MyGeneFriends : vers un nouveau rapport entre chercheurs et mégadonnées

Alexis Allot

UNIVERSITÉ DE STRASBOURG

ÉCOLE DOCTORALE DES SCIENCES DE LA VIE ET DE LA SANTÉ

LBGI – BFO ICube UMR7357

THÈSE présentée par :

Alexis ALLOT

soutenue le 9 **octobre 2015**

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

Discipline Science de la vie et de la santé, Spécialité Bioinformatique

MyGeneFriends :

vers un nouveau rapport entre chercheurs et mégadonnées

THÈSE dirigée par :

Mme LECOMPTE Odile	Maître de conférences, Université de Strasbourg
M POCH Olivier	Directeur de recherche, CNRS ICUBE

RAPPORTEURS :

M CHALMEL Frédéric	Chargé de recherches, INSERM Université de Rennes 1
M XENARIOS Ioannis	Professeur, SIB, Université de Lausanne

AUTRES MEMBRES DU JURY :

M GANÇARSKI Pierre	Professeur, Université de Strasbourg
M VALLAR Laurent	Directeur de recherches, Luxembourg Institute of Health

Remerciements

Avant tout je tiens à exprimer ma sincère reconnaissance à M. CHALMEL Frédéric, M. XENARIOS Ioannis, M. GANÇARSKI Pierre et M. VALLAR Laurent pour l'honneur qu'ils me font de juger cette thèse.

Merci à mes directeurs de thèse Odile et Olivier d'avoir cru en GeneBook (renommé en MyGeneFriends par la suite, tout en échappant de peu à SoBIM et GenePathy) et de m'avoir permis de réaliser ce projet, tout en m'aiguillant dans la bonne direction. Merci à Raymond de m'avoir initié à la ligne de commande, merci à Luc de m'avoir ouvert les yeux sur la robustesse et la rapidité du Tcl, merci à Julie d'avoir toujours été là pour la correction de mon anglais approximatif. Merci à Laetitia et Wolfgang pour votre aide à comprendre les puces à ADN. Merci à Anne d'avoir été là pour me protéger et me soutenir face aux papiers administratifs.

Merci à Benjamin de m'avoir fait découvrir ce laboratoire.

Merci à Alexia, Yannis, Kirsley pour leurs contributions à MyGeneFriends. Merci à la ligue des jeunes du laboratoire, Salma, Carlos, Renaud, Arnaud, Hélène, Vincent pour les nombreuses sorties qu'on avait fait ensemble.

Merci à Marie et Anissa pour leur présence et leur soutien et enfin un très grand merci à mes parents de m'avoir soutenu et d'avoir été là pour moi.

Sommaire

Liste des figures

Liste des tables

Abréviations

(5'/3')-UTR	Région non-traduite (en 5'/3')
ADN	Acide désoxyribonucléique
API	Application Programming Interface
BLAST	Basic Local Alignment Search Tool
CDS	Séquence codante
ChIP	Immunoprécipitation de la chromatine
CST	Compressed Suffix Trees
DAVID	Database for Annotation, Visualisation and Integrated Discovery
DBD	Domaine de liaison a l'ADN
FTP	File Transfert Protocol
GO	Gene Ontology
GWAS	Genome Wide Association Studies
HPO	Human Phenotype Ontology
HTTP	Hypertext Transfer Protocol
IC	Information Content
IC	Information Content
IDF	Inverse Document Frequency
IUPAC	International union of pure and applied chemistry
JSON	JavaScript Object Notation
NCBI	National Center for Biotechnology Information
OMIM	Online Mendelian Inheritance in Man
RAR	Récepteur à l'Acide Rétinoïque

RARE	Elément de réponse à RAR
RefSeq	Reference Sequence database
SNP	Single Nucleotide Polymorphism
SQL	Structured Query Language
STRING	Search Tool for the Retrieval of Interacting Genes/Proteins
TF	Term Frequency
TSS	Transcription Start Site
UCSC	Université de Santa-Cruz Californie
UniProt	Universal Protein resource
URL	Uniform Resource Locator

Avant-propos

L'introduction de cette thèse présentera les principaux défis et opportunités liés aux mégadonnées en général et notamment, dans le contexte des réseaux sociaux et dans le contexte de la biologie. Nous verrons comment ces deux domaines, apparemment très distincts, gagnent à fusionner. Des aspects plus spécifiques de problématiques abordées durant la thèse, telles l'orthologie ou la régulation de la transcription seront introduits dans la partie Résultats et Discussion.

Je parlerai ensuite dans la partie Matériel et Méthodes des multiples sources de données et des outils qui m'ont permis de mener à bien mon projet de thèse, de leur hétérogénéité et des traitements appliqués aux données.

La partie Résultats et Discussions présentera les trois phases représentatives de mes travaux de thèse. Je présenterai d'abord ma découverte des mégadonnées et de la nécessité de leur traitement en temps réel par l'expert humain avec de le développement de Parsec (outil de recherche et de contextualisation de sites sur le génome). Nous verrons ensuite ma contribution à l'amélioration d'un programme de prédiction de relations d'orthologie, OrthoInspector, développé au laboratoire qui m'a permis de me sensibiliser aux problèmes des visualisations efficaces et interactives. Enfin, je présenterai MyGeneFriends (réseau social pour gènes, maladies et chercheurs), projet phare de ma thèse, pour lequel j'ai pu profiter de l'expérience des mégadonnées acquis dans les projets précédents pour aller plus loin dans ma volonté d'une relation symbiotique entre données et experts.

Dans la dernière partie du manuscrit, je finirai avec un regard synthétique sur ces trois années et l'évolution de mon projet de thèse, avant de présenter les perspectives et l'avenir de MyGeneFriends.

Introduction

Je vais présenter dans cette introduction les aspects transversaux concernant le contexte ainsi que le cheminement et les objectifs de ma thèse, à savoir d'une part, optimiser les interactions entre chercheurs et mégadonnées biologiques et d'autre part, intégrer l'expert humain dans le réseau d'informations biologiques par la création d'un réseau bio-social.

Les mégadonnées seront au centre de cette thèse. Les mégadonnées ne sont pas seulement de larges quantités de données. Ce sont des données dont l'ensemble des propriétés intrinsèques font qu'elles ne peuvent plus être assimilées avec les technologies et approches habituelles.

Nous verrons comment les mégadonnées, après avoir conquis le monde des sciences exactes, ont fait une entrée remarquée dans les réseaux sociaux (le « *Big Social Data* »), pour enfin se faire un chemin en biologie. Nous verrons comment la biologie a réagi à ce nouveau défi, en profitant des avancées dans d'autres domaines et en apportant ses propres solutions. Nous verrons enfin les tendances actuelles pour faire face aux enjeux des mégadonnées en biologie et comment ma thèse s'inscrit dans ce contexte.

1 Les mégadonnées : un enjeu global

"Information will be 'the 21th Century oil"

Gartner, 2010

Lorsque j'avais sept ans, ce qui n'est pourtant pas une époque si lointaine, un lecteur de disquettes était une énorme boite dépassant vingt centimètres de haut, et je pouvais sauvegarder mes programmes Basic sur des cassettes audio qui faisaient un bruit strident à la lecture et possédaient l'incroyable quantité de stockage de 660KB par côté.

Aujourd'hui, le paysage a bien changé. La loi de Moore s'est vérifiée et l'on a assisté à une explosion de la puissance des processeurs, de la capacité de la mémoire, et à une réduction du coût du matériel. La miniaturisation a fait rentrer un ordinateur qui occupait des salles immenses dans les quelques centimètres de l'Apple Watch. Des logiciels, tels Siri, Google Now ouCortana sont à présent capables de servir d'assistant et de répondre à des requêtes hétéroclites en tenant compte du contexte de la conversation. L'informatique est désormais partout, dans les voitures, les montres, les téléphones ou les puces des ascenseurs.

La quantité de données générées en deux jours en 2011 était de 1.8ZB (pour cet acronyme sur la taille des données et les suivants, voir Annexe 3), plus que la quantité totale d'information générée par la civilisation humaine depuis ses origines jusqu'au début du vingt et unième siècle (Chen et al., 2014).

Ce changement est dû à un nouveau type de données appelé mégadonnées, car il se différencie de données simplement grandes, du fait qu'il ne peut être acquis, géré, analysé par des outils logiciels ou matériels habituels dans un temps raisonnable (Chen et al., 2014). Son Volume, sa Vélocité, sa Variété et sa Véracité (les 4V classiques décrivant les mégadonnées) impliquent l'émergence d'outils et d'approches radicalement nouveaux, la simple augmentation de la puissance du matériel n'étant plus suffisante pour répondre aux problématiques d'acquisition, stockage, traitement, visualisation de ces données.

La notion de *Cloud* et le modèle *MapReduce* font partie des nouvelles approches les plus connues pour affronter cette nouvelle situation. Le *Cloud* permet un accès souple à de la mémoire, de la puissance de calcul ou des logiciels (Saas, *Software as a service*) permettant d'exécuter des calculs importants sur des serveurs distants. *MapReduce* quant à lui est un modèle de programmation proposé par Google, particulièrement adapté pour des calculs parallèles et distribués de grandes quantités de données (les données sont transformées en un dictionnaire clef-valeur qui peut être découpé en plusieurs parties, chacune traitée en parallèle pour produire des résultats intermédiaires qui seront réunis en un résultat final). La généralisation de ces approches

pour un grand nombre de problématiques peut être illustrée par les nouveautés majeures de la nouvelle version (Java 8) du langage de programmation généraliste et très populaire, Java, qui comprennent l'introduction du flux « java.util.stream » facilement parallélisable, et l'implémentation de la fonctionnalité MapReduce. Le « *machine learning* » enfin, permet aux ordinateurs d'apprendre sur la base d'informations existantes (Hotho et al., 2005) et d'extraire des connaissances ou des motifs sur la base de textes ou de comportements humains.

La généralisation de techniques adaptées aux mégadonnées s'explique également par la nature globale des enjeux liés aux mégadonnées qui ont le potentiel de favoriser la croissance de l'économie mondiale en révolutionnant des domaines entiers (Chen and Zhang, 2014) :

- le domaine de la santé publique (traitement personnalisé, réduction des coûts grâce à des traitements optimisés, …),
- l'administration du secteur publique (réduction des coûts des activités administratives),
- les grandes chaines commerciales (analyse des transactions, prédictions, optimisation des stocks, etc…),
- la production de biens à échelle globale.

Nous allons voir comment les mégadonnées, cantonnées pendant un temps à des domaines spécialisés sont entrées dans notre quotidien avec la révolution du web 2.0 et des réseaux sociaux.

1.1 Origine des Mégadonnées

Les technologies à haut débit produisant de grandes quantités de données ont été longtemps réservées aux sciences exactes comme la physique ou l'astronomie. La physique des hautes énergies est un exemple emblématique. Les 600 millions de collisions de particules par seconde du Grand collisionneur de hadrons du CERN (Pop, 2014), le plus grand et puissant accélérateur de particules au monde, génèrent plus de 22 PB chaque année.

L'astronomie quant à elle enregistre les positions et intensités de milliards d'étoiles, de galaxies, de quasars, les spectres de millions d'objets, auxquels s'ajoutent les images dans le spectre visible ou invisible avec des capteurs CCD de milliards de pixels (Feigelson and Babu, 2012). Le réseau de télescopes Pan-STARRS génère à lui seul plusieurs PB de données chaque jour (Hey, 2012).

Les simulations sociales, la météorologie, la finance (prédiction des cours de la bourse, achats et ventes d'actions automatisés) sont également de grandes productrices et consommatrices de mégadonnées. L'exemple de la météorologie est extrêmement intéressant, car face à un système très complexe (les interactions des 10^{44} molécules de l'atmosphère terrestre produisant le climat) les modèles actuels et la puissance des super ordinateurs ne suffisent pas à prédire correctement le climat. Cela est dû à sa grande sensibilité aux conditions initiales (quelques millièmes de degrés peuvent totalement inverser le résultat de la simulation), et à son caractère dynamique (dépendance de l'état précédant) et non linéaire (changements abrupts, exponentiels). Le centre national des prédictions environnementales américaines emploie un grand nombre d'experts

humains qui interprètent, complètent et modifient les prédictions du super-ordinateur, améliorant la précision des prédictions jusqu'à 25% (Silver, 2012). De plus, les météorologistes ont compris l'importance de communiquer sur l'incertitude de leurs prédictions, qui est désormais devenue une composante fondamentale des prédictions.

Les entreprises de nombreux domaines (vente, transports, loisirs) sont de plus en plus nombreuses à récolter massivement des données dans le but d'optimiser leurs performances, comme Wal-Mart (géant de la grande distribution) qui a investi dans 4 PB de stockage pour emmagasiner et analyser toutes ses transactions (Chen and Zhang, 2014).

Cette nouvelle ruée vers l'or entraine également quelques dérives. Les avantages et gains que font miroiter les mégadonnées ont convaincu de nombreuses entreprises à faire du « BigData » sans même réellement savoir ce que c'est, ni comment cela peut leur servir.

Internet est passé en première ligne de cette révolution avec plus de 700 EB de trafic annuel (Nabi, 2013). Le volume croissant de ce trafic est de plus en plus lié aux données créées et partagées par les utilisateurs des services web : photos, vidéos, messages, commentaires, SMS. Ces données transitent par un système de réseaux sociaux qui s'élargissent constamment. Les géants du web accumulent les records : 1 milliard de photos ajoutées chaque jour sur Facebook, 400 millions de tweets journaliers (Nabi, 2013), chaque minute plus de 72 heures de vidéos téléchargés sur YouTube, et les exemples ne manquent pas (Figure 1).

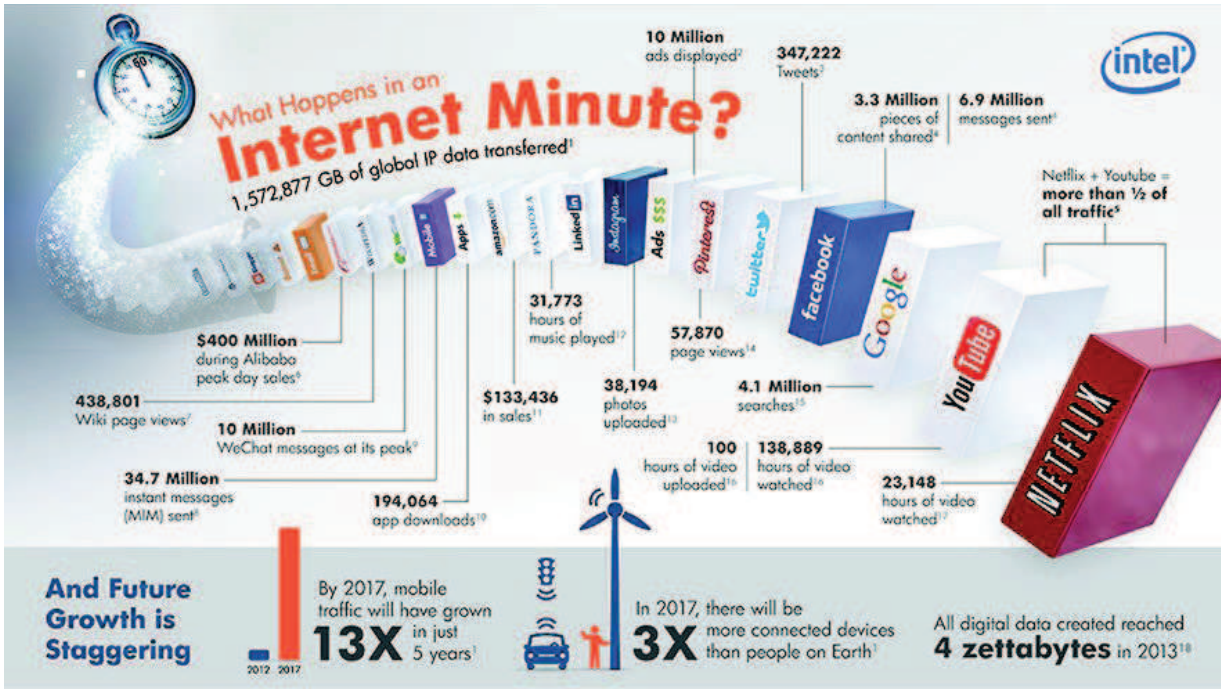


Figure 1 : Les flux de données sur Internet pendant une minute, illustration créée par Intel (http://www.intel.com/content/www/us/en/communications/internet-minute-infographic.html)

L'internet des objets (IoT, *Internet of Things,* qu'on considère également comme le Web 3.0) est la prochaine grande révolution qui pointe le bout de son nez. Visible aujourd'hui dans les ampoules connectées, les bracelets de sport ou les box de domotique proposées par plusieurs grands fournisseurs d'accès à internet, son objectif est de connecter à internet tous les objets de la vie quotidienne (machine à café, microonde, télévision, voiture, brosse à dents, …) pour les rendre plus « intelligents » et donc plus utiles. Ce réseau d'objets physiques crée de nombreux flux de données supplémentaires et repose sur l'identification des objets, le recueil d'informations par divers objets, la communication des informations entre objets, le traitement des données permettant la prise de décision ou le déclenchement d'actions.

Les données accumulées n'ont que peu d'intérêt en tant que telles, c'est leur interconnexion, leur analyse et la génération de connaissances et de nouveaux services à partir de ces données qui changent réellement notre monde, et les réseaux sociaux démontrent de plus en plus leur capacité à faire fructifier les mégadonnées.

1.2 Le Web 2.0 et les réseaux sociaux au cœur des mégadonnées

1.2.1 Le Web 2.0

Les premiers sites web se caractérisaient par une relation unidirectionnelle entre le créateur de contenu (le créateur d'un site web, un blog par exemple, qui se chargeait d'y conter sa vie passionnante) et les utilisateurs de ce site qui en étaient purement consommateurs et ne pouvaient donner aucun retour sur son contenu.

L'avènement du Web 2.0 transforma les utilisateurs d'un site web jusqu'ici entièrement passifs en utilisateurs actifs (Figure 2).

Figure 2: En 2006, « You » a été élue la « personne de l'année » par Time Magazine afin de reconnaitre la contribution anonyme de millions de personnes au contenu de nombreux sites web participatifs (Wikipedia, YouTube, Facebook, etc…)

Les utilisateurs peuvent désormais interagir avec le site web, donner leur retour et apporter du contenu. Ce nouveau paradigme s'étend d'une interactivité minimaliste (commenter une publication sur un blog) jusqu'à la conception des utilisateurs comme principaux générateurs de contenu. Cette forte interactivité est à la base des sites de réseaux sociaux ou de sites comme Wikipédia ou YouTube (permettant le partage massif de vidéos).

1.2.2 Le réseau social

Un réseau social peut être défini comme un service web permettant aux utilisateurs de construire un profil public ou semi-public, de se connecter à d'autres utilisateurs, et de traverser le réseau de connections ainsi créé (Boyd and Ellison, 2007). Ces fonctionnalités existaient déjà dans le tout premier réseau social créé en 1997 : SixDegrees.com. Les réseaux sociaux actuels (quelques exemples représentatifs sont présentés dans le Tableau 1) y ajoutent également de nombreuses fonctionnalités supplémentaires que nous verrons plus en détail plus tard (page de profil, flux de *news*, *micro-blogging*, suggestions…).

Tableau 1: Quelques réseaux sociaux majeurs et leurs descriptions. Données provenant de (Boyd and Ellison, 2007) et de Wikipédia. En orange apparaissent les réseaux sociaux généralistes, en vert celui orienté vers la photographie, et en bleu ceux orientés vers les chercheurs ou professionnels.

Nom	Description	Date de création	Nombre d'utilisateurs actifs
Facebook	Réseau social le plus populaire et orienté vers une large communauté d'utilisateurs	**2004**	1,44 milliards
Google+	Le réseau social de Google qui devait supplanter Facebook	**2013**	540 millions
Twitter	Réseau social fonctionnant autour de courts messages et de *hashtags* (marqueurs de métadonnées) dans ces messages	**2006**	300 millions
Instagram	Réseau social orienté autour de partage de photos	**2010**	Plus de 300 millions
ResearchGate	Réseau social orienté chercheurs	**2008**	7 millions
LinkedIn	Réseau social professionnel	**2003**	364 millions

1.2.3 Les réseaux sociaux introduisent de nouvelles notions…

Les réseaux sociaux ont introduit ou popularisé des notions révolutionnant les interactions humaines (Nadkarni and Hofmann, 2012). Certaines notions comme l'échange (de photos, de vidéos ou d'évènements) sont secondaires, d'autres sont cruciales et nous allons les décrire plus en détail.

1.2.3.1 Le profil

Les profils des utilisateurs sont une partie importante de tout réseau social et peuvent contenir de grandes quantités d'informations soumises par l'utilisateur. Plus qu'une simple vitrine, ils jouent un rôle important dans l'établissement de nouvelles connections entre utilisateurs et l'enrichissement du réseau. Des études montrent que la quantité et le type d'informations présents sur un profil utilisateur impactera directement le nombre de connections (Lampe et al., 2007).

La théorie des signaux considère que les différents éléments présents sur un profil peuvent agir comme des signaux dévoilant des points précis sur l'identité de l'utilisateur. Ces signaux peuvent être manipulés par l'émetteur et interprétés par le receveur pour émettre des jugements. Le système de signaux évolue de sorte qu'il devient bénéfique pour les utilisateurs de produire des signaux vrais et qu'il est couteux de produire de faux signaux. Une étude tend à montrer que la structure en réseau favorise des profils plus véridiques grâce à des vérifications explicites et implicites des informations fournies (Donath and Boyd, 2004).

1.2.3.2 Le Post ou micro-blogging

Le *micro-blogging* est un système permettant de diffuser de courts textes à un public donné, popularisé par Twitter dont c'est la fonctionnalité principale, et repris par d'autres réseaux sociaux comme Facebook ou Google+. Cela constitue des *news* permettant aux utilisateurs de se tenir informé de l'activité ou des pensées des uns et des autres. Sur Facebook par exemple, on peut poster ces courts textes sur son propre mur ou le mur de quelqu'un d'autre. En complément des humains, le système peut également poster des *news* sur l'activité de l'utilisateur (nouvelle amitié, modification du profil, etc…).

1.2.3.3 Les amis

Les réseaux sociaux permettent à leurs membres d'établir des connections explicites avec d'autres utilisateurs (nœuds) du réseau, Facebook désigne ces connections par le terme d'amitiés. Tous les nœuds connectés à un nœud du réseau sont donc considérés comme amis de ce nœud. La création de nouvelles connexions passe par l'envoi de demandes d'amitié d'un nœud à un autre. Si le nœud cible accepte la demande, un lien d'amitié bidirectionnel est créé et chaque nœud apparait alors dans la liste d'amis de l'autre nœud. Il existe également des connexions unidirectionnelles introduisant les notions de « fan » ou « *follower* ».

L'amitié une composante cruciale de Facebook car elle permet aux visiteurs de traverser un réseau de connexions entre utilisateurs, en naviguant de profil à profil. De plus, comme précisé

précédemment, l'évaluation de la véracité des informations présentes sur un profil est une part importante dans l'utilisation des réseaux sociaux, et le réseau d'amitiés mutuelles joue un rôle majeur dans la confirmation ou non des informations présentes.

Ce réseau d'amitiés appelé aussi « Graphe Social » fait l'objet de nombreuses études sociologiques ou médicales basées sur sa structure (nous verrons dans le chapitre suivant quelques exemples).

1.2.3.4 Commentaires

Les commentaires sont de courts textes écrits par les utilisateurs en réaction à un *post*, et peuvent s'enchainer pour constituer une discussion sur le sujet décrit par le *post* de base. En plus des *likes*, le nombre de commentaires est une mesure de la popularité d'un contenu. Les commentaires, aussi bien leur contenu que leurs connexions, font partie de nombreuses études, que ce soit l'analyse de réseaux de co-participation (on crée un lien entre deux utilisateurs s'ils ont commenté le même *post*) permettant de prédire la popularité future du contenu d'un *post* quelques heures à peine après sa publication (Jamali and Rangwala, 2009) ou l'analyse de la corrélation entre nombre de vues, de commentaires et de fans d'un *post* (Cha et al., 2009).

1.2.3.5 Le Like

Le *like* est parmi les fonctionnalités les plus connues des réseaux sociaux et d'autres sites Web2.0 comme YouTube. Il permet en un clic de donner un retour positif sur un contenu. Certains sites comme Facebook n'acceptent que le retour positif (le bouton *dislike* n'existe plus sur Facebook) tandis que d'autres sites comme YouTube permettent de donner des avis positifs ou négatifs.

1.2.3.6 La messagerie

Le *chat* ou messagerie instantanée existait bien avant les réseaux sociaux avec quelques clients phares comme MSN par exemple. Mais avec la croissance exceptionnelle d'utilisateurs de réseaux sociaux, les flux d'informations en temps réel autorisés par la messagerie furent pratiqués à une échelle sans précédent. Il devint inutile d'installer une application de messagerie supplémentaire alors que pratiquement tous les interlocuteurs potentiels sont déjà sur le réseau social. La messagerie a également évolué en apportant des fonctionnalités contextuelles, comme la reconnaissance et mise en forme des liens par exemple. Un lien vers une image se transforme en image, un lien vers une vidéo en vidéo, etc…

1.2.4 **… et révolutionnent le monde…**

Avec l'usage grandissant et la multiplication de réseaux sociaux, internet cesse d'être seulement un réseau de documents et devient de plus en plus un réseau de personnes (Ugander et al., 2011). Les réseaux sociaux comme Facebook possèdent des propriétés de structure intéressantes parmi lesquelles figurent :

- Le phénomène du « **petit monde** » qui désigne les réseaux dans lesquels un nœud est relié à n'importe quel autre nœud par une courte chaine de nœuds. Sur Facebook, la distance moyenne entre deux personnes étant de 4.7 nœuds (Figure 3), les contenus partagés peuvent facilement toucher une fraction importante de la population mondiale (Ugander et al., 2011).

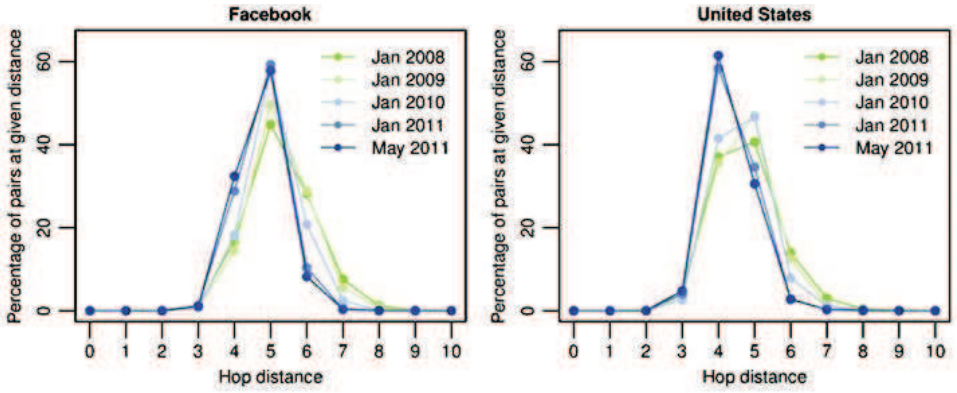


Figure 3 : Degré de séparation (hop distance, ou nombre de nœuds intermédiaires) entre toutes les paires d'utilisateurs sur Facebook (https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859)

- Le principe de l'**Homophilie** repose sur le fait que la similarité fait naitre des connections (McPherson et al., 2001). En d'autres termes, cela signifie qu'un contact entre personnes similaires est bien plus fréquent qu'entre personnes dissemblables. Ainsi les réseaux créés par les personnes (amitiés, réseaux professionnels, etc…) ont tendance à être homogènes du point de vue de divers facteurs sociodémographiques (âge, religion, éducation, métier, etc… Figure 4), ce qui limite la diversité des informations échangées.

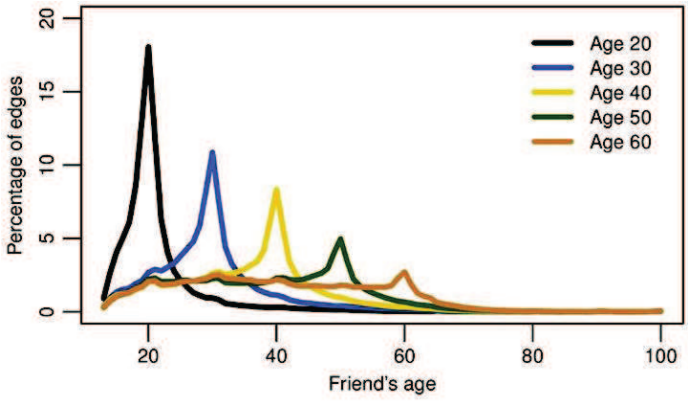


Figure 4: Les utilisateurs de Facebook ont plus de connections avec des utilisateurs de leur âge (https://www.facebook.com/notes/facebook-data-team/anatomy-of-facebook/10150388519243859)

- L'effet de ***Clustering*** signifie qu'en observant les amis d'un utilisateur, on pourra se rendre compte qu'une partie non négligeable de ces amis sont déjà amis entre eux. Cet effet est diminué pour les utilisateurs ayant des milliers d'amis et donc un réseau d'amis moins cohérent. Le coefficient de *clustering* permet de comparer les caractéristiques sociales des différents réseaux, par exemple Facebook possède un coefficient cinq fois plus élevé que ne possédait le réseau de « MSN messenger » (ancien réseau de messagerie instantanée développé par Microsoft).

Les média sociaux et les réseaux sociaux en particulier ont également transformé la nature des échanges en plus de les enrichir et les rendre plus rapides (Hawn, 2009). L'ancien modèle de communication *un à un* caractéristique du télégraphe ou du téléphone a laissé place aux modèles *un à plusieurs* qu'on retrouve dans un *tweet* ou un *post* sur un *blog* et *plusieurs à plusieurs* comme le mur de Facebook.

Les propriétés, la taille, et les connections de ces réseaux révolutionnent le monde de l'échange et de l'exploitation des communications entre humains, de l'échelle de l'individu (profilage individuel) à l'échelle de toute une population, autorisant des études massives.

1.2.4.1 Profilage individuel

1.2.4.1.1 Publicité ciblée

La fonctionnalité la moins appréciée des utilisateurs des réseaux sociaux mais celle qui permet leur gratuité est la publicité ciblée. La publicité ciblée la plus courante est associée aux grands moteurs de recherche qui vont utiliser les mots clefs saisis par un utilisateur lors de la recherche d'un contenu afin de le cataloguer (profiler) et lui proposer de la publicité pour des produits qui seraient les plus à même de l'intéresser. C'est donc une publicité personnalisée qui va essayer de deviner et profiter des centres d'intérêts de la personne. Les réseaux sociaux permettent d'aller encore plus loin grâce aux multiples informations personnelles que les utilisateurs fournissent d'eux-mêmes et au réseau de connexions dont ils font partie qui permettent d'améliorer nettement la qualité des suggestions. Un brevet a par exemple été déposé par Yahoo afin d'utiliser le contenu du *post* d'un utilisateur et du contexte associé (spatial, temporel, etc…) afin de proposer une publicité ciblée (King et al., 2010).

Le réseau peut également être utilisé pour évaluer la susceptibilité des différents types d'individus à influencer les autres, par exemple de savoir si la joie, l'obésité ou le fait de fumer peut être contagieux. Ces réseaux d'influence au sein du réseau social permettent d'optimiser la diffusion de produits ou d'évaluer la sensibilité des différents groupes d'individus à la publicité ciblée (Aral and Walker, 2012).

1.2.4.1.2 Les suggestions

Le principe des suggestions est de se baser sur des informations que l'on possède sur l'utilisateur afin de trouver des informations qui s'en rapprochent le plus. Il peut s'agir de suggestions de films comme sur Netflix, de produits sur Amazon ou de musique sur iTunes. L'importance économique des suggestions de qualité est colossale ; ainsi Netflix créa le « Netflix Prize » (http://www.netflixprize.com/), un concours visant à récompenser à hauteur de 1 million de dollars l'algorithme permettant d'améliorer de manière substantielle la précision à laquelle on pourra évaluer à quel point une personne appréciera un film donné.

Les réseaux sociaux apportent à ces systèmes de recommandation classiques leur réseau, en permettant des « *social recommandations* », basées non seulement sur les goûts de la personne ciblée mais également sur les goûts de ses amis (Ma et al., 2011).

1.2.4.1.3 Score d'affinité

Le score d'affinité est utilisé par certains réseaux sociaux pour mesurer l'affinité ou la compatibilité entre deux personnes. OkCupid par exemple, fondé par deux mathématiciens, se base sur les réponses à un quizz pour calculer la compatibilité entre deux personnes. Il est intéressant de noter que ce système ne calcule pas la similarité des profils. Ainsi, à chaque question à choix multiples, l'utilisateur peut choisir sa propre réponse et les réponses acceptées par l'autre personne, en y associant des poids. Cela permet au système de calculer à quel point l'utilisateur A est compatible avec les attentes de l'utilisateur B, et l'utilisateur B compatible avec les attentes de l'utilisateur A (en.wikipedia.org/wiki/OkCupid).

La capacité d'associer des personnes compatibles est également un enjeu algorithmique pour les modes multi-joueurs des jeux vidéo. Cette mise en relation ou *matchmaking*, repose souvent sur le niveau des joueurs. Cette information peut être complétée par la mesure des différentes compétences des joueurs et la déduction du rôle qui leur conviendrait le mieux en appliquant des techniques de classification et de *clustering* (Jiménez-Dıaz and Menéndez, 2011). L'importance économique d'un bon système de *matchmaking* est montrée par l'existence de nombreux brevets : *Player character matchmaking with distributed peer-to-peer functionality*, *Dynamic battle session matchmaking in a multiplayer game* ou *Simplified matchmaking* …

Dans un tout autre domaine, la formation de groupes dans les MOOCs (*Massive Open Online Courses*) peut faire appel à des outils de *peer-matching* comme APMatch (Verburg et al., 2014) pour créer des groupes d'étudiants et faire des recommandations de groupes de travaux pratiques, en se basant sur les compétences définies par les étudiants et le professeur.

1.2.4.1.4 Le flux de *news* (*News feed*)

Tous les grands réseaux sociaux disposent d'une page qui va afficher le flux de *news*, à savoir les textes écrits par l'utilisateur et ses amis ainsi que ceux générés par le système lui-même. Ce système offre un flux de *news* personnalisé et adapté aux intérêts de l'utilisateur. La diffusion de l'information à une large proportion d'utilisateurs est assurée par la possibilité de partager un lien « posté » par un autre utilisateur (le recopier sur son propre mur), et des études montrent la

contagion du comportement visant à partager (Bakshy et al., 2012) favorisant la propagation de l'information sur le réseau.

1.2.4.2 Etudes massives

1.2.4.2.1 Etudes scientifiques

Les réseaux de connections entre utilisateurs ou *Social Graphs* sont une ressource de grande importance autant pour les réseaux sociaux eux-mêmes que pour les scientifiques qui peuvent les explorer dans de nombreux projets centrés sur les mégadonnées. Les questions que ces projets scientifiques posent ne sont pas nouvelles, ce qui est révolutionnaire c'est l'échelle à laquelle ces questions peuvent désormais se poser (Burgess and Bruns, 2012).

Parmi les études les plus intéressantes, on peut citer l'utilisation d'êtres humains comme détecteurs vivants d'évènements sociaux (buts marqués dans un jeu de football, élections présidentielles, …) ou physiques (tremblements de terre, incendies de forêts, inondations,…) (Zhao et al., 2011).

1.2.4.2.2 Suivi en temps réel de la Grippe

Les pandémies sont un autre enjeu de taille auxquels s'attaquent les géants des mégadonnées. Le virus Influenza, responsable de la grippe se transmet très facilement et des épidémies de grippe fleurissent annuellement un peu partout dans le monde. Le service Google *Flux Trends* permet d'estimer l'importance de l'épidémie dans différents états des Etats Unis avec, selon Google, une réactivité de deux semaines supérieure aux techniques traditionnelles. Cette estimation est basée sur la corrélation entre le nombre de personnes effectuant des recherches sur des sujets relatifs à l'influenza et ceux ayant des symptômes de cette grippe (Carneiro and Mylonakis, 2009).

Figure 5: Un site Google Trend est dédié au suivi en temps réel de la propagation de la grippe (https://www.google.org/flutrends/about/how.html). La figure ci-dessus montre la corrélation entre les recherches Google et les données officielles.

Une autre étude montre que l'utilisation des données proposées publiquement par Twitter *via* sa *Stream API* permet de suivre la grippe avec une précision encore meilleure grâce au contexte apporté par le texte contenu dans les *tweets* (Signorini et al., 2011). Ce contexte permet également d'étudier les effets secondaires des médicaments ou leur indisponibilité dans certaines régions.

1.2.4.2.3 Prédiction des élections et du cours de la bourse

Le potentiel prédictif de Twitter a également été évalué avec succès dans le cadre des élections en Allemagne (en analysant les mentions des partis politiques dans les tweets et les émotions associées) (Tumasjan et al., 2010) ou dans le contexte des cours de la bourse qui ont pu être prédits avec une précision de plus de 86% (Bollen et al., 2011). Il est intéressant de noter ici que l'étude ne se base pas sur les *posts* individuels très courts, mais sur l'agglomération de millions de *tweets* qui permettent de prédire l'humeur et le climat émotionnel de la population, selon six états : Calme, en Alerte, Sûr, Vital, Gentil et Heureux.

1.2.5 … et la technologie…

1.2.5.1 Generative Internet

Le terme *Generative Internet* signifie que l'utilisateur n'est plus simple fournisseur de données, mais également fournisseur de fonctionnalités.

Lorsque le premier iPhone est apparu en 2007, il était livré avec un ensemble de programmes préinstallés (Zittrain, 2009). Aujourd'hui il est difficile d'imaginer un smartphone sans que ses fonctionnalités puissent être étendues par des bouts de code créés par les utilisateurs eux-mêmes et disponibles à la communauté *via* des magasins d'applications : les *apps* ou applications. Loin d'être réservée au monde du *hardware*, cette modularité de nouvelle génération se retrouve sur de grandes plateformes web comme Facebook et révolutionne l'accès aux nouvelles fonctionnalités.

