Bases de données et outils bioinformatiques utiles en génétique

Collège National des Enseignants et Praticiens de Génétique Médicale

C. Beroud

Date de création du document 2010-2011

Table des matières

Avec le développement de la génétique et des nouvelles technologies à très haut débit, nous faisons actuellement face à la production de données à un niveau encore jamais atteint. En effet, il est aujourd'hui démontré que les données produites par les technologies de séquençage à haut débit seront plus importantes que tout ce qui a jamais été produit dans le passé y compris le web lui même ! Nous faisons donc face à de multiples challenges tant pour le stockage de ces données (les nouvelles plateformes de séquençage peuvent produire jusqu'à 0,1 téraoctets de données par heure) que pour leur analyse.

Heureusement, nous ne partons pas de zéro. La communauté scientifique a depuis longtemps compris que la bonne utilisation des données pouvait permettre d'accélérer les découvertes scientifiques et ceci a rapidement conduit à l'émergence d'une nouvelle discipline : la bioinformatique.

L'objectif de ce cours est donc de faire le point sur les apports de la bioinformatique notamment par les différentes bases de données et outils bioinformatiques qu'elle a permis de créer ces dernières années et qui sont aujourd'hui autant d'outils incontournables pour les généticiens.

I CONCEPTS

I.1 LA BIOINFORMATIQUE

Lors de sa création, la bioinformatique correspondait à l'utilisation de l'informatique pour stocker et analyser les données de la biologie moléculaire. Cette définition originale a maintenant été étendue et le terme bioinformatique est souvent associé à l'utilisation de l'informatique pour résoudre les problèmes scientifiques posés par la biologie dans son ensemble. Il s'agit dans tous les cas d'un champ de recherche multidisciplinaire qui associe informaticiens, mathématiciens, physiciens et biologistes.

Comme le décrit très bien **Jean-Michel Claverie** : "*La bioinformatique est constituée par l'ensemble des concepts et des techniques nécessaires à l'interprétation de l'information génétique (séquences) et structurale (repliement 3-D). C'est le décryptage de la "bio-information" ("Computational Biology" en anglais). La bioinformatique est donc une branche théorique de la Biologie. Son but, comme tout volet théorique d'une discipline, est d'effectuer la synthèse des données disponibles (à l'aide de modèles et de théories), d'énoncer des hypothèses généralisatrices (ex. :comment les protéines se replient ou comment les espèces évoluent), et de formuler des prédictions (ex. : localiser ou prédire la fonction d'un gène)*".

Pour aboutir à la formulation de ces modèles et à ces prédictions, il est indispensable de tout d'abord collecter et organiser les données à travers la création de bases de données.

I.2 LES BASES DE DONNÉES

Une base de données est un ensemble structuré et organisé permettant le stockage de grandes quantités d'informations afin d'en faciliter leur utilisation (ajout, mise à jour, recherche et éventuellement analyse dans les systèmes les plus évolués que nous verrons par la suite).

Elles sont toutes organisées en fonction d'un modèle de données (*data model*) qui peut être de différents types : modèle hiérarchique (*hierarchical model*), modèle en réseau (*network model*), modèle relationnel (*relational model*), modèle orienté objet (*object-oriented model*), modèle semi structuré *(semistructured model*), modèle associatif (*associative model*), modèle EAV (*Entity-Attribute-Value data model*) ou encore modèle contextuel (*context model*). [Pour en savoir plus : ***database models***].

L'un des modèles les plus utilisés aujourd'hui est le modèle de bases de données relationnelles qui a été inventé un 1970 par **Edgar Frank Codd.**
Ce modèle repose ainsi sur les 12 règles de Cood *(source Wikipédia)*:

Règle 1 : **Unicité** : Toute l'information dans la base de données est représentée d'une et une seule manière, à savoir par des valeurs dans des champs de colonnes de tables.

Règle 2 : **Garantie d'accès** : Toutes les données doivent être accessibles sans ambiguïté. Cette règle est essentiellement un ajustement de la condition fondamentale pour des clefs primaires. Elle indique que chaque valeur scalaire individuelle dans la base de données doit être logiquement accessible en indiquant le nom de la table contenante, le nom de la colonne contenante et la valeur principale primaire de la rangée contenante.

Règle 3 : **Traitement des valeurs nulles** : Le système de gestion de bases de données doit permettre à chaque champ de demeurer nul (ou vide). Spécifiquement, il doit soutenir une représentation "d'information manquante et d'information inapplicable" qui est systématique, distincte de toutes les valeurs régulières (par exemple, "distincte de zéro ou tous autres nombres," dans le cas des valeurs numériques), et ce indépendamment du type de données. Cela implique également que de telles représentations doivent être gérées par le système de gestion de bases de données d'une manière systématique.

Règle 4 : **Catalogue lui-même relationnel** : Le système doit supporter un catalogue en ligne, intégré, relationnel, accessible aux utilisateurs autorisés au moyen de leur langage

d'interrogation régulier. Les utilisateurs doivent donc pouvoir accéder à la structure de la base de données (catalogue) employant le même langage d'interrogation qu'ils emploient pour accéder aux données de la base de données.

