 (0.0208), Neg cover = 5 (0.0025), Rank = 24/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_increase),
      is_in_site(A, yes),
      freq_at_pos(A, B), B@>=2, gain_contact(A, dc).

% humvar_34 [Pos cover = 83 (0.014), Neg cover = 4 (0.002), Rank = 48/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged),
      conservation_class(A, global_conservation_rank_2),
      gain_contact(A, phob), stability(A, decrease).

% humvar_35 [Pos cover = 249 (0.0415), Neg cover = 5 (0.0025), Rank = 3/111]
deleterious(A) :-
      freq_at_pos(A, B), B@>=3,
      secondary_struc(A, other).

% humvar_36 [Pos cover = 48 (0.008), Neg cover = 5 (0.0025), Rank = 81/111]
deleterious(A) :-
      modif_polarity(A, polarity_decrease), g_or_p(A, g_or_p_unchanged),
      conservation_class(A, sub_family_conservation),
      mut_accessibility(A, intermediate).

% humvar_37 [Pos cover = 211 (0.035), Neg cover = 3 (0.0015), Rank = 5/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged),
      modif_polarity(A, polarity_increase),
      conservation_class(A, global_conservation_rank_1).

% humvar_38 [Pos cover = 84 (0.014), Neg cover = 5 (0.0025), Rank = 48/111]
deleterious(A) :-
      modif_size(A, size_decrease), freq_at_pos(A, B), B@>=2,
      secondary_struc(A, helix), wt_accessibility(A, intermediate),
      stability(A, decrease).

% humvar_39 [Pos cover = 88 (0.015), Neg cover = 5 (0.0025), Rank = 46/111]
deleterious(A) :-
      modif_polarity(A, polarity_increase), g_or_p(A, g_or_p_disparition),
      wt_accessibility(A, buried), mut_accessibility(A, intermediate).

% humvar_40 [Pos cover = 136 (0.023), Neg cover = 5 (0.0025), Rank = 22/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_increase),
      is_in_site(A, yes),
      wt_accessibility(A, buried), mut_accessibility(A, buried),
      stability(A, decrease).

% humvar_41 [Pos cover = 90 (0.015), Neg cover = 0 (0), Rank = 39/111]
deleterious(A) :-
      fold(A, 'ATC-like').

% humvar_42 [Pos cover = 397 (0.067), Neg cover = 2 (0.001), Rank = 2/111]
deleterious(A) :-
      freq_at_pos(A, B), B@>=2,
      secondary_struc(A, other), wt_accessibility(A, buried).

% humvar_43 [Pos cover = 81 (0.0135), Neg cover = 0 (0), Rank = 47/111]
deleterious(A) :-
      fold(A, 'HIT-like').
```

```
% humvar_44 [Pos cover = 191 (0.0318), Neg cover = 5 (0.0025), Rank = 11/111]
deleterious(A) :-
      modif_charge(A, charge_increase),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      secondary_struc(A, helix),
      wt_accessibility(A, buried), mut_accessibility(A, buried).

% humvar_45 [Pos cover = 92 (0.015), Neg cover = 5 (0.0025), Rank = 44/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged), lost_contact(A, arom).

% humvar_46 [Pos cover = 81 (0.0135), Neg cover = 4 (0.002), Rank = 50/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_decrease),
      g_or_p(A, g_or_p_apparition),
      conservation_class(A, no_conservation_typification),
      secondary_struc(A, helix),
      mut_accessibility(A, buried).

% humvar_47 [Pos cover = 103 (0.017), Neg cover = 5 (0.0025), Rank = 33/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      conservation_class(A, sub_family_conservation),
      wt_accessibility(A, buried).

% humvar_48 [Pos cover = 38 (0.006), Neg cover = 5 (0.0025), Rank = 92/111]
deleterious(A) :-
      modif_charge(A, charge_decrease), modif_polarity(A, polarity_decrease),
      g_or_p(A, g_or_p_unchanged), is_in_site(A, yes),
      conservation_class(A, no_conservation_typification),
      secondary_struc(A, other),
      wt_accessibility(A, intermediate), stability(A, decrease).

% humvar_49 [Pos cover = 42 (0.007), Neg cover = 5 (0.0025), Rank = 86/111]
deleterious(A) :-
      modif_charge(A, charge_increase),
      conservation_class(A, global_conservation_rank_2),
      secondary_struc(A, other), wt_accessibility(A, intermediate).

% humvar_50 [Pos cover = 208 (0.03), Neg cover = 5 (0.0025), Rank = 7/111]
deleterious(A) :-
      g_or_p(A, g_or_p_disparition),
      freq_at_pos(A, B), B@>=2,
      secondary_struc(A, other).

% humvar_51 [Pos cover = 59 (0.01), Neg cover = 4 (0.002), Rank = 67/111]
deleterious(A) :-
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_2),
      lost_contact(A, phob), lost_contact(A, dc).

% humvar_52 [Pos cover = 94 (0.016), Neg cover = 5 (0.0025), Rank = 42/111]
deleterious(A) :-
      modif_charge(A, charge_increase),
      modif_polarity(A, polarity_increase),
      conservation_class(A, sub_family_conservation).

% humvar_53 [Pos cover = 109 (0.018), Neg cover = 4 (0.002), Rank = 30/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      g_or_p(A, g_or_p_apparition),
```

```
      secondary_struc(A, helix), stability(A, decrease).

% humvar_54 [Pos cover = 117 (0.0195), Neg cover = 3 (0.0015), Rank = 27/111]
deleterious(A) :-
      conservation_class(A, global_conservation_rank_2),
      freq_at_pos(A, B), B@>=2, secondary_struc(A, helix).

% humvar_55 [Pos cover = 196 (0.03), Neg cover = 5 (0.0025), Rank = 9/111]
deleterious(A) :-
      modif_charge(A, charge_increase),
      modif_polarity(A, polarity_increase),
      conservation_class(A, global_conservation_rank_1).

% humvar_56 [Pos cover = 59 (0.01), Neg cover = 3 (0.0015), Rank = 65/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_hydrophobicity(A, hydrophobicity_increase),
      modif_polarity(A, polarity_unchanged),
      is_in_site(A, yes), secondary_struc(A, other),
      mut_accessibility(A, intermediate).

% humvar_57 [Pos cover = 25 (0.004), Neg cover = 2 (0.001), Rank = 101/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      modif_polarity(A, polarity_increase), g_or_p(A, g_or_p_unchanged),
      conservation_class(A, no_conservation_typification).

% humvar_58 [Pos cover = 51 (0.0085), Neg cover = 5 (0.0025), Rank = 74/111
deleterious(A) :-
      modif_polarity(A, polarity_decrease), g_or_p(A, g_or_p_unchanged),
      is_in_site(A, yes), secondary_struc(A, other), lost_contact(A, dc),
      wt_accessibility(A, intermediate).

% humvar_59 [Pos cover = 65 (0.0108), Neg cover = 5 (0.0025), Rank = 59/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_unchanged), is_in_site(A, yes),
      freq_at_pos(A, B), B@>=2,
      secondary_struc(A, helix),
      wt_accessibility(A, buried), stability(A, decrease).