Jusqu'ici la notion de programmes modulaires signifiait que de nouvelles fonctionnalités pouvaient être créées par des développeurs tiers sous forme de modules. Si un utilisateur voulait avoir accès à cette nouvelle fonctionnalité, il devait installer chez lui le programme en question, télécharger le module et l'ajouter au programme. Les *apps* permettent aux développeurs de créer des programmes et de les soumettre à Facebook. Facebook profite ainsi de l'ajout de nouvelles fonctionnalités qui vont plaire à ses utilisateurs et les développeurs profitent de l'énorme base d'utilisateurs que leur offre Facebook. Ces *apps* tournent sur une Plateforme qui sert d'interface entre les applications et Facebook (Grimmelmann, 2008), permettant aux applications de récupérer des informations de Facebook et de lui envoyer des instructions pour créer de nouveaux

posts sur le mur des utilisateurs, partager des liens vers d'autres pages web, générer des notifications, etc…

Il existe ainsi à l'heure actuelle des milliers d'applications, incluant des jeux qui peuvent être joués avec d'autres utilisateurs du réseau (scrabble, zombies, FarmVille, etc..), des applications comme *Causes* permettant à l'utilisateur de trouver d'autres personnes faisant des dons aux mêmes causes caritatives, etc…

1.2.5.2 Programmation multilingue

La programmation multilingues (*polyglot programming*) (Fjeldberg, 2008), à savoir la réunion de plusieurs langages de programmation dans un même projet est un concept assez ancien qui a été utilisé dans plusieurs projets dont le toujours populaire, Emacs (code C avec un interpréteur lisp pour ajouter des scripts). La JNI (*Java Native Interface*, décrite plus bas) permet de faire « facilement » interagir du code Java avec des librairies C ou C++.

Mais c'est avec le besoin grandissant d'efficacité et de rapidité des nouveaux acteurs du web social que l'on assiste à une véritable explosion de l'usage de la programmation multilingue, afin de profiter des avantages de différents langages de programmation ainsi que de *frameworks* et librairies écrits dans ces langages lors de la création de logiciels. Ainsi Google+ est un mélange de Java et de JavaScript, Twitter un mélange de Java, Ruby, Scala, JavaScript, Facebook un mélange de C++, PHP et D, et pour corser le tout, rien que le système de messagerie instantanée de Facebook utilise un ensemble de composants codés en quatre langages différents (PHP, Javascript, C++, Erlang).

1.2.5.3 Les API et l'architecture orientée service

En 2002, Jeff Bezos, le patron d'Amazon, a ordonné que chacune de ses équipes expose toutes ses données et ses fonctionnalités aux autres équipes à travers des interfaces de programmation (API, *Application Programming Interface*), en menaçant de renvoyer tous ceux qui ne le feraient pas. Depuis, les API sont devenues incontournables chez les géants du web, qu'elles soient à usage interne ou externe. Plus de la moitié du trafic des sites comme Twitter, Google, Netflix ou eBay passe par des API (Jacobson et al., 2011).

Une API est un moyen pour les développeurs d'une application de donner accès aux données et fonctionnalités de cette application à d'autres développeurs, qu'ils soient internes ou externes. Lorsqu'un programme client demande l'accès à une ressource d'un serveur par le biais d'une URL (*Uniform Resource Locator*), l'API va lui renvoyer une réponse structurée facilement exploitable (généralement sous forme de JSON (*JavaScript Object Notation*)). Tout comme les clients de pages web classiques sont des humains, les clients des API sont d'autres programmes.

De nombreuses études scientifiques exploitent les fonctionnalités *Streaming API* et *Search API* de Twitter pour récolter facilement les données. Un exemple d'API devenue célèbre est la Graph API de Facebook permettant d'accéder à des objets du graphe social de Facebook (utilisateurs, pages, photos, évènements, etc…) et aux connections entre ces objets (amitiés, contenu partagé, photos annotées (tags), etc…) (Weaver and Tarjan, 2013). Cette API permet de répondre à des questions comme « est-ce que ces deux personnes sont amies ? », « Est-ce que cette personne aime cette page Facebook ? ».

1.2.6 …malgré des dangers et des dérives réelles

Pourtant, les exigences de sécurité et de confidentialité des données que l'on retrouve dans tous les domaines avec le développement de réseaux sociaux, l'interconnexion des informations et l'informatique « en nuage » (*cloud computing*) ne sont pas toujours satisfaites. Certains réseaux sociaux peuvent faillir à faire la distinction entre ce qui est de l'ordre du privé et ce qui est de l'ordre du public. Ils peuvent également présenter et disséminer des informations à véracité discutable, être sujets à des failles techniques ou devenir le relais de code malicieux.

Ainsi des controverses ont éclaté pour questionner la propriété des informations envoyées aux réseaux sociaux et leur confidentialité. Par exemple, Instagram avait en 2013 modifié son règlement afin de s'attribuer le droit de vendre les photographies postées par les utilisateurs sans les notifier ou les payer. Un bug dans Facebook découvert en 2013 permettait à n'importe qui de publier sur le mur de n'importe qui (fonctionnalité normalement réservée exclusivement aux amis de la personne). Egalement, une gestion confuse des paramètres du compte ne permettait pas aux utilisateurs de Facebook de spécifier clairement la visibilité des différentes informations.

Ces controverses ont nourri la création de réseaux sociaux décentralisés comme Diaspora, qui ont pour but de garder chaque utilisateur propriétaire de son information. Pourtant, comparé à des géants comme Facebook, Google+ ou Twitter, ils représentent une faible proportion d'utilisateurs, et le modèle centralisé risque de rester encore longtemps au-devant de la scène.

2 Les mégadonnées rencontrent la biologie

La biologie est un acteur relativement nouveau dans le monde des mégadonnées. Son entrée fracassante dans ce nouvel univers est essentiellement due aux nouvelles technologies à haut débit, une date charnière à citer serait le changement de millénaire qui vit le génome humain séquencé.

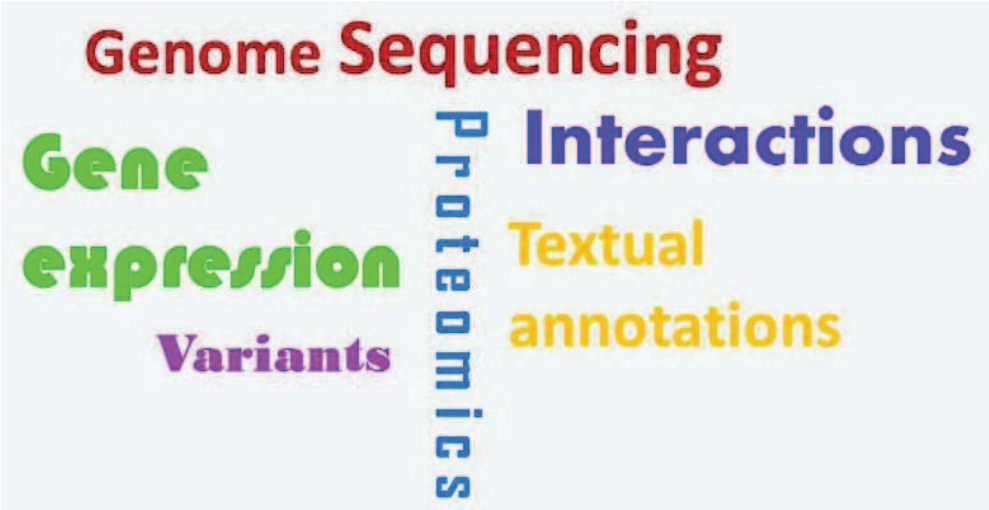


Figure 6: Un cliché des aspects massifs de la biologie qui ont été au cœur de ma thèse

Son rôle de nouveau venu permet à la biologie de profiter des avancées réalisées dans les autres domaines en termes de gestion et d'exploitation des mégadonnées. Ces avancées se retrouvent aussi bien au niveau matériel (*hardware*) que logiciel (*software*). Suivant la loi de Moore, le matériel informatique, progresse de manière exponentielle en quantité de mémoire, puissance de calcul et en nombre de processeurs, offrant ainsi des capacités d'analyse de plus en plus impressionnantes. Le développement logiciel quant à lui adopte de plus en plus la parallélisation des traitements, la logique *cloud*, l'architecture orientée service, la modularité. Enfin, des algorithmes développés dans un vaste éventail de domaines peuvent maintenant s'appliquer à la biologie pour des tâches comme la fouille de données (extraction de connaissances, de textes ou d'ontologies par exemples) ou l'apprentissage automatique (classifier, déduire et appliquer des règles, etc…).

De grands projets reposant sur ces infrastructures sont financés du fait de l'enjeu économique majeur que représentent les mégadonnées en médecine. Leur utilisation efficace pourrait en effet

rapporter plus de 300 milliards rien qu'à l'industrie médicale américaine (Chen and Zhang, 2014).

Comme on l'a vu, bien avant la biologie, les grands acteurs du web, en particulier les réseaux sociaux, se sont retrouvés confrontés aux grands volumes de données et à la nécessité d'en extraire une information personnalisée d'intérêt. Les publicités ciblées de Google, les suggestions d'amitié de Linkedin et Facebook, les suggestions de comptes à suivre de Twitter en sont quelques exemples emblématiques et vont inspirer la biologie face à des défis similaires. Les problématiques de cloisonnement et de gestion de la confidentialité des données qui touchent les réseaux sociaux sont plus cruciales encore dans le domaine biomédical, avec l'avènement de la médecine personnalisée et la multiplication des projets de recherche biomédicale centrés sur des cohortes de patients (Scheidecker et al., 2014). La gestion de la sécurité des données (Jee and Kim, 2013) et le cloisonnement entre informations publiques, privées ou partagées par un groupe de chercheurs plus ou moins étendu, deviennent en effet essentiels pour éviter des conséquences dramatiques.

Enfin, la biologie, de par la nature hautement complexe de ses objets d'étude, pose également de nouvelles questions, dont les réponses font avancer à la fois la biologie et les autres domaines.

Et tout cela n'a pourtant commencé qu'avec une poignée de séquences et quelques protéines…

2.1 Data POOR à data RICH

> *There is a tremendous amount of information regarding evolutionary history and biochemical function implicit in each sequence and the number of known sequences is growing explosively. We feel it is important to collect this significant information, correlate it into a unified whole and interpret it."*
>
> *Margaret Dayhoff, 1967*

En 1965, la première édition de l'Atlas de Séquences protéiques et de leurs structures de Margaret Dayhoff (Dayhoff, 1965) ne comportait que 65 séquences. La Protein DataBank ne contenait en 1972 que 10 structures de protéines déterminées par cristallographie aux rayons X, et en 1985 se tenait la première conférence pour évaluer la faisabilité d'une initiative de séquençage du génome humain.

Une étape importante pour la bio-informatique fut la création en 1980 d'une banque de données nationale de séquences d'ADN, financée par le NIH (*National Institute of Health*). Ce projet posa des principes qui sont à la base de la bio-informatique actuelle :

- la facilité d'accès et de partage des informations de cette banque par le réseau ARPANET (*Advanced Research Projects Agency Network*),

- la gratuité de l'accès aux informations,

- l'intégration de cette base de données bio-informatique dans *l'économie morale* des sciences du vivant. Ce système repose sur la récompense du chercheur par le fait d'être crédité pour son travail. Afin d'appliquer ce principe, les responsables du projet ont convaincu vingt journaux scientifiques majeurs de rendre la soumission de séquences à cette base de données obligatoire pour publier un article relatif à une ou des séquences.

Lorsqu'à la suite d'un effort sans précédent (13 ans et 3 milliards de dollars), le génome humain fut finalement séquencé à l'aube du 21ème siècle par un consortium international (Lander et al., 2001), la biologie avait franchi une étape clef pour passer du domaine des *data-poor* à celui des *data-rich*.

D'autres grandes initiatives ont pris le relais, grâce aux progrès de la biologie moléculaire et la construction d'appareils de plus en plus performants, permettant l'émergence d'une « biologie à haut débit ». Cette nouvelle biologie automatise des techniques classiques de biologie moléculaire permettant de répondre à des questions à des échelles jamais atteintes. Elle intègre pour cela les progrès de l'informatique, de la chimie, de l'optique et de l'analyse d'images et permet de s'attaquer à des répertoires entiers de molécules biologiques.

Ces techniques à haut débit ont, entre autres, révolutionné la génomique, l'étude des gènes et de leur régulation, permis la découverte de l'importance des variations entre individus et donné une base moléculaire aux maladies génétiques. Parmi les exemples les plus marquants, on pourrait citer les projets de séquençage de génomes de multiples espèces et souches, les études GWAS (*Genome Wide Association Studies*), le projet 1000 génomes (Genomes Project et al., 2010) (détection de variations entre individus de différentes populations humaines) et le séquençage massif d'exomes.

Le séquençage des génomes a connu un boom sans précédent avec un processus de plus en plus rapide et de moins en moins cher, permettant grâce à des technologies de séquençage à haut débit de séquencer un nombre croissant d'espèces (Figure 7).

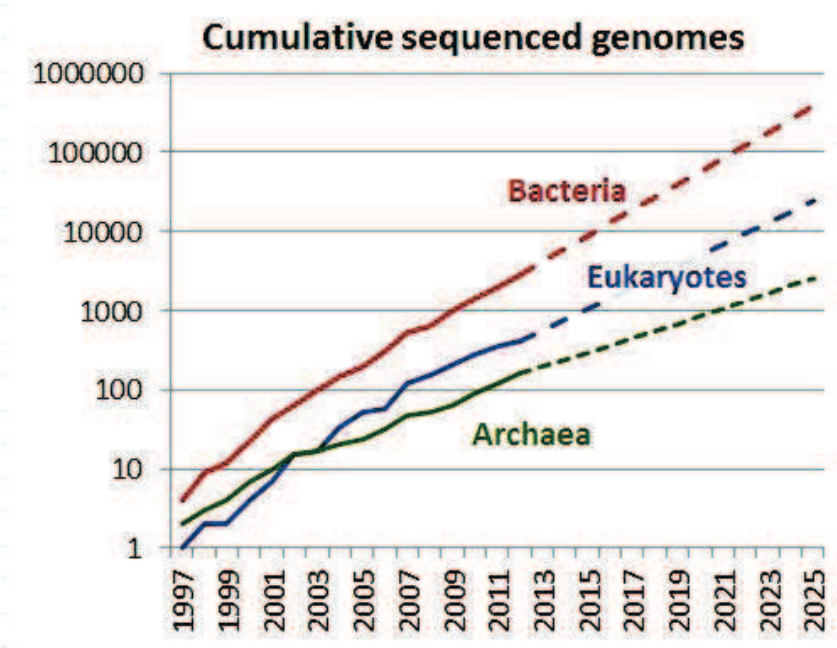


Figure 7 : Le nombre cumulatif de génomes séquencés selon les années augmente exponentiellement (banque de données GOLD, figure réalisée par Su, Andrew (2013))

Les approches GWAS (*Genome Wide Association Study*) étudient les relations entre variations et phénotypes, dans le but, par exemple, d'associer statistiquement des variations à la présence d'une maladie donnée dans un groupe de patients. Outre son intérêt pour l'examen médical, il permet de mieux comprendre la fonction de certains gènes en comprenant quel gène est lié à quel phénotype. Son premier grand succès en 2005 permit de relier des SNP (*Single Nucleotide Polymorphism*, mutation ponctuelle dans l'ADN) à la DMLA (Dégénérescence Maculaire Liée à l'âge) (Klein et al., 2005).

Le projet 1000 génomes (initié en 2008) avait pour but de séquencer le génome de 1000 individus représentatifs des populations humaines à travers le monde (Clarke et al., 2012). L'objectif final était de créer un catalogue des variations humaines et de déterminer leur fréquence afin de faciliter le diagnostic médical (des variations rares sont de meilleurs candidats pour expliquer les maladies que les variations communes, souvent neutres hors contexte particulier).

Le séquençage d'exomes enfin, fournit un outil puissant et peu couteux pour le diagnostic de maladies génétiques et l'évaluation de risques de maladies (Bamshad et al., 2011). Par ailleurs, les techniques de séquençage dites de RNA-seq qui remplacent les *microarray*, fournissent des données de plus en plus détaillées sur l'expression des gènes dans différents tissus ou à différents stades de développement (Su et al., 2004).

Ces grands projets s'inscrivent dans l'ère des 'OMICS', c'est-à-dire l'étude à grande échelle de gènes, de transcrits, de métabolites… Ces OMICS constituent une part prépondérante des mégadonnées dans la biologie d'aujourd'hui. Parmi le vaste éventail de domaines étudiés (qui

poussent comme des champignons, quelques exemples majeurs sont présentés dans le Tableau 2), ma thèse m'a surtout confronté aux données de la génomique, transcriptomique, protéomique, phénomique, variomique ou bibliomique, et ce, dans le contexte du « diseasome » (Goh and Choi, 2012) qui replace les maladies génétiques humaines dans une approche de réseau biologique (voire social ! (Barabasi, 2007)) global.

Tableau 2: Une présentation non exhaustive des OMICS. Sur fond bleu, ceux qui vont nous intéresser plus en détail.

Nom	Description
Genomics	Etude de la séquence de l'ADN
Transcriptomics	Etude des niveaux d'expression des mRNA au niveau de génome
Proteomics	Etude des protéines, leurs interactions, leur structure tridimensionnelle…
Metabolomics	Etude quantitative des petites molécules et de leur dynamique dans les organismes vivants
Epigenomics	Etude d'informations transmissibles, supplémentaires aux gènes et n'impliquant pas la modification de la séquence de l'ADN
Lipidomics	Etude des lipides (matières grasses)
Bibliomics	Etude des publications biologiques (bibliome)
Metagenomics	Etude d'un échantillon de l'environnement contenant de multiples génomes
Diseasome	Etude des interconnections entre les maladies et leurs liens avec les gènes
Variomics	Etude des variations dans l'ADN, l'ARN et les protéines
Phenomics	Etude de l'ensemble des phénotypes observés dans un organisme
Connectomics	Etude de l'ensemble des connections des neurones du cerveau humain

2.2 Complexité des mégadonnées biologiques

Les données générées par les technologies à haut débit possèdent les quatre caractéristiques majeures des mégadonnées définies par les 4V : Volume, Vélocité, Variété, Véracité. Ainsi, ces dernières années, la « biologie à haut débit » a fait croître le volume des données en biologie de manière vertigineuse. Pour l'homme par exemple, on dispose à présent de dizaines de milliers de gènes codants et non codants, de centaines de milliers de transcrits, de millions de relations d'interaction entre protéines, d'une multitude de séquences génomiques auxquelles s'ajoutent les informations d'expression, de régulation, de protéomique, etc... Ainsi, la quantité de données disponible publiquement sur le site ftp du NCBI dépasse 3 PB et le nombre de publications accessibles sur PubMed (Giglia and Spinelli, 2009) dépasse les 24 millions !

La vélocité est également au cœur de la biologie moderne avec des mises à jour constantes et désynchronisées des multiples bases de données : plusieurs fois par an pour Ensembl (Cunningham et al., 2015), tous les mois pour GO (Gene Ontology, 2015), plus souvent encore pour Orphanet (Weinreich et al., 2008) ou OMIM (Amberger et al., 2015) et tous les jours pour

Pubmed. Cela entraîne non seulement, le défi pour les chercheurs et cliniciens d'être constamment à jour mais également, de pouvoir suivre l'évolution de l'information et comprendre ce qui a été modifié, ajouté ou supprimé.

Les données sont également très variées, par leur nature (image, texte, ontologie, etc…), leur format (FASTA, fichier CSV, VCF, etc…) ou leur caractère plus ou moins structuré (de la base de données au texte plat).

Enfin, la véracité reflète tout à la fois le bruit inhérent à la biologie haut débit et l'état de nos connaissances partielles et partiales des entités biologiques (gènes, éléments régulateurs, protéines, ARN non-codants). Le biologiste est ainsi confronté quotidiennement aux données manquantes et à des éléments incorrects ou de qualité insuffisante. Par exemple, seuls 60 % des gènes humains seraient correctement prédits (Nagy and Patthy, 2014) et la quantité de rétractation de publications (pour cause d'erreur ou de fraude) grimpe en flèche (Steen, 2011) comme le montre la Figure 8. Tous ces facteurs peuvent impacter dramatiquement les études d'analyse et de comparaison (Prosdocimi et al., 2012).

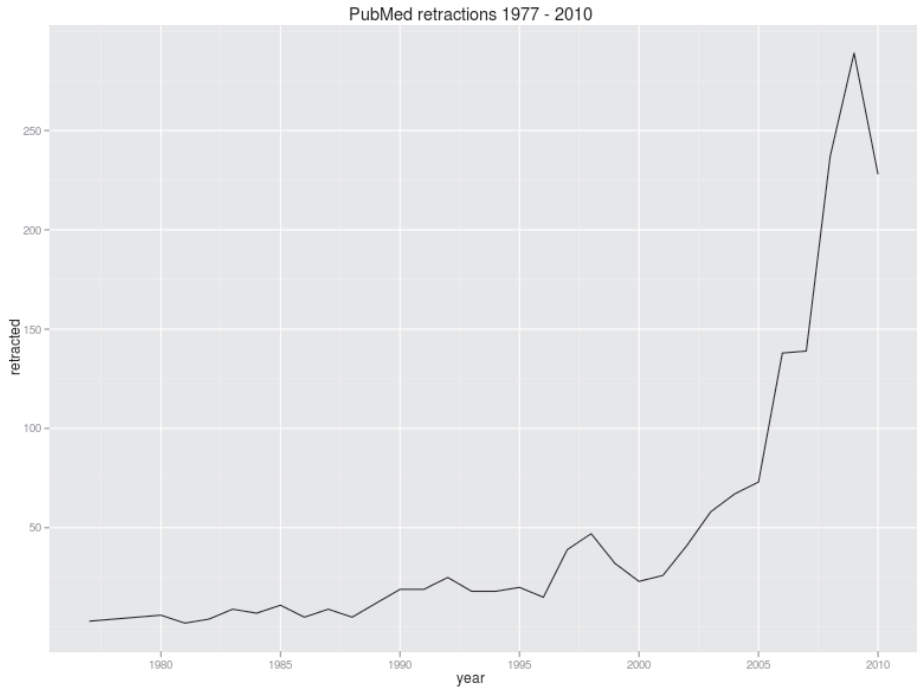


Figure 8 : Evolution du nombre de rétractations de publications (non cumulatif !) par an, calculé par Neil Saunders : https://nsaunders.wordpress.com/2010/11/30/analysis-of-retractions-in-pubmed/

Dans ce contexte, nous allons voir les défis que les mégadonnées biologiques posent à leur exploitation bio-informatique, de par leur complexité de séquence avec le génome, la richesse des types de transcrits et d'expression avec le transcriptome. Nous les confronterons aux problématiques liées à l'étude de maladies génétiques, de par la difficulté de leur définition et la variabilité des données les concernant. Nous verrons enfin comment les variations et les phénotypes aident à donner une base plus objective à ces maladies, tout en apportant de nouvelles difficultés.

Ce chapitre n'est pas exhaustif, et se concentrera sur les aspects les plus importants abordés au cours de ma thèse.

2.2.1 Génome

Chaque organisme vivant possède l'information biologique qui lui permet de se construire et de se maintenir en vie (Brown, 2002), c'est ce qu'on appelle le génome. Chez la plupart des organismes le génome est composé de molécules d'ADN (acide désoxyribonucléique), de longues chaines de nucléotides.

L'exploitation des données de séquençage de ces génomes est complexe, et se heurte à de nombreuses problématiques. Par exemple, le séquençage de nombreuses souches de bactéries ou de virus entraine un grand nombre de données redondantes, difficiles à gérer et analyser. Par ailleurs, l'utilisation des données de séquençage doit se faire avec prudence, car un certain nombre de génomes disponibles publiquement (le poulet par exemple) sont de très mauvaise qualité et incomplets.

Bien qu'étant un exploit majeur, la simple détermination de la séquence des chromosomes n'est qu'un point de départ. Un génome possède des gènes qui vont être transcrits en ARN et, chez bon nombre d'eucaryotes, un gène aboutira à de multiples transcrits se distinguant par un site d'initiation ou de terminaison de transcription diffèrent, ou un épissage alternatif.

A ce jour, la dernière version d'Ensembl du génome humain de référence (Ens80) liste plus de 64000 gènes, mais cette liste est incomplète car une partie du génome reste inaccessible (typiquement télomères et centromères). Ainsi plus de 900 gènes étaient estimés inaccessibles en 2013 (Topol, 2014).

Pourtant, une large partie de la séquence de l'ADN est constituée par ce que l'on a longtemps appelé l'ADN poubelle, contenant de larges régions répétées et dont le rôle n'est pas entièrement élucidé. Parmi ces éléments répétés, les transposons et rétrotransposons qui peuvent s'intégrer en de multiples sites du génome, sont massivement présents dans les grands génomes eucaryotes. Un seul rétrotransposon nommé *Alu* est par exemple présent des millions de fois dans le génome humain, occupant 10% de sa séquence (Eddy, 2012), et au moins 45% du génome humain est clairement dérivé d'éléments transposables. Des outils bio-informatiques comme RepeatMasker

(Tempel, 2012) vont alors avoir pour tâche de détecter ces éléments répétitifs pour les masquer (les remplacer par des N dans la séquence d'ADN).

En dehors des gènes, les sites de fixation des protéines sont des régions de l'ADN particulièrement importantes permettant la transcription des gènes et la régulation de cette transcription. Des méthodologies expérimentales ont été créées afin de les détecter. La technique de ChIP-seq est basée sur l'immunoprécipitation sélective du complexe ADN-protéine, grâce à des anticorps dirigés contre la protéine d'intérêt puis au séquençage de masse des fragments d'ADN reconnues par cette protéine d'intérêt (Park, 2009). ChIP-exo est une amélioration de ChIP-seq permettant une localisation plus précise du site de liaison, grâce à une étape lors de laquelle une exonucléase digère les fragments d'ADN qui ne sont pas liés à la protéine (Bailey et al., 2013). Cette technique particulièrement utile pour déterminer les sites de fixation de facteurs de transcription (TFBS pour *Transcription Factor Binding Site*) ou d'amplificateurs par exemple, est largement utilisée, mais présente certaines limites. Ainsi, la qualité des résultats obtenus dépend de l'obtention d'un anticorps dirigé contre la protéine d'intérêt et du degré de sensitivité et spécificité de celui-ci. Les erreurs de séquençage ou le dosage correct de l'échantillon peuvent également impacter la qualité des résultats (Park, 2009). L'énorme quantité de données produites (plusieurs tera-octets) doit être traitée statistiquement avant d'être exploitable et on relève de nombreux faux positifs et faux négatifs (Chen et all., 2012). Dans ce contexte, la localisation *in silico* de TFBS constitue souvent un complément indispensable en termes d'étude exploratoire, de comparaison et de validation d'hypothèses.

2.2.2 Transcrits

Les récentes découvertes en biologie couplées aux technologies de séquençage à haut débit performantes, ont montré le nombre important de gènes et de transcrits différents qui existent chez les eucaryotes. Les transcrits les plus connus et les plus étudiés sont ARN messagers (mRNA) codant pour des protéines. Mais l'importance des autres types de transcrits, souvent appelés transcrits non codants, devient de plus en plus manifeste, notamment au regard de la découverte de leurs rôles dans diverses maladies (Esteller, 2011). Les plus connus de ces ARN non codants sont les ARN ribosomiques (rRNA) et les ARN de transfert (tRNA). Les rRNA vont s'assembler à des protéines pour former le ribosome, structure essentielle permettant la traduction, et les tRNA jouent également un rôle crucial, en permettant d'associer un acide aminé donné à un triplet de bases, mécanisme à la base de la traduction.

D'autres transcrits, comme les miRNA, vont se fixer par complémentarité de bases sur des mRNA cibles et réprimer leur traduction. L'incapacité de produire des miRNA corrects (par défaut de régulation épigénétique par exemple dans le cas du miR-200 (Sun et al., 2015)), peut aboutir à l'apparition de cancers, mais aussi, à des désordres neurologiques.

Les lincRNA ont été découverts plus récemment. Ce sont de longs ARN non codants trouvés dans les régions intergéniques conservées et jouant un rôle dans la régulation d'expression de

certains gènes, agissant par exemple sur la réponse aux dommages subis par l'ADN (p53) (Luo et al., 2015).

On trouve également des transcrits issus de pseudogènes, c'est-à-dire de copie non fonctionnelle de gènes. Les pseudogènes simples (*non processed pseudogenes*) dérivent d'un gène par dégénérescence tandis que les rétropseudogènes (*processed pseudogenes*) proviennent de la retrotransposition d'un ARNm et sont généralement dépourvus d'introns.

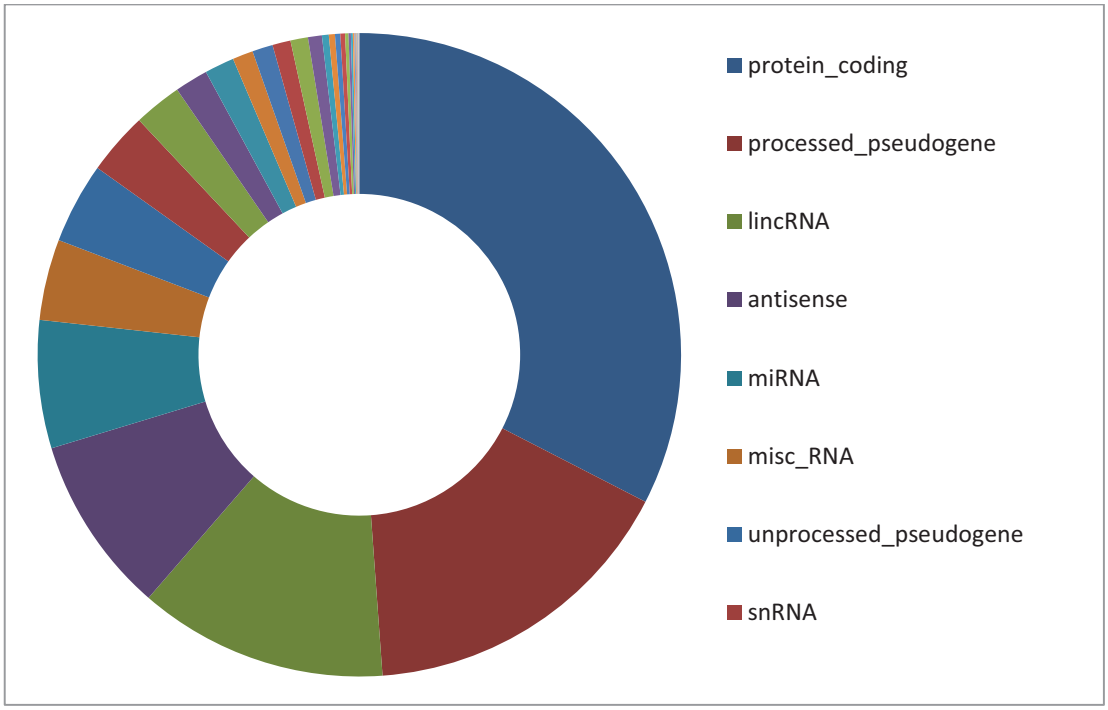


Figure 9: Principaux types de transcrits disponibles dans la banque Ensembl

2.2.3 Expression et transcriptome

L'étude des transcrits d'une cellule (transcriptome) selon le type cellulaire ou l'organe auquel elle appartient ou selon les conditions extérieures auxquelles elle est soumise offre de nombreuses informations sur :

- L'expression d'un gène dans un type cellulaire ou tissu donné,
- L'évolution de l'expression des gènes pendant le développement de l'organisme,
- L'altération de l'expression des gènes lors du traitement par un médicament,
- La modification de transcription d'un gène due à une maladie génétique…

Parmi les méthodes d'évaluation de l'expression des gènes (profil d'expression), les *microarrays* sont parmi les plus anciennes (Augenlicht et al., 1987) et les plus utilisées, bien qu'elles commencent à être détrônées par les méthodes de séquençage RNAseq.

Schématiquement, la technique des *microarrays* repose sur la complémentarité des brins d'ADN et la faculté de deux brins complémentaires à se rassembler en double brin (Southern et al., 1999). Les *microarrays* possèdent des centaines de milliers d'alvéoles, chacune comportant des millions de courts brins d'ADN spécifiques (*probes*) susceptibles de s'hybrider à des fragments de cDNA liés à des sondes fluorescentes et obtenus à partir de la population des mRNA de la cellule. La quantité de fluorescence détectée dans les cellules par un scanner sera corrélée au degré d'expression des différents gènes. Cette technique possède plusieurs inconvénients notamment, une spécificité insuffisante pour distinguer tous les mRNA dont les séquences sont proches et donc, pour distinguer les variants d'épissage. De plus, le bruit inhérent à cette technologie pose de nombreux problèmes qui nécessitent le développement de méthodes statistiques élaborées afin d'associer au mieux les intensités de pixels de l'image à l'expression réelle des gènes. Elle reste pourtant une méthode mature, fiable et très utilisée.

Les *microarrays* sont de plus en plus supplantés par la méthode de RNAseq. Cette méthode repose sur le séquençage à haut débit des fragments de cDNA obtenus à partir de la population de mRNA de la cellule. Les lectures sont alignées avec le génome de référence et comptées pour déterminer le profil d'expression de chaque transcrit. Cette méthode fournit une mesure plus exacte de l'expression des gènes ainsi qu'une précision des séquences lues qui autorise la discrimination des variants d'épissage. Le nombre d'expériences RNAseq disponibles publiquement est bien plus faible que celles provenant des *microarrays*, mais est en constante augmentation.

Cependant, il faut noter que notre connaissance des transcrits reste encore partielle et partiale ; d'une part, à cause de la limite des méthodes de détection (les transcrits très peu exprimés ne sont pas aisément détectés) et d'autre part, parce que l'expression des différents transcrits dépend de conditions (intra- ou extra-cellulaires, stades de développement, pathologies, etc…) qui n'ont pas toutes été étudiées, loin s'en faut !

Dès lors, bien que les données d'expression des gènes dans des conditions 'normales' soient souvent disponibles et fournissent une information précieuse pour l'élucidation de leur fonction, le degré de perturbation de leur expression dans le contexte d'une maladie génétique, qui apporterait un éclairage complémentaire essentiel à la fois pour la recherche et la médecine, reste encore largement inconnu.

2.2.4 Maladies, points d'entrée du « diseasome »

2.2.4.1 La maladie, un concept difficile à cerner

Le grec Alcmaeon de Croton (500 avant J.C.) fut le premier à pratiquer des dissections pour produire de la connaissance anatomique à partir d'observations expérimentales (Bestetti et al., 2014). Sa définition de la maladie reposait sur l'égalité (isonomie) des "puissances opposées composant" un corps humain (l'humide et le sec, le froid et le chaud, etc…). Lorsque l'isonomie est rompue, la maladie survient. Selon lui, la maladie pouvait provenir du sang, de la moelle ou du cerveau. Mais elle pouvait également provenir de facteurs extérieurs comme l'eau, l'environnement ou la violence. La médecine traditionnelle chinoise est également basée sur l'idée que tout organisme sain était en équilibre entre deux forces contradictoires le Yin et le Yang (Wang et al., 2005). Cet équilibre était considéré comme un jeu complexe entre le corps et l'esprit.

La définition même d'une maladie, qui semble évidente à première vue, pose des problèmes de définition spécifique (Scully, 2004). Elle peut varier en fonction du contexte ethnique, de la classe sociale ou du genre ou même, du contexte historique. Ainsi l'ostéoporose (fragilité excessive du squelette), reconnue officiellement comme maladie par l'organisation mondiale de la santé en 1994, est passée d'une caractéristique normale du vieillissement à une pathologie. A l'inverse, l'homosexualité, après avoir été considérée au début du vingtième siècle comme un dérèglement endocrinien soigné par traitement hormonal, fut reconsidérée comme un trouble mental traitable par électrochocs, puis catégorisée comme maladie en 1974. L'hyperlaxité, ou l'élasticité excessive de certains tissus, considérée comme normale et même, comme un atout pour les contorsionnistes, gymnastes, pianistes ou ballerines, est maintenant considérée comme un dysfonctionnement.

De plus, la « découverte » de nouvelles maladies peut être en partie motivée par des aspects financiers en terme de création de nouveaux marchés, ainsi voyant le succès commercial du viagra pour traiter les dysfonctionnements sexuels chez l'homme, les compagnies pharmaceutiques se sont mises à la recherche d'un trouble semblable chez la femme (Moynihan, 2003). L'aspect financier peut donc rendre la notion de maladie encore moins objective.

Enfin, on peut relever que beaucoup de personnes sourdes ne se considèrent pas comme infirmes, mais comme une minorité linguistique (Scully, 2004).

Dans ce contexte disparate des définitions d'une maladie, les nouvelles capacités d'identification des variations génétiques et les notions de marqueurs génétiques qui en résultent introduisent de nouvelles problématiques, telle celle de savoir si une personne avec une prédisposition génétique à une maladie doit être considérée comme malade (Scully, 2004).

Un autre éclairage concerne les diverses perceptions d'une maladie et la volonté qui en découle d'utiliser l'argent publique pour la soigner. Ainsi, une étude a été menée en Finlande auprès de 3000 'profanes', 1500 médecins, 1500 infirmiers et 200 membres du parlement afin d'étudier la perception de différents phénotypes ou comportements comme correspondant ou non à une maladie (Tikkinen et al., 2012). Cette étude révèle qu'à quelques exceptions près (cancer du sein, pneumonie, malaria sont considérés à la quasi-unanimité comme maladies tandis que fumer, vieillir sont considérés à la quasi-unanimité comme n'étant pas une maladie), il y avait de profonds désaccords sur la classification de différents états comme étant maladifs ou pas (intolérance au lactose, boulimie, anorexie, insomnie, infertilité, alcoolisme…).

Les noms mêmes des maladies n'ont aucune structure ou ontologie rigoureuse et peuvent procéder des noms des découvreurs de la maladie (Bardet-Biedl (Bardet, 1995; Biedl, 1995)) ou de la simple énumération de phénotypes (*Mental retardation, epileptic seizures, hypogonadism and hypogenitalism, microcephaly, and obesity*). Ces noms peuvent également évoluer au fil du temps. Ainsi, après avoir signalé à la base OMIM en mai 2014 que deux entrées de maladies dans la base portaient exactement le même nom « *ALACRIMA, CONGENITAL* », j'ai pu constater peu après que la deuxième entrée a été renommée « *ALACRIMA, CONGENITAL, AUTOSOMAL RECESSIVE* ».

Enfin, question qui s'est avérée fondamentale dans le cadre de MyGeneFriends, comment distinguer des maladies distinctes des sous-types d'une même maladie ? C'est une notion en pleine évolution, comme le souligne l'éclatement récent de l'entrée Bardet-Biedl dans la base OMIM qui regroupait tous les sous-types (BBS1, BBS2, BBS3…), en plusieurs entrées, une par sous-type, devant maintenant être considérées comme des maladies distinctes.

Ces problématiques liées à la nature et à l'identité des maladies, risquent d'être encore plus exacerbées avec l'introduction du paradigme des 4P.