Règle 5 : **Sous-langage de données** : Le système doit soutenir au moins un langage relationnel qui : a une syntaxe linéaire ; peut être employé interactivement et dans des programmes d'application ; supporte des opérations de définition d'informations supplémentaires (incluant des définitions de vues), de manipulation de données (mise à jour aussi bien que la récupération), de contraintes de sécurité et d'intégrité, et des opérations de gestion de transaction (commencer, valider et annuler une transaction).

Règle 6 : **Mise à jour des vues** : Toutes les vues pouvant théoriquement être mises à jour doivent pouvoir l'être par le système.

Règle 7 : **Insertion, mise à jour, et effacement de haut niveau** : Le système doit supporter les opération par lot d'insertion, de mise à jour et de suppression. Ceci signifie que des données peuvent être extraites d'une base de données relationnelle dans des ensembles constitués par des données issues de plusieurs tuples et/ou de multiples table. Cette règle explique que l'insertion, la mise à jour, et les opérations d'effacement devraient être supportées aussi bien pour des lots de tuples issues de plusieurs tables que juste pour un tuple unique issu d'une table unique.

Règle 8 : **Indépendance physique** : Les modifications au niveau physique (comment les données sont stockées, si dans les rangées ou les listes liées etc...) ne nécessitent pas un changement d'une application basée sur les structures.

Règle 9 : **Indépendance logique** : Les changements au niveau logique (tables, colonnes, rangées, etc) ne doivent pas exiger un changement dans l'application basée sur les structures. L'indépendance de données logiques est plus difficile à atteindre que l'indépendance de donnée physique.

Règle 10 : **Indépendance d'intégrité** : Des contraintes d'intégrité doivent être indiquées séparément des programmes d'application et être stockées dans le catalogue. Il doit être possible de changer de telles contraintes au fur et à mesure sans affecter inutilement les applications existantes.

Règle 11 : **Indépendance de distribution** : La distribution des parties de la base de données à de diverses localisations doit être invisible aux utilisateurs de la base de données. Les applications existantes doivent continuer à fonctionner avec succès : quand une version distribuée du système de gestion de bases de données est d'abord présentée ; et quand des données existantes sont redistribués dans le système.

Règle 12 : **Règle de non-subversion** : Si le système fournit une interface de bas niveau, cette interface ne doit pas permettre de contourner le système (par exemple une contrainte relationnelle de sécurité ou d'intégrité).

Afin de créer ces banques de données relationnelles, il est nécessaire d'avoir recours à un système informatique nommé Système de Gestion de Bases de Données Relationnel (SGBDR) dont les plus connus sont : ***Oracle, Access, SQLServer, Informix, Sybase, DB2, MySQL, 4D, Filmaker…***

Ces SGBDR permettent alors d'accéder à la base de données directement via Internet afin d'en assurer la diffusion la plus large possible.

II LES BANQUES DE DONNÉES UTILES DANS LE DOMAINE DE LA GÉNÉTIQUE

II.1 LES "GENOME BROWSERS"

Ils correspondent à différentes bases de données qui permettent d'accéder aux données du génome humain (et de celui d'autres espèces) à l'aide d'une interface graphique. En plus des données de séquence, ces navigateurs permettent d'accéder à de nombreuses données d'annotation (gènes avec exons et introns, sites de fixation, régions d'homologie) ((cf. 3.1 :)).

Les plus populaires sont :

- **Ensembl** *(European Bioinformatics Institute / Wellcome Trust Sanger Institute)*
- **NCBI** *(National Cancer for Biology Information)*
- **UCSC** *(University of California Santa Cruz)*

D'autres méritent également le détour :

- **Vista** *(University of California)*
- **Argo** *(BROAD Institute)*
- **Mochiview** *(University of California Santa Cruz)*
- **X :map** *(Paterson Institute for Cancer Research)*
- **DiProGB** *(Leibniz Institute for Age Research)*
- **Genatlas** *(Université René Descartes - Paris)*

Si l'ensemble des "*Genome Browsers*" permet d'accéder à de très nombreuses données, aucun d'entre eux ne génère ces données. Ils sont donc dépendants d'autres centres ou laboratoires de recherche qui eux les produisent. Ceci explique pourquoi les mêmes données sont partagées par ces différents navigateurs et c'est souvent l'interface qui oriente vers l'un plutôt que l'autre ou la richesse des outils d'analyse associés.

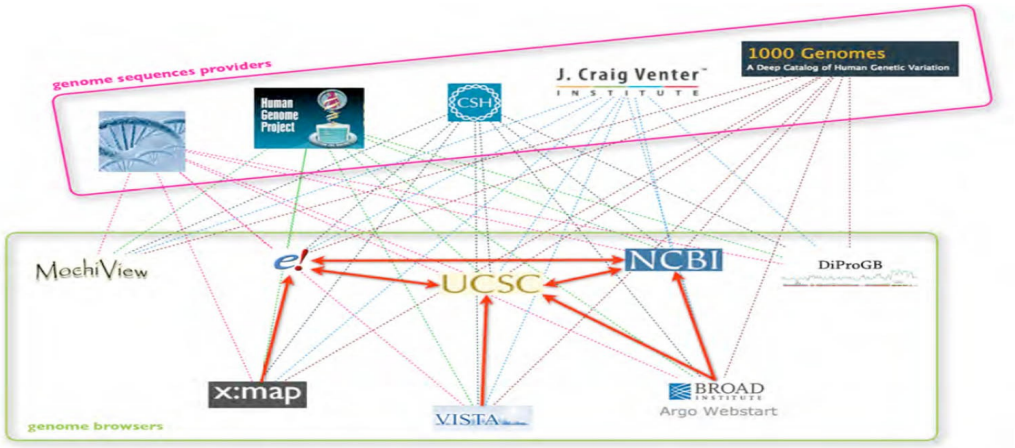
Il existe cependant des "*Genome Browsers*" dédiés à un projet de recherche particulier. Dans ce cas, leur champ d'action est plus réduit mais ils fournissent directement les données et sont donc responsables de leur qualité. Il est en effet critique de s'assurer de la qualité des données collectées dans une base de données car si elle est ouverte à tous, sa qualité ne pourra être assurée et les données qu'elle contient seront vite d'une utilité limitée comme nous le verrons dans le chapitre dédiée aux banques de données de mutations ((cf. 2.5.1 :)).