% humvar_60 [Pos cover = 47 (0.008), Neg cover = 4 (0.002), Rank = 81/111]
deleterious(A) :-
      modif_size(A, size_unchanged),
      modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      is_in_site(A, yes), secondary_struc(A, sheet).

% humvar_61 [Pos cover = 114 (0.019), Neg cover = 5 (0.0025), 29/111]
deleterious(A) :-
      modif_charge(A, charge_increase),
      modif_polarity(A, polarity_increase),
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, no_conservation_typification),
      wt_accessibility(A, buried), mut_accessibility(A, buried).

% humvar_62 [Pos cover = 50 (0.008), Neg cover = 5 (0.0025), Rank = 76/111]
deleterious(A) :-
      modif_polarity(A, polarity_unchanged),
      g_or_p(A, g_or_p_apparition),
      secondary_struc(A, other),
      mut_accessibility(A, buried).
```

```
% humvar_63 [Pos cover = 102 (0.017), Neg cover = 4 (0.002), Rank = 33/111]
deleterious(A) :-
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_1), stability(A, increase).

% humvar_64 [Pos cover = 15 (0.0025), Neg cover = 5 (0.0025), Rank = 111/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      is_in_site(A, no),
      secondary_struc(A, helix), mut_accessibility(A, buried).

% humvar_65 [Pos cover = 50 (0.008), Neg cover = 5 (0.0025), Rank = 76/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_charge(A, charge_unchanged),
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_2),
      secondary_struc(A, sheet), wt_accessibility(A, buried).

% humvar_66 [Pos cover = 58 (0.01), Neg cover = 2 (0.001), Rank = 65/111]
deleterious(A) :-
      modif_size(A, size_increase),
      g_or_p(A, g_or_p_apparition), secondary_struc(A, helix).

% humvar_67 [Pos cover = 43 (0.007), Neg cover = 5 (0.0025), Rank = 85/111]
deleterious(A) :-
      modif_charge(A, charge_increase),
      modif_hydrophobicity(A, hydrophobicity_unchanged), gain_contact(A, dc).

% humvar_68 [Pos cover = 76 (0.013), Neg cover = 3 (0.0015), Rank = 53/111]
deleterious(A) :-
      modif_size(A, size_unchanged), is_in_site(A, yes),
      freq_at_pos(A, B), B@>=2,
      wt_accessibility(A, intermediate).

% humvar_69 [Pos cover = 173 (0.029), Neg cover = 5 (0.0025), Rank = 13/111]
deleterious(A) :-
      modif_polarity(A, polarity_decrease), g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_1), stability(A, decrease).

% humvar_70 [Pos cover = 119 (0.0198), Neg cover = 5 (0.0025), Rank = 27/111]
deleterious(A) :-
      modif_size(A, size_decrease), modif_polarity(A, polarity_decrease),
      is_in_site(A, yes), secondary_struc(A, other),
      wt_accessibility(A, buried).

% humvar_71 [Pos cover = 62 (0.0103), Neg cover = 3 (0.0015), Rank = 60/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      g_or_p(A, g_or_p_unchanged),
      is_in_site(A, yes), secondary_struc(A, helix),
      wt_accessibility(A, buried), mut_accessibility(A, buried).

% humvar_72 [Pos cover = 64 (0.0107), Neg cover = 5 (0.0025), Rank = 60/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_polarity(A, polarity_decrease), g_or_p(A, g_or_p_unchanged),
      is_in_site(A, yes), conservation_class(A, no_conservation_typification),
      mut_accessibility(A, buried).

% humvar_73 [Pos cover = 95 (0.0158), Neg cover = 5 (0.0025), Rank =39/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_hydrophobicity(A, hydrophobicity_increase),
```

```
      g_or_p(A, g_or_p_unchanged), freq_at_pos(A, B), B@>=2.

% humvar_74 [Pos cover = 20 (0.003), Neg cover = 3 (0.0015), Rank = 106/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged), is_in_site(A, yes),
      conservation_class(A, no_conservation_typification),
      secondary_struc(A, other), lost_contact(A, hb),
      mut_accessibility(A, intermediate).

% humvar_75 [Pos cover = 108 (0.018), Neg cover = 4 (0.002), Rank = 32/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged),
      conservation_class(A, sub_family_conservation),
      freq_at_pos(A, B), B@>=2.

% humvar_76 [Pos cover = 19 (0.003), Neg cover = 4 (0.002), Rank = 108/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      conservation_class(A, sub_family_conservation), lost_contact(A, dc).

% humvar_77 [Pos cover = 39 (0.0065), Neg cover = 3 (0.0015), Rank = 88/111]
deleterious(A) :-
      modif_charge(A, charge_opposite),
      conservation_class(A, sub_family_conservation).

% humvar_78 [Pos cover = 211 (0.04), Neg cover = 4 (0.002), Rank = 6/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_decrease), is_in_site(A, yes),
      freq_at_pos(A, B), B@>=2, secondary_struc(A, other).

% humvar_79 [Pos cover = 71 (0.012), Neg cover = 5 (0.0025), Rank = 56/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      modif_polarity(A, polarity_decrease), g_or_p(A, g_or_p_unchanged),
      mut_accessibility(A, buried).

% humvar_80 [Pos cover = 180 (0.03), Neg cover = 4 (0.002), Rank = 12/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_polarity(A, polarity_unchanged),
      conservation_class(A, global_conservation_rank_1),
      wt_accessibility(A, buried).

% humvar_81 [Pos cover = 41 (0.007), Neg cover = 5 (0.0025), Rank = 88/111]
deleterious(A) :-
      modif_charge(A, charge_decrease), modif_polarity(A, polarity_decrease),
      is_in_site(A, yes), conservation_class(A, no_conservation_typification),
      freq_at_pos(A, B), B@>=2, wt_accessibility(A, accessible).

% humvar_82 [Pos cover = 49 (0.008), Neg cover = 4 (0.002), Rank = 76/111]
deleterious(A) :-
      modif_size(A, size_unchanged), modif_polarity(A, polarity_unchanged),
      is_in_site(A, yes), secondary_struc(A, sheet),
      wt_accessibility(A, buried), mut_accessibility(A, buried).

% humvar_83 [Pos cover = 32 (0.005), Neg cover = 5 (0.0025), Rank = 100/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_charge(A, charge_unchanged),
      modif_polarity(A, polarity_unchanged),
      conservation_class(A, global_conservation_rank_2),
      secondary_struc(A, other), wt_accessibility(A, intermediate).

% humvar_84 [Pos cover = 58 (0.01), Neg cover = 5 (0.0025), Rank = 69/111]
```

```
deleterious(A) :-
      modif_size(A, size_decrease), g_or_p(A, g_or_p_apparition),
      is_in_site(A, yes), conservation_class(A, no_conservation_typification),
      lost_contact(A, dc), wt_accessibility(A, intermediate).

% humvar_85 [Pos cover = 37 (0.006), Neg cover = 5 (0.0025), Rank = 94/111]
deleterious(A) :-
      modif_charge(A, charge_decrease),
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      conservation_class(A, sub_family_conservation).