2.2.4.2 Les maladies à l'ère de la biologie des systèmes

Suite à l'introduction de la biologie des systèmes et au développement des technologies à haut débit à prix abordable, la médecine connaît un tournant avec l'émergence d'un nouveau paradigme, dit des 4P (Tian, 2011) (Cano et al., 2014), qui fait référence au fait que la médecine moderne est :

- **Prédictive**, en faisant interagir la pratique clinique et la recherche biomédicale afin d'élaborer des simulations et des modèles prédictifs,
- **Personnalisée**, car l'accès aux données génomiques individuelles (Fernald, et al., 2011) (séquence, variations, phénotypes) permet d'adapter les thérapies (Hendlisz, 2015), traitements et médicaments à chaque patient. «Le bon médicament pour la bonne personne » (Allison, 2008),

- **Préventive**, en gérant les facteurs de risques de maladie au cours du temps pour réduire les maladies et handicaps potentiels et augmenter la durée de la vie en bonne santé des populations,
- **Participative**, en cherchant à obtenir la participation active des patients à leur traitement pour améliorer la communication entre patients et corps médical, pour partager des expériences et perceptions d'un traitement, pour régler des problèmes de régulation et de confidentialité des données, pour rendre accessibles à la société des données médicales…

De plus, étant imprégnée de la biologie des systèmes, cette médecine ne se concentre plus sur une cible spécifique à la fois mais tente d'identifier les perturbations dans le contexte des réseaux biologiques. Cette approche permet d'avancer sur des questions pour lesquelles des approches plus conventionnelles échouent.

2.2.4.3 Les variations, une clef des maladies génétiques

L'accès massif aux séquences de génomes humains a permis de révolutionner l'approche des maladies génétiques en accélérant l'identification des mutations et gènes causaux.

Ainsi, seuls 99.9% du génome sont identiques entre deux individus, et une centaine de nouvelles différences, dues à des imperfections de réplication du génome, apparaissent à chaque génération. Ces variations peuvent concerner des insertions ou délétions de nucléotides dans le génome, appelées INDEL ou la substitution d'un seul nucléotide appelée SNV (*Single Nucleotide Variation*) ou SNP (*Single Nucleotide Polymorphism*). Ces variations entre individus sont appelées *polymorphismes,* si l'évènement est observé dans plus d'1% de la population et *mutations,* pour des fréquences extrêmement rares. Certaines variations sont directement associées aux maladies alors que d'autres influencent la probabilité ou susceptibilité d'avoir une maladie (en l'augmentant ou la diminuant), ainsi les allèles E2 et E4 du gène ApoE influencent la susceptibilité à l'Alzheimer (Serrano-Pozo et al., 2015). Dès lors, la présence d'une variation liée à une maladie donnée ne suffit pas toujours à garantir que la personne sera malade. Pour décrire le lien entre la mutation et son expression, on utilise la notion de pénétrance qui décrit la proportion d'individus malades parmi tous les individus possédant la mutation.

Il est bien plus simple de détecter la causalité entre une variation et une maladie que de comprendre le cheminement logique et fonctionnel qui va du nucléotide aux phénotypes observables dans une maladie. Ce constat a placé les technologies de séquençage de dernière génération au cœur de l'étude des maladies génétiques entrainant la constitution de nombreux projets à grande échelle et la découverte de millions de variations génétiques dans les génomes humains (Henn et al., 2015) .

Le projet 1000 génomes en particulier (Clarke et al., 2012) a pour but de caractériser de manière la plus exhaustive possible les variations dans les populations humaines. Il permet d'établir la fréquence des variations et de définir la MAF (*Minor Allele Frequency*), permettant de

discriminer un 'simple' polymorphisme d'une mutation. En 2012, ce projet proposait plus de 260To de données disponibles au public défiant les limites du volume de données traitées. En effet, pendant la génération de données, la vitesse de transferts FTP a été jugée insuffisante pour les échanges de données entre les centres de séquençage, les centres de stockage (NCBI et EBI), les groupes effectuant un contrôle qualité et les groupes chargés d'exploiter les séquences. Dès lors, certains échanges se sont faits *via* l'envoi de disques durs. Finalement, le changement de protocole (UDP au lieu de FTP) ainsi que l'utilisation de formats de fichiers plus compacts (BAM et FASTQ au lieu de SRF) permirent de faire face à ce problème de volume et vélocité.

Les variations détectées sont cataloguées et stockées dans des bases de données comme dbSNP (Sherry et al., 2001) ou Clinvar (Landrum et al., 2014). La base dbSNP est centrée sur l'exhaustivité et possède des dizaines de millions de variations, dont une partie importante est du bruit. ClinVar possède beaucoup moins de mutations, mais se concentre sur celles liées à des maladies, avec preuves à l'appui.

2.2.4.4 Les phénotypes, l'expression macroscopique des maladies

On désigne généralement par phénotype, les traits observables au-delà du niveau moléculaire, comme l'anatomie, le comportement, etc…(Deans et al., 2015). Le phénotype fut longtemps utilisé pour classifier les espèces ou observer les effets de croisements entre individus et déduire la fonction des gènes. Les variations phénotypiques entre individus, liées aux variations génétiques, sont également à la base de la sélection naturelle.

Alors que les variations vont provoquer la maladie génétique, les phénotypes vont permettre de la détecter et l'étudier par l'expert humain. Les phénotypes sont généralement utilisés pour reconnaitre, définir et diagnostiquer les conditions pathologiques ou maladies.

Un premier lien entre phénotype et maladie (Bestetti et al., 2014) nous vient d'un papyrus Egyptien datant de 1500 avant J.C. dans lequel nous découvrons que des anomalies dans le pouls périphérique étaient considérés comme liés à une maladie cardiaque. Depuis, les phénotypes sont systématiquement utilisés en médecine pour diagnostiquer un patient et lui associer une maladie. Cependant, le lien entre phénotypes et maladie est complexe. Certains phénotypes se retrouvent dans un grand nombre de maladies, d'autres sont très spécifiques. On distingue également des phénotypes majeurs, essentiels pour diagnostiquer la maladie, et des phénotypes mineurs qui sont plus facultatifs car moins fréquemment observés.

Des données phénotypiques sont disponibles pour un large spectre d'organismes : plantes, insectes, mammifères. Mais elles sont décentralisées et présentes majoritairement sous forme non structurée : textes des publications, données supplémentaires, images, etc… Même les données plus structurées peuvent être hétérogènes par leur nature : qualitatives (« *large head* », « *rare* ») ou quantitatives avec des mesures exactes. Le phénome est loin derrière le génome en termes de standardisation et de structuration (Deans et al., 2015). Des initiatives ont donc été mises en place

pour structurer les phénotypes, et la plus connue, qui est aussi la plus récente, est l'ontologie HPO (*Human Phenotype Ontologie*) (Kohler et al., 2014).

2.3 Exploitation des données

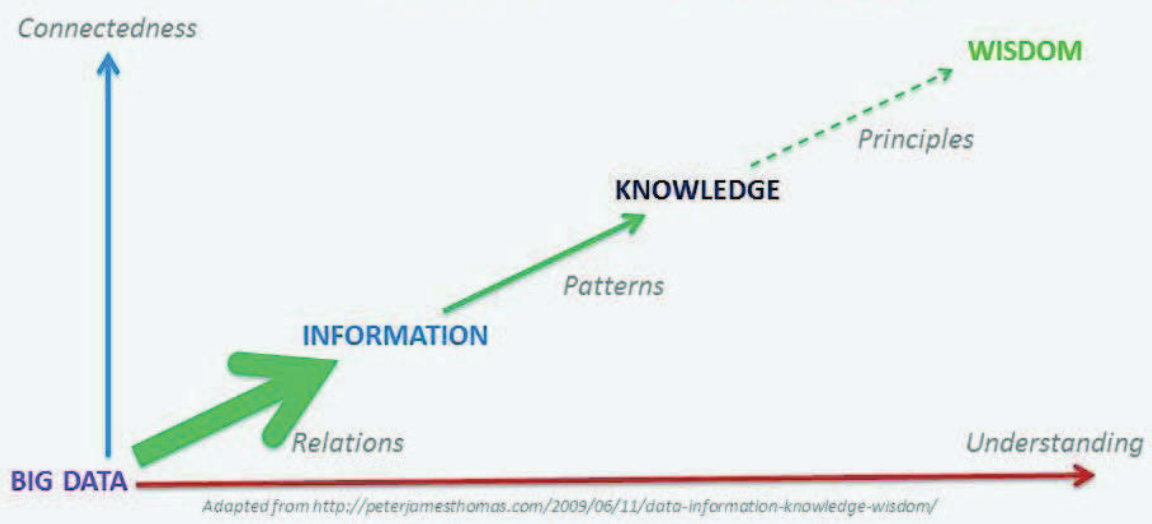


Figure 10: Evolutions successives et conditions pour passer des mégadonnées à la 'sagesse'.

Les nouvelles technologies à haut débit ont permis à la biologie de collecter dans des banques d'énormes quantités de données brutes ou traitées et un taux de publication soutenu alimente les archives en textes non structurés. La majorité de ces données sont difficilement exploitables pour générer de la connaissance, aussi bien par les humains que par les machines. Afin de relever ces défis, de nombreuses méthodes ont vu le jour pour extraire de l'information, la classifier et la structurer en ontologies et enfin, la visualiser efficacement afin de permettre à l'expert humain de l'analyser.

2.3.1 Structurer les connaissances.

La structuration des connaissances est une composante essentielle de toute science, elle organise les informations en les hiérarchisant, les définissant et créant des liens entre elles. L'acceptation par une communauté de cette standardisation et du vocabulaire contrôlé qu'elle impose est une étape essentielle et un enjeu complexe (Tenenbaum et al., 2014).

Un exemple célèbre de structuration et de standard en biologie est la classification des organismes vivants par Carl von Linné en 1735 dans son « *Systema naturae* ». Depuis, de

nombreuses classifications et ontologies ont vu le jour pour hiérarchiser et systématiser diverses informations biologiques.

Le mot ontologie nous vient du grec *onto* (être) et *logos* (rationalité, intelligence) pour désigner l'étude des propriétés générales de ce qui existe. En informatique, ce terme désigne un ensemble de concepts reliés et hiérarchisés décrivant un domaine. Le développement d'ontologies est gourmand en temps, nécessite la collaboration d'experts et surtout, l'ontologie doit être acceptée et utilisée par la communauté scientifique (Deans et al., 2015). L'extraction de termes est réalisée manuellement ou par des techniques de fouille de données. Les ontologies biologiques ont deux grands avantages (Blake, 2004), d'une part, elles fournissent aux experts un vocabulaire partagé permettant de clarifier la communication et d'autre part, elles permettent aux systèmes informatisés d'extraire les données et d'y appliquer diverses algorithmes pour les explorer.

Une des bases des ontologies est le vocabulaire contrôlé. Le langage naturel possède plusieurs caractéristiques qui rendent difficile son appropriation par des systèmes automatisés. Parmi ces caractéristiques, on peut citer :

- La différente forme des mots (pluriels, temps, genre…),
- L'utilisation de synonymes (par exemple « physician » et « doctor »).

Un vocabulaire contrôlé est essentiel pour désigner les mêmes choses de la même façon et permettre la récupération, avec une grande sensibilité et spécificité, des informations de documents d'intérêt. Les documents peuvent alors être indexés avec un nombre réduit de mots appartenant à ce vocabulaire contrôlé (Dumais, 2004).

Des ontologies spécifiques à la biologie ont été créées pour structurer les termes cliniques (MESH), les phénotypes (HPO), la fonction ou localisation des gènes (GO), les séquences nucléiques (SO), la relation hôte-pathogène etc… Les nombreuses ontologies existantes ne sont pas toujours compatibles entre elles ou facilement intégrables dans un même traitement analytique. Pour y faire face, plusieurs démarches ont étés entreprises pour faciliter leur interopérabilité et relever le défi de réunir des ontologies dispersées. On peut citer la définition de termes avec des règles logiques ou des standards comme OWL (*Web Ontology Language*) ou OBO (*Open Biomedical Ontologies*).

2.3.2 Fouille de textes

Seule une petite partie des informations disponibles dans le monde est présente sous forme structurée (base de données, ontologies, etc…), la plupart est sous forme de textes écrits par des humains pour d'autres humains (Cunningham et al., 2013). La fouille de textes (ou *text mining*) est une approche pluridisciplinaire faisant appel à la linguistique, aux statistiques, à la fouille de données et à l'apprentissage automatisé (Hotho et al., 2005) afin d'extraire des informations intéressantes de ces textes sous forme structurée et compréhensible par l'ordinateur.

La première étape est généralement la *tokenization*, qui correspond au processus de transformation d'un texte ou document en une liste de phrases, mots et symboles. Afin de ne plus manipuler de chaines de caractères par la suite, certaines librairies comme gensim créent un dictionnaire qui associe chaque mot à un identifiant numérique entier. Puis les mots sont généralement filtrés en utilisant un fichier *stopwords* qui contient une liste de mots peu informatifs comme *the*, *to*, *for*… Un filtre par fréquence peut également être appliqué pour enlever les mots très fréquents ou très rares. Mais le filtrage par fréquence impose de définir des seuils précis. La 'lemmatisation' et la 'racinisation' (*stemming* en anglais) sont deux des méthodes permettant de normaliser les mots. La 'lemmatisation' réduit les verbes à leur forme infinitive et les noms à leur forme singulière, mais comme elle repose sur des informations grammaticales, elle est lente et imparfaite, on lui préfère généralement la 'racinisation' qui se base sur des algorithmes et des règles appliquées aux mots, le 'racinisateur' le plus connu reste celui de Porter (Porter, 1980). Certaines notions ne peuvent être correctement représentées par des mots uniques, et des processus comme le *chunking* (détection de groupes nominaux) (Tellier et al., 2012) permettent de créer des mots clefs de plusieurs mots. Enfin d'autres méthodes, comme la reconnaissance d'entités spécifiques dans un texte (*Named Entity Recognition* ou NER) viennent compléter les multiples techniques de fouille de textes.

La fouille de textes peut avoir des usages variés et plus ou moins complexes, allant de la simple détection de termes précis, à la détection de relations complexes entre termes (interactions protéine-protéine (Marcotte et al., 2001), régulation…) en passant par l'indexation et la classification de documents avec labellisation automatique. Dans ce dernier cas, on peut associer des valeurs d'entropie (Lochbaum and Streeter, 1989) à chaque mot et récupérer les mots aux plus hautes valeurs d'entropie qui couvrent tout le document. Dans certains cas de classification de documents, les techniques d'apprentissage automatique ou *machine learning* pourront être utilisées, reposant souvent sur un apprentissage à partir d'un ensemble d'entités (ici des documents) déjà annotés par des humains. Des classifieurs, dont un des plus simples à mettre en œuvre et le plus connu est le « *Naive Bayesian Classifier* », permettent alors d'annoter automatiquement un série de documents test.

UIMA (*Unstructured Information Management Architecture*) , GATE (*General Architecture for Text Engineering*) (Cunningham et al., 2013) et NLTK (*Natural Language Toolkit*) font partie des *frameworks* gratuits les plus utilisés pour la fouille de données. UIMA a, par exemple, été utilisé par le superordinateur Watson d'IBM pour analyser et annoter de grandes quantités d'informations textuelles.

Une sous-discipline de la fouille de textes est la fouille de publications scientifiques dans le domaine de la biologie, discipline à laquelle on se réfère sous le nom de BioNLP (*Biological natural language processing*). L'intérêt de cette discipline est facilement compréhensible étant donnée l'énorme quantité de connaissances scientifiques enfermées dans les dizaines de millions

de publications biologiques ou médicales disponibles dans la base PubMed. Le but de la BioNLP est d'une part, d'identifier la référence à des entités biologiques dans un texte (BER ou *Biological Entity Recognition*) tels que les gènes, les protéines, les variations ou les maladies et d'autre part, de découvrir et d'extraire des interactions entre protéines ou avec des médicaments. De nombreux logiciels ont vu le jour pour relever ces défis, comme GNormPlus (Wei et al., 2015) spécialisé dans la reconnaissance de gènes et de protéines.

Enfin, plusieurs outils comme Chilibot (Chen and Sharp, 2004) ou Evex (Van Landeghem et al., 2013) tentent d'extraire des relations entre les entités biologiques détectées, mais la complexité du langage naturel rend cet exercice très difficile comme le montre la Figure 11.

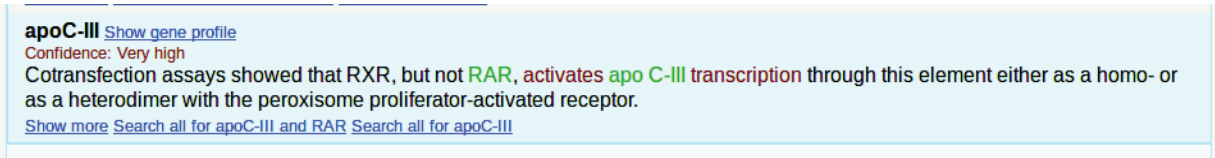


Figure 11: L'application web EVEX détecte les relations de régulation entre les gènes. On peut voir dans cet exemple la difficulté de cette tâche. Lorsqu'on demande à EVEX de détecter les gènes régulés par RAR elle détecte avec grande certitude, mais à tort, que RAR régule apo C-III, alors que le texte dit clairement « ...ont montré que RXR, et non RAR, active la transcription de apo C-III... »

2.3.3 Visualisation des données

La visualisation est une étape clef pour l'analyse et la compréhension des données à haut débit. Une visualisation doit être claire, pertinente, transmettre une information permettant l'interprétation d'évènements biologiques. Une bonne visualisation doit être en mesure de faire passer le ou les messages principaux des données sans que l'utilisateur ne soit submergé par leur nombre ou leur complexité. Elle peut également permettre d'interpréter facilement l'évolution de l'information (par exemple, dans le temps), les liens entre entités ou groupes d'entités, les métadonnées décrivant information comme sa force descriptive, source, seuil de confiance, etc… Enfin, sur un plan plus pragmatique, une visualisation efficace permet à l'utilisateur d'exercer facilement des tâches analytiques basiques comme le tri, filtrage, la détection d'extrêmes ou d'anomalies, la classification et corrélation des informations (Amar et al., 2005)…. Pour résumer, une visualisation de données adaptée fournira une valeur ajoutée en facilitant l'interprétation par l'expert humain (Chen et al., 2014), tandis qu'une mauvaise visualisation réduira la valeur de ces données.

Ce sont les techniques de visualisation de données, ou DataViz, qui vont créer une interface d'échange entre l'expert humain et les longues listes ou tables de chiffres obscurs. Le rôle de la visualisation pour faciliter l'accès à ces données est étudié par le champ d'études de l'exploration

visuelle des données (*visual analytics*) (Cook and Thomas, 2005), la « science de raisonnement analytique facilité par des interfaces visuelles interactives ».

Le domaine de la visualisation des données fait face à plusieurs défis définis récemment par Wong et collaborateurs (Wong et al., 2012) :

1. Les données sont souvent stockées sur un disque puis analysées. Avec l'augmentation constante de la quantité de données, celles-ci devront être visualisées et analysées directement en mémoire vive,
2. Les capacités cognitives humaines restant constantes alors que les données croissent, les interfaces utilisateurs devront s'adapter pour résumer judicieusement les données et les présenter de manière optimale,
3. Le fait que certains algorithmes très utilisés d'exploration visuelle de données sont inadaptés aux mégadonnées et provoquent une utilisation exponentielle des mémoires et ressources processeurs, requiert de les identifier et de trouver des solutions alternatives,
4. Des techniques intuitives de quantification de l'incertitude sur les données sont à élaborer, ainsi que la capacité de gérer les données manquantes,
5. Enfin, il devient urgent de de développer des librairies et *frameworks* pour faciliter l'incorporation des techniques de visualisation dans les outils et sites web.

Très schématiquement, en biologie, on peut distinguer plusieurs grands types d'usages des techniques de visualisation (Gehlenborg et al., 2010) :

- les *genome browsers* comme ceux proposés par Ensembl (Cunningham et al., 2015) ou UCSC (Rosenbloom et al., 2015) qui permettent de parcourir un génome entier en modifiant l'échelle observée (du chromosome au nucléotide) et surtout, de corréler de nombreuses informations présentes sous forme de *track* (transcrits de différentes sources de données, informations de régulation et d'expression, des variations et des éléments répétés…),
- les outils de visualisation des réseaux et liens entre entités biologiques. De nombreux outils ont étés développés pour visualiser et analyser les réseaux biologiques dont le plus connu est Cytoscape (Kohl et al., 2011). Des sites web comme celui de la banque de liens fonctionnels STRING (Szklarczyk et al., 2015) représentent les interactions entre protéines sous forme de réseau. La visualisation des *pathways* par des outils comme KEGG Atlas (Okuda et al., 2008) permet d'explorer de nombreuses voies cellulaires, comme les voies métaboliques,
- les *heatmaps*, *dendogrammes*, *scatter plots* ou *profile plots* sont communément utilisés pour l'analyse de données d'expression des transcrits ou l'évaluation des proximités phylogénétiques,
- enfin, la visualisation d'alignements multiples de séquences d'objets biologiques (génomes, transcrits, protéines). Un logiciel emblématique est Jalview (Waterhouse et al., 2009) qui fournit de nombreuses fonctionnalités visuelles et permet aux experts de repérer des motifs,

de détecter des erreurs d'alignement, de manipuler les alignements pour permettre des améliorations manuelles….

A cela s'ajoute de nombreux développements de représentation et visualisation dans un cadre tridimensionnel, développements initiés autour des structures de protéines qui s'étendent à présent à une multitude de données (génomes, gènes, réseaux protéiques ou biologiques, maladies…). Des initiatives, orientées vers les développeurs, voient le jour, pour éviter de « réinventer la roue »et utiliser des bibliothèques de composants graphiques pré-établis, permettant d'afficher des protéines sous forme schématique ou tridimensionnelle, des chromosomes, des interactions... comme BioJS (Yachdav et al., 2015), ou des magasins d'applications comme celui de Cytoscape (Lotia et al., 2013).

3 L'avenir des mégadonnées en biologie

Nous allons conclure cette présentation du contexte dans lequel s'est inscrit mon travail de thèse en montrant comment les réseaux sociaux influencent les nouveaux outils de la biologie moderne, puis nous verrons que la biologie est un domaine particulier au sein des mégadonnées qui nécessite une approche adaptée. Nous nous demanderons alors si l'avenir des mégadonnées biologiques se trouve exclusivement dans la machine ou dans une association machine-expert, association qui est au cœur des travaux réalisés dans MyGeneFriends.

3.1 De nouvelles opportunités dans une nouvelle ère

La bio-informatique se développe intensément sur de nombreux fronts et notamment, sur celui de l'amélioration de l'accès des biologistes aux mégadonnées. Ainsi, l'intégration de nombreuses composantes créées ou popularisées par les grands acteurs du Web 2 .0 a fait émerger de nouvelles ressources bio-informatiques (ou des évolutions des acteurs historiques) qui s'inscrivent dans des normes et attentes généralement présentes sur des sites non-dédiés à la science. Cependant, on peut noter que, souvent, les sites web bio-informatiques se concentrent sur un ou deux aspects qu'ils vont mettre en avant.

Emblématique du Web 2.0, l'aspect participatif et la mobilisation d'utilisateurs créateurs d'information, se retrouve dans des sites tels, Proteopedia (Hodis et al., 2010) ou WikiGene (Hoffmann, 2008). GeneTalk (Kamphans and Krawitz, 2012) propose à ses utilisateurs de voter pour la 'pertinence' de différentes variations génomiques par rapport à une maladie, et WebApollo (Lee et al., 2013), etc…) généralise la collaboration autour de l'annotation de séquences. Des aspects sociaux et d'interaction entre utilisateurs se retrouvent aussi sur un site comme Coremine (www.coremine.com), qui permet d'explorer les connections entre concepts biomédicaux.

L'accès aux données sous forme d'APIs se généralise. Les acteurs les plus connus du secteur, comme Ensembl ou NCBI (avec Entrez), permettent d'accéder à une grande partie de leurs informations à travers des APIs. D'autres services, comme MyGene.info (Wu et al., 2014), se spécialisent dans l'offre d'accès par API à des informations diverses.

Enfin, l'instantanéité est devenue une exigence, avec des sites web effectuant des analyses et leur visualisation en temps réel (comme gene.iobio.io par exemple), tout comme l'est l'interconnexion massive des informations et des gènes, avec un traitement et affichage sous forme de réseaux proposée par des sites comme GeneMania (Zuberi et al., 2013).

Mais c'est le rapprochement des réseaux sociaux avec la biologie qui promet les avancées les plus excitantes.

3.2 Les influences des réseaux sociaux sur la biologie moderne et la médecine

Les réseaux sociaux ont une place importante dans la vie de millions de personnes, et leur place dans le monde de la recherche, et en particulier en biologie, ne cesse de croitre. On peut relever, à ce sujet, l'émergence de plusieurs grandes tendances. Premièrement, les chercheurs utilisent de plus en plus les réseaux sociaux pour communiquer, créer de nouvelles collaborations, ou se tenir informés (Van Noorden, 2014), et cela devient encore plus flagrant avec la création de réseaux sociaux académiques comme ResearchGate. Deuxièmement, les patients eux-mêmes se servent des réseaux sociaux pour échanger des informations pertinentes sur leur condition et les moyens de la soulager. Troisièmement, le concept de réseau social est mis en avant pour favoriser la connexion entre médecins et patients. Quatrièmement, l'ajout de fonctionnalités sociales aux sites web bio-informatiques stimule l'échange de protocoles et de données. Enfin, l'utilisation d'algorithmes et d'approches issus des réseaux sociaux permet de révolutionner la réponse à des problématiques bio-informatiques.

3.2.1 Les chercheurs et les réseaux sociaux

Après quelques échecs pour créer un « Facebook pour la science », on a vu apparaître des réseaux sociaux orientés vers la communauté de chercheurs (Tableau 1 du premier chapitre) offrant, en plus des conventions habituelles des réseaux sociaux, la possibilité de partager des publications, de suivre les statistiques de visites de profil, de visualisation ou téléchargement de publications… (Van Noorden, 2014). Les membres peuvent également poser des questions scientifiques comme « *Can anyone suggest a NGS service provider for small RNAs?* », poster des rapports sur les publications avec la fonctionnalité *Open Review*, recevoir des suggestions d'emploi en fonction des informations fournies au système. Un équivalent français moins connu, MyScienceWork (www.mysciencework.com) permet même de suivre les différents évènements organisés dans le monde de la recherche.

Pourtant les statistiques ont montré que la majorité des utilisateurs sont passifs sur ResearchGate ou Academia.edu et se contentent de mettre à jour leur profil au cas où quelqu'un voudrait les contacter. Relativement peu d'utilisateurs profitent de la suggestion de personnes aux centres d'intérêt proches ou de publications à lire. Toutefois, le bilan est plus contrasté qu'il n'y paraît, ainsi l'utilisation des réseaux sociaux traditionnels, moins populaires chez les chercheurs que les réseaux spécialisés comme LinkedIn ou ResearchGate, est pourtant beaucoup plus interactive. La moitié de chercheurs utilisant Twitter suivent les discussions sur des sujets liés à la recherche et un peu moins de la moitié l'utilisent pour poster des commentaires sur des sujets relatifs à leur champ de recherches (Van Noorden, 2014).

3.2.2 Facebook comme plateforme d'échanges d'informations médicales

De plus en plus de patients utilisent Facebook pour échanger des informations médicales personnelles et recevoir un retour et un support émotionnel, c'est notamment le cas de patients atteints de diabète (Greene et al., 2011). Les patients privilégient même internet au médecin, lorsqu'il s'agit de rechercher des réponses à des questions de santé. Hormis la demande de conseils pour se soigner et le partage d'informations, les patients cherchent souvent, chez les autres membres du groupe, des corrélations entre un médicament et des effets secondaires qu'ils ont ressentis. Selon l'étude, presque un quart des *posts* contenait des informations sensibles sur la gestion du diabète qui auraient peu de chances d'être communiquées par le patient à son médecin, car liées à l'alcool ou aux triathlètes diabétiques (pratiquants diabétiques du triathlon : natation, cyclisme, course), par exemples.

3.2.3 Création de réseaux sociaux spécialisés pour les médecins, les patients, les chercheurs

> *Health 2.0 is participatory healthcare. Enabled by information, software, and community that we collect or create, we the patients can be effective partners in our own healthcare, and we the people can participate in reshaping the health system itself.*
>
> *Dr. Eytan, Tweet, 13 Juin 2008*

Après l'émergence d'un Web 2.0, on voit celle d'un Healh 2.0 répondant au concept Participatif des nouveaux paradigmes (4P) de la médecine moderne.

Des sites web comme hellohealth.com sont des réseaux sociaux spécialisés dans l'interaction entre médecins et patients (Hawn, 2009). Les patients peuvent dialoguer avec leur médecin par messagerie instantanée, visiter les pages de profil de leur médecin. En cas de symptômes inquiétants, les utilisateurs de ce réseau social peuvent demander l'avis d'un médecin ou même effectuer une « cyber-visite ». Cette nouvelle approche rencontre cependant de nouvelles problématiques à gérer, telle que la garantie de la nature privée des données (plusieurs procès ont été lancés par des patients accusant les médecins de violer la nature privée de leurs informations médicales).

3.2.4 L'aspect social en recherche scientifique

L'aspect social et la notion de partage pénètre également les ressources bioinformatiques. Par exemple, myExperiement est une banque de flux de travaux (*workflows*) possédant des aspects sociaux (groupes, profils) et qui se considère comme une expérience destinée à découvrir si la communauté scientifique est suffisamment disposée à échanger pour bénéficier du réseau constitué par un site web social (Roure et al., 2008). Après analyse des retours utilisateurs, ses créateurs ont remarqué que le crédit pour les informations/objets partagés et un contrôle précis de la visibilité et du partage de ces objets étaient essentiels pour que les scientifiques acceptent d'utiliser un site web social.

3.2.5 L'utilisation des algorithmes et approches issus des réseaux sociaux pour retirer des informations des réseaux biologiques

Les réseaux sociaux ont également des implications plus directes en sciences. Des chercheurs s'inspirent des algorithmes propres aux réseaux sociaux pour les appliquer aux réseaux biologiques et découvrir de nouvelles connections entre gènes et maladies (Singh-Blom et al., 2013). La mesure de Katz (Katz, 1953) par exemple, utilisée pour mesurer le degré d'influence d'un nœud dans un réseau et prédire avec succès des amis dans les réseaux sociaux, a été adaptée et exploitée pour recommander des gènes en considérant un phénotype donné. L'analyse du réseau cellulaire du point de vue de l'influence des différents acteurs permet de mieux comprendre la causalité d'une tumeur (Crespo et al., 2015) et la détection de communautés (groupes dont les membres sont densément connectés, par exemple la famille, les collègues de travail, cercle d'amis, etc…) a permis d'identifier des modules fonctionnels dans le réseau de régulation de transcription de *E. coli.* (Liu et al., 2014).

3.3 Biologie et Informatique : c'est donnant – donnant

L'informatique a pris la relève de la physique pour nourrir et faire avancer la biologie avec des solutions aussi bien matérielles que logicielles pour :

- Acquérir les données (*scanners*, séquenceurs…),
- Stocker les données (disques, *cloud,* compression…),
- Analyser les données (algorithmes, parallélisation),
- Visualiser les données (cartes graphiques, 3D, techniques de DataViz).

Mais les connaissances en biologie commencent également à révolutionner l'informatique. On peut citer i) les algorithmes évolutionnaires qui s'inspirent de l'évolution pour apporter des

solutions à des problématiques énergétiques (Mahela et al., 2015) et qui retournent en biologie en s'appliquant notamment, au problème de l'alignement multiple de séquences (Choudhury, 2003) ou à l'analyse de l'activité neuronale (Zarifia et al., 2015), ii) les systèmes immunitaires artificiels (algorithmes inspirés par les principes du système immunitaire comme l'apprentissage et la mémorisation) qui permettent de détecter l'infection des serveurs par des virus (Rasheed and Ghazali, 2010) ou de créer des logiciels résistants au *crashs* et bugs (Sidiroglou et al., 2005) et iii) les algorithmes mathématiques (Bonabeau et al., 2000) qui s'inspirent de l'étude du comportement social des insectes afin d'aborder les problèmes d'amarrage (*docking*) entre ligand et protéine (Korb et al., 2010).

3.4 La biologie : un domaine hors norme pour les mégadonnées

Malgré ce 'jeu de ping-pong' entre informatique et biologie pour résoudre des problématiques communes, la massification en biologie ouvre de nouveaux perspectives dus à la complexité des systèmes qu'elle révèle. En effet, pendant longtemps, la biologie a du se cantonner à une approche réductionniste (c'est-à-dire : se concentrer sur l'étude des éléments du système et non sur le système en entier). Cette approche s'est appuyée sur des principes fondateurs (Vidal, 2009), chacun à la base de disciplines spécifiques :

- Le gène est à la base de l'hérédité (Génétique et biologie moléculaire),
- La cellule est l'unité fondamentale de chaque organisme vivant (Biologie cellulaire),
- La chimie est à la base de la biologie (Biochimie),
- L'évolution des espèces s'effectue par sélection naturelle (Biologie évolutionnaire).

Les limites de cette approche réductionniste deviennent de plus en plus évidentes et ont entrainé, en 2009, lors du Nobel Symposium 146 dédié à la biologie des systèmes, la proposition d'un cinquième grand principe qui introduit les systèmes et réseaux complexes comme unités fondatrices de l'émergence, la maintenance et l'évolution du vivant (Ehrenberg et al., 2009).

Ce constat s'appuie sur une nouvelle discipline, la biologie des systèmes, qui étudie les systèmes biologiques constitués par les interactions, complexes, dynamiques et à échelles multiples, des macromolécules, métabolites, cellules, organes et organismes entiers ; systèmes à la base de l'ensemble des processus biologiques (Vidal, 2009). Ainsi, est mis en exergue, que quelques dizaines de milliers de gènes, 10^{13} cellules, quelques milliers d'enzymes ne peuvent décrire le phénomène de la vie sans qu'on y adjoigne les réseaux et connexions complexes, artisans essentiels des comportements et propriétés émergents du système. Le réseau biologique peut également être vu comme un ensemble de modules discrets qui interagissent entre eux. C'est le cas de la fanjiomique, la science qui étudie les formules de la médecine traditionnelle chinoise (TCM). Ainsi, à la différence des approches occidentales qui se basent sur l'identification de cibles spécifiques et l'action sur elles, la TCM ambitionne d'agir sur le système biologique dans son ensemble (Tian, 2011).

3.5 Machine toute puissante ou expert indispensable ?

Comme on vient de le voir, grâce aux biotechnologies à haut débit, le vivant est en train de révéler une complexité phénoménale. Dès lors, se posent de multiples questions sur les stratégies futures pour faire face à cette complexité et notamment, sur les places respectives de l'homme et des machines dans ces stratégies?

Les tâches, dont on croyait jusqu'à présent seul l'humain capable, deviennent peu à peu le terrain de jeu des ordinateurs qui franchissent les frontières, les unes après les autre. Ainsi, en 1997 l'ordinateur Deep Blue II construit par IBM a battu le champion du monde d'échecs Garry Kasparov (Campbell et al., 2002). Deep Blue II (comme ses prédécesseurs Deep Blue I, Deep Thougth and ChipTest) combinait la recherche matérielle (notamment, un *hardware* spécialisé qui apportait la vitesse et le parallélisme massif) et la recherche logicielle qui apportait la flexibilité.

En 2011, l'ordinateur IBM Watson a remporté la victoire au célèbre jeu télévisé américain *Jeopardy* (Lally and Fodor, 2011). Ce jeu consiste à trouver la question correspondant aux réponses fournies par le présentateur et a donc nécessité de la part de Watson, la capacité de :

- Comprendre des questions complexes (grâce à un analyseur syntaxique écrit en Prolog),
- Posséder un vaste domaine de connaissances (sous forme non structurée ou structurée par la librairie UIMA),
- Evaluer sa confiance dans ses réponses,
- Effectuer tous ces traitements en très peu de temps (par une parallélisation massive).

Pour la firme IBM, la victoire de Watson était surtout une démonstration des capacités d'une intelligence artificielle à prendre des décisions. Ainsi, peu après avoir conquis les jeux télévisés, Watson s'est lancé dans le diagnostic médical (Ferrucci et al., 2013).

L'idée de diagnostic médical automatique n'est pas nouvelle, et d'autres programmes ont étés construits par le passé pour s'y essayer. Dans ce cadre, le réseau bayésien est assez populaire, on le retrouve dans le diagnostic des douleurs abdominales ou le réseau HEPAR II (réseau bayésien spécialisé dans les maladies du foie dont la structure est construite à partir des connaissances de diagnosticiens experts et dont les paramètres sont appris automatiquement à partir d'une base de données de cas réels) (Onisko et al., 2001). Les créateurs de HPO (*Human Phenotype Ontology*) proposent également un outil appelé Phenomyzer qui, une fois un ensemble de phénotypes sélectionnés, classe les maladies les plus susceptibles de les expliquer. Mais les systèmes experts basés sur des règles définies par les êtres humains sont lents et chers à mettre en place. L'avantage de Watson, et de son approche DeepQA, est dans la génération d'hypothèses, la collection de preuves et l'évaluation des réponses possibles (comme le montre la Figure 12), tout en restant flexible et à jour par rapport aux informations les plus récentes. En médecine, de telles

preuves pourraient inclure les symptômes, l'histoire médicale du patient, l'histoire médicale de la famille du patient, les données démographiques, les médicaments disponibles, etc…

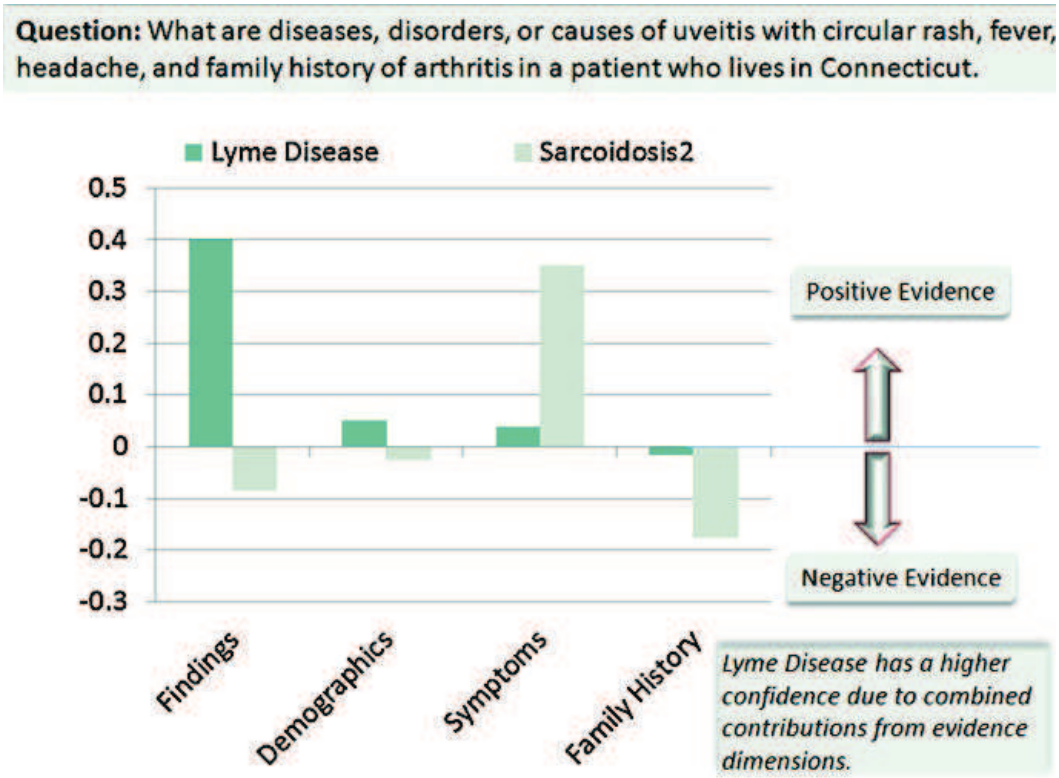


Figure 12: Application de Watson aux problématiques de diagnostic médical (figure reprise de (Ferrucci et al., 2013)) . Pour trouver la bonne réponse à la question, Watson s'appuie sur un ensemble de preuves et d'informations appartenant à diverses catégories, comme les symptômes, l'histoire familiale médicale du patient, les données statistiques sur la population générale, etc…

L'intérêt pour un système comme Watson de supplanter le diagnostic humain est d'éviter les erreurs de jugement faites par des humains et responsables d'une proportion non négligeable d'erreurs médicales. Ces erreurs de jugement sont dues à plusieurs causes:

- Non considération de diagnostics alternatifs, une fois le diagnostic initial posé,
- Problème de contextualisation des informations,
- Difficulté à juger correctement de l'importance d'une découverte récente,
- Perception erronée,
- Mauvaise utilisation des règles de la logique.