Trois bases de données illustrent bien cette catégorie :

- ***James Watson's Personal Genome Sequence*** *(Baylor College of Medicine)*
- ***Craig Venter's Personal Genome Sequence*** *(Craig Venter Institute)*
- ***1000 genomes project*** *(Projet international)*

Comme nous l'avons vu, les différents "*Genome Browsers*" partagent des données brutes (séquence de référence) mais également des données d'annotation. Comme le montre **la figure 1**, il existe ainsi des relations complexes entre les fournisseurs de données et les "*Genome Browsers*".

Figure 1 : Représentation des liens entre les "Genome Browsers" et les fournisseurs de données.



Rectangle rose = fournisseurs de données : centres de séquençage académiques et privés, centres de séquençage et d'assemblage du projet génome humain, projets de séquençage de génomes personnels (James Watson, Craig Venter …), projet 1 000 génomes. Rectangle vert

= Genome Browsers. Lignes pointillées = données utilisées par les génomes Browsers. Flèches rouges = liens entre les différents Genome Browsers.

II.2 L'ANNOTATION : OUTILS ET BASES DE DONNÉES

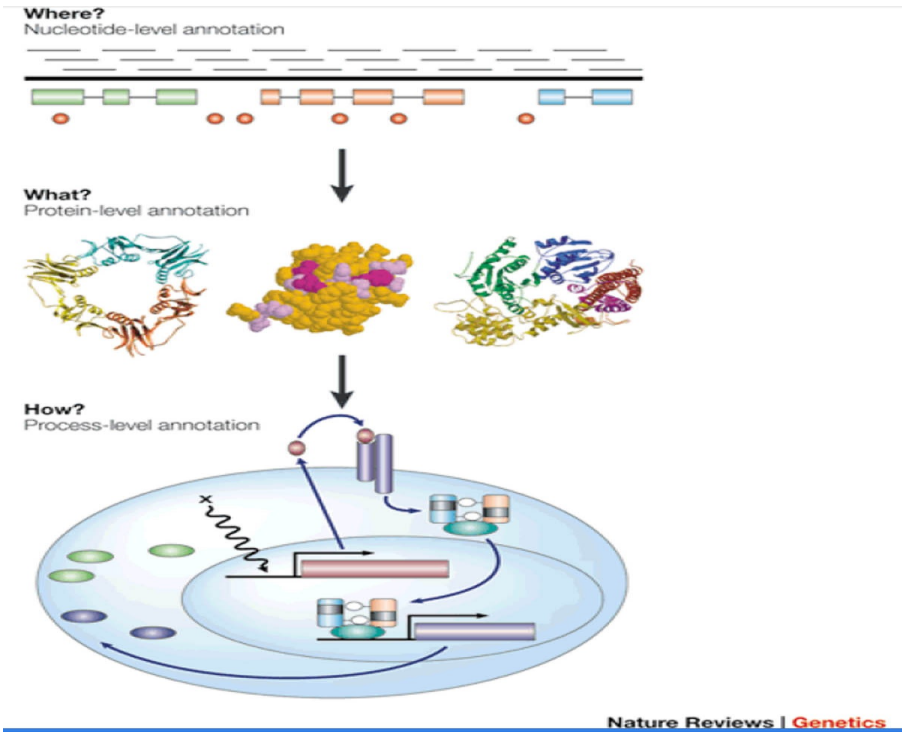
La connaissance de la séquence du génome humain n'aurait qu'une portée limitée si elle n'était annotée à différents niveaux. Ainsi l'annotation est un processus complexe qui peut être subdivisé en trois catégories : l'annotation syntaxique, l'annotation fonctionnelle et l'annotation relationnelle (figure 2) :

L'annotation syntaxique qui permet d'identifier les séquences présentant une pertinence biologique (gènes, signaux, répétitions, …)

L'annotation fonctionnelle qui permet de prédire les fonctions et produits potentiels des gènes préalablement identifiés (similitudes de séquences, motifs, structures, ...) et de collecter d'éventuelles informations expérimentales (littérature, jeux de données à grande échelle, …)

L'annotation relationnelle qui permet enfin de déterminer les interactions que les objets biologiques préalablement identifiés sont susceptibles d'entretenir (familles de gènes, réseaux de régulation, réseaux métaboliques, …).