% humvar_86 [Pos cover = 50 (0.008), Neg cover = 5 (0.0025), Rank = 76/111]
deleterious(A) :-
      modif_charge(A, charge_decrease),
      modif_polarity(A, polarity_decrease),
      is_in_site(A, yes), mut_accessibility(A, intermediate),
      stability(A, increase).

% humvar_87 [Pos cover = 16 (0.003), Neg cover = 3 (0.0015), Rank = 110/111]
deleterious(A) :-
      modif_size(A, size_unchanged),
      g_or_p(A, g_or_p_unchanged),
      is_in_site(A, no), secondary_struc(A, helix).

% humvar_88 [Pos cover = 93 (0.016), Neg cover = 5 (0.0025), Rank = 43/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      modif_polarity(A, polarity_unchanged), is_in_site(A, yes),
      secondary_struc(A, sheet), stability(A, decrease).

% humvar_89 [Pos cover = 34 (0.006), Neg cover = 4 (0.002), Rank = 97/111]
deleterious(A) :-
      modif_polarity(A, polarity_increase),
      conservation_class(A, global_conservation_rank_2),
      wt_accessibility(A, accessible).

% humvar_90 [Pos cover = 38 (0.006), Neg cover = 5 (0.0025), Rank = 92/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged), g_or_p(A, g_or_p_unchanged),
      secondary_struc(A, helix), gain_contact(A, hb).

% humvar_91 [Pos cover = 20 (0.003), Neg cover = 5 (0.0025), Rank = 108/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_increase),
      g_or_p(A, g_or_p_disparition),
      conservation_class(A, no_conservation_typification),
      secondary_struc(A, other),
      mut_accessibility(A, accessible), stability(A, decrease).

% humvar_92 [Pos cover = 38 (0.006), Neg cover = 3 (0.0015), Rank = 90/111]
deleterious(A) :-
      fold(A, 'TIM beta/alpha-barrel').

% humvar_93 [Pos cover = 24 (0.004), Neg cover = 2 (0.001), Rank = 102/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_charge(A, charge_opposite),
      conservation_class(A, no_conservation_typification),
      wt_accessibility(A, buried).

% humvar_94 [Pos cover = 20 (0.003), Neg cover = 3 (0.0015), Rank = 106/111]
deleterious(A) :-
      modif_size(A, size_increase),
```

```
      fold(A, 'NAD(P)-binding Rossmann-fold domains').

% humvar_95 [Pos cover = 96 (0.016), Neg cover = 4 (0.002), Rank = 37/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_increase),
      modif_polarity(A, polarity_unchanged),
      conservation_class(A, global_conservation_rank_1),
      secondary_struc(A, helix).

% humvar_96 [Pos cover = 66 (0.011), Neg cover = 4 (0.002), Rank = 57/111]
deleterious(A) :-
      g_or_p(A, g_or_p_disparition), is_in_site(A, yes),
      conservation_class(A, no_conservation_typification), gain_contact(A, dc),
      mut_accessibility(A, buried).

% humvar_97 [Pos cover = 77 (0.013), Neg cover = 5 (0.0025), Rank = 54/111]
deleterious(A) :-
      modif_polarity(A, polarity_increase), g_or_p(A, g_or_p_disparition),
      conservation_class(A, no_conservation_typification),
      mut_accessibility(A, buried),
      stability(A, decrease).

% humvar_98 [Pos cover = 23 (0.004), Neg cover = 4 (0.002), Rank = 104/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged), modif_polarity(A, polarity_unchanged),
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_2),
      wt_accessibility(A, accessible), stability(A, decrease).

% humvar_99 [Pos cover = 54 (0.009), Neg cover = 5 (0.0025), Rank = 71/111]
deleterious(A) :-
      modif_polarity(A, polarity_unchanged),
      conservation_class(A, global_conservation_rank_1),
      secondary_struc(A, other), mut_accessibility(A, accessible).

% humvar_100 [Pos cover = 23 (0.004), Neg cover = 5 (0.0025), Rank = 105/111]
deleterious(A) :-
      modif_hydrophobicity(A, hydrophobicity_decrease),
      modif_polarity(A, polarity_decrease), is_in_site(A, yes),
      secondary_struc(A, other), wt_accessibility(A, intermediate).

% humvar_101 [Pos cover = 73 (0.012), Neg cover = 4 (0.002), Rank = 55/111]
deleterious(A) :-
      modif_size(A, size_unchanged),
      modif_hydrophobicity(A, hydrophobicity_increase),
      conservation_class(A, global_conservation_rank_2).

% humvar_102 [Pos cover = 53  (0.009), Neg cover = 5 (0.0025), Rank = 73/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      conservation_class(A, global_conservation_rank_1),
      secondary_struc(A, helix),
      stability(A, decrease).

% humvar_103 [Pos cover = 35 (0.006), Neg cover = 5 (0.0025), Rank = 97/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      modif_polarity(A, polarity_unchanged),
      conservation_class(A, global_conservation_rank_2),
      secondary_struc(A, helix).

% humvar_104 [Pos cover = 66 (0.011), Neg cover = 5 (0.0025), Rank = 58/111]
```

```
deleterious(A) :-
      modif_size(A, size_decrease), modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_decrease),
      modif_polarity(A, polarity_unchanged), mut_accessibility(A, intermediate).

% humvar_105 [Pos cover = 50 (0.008), Neg cover = 5 (0.0025), Rank = 76/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_polarity(A, polarity_increase),
      g_or_p(A, g_or_p_disparition),
      is_in_site(A, yes), secondary_struc(A, other),
      wt_accessibility(A, intermediate), mut_accessibility(A, intermediate).

% humvar_106 [Pos cover = 26 (0.004), Neg cover = 5 (0.0025), Rank = 103/111]
deleterious(A) :-
      modif_charge(A, charge_unchanged),
      modif_hydrophobicity(A, hydrophobicity_increase),
      g_or_p(A, g_or_p_unchanged),
      conservation_class(A, global_conservation_rank_2),
      wt_accessibility(A, intermediate), stability(A, decrease).

% humvar_107 [Pos cover = 160 (0.027), Neg cover = 5 (0.0025), Rank = 17/111]
      deleterious(A) :-
      modif_size(A, size_increase), modif_charge(A, charge_increase),
      conservation_class(A, global_conservation_rank_2),
      wt_accessibility(A, buried).

% humvar_108 [Pos cover = 131 (0.022), Neg cover = 5 (0.0025), Rank = 23/111]
deleterious(A) :-
      modif_charge(A, charge_decrease),
      conservation_class(A, global_conservation_rank_2),
      mut_accessibility(A, buried).

% humvar_109 [Pos cover = 99 (0.0165), Neg cover = 5 (0.0025), Rank = 36/111]
deleterious(A) :-
      modif_size(A, size_increase), modif_polarity(A, polarity_increase),
      g_or_p(A, g_or_p_unchanged), is_in_site(A, yes),
      secondary_struc(A, other), mut_accessibility(A, buried).