Ces problématiques de diagnostic médical s'appliquent tout autant en recherche théorique ou appliquée.

Les exploits de Watson sont pourtant nuancés par certains observateurs (Norman, 2011) qui opposent la force brute et les probabilités mathématiques de Watson à la reconnaissance de similarités et de motifs par l'expert humain et rappellent que les tentatives de diagnostic automatisé ont toujours dû se transformer en « aide au diagnostic », un objectif nettement plus modeste, les systèmes d'aide au diagnostic implémentés jusqu'à présent n'ayant augmenté la performance des médecins que de 2.5% !

En recherche biologique, à l'ère des mégadonnées, ces observations et débats posent âprement la question de savoir : si une machine peut effectuer, seule, les tâches complexes d'analyse de données biologiques pour en retirer de l'information, de la connaissance ou de la sagesse, si un humain doit/peut le faire seul ou si une association entre expert humain et machine est possible et souhaitable.

C'est à ces questions fondamentales pour l'avenir de la recherche, que les travaux effectués durant ma thèse ont essayé d'apporter une modeste contribution et ce, dans une perspective qui intègre à la fois i) les limites humaines à l'assimilation de quantités astronomiques d'informations biologiques en perpétuelle croissance et ii) les limites des machines et leur capacité effective à extraire '*ex-nihilo'* de la connaissance des mégadonnées.

Ces années ont forgé ma conviction profonde que ni l'expert humain isolé, ni l'ordinateur seul ne peuvent exploiter efficacement les mégadonnées. Il en résulte qu'à mes yeux, il appartient en priorité aux bioinformaticiens de faciliter par tous les moyens l'échange entre ces deux acteurs incontournables en s'inspirant au mieux des évolutions informatiques récentes en matière de réseaux sociaux, évolutions qui transforment drastiquement nos modes de communications et notre conception du monde.

Matériels et méthodes

Durant ma thèse, je me suis appuyé sur un certain nombre de ressources, pour extraire des données, effectuer des traitements et créer des infrastructures web.

Je présenterai dans un premier temps les ressources bio-informatiques choisies comme sources de données dans le cadre de mes développements, et les traitements que j'ai appliqués à ces données. Puis je présenterai les structures logicielles (*frameworks*) et librairies qui m'ont permis de créer ces infrastructures. Enfin, je parlerai des outils bio-informatiques que j'ai soit intégrés dans mes projets soit utilisés pour prétraiter les données.

Par souci de clarté, certains développements originaux ou adaptations d'algorithmes que j'ai réalisés durant ma thèse seront décrits plus en détail dans les chapitres connexes de la section Résultats et Discussion.

1 Ressources bio-informatiques et traitements appliqués

Pour construire et mettre à jour la base de données de MyGeneFriends et de Parsec, je me suis servi d'un ensemble de banques de référence, complétées par des banques spécialisées dans les transcrits (expression, rôle) ou les maladies. Les données relatives aux maladies (ainsi que les phénotypes et variations) et les données textuelles sous forme de publications ont permis de fournir des informations supplémentaires.

1.1 Banques de référence

Les sites de référence comme Ensembl (Cunningham et al., 2015), NCBI (*National Center for Biotechnology Information*), le site *Genome Bioinformatics* de l'UCSC (*University of California Santa Cruz*) ou UniProt (Magrane and Consortium, 2011), regroupent les données généralistes concernant les séquences, génomes et protéomes ainsi que leur annotations et ce, pour tous les organismes connus. Ce sont des ressources incontournables en biologie et plus encore en bio-informatique.

1.1.1 Ensembl

Ensembl (Cunningham et al., 2015) est une ressource web permettant l'accès aux informations génomiques. C'est une des ressources majeures en biologie qui génère et agglomère un grand nombre de données sur beaucoup d'organismes. Les informations d'Ensembl sont accessibles directement en web sur le site (http://www.ensembl.org), mais sont également téléchargeables sur un site FTP, et accessibles dans une base de données publique.

Chaque nouvelle version d'Ensembl pour un organisme donné est accessible dans une base de données dédiée. Ainsi *homo_sapiens_core_78_38* contient des informations de base de la version 38 (h38) du génome humain couplée à la version 78 des annotations. Dans le cadre de MyGeneFriends, la liste de gènes et de transcrits humains, les relations entre transcrits et *probeset* et les données de *Gene Ontology* (décrites plus en détail plus bas) ont ainsi été récupérées de la base de données de Ensembl.

1.1.2 NCBI Gene

NCBI Gene est un portail du NCBI (*National Center for Biotechnology Information*) intégrant diverses informations relatives aux gènes de différentes espèces.

J'utilise ce portail dans MyGeneFriends afin de récupérer :

- le fichier de *stopwords* (liste de mots couramment utilisés dans une langue et ne présentant pas d'intérêt lors de la fouille de données),
- le fichier *gene2pubmed* recalculé chaque jour et contenant les relations entre les gènes et les publications dans lesquelles ils apparaissent,
- le fichier *gene2ensembl* que j'utilise pour établir les correspondances entre identifiants NCBI et identifiants Ensembl,
- le fichier *geneinfo* que j'utilise pour faire correspondre les identifiants NCBI avec les symboles de gènes.

1.1.3 UCSC

Le site *Genome Bioinformatics* de l'UCSC (*University of California, Santa Cruz*) agglomère des informations génétiques, génomiques, épi-génomiques… provenant de multiples sites (Ensembl, NCBI, Sanger Institute, ENCODE...). Ces données sont visualisables dans un *Genome Browser,* mais sont aussi accessibles en se connectant à la base de données mysql de l'UCSC ou en téléchargeant des fichiers par ftp. Un atout indéniable est la possibilité d'ajouter dans l'URL (*Uniform Resource Locator*), des instructions pour l'affichage dans le *genome browser,* de données personnelles en parallèle des données standard.

L'UCSC (Dreszer et al., 2012) propose les génomes masqués et non masqués. Les génomes masqués sont filtrés par RepeatMasker (Smit et al., 2008) qui identifie les régions de faible complexité et les éléments répétés simples et dispersés (transposons, rétrotransposons…). Ces régions sont remplacées par des minuscules ou des N. L'utilisation d'un génome masqué est particulièrement importante lors de la recherche de sites de fixation de facteurs de transcription. Cela permet de limiter la recherche à la partie non répétée du génome et de diminuer ainsi considérablement le nombre de faux positifs liés à la petite taille (à partir de 6 pb) de ces signaux génomiques majeurs.

L'UCSC possède également plusieurs tables répertoriant les gènes d'un organisme donné. Ces tables proviennent de sources différentes (ensGene d'Ensembl, refGene du NCBI, etc), et se distinguent par leur contenu en organismes et en types de gènes.

J'ai utilisé UCSC afin de récupérer des séquences génomiques, des alignements et des gènes dans le cadre du projet Parsec, et des descriptions RefSeq de gènes (accessibles simplement dans une table dédiée) dans le cas du projet MyGeneFriends.

1.1.4 UniProt

UniProt (*Universal Protein ressource*) (Magrane and Consortium, 2011) est une ressource très complète permettant d'accéder aux séquences et annotations de protéines. Elle est composée des banques UniProtKB (*UniProt Knowledgebase*) qui est la banque principale proposant les

séquences et annotations des protéines, UniRef (*UniProt Reference clusters*) qui regroupe les protéines ayant 100%, 90% ou 50% d'identité de séquence en commun, UniMES *(UniProt Metagenomic and Environmental Sequences*) spécialisée dans les données méta-génomiques et UniParc (*UniProt Archive*) qui contient la plupart des séquences protéiques disponibles dans le monde.

La banque UniProtKB est composée de deux banques. La banque SwissProt comporte uniquement les annotations manuelles de haute qualité. La banque TrEMBL fut créée en réponse au flux croissant de données provenant de projets de génomique (Bairoch and Apweiler, 2000). Il fut alors évident que la curation manuelle ne pouvait pas suivre le rythme de nouvelles séquences protéiques, qui furent alors classifiées et annotées automatiquement pour être stockées dans la banque TrEMBL.

Au cours de ma thèse, je me suis servi des séquences SwissProt dans le protocole de détermination des transcrits canoniques (chapitre 3.3.1.2), et dans la construction du fichier de correspondance entre identifiants UniProtKB et identifiants Ensembl.

1.2 Transcriptome et protéome

Les sources de données présentées dans ce chapitre permettent d'apporter des annotations supplémentaires aux gènes et aux relations entre gènes.

1.2.1 GEO

GEO (Barrett et al., 2013) ou *Gene Expression Omnibus* est une base de données gratuite regroupant des données d'expression provenant d'expériences de puces à ADN (*microarray)* à haut débit ou de séquençage de nouvelle génération. Elle comporte à ce jour des dizaines de TeraOctets de données. Nous utilisons GEO comme source de données d'expression publiques. Les données d'intérêt sont récupérées par notre équipe afin de retraiter et normaliser les intensités de signal par des méthodes statistiques existantes ou développées au sein du LBGI.

J'intègre dans MyGeneFriends les données de l'expérience HGA (*Human Genome Atlas*) (Su et al., 2004) présentant l'expression normale des gènes dans 79 tissus humains.

Afin de permettre un affichage en carte de chaleur (*heatmap*) de l'expression d'un gène dans différents tissus, je commence par effectuer une transformation logarithmique pour une distribution plus uniforme des intensités I de signal et obtenir la valeur I_L :

$$I_L = \log(I)$$

Puis, j'effectue une transformation min-max pour replacer les valeurs dans un intervalle [0;1] :

$$I_{LR} = \frac{I_L - \min(I_L)}{\max(I_L) - \min(I_L)}$$

Enfin, j'associe à cette intensité une couleur HSL (*Hue, Saturation, Luminosity*) entre le bleu (peu exprimé) et le rouge (très exprimé) afin d'attribuer une couleur à chaque tissu en fonction de l'expression du gène considéré. Pour ce faire, j'utilise les formules données par

(http://stackoverflow.com/questions/12875486/what-is-the-algorithm-to-create-colors-for-a-heatmap) :

$$H = round(\,(1 - I_{LR}) * 100\,)$$

$$S = 100$$

$$L = round(\,I_{LR} * 50\,)$$

1.2.2 Gene Ontology (GO)

Le consortium à l'origine de l'ontologie GO (*Gene Ontology*) a eu pour but de produire un vocabulaire contrôlé susceptible de s'appliquer à tous les organismes (Ashburner et al., 2000) et permettant de créer des annotations pour décrire les rôles biologiques des produits du génome (Gene Ontology, 2015). Cette ontologie est constituée de trois sous-ontologies indépendantes (Tableau 3).

Tableau 3: Sous-ontologies créées et maintenues par le consortium Gene Ontology

Nom	Description	Exemple
Biological process	Un processus biologique est une série d'actions accomplies par des assemblages de molécules biologiques	Transduction de signal
Molecular function	Activités au niveau moléculaire	Activité catalytique
Cellular component	Un composant d'une cellule, partie d'un ensemble plus grand	Noyau cellulaire

Une annotation GO décrit l'association entre un terme de l'ontologie et le produit d'un gène, en précisant les types de sources (expérience, prédiction, fouille de textes, etc…) à l'origine de cette association. GO est continuellement en développement. Initié autour de trois organismes modèles (levure, drosophile et souris) (Blake, 2004) et 18000 termes, l'ontologie comporte désormais plus

de 40000 termes (Gene Ontology, 2015). Des projets permettant de mieux modéliser le cycle cellulaire ou les processus multi-organismes (comme l'interaction entre un virus et son hôte par exemple) sont en cours. En plus de gérer l'ontologie et d'annoter les produits du génome avec cette ontologie, le consortium GO met en place un certain nombre d'outils, comme AmiGO permettant de rechercher et visualiser les termes GO dans l'arbre de l'ontologie.

J'ai utilisé l'ontologie GO dans le cadre de Parsec (voir chapitre 1.2.3) et de MyGeneFriends (voir chapitre 3.2) afin d'annoter les gènes. MyGeneFriends utilise également cette ontologie pour établir des liens entre les gènes, et ce de deux manières. Le premier type de lien consiste à établir un score entre deux gènes en se basant sur le nombre de termes GO qu'ils ont en commun, c'est la mesure la plus simple. Un autre moyen d'évaluer le lien entre deux gènes en profitant des informations fournies par GO est de calculer la FSS (*Functional Semantic Similarity*) décrite par (Reyes-Palomares et al., 2013) et qui consiste en trois phases :

1. Calcul, pour chaque terme GO, de son *information content* (IC) qui est une mesure de sa spécificité :

   $IC(c) = -logP(c)$ (avec 'c' le concept évalué)

   Ce qui dans le cas de termes GO correspond à :

   $$IC(term) = -\log \frac{nombre\ d'occurences\ du\ terme\ et\ de\ ses\ descendants}{nombre\ d'occurences\ de\ tous\ les\ termes}$$

2. Calcul de la similarité entre deux termes, qui correspond à la spécificité de leur parent commun qui a la spécificité maximale, ou le MICA (*most informative common ancestor*) :

   $$sim(t_1, t_2) = \max_{p \epsilon S(t1,t2)} IC(p)$$

3. Calcul de la similarité entre deux gènes en considérant la similarité entre les termes les décrivant pris deux à deux. Reyes_Palomares et collaborateurs (Reyes-Palomares et al., 2013) proposent d'utiliser la similarité maximale :

   $$sim(g1, g2) = \max sim(ti, tj)\ \ sachant\ que\ ti \epsilon g1\ et\ tj \epsilon g2$$

Une autre publication (Kohler et al., 2009) suggère d'utiliser la moyenne.

J'ai choisi d'utiliser le maximum car il est insensible à l'absence de valeurs de MICA trop faibles (et donc termes très peu similaires), ce qui permet de ne pas avoir à générer tous les couples terme-terme et de définir un seuil qui donne le meilleur rapport entre le temps de calcul nécessaire et le nombre de relations entre gènes.

1.2.3 STRING

La connaissance des partenaires d'interaction d'une protéine est une source précieuse d'informations pour définir sa fonction, replacer sa structure dans le contexte d'un complexe ou replacer son rôle dans le contexte d'une voie métabolique ou d'un module (Szklarczyk et al., 2015).

Dans ce but a été créé une base de données de référence des associations fonctionnelles entre protéines : STRING (Szklarczyk et al., 2015) (*Search Tool for the Retrieval of Interacting Genes/Proteins*). STRING intègre des données d'interaction entre protéines de nombreuses sources et espèces :

- Données expérimentales (biochimiques, biophysiques, techniques de génétique),
- Données sur les voies métaboliques,
- Fouille de données automatique sur les résumés et articles complets provenant de Medline,
- Interactions prédites,
- Interactions observées dans un organisme et transférées vers un autre par des relations d'orthologie.

A chaque interaction est associé un score global (entre 0 et 1), lui-même calculé à partir de scores associés à chaque type d'information. Dans un but de simplification, j'intègre dans MyGeneFriends uniquement le score global d'association. De plus, j'utilise le seuil de scores de 0.7 (proposé par défaut par STRING), afin de considérer que deux gènes sont bien reliés.

1.3 Données relatives aux maladies

Les données relatives aux maladies m'ont permis d'intégrer et fusionner des définitions de maladies et d'annoter ces maladies par des informations phénotypiques et les variations génétiques qui les causent.

1.3.1 OMIM

OMIM est une base de données de gènes humains et de maladies génétiques humaines. Elle est utilisée dans MyGeneFriends pour incorporer des informations sur les maladies car elle possède des fiches très détaillées avec notamment, des descriptions des maladies.

OMIM ne propose pas de sections séparées pour les maladies et les gènes ; elle propose plusieurs types d'entrées, spécifiés par un symbole devant chaque nom de fiche (Tableau 4).

Tableau 4: Types d'entrées disponibles dans OMIM avec sur fond bleu les entrées intégrées dans MyGeneFriends

Symbol	Description	Nombre
*	Un gène	14 985
#	Une entrée descriptive, généralement une maladie, ne représente pas un locus unique sur le chromosome	4 518
+	Gène et maladie combinés	86
%	Phénotype/maladie dont la base moléculaire n'est pas connue	1 653
	Phénotype/maladie dont la base mendélienne n'a pas été clairement établie	1 821
^	L'entrée a été supprimée ou déplacée vers une autre	

Le contenu de la banque OMIM est gratuitement accessible sous forme de fichier plat (peu structuré) qu'il faut traiter soi-même pour extraire des informations structurées.

1.3.2 Orphanet

Orphanet (Ayme, 2003) est une base de données spécialisée dans les maladies rares humaines. Elle regroupe les informations concernant les maladies (Tableau 5), les gènes qui y sont reliés, les médicaments, des articles scientifiques… Les fiches de maladies sont souvent mises en correspondance avec des équivalences dans d'autres bases de données comme OMIM, MeSH, UMLS, etc…

Tableau 5: Contenu d'Orphanet en Décembre 2014 (selon le dernier rapport d'activité)

Caractéristique	Nombre de fiches
Maladies avec au moins un lien vers OMIM	4229
Maladies liées à des gènes	3312
Maladies avec descriptions textuelles en anglais	3969
Maladies avec signes cliniques	2689

En plus d'un accès par site web (http://www.orpha.net), des informations sont disponibles gratuitement sous forme de fichiers XML à télécharger sur http://www.orphadata.org. Ces fiches ne permettent pas d'accéder à toutes les informations du site, les descriptions des maladies ne sont par exemple pas téléchargeables.

J'ai choisi de compléter la liste de maladies récupérées dans OMIM par celles disponibles sur Orphanet, d'une part pour avoir une liste de maladies plus exhaustive, et d'autre part pour nuancer la définition des maladies d'OMIM.

1.3.3 HPO

Human Phenotype Ontology (HPO) est une ontologie de plus de dix mille phénotypes (termes) structurés hiérarchiquement et décrivant des maladies (Kohler et al., 2014). C'est un effort allant dans le sens d'une plus grande structuration et moindre subjectivité dans la description des maladies. L'annotation des maladies avec des phénotypes pour OMIM par exemple provient à la fois de l'annotation manuelle par les équipes de HPO et de la détection automatique des termes HPO dans le « *Clinical Synopsis* » des entrées OMIM.

Les relations entre phénotypes et maladies proviennent de sources de nature diverse, décrites par des « *evidence code* », présenté dans le Tableau 6 du moins fiable au plus fiable. Afin d'annoter le plus de gènes, les méthodes expérimentales sont en effet souvent complétées par des méthodes prédictives, basées sur la génomique comparative, ou basées sur la fouille de textes.

Tableau 6: Sources des relations phénotypes-maladies dans HPO.

Code	Description
ITM	Provenant de la fouille de textes
IEA	Annotation automatique
PCS	Etude clinique publiée
ICE	Expérience clinique individuelle
TAS	Affirmation traçable d'un auteur

Malgré les efforts de structuration, certaines informations sont toujours présentes de manière hétérogène comme la fréquence d'un phénotype dans une maladie qui peut être décrite qualitativement (« *rare* », « *frequent* », « *very rare* »), par un pourcentage (30%) ou par le ratio d'un échantillon (*5 of 16*).

Les informations HPO sont souvent mises à jour et accessibles sur le serveur public Hudson (http://compbio.charite.de/hudson/).

J'intègre les informations de HPO afin d'annoter les maladies et d'établir des liens entre maladies possédant des phénotypes en commun. De plus, j'utilise un fichier construit par HPO afin de récupérer les liens entre gènes et maladies car ces relations sont plus à jour que celles proposées par Ensembl.

1.3.4 Clinvar

ClinVar (Landrum et al., 2014) est une base de données contenant des variations reportées comme ayant une valeur médicale. Une entrée ClinVar SCV (*Submitted ClinVar*) est composée

d'une variation, d'une maladie et de la source de provenance des données. ClinVar réunit les entrées ayant le même couple variation/maladie pour crées des entrées RCV (*Reference ClinVar*). ClinVar accepte uniquement les variations identifiées par tests cliniques, ou décrits dans la littérature. A chaque variation est associé un identifiant provenant des banques dbSNP (base de données très exhaustive de petites variations génomiques) (Sherry et al., 2001) ou dbVar (base de données de variations structurelles) (Lappalainen et al., 2013).

Les variations sont caractérisées par leur conséquence clinique avec la notion de « *Clinical significance* » (Landrum et al., 2014). Ainsi, *Pathogenic* et *probably pathogenic* signifient que la variation va être liée à une maladie. *Drug-response* indique que la variation impactera la réponse des individus à certains médicaments. Cela peut s'exprimer par la variabilité du dosage nécessaire, l'absence de réponse, ou la toxicité (Pereira et al., 2015). *Histocompatibility* signifie que la variation va influencer la compatibilité entre donneur et receveur dans le cas d'une greffe. Enfin, *benign* et *propably benign* permettent d'annoter le variant comme ne provoquant pas de maladie.

Les données de ClinVar peuvent être téléchargées sur le site FTP sous plusieurs formats, dont le format VCF (format officiel sur les données de variations) que je traite. Le format de fichier VCF (*Variant Call Format*) (Danecek et al., 2011) est un format standardisé permettant de stocker et communiquer les types de variations les plus courantes comme les SNP, indels (insertions et délétions), variations structurales, etc… et de les annoter. Les premières lignes d'un fichier VCF constituent l'en-tête et décrivent les différents champs des fichiers. Une ligne d'un fichier décrit une variation à un endroit spécifique du génome et est composée de 8 champs obligatoires (Tableau 7).

Tableau 7 : Les champs obligatoires d'un fichier VCF.

Nom	Description	Exemple
CHROM	Le chromosome	1
POS	La position sur le chromosome du début d'un variant. La première position sur le chromosome est 1.	12304579
ID	L'identifiant unique du variant. C'est généralement un identifiant dbSNP	rs267597962
REF	L'allèle de référence	A
ALT	Une liste d'allèles alternatifs	T,C
QUAL	Qualité de l'échantillon	3
FILTER	Si l'échantillon a passé tous les filtres, ou liste de filtres qu'il n'a pas passés	PASS
INFO	Annotation additionnelle, dépendante du système qui a généré le fichier	

La banque de variations ClinVar propose beaucoup d'informations importantes complémentaires dans le champ INFO du fichier VCF telles que le gène concerné, les maladies reliées aux allèles alternatifs, etc… J'extrais de ce champ les associations entre allèles et maladies

1.4 Données de publications

PubMed est une base de publications biomédicales maintenue par le NCBI et regroupant plus de 25 millions d'articles en anglais. Certains de ces articles sont accessibles gratuitement dans leur intégralité, mais une grande majorité n'est présente que sous forme de résumé.

Les articles Pubmed peuvent être recherchés sur le site web ou récupérés à travers une API très complète. Les deux principales composantes de cette API sont les services E-Search et E-Fetch. Le service E-Search permet d'effectuer une recherche avec des mots clefs, et renvoie une liste d'identifiants correspondant aux articles sélectionnés. Cette liste d'identifiants peut ensuite servir à récupérer les documents associés avec le service EFetch, dans le format spécifié. On fournit pour cela une liste d'identifiants, la base de données à interroger ('pubmed' pour les résumés et 'pmc' pour les publications entières) et le format dans lequel on désire recevoir le résultat (XML par exemple).

J'utilise ces deux services afin de récupérer les publications liées à des gènes de MyGeneFriends, mais également pour récupérer les publications écrites par les utilisateurs de MyGeneFriends.

2 Ressources informatiques

L'architecture d'une application web comprend l'interconnexion d'un certain nombre de composants qu'on pourrait représenter sous forme d'un flux allant des donnés au client, avec à chaque étape un ensemble d'outils et de méthodes nécessaires à l'accomplissement des tâches spécifiques (Figure 13). Nous allons présenter ces différentes étapes.

Figure 13: Le flux de données entre les sources et l'utilisateur du site web peut être découpé en plusieurs grandes étapes.

2.1 Intégration de données

L'intégration de données comprend les outils de récupération et de traitement, mais également, ceux de pilotage de la machinerie d'intégration dans la base de données.

2.1.1 Jenkins

Jenkins est un outil d'intégration continue *open source* écrit en Java (Smart, 2011) avec de nombreux *plugins* disponibles. Il est utilisé notamment par HPO (Kohler et al., 2014) et GO (Gene Ontology, 2015) pour mettre à jour les ontologies et rendre publiquement disponible chaque version (*build*).

Jenkins permet de lancer des tests sur le code soumis à un système de gestion de versions, lancer la construction et compilation de gros projets, etc… En d'autres termes, Jenkins exécute une liste prédéfinie d'étapes suivant un élément déclencheur (*trigger*) (par exemple, une heure fixe). Des nœuds peuvent être créés sur plusieurs serveurs afin de distribuer la charge et des scripts peuvent être lancés. Un espace de travail associé à chaque projet ou chaque *build* permet le stockage des résultats intermédiaires ou finaux produits par différentes étapes.

Il est utilisé au sein de notre laboratoire pour des tâches diverses comme l'intégration continue (exécution de tests sur le code de développement et passage de ce code en production si tous les tests sont concluants), le téléchargement de ressources ou le pilotage de scripts.

2.1.2 NLTK

NLTK (*Natural Language ToolKit*) (Bird, 2006) fait partie des librairies les plus utilisées pour la fouille de données textuelles. C'est une librairie écrite en python qui propose des outils pour la classification, *tokenization*, racinisation, identification de la catégorie grammaticale des mots, analyse syntaxique, etc… des textes écrits en langues multiples. Cette librairie est bien documentée et facilement utilisable pour traiter des textes en langage naturel.

Parmi les fonctionnalités les plus intéressantes de NLTK, on citera :

- la décomposition d'un texte en phrases et en mots,
- l'élimination des mots peu informatifs,
- la racinisation des mots.

J'utilise cette librairie pour l'extraction de mots clefs caractérisant les acteurs de MyGeneFriends.

2.1.3 Gensim

Gensim (Řehůřek and Sojka, 2011) est un *framework* python de traitement du langage naturel simple à prendre en main et capable de traiter de larges quantités de données. Il implémente un certain nombre d'algorithmes populaires comme TF-IDF, LSA, LDA, etc… Je me suis principalement servi de TF-IDF qui est une solution toujours très utilisée pour attribuer un score à un mot par rapport à un document dans le cadre de l'intégration de mots clefs dans MyGeneFriends.

2.2 Bases de données

2.2.1 PostgreSQL

J'ai choisi Postgres comme gestionnaire de base de données pour Parsec et MyGeneFriends pour plusieurs raisons. Postgres, libre et *open source*, est historiquement connu pour son respect des contraintes d'intégrité (unicité par exemple, ou le fait qu'une colonne ou un ensemble de colonnes doivent contenir des valeurs uniques). Postgres permet également de traiter du JSON (*JavaScript Object Notation*). JSON est un encodage en format texte de données structurées très utilisé pour les échanges entre processus distants écrits en langages souvent différents. Ainsi Postgres possède un type de champ « JSON » sur lequel il est possible de stipuler des conditions lors d'une requête dont le résultat sera aussi sous forme JSON. Par ailleurs l'affichage du plan d'exécution d'une requête est très clair, ce qui facilite les optimisations de la requête et le choix des index. Enfin, Postgres permet une bonne évolutivité horizontale (basée sur l'ajout de machines supplémentaires à l'infrastructure, à la différence de l'évolutivité verticale qui consiste

à ajouter plus de puissance à la machine déjà présente) grâce à des mécanismes de réplication avec des serveurs maître et esclaves et une répartition de charge en lecture.

2.2.2 ElasticSearch

ElasticSearch (www.elastic.co) est un gestionnaire de base de données NoSQL et possède un moteur de recherche principalement utilisé pour indexer les données textuelles. Il est utilisé par des grands acteurs comme Wikipedia, The Guardian, StackOverflow, GitHub…

ElasticSearch est basé sur Apache Lucene (lucene.apache.org/core/), un moteur de recherche textuel puissant mais peu pratique à prendre en main. Afin d'avoir des temps de réponse très courts et une bonne évolutivité horizontale, ElasticSearch utilise une architecture parallèle et redondante. Ainsi, ce qu'on appelle, un groupe (*cluster*) ElasticSearch est composé de plusieurs nœuds, chaque nœud étant une instance d'ElasticSearch qui est généralement exécuté sur une machine séparée. Un nœud peut contenir des *shards* primaires (une partie de l'index) et des *replica* de *shards* primaires (des copies). Cette architecture déployant l'information en plusieurs copies sur des machines différentes, assure la tolérance aux pannes et a la capacité de servir beaucoup de requêtes parallèles.

L'organisation de l'information dans ElasticSearch a quelques similitudes avec les bases de données SQL, et certains termes bien que différents représentent la même chose, comme l'indique le Tableau 8.

Tableau 8: Comparaison entre l'organisation de l'information dans ElasticSearch et dans une base de données SQL

Notion ElasticSearch	Equivalent SQL
Index	Database
Type	Table
Document	Row

Dans ElasticSearch, la structure de données d'une entrée (Document) n'a pas de schéma prédéfini ou de colonnes et correspond à un document JSON.

ElasticSearch dispose d'une API REST (*Representational State Transfer*) puissante qui permet de manipuler ces données avec les requêtes permettant les opérations CRUD (*Create, Read, Update, Delete*) (Tableau 9) et les résultats sont présentés sous forme de JSON.

Tableau 9: Les fonctionnalités CRUD accessibles par l'API d'ElasticSearch

Méthode HTTP	Fonction
GET	Faire une recherche et récupérer des documents
PUT	Ajouter un nouveau document
POST	Modifier un document existant
DELETE	Effacer un document d'identifiant donné

Parmi les nombreux modes de recherche disponibles (*more like this*, *fuzzy like this*, *term*, *text*, *wildcard*, *range*, etc…), on peut citer *QueryString* qui permet d'utiliser la syntaxe Lucene afin de construire des requêtes puissantes en utilisant des symboles spéciaux présentés dans le Tableau 10.

Tableau 10: Symboles permettant de construire une requête de recherche complexe, qui sera traitée par Lucene

Symbole	Signification	Exemple
:	Permet de faire la recherche dans un champ spécifique	title:java AND python
+	Le terme après le + doit obligatoirement être présent dans le document	java +python
OR	Opération OU, les deux termes **peuvent** être présents	java OR python
AND	Opération ET, les deux termes **doivent** être présents	java AND python
-	Exclut les documents qui contiennent le terme après le -	java -tcl
()	Permettent de regrouper des termes	(java OR python) AND programming
^	Permet de « booster » ou d'augmenter le poids d'un mot	java^5
*	Replace 0 ou plusieurs caractères. Placé à la fin d'un mot, permet de considérer ce mot comme un préfixe	Progr*
?	Remplace un seul caractère	ja?a
~	Permet une recherche approximative du mot	python~

La chaine « java^5 python –tcl » permettra par exemple de chercher tous les documents qui contiennent les mots « java » et « python » mais ne contiennent pas le mot « tcl », en donnant un meilleur score aux documents qui contiennent le mot « java » (car on y a applique un *boost* de 5).

Une fois la recherche effectuée, les résultats retournés sont évalués par pertinence. ElasticSearch ne retourne par défaut que les dix premiers résultats, en précisant le nombre total de résultats

trouvés et en permettant de les récupérer en totalité. Pour chaque document retourné, on peut accéder aux détails du calcul de son score.

La synchronisation entre une base de données classique et ElasticSearch se fait par un module supplémentaire (*plugin)* de type « *river* ». Celui-ci est configuré avec les informations de connexion à la base de données, la requête à effectuer pour récupérer les données et une expression *cron* décrivant la périodicité à laquelle la synchronisation doit s'appliquer. Pour chaque ligne retournée par la requête, le *river* crée un document JSON qu'il ajoute à ElasticSearch.

En plus d'être interrogeable par REST, ElasticSearch peut être utilisé directement à partir du code d'un programme java grâce à une librairie dédiée. C'est cette librairie que j'utilise pour communiquer avec ElasticSearch dans MyGeneFriends.

2.3 Les ORM

Le principe des ORM (*Object Relational Mapping*) consiste à créer des classes dans un langage objet qui vont, chacune, représenter une table de la base de données d'intérêt. Cela signifie que chaque attribut de la classe correspondra par son nom et son type à une colonne de la table qu'il représente. Ainsi, lorsqu'on interroge une base de données par l'intermédiaire d'un ORM, on reçoit les résultats directement sous forme d'une liste ou d'une série d'objets. Les ORM permettent également d'éviter d'écrire les requêtes SQL directement dans le code en utilisant des fonctions prédéfinies.

2.3.1 EBean

EBean est un ORM écrit en Java permettant d'interagir avec une base de données relationnelle et de récupérer les résultats de requêtes sous forme d'objets Java. Il a été conçu pour être plus simple à utiliser que la JPA (*Java Persistance API*) et ne nécessite quasiment aucune configuration, mis à part les informations nécessaires pour se connecter à la base de données.

J'utilise EBean pour interroger la base de données à partir du serveur web MyGeneFriends.

2.3.2 Peewee

Peewee (github.com/coleifer/peewee) est un petit ORM python, que je trouve simple et intuitif, gérant des bases de données SQL (MySQL, Postgres, Sqlite). Il est beaucoup plus simple qu'un autre ORM python très connu, SQLAlchemy (http://www.sqlalchemy.org/), bien qu'il offre un peu moins de fonctionnalités.

J'utilise peewee pour intégrer les données dans la base de données MyGeneFriends et pour mettre à jour ces données.

2.4 *Frameworks* Web

Les *frameworks* web facilitent énormément le développement des sites web en se chargeant des mécanismes dits de 'bas niveau', et en proposant souvent un ensemble de librairie intégrées.

2.4.1 Apache Tomcat

Apache Tomcat est un conteneur de servlet. C'est un projet écrit en java, maintenu par *l'Apache Software Fondation*, et qui permet de déployer facilement des applications web. Le projet doit être déployé sous forme d'archive WAR (*Web application ARchive*), puis importé dans Tomcat.

2.4.2 Flask

Flask est un micro-*framework* web écrit en python, permettant de créer facilement et rapidement des applications web. Il est utilisé par des acteurs comme Pinterest ou LinkedIn. Sa nature de micro-*framework* signifie qu'il est très léger, et qu'il n'embarque aucun outil supplémentaire. Pouvoir gérer une base de données implique, par exemple, d'ajouter un *plugin* intégrant Peewee. Cette modularité permet de construire le *framework* selon ses besoins. Flask permet de créer un site web en cinq lignes de code python et en cinq minutes.

J'utilise Flask pour créer un serveur API permettant au serveur de MyGeneFriends codé en java d'accéder en couplage faible à des fonctionnalités codées en python, mais également pour exécuter des programmes sur le serveur.

2.4.3 Play 2 Framework

Play Framework (www.playframework.com) est un *framework* web écrit en Scala (langage de programmation compatible avec Java), permettant de développer des services web en Java ou Scala. Les applications développées suivent l'architecture MVC (Modèle, Vue, Contrôleur), ce qui donne une colonne vertébrale au code et implique une organisation structurée (Figure 14). La configuration est réduite au minimum (le paradigme de Convention plutôt que configuration) et est concentrée dans un seul fichier : *application.conf*. C'est un *framework* robuste et stable, utilisé par de grands acteurs du web comme LinkedIn ou The Guardian.

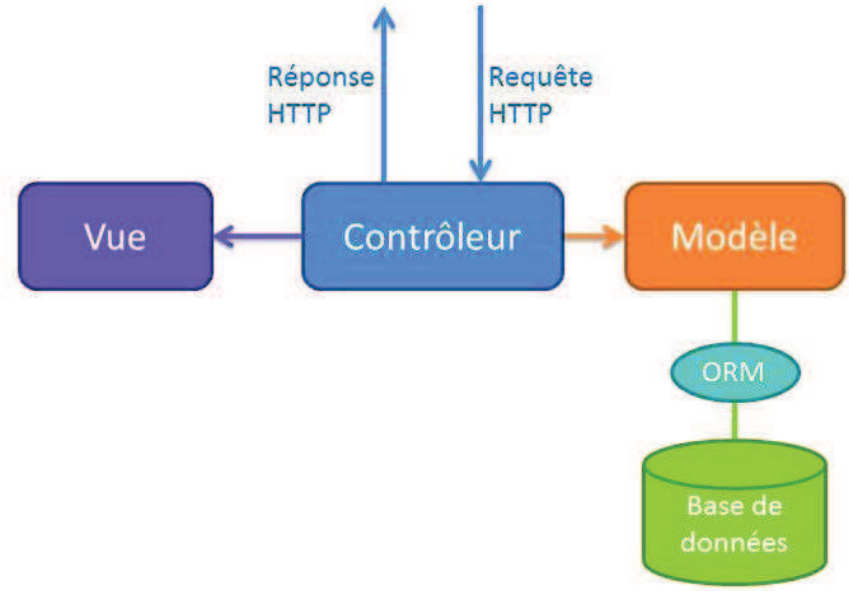


Figure 14: Le patron de conception (design pattern) MVC tel qu'il est implémenté dans le Play Framework. Lorsque l'utilisateur demande au serveur une page web, un contrôleur dédié à l'URL est saisi. Il effectue les modifications du Modèle si nécessaire et renvoie la réponse sous forme d'une Vue.

Play Framework met à disposition du développeur un certain nombre de librairies tierces profondément intégrées, permettant de gérer une base de données (EBean), du JSON (Jackson) et de compiler à la volée du CoffeeScript.

2.4.3.1 Jackson

Jackson (github.com/FasterXML/jackson) est une librairie de traitement du JSON, permettant de parser du JSON, mais aussi de générer le JSON à partir d'objets (les attributs et les fonctions sans arguments seront transformés en champs JSON). Les annotations disponibles permettent d'ignorer certains attributs pour éviter les boucles infinies dues aux références cycliques par exemple.

Cette librairie permet d'exposer très facilement des données sous forme de JSON dans le cadre d'une API.