Figure 2 : Représentation des différents niveaux d'annotation (d'après Lincoln Stein, Nature Reviews Genetics 2, 493-503 2001).



Where? = annotation syntaxique ; What? = annotation fonctionnelle; How? = annotation relationnelle.

II.3 STRUCTURE DES PROTÉINES

Parmi les différents outils d'annotation fonctionnelle, attachons nous à ceux en relation avec la structure des protéines puisque cette connaissance sera d'un apport primordial pour l'interprétation des mutations responsables de maladies génétiques.

Nous pouvons distinguer plusieurs niveaux dans la description de la structure des protéines :

- **La structure primaire :** elle correspond à la séquence des acides aminés constituant la protéine. Il s'agit d'un assemblage linéaire des acides aminés codés par l'ARN messager.
- **La structure secondaire** : elle décrit un niveau structural plus complexe : les structures secondaires qui sont représentées par les repliements locaux de la protéine. Elle comporte les structures en hélices (α, 310, π, type II) et les feuillets (β parallèles et antiparallèles) et enfin les coudes (types I, II, III et γ).
- **La structure tertiaire** : décrit la structure tridimensionnelle de la protéine ou plus précisément d'une forme particulière que peut prendre dans l'espace la protéined'intérêt dans des conditions expérimentales données et ceci à un temps t.
- **La structure quaternaire** : permet de décrire les interactions entre protéines.

Les différents outils et bases de données que nous avons sélectionnés permettent de collecter les informations en relation avec le protéines à ces différents niveaux (lorsque des informations sont disponibles ce qui est toujours vrai pour la séquence primaire mais peu fréquent pour la séquence tertiaire et encore plus rare pour la séquence quaternaire). Parallèlement à ces données classiques, des annotations complémentaires sont de plus en plus fréquemment disponibles (domaines protéiques en relation avec une structure ou une fonction particulières, structure de protéines mutantes …). Comme vous le constatez, nous associons ici outils et bases de données qui sont en effet indissociables dans le cas des structures puisque les données brutes ne sont pas directement interprétables par l'homme et nécessitent l'utilisation d'outils de visualisation.

Les plus populaires sont :

- **Uniprot/Swiss Prot/Expasy** *(Uniprot Consortium)*
- **Protein Data Bank** *(Research Collaboratory for Structural Bioinformatics)*
- **Topspan** *(Open Protein Structure Annotation Network)*
- **NCBI** *(National Cancer for Biology Information)*
- **PDBsum** *(European Bioinformatics Institute)*

D'autres bases de données sont particulièrement utiles pour identifier des domaines protéiques présents chez plusieurs protéines et ainsi définir des familles et des superfamilles de protéines :

- **CATH protein structure classification** *(University College London)*
- **Pfam** *(Wellcome trust Sanger Institute)*
- **Protein Information Resource** *(University of Delaware / Georgetown University Medical Center)*
- **Structure Function Linkage Database** *(University of California, San Francisco)*

II.4 LES BASES DE DONNÉES DÉDIÉES AUX MALADIES GÉNÉTIQUES

La base de données de référence pour les maladies génétiques est sans conteste **OMIM** *(Online Mendelian Inheritance in Man)*. Cette base de données est née dans les années 1960 grâce au travail de Victor McKusick qui est souvent surnommé "the father of medical genetics" et qui a patiemment et sans relâche démontré l'importance de l'étude des bases génétiques des maladies :"*I like to say that the arrangement of genes on chromosomes is part of the micro-anatomy, just as the gross anatomy in the Middle Ages was important to medicine, every medical specialty now uses mapping genes for diseases*".

Il a également été l'un des premiers à comprendre la puissance de la bioinformatique et la nécessité d'organiser le savoir médical sous la forme de bases de données. La version Internet de son oeuvre a été créée en 1985 et est aujourd'hui encore la référence internationale. Il nous a quittés en 2008.

Parallèlement à OMIM, il existe d'autres bases de données dédiées aux maladies génétiques. Citons par exemple :

- ***GeneCards*** *(Weizmann Institute of Science) qui a pour porte d'entrée le gène mais qui permet également d'obtenir des données sur les maladies associées (5551 gènes sont associés à un phénotype clinique).*
- ***Office of Rare Diseases Research (National Institute of Health). Ce site est dédié aux maladies rares et a un champ d'utilisation non restreint aux scientifiques puisqu'il s'adresse aussi bien aux chercheurs qu'aux cliniciens ou aux patients.***
- ***Orphanet*** *(INSERM)*
- ***MEDGENE*** *(Harvard Medical School) Medline*

A côté de ces bases de données généralistes, il existe nombre de bases de données dont le champ d'application est plus étroit. Citons par exemple :

- ***HuGE Navigator*** *(National Office of Public Health Genomics Centers for Disease Control and Prevention)*
- ***Infevers*** *(Institut de Génétique Humaine - Montpellier* **)**

II.5 VARIABILITÉ DU GÉNOME HUMAIN

Avec l'essor des nouvelles technologies, le nombre de variations de la séquence du génome humain ne cesse de croître. Ainsi le séquençage du génome complet d'un individu permet aujourd'hui d'identifier environ 3 millions de SNPs (*Single Nucleotide Polymorphisms*) dont 20 à 25% n'ont jamais été décrits auparavant. La collection de ces informations est d'un intérêt majeur, non seulement pour la recherche mais également pour le diagnostic des maladies génétiques.