% humvar_110 [Pos cover = 35 (0.006), Neg cover = 5 (0.0025), Rank = 97/111]
deleterious(A) :-
      modif_size(A, size_decrease),
      modif_hydrophobicity(A, hydrophobicity_unchanged),
      is_in_site(A, yes), secondary_struc(A, other),
      lost_contact(A, phob), mut_accessibility(A, intermediate).

% humvar_111 [Pos cover = 54, (0.009, Neg cover = 5 (0.0025), 71/111]
deleterious(A) :-
      modif_size(A, size_increase),
      modif_charge(A, charge_unchanged),
      conservation_class(A, no_conservation_typification),
      secondary_struc(A, other), gain_contact(A, phob),
      mut_accessibility(A, buried).
```

LISTE DES PUBLICATIONS PERSONNELLES

a. Publications :

- Luu TD, Rusu AM, Walter V, Linard B, Poidevin L, Ripp R, Moulinier L, Muller J, Raffelsberger W, Wicker N, Lecompte O, Thompson JD, Poch O, Nguyen NH. KD4v: Comprehensible Knowledge Discovery System For Missense Variant. Nucleic Acids Res. 2012.

- Luu TD, Rusu AM, Walter V, Ripp R, Moulinier L, Muller J, Toursel T, Thompson JD, Poch O, Nguyen NH. MSV3d: database of human MisSense variants mapped to 3D protein structure. Database (Oxford). 2012.

- Audo I, Bujakowska K, Orhan E, Poloschek CM, Defoort-Dhellemmes S, Drumare I, Kohl S, Luu TD, Lecompte O, Zrenner E, Lancelot ME, Antonio A, Germain A, Michiels C, Audier C, Letexier M, Saraiva JP, Leroy BP, Munier FL, Mohand-Saïd S, Lorenz B, Friedburg C, Preising M, Kellner U, Renner AB, Moskova-Doumanova V, Berger W, Wissinger B, Hamel CP, Schorderet DF, De Baere E, Sharon D, Banin E, Jacobson SG, Bonneau D, Zanlonghi X, Le Meur G, Casteels I, Koenekoop R, Long VW, Meire F, Prescott K, de Ravel T, Simmons I, Nguyen H, Dollfus H, Poch O, Léveillard T, Nguyen-Ba-Charvet K, Sahel JA, Bhattacharya SS, Zeitz C. Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness. Am J Hum Genet 90, 321-330. 2012.

- Zeitz C, Jacobson SG, Hamel CP, Bujakowska K, Orhan E, Zanlonghi X, Lancelot ME, Michiels C, Schwartz SB, Bocquet B, CSNB consortium, Antonio A, Audier C, Letexier M, Saraiva JP, Luu TD, Sennlaub F, Nguyen H, Poch O, Dollfus H, Lecompte O, Kohl S, Sahel JA, Bhattacharya SS, Audo I. *Whole exome sequencing identifies mutations in LRIT3 as a cause for autosomal recessive complete congenital stationary night blindness*. Am J Hum Genet, accepté.

b. Communications orales :

- Luu TD, Nguyen NH, Friedrich A, Muller J, Moulinier L, Poch O. Extracting Knowledge from a Mutation Database Related to Human Monogenic Disease Using Inductive Logic Programming. In International Conference on Bioscience, Biochemistry and Bioinformatics; Singapore, Frévrier 2011. IEEE Catalog Number: CFP1134M-PRT. ISBN: 978-1-4244-9388-3.

- Luu TD, Nguyen NH, Friedrich A, Muller J, Moulinier L, Poch O. Discovering knowledge hidden in mutation data using Inductive Logic Programming. In Les assises du GdR I3; Strasbourg. Juin 2010.

c. Posters :

- Luu TD, Poch O, Nguyen NH. KD4v: Comprehensible Knowledge Discovery System For Missense Variants. In European Conference on Computational Biology; Basel. Septembre 2012.

- Luu TD, Poch O, Nguyen NH. MSV3d: Database of human MisSense Variants mapped to 3D protein structure. In European Conference on Computational Biology ; Basel. Septembre 2012.

- Luu TD, Nguyen NH, Friedrich A, Muller J, Moulinier L, Poch O. SM2PH-kb: Data Warehouse Intelligence for the Integrated Study of Human Structural Mutation to

Phenotypes Relationships. Journées Ouvertes en Biologie, Informatique et Mathématiques ; Paris. Juin 2011.

- Luu TD, Nguyen NH, Friedrich A, Muller J, Moulinier L, Mandel J, Poch O. A novel tool for the integrated study of human missense variants to phenotypes relationships. In European Human Genetics Conference; Amsterdam. Mai 2011.
- Luu TD, Nguyen NH, Friedrich A, Muller J, Moulinier L, Poch O. Human-comprehensible rule generator for identifying deleterious amino acid variants. In Theoretical Approaches for the Genome and the proteins; Annecy-le-Vieux. Octobre 2010. Young Fellowship
- Luu TD, Nguyen NH, Friedrich A, Muller J, Moulinier L, Poch O. Development of knowledge-based system for analysing the effects of single nucleotide polymorphisms on the protein function. In Journées Ouvertes en Biologie, Informatique et Mathématiques; Montpellier. Septembre 2010.
- Luu TD, Nguyen NH, Friedrich A, Muller J, Moulinier L, Poch O. Discovering knowledge hidden in mutation data using Inductive Logic Programming. In Intelligent Systems for Molecular Biology; Boston. Juillet 2010.

BIBLIOGRAPHIE

Aartsma-Rus, A., Van Deutekom, J.C., Fokkema, I.F., Van Ommen, G.J., and Den Dunnen, J.T. (2006). Entries in the Leiden Duchenne muscular dystrophy mutation database: an overview of mutation types and paradoxical cases that confirm the reading-frame rule. Muscle Nerve *34*, 135-144.

Adzhubei, I.A., Schmidt, S., Peshkin, L., Ramensky, V.E., Gerasimova, A., Bork, P., Kondrashov, A.S., and Sunyaev, S.R. (2010). A method and server for predicting damaging missense mutations. Nat Methods *7*, 248-249.

Aerts, M., Van Holsbeke, C., de Ravel, T., and Devlieger, R. (2006a). Prenatal diagnosis of type II osteogenesis imperfecta, describing a new mutation in the COL1A1 gene. Prenat Diagn *26*, 394.

Aerts, S., Lambrechts, D., Maity, S., Van Loo, P., Coessens, B., De Smet, F., Tranchevent, L.C., De Moor, B., Marynen, P., Hassan, B.*, et al.* (2006b). Gene prioritization through genomic data fusion. Nat Biotechnol *24*, 537-544.

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*, D418-424.

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.

Amberger, J., Bocchini, C.A., Scott, A.F., and Hamosh, A. (2009). McKusick's Online Mendelian Inheritance in Man (OMIM). Nucleic Acids Res *37*, D793-796.

Amini, A., Shrimpton, P.J., Muggleton, S.H., and Sternberg, M.J. (2007). A general approach for developing system-specific functions to score protein-ligand docked complexes using support vector inductive logic programming. Proteins *69*, 823-831.

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.

Audo, I., Bujakowska, K., Orhan, E., Poloschek, C.M., Defoort-Dhellemmes, S., Drumare, I., Kohl, S., Luu, T.D., Lecompte, O., Zrenner, E.*, et al.* (2012). Whole-exome sequencing identifies mutations in GPR179 leading to autosomal-recessive complete congenital stationary night blindness. Am J Hum Genet *90*, 321-330.