2.4.3.2 CoffeeScript

CoffeeScript (coffeescript.org) est un langage de programmation qui est compilé en JavaScript (MacCaw, 2012). Sa syntaxe ressemble à celle du python (blocs définis par des tabulations et non

par des «} » par exemple). J'ai décidé d'utiliser CoffeeScript au lieu de programmer directement en JavaScript car CoffeeScript possède plusieurs avantages :

- il est plus succinct (moins de code à écrire pour réaliser la même tâche, la taille du programme peut être réduite de moitié),
- les mécanismes de classe et d'héritage sont simplifiés.

De plus, CoffeeScript est compilé en JavaScript avant d'être exécuté par le navigateur, et il est entièrement compatible avec les librairies JavaScript.

2.5 Visualisation des données

2.5.1 vis.js

Cette librairie JavaScript (visjs.org) de visualisation de données est simple à prendre en main et capable de gérer de grands volumes de données. De plus, elle permet d'interagir avec les données et offre plusieurs types de visualisations :

- Network, un graphe dynamique affichant des nœuds et des liens entre ces nœuds,
- Timeline, la visualisation d'évènements au cours du temps,
- Graph2d, représentations standards comme des barres, des camemberts, etc…
- Graph3d : représentation des données en trois dimensions.

La visualisation *Network* que j'utilise dans MyGeneFriends simule en temps réel les forces de répulsion entre les nœuds, ce qui permet aux acteurs symbolisés par ces nœuds de former naturellement des groupes en temps réel. Plusieurs modèles de simulation de la physique de répulsion des nœuds sont disponibles :

- *Barnes Hut* : modèle de simulation à n corps très rapide, utilisé notamment pour simuler la collision de deux galaxies.
- *Force Atlas 2* : modèle utilisé notamment dans Gephi, une plateforme interactive d'exploration de réseaux.
- *Repulsion* : modèle très simple basée sur l'idée d'un champ de répulsion autour des nœuds.
- *Hierarchical Repulsion*

Après avoir testé les différents modèles de simulation physique, j'ai sélectionné le modèle Barnes Hut car il favorise le mieux la formation de groupes de nœuds. Ce modèle peut être configuré avec les options décrites dans le Tableau 11. J'ai diminué la valeur de la constante de gravitation (qui est -2000 par défaut) à -40000 pour une stabilisation plus rapide du réseau.

Tableau 11: Principales options de la simulation Barnes Hut dans Vis.js

Option	Description	Valeur
gravitationalConstant	C'est presque l'option la plus importante de la simulation car elle représente la force de la gravitation. Pour que les nœuds ne s'attirent pas mais se repoussent, cette valeur est négative.	-40000
centralGravity	Gravité qui attire les nœuds vers le centre du réseau	0.3 par défaut
springLength	Longueur des arrêtes qui connectent les nœuds	95 par défaut
springConstant	Elasticité des arrêtes	0.04 par défaut

2.5.2 Highcharts

Highcharts (www.highcharts.com) est une autre librairie javascript de visualisation de données très utilisée, orientée vers les graphes 2D et 3D, radars, *heatmaps*, etc… Je l'utilise afin de produire des *Timelines* interactives de la popularité des acteurs de MyGeneFriends. Cette librairie permet à l'utilisateur de sauvegarder facilement les graphiques produits en image ou PDF, choisir les séries de données à afficher ou masquer, etc…

2.6 Autres

2.6.1 YouTrack

YouTrack (www.jetbrains.com/youtrack) est un système de gestion de bugs développé par JetBrains qui propose une licence d'utilisation gratuite jusqu'à dix utilisateurs. Il dispose d'une riche API, suffisamment complète pour permettre une utilisation en API presque aussi complète qu'en passant par l'interface graphique.

MyGeneFriends envoie à YouTrack par le biais de cette API les bugs signalés par les utilisateurs, mais aussi toutes les exceptions rencontrées pendant son fonctionnement.

2.6.2 JNI

La JNI ou *Java Native Interface* est une technologie Java permettant de faire communiquer du code Java avec une librairie écrite en C ou C++. Les deux usages et avantages principaux de cette technologie sont l'amélioration des performances en exportant le code crucial ou très utilisé dans la librairie C/C++ et la possibilité de réutiliser ou intégrer dans un programme Java du code C++ existant, ce sera mon cas dans le projet Parsec.

Un fichier d'entêtes .h généré à partir de méthodes natives déclarées dans le code java permet de faire communiquer la partie Java et native du programme.

Notons enfin que la JNI complexifie la portabilité du programme Java résultant, puisque la librairie native devra être recompilée pour chaque système (Windows, Linux…).

3 Outils bioinformatiques

3.1.1 Vep

VEP (*Variant Effect Predictor*) (McLaren et al., 2010) est un outil développé par Ensembl permettant de prédire les effets des variations génomiques sur les transcrits et protéines. Cet outil est accessible en API sur le site de Ensembl ou téléchargeable sous forme de scripts perl. VEP accepte en entrée un fichier tabulé de positions génomiques ou bien un fichier au format VCF.

L'outil en ligne de commande accepte un grand nombre d'options et supporte l'ajout de *plugins*. Lors de l'installation de l'outil, il est possible de créer un cache avec des génomes et leurs annotations ce qui accélère énormément l'analyse.

Un de ses points forts est également l'option permettant de récupérer la sortie sous format JSON, ce qui permet de la traiter facilement. J'utilise VEP afin d'annoter les fichiers ClinVar et récupérer les transcrits Ensembl reliés aux différentes variations, mais également pour annoter les variations que l'utilisateur ajoute à sa liste de variations personnelles sur MyGeneFriends. Les options que j'utilise pour exécuter VEP sont présentées dans le Tableau 12.

Tableau 12: Les options de VEP utilisés par MyGeneFriends

Option	Description
-fork	Permet d'accélérer considérablement le traitement en le répartissant sur plusieurs *threads*.
-cache	Utilise les annotations et génomes en cache (sur le disque local) plutôt que de les récupérer en se connectant au site Ensembl
--sift	Prédit si une substitution d'acide aminé va affecter la fonction de la protéine, utilise la méthode SIFT (Ng and Henikoff, 2003)
--polyphen	Prédit si une substitution d'acide aminé va affecter la fonction de la protéine, utilisant la méthode Polyphen (Adzhubei et al., 2010)
--domains	Affiche les noms des domaines protéiques concernés
--JSON	Génère un fichier JSON comme résultat. Le fichier généré n'est pas une liste JSON valide, en fait chaque ligne de ce fichier doit être traitée comme un JSON indépendant.
--hgvs	Affiche la nomenclature HGVS (*Human Genome Variation Society*) de la variation sur le transcrit et la protéine
--assembly	Précise quelle version du génome en cache doit être utilisée

3.1.2 Clustalw

ClustalW (Thompson et al., 2002) est un programme d'alignement multiple basé sur l'algorithme d'alignement global progressif. Je l'ai utilisé afin de générer des alignements deux à deux entre les séquences protéiques d'un même transcrit appartenant à deux version différentes de la base de données Ensembl. Cet alignement peut être affiché par l'utilisateur ou intégré dans la base de données MyGeneFriends.

3.1.3 GoMiner

La *Gene Ontology* permet de replacer un gène dans le contexte du *processus*, de la *fonction* et de la *localisation cellulaire*. Mais lors de l'analyse de données haut débit (provenant des *microarray* par exemple), l'enjeu n'est plus de décrire un seul gène, mais d'extraire les points communs d'un ensemble de gènes. GoMiner est un programme écrit en Java (Zeeberg et al., 2003) qui permet de catégoriser avec des termes GO une liste de gènes, en calculant l'enrichissement en termes GO de ces gènes par rapport à une autre liste de gènes ou à l'ensemble des gènes de l'organisme. L'enrichissement fonctionnel d'un ensemble de gènes revient à chercher des caractéristiques fonctionnelles ou cellulaires les plus précises partagées par le plus grand nombre des gènes considérés.

La version téléchargeable de GoMiner présente plusieurs avantages. Son utilisation est assez simple, il supporte l'utilisation d'une base de données GO locale ce qui permet d'avoir une analyse rapide. Par ailleurs, il peut être exécuté sans qu'on lui fournisse une liste de gènes de référence (qu'il construit alors lui-même, en prenant un échantillon aléatoire dans le génome de l'espère considérée). Enfin GoMiner, parce qu'il est écrit en Java, s'intègre naturellement dans notre service web.

Le score utilisé pour cet enrichissement est le FDR (*False Discovery Rate*), qui permet de comparer l'enrichissement réalisé par rapport à ce qu'on attendrait d'un résultat aléatoire.

De plus, un filtrage par type de source d'association entre gène et terme GO (expérimental, similarité, alignement, annotation électronique, etc…) permet de trouver l'équilibre souhaité entre qualité et sensibilité.

3.1.4 DAVID

En complément de GoMiner, j'ai également utilisé DAVID (Huang et al., 2007). C'est un logiciel populaire (plus de 2000 citations) permettant de calculer l'enrichissement de certaines catégories fonctionnelles dans une liste de gènes, ainsi que d'établir des groupes, classés par scores. Il n'est

pas disponible sous forme de programme à installer localement je n'ai donc pas pu l'inclure directement dans mes applications web. Lorsque MyGeneFriend propose à l'utilisateur d'effectuer une analyse par DAVID d'une liste de gènes, il lui propose en fait un lien sur lequel il faudra cliquer.

Résultats et Discussion

Dans une vision synthétique des travaux qui seront décrits dans les chapitres suivants, on peut dire que ma thèse s'est clairement positionnée à l'interface entre chercheurs et données biologiques. En effet très tôt, j'ai été attiré par les défis et contraintes attachés aux développements visant à créer une connexion efficiente entre l'expert humain et les mégadonnées décrivant les acteurs de la biologie moderne, Cet intérêt se retrouve aussi bien dans les travaux liés à la caractérisation en temps réel de sites génomiques couplée à des scénarios d'analyse personnalisés (Parsec), que dans ceux effectués pour obtenir une visualisation intuitive des relations d'orthologie complexes entre organismes (développements ajoutés à la nouvelle version d'OrthoInspector) ou enfin, dans la création d'un système complet permettant l'intégration de l'humain dans les réseaux biologiques de connaissances liés aux maladies génétiques humaines (MyGene Friends).

Ce chemin m'a permis d'explorer différentes échelles de la biologie, depuis le nucléotide et les sites génomiques jusqu'aux relations entre organismes *via* les gènes, maladies et réseaux complexes que ces derniers constituent.

Il m'a également fallu faire un certain nombre de choix concernant aussi bienles technologies à utiliser ou à développer que les données à manipuler, la manière de les stocker, de les visualiser et d'interagir avec elles.

J'ai appliqué les infrastructures développées à diverses problématiques biologiques afin d'évaluer leur potentiel et découvrir de nouvelles pistes de recherche. Ces problématiques impliquaient soit les centres d'intérêts de notre laboratoire (ciliopathies, syndrome de Bardet-Biedl), soit des collaborations avec d'autres laboratoires (régulation à l'acide rétinoïque chez le poisson zèbre, filtrage de siRNA).

Une étape importante de ce voyage fut la création de Parsec qui, malgré la simplicité apparente des nucléotides au regard des systèmes biologiques dont ils sont un maillon, a révélé que la découverte, l'analyse et le caractérisation de ces sites nécessitent d'intégrer un grand nombre d'informations.

1 Parsec

L'objectif de Parsec est de replacer l'expert dans l'analyse de séquences génomiques et de lui permettre d'interagir rapidement (temps d'exécution) et efficacement (réduction du bruit, données pertinentes) avec les sites génomiques en les replaçant dans leur contexte évolutif et génétique.

En effet, dans le paysage actuel de la biologie à haut débit, la contextualisation et le filtrage sont particulièrement nécessaires dans l'étude de sites génomiques où il s'agit d'identifier de petites séquences nucléotidiques fonctionnelles parmi des milliards de bases. Ces sites, malgré leur taille réduite, peuvent jouer un rôle essentiel dans l'expression des gènes, l'épissage, la réplication ou réparation de l'ADN... Grâce à de grands projets comme ENCODE (*Encyclopedia of DNA Elements*) (Consortium et al., 2007), des milliers de sites ont été identifiés dans le génome humain et dans les génomes d'organismes modèles. S'il est crucial d'identifier et annoter ces sites pour la compréhension de processus biologiques majeurs et de maladies humaines, leur recherche *in silico* dans les génomes complets à partir d'un court motif dégénéré aboutit à de nombreux faux positifs.

Afin de séparer les hits biologiquement signifiants du bruit, j'ai créé PARSEC, une plateforme logicielle de caractérisation et d'analyse guidée de sites et motifs au sein des génomes eucaryotes complets. Elle permet d'identifier très rapidement des séquences nucléotidiques strictes ou dégénérées à l'échelle d'un génome en exploitant une structure de données très efficace, les arbres de suffixes compressés. Les sites potentiels détectés sont ensuite filtrés et/ou annotés par des modules supplémentaires permettant une caractérisation par la conservation phylogénétique (sites conservés dans une liste d'organismes sélectionnés), par le voisinage génétique (distance aux gènes et types de gènes) ou par la fonction biologique afin d'effectuer un filtrage fonctionnel.

Le module de recherche repose sur une structure de données en mémoire vive, alors que les modules de contextualisation ont nécessité l'intégration de plus de 200 GigaOctets de données dans une base de données dédiée.

Les modules utilisés dans n'importe quel ordre et de façon répétée permettent de créer des scénarii complexes pour filtrer des sites d'intérêt à partir d'une liste initiale et caractériser, par exemple, les gènes proches de sites conservés (scénario du type : recherche-conservation-contexte), les sites proches de gènes ayant les mêmes fonctions (scénario : recherche-conservation-contexte-fonction-filtre) ou créer des profils à la volée (scénario : recherche-contexte-fonction-filtre-conservation).

Le développement de Parsec s'inscrit dans une collaboration entre notre laboratoire et l'Equipe de Zoologie Moléculaire de l'Université de Lyon, afin de répondre à un besoin de compléter des résultats expérimentaux de RNA-Seq avec une approche bio-informatique.

Parsec a été publié dans Bioinformatics (Allot et al., 2013) (voir chapitre 1.4)

Après une présentation des différents modules de Parsec et des stratégies et techniques mises en œuvre pour aborder les problématiques biologiques ou algorithmiques associés, je présenterai quelques applications de Parsec. Enfin, je replacerai ce travail dans le contexte des outils concurrents et proposerai des pistes d'optimisation de Parsec, tout en discutant les choix effectués.

1.1 La recherche génomique

1.1.1 La chasse aux sites sur le génome

L'ADN fourmille de signaux divers et variés, définis par exemple par un enchaînement particulier de nucléotides. Ces signaux vont modifier le comportement cellulaire soit en étant reconnus directement par des partenaires biologiques comme les ARN ou les protéines, soit sous forme transcrite voire traduite (peptide signal par exemple). La détection et l'identification du sens biologique de ces signaux est un aspect qui intéresse depuis des années les chercheurs.

Certains signaux courts, particulièrement importants dans l'expression de l'ADN, sont appelés sites. On peut citer notamment les sites de jonction entre intron et exon, les sites de reconnaissance de l'ARN polymérase ou les sites de fixation de facteurs de transcription.

Les facteurs de transcription se lient à des positions précises de l'ADN grâce à leur domaine de fixation (*DNA Binding Domain*, DBD). Ces DBD (un exemple est présenté dans la Figure 15) présentent différentes structures tridimensionnelles, les plus connues incluent le doigt de zinc, l'hélice-boucle-hélice, la glissière à leucine (*leucine zipper*). Les facteurs de transcription recrutent souvent quantité d'autres facteurs comme des coactivateurs ou des corépresseurs de la transcription. Avec ces facteurs, ils vont réguler l'expression de gènes spécifiques en fonction, par exemple, du cycle cellulaire, du tissu, de signaux reçus, du stade de développement...

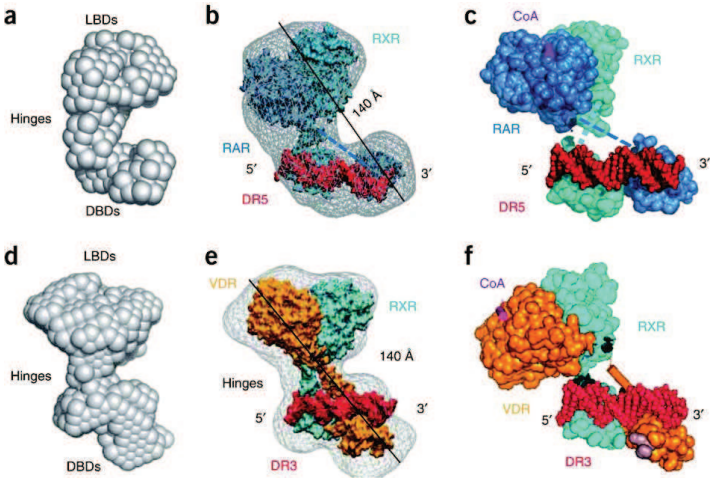


Figure 15 : Interaction entre les DBD de récepteurs nucléaires (RAR-RXR ; VDR-RXR) et différents motifs d'ADN (DR5 ; DR3) (Rochel et al., 2011)

Classiquement, on retrouve les sites de fixation des facteurs de transcription (*Transcription Factor Binding Site*, TFBS) dans le promoteur proximal couvrant -1000 pb à +200 pb par rapport au site d'initiation de la transcription (*Transcription Start Site*, TSS) (Sarkar and Maitra, 2008). Étant donné la nature flexible de l'ADN, un TFBS ne se trouve pas cependant obligatoirement à proximité d'un gène. Un site de fixation distal d'un point de vue de la séquence peut être proche spatialement, étant donné les repliements de l'ADN, qui forme des boucles (Gheldof et al., 2010). Il est donc important de ne pas réduire la recherche de sites génomiques à la région proximale du gène, mais d'offrir la possibilité de l'effectuer sur le génome entier.

Certains facteurs de transcription comme les RAR (Récepteur à l'Acide Rétinoïque) peuvent même provoquer la formation de ces boucles, une fois associés à des coactivateurs et à l'ARN Polymérase III (Rochette-Egly and Germain, 2009). Ainsi, les TFBS peuvent être situés à des distances importantes en amont du TSS mais également en aval dans des régions comme les 3'-UTR ou même les introns (Zhang and Gerstein, 2003). Les TFBS sont généralement assez courts, de 6 à 18 bases suivant les facteurs, et l'on observe de nombreuses variations entre sites reconnus par un même facteur de transcription.

Il existe des banques répertoriant ces motifs, comme Jaspar (Sandelin et al., 2004) ou Transfac (Wingender, 2008). Les TFBS peuvent être identifiés expérimentalement par de nombreuses méthodes, mais l'approche expérimentale à l'échelle de génomes entiers la plus couramment utilisée est actuellement, la méthode de *Chromatin ImmunoPrecipitation sequencing* (ChIP-Seq).

Les sites présentent souvent des variations importantes, c'est-à-dire que seules certaines bases d'un motif indispensables biologiquement sont conservées. De tels motifs sont dits dégénérés. Un site dégénéré peut être représenté sous forme de modèle de Markov caché ou de matrice de fréquences, répertoriant la fréquence d'apparition d'une base à chaque position du site. Un logo (Figure 16) permet de visualiser cette matrice. Ce type de représentation ne peut être adoptée qu'à partir d'un grand nombre de sites établis expérimentalement et ne permet pas de spécifier la conservation stricte d'une ou plusieurs positions sur la base de connaissances biologiques, le score étant établi sur l'ensemble des séquences. La représentation alternative est le motif consensus dégénéré, sans précision sur la fréquence des résidus mais qui autorise des positions conservées strictement, des positions ambiguës et des positions totalement variables (Figure 16). Ce type de motif peut être codé selon la nomenclature IUPAC (*International Union of Pure and Applied Chemistry*). La recherche de ces sites dans les séquences macromoléculaires est un problème exploré depuis longtemps en bio-informatique.

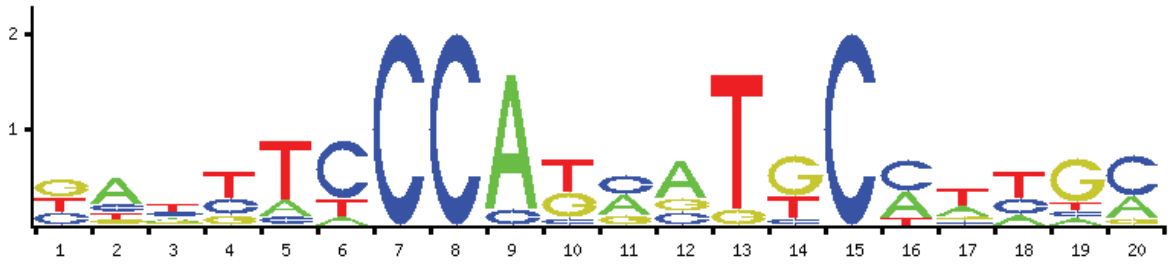


Figure 16: Le site de fixation du facteur de transcription STAF peut être représenté sous forme de logo de séquences, avec des positions fixes et des positions dégénérées (logo provenant de JASPAR)

La méthode la plus naturelle, mais aussi la plus lente de recherche de petites séquences dégénérées dans des séquences plus longues est basée sur les expressions régulières, comme implémenté dans l'outil 'dna-pattern' (van Helden et al., 2000) ou DNAid (Dardel and Bensoussan, 1988). La recherche par expression régulière est performante sur les protéines, les gènes ou les petites régions spécifiques mais devient vite limitée lorsqu'il s'agit de génomes eucaryotes complets (temps de recherche de plusieurs heures avec RSAT (*Regulatory Sequence Analysis Tool*) (Medina-Rivera et al., 2015). D'autres méthodes, comme les recherches par profils ou les outils basés sur les modèles HMM (*Hidden Markov Model*), s'appuient sur la fréquence des résidus à chaque position du motif. Ainsi, un HMM modélise la distribution de résidus autorisés à une position donnée, étant donné les caractéristiques d'un alignement multiple de séquences (Eddy, 1998). L'approche HMM est implémentée dans des logiciels comme Hammer (Finn et al., 2011) ou matrix-scan (Turatsinze et al., 2008), qui demandent en entrée une matrice qui représente le site recherché par des probabilités associées aux différentes bases du site.

1.1.2 Toujours plus vite : les arbres de suffixes compressés

La généralisation des méthodes de séquençage à très haut débit, leur rapidité, et leur prix de plus en plus accessible se traduit d'une part, par une augmentation prodigieuse du nombre de génomes séquencés, et d'autre part, par la nécessité de disposer de nouvelles méthodes d'analyse de ces génomes et de nouveaux programmes implémentant ces méthodes. Ainsi, de nouvelles méthodes ont vu le jour pour localiser de petits sites dégénérés à l'échelle de génome complet en un temps raisonnable. Beaucoup d'entre elles s'appuient sur l'indexation de la séquence dans laquelle un motif sera cherché. Parmi elles, les plus connues sont basées sur les arbres de suffixes (Ukkonen, 1995). Lorsqu'une séquence est encodée en un arbre de suffixes, tout chemin du nœud racine à n'importe quelle feuille de cet arbre correspond à un suffixe de la séquence d'intérêt. Chaque feuille contient l'information sur la position dans la séquence du suffixe donné. Cet arbre est simplement une forme condensée de la représentation de tous les suffixes d'un mot. Chaque nœud

d'un arbre de suffixes possède au moins deux fils. N'importe quel mot compris dans la séquence d'intérêt peut alors être considéré comme le préfixe d'un ou plusieurs suffixes. Ainsi comme le montre la Figure 17, pour la séquence 'GATTACA', 'A' est le préfixe des suffixes ‘A’, ‘ACA’ et 'ATTACA'.

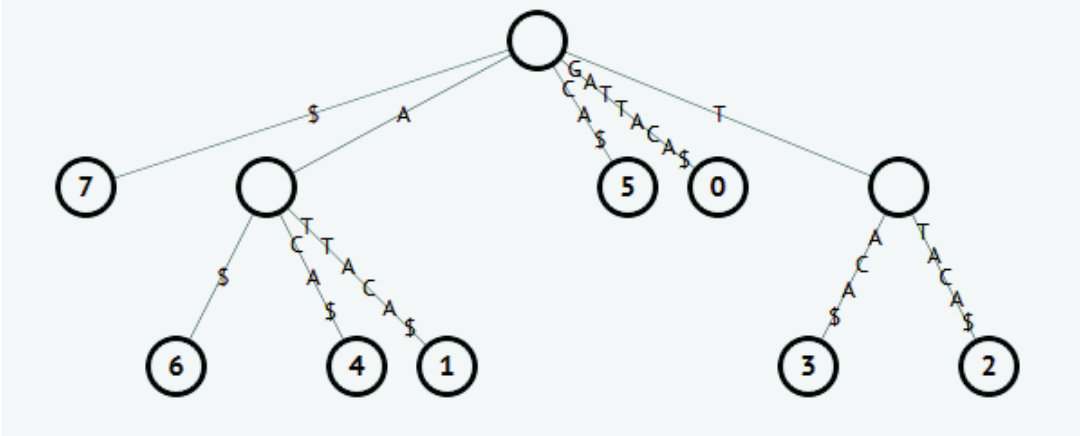


Figure 17 : Visualisation de l'arbre de suffixes du texte 'GATTACA' réalisée avec le site web http://visualgo.net/suffixtree.html

La recherche des positions de ce mot dans la séquence se décompose en deux étapes (Figure 18). D'une part, la progression dans l'arbre nœud à nœud, en partant de la racine, jusqu'à épuiser toutes les lettres du mot cherché, en choisissant à chaque fois le chemin correspondant à une partie de ce mot (étape montrée en bleu sur la figure). Si la progression devient impossible (aucun chemin ne correspond à la prochaine lettre du mot), le mot n'est pas présent dans la séquence. Lorsque toutes les lettres du mot sont épuisées, le nœud sur lequel on s'arrête possède la caractéristique unique d'être le nœud de plus bas niveau commun à tous les suffixes (de la séquence d'intérêt) dont notre mot est un préfixe.

Toutes les feuilles du sous-arbre dont ce nœud est la racine, correspondront à des positions de notre mot dans la séquence (entourées en vert sur la figure). La deuxième étape consiste donc à parcourir le sous-arbre pour récupérer les positions associées à toutes les feuilles.

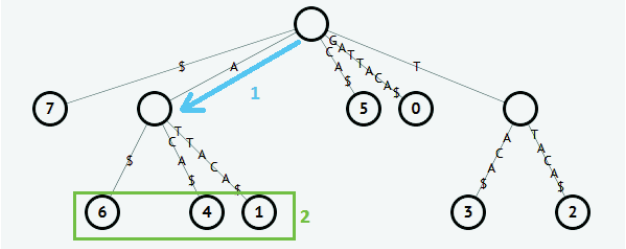


Figure 18: La recherche des positions se fait en deux étapes.

Le temps de parcours de l'arbre est fonction de la longueur du mot recherché, et le temps de récupération des positions dépend du nombre de feuilles, nombre lui-même dépendant de la taille du sous-arbre et donc du nombre d'occurrences du mot recherché dans la séquence. Le temps nécessaire pour connaître toutes les positions d'un mot dans une séquence dépend ainsi uniquement de la longueur du motif et de son nombre d'occurrences. Malgré leur rapidité, les arbres de suffixes possèdent un défaut de taille : la quantité de mémoire qu'ils occupent. Une autre structure de données fut donc proposée, associant la simplicité et puissance des arbres de suffixes avec une occupation mémoire moindre : les Arbres de Suffixes Compressés (Sadakane, 2007).

Pour trouver rapidement un motif dégénéré sur un génome, je pouvais utiliser deux stratégies :

- générer toutes les combinaisons possibles de séquences exactes à partir du motif dégénéré et les rechercher sur le génome avec un aligneur de lectures (utilisé pour aligner de courtes séquences exactes sur un génome entier tel que Bowtie (Langmead and Salzberg, 2012) par exemple),
- utiliser les embranchements 'OU' d'un motif dégénéré pour le chercher dans une structure de données en arbre également.

C'est cette deuxième approche qui m'a paru la plus intéressante pour plusieurs raisons. D'une part, la structure en arbre est naturellement adaptée à des motifs dégénérés. Chaque nœud étant l'équivalent d'un 'OU' d'une expression régulière, il n'est pas nécessaire de retraverser l'arbre à plusieurs reprises à partir de la racine pour trouver les occurrences de toutes les combinaisons associées à un motif dégénéré en des positions spécifiques (un tel motif pouvant lui même être représenté sous forme d'arbre).D'autre part, les propriétés additionnelles des arbres de suffixes, comme la recherche rapide d'une séquence correspondant à des coordonnées chromosomiques, recherche du motif le plus fréquent d'une longueur donnée, recherche du motif le plus long commun à deux séquences me semblèrent intéressantes pour de futures extensions du service web.

Enfin, les arbres permettent de détecter rapidement si un motif n'existe pas. Imaginons le motif AT[TG]CCC[CTG]NNNNN[TG]T[GA]AT[GTA]NNNTG[GA]. Si les séquences ATTCCC et ATGCCC n'existent pas dans le génome cible, la progression dans l'arbre s'arrête tout de suite après très peu d'efforts. En utilisant un aligneur de lectures, nous aurions dû générer et tester 9437184 combinaisons de séquences exactes.

Pour fournir une base au module de recherche de motif, mon choix s'est donc orienté vers les arbres de suffixes compressés, et en particulier l'implémentation (Välimäki et al., 2007) réalisée par le laboratoire des structures de données succinctes (SuDS, *Succinct Data Structures*, www.cs.helsinki.fi/group/suds/cst/ depuis renommé en GSA (*Genome-scale algorithmics*, www.cs.helsinki.fi/en/gsa/) car cette implémentation en langage C++ propose des méthodes

simples de manipulation d'arbres de suffixes compressés, comme s'il s'agissait d'arbres de suffixes ordinaires, comme illustré par le Tableau 13.

Tableau 13: Les fonctions les plus intéressantes proposées par l'implémentation des CST (Compressed Suffix Trees) réalisée par le SuDS. La complexité de la structure des données est encapsulée et des fonctions simulant le parcours d'un arbre sont exposées.

Fonction	Description
child(n,c)	Renvoie le numéro du nœud m qui est un enfant du nœud de numéro n à condition que l'arrête qui les lie commence par le caractère c. Renvoie 0 si aucun nœud ne satisfait à ces conditions.
edge(n, i)	Renvoie le caractère se trouvant à la i-ème position de l'arrête pointant vers le nœud de numéro n
numberofleaves(n)	Renvoie le nombre de toutes les feuilles du sous-arbre dont le nœud de numéro n est la racine.
isleaf(n)	Renvoie vrai si le nœud de numéro n est une feuille (n'a pas de nœuds fils)
textpos(f)	Renvoie la position dans le texte de la chaine de caractères qui correspond au parcours de l'arbre du nœud racine à la feuille de numéro f
firstChild(n)	Renvoie le numéro du premier fils (le plus à gauche) du nœud dont le numéro est spécifié.
sibling(n)	Renvoie le nœud à droite du nœud de numéro n ou 0 s'il n'y a plus de nœuds au même niveau à droite.
leftmost(n)	Renvoie le numéro de la feuille la plus à gauche du sous-arbre dont la racine est le nœud de numéro n
rightmost(n)	Renvoie le numéro de la feuille la plus à droite du sous arbre dont la racine est le nœud de numéro n

J'ai développé plusieurs fonctions afin de compléter le code de base et répondre aux exigences d'une recherche de motif dégénéré dans plusieurs espèces. Ces nouvelles fonctions écrites en C++ incluent une fonction de gestion de séquences dégénérées (Figure 19), une fonction de gestion d'un tableau de chromosomes (tous génomes confondus), des fonctions de lien avec Java ainsi qu'une fonction de « nettoyage » (excision de l'en-tête et des retours à la ligne) des fichiers contenant la séquence des chromosomes au format FASTA (fichier texte avec pour chaque séquence, une ligne d'en-tête puis les nucléotides représentés chacun par une lettre).

```
FONCTION recursivePatternSearch

INPUT :  1. motif à chercher, les positions dégénérées sont entre crochets, ex : AT[CG]A
         2. numéro du nœud (n), on commence avec 0
         3. position sur le label de l'arrête entre deux nœuds (l), on commence à 1
         4. position dans le motif (p), on commence à 0
         5. liste de positions, on commence avec une liste vide

Tant que la fin du motif n'est pas atteinte :
        Pour chaque caractère c à la position p
                Si c n'est pas un crochet :
                        Essayer de progresser dans l'arbre en fonction de n, de c et de l
                        Si la progression est possible :
                                Mettre à jour n (c'est toujours le nœud qu'on vise et non le nœud
                                courant) et l
                                Incrémenter p
                        Sinon :
                                retour
                Si c est un crochet ouvrant :
                        Incrémenter p
                        Tant que le caractère c à la position p n'est pas le crochet fermant
                                Remplacer la première paire de crochets et son contenu par c
                                Appeler recursivePatternSearch avec le nouveau motif
                                Incrémenter p
                        retour
Le nœud LCA de cette combinaison a été atteint. Stocker dans la liste de positions, les positions
contenues dans les feuilles du sous arbre.
```

Figure 19: Pseudocode de la fonction permettant de rechercher le motif dégénéré dans l'arbre de suffixes.

La partie Java comprend la gestion du chargement des chromosomes en se basant sur un fichier de configuration lu au démarrage, la recherche parallélisée dans les chromosomes, la génération de mésappariements (permettant des *hits* non exacts de la séquence dégénérée) en cas de besoin, avec une fonction récursive. Cette fonction génère un ensemble de combinaisons sous forme de motifs dégénérés, à partir d'un motif dégénéré, et d'un nombre de mésappariements demandé. Ce nombre peut être inférieur ou égal à la taille du motif, bien que pour des questions de performance de recherche, cette fonction n'est pas appelée avec un nombre de mésappariements supérieur à 2.

J'ai validé l'exactitude des résultats fournis par le module de recherche, en termes de nombre et position de sites trouvés, en comparant mes résultats à ceux d'un programme basé sur les expressions régulières développé précédemment au sein du laboratoire.

1.1.3 Une API pour une recherche massive

En plus de la recherche de sites avec une interface utilisateur, j'ai également créé une API permettant une interrogation simple, rapide et massive par un programme de PARSEC, permettant de rechercher jusqu'à 100 000 courtes séquences exactes sur le génome humain. Cette API permet de filtrer une liste de séquences par leur présence ou non sur le génome donné.

L'API demande en paramètre d'entrée une chaine de caractères sous format JSON comportant les champs détaillés dans le Tableau 14.

Tableau 14: Paramètres d'interrogation de l'API de filtrage de sites de PARSEC

Paramètre	Description
Genome	Le nom complet du génome à interroger. Tous les génomes chargés dans Parsec sont disponibles. Par exemple : *hg19_9606_Homo_sapiens_masked*
Sequences	Une liste de séquences nucléotidiques à tester. Le complémentaire inverse de chaque séquence est testé également.

L'API renvoie sous format JSON (*JavaScript Object Notation*, format permettant de représenter une information structurée) un document précisant la taille de la liste de séquences soumise, la taille de la liste de séquences non détectées, la liste de séquences filtrées (non trouvées sur le génome donné) et le temps mis à exécuter le filtrage.

Les tests de performances ont montré que Parsec pouvait rechercher 100 000 séquences exactes de 20 bases en 10 sec sur le génome humain.

1.2 Le contexte pour combattre le hasard

Effectuée seule, la détection de séquences n'est pas suffisante pour produire des résultats intéressants biologiquement. Lors de la recherche de sites dégénérés sur les génomes complets, la relative petite taille des motifs recherchés donne lieu à un nombre de *hits* très important (par exemple, le motif RGKTSA est retrouvé plus de cinq millions de fois dans le génome humain). Il

devient donc nécessaire de caractériser les vrais positifs, en complétant l'information de séquence par le contexte biologique. Ce contexte a plusieurs dimensions.

1.2.1 L'évolution conserve ce qui marche

Les régions génomiques importantes fonctionnellement (régions codantes, régions de régulation) sont soumises à une plus forte pression de sélection et évoluent moins rapidement que le reste du génome. En comparant des séquences génomiques d'organismes relativement proches tels que l'homme et la souris, on peut ainsi distinguer les régions de forte conservation qui correspondront généralement à des régions fonctionnelles. Cette approche est à la base des méthodes de *phylogenetic footprinting* et de *phylogenetic shadowing* (Wasserman and Sandelin, 2004) utilisées notamment pour la détection de TFBS (*Transcription Factor Binding Site*).

J'ai donc développé un module de contextualisation des sites par leur conservation plus ou moins stricte dans des génomes de plusieurs espèces. Pour cela, j'ai intégré dans la base de données de Parsec les alignements deux à deux générés par BLASTZ (Schwartz et al., 2003) et disponibles sur le site génomique de l'UCSC pour un grand nombre d'espèces. L'étude manuelle de la conservation de certains sites s'est appuyée sur des alignements multiples de séquences génomiques proposées à l'UCSC, construits par multiZ (Blanchette et al., 2004).

Le module de filtrage par conservation prend en entrée un ensemble de positions génomiques obtenues à partir d'un module en amont, et ne retient que les positions répondant à certains critères phylogénétiques. L'utilisateur choisit un ensemble d'organismes dans lesquels les sites devront être conservés. Par défaut, les sites seront retenus s'ils sont conservés dans au moins une des espèces sélectionnées, mais l'utilisateur peut décider de ne conserver que les positions conservées dans toutes les espèces. Trois niveaux de conservation sont proposés comme illustré par la Figure 20 :

- (i) Le niveau le plus permissif correspond à la présence du site d'intérêt dans une région alignée, donc suffisamment conservée (notée *Aligned Region*). Pour cela, on vérifie simplement si un alignement avec les séquences génomiques de l'un des organismes cibles choisis comprend les bornes du site d'intérêt.

- (ii) Le niveau suivant requiert la présence du site recherché dans la séquence alignée de l'organisme cible, sans que celui-ci soit nécessairement parfaitement aligné avec le site de l'organisme de départ. Cette conservation est notée *Conserved Site*. Pour cela, le module vérifie la présence du motif dans la séquence (sans gap) de l'organisme cible de l'alignement.

- (iii) Enfin, le niveau de conservation le plus strict (noté *Aligned Site*) demande la présence du site sur la séquence alignée du génome cible, parfaitement aligné avec le site de l'organisme de départ. Notons que nous comparons le motif, et non la séquence. Ainsi, si notre motif est RR (R

signifiant A ou G), que le site trouvé sur l'organisme source est AA, et que l'on retrouve GA en face sur l'alignement, on considère qu'il y a « match » exact.