La grande difficulté est actuellement de collecter des données très hétérogènes tant par leur mode de production (quel technologie a été utilisée ?) que par leur qualité (quels étaient les paramètres qualités employés ?). Comme nous allons le voir, il existe de nombreuses bases de données permettant d'accéder à des informations sur la variabilité de la séquence du génome humain mais il n'existe pas (encore) une base de données idéale.

Deux approches ont été retenues par différents groupes : l'approche généraliste (les données sont collectées pour l'ensemble des gènes) et l'approche spécialisées (les données sont collectées pour un gène donné).

II.5.1 Les bases de données centrales

Elles permettent d'accéder rapidement à des données relatives à la variabilité de séquence d'un gène quelconque.

Nous pouvons distinguer plusieurs types de bases de données en fonction du type de mutation (ici pris dans son sens littéral c'est à dire toute variation stable de la séquence) : celles dédiées aux SNPs *(Single Nucleotide Polymorphism*), aux CNVs (*Copy Number Variation*) et celles dédiées aux mutations pathogènes

II.5.1.1 Les bases de données centrales dédiées aux SNPs

Nous illustrerons ce type de base de données avec trois modèles complémentaires :

- **dbSNP** *(National Cancer Bioinformatics Institute)*
- **Allele FREquency Database** *(Yale University)*
- **HapMap** *(Projet international)*

II.5.1.2 Les bases de données centrales dédiées aux CNVs

Les CNVs sont connus depuis longtemps mais l'émergence des technologies à très haut débit comme l'hybridation génomique comparative (CGH) sur puces (microarray CGH) ont véritablement révélé un aspect insoupçonné de la variabilité du génome humain : des variations de fragments de séquence de plusieurs centaines de milliers de paires de bases. Ces données ainsi que leurs conséquences phénotypiques (certains CNVs sont pathogènes, d'autres pas) sont répertoriés dans plusieurs bases de données dont voici quelques exemples :

- **CNVVdb** *(Academia Sinica - Taiwan)*
- **DGV** *(Department of Genetics and Genomic Biology - Toronto)*
- **DECIPHER** *(Wellcome Trust Sanger Institute) (Database of Chromosomal Imbalance and Phenotype in Humans Using Ensembl Resource)*

II.5.1.3 Les bases de données centrales dédiées aux mutations pathogènes

Dans le domaine de la génétique humaine, ce sont bien sûr les mutations pathogènes qui sont de la plus grande importance puisqu'elles sont responsables de maladies génétiques. Leur connaissance est ainsi essentielle tant pour le conseil génétique que pour la compréhension des mécanismes moléculaires responsables de pathologies voire même pour la création de nouvelles approches thérapeutiques.

Différentes bases de données ont pour objet de collecter ces mutations pathogènes à l'échelle du génome :

- **HGMD** *(Institute of Medical Genetics - Cardiff)*
- **OMIM** *John Hopkins University – National Cancer Bioinformatics Institute*

II.5.2 Les bases de données spécifiques de locus

Plus connues sous l'acronyme de LSDB (Locus Specific DataBase), elles sont développées par des experts d'un gène ou de maladies et sont donc considérées comme les bases de données de référence pour un gène donné.

Leur qualités principales résident dans la validation des données qu'elles contiennent par des experts du domaine considéré ainsi que par leur exhaustivité (jusqu'à 50% de leur contenu peut correspondre à des soumissions directes non publiées et ainsi absent des bases de données centrales). La liste des différentes LSDBs disponibles via Internet peut être retrouvée sur le site de la Human Genome Variation Society *(HGVS)*.

III OUTILS INFORMATIQUES UTILES DANS LE DOMAINE DE LA GÉNÉTIQUE

Comme vous pouvez l'imaginer à la vue du nombre et de la diversité des bases de données disponibles via Internet, les outils bioinformatiques disponibles sont également très nombreux allant de la prédiction de gènes à partir d'une séquence quelconque à l'identification de motifs particuliers (sites de fixation de protéines, etc.) ou à la prédiction du caractère pathogène d'une mutation faux-sens.

Ne pouvant traiter ici l'ensemble des outils bioinformatiques disponibles, j'ai choisi de limiter ce paragraphe aux différents outils de prédiction pouvant être directement utiles pour apporter une aide à l'interprétation du caractère pathogène ou non d'une variation de séquence découverte dans le cadre d'un diagnostic moléculaire. En effet, la révolution génomique (séquençage complet d'un ou plusieurs gènes) aboutie à l'identification de nombreuses variations de séquence et il est souvent difficile d'identifier la ou les mutations réellement pathogènes.

La plupart des gènes humains codent pour des protéines et c'est tout naturellement que les outils de prédiction se sont attachés à la protéine plutôt qu'au gène lui même à l'exception de quelques outils comme nous le verrons par la suite. Dans une situation idéale, la structure 3D de la protéine est disponible et de nombreux orthologues ont également été décrits. Bien entendu cela est encore loin d'être le cas, limitant ainsi l'intérêt de certains outils.