Bao, L., and Cui, Y. (2005). Prediction of the phenotypic effects of non-synonymous single nucleotide polymorphisms using structural and evolutionary information. Bioinformatics *21*, 2185-2190.

Bao, L., Zhou, M., and Cui, Y. (2005). nsSNPAnalyzer: identifying disease-associated nonsynonymous single nucleotide polymorphisms. Nucleic Acids Res *33*, W480-482.

Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J., and Wheeler, D.L. (2006). GenBank. Nucleic Acids Res *34*, D16-20.

Berman, H., Henrick, K., Nakamura, H., and Markley, J.L. (2007). The worldwide Protein Data Bank (wwPDB): ensuring a single, uniform archive of PDB data. Nucleic Acids Res *35*, D301-303.

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 Research *37*, D365-D368.

Beroud, C., Hamroun, D., Collod-Beroud, G., Boileau, C., Soussi, T., and Claustres, M. (2005). UMD (Universal Mutation Database): 2005 update. Hum Mutat *26*, 184-191.

Bertone, P., and Gerstein, M. (2001). Integrative data mining: the new direction in bioinformatics. IEEE Eng Med Biol Mag *20*, 33-40.

Blagosklonny, M.V., and Pardee, A.B. (2002). Conceptual biology: unearthing the gems. Nature *416*, 373.

Blockeel, H., and Raedt, L.D. (1998). Top-down induction of first-order logical decision trees. Artif Intell *101*, 285-297.

Bootsma, D., and Hoeijmakers, J.H. (1991). The genetic basis of xeroderma pigmentosum. Ann Genet *34*, 143-150.

Bowie, J.U., Reidhaar-Olson, J.F., Lim, W.A., and Sauer, R.T. (1990). Deciphering the message in protein sequences: tolerance to amino acid substitutions. Science *247*, 1306-1310.

Breitkreutz, B.J., Stark, C., Reguly, T., Boucher, L., Breitkreutz, A., Livstone, M., Oughtred, R., Lackner, D.H., Bahler, J., Wood, V.*, et al.* (2008). The BioGRID Interaction Database: 2008 update. Nucleic Acids Res *36*, D637-640.

Bromberg, Y., and Rost, B. (2007). SNAP: predict effect of non-synonymous polymorphisms on function. Nucleic Acids Res *35*, 3823-3835.

Calabrese, R., Capriotti, E., Fariselli, P., Martelli, P.L., and Casadio, R. (2009). Functional annotations improve the predictive score of human disease-related mutations in proteins. Hum Mutat *30*, 1237-1244.

Calvo, B., Lopez-Bigas, N., Furney, S.J., Larranaga, P., and Lozano, J.A. (2007). A partially supervised classification approach to dominant and recessive human disease gene prediction. Comput Methods Programs Biomed *85*, 229-237.

Capriotti, E., and Altman, R.B. (2011). Improving the prediction of disease-related variants using protein three-dimensional structure. BMC Bioinformatics *12 Suppl 4*, S3.

Capriotti, E., Fariselli, P., and Casadio, R. (2005). I-Mutant2.0: predicting stability changes upon mutation from the protein sequence or structure. Nucleic Acids Res *33*, W306-310.

Chakravarti, A. (2001). To a future of genetic medicine. Nature *409*, 822-823.

Chasman, D., and Adams, R.M. (2001). Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. J Mol Biol *307*, 683-706.

Chatr-aryamontri, A., Ceol, A., Palazzi, L.M., Nardelli, G., Schneider, M.V., Castagnoli, L., and Cesareni, G. (2007). MINT: the Molecular INTeraction database. Nucleic Acids Res *35*, D572-574.

Chen, J., Bardes, E.E., Aronow, B.J., and Jegga, A.G. (2009). ToppGene Suite for gene list enrichment analysis and candidate gene prioritization. Nucleic Acids Res *37*, W305-311.

Chen, R., Mias, G.I., Li-Pook-Than, J., Jiang, L., Lam, H.Y., Miriami, E., Karczewski, K.J., Hariharan, M., Dewey, F.E., Cheng, Y.*, et al.* (2012). Personal omics profiling reveals dynamic molecular and medical phenotypes. Cell *148*, 1293-1307.

Cohen, J.C., Kiss, R.S., Pertsemlidis, A., Marcel, Y.L., McPherson, R., and Hobbs, H.H. (2004). Multiple rare alleles contribute to low plasma levels of HDL cholesterol. Science *305*, 869-872.

Collins, F.S., Brooks, L.D., and Chakravarti, A. (1998). A DNA polymorphism discovery resource for research on human genetic variation. Genome Res *8*, 1229-1231.

Consortium, T.U. (2009). The Universal Protein Resource (UniProt) 2009. Nucleic Acids Research *37*, D169-D174.

Cootes, A.P., Muggleton, S., Greaves, R.B., and Sternberg, M.J.E. (2001). Automatic determination of protein fold signatures from structural superpositions. Electron Trans Artif Intell *5*, 245-274.

Cootes, A.P., Muggleton, S.H., and Sternberg, M.J. (2003). The automatic discovery of structural principles describing protein fold space. J Mol Biol *330*, 839-850.

Cotton, R.G. (2000). Progress of the HUGO mutation database initiative: a brief introduction to the human mutation MDI special issue. Hum Mutat *15*, 4-6.

de Wind, N., and Hays, J.B. (2001). Mismatch repair: praying for genome stability. Curr Biol *11*, R545-548.

Dobson, R.J., Munroe, P.B., Caulfield, M.J., and Saqi, M.A. (2006). Predicting deleterious nsSNPs: an analysis of sequence and structural attributes. BMC Bioinformatics *7*, 217.

Dyson, H.J., and Wright, P.E. (2005). Intrinsically unstructured proteins and their functions. Nat Rev Mol Cell Biol *6*, 197-208.

Eswar, N., Eramian, D., Webb, B., Shen, M.Y., and Sali, A. (2008). Protein structure modeling with MODELLER. Methods Mol Biol *426*, 145-159.

Fayyad, U.M., Piatetsky-Shapiro, G., and Smyth, P. (1996). From data mining to knowledge discovery: an overview. In Advances in knowledge discovery and data mining, M.F. Usama, P.-S. Gregory, S. Padhraic, and U. Ramasamy, eds. (American Association for Artificial Intelligence), pp. 1-34.

Finn, R.D., Mistry, J., Tate, J., Coggill, P., Heger, A., Pollington, J.E., Gavin, O.L., Gunasekaran, P., Ceric, G., Forslund, K.*, et al.* (2010). The Pfam protein families database. Nucleic Acids Res *38*, D211-222.

Fokkema, I.F., den Dunnen, J.T., and Taschner, P.E. (2005). LOVD: easy creation of a locus-specific sequence variation database using an "LSDB-in-a-box" approach. Hum Mutat *26*, 63-68.

French, S., and Robson, B. (1983). What is a conservative substitution? Journal of Molecular Evolution *19*, 171-175.

Friedrich, A. (2007). De la mutation structurale aux phénotypes des pathologies humaines : vers une approche intégrative des mutations et de leurs conséquences (Strasbourg, Université Louis Pasteur).