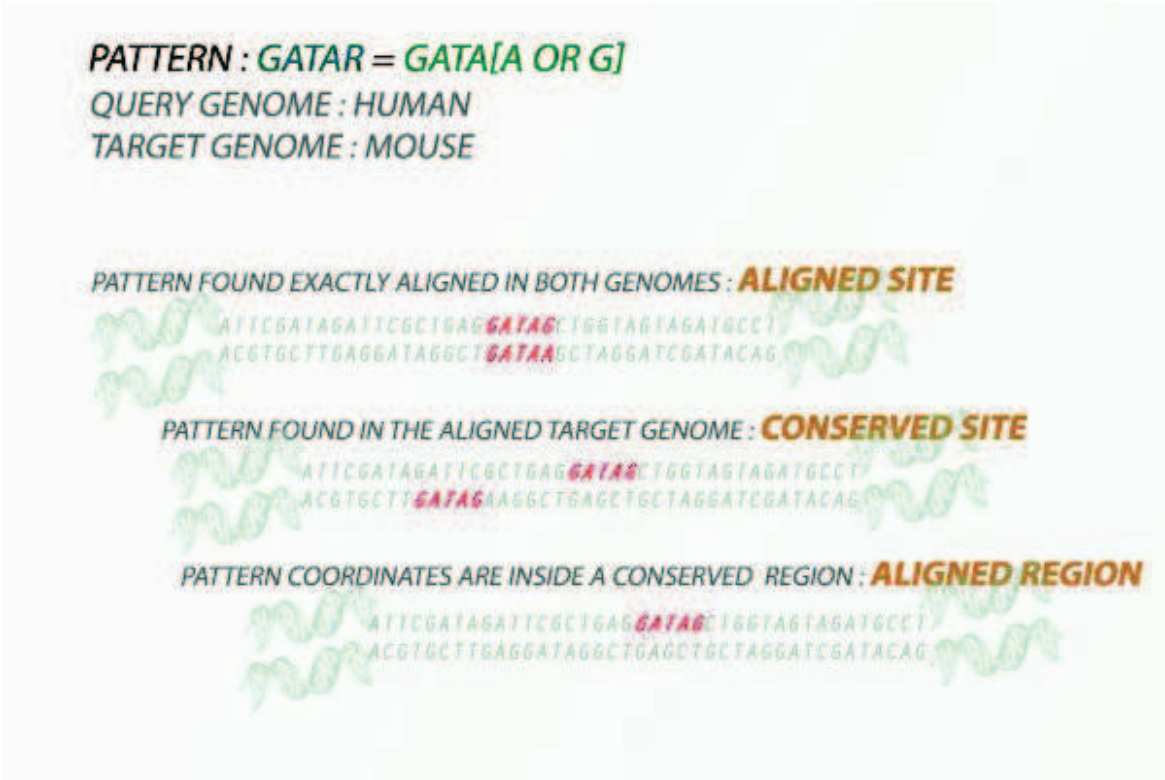


Figure 20: Parsec propose trois niveaux de conservation pour filtrer les sites par leur conservation entre espèces.

Les résultats sont présentés dans un tableau avec la liste des sites conservés et leur position. Le tableau peut être téléchargé sous format compatible Excel.

Afin d'améliorer les performances de ce module, j'ai procédé à plusieurs optimisations. Pour chaque site d'un génome donné, le module de filtrage par conservation recherche la liste des alignements avec un organisme cible qui comprennent le site d'intérêt. La taille des tables d'alignements deux à deux BLASTZ étant conséquente (> 50GO pour les alignements entre génome humain et 49 génomes eucaryotes), et le nombre de positions à tester pouvant aller jusqu'à plusieurs dizaines de milliers, j'ai dû analyser le plan d'exécution de la requête afin de l'optimiser de deux manières :

- Déterminer et évaluer les index qui donnent le meilleur temps de recherche,
- Ajouter des conditions redondantes à la requête afin de diminuer le nombre de lignes à parcourir par la base de données. Les positions de début et de fin des alignements cherchés sont cloisonnées avec un minimum et un maximum sous forme :

start_alignement <= start_site

AND

start_alignement >= start_site - max_alignement_lenght

AND

end_alignement >= end_site

AND

end_alignement <= end_site + max_alignement_lenght

J'ai également cherché à évaluer si le transfert d'une partie du code du côté de la base de données permettrait d'améliorer les performances grâce, entre autres, à la réduction des échanges réseau. J'ai donc recodé sous forme de procédures en Pg/Sql (le langage de Postgres), la partie du code chargée de filtrer les sites par leur conservation dans les génomes. Les performances obtenues étaient, contre toute attente, bien inférieures à celles sans procédures stockées. Après analyse étape par étape, de mes fonctions Pg/Sql, j'ai remarqué que l'appel à la fonction native de Postgres *substring* pour récupérer un caractère de la chaine de caractères était en grande partie responsable de cette lenteur. Après analyse du code source de cette fonction, j'ai remarqué qu'elle allouait, à chaque fois, de la mémoire pour créer la chaine de caractères renvoyée. J'ai donc codé une très petite fonction en C pour renvoyer un caractère précis d'une chaine de caractères. En utilisant cette fonction, la vitesse de la solution en Pg/Sql devint comparable à la solution où tout le code était dans la partie application. Les procédures stockées n'ayant pas permis d'améliorer la vitesse d'exécution, j'ai décidé de conserver tout le code dans la partie application.

Pour estimer et permettre la visualisation de la conservation de régions génomiques, je me suis appuyé sur deux méthodes, PhastCons et PhyloP, comme le montre la Figure 21.

- PhastCons est basé sur le modèle de Markov caché, et estime la probabilité qu'un nucléotide appartienne à un élément conservé, en tenant compte des nucléotides voisins (Siepel et al., 2005). Cette méthode est adaptée à la recherche de sites conservés.
- PhyloP (Pollard et al., 2010) mesure la conservation nucléotide par nucléotide, sans tenir compte des nucléotides adjacents. Cette approche est adaptée à une étude de nucléotides précis. Le graphique généré par PhyloP permet de comparer l'évolution réelle (en terme de vitesse et donc de conservation) avec une évolution neutre, sans pression particulière.

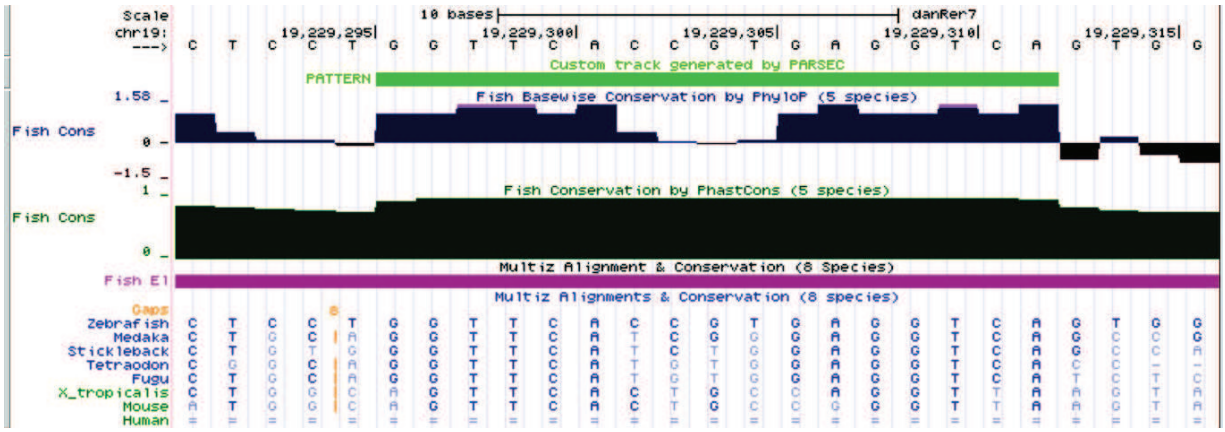


Figure 21: Parsec peut créer et envoyer au 'genome browser' de l'UCSC, un 'track' personnalisé avec le site d'intérêt (en vert clair). Une analyse visuelle permet alors de le comparer avec les données de conservation multi-espèces provenant de PhyloP (en bleu sombre) et PhastCons (en vers sombre). On peut voir que les deux parties du site DR5 (en vert clair) sont conservées (score en bleu foncé), ce qui n'est pas le cas de la région séparant les deux hémi-sites.

1.2.2 Le gène comme outil d'annotation et de filtrage

Le module de caractérisation du contexte génomique permet de rechercher les gènes à proximité de sites identifiés et de sélectionner ainsi, les sites répondant aux critères de localisation définis par l'utilisateur, comme le montre la Figure 22.

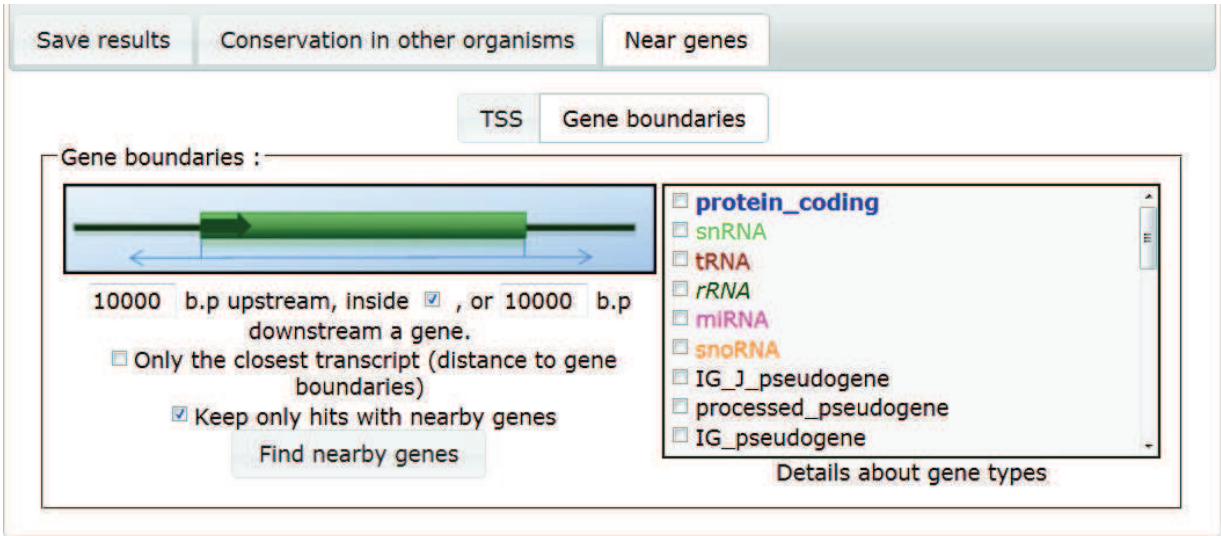


Figure 22 : Pour filtrer et annoter les sites par leur proximité avec les gènes, l'utilisateur peut choisir l'intervalle de proximité avec les gènes et le type de gènes qui l'intéresse.

Ainsi, dans le cadre d'une recherche de TFBS, les sites dans la région promotrice pourront être privilégiés par exemple. Les sites peuvent être sélectionnés de deux manières :

- en fonction de leur proximité à un site d'initiation de la transcription (*Transcription Start Site*, TSS). Les sites retenus doivent être situés dans un intervalle donné autour du TSS. L'utilisateur choisit les bornes en amont et en aval du TSS.

- en fonction de leur proximité à un gène ou plus exactement à la région génomique correspondant à un transcrit. Les sites en amont, en aval, ou à l'intérieur d'un gène sont sélectionnés. Les gènes des deux brins sont pris en compte.

Il est à noter que l'utilisateur peut choisir de considérer tous les types de gènes annotés dans l'espèce considérée ou d'en privilégier certains comme les gènes codant pour des protéines. Les statistiques concernant les types et nombre de transcrits trouvés sont affichés à côté des résultats, permettant d'avoir un aperçu global (Figure 23). Les gènes sont colorés selon leur type. Les noms de gènes correspondant aux transcrits identifiés sont affichés sur la page de résultats, à côté du site. Un lien permet d'explorer plus en détails la position du site par rapport à tous les transcrits identifiés, et un lien vers le *genome browser* de l'UCSC permet de visualiser le site dans son contexte génétique.

Figure 23: Description synthétique des transcrits trouvés à côté des sites identifiés

En pratique, le module de contextualisation par proximité des gènes possède un triple rôle, à la fois de filtrage de sites, d'annotation des sites et d'annotation des gènes. Premièrement, le filtrage par la proximité aux gènes permet d'éliminer les sites considérés comme trop éloignés pour intervenir dans la régulation de l'expression. Une entière liberté est laissée à l'utilisateur afin qu'il puisse adapter ce critère à sa problématique, en particulier à l'organisme considéré et au facteur de transcription étudié. Deuxièmement, la détection de gènes de fonction connue à proximité d'un ou plusieurs sites peut permettre de caractériser fonctionnellement les sites en question. Enfin, des sites au rôle connu présents à proximité d'un gène de fonction inconnue vont permettre de rattacher le gène à un ou plusieurs processus biologiques.

1.2.3 L'ontologie pour faire parler les gènes

Parsec utilise les données sur la fonction moléculaire, le rôle et localisation des gènes et de leur produits provenant de GO (*Gene Ontology*) afin d'annoter (ajouter de l'information) les gènes et les sites mais également, pour filtrer (enlever du bruit) les gènes et les sites non pertinents.

1.2.3.1 Annotation

Le module de recherche d'enrichissement permet la caractérisation fonctionnelle d'un ensemble de gènes. Ces gènes peuvent, par exemple, avoir été identifiés comme étant les gènes potentiellement régulés par un facteur de transcription. L'utilisateur pourra ainsi voir si une ou plusieurs fonction(s) biologique(s) commune(s) se dégagent au sein des gènes cibles du facteur de transcription.

Le module de contextualisation fonctionnelle se concentre sur l'injection et la récupération de données de GoMiner (programme déterminant l'enrichissement en termes GO d'une liste de gènes). C'est une surcouche qui permet de se servir de GoMiner comme d'une librairie java, et ainsi de s'affranchir de l'interface en ligne de commande de GoMiner. Le module se charge de créer et fournir à GoMiner, un fichier de gènes protéiques uniques et de traiter le fichier généré par GoMiner afin de créer une liste d'objets résultats. Ces résultats pourront être passés à un autre module, pour analyse supplémentaire. Comme conseillé sur le site de GoMiner, j'ai fait fonctionner GoMiner avec une base de données GO locale pour accélérer le traitement.

1.2.3.2 Filtrage

Une fois la recherche d'enrichissement fonctionnelle effectuée sur les gènes liés aux sites d'intérêt, l'utilisateur peut sélectionner les catégories GO qui lui semblent les plus intéressantes afin de ne garder que les gènes liés à ces catégories, et les sites correspondants. Ce filtrage par catégories GO permet ainsi d'affiner encore plus le set de sites trouvés sur le génome par leur fonction biologique probable.

1.3 La modularité pour mieux s'adapter

La modularité est une composante essentielle de Parsec permettant une utilisation souple et le choix de l'orientation de l'analyse à chaque étape en fonction des résultats obtenus.

Chaque module fournit obligatoirement un certain nombre de services. Il peut être remplacé par n'importe quel autre module offrant ces mêmes services. Par exemple, le module de recherche de motif fournit un ensemble de positions. Le module actuel le fait en recherchant un motif donné sur un génome mais il pourrait facilement être remplacé par un module obtenant ces positions par l'interrogation d'une URL (*Uniform Ressource Locator*), ou d'une autre source de données. Les modules sont indépendants et la communication entre eux est assurée par un gestionnaire superviseur. Ce faible couplage entre le module et le reste du système facilite l'intégration de nouveaux modules (Figure 24). A chaque module ajouté au gestionnaire est associée une liste de noms de modules qui devront traiter ses résultats. Ainsi, lorsque le module s'exécutera et affichera ses résultats, on pourra sur cette même page voir les formulaires de traitement de ces résultats par les modules enregistrés.

Chaque module génère le fragment de code HTML permettant l'interaction avec l'utilisateur, il fait les calculs, produit les résultats et fournit également le fragment de code HTML pour l'affichage de ces résultats. C'est le service web qui procède à l'affichage de la page d'attente, au lancement et à la gestion des calculs dans des *threads* séparés ainsi qu'à l'affichage des résultats.

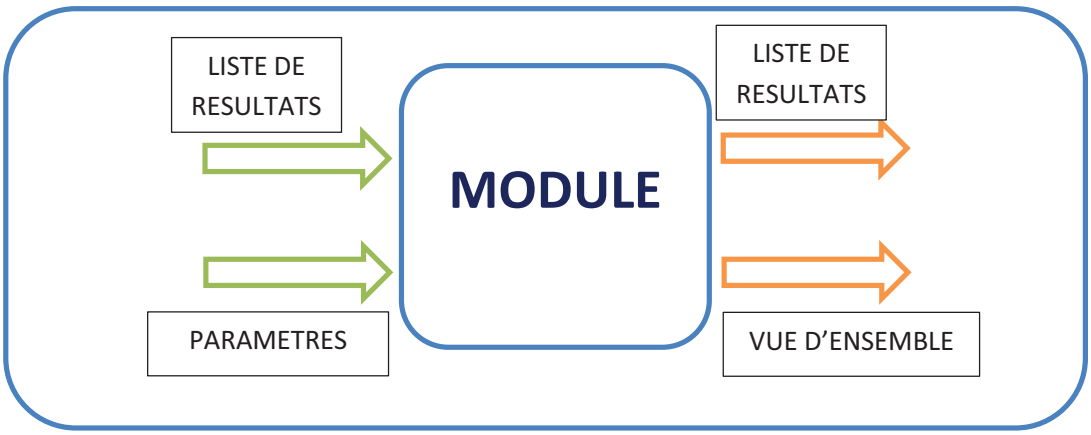


Figure 24: L'interaction d'un module avec le reste du système se fait par deux entrées (flèches vertes) et deux sorties (flèches de couleur orange).

1.4 Publication

PARSEC:
PAtteRn SEarch and Contextualization
Bioinformatics, 2013

BIOINFORMATICS APPLICATIONS NOTE Vol. 29 no. 20 2013, pages 2643–2644
doi:10.1093/bioinformatics/btt455

Genome analysis

Advance Access publication August 7, 2013

PARSEC: PAtteRn SEarch and Contextualization

Alexis Allot, Yannick-Noël Anno, Laetitia Poidevin, Raymond Ripp, Olivier Poch and Odile Lecompte*

Laboratoire de Bioinformatique et Génomique Intégratives, Institut de Génétique et de Biologie Moléculaire et Cellulaire (IGBMC) CNRS/INSERM/UDS, 67404, Illkirch, France

Associate Editor: John Hancock

ABSTRACT

Summary: We present PARSEC (PAtteRn Search and Contextualization), a new open source platform for guided discovery, allowing localization and biological characterization of short genomic sites in entire eukaryotic genomes. PARSEC can search for a sequence or a degenerated pattern. The retrieved set of genomic sites can be characterized in terms of (i) conservation in model organisms, (ii) genomic context (proximity to genes) and (iii) function of neighboring genes. These modules allow the user to explore, visualize, filter and extract biological knowledge from a set of short genomic regions such as transcription factor binding sites.

Availability: Web site implemented in Java, JavaScript and C++, with all major browsers supported. Freely available at lbgi.fr/parsec. Source code is freely available at sourceforge.net/projects/genomicparsec.

Contact: odile.lecompte@unistra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on May 6, 2013; revised on July 16, 2013; accepted on August 2, 2013

1 INTRODUCTION

Genomic sites are short genomic regions with a defined biological function (e.g. regulation of gene expression, splicing or epigenetic signals). Most genomic sites are represented by degenerated motifs with a scattered distribution and can either be specific to a given species or conserved between species. The characterization of these sites is a major challenge in the exploitation of next-generation sequencing data and in understanding genome expression. The *in silico* detection of these motifs in complete genomes is associated with a huge amount of noise. Consequently, the development of a computational platform for accurate definition of genomic sites requires the integration of various large-scale biological data resources to filter out false positives.

PARSEC (PAtteRn Search and Contextualization) is a modular web service designed for the rapid localization and characterization of genomic sites. The main program exploits an efficient data structure, namely, compressed suffix trees (CST) (Sadakane, 2007), to rapidly localize sequences or degenerated patterns in complete eukaryotic genomes (eight species are currently available: human, mouse, rat, chicken, zebrafish, drosophila, nematode and yeast). This pattern search module is linked to three in-house modules for conservation analysis, genomic context analysis and functional filtering, as well as to the GoMiner functional enrichment tool (Zeeberg *et al.*, 2005). It can be used to filter biologically meaningful genomic sequences, to complement transcriptomic data analysis and to further characterize known genomic sites.

In contrast to web services dedicated to genomic pattern searches like TagScan (Iseli *et al.*, 2007), or to functional interpretation of genomic regions like GREAT (McLean *et al.*, 2010), PARSEC proposes both search and characterization aspects in a single intuitive interface, guiding the user through the discovery path.

2 ARCHITECTURE AND IMPLEMENTATION

2.1 Web infrastructure

The PARSEC web service is a modular infrastructure composed of five modules (Fig. 1). These modules can be combined in various 'discovery pipelines', exploiting the relevant tools and results at each step of the analysis.

The pattern search module is based on a CST implementation (Valimaki *et al.*, 2007), which we extended to facilitate degenerated pattern searches (interpretation of basic regular expressions and recursive navigation of tree edges) and parallelization of chromosomal searches (management of an array of CST representing chromosomes). It is adapted for easy use as a native library for a Java program. The user can submit degenerated patterns (minimum of 5 bp) using the IUPAC code and can additionally allow up to two mismatches. The maximum number of hits for further characterization is set to 100 000 on the web server owing to memory limitations, but can be easily increased in a local installation. For the same reason, conservation analysis cannot be performed for patterns defined with one or two mismatches.

The conservation module, based on BlastZ pairwise alignments provided by University of California, Santa Cruz (UCSC) (Dreszer *et al.*, 2012), provides fast evolutionary screening and characterization of the identified genomic sites. The analysis can be customized, because of the three hierarchized levels of conservation stringency (from perfect alignment of query and target sites to conservation of the region around the query site) and the possibility to select the minimum number of organisms in which a site must be conserved. Depending on the biological question, the user can select phylogenetically close organisms for *phylogenetic shadowing* or more distant ones for *phylogenetic footprinting*.

*To whom correspondence should be addressed.

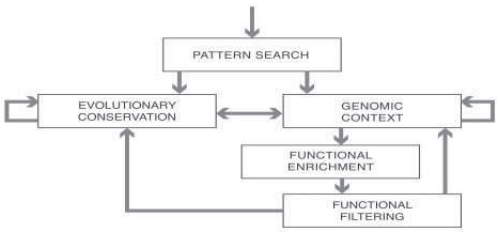


Fig. 1. Non-linear organization of modules in PARSEC, allowing multiple exploitation scenarios. The user chooses the analysis path according to results obtained at each previous step

The genomic context module, also based on data from UCSC tables (ensGene, EnsemblSource, ensemblToGeneName tRNAs), allows the detection of gene candidates spatially linked to the site of interest. Again, a high level of customization is provided. Analyses can be performed relative to the two gene boundaries or to the transcription start site. Several types of genes are available: protein coding, miRNA, snRNA, snoRNA, tRNA, etc. The user can retrieve all genes in proximity of the site or can choose only the closest gene to the site. The full set of alternative transcripts of proximal genes with their relative distances to the site is also provided and can be visualized through a link to UCSC.

For the enrichment module, we developed a layer allowing the use of GoMiner as a simple Java library. It allows the functional characterization of a previously identified set of proximal genes and query sites. This can be useful for identification of processes regulated by a transcription factor for example.

The functional filtering step allows the selection of sites potentially related to specific molecular functions, biological processes or cellular components.

At each step, the results can be saved as browser extensible data (BED) or comma-separated values (CSV) format files and include various links to external resources [UCSC, Ensembl (Flicek *et al.*, 2012)] for additional information. The algorithmic complexities and running times of the different modules are provided in Supplementary Dataset S1.

2.2 Installing PARSEC on a local server

The PARSEC web infrastructure relies on an information manager that allows automated retrieval of primary data (genomes, genes and BlastZ alignments) from UCSC and their integration in the PARSEC database. A new species can easily be added to PARSEC by specifying its NCBI taxonomy identifier, its UCSC genome version and some other parameters. This genome information is added to the PARSEC startup genome loading configuration file, so that the new genome is directly available after a simple web service restart. Genetic information is added to the PARSEC database, using a set of intelligent parsers, which check, for example, whether all the required gene types were found for the given organism and, if necessary, add the missing gene types. Each database entry is labeled with its source, and genetic or alignment entries can be easily updated by running retrieval commands on existing organisms. The source code, the information manager command line program and the .WAR archive for deployment on a Tomcat server are all freely available.

3 CASE STUDY

PARSEC can be used for various exploitation scenarios. For instance, the user can retrieve the set of genes potentially regulated by a transcription factor. To illustrate this, we searched for the DR5 (direct repeats separated by 5 bp) Retinoic Acid Response Element (RARE) in the masked human genome (version hg19), using the consensus pattern RGKTSANNNNNRGKTSA (Lalevee *et al.*, 2011). We localized 14 251 sites in 5.3 s. Using the conservation module, we then selected the sites conserved ('Aligned site' and 'Conserved site' parameters) in at least one amphibian or fish species. We found 178 sites in 20.8 s. We then filtered the 104 sites located near a transcription start site (5000 bp upstream, 5000 bp downstream) in the human genome using the genomic context module (0.13 s). We then analyzed the nearby protein coding genes with the functional enrichment module. Several GO categories were identified (false discovery rate $< 10^{-5}$), including embryo development and retinoic acid receptor signaling pathway. These categories related to retinoic acid regulation (Kumar and Duester, 2010) demonstrate PARSEC's ability to perform biologically meaningful genomic site analysis. Another case study (Supplementary Dataset S2) illustrates the usage of mismatches in PARSEC to improve the definition of the neuron-restrictive silencer factor binding site.

ACKNOWLEDGEMENTS

The authors are grateful to Vincent Laudet and Cécile Rochette-Egly for helpful discussions during the development of PARSEC. They thank Julie Thompson for critical reading of the manuscript.

Funding: This work was supported by the Agence Nationale de la Recherche [*Puzzle-Fit*: 09-PIRI-0018-02 and BIPBIP: ANR10-BINF03-05].

Conflict of Interest: none declared.

REFERENCES

Dreszer,T.R. *et al.* (2012) The UCSC genome browser database: extensions and updates 2011. *Nucleic Acids Res.*, **40**, D918–D923.

Flicek,P. *et al.* (2012) Ensembl 2012. *Nucleic Acids Res.*, **40**, D84–D90.

Iseli,C. *et al.* (2007) Indexing strategies for rapid searches of short words in genome sequences. *PLoS One*, **2**, e579.

Kumar,S. and Duester,G. (2010) Retinoic acid signaling in perioptic mesenchyme represses Wnt signaling via induction of Pitx2 and Dkk2. *Dev. Biol.*, **340**, 67–74.

Lalevee,S. *et al.* (2011) Genome-wide *in silico* identification of new conserved and functional retinoic acid receptor response elements (direct repeats separated by 5 bp). *J. Biol. Chem.*, **286**, 33322–33334.

McLean,C.Y. *et al.* (2010) GREAT improves functional interpretation of cis-regulatory regions. *Nat. Biotechnol.*, **28**, 495–501.

Sadakane,K. (2007) Compressed suffix trees with full functionality. *Theory Comput. Syst.*, **41**, 589–607.

Valimaki,N. *et al.* (2007) Compressed suffix tree—a basis for genome-scale sequence analysis. *Bioinformatics*, **23**, 629–630.

Zeeberg,B.R. *et al.* (2005) High-Throughput GoMiner, an 'industrial-strength' integrative gene ontology tool for interpretation of multiple-microarray experiments, with application to studies of common variable immune deficiency (CVID). *BMC Bioinformatics*, **6**, 168.

Supplementary Material

PARSEC: PAtteRn SEarch and Contextualization

Alexis Allot, Yannick-Noël Anno, Laetitia Poidevin, Raymond Ripp, Olivier Poch and Odile Lecompte

Supplementary dataset S1: Algorithmic complexity of the modules and running times for different analysis parameters in masked human genome

Table S1.1. Algorithmic complexity of the modules.

Pattern search	Worst case : $O(4^m+z)$ with z all occurrences of the pattern P of total length m and maximum 4 possible symbols for each position.
Conservation analysis	O(nm) where n is the number of hits and m the number of target organisms.
Genomic context	O(n) where n is the number of hits.
Functional Enrichment	O(n) where n is the number of genes.
Functional Filtering	O(n) where n is the number of hits.

Table S1.2. Running times for different analysis parameters in the masked human genome.

Pattern	Mis-matches	Pattern search	Conservation analysis	Genomic context	Functional Enrichment
RGKTSA	0	*3 sec* 5,530,746 *hits*	----	----	----
ATCRGKTSARGKTSA	0	*0.032 sec 114 hits*	*annotation in dog and cat 0.598 sec*	*annotation 0.164 sec*	*56 sec*
ATCRGKTSARGKTSA	1	*0.841 sec 4,951 hits*	----	*annotation 7 sec*	*64 sec*
ATCRGKTSARGKTSA	2	*10 sec 80,407 hits*	----	*annotation 122 sec*	*89 sec*
CAGCACCNNGCACAGNNNC	2	*74 sec 22,634 hits*	----	*annotation 52 sec*	*70 sec*

Supplementary dataset S2: Analysis of the binding site of the neuron-restrictive silencer factor (NRSF)

The neuron-restrictive silencer factor (NRSF) is reported to repress transcription of several neuronal genes in non-neuronal cells by binding to the neuron-restrictive silencer element (NRSE) defined by the following degenerated pattern: CAGCACCNNGCACAGNNNC (Schoenherr, et al., 1996). We searched for this pattern in the masked human genome to identify the target genes of NRSF. We then filtered these hits using the genomic context module (only hits located at +/10 kb of a protein gene TSS were selected) and retrieved the closest protein gene for each selected hit. Only 36 genes were retrieved using these parameters. Since the reported pattern is moderately degenerated, we extended our search by allowing 1 or 2 mismatches (Table S2.1).

Table S2.1. Search for NRSE in the masked human genome

Mis-matches	Hits	Hits filtered by genomic context	Distinct closest protein genes	Functional enrichment (evidence: all)
0	96	37	36	No significant enrichment
1	2,249	789	741	179 genes in GO categories related to neurons, synapse, neurotransmitters...
2	22,634	7,272	5,110	829 genes in GO categories related to neurons, synapse, neurotransmitters...

With 2 mismatches, we obtained a very high number of potential target genes suggesting a large amount of false positive hits. With 1 mismatch, we retrieved 741 potential target genes. Using the functional filtering module, we selected 179 of these target genes belonging to GO categories unambiguously related to neurons, synapses and neurotransmitters (Figure S2.2) and retrieved the corresponding 191 potential NRSE.

Figure S2.2. GO categories with significant enrichment (p-value<10^{-5}, FDR<10^{-5}). Categories directly related to neurons, neurotransmitters and synapses are indicated in bold.

cell periphery
regulation of biological quality
biological regulation
cellular response to stimulus
localization
neurological system process
regulation of cellular process
regulation of biological process
system development
generation of neurons
regulation of synaptic plasticity
synapse
plasma membrane part
synapse part
cell body
regulation of system process
neuronal cell body
neuron projection
cell projection
multicellular organismal signaling
potassium channel complex
cation channel complex
multicellular organismal process

intrinsic to plasma membrane
dendrite
neuron differentiation
cell junction
signaling
regulation of signaling
voltage-gated cation channel activity
neurogenesis
transmission of nerve impulse
lipid binding
voltage-gated potassium channel complex
nervous system development
multicellular organismal development
synaptic transmission
cell-cell signaling
cell communication
neurotransmitter transport
potassium ion transport
integral to plasma membrane
plasma membrane
phospholipid binding
system process
regulation of neurotransmitter levels

Sequence analysis of these NRSE showed that the 11th position is the most variable among previously non-degenerated positions (Figure S2.3). C is replaced by G in 31% of the 191 potential NRSE. Thus, we propose a new consensus pattern: CAGCACCNNGSACAGNNNC (S = C or G).

Figure S2.3. Sequence logo representation (Crooks et al, 2004) of the 189 potential NRSE

The new consensus pattern CAGCACCNNGSACAGNNNC (without mismatch) retrieved 471 hits. 201 potential sites are located at +/- 10 kb of a protein gene TSS. The functional enrichment analysis of the 194 corresponding candidate genes clearly highlights the prevalence of neuron-related processes (Figure S2.4), in agreement with NRSF function.

Figure S2.4. GO categories with significant enrichment (p-value<10^{-5}, FDR<10^{-5}). Categories directly related to neurons, neurotransmitters and synapses are indicated in bold.

regulation of transmission of nerve impulse
neurological system process
regulation of synaptic transmission
generation of neurons
regulation of synaptic plasticity
secretion
synapse
synapse part
cell body
regulation of system process
neuronal cell body
neuron projection
cell projection
presynaptic membrane
multicellular organismal signaling
axon part
multicellular organismal process
regulation of neurological system process
intrinsic to plasma membrane

synaptic vesicle membrane
clathrin coated vesicle membrane
dendrite
axon
signal release
signaling
transmission of nerve impulse
synaptic vesicle
locomotory behavior
learning or memory
neurotransmitter secretion
synaptic transmission
cell-cell signaling
neurotransmitter transport
integral to plasma membrane
neurofilament
system process
generation of a signal involved in cell-cell signaling
regulation of neurotransmitter levels

References

Schoenherr, C.J., *et al.* (1996) Identification of potential target genes for the neuron-restrictive silencer factor, *Proc Natl Acad Sci,* **93**(18):9881-6
Crooks, G.E., *et al.* (2004) WebLogo: A sequence logo generator, *Genome Research*, **14**:1188-1190

1.5 Recul sur le travail effectué

Parsec est accessible à l'adresse http://lbgi.fr/parsec/. Son fonctionnement presque continu sur près de deux ans (interrompu par quelques maintenances de la machine) a montré sa robustesse au fil du temps. Il a subi plusieurs mises à jour afin d'ajouter des fonctionnalités lors des collaborations et améliorer la navigation dans les différentes étapes du scénario d'analyse.

De plus, l'équipe à l'origine de l'implémentation des arbres de suffixes compressés sur laquelle est basé le module de recherche de Parsec a depuis travaillé sur de nouvelles structures de données encore plus intéressantes, et orientées vers un usage en génomique (www.cs.helsinki.fi/en/gsa/publications). Une évolution de Parsec pourra donc intégrer ces avancées.

Je vais présenter dans un premier temps les trois projets biologiques utilisant Parsec auxquels j'ai participé, puis je le comparerai à des solutions concurrentes, pour enfin prendre un recul sur les choix techniques et algorithmiques effectués et les évolutions possibles.

1.5.1 Applications

Parsec a été appliqué à trois problématiques biologiques distinctes : l'analyse de gènes régulés ou potentiellement régulés par l'acide rétinoïque, la détermination d'un set de gènes candidats ayant un rôle dans les ciliopathies et enfin, une étape de filtrage pour la génération de librairies de siRNA.

Dans les trois cas, la question s'est posée d'effectuer les recherches sur les génomes dans leur ensemble ou d'utiliser la séquence génomique masquée (c'est-à-dire en excluant les régions répétées). On peut privilégier la sensibilité en faisant la recherche sur un génome non masqué, ou la spécificité en faisant la recherche sur un génome masqué. Nous avons choisi d'utiliser un génome masqué pour privilégier la qualité des sites caractérisés à l'exhaustivité de la recherche.

1.5.1.1 Application à la régulation par l'acide rétinoïque

Dans le cas d'une collaboration avec l'Equipe de Zoologie Moléculaire de l'Université de Lyon, nous avons utilisé Parsec afin de répondre au besoin de recherche, filtrage et caractérisation rapide des sites génomiques. Cette collaboration s'intéressant aux sites reconnus par le récepteur à l'acide rétinoique chez le poisson zèbre, c'est sur cette problématique biologique que j'ai décidé d'évaluer la puissance de Parsec à grande échelle et la cohérence des résultats produits.

1.5.1.1.1 Contexte biologique

Les récepteurs nucléaires sont des facteurs de transcription activés par un ligand (Aagaard et al., 2011). On distingue généralement plusieurs domaines dans un récepteur nucléaire, comme le montre la Figure 25. Le DBD (*DNA Binding Domain*) très conservé assurera la reconnaissance de l'élément de réponse (*Response Element*, RE), le LBD (*Ligand Binding Domain*) fixera le ligand, le NTD (*N-Terminal Domain*) assurera la sélectivité des gènes régulés, et le CTE (*C-Terminal Extension*) participera à la spécificité de reconnaissance du RE.

Figure 25: Structure schématique d'un récepteur nucléaire

Un des représentants des récepteurs nucléaires est le Récepteur à l'Acide Rétinoïque (RAR) (Aagaard et al., 2011). Ce récepteur forme un hétéro-dimère (deux protéines différentes) en tête à queue avec un RXR (*Retinoid X Receptor*) pour pouvoir se lier à l'ADN. L'hétérodimère RAR:RXR peut se lier à l'ADN en présence et en absence de son ligand (l'Acide Rétinoïque, AR). Sans ligand, RAR:RXR se fixe sur le génome, recrute des corépresseurs et réprime l'expression du gène. En présence du ligand, les corépresseurs sont relâchés, RAR:RXR recrute des coactivateurs et active l'expression du gène cible, en recrutant la machinerie transcriptionnelle et les facteurs généraux de transcription. L'une des conséquences de l'action des coactivateurs sera une décompaction de la chromatine.

Le RAR est connu pour avoir deux types de cibles : les gènes codant pour des protéines qui auront une action donnée, et les gènes codant eux-mêmes pour des facteurs de transcription (les gènes *hox* par exemple) qui réguleront l'expression d'autres gènes. L'action des RAR peut donc être directe ou indirecte. Enfin, les RAR subissent un certain nombre de modifications post-traductionelles, comme la phosphorylation ou l'ubiquitination (Rochette-Egly and Germain, 2009). Il a également été récemment montré que le RAR alpha pouvait lier l'ARN et réprimer la traduction.

La vitamine A, ou rétinol, est une substance connue entre autres pour son importance dans le développement oculaire et embryonnaire, l'organogénèse, les fonctions immunitaires ou reproductives (Samarut and Rochette-Egly, 2012). Elle régule divers gènes responsables de la formation des axes du corps, en particulier les gènes *hox* (Cunningham and Duester, 2015). Les cellules possèdent des protéines capables de transformer cette vitamine en rétinaldéhyde, puis en acide rétinoïque, qui est sa forme active. Ces composés font partie de la famille des rétinoïdes, de petites molécules hydrophobes et liposolubles, passant facilement par les membranes cellulaires. L'acide rétinoïque sert de ligand à l'hétérodimère RAR:RXR. Un rôle important dans la régulation de transcription par l'acide rétinoïque est joué par la protéine CYP26A1 codée par le gène du même nom, qui régule la quantité d'acide rétinoïque dans le milieu, en l'oxydant et le rendant inactif. Le gène de CYP26A1 est également régulé par l'acide rétinoïque.