III.1 PRÉDICTION DES CHANGEMENTS DE STABILITÉ DES PROTÉINES

Tous les outils de cette catégorie nécessitent la disponibilité d'une structure 3D de la protéine elle-même ou de l'un de ses orthologues. Les algorithmes utilisés ont des performances hétérogènes tant en terme de rapidité que de prédiction. Les plus connus sont

- **Cupsat** *(Cologne University)*
- **FoldX** *(European Molecular Biology Laboratory – Heidelberg*

III.2 PRÉDICTION DE L'AGRÉGATION DES PROTÉINES

L'agrégation est un terme général qui regroupe différents types d'interactions ou caractéristiques. Ainsi l'agrégation des protéines peut survenir via différents mécanismes et peut être classée de différents façons : soluble/insoluble, covalence/non-covalence, réversible/irréversible, natif/dénaturé. Elle survient par la formation d'un lien chimique entre 2 (ou plus) monomères : la création de ponts disulfures est un mécanisme fréquent mais d'autres liens peuvent également être

observés comme la formation de bi-tyrosines après un phénomène d'oxydation des tyrosines, etc. Deux outils de prédiction sont souvent utilisés dans ce domaine :

- **Aggrescan** *(Universitat Autónoma de Barcelona)*
- **Tango** *European Molecular Biology Laboratory – Heidelberg)*

III.3 PRÉDICTION DES RÉGIONS DÉSORDONNÉES

Les régions désordonnées (DR) correspondent à des régions protéiques qui ne possèdent pas de structure tertiaire fixe. Elles sont ainsi partiellement ou totalement non repliées. Il a été démontré que de telles régions étaient impliquées dans une grande variété de fonctions comprenant la reconnaissance de l'ADN, la modulation de la spécificité ou de l'affinité de la liaison à d'autres protéines, l'activation par protéolyse, le contrôle de la demi-vie des protéines etc. Bien que ces régions ne possèdent pas de structure 3-D fixe dans leur état natif, elles vont souvent faire l'objet de transitions entre divers états (DR/3-D) lors d'interactions.

Deux outils peuvent être utilisés pour ces prédictions : PONDR *(Molecular kinetics - Indianapolis)* et Disprot *(Indiana University school of medicine).*

III.4 PRÉDICTION DU CARACTÈRE PATHOGÈNE DES MUTATIONS FAUX-SENS

Les mutations faux-sens représentent plus de la moitié des mutations pathogènes décrites dans les maladies génétiques humaines et plus de la moitié des variations de séquence non-pathogènes. Leur interprétation est souvent délicate ce qui a conduit à la création d'outils de prédiction dont les principaux sont présentés ici :

- **SIFT** *(Craig Venter Institute)*
- **Polyphen** *(Harvard University)*
- **UMD-Predicto** *(INSERM*

III.5 PRÉDICTION DU CARACTÈRE PATHOGÈNE DES MUTATIONS INTRONIQUES

Il existe nombre de mutations qui sont localisées aux jonctions intron/exon/intron et il a été démontré qu'elles altèrent l'épissage des introns en détruisant certains signaux clés : les sites donneurs et accepteurs d'épissage. De la même façon, des mutations introniques localisées à distance des exons peuvent être pathogènes par la création de nouveaux signaux d'épissage reconnus par la machinerie cellulaire. Ces sites nouveaux sont nommés sites cryptiques. Enfin, il serait trop restrictif de limiter les signaux d'épissage aux simples sites donneurs et accepteurs d'épissage. Il existe en effet

d'autres signaux qui jouent un rôle clé comme le point de branchement situé en 5' du site accepteur, les ESE (*Exonic Splicing Enhancer)* et ESS *(Exonic Splicing Silencer)* localisés dans les exons, ou les ISE (*Intronic Splicing Silencer)* et ISS *(Intronic Splicing Silencer)* localisés dans les introns.

La connaissance de ces signaux est encore incomplète mais il existe d'ores et déjà des outils de prédiction de ces signaux qui peuvent également prédire l'impact d'une mutation quelconque
(exonique ou intronique) sur les signaux d'épissage. L'outil le plus utilisé est aujourd'hui **HSF** (*Human Splicing Finder*) qui intègre l'ensemble des algorithmes et matrices de prédiction et permet ainsi de disposer d'un large éventail de prédictions en un seul endroit.

IV EXEMPLES

IV.1 INTERPRÉTATION D'UNE MUTATION SYNONYME

Vous travaillez dans un laboratoire de diagnostic qui s'intéresse au gène LAMA2 et plus particulièrement à la mutation c.7572G>A (p.Glu2534Glu). Vous devez tout d'abord rechercher des données sur le gène LAMA2 et sur les pathologies associées à des mutations du gène LAMA2 et ensuite évaluer le caractère pathogène de cette mutation.