Friedrich, A., Garnier, N., Gagnière, N., Nguyen, H., Albou, L.P., Biancalana, V., Bettler, E., Deléage, G., Lecompte, O., Muller, J.*, et al.* (2010). SM2PH-db: an interactive system for the integrated analysis of phenotypic consequences of missense mutations in proteins involved in human genetic diseases. Human Mutation *31*, 127-135.

Friedrich, A., Ripp, R., Garnier, N., Bettler, E., Deleage, G., Poch, O., and Moulinier, L. (2007). Blast sampling for structural and functional analyses. BMC Bioinformatics *8*, 62.

Garnier, N. (2008). Mise en place d'un environnement bioinformatique d'évaluation et de prédiction de l'impact de mutations sur le phénotype de pathologies humaines (Lyon, Université Claude Bernard).

Hamosh, A., Scott, A.F., Amberger, J.S., Bocchini, C.A., and McKusick, V.A. (2005). Online Mendelian Inheritance in Man (OMIM), a knowledgebase of human genes and genetic disorders. Nucleic Acids Res *33*, D514-517.

Han, J., and Kamber, M. (2001). Data Mining : Concepts and Techniques.

Henrick, K., Feng, Z., Bluhm, W.F., Dimitropoulos, D., Doreleijers, J.F., Dutta, S., Flippen-Anderson, J.L., Ionides, J., Kamada, C., Krissinel, E.*, et al.* (2008). Remediation of the protein data bank archive. Nucleic Acids Res *36*, D426-433.

Hernandez, T., and Kambhampati, S. (2004). Integration of biological sources: current systems and challenges ahead. SIGMOD Rec *33*, 51-60.

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*, D577-581.

Hutz, J.E., Kraja, A.T., McLeod, H.L., and Province, M.A. (2008). CANDID: a flexible method for prioritizing candidate genes for complex human traits. Genet Epidemiol *32*, 779-790.

IGHSC (2004). Finishing the euchromatic sequence of the human genome. Nature *431*, 931-945.

Jensen, L.J., Lagarde, J., von Mering, C., and Bork, P. (2004). ArrayProspector: a web resource of functional associations inferred from microarray expression data. Nucleic Acids Res *32*, W445-448.

Jirtle, R.L., and Skinner, M.K. (2007). Environmental epigenomics and disease susceptibility. Nat Rev Genet *8*, 253-262.

Kabsch, W., and Sander, C. (1983). Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers *22*, 2577-2637.

Kaminker, J.S., Zhang, Y., Waugh, A., Haverty, P.M., Peters, B., Sebisanovic, D., Stinson, J., Forrest, W.F., Bazan, J.F., Seshagiri, S.*, et al.* (2007). Distinguishing cancer-associated missense mutations from common polymorphisms. Cancer Res *67*, 465-473.

Kampa, D., Cheng, J., Kapranov, P., Yamanaka, M., Brubaker, S., Cawley, S., Drenkow, J., Piccolboni, A., Bekiranov, S., Helt, G.*, et al.* (2004). Novel RNAs identified from an in-depth analysis of the transcriptome of human chromosomes 21 and 22. Genome Res *14*, 331-342.

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.

Kelley, L.A., Shrimpton, P.J., Muggleton, S.H., and Sternberg, M.J. (2009). Discovering rules for protein-ligand specificity using support vector inductive logic programming. Protein Eng Des Sel *22*, 561-567.

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*, D561-565.

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. Nucleic Acids Res *37*, D464-470.

King, R.D. (2004). Applying inductive logic programming to predicting gene function. AI Mag *25*, 57-68.

Kouranov, A., Xie, L., de la Cruz, J., Chen, L., Westbrook, J., Bourne, P.E., and Berman, H.M. (2006). The RCSB PDB information portal for structural genomics. Nucleic Acids Res *34*, D302-305.

Kunkel, T.A. (2004). DNA replication fidelity. J Biol Chem *279*, 16895-16898.

Kussie, P.H., Gorina, S., Marechal, V., Elenbaas, B., Moreau, J., Levine, A.J., and Pavletich, N.P. (1996). Structure of the MDM2 oncoprotein bound to the p53 tumor suppressor transactivation domain. Science *274*, 948-953.

Lauer, F., and Guermeur, Y. (2011). MSVMpack: a Multi-Class Support Vector Machine Package. Journal of Machine Learning Research *12*, 2269-2272.

Lavrac, N., and Dzeroski, S. (1994). Inductive Logic Programming: Techniques and Applications (New York, Ellis Horwood).

Lee, T.J., Pouliot, Y., Wagner, V., Gupta, P., Stringer-Calvert, D.W., Tenenbaum, J.D., and Karp, P.D. (2006). BioWarehouse: a bioinformatics database warehouse toolkit. BMC Bioinformatics *7*, 170.

Lejeune, J., Turpin, R., and Gautier, M. (1959). [Mongolism; a chromosomal disease (trisomy)]. Bull Acad Natl Med *143*, 256-265.

Letourneau, I.J., Deeley, R.G., and Cole, S.P. (2005). Functional characterization of non-synonymous single nucleotide polymorphisms in the gene encoding human multidrug resistance protein 1 (MRP1/ABCC1). Pharmacogenet Genomics *15*, 647-657.

Li, B., Krishnan, V.G., Mort, M.E., Xin, F., Kamati, K.K., Cooper, D.N., Mooney, S.D., and Radivojac, P. (2009). Automated inference of molecular mechanisms of disease from amino acid substitutions. Bioinformatics *25*, 2744-2750.

Linard, B., Nguyen, N.H., Prosdocimi, F., Poch, O., and Thompson, J.D. (2012). EvoluCode: Evolutionary Barcodes as a Unifying Framework for Multilevel Evolutionary Data. Evol Bioinform Online *8*, 61-77.

Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics *12*, 11.

Lloyd, J.W. (1987). Foundations of logic programming.

Lopez-Bigas, N., and Ouzounis, C.A. (2004). Genome-wide identification of genes likely to be involved in human genetic disease. Nucleic Acids Res *32*, 3108-3114.

Luu, T.D., Rusu, A., Walter, V., Linard, B., Poidevin, L., Ripp, R., Moulinier, L., Muller, J., Raffelsberger, W., Wicker, N.*, et al.* (2012). KD4v: comprehensible knowledge discovery system for missense variant. Nucleic Acids Res.

Masso, M., and Vaisman, II (2008). Accurate prediction of stability changes in protein mutants by combining machine learning with structure based computational mutagenesis. Bioinformatics *24*, 2002-2009.

Matthews, L., Gopinath, G., Gillespie, M., Caudy, M., Croft, D., de Bono, B., Garapati, P., Hemish, J., Hermjakob, H., Jassal, B.*, et al.* (2009). Reactome knowledgebase of human biological pathways and processes. Nucleic Acids Res *37*, D619-622.

Meyers, L.A., Benedikt, H.m., and Brian, K.H. (2005). Chapter 6 - Constraints on Variation from Genotype through Phenotype to Fitness. In Variation (Burlington, Academic Press), pp. 87-111.

Milan, Z. (1987). Management Support Systems: Towards Integrated Knowledge Management. Human Systems Management *7*, 59-70.

Mornet, E., and Simon-Bouy, B. (2004). [Genetics of hypophosphatasia]. Arch Pediatr *11*, 444-448.

Muggleton, S. (1991). Inductive logic programming. New Generation Computing *8*, 295-318.