La fixation de RAR:RXR sur le génome se fait au niveau d'éléments de réponse à l'acide rétinoïque appelés RAREs (*Retinoic Acid Response Elements*). Du fait de la configuration en hétérodimère de RAR:RXR, le complexe reconnaîtra des sites répétés directement (*direct repeat*, ou DR), c'est-à-dire des motifs orientés dans le même sens et éventuellement séparés par quelques bases. Le premier hémi-site du RARE est généralement reconnu par le DBD du RXR tandis que le second est reconnu par celui du RAR. Il existe plusieurs types de RAREs. Par exemple, RGKTSANNRGKTSA est un DR2, car les deux répétitions directes du motif RGKTSA sont séparées par deux bases. Le RARE le plus connu est de type DR5 et régule un grand nombre de fonctions. On retrouve aussi le DR2, plus rarement le DR1 voire, le DR8. Une majorité de RAREs se trouveraient dans les régions intergéniques ou introniques, loin du TSS du gène régulé. Bien qu'il soit classiquement convenu de considérer le motif du RARE comme RGKTCA, des études récentes (Lalevee et al., 2011) ont montré que ce motif pouvait être élargi au motif RGKTSA. Ce sont les RAREs basés sur cet hémi-site que j'ai été amené à rechercher dans différents génomes complets d'Eucaryotes supérieurs, en particulier chez le poisson zèbre, afin d'évaluer l'utilité de Parsec pour cette question biologique.

1.5.1.1.2 Evaluation de Parsec sur la problématique biologique

La cible prioritaire de mon analyse était les RAREs de type DR5 (dont le logo est visible sur la Figure 26), site principal (Rochette-Egly and Germain, 2009) de fixation du RXR:RAR en privilégiant le motif RGKTSA retrouvé près de certains gènes connus pour être régulés par les RAR (Lalevee et al., 2011), mais j'ai également recherché les RARE de type DR2 (à savoir RGKTSANNRGKTSA) et DR8.

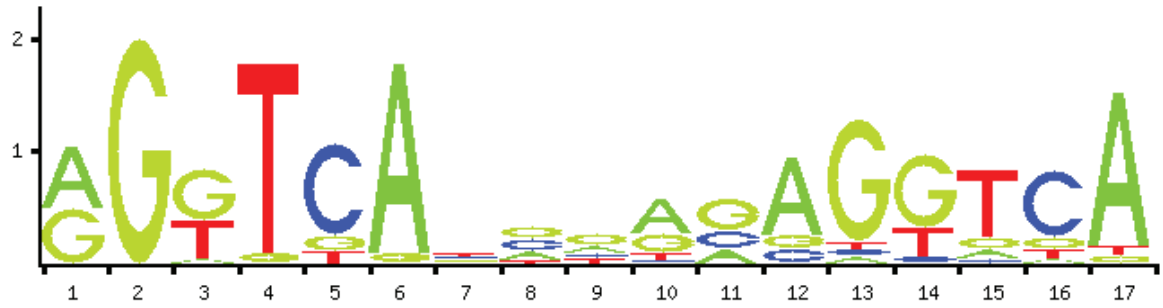


Figure 26 : Logo provenant de JASPAR et caractérisant le site DR5, la taille des caractères reflète le degré de conservation de chaque position.

Les recherches ont été menées dans le génome du poisson zèbre et dans plusieurs organismes modèles à titre de comparaison (voir chapitre 1.4 pour un exemple chez l'homme). L'analyse avec Parsec dénombre 7365 sites RAREs potentiels de type DR5 chez le poisson zèbre. On remarque également un lien proportionnel entre la taille du génome et le nombre de sites trouvés. Ainsi l'homme, le rat et la souris, dont le génome séquencé fait entre 1,5 et 2 fois la taille de celui

du poisson zèbre et du poulet, possèdent également deux fois plus de motifs RAREs. De plus, comme attendu, il existe davantage d'occurrences de RARE DR5 que DR2 et DR8.

On voit que l'identification de sites avec le module de recherche génère énormément de sites potentiels parmi lesquels se trouvent beaucoup de faux positifs, c'est-à-dire beaucoup de régions génomiques présentant le motif recherché mais ne correspondant pas biologiquement à des RAREs.

Afin de limiter le nombre de faux positifs, j'ai filtré les résultats par la méthode du *phylogenetic footprinting* grâce au module de contextualisation par conservation. Pour cela, j'ai recherché les motifs RAREs DR5 du poisson zèbre (*Danio rerio*) conservés dans les génomes de quatre poissons : l'épinoche (*Gasterosteus aculeatus*), le fugu (*Takifugu rubripes*), le médaka (*Oryzias latipes*) et le poisson-globe (*Tetraodon nigroviridis*). Ces poissons sont suffisamment éloignés du poisson zèbre pour diminuer les conservations dues au hasard.

La majeure partie des sites retrouvés sont en fait situés dans une région conservée sans pour autant être conservés eux-mêmes. Pour la suite de l'étude, j'ai gardé uniquement les sites présentant un « match exact » dans un souci de limiter au mieux le nombre de faux positifs. 29 sites sont conservés dans l'ensemble des 5 espèces de poissons considérées et constituent de très bons candidats RARE.

Pour plusieurs des *hits* conservés, j'ai également vérifié visuellement sur le *genome browser* de l'UCSC si la conservation est large (le site se retrouve par hasard dans une vaste région conservée) ou si la conservation est localisée précisément autour du site (forte probabilité d'importance de ce site dans la régulation de la transcription). Dans plusieurs cas, le site se trouve dans un îlot de conservation localisé, sa conservation est donc probablement liée à des contraintes biologiques et non au simple hasard.

En analysant les conservations avec un niveau de précision de l'ordre de la base, plusieurs cas de figure apparaissent.

La plupart des sites sont conservées au-delà des hémi-sites. Cela est compatible avec ce qu'on sait sur la reconnaissance du RARE par les RAR, lorsque l'extension C-terminale participe à la reconnaissance, en établissant des interactions avec des bases à côté du hémi-site.

Dans certains cas, comme le montre la Figure 21, la signature phylogénétique présente une conservation très nette, on remarque une absence de conservation localisée dans la région entre les deux hémi-sites, ainsi que la non-conservation des régions en amont et en aval.

Parmi les 29 sites conservés, j'ai sélectionné les sites situés dans un gène ou à proximité d'un gène (entre 10 000 pb en amont du TSS et 10 000 pb en aval de la fin du gène, intervalle comportant le plus de RARE selon (Lalevee et al., 2011)). 19 sites répondent aux critères et seraient potentiellement impliqués dans la régulation de 20 gènes codant pour des protéines (avec un total de 48 transcrits) et de 4 transcrits codant pour des micro-ARN.

J'ai analysé les 20 gènes cibles potentiels en utilisant le module de recherche d'enrichissement fonctionnel basé sur GoMiner ainsi que l'outil DAVID (Huang et al., 2007). L'analyse par GoMiner en sélectionnant le niveau 3, c'est-à-dire uniquement les annotations expérimentales, a mis en évidence un enrichissement significatif (FDR < 10^{-5}) en gènes impliqués dans le développement du rhombencéphale, une région du cerveau connue pour être régulée par l'acide rétinoïque (Linville et al., 2004).

Enfin, j'ai fait une analyse des publications relatives à ces 20 gènes, ainsi que des annotations GO qui les caractérisent. Au total, 10 des 20 gènes sont des facteurs de transcription ce qui est en accord avec ce que l'on sait sur RAR, connu pour être un facteur de transcription de haut niveau régulant d'autres facteurs de transcriptions (Rochette-Egly and Germain, 2009). Parmi les facteurs de transcription retrouvés, se trouvent les gènes *hox*, des gènes caractérisés par la présence d'une homéobox, codant pour des facteurs de transcription essentiels dans le développement de l'organisme. Il y a également des gènes au rôle plus direct comme le gène Cyp26a. Ce gène régule la concentration en acide rétinoïque de la cellule, en codant pour une protéine qui oxyde les molécules d'acide rétinoïque. Au total, 14 des 20 gènes protéiques identifiés possèdent un lien clairement établi attesté par une publication avec la régulation par l'acide rétinoïque dans au moins un organisme. Ce résultat valide l'intérêt de Parsec pour l'identification de gènes cibles de RAR. Les 6 autres gènes, Znf503, gria3b, zgc:162576, scap1, ophn1, et C13H10orf11 constituent des candidats potentiels.

1.5.1.1.3 Application de Parsec à l'étude de la régulation par l'acide rétinoïque de l'embryon du poisson zèbre

Les vertébrés possèdent trois sous-types de RARs, qui jouent à la fois un rôle redondant et spécifique. Afin de déterminer si les sous-types des RARs peuvent avoir une activité spécifique lorsqu'ils sont tous exprimés, l'Equipe de Zoologie Moléculaire de l'Université de Lyon a mené une étude des transcriptomes d'embryons du poisson zèbre en désactivant sélectivement des sous-types de RARs. L'approche expérimentale fut complémentée par une approche bio-informatique avec Parsec et ce, de deux manières.

D'une part, j'ai généré avec Parsec des listes de sites RARE (DR0 à DR10, IR0 (*Inverse Repeat*) à IR10) annotés par les gènes proximaux, avec ou sans filtrage par conservation. Pour cela, j'ai utilisé la flexibilité de Parsec permettant de choisir pour chaque étape de contextualisation si un filtrage doit être appliqué en plus de l'annotation. Ces résultats ont été envoyés à l'équipe de Lyon et ont permis d'identifier plusieurs nouveaux sites reconnus par le récepteur à l'acide rétinoïque dans la région promotrice des gènes *cyp26a1* et *rara*, dont la régulation par l'acide rétinoïque est très robuste et résiste à la désactivation de tous les RARs.

D'autre part, afin d'étudier si un sous-type de RAR avait une préférence particulière pour un DR ou un consensus RARE particulier, j'ai créé un script python pour grouper les sites identifiés par Parsec en fonction des listes de gènes envoyées par l'équipe de Lyon. Le script réalisait entre autre un Logo de la séquence consensus des différents hémi-sites (DR0 à DR10 et IR0 à IR10) à

proximité des gènes de chaque liste en utilisant le librairie python WebLogo (Crooks et al., 2004). Aucun enrichissement particulier en DR n'a malheureusement pu être identifié.

Cette étude a donné lieu à une publication (Annexe 2).

1.5.1.2 Application aux ciliopathies

Une des thématiques majeures de notre laboratoire sont les ciliopathies (maladies génétiques liées à des anomalies dans la formation ou fonction des cils cellulaires). Les cils sont des organites que l'on retrouve à la surface des cellules de beaucoup d'eucaryotes. Ils sont impliqués dans plusieurs fonctions essentielles comme la motilité, la reproduction, la signalisation ou les voies sensorielles. Prises dans leur ensemble, les ciliopathies sont des maladies graves qui concernent beaucoup de personnes (on estime qu'une personne sur 1000 est affectée), l'identification de gènes impliqués dans la genèse et le fonctionnement du cil est donc une tâche importante pour la recherche biomédicale. Plusieurs approches sont utilisées, comme les profils d'expression de cellules ciliées et non ciliées, les techniques de génomique comparative (organismes ciliés et non ciliés), ou la recherche de sites de fixation de facteurs de transcriptions liés au cil par des outils bio-informatiques. Parmi ces facteurs, le groupe de protéines RFX (*regulatory factor X*) est essentiel pour la ciliogénèse, et se fixe au site X-box de l'ADN afin d'exercer son activité de régulation de transcription. Le site X-box est un palindrome de 12 à 15 bases, composé de deux hémi-sites de 6 bases.

Julien Lision a effectué un stage dans notre laboratoire afin de rechercher et caractériser avec Parsec le site X-box GTYBYC NN GRMAAC (Piasecki et al., 2010) et les gènes qu'il régule potentiellement. Après recherche du site dégénéré dans le génome humain, filtrage par proximité aux gènes et par type des gènes (uniquement ceux codant pour des protéines), il a effectué un filtrage des sites par conservation dans le génome de la souris. Cela a produit comme résultat 663 gènes autour de 628 sites. Le module d'enrichissement fonctionnel a permis de mettre en évidence que les catégories GO les plus enrichies en gènes détectés (FDR inférieure à 10^{-5}) avaient un rapport direct avec le cil (*cilium assembly, cilium morphogenesis, cilium parts, cilium axoneme*, etc…). Parsec a donc permis de produire une liste de gènes potentiellement ciliaires qui servira dans nos études ultérieures.

1.5.1.3 Application aux barcodes

La régulation post-transcriptionelle de l'expression des gènes est en partie assurée par des miRNA et des siRNA. C'est le processus de RNA interférence, ou RNAi. Ces molécules se lient toutes les deux à un mRNA (entrainant dans le cas du siRNA et parfois dans le cas du miRNA sa dégradation par la machinerie cellulaire), et empêchant l'expression du gène. Cependant, alors que les siRNA exigent une complémentarité exacte des séquences, ce n'est pas le cas des miRNA qui peuvent donc réguler un plus grand nombre de gènes (Novina and Sharp, 2004).

Les siRNA sont de plus en plus utilisées à des fins thérapeutiques ou de recherche, permettant d'empêcher ou réduire l'expression de certains gènes. Le *design* de siRNA efficaces est une des difficultés des expériences de RNAi, et des méthodes bio-informatiques sont utilisées afin de créer des librairies de séquences respectant des contraintes de taille, de température de fusion, etc… (Devi, 2006)

J'ai été contacté par un étudiant post-doctoral travaillant au Commissariat à l'énergie atomique et aux énergies alternatives (CEA) qui cherchait à créer un générateur de courtes séquences nucléotidiques pour créer des librairies de siRNA. Il voulait utiliser Parsec pour filtrer les séquences générées pour ne garder que celles qui ne seraient pas présentes sur le génome de l'organisme d'intérêt. Pour cette collaboration, j'ai donc rajouté un accès API à Parsec (décrit plus haut), permettant d'y faire appel à partir d'un pipeline automatisé, ou même d'y accéder par AJAX dans un script JavaScript. J'ai également créé un client (Page HTML avec code JavaScript) permettant d'évaluer et d'utiliser ce service.

1.5.1.4 Infrastructures concurrentes

Afin d'évaluer la pertinence de Parsec, je l'ai comparé à plusieurs autres services web concurrents comme le montre le Tableau 15.

Tableau 15: Plusieurs concurrents de Parsec

Programme	Taille minimale du motif	Caractérisation des sites	Support d'un motif IUPAC dégénéré	Scénario personnalisé	Génome entier
PARSEC	5	Contexte génomique, évolutionnaire, fonctionnel	OUI	OUI	OUI
BLAT (Kent, 2002)	20 (25 pour une certitude de hit)	AUCUNE	NON	NON	OUI
TagScan (Iseli et al., 2007)	10	AUCUNE	OUI	NON	OUI
Great (McLean et al., 2010)	Ne cherche pas de motifs sur le génome	Contexte génomique et fonctionnel	NON	NON	NON
GOMO (MEME) (Bailey et al., 2009)	AUCUNE	Contexte fonctionnel	OUI	NON	NON
FIMO (MEME) (Bailey et al., 2009)	AUCUNE	AUCUNE	OUI	NON	OUI

On remarque que Parsec offre une grande flexibilité sur la taille et la nature du motif à chercher, permet une riche contextualisation des sites retrouvés et est le seul à offrir une exécution non linéaire et personnalisée.

1.5.1.5 Revue critique des choix effectués

J'ai utilisé Apache Tomcat pour développer Parsec car il fait partie des *framework*s de développement web les plus connus et les plus utilisés. Cela m'a permis d'apprendre beaucoup de choses sur la construction d'un service web, mais Tomcat ne propose aucune aide pour structurer convenablement le code, ce qui peut poser problème dans un grand projet. De plus, le temps de développement sur Tomcat est beaucoup plus long que sur un framework de plus haut niveau comme Play! ou Flask. J'ai tenu compte de ce fait lorsqu'il s'est agi de développer MyGeneFriends.

Par ailleurs, l'utilisation de positions complètement dégénérées (N) dans le motif soumis à Parsec ralentit beaucoup son exécution lorsque le nombre de N est grand, car cela oblige la fonction de parcours de l'arbre à explorer toutes les branches partant d'un nœud donné. Une région constituée par plusieurs N est généralement utilisée pour représenter l'espace entre deux sites. Pour optimiser les performances de Parsec, il faudrait donc détecter lorsque le nombre de N est supérieur à deux et traiter les motifs séparés par les N comme des sites cherchés indépendamment, puis filtrés par la distance qui les sépare.

Les mésappariements sont également un facteur qui ralentit beaucoup Parsec, car je génère toutes les combinaisons de motifs dégénérés correspondantes au nombre de mésappariements stipulés. Pour optimiser cet aspect, il faudrait traiter les mésappariements au moment du parcours de l'arbre en décrémentant un compteur à chaque fois qu'une progression dans l'arbre passe par un chemin ne correspondant pas à la position évaluée dans le motif.

2 OrthoInspector

La collaboration au développement de la nouvelle version (2.0) (Linard et al., 2015) (voir Annexe 1) d'un logiciel d'analyse d'orthologie au sein de notre laboratoire m'a permis de me sensibiliser plus avant aux problèmes de la visualisation interactive de l'information et de l'interaction entre experts humains et mégadonnées.

Nous allons d'abord voir le contexte biologique de cette application et les caractéristiques des mégadonnées liées à l'orthologie. Puis, nous verrons en détails les deux outils de visualisation sur lesquels j'ai travaillé pour l'interface graphique de l'application OrthoInspector.

2.1 Contexte

L'homologie est un concept très important en biologie car il désigne une origine évolutive commune entre deux caractéristiques ou deux gènes. Lorsqu'elle est appliquée aux gènes, l'homologie peut être divisée en deux sous-classes : la paralogie et l'orthologie (Fitch, 1970).

Sans entrer dans une description exhaustive de ces notions centrales, on peut rappeler que dans les deux cas, les gènes considérés descendent d'un même gène ancestral : dans le cas de la paralogie, le gène a été dupliqué tandis que dans le cas de l'orthologie, les deux gènes sont séparés par un évènement de spéciation. Le profil de conservation des paralogues et orthologues est sensiblement différent. Les orthologues conservent généralement la fonction qu'ils avaient dans l'organisme d'origine et auront tendance à être conservés tandis que les paralogues auront plus facilement tendance à diverger fonctionnellement. Parce qu'après spéciation, des gènes peuvent être perdus ou dupliqués, les relations d'orthologie entre organismes peuvent être complexes, et un gène n'aura pas forcément un seul orthologue dans un autre organisme (relation *OneToOne*) et on pourra observer des relations *OneToMany* et *ManyToMany* (Theissen, 2002). Enfin, parce qu'ils remplissent souvent des fonctions comparables entre espèces et auront tendance à être conservés, les gènes orthologues sont utilisés pour construire un arbre phylogénétique des espèces ou prédire la fonction d'un gène dans une espèce, si la fonction du ou des gènes orthologues dans d'autres espèces est connue. L'orthologie est également centrale en génomique comparative, par exemple pour établir des corrélations génotype-phénotype, c'est-à-dire identifier les gènes avec des orthologues uniquement dans des espèces présentant un phénotype d'intérêt. Les gènes avec le profil de présence/absence recherché seront potentiellement impliqués dans la caractéristique phénotypique.

Dans le cadre de ses recherches en génomique comparative, l'équipe a développé le logiciel OrthoInspector (Linard et al., 2011) qui améliore sensiblement la caractérisation des relations d'orthologie. Ce logiciel se base sur l'analyse du graphe des relations de similarité fournies par les alignements BLAST afin d'inférer les liens d'orthologie entre protéines. Rapidement, dans un premier temps, l'homologie est détectée en réalisant des comparaisons de séquences protéiques

tous-contre-tous entre les protéomes des espèces considérées (chaque séquence est comparée à toutes les autres) en utilisant le programme BLAST. BLAST associe à chaque alignement (appelé *hit*) entre la séquence soumise et les séquences de la banque, un score et un *expect*. Le score décrit la qualité de l'alignement réalisé, plus le score est élevé et plus les deux séquences alignées sont similaires. L'*expect* vise à évaluer l'alignement au regard du nombre de hits dus au hasard c'est-à-dire auxquels on pourrait s'attendre dans le contexte des séquences et compositions de la séquence soumise et des séquences de la banque. Plus l'*expect* est proche de zéro, et plus le *hit* est significatif. OrthoInspector infère et caractérise ensuite les relations d'orthologie (*OneToOne*, *OneToMany* et *ManyToMany*) en analysant les hits intra et interspécifiques. Le programme OrthoInspector est disponible sous forme de package permettant à l'utilisateur d'analyser les relations d'orthologie entre les protéomes de son choix (notamment ses propres protéomes). Les relations d'orthologie entre 259 eucaryotes et 1688 procaryotes sont par ailleurs pré-calculées et disponibles sur le site lbgi.fr/orthoinspector.

2.2 Intégration de l'expert

Les données pour les 259 organismes eucaryotes, présents à ce jour dans la base de données d'OrthoInspector, représentent presque quatre millions de séquences, plus de cent millions de relations *OneToOne*, des dizaines de millions de relations *OneToMany* et des millions de relations *ManyToMany*. Le nombre élevé de protéines et relations inter-protéines nécessite des outils efficaces pour offrir un accès clair, synthétique et ergonomique à l'utilisateur humain. Les deux outils sur lesquels j'ai travaillé visent à répondre à ces besoins. Pour cela, ces outils ont été centrés sur l'obtention d'une visualisation efficace et adaptée et sur l'interactivité avec cette visualisation.

2.2.1 Visualisation en diagrammes de Venn

L'orthologie permet à l'expert de comparer les répertoires de gènes de plusieurs organismes. Ce type de comparaison, quand elle est appliquée à trois espèces (ou souches), est typiquement représentée sous forme de diagramme de Venn permettant de visualiser très rapidement la proximité entre les organismes et leur spécificité.

Afin d'effectuer visuellement et de manière intuitive ce type d'analyse à la base de nombreuses applications en génomique comparative, j'ai privilégié l'emploi et l'adaptation des diagrammes de Venn et d'Euler afin de développer un outil de visualisation des orthologues multi-taches pour OrthoInspector.

A la base de cet outil, on trouve la librairie VennEuler (Wilkinson, 2012) qui permet de créer les diagrammes de Venn et d'Euler. Cette librairie est écrite en Java facilitant son intégration à OrthoInspector sous forme de JAR, mais cela permet également de modifier et compléter le code de cette librairie pour rendre la visualisation plus interactive et adaptée à notre problématique.

2.2.1.1 Adaptation et intégration

L'utilisateur commence à sélectionner trois organismes en mode classique (diagramme de Venn), ou un nombre personnalisé d'organismes en mode expert ou avancé (diagramme d'Euler). En mode classique, toutes les intersections entre organismes (relations d'orthologie) seront considérées, tandis qu'en mode expert, l'utilisateur est sollicité pour sélectionner dans une matrice les intersections qui l'intéressent.

En utilisant cette matrice de relations, je génère toutes les combinaisons possibles entre les organismes, puis je récupère grâce à une fonction d'OrthoInspector (*getValidatedProteins()* de la classe *OccurencesProteins*), le nombre de protéines orthologues communes à un ensemble d'organismes et absentes des autres. Je soustrais les protéines présentes dans les intersections pour obtenir le nombre de protéines spécifiques à chaque organisme, c'est-à-dire sans orthologue dans les espèces considérées. Pour associer à chaque intersection un poids relatif à son aire, je divise le nombre de protéines de l'intersection par la somme des protéines des organismes impliqués dans l'intersection.

J'ai également modifié la librairie VennEuler afin de permettre à l'utilisateur d'interagir avec les cercles représentant les protéines d'un organisme et les intersections de ces cercles. L'objectif est qu'en cliquant sur une intersection, l'utilisateur puisse télécharger toutes les protéines correspondantes. Pour cela, je détecte la position de la souris par rapport aux centres des cercles, ce qui fait qu'en tenant compte de leur diamètre je peux savoir quelle intersection d'organismes l'utilisateur est en train de sélectionner.

La représentation d'Euler n'étant pas parfaite, afin de vérifier la qualité du diagramme obtenu, j'ai également ajouté des fonctions (*getFalsePositives()* et *getFalseNegatives()*) permettant de détecter les faux positifs (intersections dessinées alors que non souhaitées) et les faux négatifs (intersections souhaitées mais non dessinées). Pour cela, je teste l'intersection de tous les cercles (distance entre les centres inférieure ou égale à la somme des rayons), et je compare la liste des intersections détectées aux intersections calculées.

2.2.1.2 Evaluation dans le cadre des ciliopathies

La caractérisation de gènes liés aux cils primaires et motiles est une étape essentielle dans la compréhension et la lutte contre les ciliopathies (comme expliqué dans la partie Parsec) et l'analyse soustractive est une des principales méthodes utilisées pour cela (Li et al., 2004).

L'analyse soustractive est souvent utilisée pour étudier le lien entre des groupes de gènes et des phénotypes. Pour cela, on doit disposer de l'ensemble des gènes des espèces impliquées (génome complet) et si possible, des génomes et protéomes déduits de qualité. Lorsqu'on s'intéresse à un phénotype commun à des espèces et absent d'autres espèces, on va, dans un premier temps, 'soustraire' (ne pas considérer) tous les gènes spécifiques à chaque espèce et obtenir une liste de gènes communs de laquelle on 'soustraira' les gènes qui sont communs aux organismes qui ne partagent pas le phénotype d'intérêt.

Appliqué au cil, l'outil que j'ai développé permet de sélectionner facilement les protéines communes à des organismes ciliés et absentes des organismes non ciliés (Figure 27).

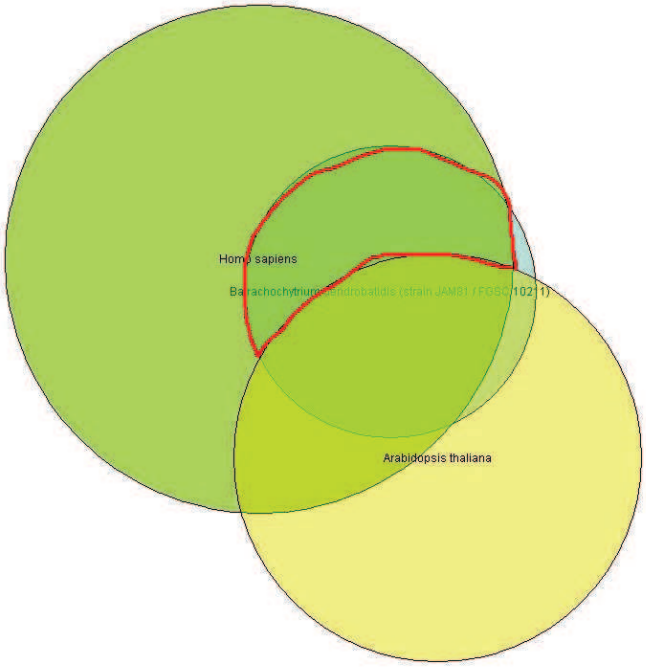


Figure 27: Diagrammes de Venn et analyse soustractive entre l'homme (cilié, vert clair), Batrachochytrium dendrobatidis (cilié, bleu) et Arabidopsis thaliana (non cilié, jaune). L'intersection contenant les protéines d'intérêt est entourée en rouge.

Cette analyse préliminaire permet d'isoler une liste de protéines potentiellement ciliaires, cependant elle comporte du bruit. En effet, dans le cadre de l'étude illustrée dans la figure 26, le cil n'est pas l'unique caractéristique commune à l'homme et *Batrachochytrium dendrobatidis* et absente d'*Arabidopsis thaliana*. Ainsi, l'homme et *B. dendrobatidis*, à la différence d'*A. thaliana*, sont également tous deux des opisthocontes impliquant que la liste de gènes obtenue ne permettra pas de discriminer les gènes spécifiques des ciliés de ceux, propres aux opisthocontes (Tableau 16). Les opisthocontes (du grec *opisthios* qui signifie postérieur, et de *kontos* qui désigne flagelle) sont un groupe d'eucaryotes réunissant champignons et métazoaires. Un élément essentiel commun à tous les opisthocontes est que leurs cellules flagellées possèdent un seul flagelle à l'arrière assurant la propulsion.

Tableau 16: Espèces utilisées dans l'analyse soustractive

Espèce	Description	Cilié ?	Opisthoconte ?
Homo sapiens	Un être humain (cilié)	Oui	Oui
Batrachochytrium dendrobatidis	Champignon parasite des amphibiens (cilié)	Oui	Oui
Arabidopsis thaliana	Arabette des dames, plante (non ciliée)	Non	Non
Saccharomyces cerevisiae	Champignon utilisé pour la fabrication de boissons fermentées (non cilié)	Non	Oui

Dès lors, afin d'écarter de notre liste, les gènes communs aux opisthocontes mais non liés aux cils, il suffit d'ajouter à l'analyse un organisme qui soit un opisthoconte non cilié, en l'occurrence *Saccharomyces cerevisiae (*Figure 28), la levure de bière.

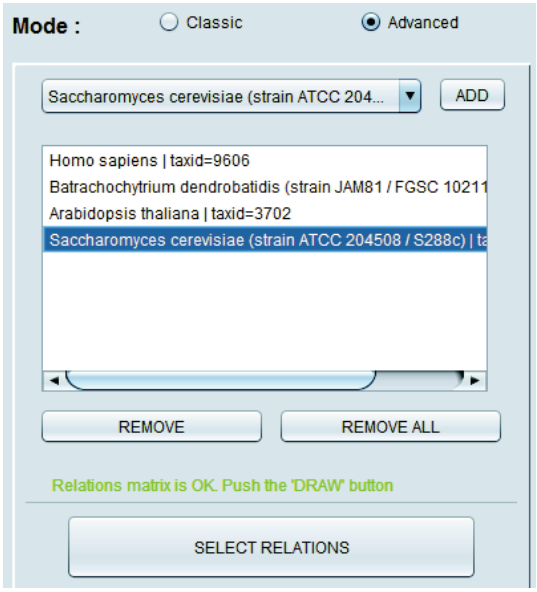


Figure 28: Illustration du mode avancé de l'outil de visualisation autorisant l'ajout de Saccharomyces cerevisiae à la liste initiale d'organismes.

Lorsque plus de trois organismes sont sélectionnés, toutes les intersections entre les cercles ne pouvant être représentées, il faut remplir une matrice pour préciser les intersections que l'utilisateur souhaite conserver dans la visualisation (Figure 29). Trop peu d'intersections sélectionnées peuvent aboutir à des faux négatifs (intersection souhaitée mais qui n'a pu être représentée pour des raisons géométriques) et trop d'intersections sélectionnées peuvent aboutir à des faux positifs.

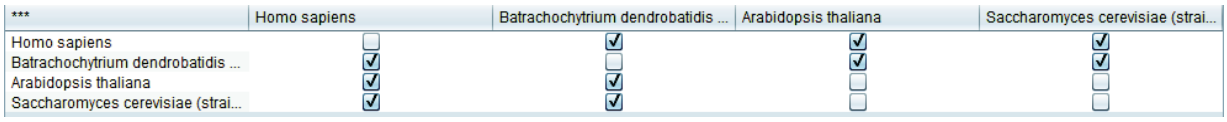


Figure 29: Lorsque l'utilisateur choisit plus de trois organismes, toutes les intersections entre les cercles ne peuvent pas être représentées. Il convient donc à l'utilisateur de sélectionner les intersections qui l'intéressent, en cochant les cases sur une matrice représentant toutes les intersections possibles entre les organismes.

Cette fois, la visualisation révèle une intersection plus réduite (entourée en rouge sur la figure) (Figure 30) sur laquelle il suffit de cliquer pour télécharger les protéines correspondantes (on passe des 2696 protéines de l'étape précédente à 1992 protéines). Il est également possible de continuer à affiner la liste de gènes en ajouter des organismes supplémentaires.

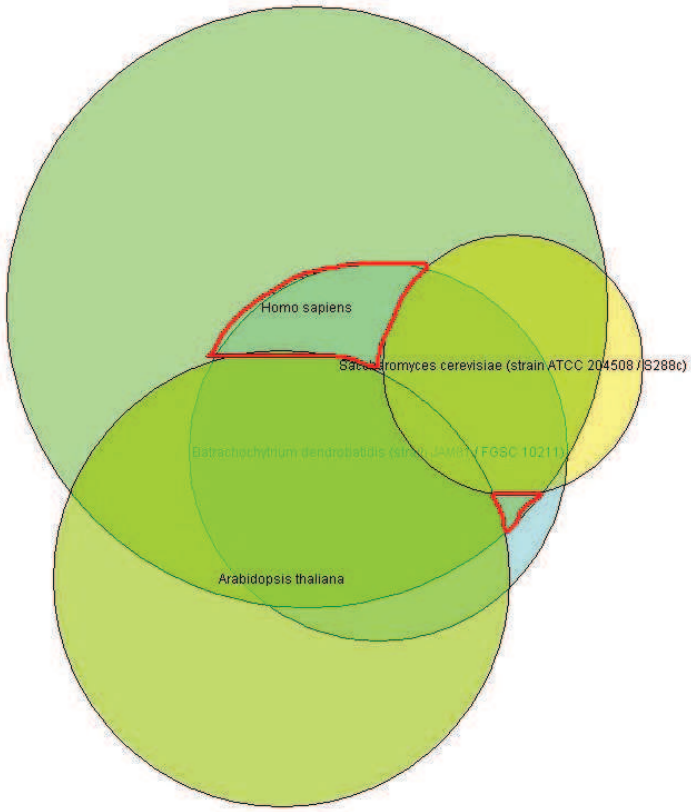


Figure 30: Résultat obtenu après l'ajout, comme organisme de filtre supplémentaire, de Saccharomyces cerevisiae. L'intersection correspondant au profil phylogénétique d'intérêt est entourée en rouge.

2.2.2 Visualisation du réseau de relations d'homologie

Afin d'améliorer la visualisation des relations d'homologie au sein d'une famille ou super-famille de protéines, j'ai également travaillé sur un outil permettant l'affichage sous forme de réseau des protéines reliés par des relations d'orthologie. Cet outil est basé sur une librairie de visualisation java très utilisée, Prefuse (Heer et al., 2005), qui permet de générer des graphes dynamiques (propriétés modifiables en temps réel) et interactifs (possibilité de sélectionner et déplacer des nœuds). De plus, celle librairie gère une simulation physique permettant la formation dynamique de groupes de nœuds. Cette simulation, utilisant une constante de gravité variable à façon, est basée sur deux principes ; les nœuds se repoussent et les arrêtes fonctionnent comme des ressorts (elles ont une longueur et un coefficient d'élasticité).

2.2.2.1 Adaptation et intégration

Les protéines proches (le score est grand ou *l'expect* est petit) doivent apparaître plus proches sur le graphe. Pour cela, avant que la visualisation ne soit lancée, je calcule les valeurs de longueur et largeur des différentes arêtes que j'enregistre dans des tableaux. Les quatre valeurs que je calcule sont E_{SL} (longueur en fonction du score), E_{EL} (longueur en fonction de l'exposant de l'*expect*, en fait le log en base 10), E_{SW} (largeur en fonction du score), E_{EW} (largeur en fonction de l'exposant de l'*expect*). Ces calculs reposent sur des valeurs prédéfinies comme les longueurs (EL) et largeurs (EW) minimales et maximales des arrêtes (EL_{min}, EL_{max}, EW_{min}, EW_{max}) mais également des valeurs minimales et maximales calculées du score et de l'exposant de l'*expect* (S_{max}, S_{min}, EE_{max}, EE_{min}).

$$E_{sL} = EL_{min} + (EL_{max} - EL_{min}) \times \frac{\frac{1}{score} - \frac{1}{S_{max}}}{\frac{1}{S_{min}} - \frac{1}{S_{max}}}$$

$$E_{EL} = EL_{min} + (EL_{max} - EL_{min}) \times \frac{expect\_exponent - EE_{min}}{EE_{max} - EE_{min}}$$

$$E_{sW} = EW_{min} + (EW_{max} - EW_{min}) \times \frac{\frac{1}{score} - \frac{1}{S_{max}}}{\frac{1}{S_{min}} - \frac{1}{S_{max}}}$$

$$E_{EW} = EW_{min} + (EW_{max} - EW_{min}) \times \frac{expect\_exponent - EE_{min}}{EE_{max} - EE_{min}}$$

La largeur de l'arrête est un indicateur visuel, alors que sa longueur sera utilisée par le moteur de simulation physique pour ajouter une contrainte sur la proximité des organismes (nœuds).

2.2.2.2 Isolation de sous-groupes de protéines

L'utilisateur peut choisir si c'est le score ou l'*expect* associé au meilleur *hit* blast qui doit influencer la longueur de l'arrête qui relie les séquences protéiques orthologues appartenant aux différents organismes (Figure 31). Cette visualisation permet, en manipulant en temps réel, le seuil d'affichage des liens, de découvrir les sous-groupes de protéines en fonction du profil de conservation des différents gènes et de déterminer de manière intuitive les valeurs de score ou d'*expect* séparant des sous-familles taxonomiques ou fonctionnelles.

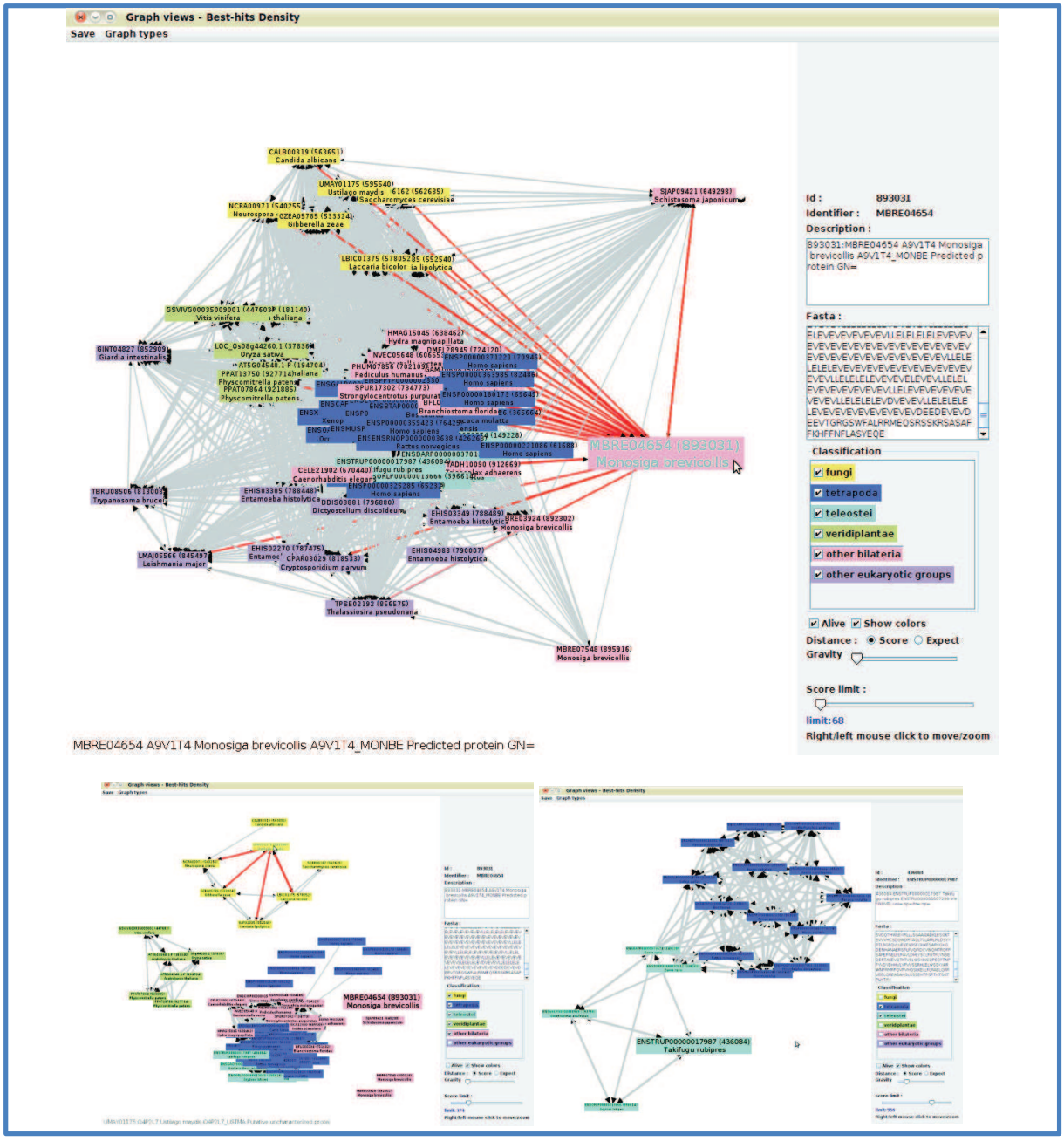


Figure 31: Visualisation des liens d'orthologie entre différentes espèces. Les noms d'espèces sont colorés en fonction du groupe taxonomique auquel elles appartiennent. On peut voir que les protéines d'un groupe taxonomique sont souvent plus similaires (arête plus courte entre les nœuds) comparées à celles des autres groupes.