a) Recherche des informations relatives au gène LAMA2. Nous pouvons pour cela utiliser différents navigateurs, nous choisirons dans cet exemple le site du NCBI
cliquez sur le lien : (http://www.ncbi.nlm.nih.gov/) , tapez LAMA2 dans le critère de recherche, vous obtenez :
Résultat : http://www.ncbi.nlm.nih.gov/gquery/?term=lama2

b) Vous avez maintenant accès à de nombreuses bases de données. Cliquez sur OMIM (19 résultats) en haut à droite. Vous obtenez :
Résultat : http://www.ncbi.nlm.nih.gov/omim?term=lama2

c) Cliquez sur le lien *156225 (premier de la liste), vous obtenez :
Résultat : http://omim.org/entry/156225)

d) Vous apprenez ainsi que le gène LAMA2, localisé sur le brin + du chromosome 6 sur un fragment de 633 kb (129,204,285-129,837,710) code pour la chaine alpha-2 de la laminie-2. Les mutations de ce gène sont responsables de près de 50% des dystrophies musculaires congénitales … Pour en savoir plus sur le gène LAMA2 lui même, cliquez sur Genome dans la table à droite et sélectionner le navigateur),

vous obtenez :
Résultat : http://www.ensembl.org/Homo_sapiens/Location/View?db=core;g=ENSG000001 96569;r=6:129204286-129837714;t=ENST00000421865)

e) Dans le cadre "Region in details", cliquez sur LAMA2, vous pouvez maintenant accéder à des informations plus détaillées sur le gène et sa structure via le lien ENSG00000196569. Vous obtenez :
Résultat : http://www.ensembl.org/Homo_sapiens/Gene/Summary?db=core;g=ENSG0000 0196569;r=6:129204286-129837714;t=ENST00000421865)

f) Dans la liste des différents transcrits, repérez celui qui possède un CCDS (le premier). Cliquez sur le lien ENST00000421865, vous obtenez :
Résultat : http://www.ensembl.org/Homo_sapiens/Transcript/Summary?db=core;g=ENSG0 0000196569;r=6:129204286-129837714;t=ENST00000421865)

g) Dans la liste de gauche "transcript-based displays", vous pouvez maintenant accédez à la séquence des 65 exons du gène. Si vous avez été attentifs, vous constaterez que ce nombre est différent de celui de OMIM qui rapportait 64 exons, nombre erroné décrit lors de l'identification du gène. Cliquez sur exons,vous obtenez alors la séquence de référence du gène LAMA2 ainsi que sa structure intron/exon :
Résultat : http://www.ensembl.org/Homo_sapiens/Transcript/Exons?db=core;g=ENSG000 00196569;r=6:129204286-129837714;t=ENST00000421865)

Dans la liste de gauche "transcript-based displays", si vous cliquez sur "Variations", vous obtenez une liste de variations rapportées dans HGMD et dans dbSNP. Vous constaterez alors que votre mutation est inconnue. Que faire pour avancer ?

Premier élément, positionner la mutation dans le gène. La nomenclature de la mutation vous permet de positionner facilement la mutation c.7572G>A (p.Glu2534Glu) sur la dernière base de l'exon 54. Vous pouvez maintenant utiliser les outils de prédiction. La mutation étant une mutation synonyme, les outils tels que *SIFT et Polyphen* ne vous apporteront rien. Vous pouvez éventuellement utiliser le logiciel *UMD* pour bénéficier des prédictions d'UMD Predictor qui vous indiquera alors qu'il s'agit d'une mutation pathogène mais il y a plus simple. Pour cela, utilisez HSF, la dernière base d'un exon faisant en effet parti du site donneur d'épissage, vous pouvez bénéficier des prédictions de cet outil.

Pour cela rendez-vous sur (http://umd.be/HSF/)

Choisissez les paramètres suivants :
"Analysis type" = "Analyze mutation(s)"

"Number of nucleotide surrounding the exon" = 50
"Choose a sequence by" = "Gene name (e.g. DMD)" et tapez LAMA2 puis dans la boîte, saisissez le nom de la mutation (c.7572G>A)

Vous obtenez de nombreuses prédictions, concentrez-vous sur celles dédiées aux sites d'épissage :

Figure 3 : prédictions de l'impact de la mutation c.7572G>A du gène LAMA2 sur les signaux d'épissage (sites donneurs et accepteurs)

Potential splice sites ↕

HSF Matrices

Sequence Position	cDNA Position	Splice site type	Motif	New splice site	Wild Type	Mutant	If cryptic site use, exon length variation	Variation (%)
160	c.7561	Acceptor	TGTTCCCTGGAGgt	tgttccctggaaGT	85.42	56.48	NA	Site broken -33.89
161	c.7562	Acceptor	GTTCCCTGGAGgtt	gttccctggaagTT	48.12	77.07	NA	New site +60.15
169	c.7570	Donor	GAGgttggt	GAAgttggt	84.88	74.3	0	WT site broken -12.46

MaxEnt

Threshold values:
5' Motif: 0 3' Motif: 0

Sequence Position	cDNA Position	5' Motif					3' Motif				
		Ref Motif	Ref Score	Mut Motif	Mut Score	Variation (%)	Ref Motif	Ref Score	Mut Motif	Mut Score	Variation (%)
169	c.7570	GAGgttggt	6.36	GAAgttggt	-0.08	-101.26					
171	c.7572						Ggttggtctgtttttgatagttc	4.81	agttggtctgtttttgatagTTC	4.94	+2.7

Le site donneur d'épissage sauvage est présenté sur fond blanc (sequence position 169). Il est très important de connaître les paramètres de chaque outil pour évaluer la conséquence de la mutation sur les sites d'épissage. Ainsi pour l'algorithme HSF une variation de ± 10% est significative.