Muggleton, S. (1995). Inverse entailment and progol. New Generation Computing *13*, 245-286.

Muggleton, S., and Feng, C. (1990). Efficient Induction Of Logic Programs. New Generation Computing.

Muggleton, S., King, R.D., and Stenberg, M.J.E. (1992). Protein secondary structure prediction using logic-based machine learning. Protein Engineering *5*, 647-657.

Muggleton, S., and Raedt, L.D. (1994). Inductive Logic Programming: Theory and Methods. Journal of Logic Programming *19/20*, 629--679.

Muilu, J., Peltonen, L., and Litton, J.E. (2007). The federated database--a basis for biobank-based post-genome studies, integrating phenome and genome data from 600,000 twin pairs in Europe. Eur J Hum Genet *15*, 718-723.

Mukherjee, A.K., Basu, S., Sarkar, N., and Ghosh, A.C. (2001). Advances in cancer therapy with plant based natural products. Curr Med Chem *8*, 1467-1486.

Murzin, A.G., Brenner, S.E., Hubbard, T., and Chothia, C. (1995). SCOP: a structural classification of proteins database for the investigation of sequences and structures. J Mol Biol *247*, 536-540.

Nair, R., Liu, J., Soong, T.T., Acton, T.B., Everett, J.K., Kouranov, A., Fiser, A., Godzik, A., Jaroszewski, L., Orengo, C.*, et al.* (2009). Structural genomics is the largest contributor of novel structural leverage. J Struct Funct Genomics *10*, 181-191.

Needleman, S.B., and Wunsch, C.D. (1970). A general method applicable to the search for similarities in the amino acid sequence of two proteins. J Mol Biol *48*, 443-453.

Ng, P.C., and Henikoff, S. (2003). SIFT: Predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812-3814.

Nguyen, H., Friedrich, A., Berthommier, G., Poidevin, L., Ripp, R., Moulinier, L., and Poch, O. (2008). Introduction du nouveau centre de données biomédicales Décrypthon. Paper presented at: CORIA.

Nguyen, T.P., and Ho, T.B. (2008). An integrative domain-based approach to predicting protein-protein interactions. J Bioinform Comput Biol *6*, 1115-1132.

Olund, G., Lindqvist, P., and Litton, J.-E. (2007). BIMS: an information management system for biobanking in the 21st century. IBM Syst J *46*, 171-182.

Oti, M., Huynen, M.A., and Brunner, H.G. (2008). Phenome connections. Trends Genet *24*, 103-106.

Oti, M., Huynen, M.A., and Brunner, H.G. (2009). The biological coherence of human phenome databases. Am J Hum Genet *85*, 801-808.

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.

Plewniak, F., Thompson, J.D., and Poch, O. (2000). Ballast: blast post-processing based on locally conserved segments. Bioinformatics *16*, 750-759.

Plotkin, G. (1970). A Note on Inductive Generalization. Machine Intelligence *5*, 153-163.

Prasad, T.S., Kandasamy, K., and Pandey, A. (2009). Human Protein Reference Database and Human Proteinpedia as discovery tools for systems biology. Methods Mol Biol *577*, 67-79.

Pruitt, K.D., Tatusova, T., and Maglott, D.R. (2005). NCBI Reference Sequence (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins. Nucleic Acids Res *33*, D501-504.

Quinlan, J.R., and Cameron-Jones, R.M. (1993). FOIL: A Midterm Report. In Proceedings of the European Conference on Machine Learning (Springer-Verlag).

Ramensky, V., Bork, P., and Sunyaev, S. (2002). Human non-synonymous SNPs: server and survey. Nucleic Acids Res *30*, 3894-3900.

Remm, M., Storm, C.E., and Sonnhammer, E.L. (2001). Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. J Mol Biol *314*, 1041-1052.

Robinson, P.N., and Mundlos, S. (2010). The human phenotype ontology. Clin Genet *77*, 525-534.

Rose, P.W., Beran, B., Bi, C., Bluhm, W.F., Dimitropoulos, D., Goodsell, D.S., Prlic, A., Quesada, M., Quinn, G.B., Westbrook, J.D.*, et al.* (2011). The RCSB Protein Data Bank: redesigned web site and web services. Nucleic Acids Res *39*, D392-401.

Salwinski, L., Miller, C.S., Smith, A.J., Pettit, F.K., Bowie, J.U., and Eisenberg, D. (2004). The Database of Interacting Proteins: 2004 update. Nucleic Acids Res *32*, D449-451.

Saunders, C.T., and Baker, D. (2002). Evaluation of structural and evolutionary contributions to deleterious mutation prediction. J Mol Biol *322*, 891-901.

Schaefer, C.F., Anthony, K., Krupa, S., Buchoff, J., Day, M., Hannay, T., and Buetow, K.H. (2009). PID: the Pathway Interaction Database. Nucleic Acids Res *37*, D674-679.

Shannon, P., Markiel, A., Ozier, O., Baliga, N.S., Wang, J.T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res *13*, 2498-2504.

Shapiro, E.Y. (1981). An algorithm that infers theories from facts. In Proceedings of the 7th international joint conference on Artificial intelligence - Volume 1 (Vancouver, BC, Canada, Morgan Kaufmann Publishers Inc.).

Sherry, S.T., Ward, M.H., Kholodov, M., Baker, J., Phan, L., Smigielski, E.M., and Sirotkin, K. (2001). dbSNP: the NCBI database of genetic variation. Nucleic Acids Res *29*, 308-311.

Smith, C.W., Patton, J.G., and Nadal-Ginard, B. (1989). Alternative splicing in the control of gene expression. Annu Rev Genet *23*, 527-577.

Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., and Edelman, M. (1999). Automated analysis of interatomic contacts in proteins. Bioinformatics *15*, 327-332.

Srinivasan, A. (2004). The Aleph Manual.

Stenson, P.D., Ball, E., Howells, K., Phillips, A., Mort, M., and Cooper, D.N. (2008). Human Gene Mutation Database: towards a comprehensive central mutation database. J Med Genet *45*, 124-126.

Stephen, M., Huma, L., Ata, A., and Michael, J.E.S. (2005). Support vector inductive logic programming (Springer-Verlag).

Sunyaev, S.R., Eisenhaber, F., Rodchenkov, I.V., Eisenhaber, B., Tumanyan, V.G., and Kuznetsov, E.N. (1999). PSIC: profile extraction from sequence alignments with position-specific counts of independent observations. Protein Eng *12*, 387-394.

Suzek, B.E., Huang, H., McGarvey, P., Mazumder, R., and Wu, C.H. (2007). UniRef: comprehensive and non-redundant UniProt reference clusters. Bioinformatics *23*, 1282-1288.

Szklarczyk, D., Franceschini, A., Kuhn, M., Simonovic, M., Roth, A., Minguez, P., Doerks, T., Stark, M., Muller, J., Bork, P.*, et al.* (2011). The STRING database in 2011: functional interaction networks of proteins, globally integrated and scored. Nucleic Acids Res *39*, D561-568.

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, W.R. (1986). The classification of amino acid conservation. J Theor Biol *119*, 205-218.

Thompson, J.D., Higgins, D.G., and 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 *22*, 4673-4680.

Thompson, J.D., Holbrook, S.R., Katoh, K., Koehl, P., Moras, D., Westhof, E., and Poch, O. (2005). MAO: a Multiple Alignment Ontology for nucleic acid and protein sequences. Nucleic Acids Res *33*, 4164-4171.

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.

Thompson, J.D., Plewniak, F., Ripp, R., Thierry, J.C., and Poch, O. (2001). Towards a reliable objective function for multiple sequence alignments. J Mol Biol *314*, 937-951.

Thompson, J.D., Plewniak, F., Thierry, J., and Poch, O. (2000). DbClustal: rapid and reliable global multiple alignments of protein sequences detected by database searches. Nucleic Acids Res *28*, 2919-2926.

Thompson, J.D., Prigent, V., and Poch, O. (2004). LEON: multiple aLignment Evaluation Of Neighbours. Nucleic Acids Res *32*, 1298-1307.

Thompson, J.D., Thierry, J.C., and Poch, O. (2003). RASCAL: rapid scanning and correction of multiple sequence alignments. Bioinformatics *19*, 1155-1161.

Thusberg, J., Olatubosun, A., and Vihinen, M. (2011). Performance of mutation pathogenicity prediction methods on missense variants. Hum Mutat *32*, 358-368.

Townson, S.M., Kang, K., Lee, A.V., and Oesterreich, S. (2006). Novel role of the RET finger protein in estrogen receptor-mediated transcription in MCF-7 cells. Biochem Biophys Res Commun *349*, 540-548.

Tranchevent, L.C., Barriot, R., Yu, S., Van Vooren, S., Van Loo, P., Coessens, B., De Moor, B., Aerts, S., and Moreau, Y. (2008). ENDEAVOUR update: a web resource for gene prioritization in multiple species. Nucleic Acids Res *36*, W377-384.

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. Nucleic Acids Res *37*, D555-559.

Waterhouse, A.M., Procter, J.B., Martin, D.M., Clamp, M., and Barton, G.J. (2009). Jalview Version 2--a multiple sequence alignment editor and analysis workbench. Bioinformatics *25*, 1189-1191.

Wicker, N., Dembele, D., Raffelsberger, W., and Poch, O. (2002). Density of points clustering, application to transcriptomic data analysis. Nucleic Acids Res *30*, 3992-4000.

Wicker, N., Perrin, G.R., Thierry, J.C., and Poch, O. (2001). Secator: a program for inferring protein subfamilies from phylogenetic trees. Mol Biol Evol *18*, 1435-1441.

Wu, C.H., Apweiler, R., Bairoch, A., Natale, D.A., Barker, W.C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R.*, et al.* (2006). The Universal Protein Resource (UniProt): an expanding universe of protein information. Nucleic Acids Res *34*, D187-191.

Yue, P., Melamud, E., and Moult, J. (2006). SNPs3D: candidate gene and SNP selection for association studies. BMC Bioinformatics *7*, 166.

Tien Dao LUU

Développement d'une infrastructure d'analyse multi-niveaux pour la découverte des relations entre génotype et phénotype dans les maladies génétiques humaines

Résumé

Répondant au besoin de mieux comprendre les relations qui lient un génotype aux phénotypes moléculaires et cliniques associés, nous avons développé une nouvelle infrastructure bioinformatique qui unit, dans un même système, la collecte, la gestion, la maintenance et le traitement de multiples données ou informations. La première contribution de cette thèse est SM2PH Central et sa capacité de générer des instances. SM2PH Central constitue notre centre de référence en ligne pour toutes les protéines humaines intégrant des niveaux d'informations qui vont des aspects génomiques, structuraux, fonctionnels ou évolutifs aux aspects de transcriptomique, interactomique, protéomique ou métabolomique. La deuxième contribution est MSV3d, une ressource d'annotation multi-niveau (propriétés physico-chimiques, fonction, évolution, structure) des mutations humaines connues. MSV3d fournit l'ensemble des connaissances exploitées par la troisième contribution de cette thèse à savoir KD4v, notre base d'extraction de connaissances pour prédire l'impact phénotypique d'une mutation. La base de connaissances de KD4v induite par la Programmation Logique Inductive contient des règles exploitables par un humain ou un ordinateur et des facteurs prédictifs caractérisant les mutations neutres ou délétères. Enfin, l'ultime contribution de cette thèse est liée au développement de GEPeTTO, un prototype de priorisation de gènes. Une application biologique a été réalisée. Nous avons étudié la cécité nocturne en utilisant SM2PH Central, en combinaison avec le service d'annotation de MSV3d et la méthode de prédiction KD4v pour analyser le gène GPR179 et ses deux mutations nouvellement identifiées.

Keywords : infrastructure bioinformatique, relations génotype–phénotype, SM2PH, MSV3d, KD4v

Summary

Responding to the need to better understand the relationships linking the genotype to the molecular and clinical phenotype, we have developed a new bioinformatics infrastructure that unites, in a single system, the collection, the management, the maintenance and the processing of multiple data or information. The first contribution of this thesis is SM2PH Central and its ability to generate instances. SM2PH Central is our online reference center for all human proteins including many levels of information such as genomics, structural, functional and evolutionary aspects of transcriptomics, interactomics, proteomics or metabolomics. The second contribution is MSV3d, a multi-level annotation resource (physico-chemical properties, function, evolution, structure) of known human mutations. MSV3d provides the knowledge used by the third contribution of this thesis namely KD4v, our knowledgebase extraction to predict the phenotypic effect of a mutation. The KD4v knowledgebase computed by Inductive Logic Programming contains the rules describing the information that can be either exploited by a human or a computer, and the predictors characterizing neutral or deleterious mutations. The last contribution of this thesis is related to the development of GEPeTTO, a prototype of the prioritization of genes. Finally, these tools (SM2PH Central, MSV3d, KD4v) allowed us in the context of patients data analysis to confirm the implication of GPR179 as a new gene responsible for congenital stationary night blindness.

Keywords : bioinformatics infrastructure, genotype-phenotype relationships, SM2PH, MSV3d, KD4v