Vous constatez qu'avec les 2 algorithmes utilisés (HSF Matrices et MaxEnt) le site donneur sauvage est inactivé par la mutation. Il ne vous reste plus qu'à étudier l'ARN messager pour confirmer ces prédictions.

Avec cet exemple, vous avez vu qu'il était très simple de naviguer entre les différentes bases de données (du NCBI à OMIM et à ENSEMBL) afin d'y collecter des informations à de très nombreux niveaux. Vous avez également constaté qu'il était très simple d'obtenir une prédiction (juste pour cet exemple mais n'oubliez pas qu'il s'agit de prédictions) du caractère pathogène d'une mutation synonyme.

IV.2 INTERPRÉTATION DE MUTATIONS FAUX-SENS

Vous travaillez dans un laboratoire de diagnostic qui s'intéresse au gène FBN1 et vous avez identifié chez 5 malades une série de 6 mutations faux-sens : c.3623T>C

(p.Cys875Arg) ; c.1147G>A (p.Glu383Lys) ; c.8339T>C (p.Leu2780Pro) ; c.3413G>C (p.Cys1138Ser) ; c.6881A>C (p.Glu2294Ala) et c.7016G>A (p.Cys2339Tyr).

Vous devez bien sûr évaluer le caractère pathogène de ces mutations, 5 d'entre elles devant être pathogènes et une non.

a) Vous pouvez dans un premier temps rechercher si certaines de ces mutations ont déjà été décrites. Rendez-vous sur HGMD puis saisissez FBN1 et sélectionner "missense mutations". Vous obtenez les informations suivantes :

Tableau 1: présence des mutations du gène FBN1 dans la banque centrale HGMD

Mutation	**HGMD**
c.3623T>C (p.Cys875Arg)	Non
c.1147G>A (p.Glu383Lys)	Non
c.8339T>C (p.Leu2780Pro)	Oui
c.3413G>C (p.Cys1138Ser)	Non
c.7048A>G (p.Ile2350Val)	Non
c.7016G>A (p.Cys2339Tyr)	Oui

b) Vous ne disposez que d'informations pour 33% de vos mutations. Si vous vous rendez sur Ensembl comme dans l'exemple précédent, vous ne récupérez aucune information nouvelle, c'est à dire qu'aucune des 4 mutations restantes n'est documentée dans dbSNP. Vous n'en saurez pas plus en passant par d'autres bases de données centrales (Swissprot, etc.). Rendez-vous alors sur HGVS ((http://www.hgvs.org/dblist/glsdb.html) - F). Vous découvrez qu'il existe une LSDB dédiée au gène FBN1, rendez-vous y (http://www.umd.be/FBN1). Vous constaterez alors que toutes les mutations y sont rapportées, 5 étant considérées comme pathogènes et une comme polymorphisme. Cela simplifie considérablement votre travail !
Attention, si vous ne connaissez pas la qualité de la LSDB à laquelle vous venez de vous connecter, vous pouvez néanmoins aller plus loin et utiliser les logiciels de prédiction.

Dans tous les cas, reportez-vous à la publication d'origine pour évaluer la pertinence des arguments qui ont permis aux auteurs de décrire la mutation comme une mutation pathogène ou un polymorphisme. Utilisez donc les trois outils de prédiction les plus connus : SIFT, Polyphen et UMD Predictor (via la LSDB UMD-FBN1).

Utilisez le "NCBI GI number" 281485550 pour SIFT (http://sift.bii.astar.edu.sg/www/SIFT_BLink_submit.html)
Utilisez le "Protein identifier" 281485550 pour Polyphen (http://genetics.bwh.harvard.edu/pph)
Pour UMD Predictor suivez le lien http://www.umd.be/FBN1/ puis consultez chaque fiche individuellement, au total, vous devriez obtenir :

Tableau 2 : bilan des prédictions obtenues avec les systèmes SIFT, Polyphen et UMD-Predictor.

Mutation	**SIFT**	**Polyphen**	**UMD-Predictor**
c.3623T>C (p.Cys875Arg)	**Non pathogène**	*Probablement pathogène*	*Pathogène*
c.1147G>A (p.Glu383Lys)	**Non pathogène**	**Non pathogène**	*Pathogène*
c.8339T>C (p.Leu2780Pro)	*Pathogène*	*Probablement pathogène*	*Probablement pathogène*
c.3413G>C (p.Cys1138Ser)	*Pathogène*	*Probablement pathogène*	*Pathogène*
c.7048A>G (p.Ile2350Val)	*Non pathogène*	*Benign*	*Non pathogène*
c.7016G>A (p.Cys2339Tyr)	**Non pathogène**	*Probablement pathogène*	*Probablement pathogène*

En gras sont présentées les prédictions incorrectes, en italique, les prédictions correctes

Comme vous le constatez à la vue de ce tableau, les prédictions sont différentes d'un système à l'autre. Dans cet exemple très limité, seul l'un des 3 systèmes atteint 100% de bonnes prédictions.

Attention, cela ne veut pas dire que les autres systèmes ne peuvent pas fournir de meilleures prédictions dans d'autres situations.
En conclusion, l'utilisation de bases de données couplée aux outils informatiques librement accessibles via Internet a permis d'obtenir une aide précieuse à l'interprétation des résultats générés dans le cadre d'un diagnostic.