2.3 Expérience acquise

Le travail sur ces deux outils de visualisation, réalisé dans le cadre du développement d'une nouvelle version d'OrthoInspector, m'a fait réellement prendre conscience des implications de la taille des données produites en biologie et de la nécessité de les exposer clairement et visuellement.

Lorsqu'il s'agit de visualiser les données, trois solutions peuvent être envisagées :

- Pré-calculer toutes les visualisations possibles et les stocker sous forme d'image,
- Créer une image à la volée et l'afficher,
- Dessiner en temps réel la visualisation et permettre l'interactivité (visualisation interactive).

Les première et deuxième solutions ont l'avantage de la rapidité, mais sont statiques, ce qui empêche un retour utilisateur et une manipulation des données intuitive et en temps réel. J'ai privilégié la troisième approche, plus gourmande en ressources, mais offrant plus de possibilités aux utilisateurs notamment.

Ce travail m'a fait comprendre deux choses essentielles. D'une part, la visualisation interactive permet la manipulation de grandes quantités de données sans avoir à montrer les données brutes, mais surtout, elle peut faire ressortir de nouvelles informations visuelles (*i.e. via* l'utilisation de contraintes de simulation physique) voire, produire de l'information numérique (*i.e.* détection d'une limite de score ou d'*expect* optimale).

D'autre part, j'ai pris conscience de l'importance de l'encapsulation des données (qui sera présentée en détails dans le chapitre 3.1.1), concept qui a été poussé beaucoup plus loin dans MyGeneFriends. Cette encapsulation transparait déjà dans les diagrammes de Venn et d'Euler, car ils représentent simplement par un cercle un organisme avec son répertoire de protéines. La manipulation et visualisation de ces cercles est bien plus intuitive que ne serait la manipulation de listes de protéines.

Toutefois, il est à noter que d'autres niveaux de complexités existent ainsi, la création de visualisations interactives n'obéit pas aux mêmes règles, ni aux mêmes contraintes, dans une application avec interface graphique et sur un site web.

Ces réflexions combinées à l'expérience acquise lors des développements des scénarios d'analyse personnalisée de Parsec, allaient me conduire à développer MyGeneFriends.

3 MyGeneFriends

Après m'être intéressé à la contextualisation de l'information génomique en donnant à l'expert la liberté de créer des scénarios d'analyse non linéaires au sein de Parsec, puis en découvrant et employant la puissance de la visualisation interactive de données en ajoutant deux nouvelles visualisations à OrthoInspector, j'ai décidé de réunir l'expérience acquise dans ces deux projets pour aller encore plus loin dans l'amélioration du rapport entre information biologique et biologiste en créant MyGeneFriends, un réseau social interconnectant gènes, maladies et chercheurs. Ce projet qui représente une grande partie de mon travail de thèse a nécessité une réflexion profonde et constante non seulement, sur des choix pratiques et solutions informatiques à adapter ou à développer à façon mais aussi, et surtout, sur les parallèles à établir entre les concepts, éléments et fonctionnalités disponibles dans des réseaux sociaux du type Facebook et ceux d'un réseau social dédié à la biologie, et notamment aux maladies génétiques. MyGeneFriends, dont la publication est présentée dans le chapitre 3.2, est le fruit de ces multiples réflexions et choix dont les plus emblématiques seront présentés succinctement dans les chapitres suivants.

Ainsi, la notion d'acteur et le choix des acteurs à inclure dans MyGeneFriends fut un premier défi. Après cette entrée en matière, viendra le manuscrit qui présentera la plupart des aspects essentiels de MyGeneFriends. Je parlerais ensuite des choix techniques permettant d'intégrer des données complexes et d'assurer la pérennité de MyGeneFriends. Puis je présenterais le flux d'informations personnalisé que j'ai mis en place, et les démarches pour faire de l'expert une partie intégrante du réseau MyGeneFriends. Finalement, je montrerais que l'archive de News produites par MyGeneFriends depuis presque un an permet de ressortir des tendances décrivant l'évolution de notre état de connaissances sur les entités biologiques.

3.1 Les acteurs

Dans le contexte de la médecine personnalisée et du flux de données individuelles que nous devons gérer (Fernald et al., 2011), la notion de point de vue ou de prisme deviendra vitale. Nous ne voudrons plus simplement consulter la fiche généraliste d'un gène par exemple, mais plutôt, accéder aux informations de ce gène via son comportement chez un patient ou une cohorte de patients ou via son implication dans un projet de recherche. Pour ce faire, MyGeneFriends introduit la notion d'acteur : une information biologique active et encapsulée, s'adaptant à la personne qui interagit avec elle. Un acteur est un membre du réseau bien défini qui va posséder des propriétés spécifiques et des comportements propres et qui sera relié à d'autres acteurs de même type ou de type différent.

3.1.1 Qu'est-ce que l'encapsulation ?

En informatique, l'encapsulation est l'idée de cacher le contenu d'un objet. Cela permet de dissimuler la complexité d'une implémentation en n'exposant que des méthodes simples de manipulation de notre objet. L'utilisateur n'a alors plus à se soucier de son fonctionnement interne, seulement à l'utiliser.

Si l'on voulait faire un parallèle osé, on pourrait dire que les êtres humains sont également encapsulés. Nous n'avons pas accès à l'ensemble des données d'un être humain, de plus, notre perception (connaissance) sera influencée par notre relation avec l'autre personne et par ce que l'autre personne sait de nous. L'échange d'informations est contextualisé et adaptatif ainsi, un enfant ne recevra pas la même réponse à une question, qu'un adulte par exemple, un spécialiste du BBS pas la même réponse qu'un chercheur travaillant sur le virus de la rage.

Par le biais de l'encapsulation, l'interaction avec les mégadonnées et l'information biologique en général doit être aussi naturelle que l'interaction avec un autre être humain.

3.1.2 Le choix des acteurs

Le choix des acteurs fut essentiel pour la mise en place de MyGeneFriends. Ce réseau social, orienté maladies génétiques, devait rester simple et pour cela ne pas être submergé de différents types d'acteurs. D'autre part, il fallait prendre en compte la facilité pour les utilisateurs d'interagir avec un concept biologique comme avec une entité. Peu de personnes trouveraient par exemple naturel de devenir amie avec une interaction protéine-protéine ou avec l'expression d'un gène dans un tissu.

Je me suis donc limité à trois acteurs majeurs incontournables en biologie et en médecine moderne :

- ✓ **Le gène** : Entités d'information génétique discrètes définis dès les débuts de la génétique, avant même que leur base moléculaire ne soit connue. De nombreuses ressources bio-informatiques articulent leurs informations autour de fiches dédiées aux différents gènes (Ensembl, NCBI, GeneCards, etc…). Bien que beaucoup d'expériences concernent des entités plus spécifiques encore, les transcrits par exemple, une large part de la littérature scientifique raisonne et présente les découvertes autour du gène et de ses produits.

- ✓ **La maladie** : dans MyGeneFriends, les maladies génétiques humaines, permettent de lier plus directement les gènes et leurs variations à la société humaine (les patients) et aux recherches associées. On peut noter que d'autres types de maladies existent telles que : les maladies infectieuses qui impliquent de considérer des réseaux découlant des rapports hôtes-pathogènes ou les cancers qui peuvent dépendre de paramètres génétiques, mais qui,

pour l'essentiel, découlent de facteurs environnementaux ou comportementaux difficiles à saisir. Nous nous sommes concentrés sur les maladies génétiques d'une part, par souci d'efficacité et de simplification et d'autre part, pour pouvoir profiter des connaissances accumulées depuis plusieurs années par le laboratoire et nos collaborateurs (Hélène Dollfus, José Sahel, Thierry Lévéillard…) dans le domaine des maladies génétiques rares.

- ✓ **L'humain** : L'humain est évidemment l'utilisateur principal de MyGeneFriends, mais il est également un acteur (actif donc, et non un observateur passif) de notre réseau social, qui crée des interactions et amitiés avec les autres acteurs et influence le réseau. L'homme est donc entièrement intégré à notre système qui va essayer de le comprendre (pour mieux l'aider), de le décrire (pour attirer à lui d'autres acteurs, humains compris), et d'extraire de ses comportements publics des informations d'intérêt biologique (nouveaux liens entre gènes ou entre maladies par exemple).

3.2 Manuscrit

MY GENE FRIENDS:
A social network linking genes, genetic diseases and researchers

MY GENE FRIENDS: A SOCIAL NETWORK LINKING GENES, GENETIC DISEASES AND RESEARCHERS

Alexis Allot[1], Kirsley Chennen[1], Yannis Nevers[1], Alexia Rohmer[1], Laetitia Poidevin[1], Arnaud Kress[1], Raymond Ripp[1], Julie Thompson[1], Olivier Poch[1] and Odile Lecompte[1]

[1]Computer Science Department, ICube, UMR 7357, 4 rue Kirschleger, 67085 Strasbourg, France

Correspondence should be addressed to Odile Lecompte (odile.lecompte@unistra.fr)

Running title: Socializing genes, genetic diseases, researchers

Keywords: network, social, genomics, suggestions, web framework

ABSTRACT

The constant and massive increase of biological data offers unprecedented opportunities to decipher the function and evolution of genes and their roles in human diseases. However, the multiplicity of sources and the huge data flow mean that efficient access to useful information and knowledge production has become a major challenge.

Here, we present MyGeneFriends, a social network that allows three types of actors – genes, genetic diseases and humans – to interact. The goal is to optimize current research processes by leveraging conventions and practices arising from popular social networks like Facebook, in order to provide intuitive retrieval, annotation and exploration of data related to genes (expression, localization, etc…), diseases (phenotypes, variations, etc…) and researchers (collaborators, interests, etc…).

We introduce the notion of friendships to represent proximity between actors, who can then manage these friendships, view new friendship suggestions, and follow the activity of their friends (a gene befriends a new disease, personal information for a disease is updated...). Human related friendships (public, collaborative or private) can be organized into "Topics" allowing users to specify active research interests, to collaborate with other researchers within a specific field, and to add genes, diseases, keywords, etc… on which automated analyses can be performed and visualized.

MyGeneFriends ensured a daily mining of public databases, friendships and Topics to suggest new friends and pertinent publications, and dynamically adapt information visualization and actors textual descriptions and friendships.

MyGeneFriends can be accessed at: lbgi.fr/mygenefriends

INTRODUCTION

Biology is evolving and adapting at a tremendous rate, particularly in response to the widespread use of high throughput methods and the rise of personal genomics. As in many other fields, biology is now facing major challenges related to the management and analysis of the 'Big Data' produced by these technologies. Indeed, Big Data has been described by 4 dimensions (Bellazzi 2014), known as the 4V: Volume, Variety, Velocity, Veracity, that complicate any analysis by the mass of data to consider (Volume), the heterogeneity (Variety) of the data in terms of resources or formats, the evolution of these data with continuous updates (Velocity), and the large number of incomplete, untraceable or false data (Veracity). As a consequence, new data storage platforms and workflow infrastructures are being developed to deal with the ever growing data volume and flow. In addition, important efforts have been devoted to the creation of standards and ontologies to structure and formalize the data, which includes both structured (sequences, relational databases, csv files) and unstructured formats (text, figures, MRI images, etc.). These ontologies facilitate the integration and curation of data from different resources, resulting in a reduction of the noise inherent to high throughput biology (Prosdocimi et al. 2012), and an increase in the accuracy of analysis results.

Nevertheless, for the end-user of biological data, the paradigm shift initiated by the emergence of Big Data (Nielsen 2009) goes beyond the 4 ‘V’. Big data is changing the way we think about research and the means we use to deal with information (Kitchin 2014). Indeed, it is becoming challenging to keep up with all the available information, even in a specialized research field. In today’s data-intensive and highly connected context, it is essential for users to easily and intuitively access information of personal interest, to communicate and network around it. A number of solutions have been developed aimed at improving information personalization in our daily lives, including targeted advertising by Google, or friendship suggestions from social networks such as Linkedin and Facebook. In biology too, creating new ways to efficiently access useful and specific information has become a major challenge for research (Hofker et al. 2014) and several trends are starting to emerge in the bioinformatics domain. Notably, a number of traditional resources have introduced new personalized data flow management tools for their users. For example, the authors of the Online Mendelian Inheritance in Man (OMIM (Hamosh et al. 2000)) resource have created MIMmatch (Amberger et al. 2015), a service allowing to users to receive email notifications when entries for selected genes or diseases have changed. MyNCBI (Giglia 2011) can be used to create lists of genes or publications, receive email notifications, etc. The Uniprot (Apweiler et al. 2004) website has been updated to allow users to select only categories they are interested in, to mask large scale publications, and use a basket to store proteins of interest.

Various tools aimed at efficiently managing and exploiting the increasing flow of publications have also been developed. Bibsonomy (Benz et al. 2009) allows a researcher to collect and manage publications and collaborate with colleagues, while PubChase (www.pubchase.com) and

ReadCube (www.readcube.com) recommend new publications given the content of an existing library. BioTextQuest (Papanikolaou et al. 2014) is a tool that clusters related documents, resulting from a keyword search in PubMed (Giglia and Spinelli 2009) or OMIM, with four different clustering algorithms. GoPubMed (Doms and Schroeder 2005) allows more intelligent publication searches by using background knowledge in the form of ontologies (GO, MeSH, etc.), and thus not only considering the user's keywords, but also synonyms and child concepts. Medie (www.nactem.ac.uk/medie/) allows users to formulate simple questions in order to search for abstracts like *"What activates p53?"* or *"What causes colon cancer?"*. Finally, FACTA+ (Tsuruoka et al. 2008) returns concepts associated with the user's keywords in different publications, Pubtator (Wei et al. 2013) automatically identifies specific bioconcepts (Disease, Species, Mutation, Chemical, Gene) in publication abstracts, EuropePMC (Consortium 2014) searches multiple resources including PubMed and identifies concepts such as Disease, Organism, GO (Gene Ontology) and EFO (Experimental Factor Ontology) terms, and PAML-IST (Mandloi and Chakrabarti 2015) adds co-occurrence networks for genes, drugs, diseases and authors. iHOP (Hoffmann and Valencia 2004) on the other hand provides a list of most recent publications linked to a gene of interest, together with detection of related biomedical concepts.

Other tools aim to perform a more complex task of relationship extraction between entities. Chilibot (Chen and Sharp 2004) for example searches interactive (stimulation, inhibition, etc.) or parallel (studied together, co-existence, homology, etc.) relationships between user-submitted genes or proteins. EvexDB (Van Landeghem et al. 2013) extracts specific events: regulatory control, coregulation or binding to a given gene. After identifying a gene or list of genes of interest, GeneMania (Khalid et al. 2013) and GenesLikeMe (Stelzer et al. 2009) identify and score related genes that may also interest the user, based on ontologies, disorders, compounds, phenotypes, expression, domains, sequence, and other data.

Once information is retrieved and integrated in a database, it has to be displayed using efficient visualization techniques, which are becoming more and more common for the easy communication of complex data. They facilitate understanding of information updates, the links between entities and groups of entities, metadata information such as source, confidence, etc. For example, the ExAC browser (exac.broadinstitute.org) provides clear visualization of variations in a gene, javascript libraries like BioJS (Gonzalez-Alcaide et al. 2013) provide reusable components for visualization of biological information (3D structures, phylogenetic trees, proteomes, pathways, multiple sequence alignments), contributed by users in a registry. Coremine (www.coremine.com) allows exploration of various biomedical concepts and the connections between them, addition of private or public comments, alerts on new articles or connections and bookmarking.

Finally, some bioinformatics resources have a collaborative and social component, with wiki approaches like Proteopedia (Hodis et al. 2010) or WikiGene (Hoffmann 2008), collaborative sequence annotations such as WebApollo (Lee et al. 2013) or voting for medical relevance and scientific evidence of variations with GeneTalk (Kamphans and Krawitz 2012).

To further enhance the relationship between researchers and Big Data, and exploit the flood of information related to human genes and human genetic diseases, we have developed MyGeneFriends, a social network aimed at leveraging the conventions and practices arising from popular networks such as Facebook (like, wall, profile, friendships, friendship suggestions, affinity score, etc.) in order to provide more intuitive interactions with biological information, to simplify access to complex information by organizing it around three actors (Genes, Humans, Diseases) of a social network, and to extract additional knowledge from the resulting network. MyGeneFriends can be used to manage, suggest, visualize actors in their network context, and personalize actor-related information for the human expert. It can be accessed at: lbgi.fr/mygenefriends

RESULTS

MyGeneFriends is a social network interconnecting three kinds of actors: Genes, Humans and Genetic diseases. It provides researchers and clinicians with a more natural way to explore genetic and biomedical data. It exploits conventions from popular social networks to facilitate data navigation and networking. These conventions include a profile page to learn about an actor, friendships to connect to other actors and navigate through connections between actors, friendship suggestions to help find new relevant friends, affinity scores for actor profiles and a "wall" to display news about a user's friends.

After a brief overview of the MyGeneFriends platform, the presentation of the different types of actors and their profile pages will highlight the basics of the platform, and friendships between these actors will illustrate the underlying network. The concept of "Topics" in MyGeneFriends introduces new connections into this network and personalizes information for the human user. Finally, because MyGeneFriends is aimed at a large and broad audience, some examples of possible usages by human experts will be described in detail.

Overview of the platform

Organizing biological information around actors in a social network simplifies, standardizes and personalizes access to specific information. This is illustrated by several points. First, all actors belong to the same biosocial network and are linked by millions of friendships that are used to generate new knowledge, by extracting emergent properties from node interconnections such as automated annotation, suggestion, or clustering. Second, the same architecture is used for the profile pages of all actors, regardless of their type. Third, the same search bar allows access to any actor on the network: humans (by name), genes (by symbol, synonym or fragment of protein sequence), genetic diseases (by words in name, in any order), and also public Topics (by name).

Nevertheless, some aspects of MyGeneFriends are specific to human actors. Humans can sign up for MyGeneFriends, find new friends using the *friends of friends* network or the search bar, and create Topics to express their current research interests (see below). In addition, humans can

initiate new friendships, while friendships between non-human actors result from dynamic data mining processes. Finally, all data related to genes and diseases are public, while data submitted by humans to MyGeneFriends is 'private' (visible only by the owner) by default, unless the human decides to make it 'protected' (visible by owner and selected collaborators) or 'public' (visible to anyone). Similarly, friendships with genes or diseases can be private, protected or public, although they are public by default in order to encourage networking. This data privacy management (Jee and Kim 2013) is essential in our highly interconnected world in order to keep essential data private, while being open enough to "attract" new information.

Non-human actors (genes and genetic diseases) are automatically updated using a daily data mining and integration process (**Figure 1**). This involves retrieving, cleaning, adapting, comparing and integrating data from various asynchronous public data sources (Ensembl, NCBI, Uniprot, HPO, OMIM, etc.) as well as data produced in-house (Orthoinspector, CilioCarta). This process created more than 300 000 news events during the beta-testing phase of the platform, which lasted 6 months. The database underlying MyGeneFriends consists mainly of tables related to actors (genes, humans, diseases) and friendships between these actors. At the time of writing, more than 63 000 human coding and non-coding genes have been integrated from Ensembl, and more than 14 000 diseases have been integrated from OMIM and Orphanet.

As for any social network, the best place to learn about members/actors is by viewing their profiles (**Figure 2**). The different actors in MyGeneFriends share the same profile organization to simplify code maintenance and addition of new features, but also to facilitate intuitive navigation through the network. An actor's profile is meant to provide simple and personalized information about a gene, human or disease. The top keywords describing the actor, a visualization panel (including *word clouds* to highlight specific information for genes and diseases), and a news feed related to the actor allow rapid access to the main information. The links to public friends (genes, diseases and humans) on MyGeneFriends and to external databases place the actor in a larger network context.

The general display has been adapted to the human user. First, humans can expand the official description of a non-human actor by adding their own private annotations, which can then be accessed at any time or shared with collaborators. Second, keywords inferred as important for the human are highlighted in the description. For example, if the user is friends with *cilia* related genes and another human, gene or disease has the word *cilia* in the description, the word *cilia* will be highlighted. Finally, an affinity score is shown on gene and disease profiles to reflect the proximity between the actor and the content of user's active Topic (see below), and thus the interest of befriending this actor.

Gene profile

The Gene profile allows fast gene identification with different identifiers and provides an evolutionary tag reflecting the phylogenetic distribution of the gene according to its presence in seven pre-defined clusters: Universal, Metazoa, Deuterostomia, Vertebrate, Primate, Opisthokonta, Ciliate. The visualization panel illustrates different aspects of genes. First, it presents gene transcripts with their properties (sequence, biotype, reliability, corresponding protein sequence if any). Users can compare their own sequences with the different isoforms by performing a pairwise alignment on the fly. Second, gene expression is shown in more than 40 tissues as a *heatmap* through an interactive schematic view of the human body (male and female), brain and fetus. Pan and zoom capabilities (jquery.panzoom.js) allow users to navigate through the schematic view and visualize even the smallest tissues. Additional information, such as tissue descriptions and probeset signal intensities are provided. In addition to expression visualization, the "Expression filter" tool allows users to find genes of interest based on their expression or lack of expression in a defined set of tissues.

For protein-coding genes, a third tab of the visualization panel shows subcellular localization(s) of the encoded proteins as a wordcloud based on the frequencies of GO terms (2015). The emphasis is put on the specificity (rareness) of the cellular component to provide the relevant information at a glance. Finally, publications associated with the gene are listed, allowing access to additional information, such as a personalized abstract with words highlighted according to the user's profile, links to PubMed and the possibility to Like/Dislike the publication. For each publication, the number of related genes is provided to estimate the relevance of the publication for the considered gene. Moreover, genes related to a publication can be visualized as a network of established friendships, thus facilitating further networking and identification of supplementary genes of interest.

Disease profile

MyGeneFriends is currently focused on human heritable diseases, as they represent major clinical challenges and provide a simplified context to shed light on common major diseases. In addition to the general description retrieved from OMIM, two main features have been selected to characterize a disease on the visualization panel: variations explaining the causes of a disease and phenotypes describing its consequences. The description of variants is generated by the integration of more than 100 000 ClinVar (Landrum et al. 2014) curated variations directly linked to diseases by OMIM (Amberger et al. 2015), GeneReviews (http://www.ncbi.nlm.nih.gov/books/NBK1116/), dbSNP (Sherry et al. 2001) or submitters, as well as the effects of these variants that can vary depending on the given transcript (more than a million effects on different transcripts detected by VEP (McLaren et al. 2010)). To describe the complex relationships between variants, transcripts and disease-causing genes, we have developed three synoptic and interactive views with variants grouped according to the affected gene. For each gene associated with the disease, the first and second views show the relative frequency of variants classified according to their location in gene elements (UPSTREAM, 5'UTR, CDS, INTRON, Splicing region, 3'UTR, DOWNSTREAM) or their effect on proteins

(SYNONYMOUS, MISSENSE, FRAMESHIFT, START, STOP, INFRAME). This representation by gene and color coded to show the relative density of each group constitutes a 2D barcode (**Figure 3**) that allows easy comparison of ClinVar variations linked to a disease through the structure of the affected genes or the effects on protein sequences. For some genes, the variations are concentrated in a specific region (e.g. the CDS), while for others the variations are distributed over the entire gene and upstream/downstream regions. In addition to the distribution of variations along a gene, the barcode allows the user to compare the distribution for all affected genes. Users can select a specific category of variants related to a disease (for instance all non-sense variants) and obtain a detailed view of the variants with their transcripts and effects. The third view focuses on differential variant effects by distinguishing the transcripts affected or not affected by a given variant and specifically displaying the unaffected protein coding transcripts.

The heterogeneity of many disease phenotypes means that the definition of a disease can vary from one data source to another. Sources can have different names for the same disease ("Visceral myopathy" on OMIM (Hamosh et al. 2000) corresponds to "Familial visceral myopathy" on Orphanet (Ayme 2003)) or even a different point of view on what constitutes a single disease (one entry for "Bardet-Biedl syndrome" in Orphanet corresponds to multiple entries in OMIM, for each Bardet Biedl subtype (BBS1, BBS2, etc.)). To take into account this complexity, we group closely related diseases into a "metadisease", which can be viewed in the form of a network. To better define this newly created entity, we sort the genes and phenotype information related to the grouped diseases, in order to highlight the phenotypes and genes related to many diseases and thus identify the most representative for each metadisease. There are approximately 14000 genetic diseases present in MyGeneFriends, and 3418 of them are organized into 725 metadiseases.

Human profile

The last actor in MyGeneFriends is the human user, whose profile page contains information provided by the owner: a short description, his affiliation and geographic localization. The user is also asked to provide a publication list (PMID identifiers). MyGeneFriends retrieves the publications from PubMed and integrates them into the database, creating links between the human and the publications.

If users do not provide a description, an automatic process extracts the main keywords associated with their public gene and disease friends and displays them on the profile, in order to introduce them to other humans.

The private part of the profile page is dedicated to managing the user's research interests, represented by the concept of Topics (see below). A list of Topics created by the user is displayed in the "My Topics" section. One of these Topics is selected as 'active' by the user and is used for personalization and automated suggestion processes. In addition to their own Topics, users have access to Topics shared by their human friends in the "My Collaborations" section.

Friendships and networking

The creation and management of new friendships and networking through existing friendships are at the core of MyGeneFriends processes.

Humans can become friends with genes and diseases by a simple click. These friendships can be 'private' (visible only by the user), 'protected' (visible to the human and collaborators he selected) or 'public' (visible to anyone). They can also send a request to another human member of MyGeneFriends to become friends or send an email invitation to a non-member.

Other friendships in the network (gene-disease, gene-gene and disease-disease) are generated automatically based on data mining and relationships. Genes and diseases are linked by the implication of a gene in a disease: these relationships are integrated from HPO and based on data from OMIM and Orphanet.

Friendships between diseases are primarily based on shared phenotypes. For each phenotype-based friend, the number (and a list) of shared phenotypes, and the number of phenotypes specific to each of the two friends are indicated. Considering the heterogeneity of disease definitions, this information is useful for assessing the relatedness and possible overlap between two diseases. In order to relate even more different diseases, we include medical (shared genes) and social friendship (publicly shared humans) information.

For friendships between genes, six complementary views are displayed, comparing functional (Shared GO terms, FFS on GO terms), medical (shared diseases), interaction (STRING database (Szklarczyk et al. 2015)), evolutionary (similar evolution profiles) and social (publicly shared humans) aspects.

An interactive graph view with repulsion physics (visjs.org) is used for intuitive visualization of friendships in a group/community of actors. Whether the community contains genes appearing in a publication, actors associated with a Topic, or diseases in a metadisease, this view allows selection and observation of different types of friendships (common friends, common features, co-ocurrence, etc.). Highly connected actors will naturally form subgroups corresponding to biologically relevant categories (**Figure 4**).

Topic: interactive and collaborative workbench and personalization tool for humans

The concept of Topics represents a multidimensional workspace that is used to manage gene and disease friends related to a specific research project, to offer visualizations and analyses, to allow collaborations between humans, and to organize information used for personalization and suggestion.

When a human becomes friends with a new gene or disease, they are added to the active Topic. Friendship privacy with these actors can be modified at any time and actors can be removed, moved, or copied to other Topics. Topics provide useful visualization and analysis facilities such as the identification of highly connected subgroups of actors that can be displayed as a network. If a network is too compact and homogeneous, DAVID (Dennis et al. 2003) is used to obtain additional functional clustering. This representation can be complemented by a "*friends of friends"* network with access to diseases related to at least one gene in the Topic, or genes related

to at least one disease in the Topic. Finally, in order to replace the Topic in a global research perspective, a timeline representation shows the number of publications associated with the user's gene friends. This feature is interactive, allowing selection of a subset of genes, or visualization of publications related to a gene during a given period. In order to help the user identify authors of interest, authors who collaborated on most of the publications (as first, second or last author) are highlighted.

Collaborations allow users to share specifically defined information with their human friends. Collaborators will see all the protected information related to the Topic (including protected friends) and the user's protected annotations on the profile pages of genes and diseases related to this Topic. Moreover, users can choose to make their Topics public; in this case all humans will be able to find it with the search bar and integrate it into an existing Topic or import it as a new one.

A Topic contains information used to suggest new friends (genes and diseases) or publications, and to personalize information by highlighting specific words and displaying news related to befriended actors. To enhance the existing network of friends, MyGeneFriends provides the user with new friendship suggestions. These suggestions are based on the actors already included in the Topic. Gene friendship suggestions are based on the sum of STRING interaction scores between a gene and all genes in the Topic. Disease suggestions are based on the number of shared phenotypes between a disease and all diseases in the Topic. To help users select publications relevant to their research, MyGeneFriends uses text mining of keywords from textual data related to the actors in the active Topic. Elasticsearch (Gormley and Tong 2015) provides powerful and rapid search features to select publications best matching the user's interests. For example, keywords related to many friends will have a higher weight that those related to only a few of them, rare keywords will have more weight than frequent ones, etc. MyGeneFriends uses these keywords to help the user read textual data, by highlighting them in all texts (actor descriptions, abstracts of publications, etc.). Finally, the news feed provides new and relevant information about the actors in the user's active Topic. This concerns news such as a gene becoming friends with a disease, a disease updating its description, a new publication related to a gene, etc. News is presented in a context-defined form (sequence alignment, textual differences, publication abstracts, etc.).

To obtain a synoptic view of their projects, users can extract a summary of their Topics and export them as a PDF file with 3 main sections. The cover page section presents the Topic with a name and description, the next section lists all friends with their official descriptions and user added annotations, and the last section presents the 'liked' publication abstracts for further reading, with important keywords highlighted.

Usage scenarios

When users start a new Topic with no prior list of genes or diseases, they can add keywords to describe it. This will be sufficient for MyGeneFriends to make suggestions for publications and provide networks of genes linked to each publication, thus helping the user to identify interesting

hubs. By selecting an initial set of genes and related diseases as friends, a news flow will provide them with interesting news about their friends, and new friendship suggestions will appear.

Alternatively, users can start with an existing set of genes and/or diseases. Once these actors are added to the main Topic, MyGeneFriends will suggest new friendships, and publications linked to them. The size of the Topic will have important implications for the nature and quality of the suggestions and the personalization of information. A Topic containing dozens of poorly related genes and diseases will result in a huge newsfeed and unrelated publications, but the network view may distinguish two or three groups of highly related actors. These groups can be used to create two or three more specifically oriented Topics, each of which will have associated newsfeed and suggestions.

Users may even start with a gene expression approach. For example, a user is interested in genes expressed in liver, lung and heart, but not expressed in fetal brain. The "expression filter" tool can be used to show an ordered list of genes of interest, allowing users to visit their profiles and become friends with them.

Finally, if a new variant is identified in a patient, clinicians can enter its coordinates and alleles in their personal variants list on MyGeneFriends to check whether the variant has already been characterized. VEP (McLaren et al. 2010) is used to annotate human submitted variations and link them to actors from MyGeneFriends network, so the user can immediately start networking on MyGeneFriends through related genes and diseases.

DISCUSSION

By leveraging the conventions and practices used on popular social networks, MyGeneFriends aims to change the way we interact with data, by providing a first step toward a framework where biological entities like genes and diseases are no longer passive entities, but are instead proactive actors of the research process, helping and collaborating with their human collaborators. MyGeneFriends allows integration of expert knowledge with BigData by connecting human, gene and disease actors and providing powerful interactions among these actors, such as annotation, visualization, suggestions and personalization, organized around the concept of Topics and bringing together genes, diseases, keywords and collaborators.

Biology is a large field in terms of scale and complexity, ranging from the nucleotide to complex molecular and cellular networks. Therefore, the choice of the most suitable set of actors raised numerous questions. Biological networks represent interesting candidates, as emerging behavior resulting from their complexity and interconnectedness could create more autonomous actors. However their complexity also makes their integration difficult in this first version of MyGeneFriends. Variants on the other hand are more restricted entities, but still have numerous connections to humans and populations, genes, and genetic diseases. Nevertheless, a high noise to signal ratio and millions of unreliable variants in databases such as dbSNP and others, limit their usefulness as actors.

Recent initiatives, such as IBM's Watson system (Lally and Fodor 2011) for example, are based on the idea that machines alone can mine knowledge from BigData and discover new concepts, however we think human expertise and cognition is essential for BigData analysis and we therefore wished to allow human experts to explore the BigData in a cooperative way, as integrated actors of the network. Human diseases thus represent an ideal case study for MyGeneFriends, as they connect human researchers to the global human society, and represent major clinical challenges. For instance, infectious diseases are part of a complex network of host – pathogen relationships and cancer related diseases depend on numerous environmental and behavioral factors. We focused on human genetic diseases because of their more direct links to genes and genomic variations. With humans and human genetic diseases selected, the choice of the third actor was obvious; most bioinformatics resources structure their information in a gene-centric manner and publications mainly refer to genes, and much less to transcripts for example.

The development of MyGeneFriends lies at the frontier between bioinformatics and social networks, and in the future, we plan to extend the functionalities in both areas. First, genes from other species (Mouse, Zebrafish, Nematode) and additional friendships will be incorporated to provide a regulatory context, such as transcription-factor based and miRNA based friendships. Second, support for apps will be implemented, allowing third-party-developers to add new customized features directly to the MyGeneFriends service.

Finally, humans remains special actors in this first version of MyGeneFriends, but we believe that in the future, the three actors will interact on the same level, with more independent and proactive genes and diseases. Research will be facilitated by better communications between the different actors, with each actor able to produce and transmit new, relevant data and knowledge. A gene could for example find itself linked to a new disease or ask to be sequenced in a human by his friend the sequencer. With this increased autonomy of non-human actors and an independent flow of information, the role of the human in the network must clearly evolve. This evolution can be viewed either as a danger or as a source of new collaborations and opportunities.

METHODS

Platform architecture

MyGeneFriends is based on the Play framework (www.playframework.com), an open framework focused on developer productivity with many useful features such as error handling, build-in support for Json, WebServices, WebSockets, CoffeeScript (coffeescript.org), EBean (www.avaje.org) object-relational mapper (ORM), localization, logging and WebJars (www.webjars.org). To execute local scripts and programs, we have developed a web service using the Flask framework (www.fullstackpython.com/flask.html), which is called by MyGeneFriends using REST requests to run analysis or integration tasks. Data integration scripts are written in python, using peewee as the ORM.

To ensure a malfunction will not erase or corrupt MyGeneFriends data, the database is backed up daily to an external server and we rely on a stateless framework (Play 2 framework) to ensure easy horizontal scaling and scalability for increasing website traffic. Finally, to allow fast bug correction and feature suggestions, an icon is provided on the MyGeneFriends website, allowing users to enter a title and description of the problem they encountered or a suggestion for a feature they would like to be included. The YouTrack API (www.jetbrains.com/youtrack) is used to send these issues to the YouTrack server. Moreover, MyGeneFriends will automatically send issue tickets to YouTrack when exceptions are encountered during execution.

A PostgreSQL database is used to store and manage the data, as it is free, open source, not proprietary, robust, fast, and provides features that will be very interesting for future developments, such as native Json support. Elasticsearch (www.elastic.co) is used for powerful, complex and fast plain text queries of publications, and is synchronized daily with our main database using a river plugin.

Data sources

The Ensembl (Cunningham et al. 2015) database is our main data source for gene related data, including gene symbol, short description, biotype, protein sequence and other data. To ensure accurate mapping between Ensembl and NCBI gene identifiers, we combine the mappings performed by Ensembl and NCBI, together with symbol mapping. UCSC provides RefSeq (Pruitt et al. 2014) annotations to transcripts, and the list of stopwords (words that occur frequently in texts but are not informative) from the NCBI is used to extract keywords from biological texts. Relationships between genes and publications are defined using gene2pubmed file from NCBI. Publication abstracts are then downloaded from Pubmed.

Gene expression data is obtained from the Human Genome Atlas microarray data (Andrew et al. 2004) available in the GEO (Barrett et al. 2013) database, and validated using in-house statistical methods. In addition, a relative signal intensity is calculated for heatmap visualization, using log signal intensities normalized in the range [0-1].

Gene Ontology (GO (2015)) annotations of gene products in terms of cellular localization are used to create friendships between genes based on the number of shared GO terms, or their Functional Semantic Similarity (FSS) (Reyes-Palomares et al. 2013). The Information Content (IC) (Reyes-Palomares et al. 2013) score is used to represent the specificity and thus importance of localization terms. The STRING global score is also used as a metric of friendship between genes. In-house evolution profiles using orthology data produced by OrthoInspector (Linard et al. 2015) are used to calculate the Jaccard distance between two genes.

To take into account differences in disease definitions from different data sources and propose a unified view of the current disease knowledge, we have developed a merge process with two simple rules that are explained in **Figure 5**. Once diseases are merged, they are linked to phenotypes. To do this, HPO (Human Phenotype Ontology) data files (hp.obo and phenotype_annotations.tab) are parsed for phenotypes and relationships between phenotypes and diseases. Finally, to integrate variations and relationships between variations and diseases, we use the curated set provided by ClinVar (Landrum et al. 2014) in the VCF format file (limited to records with an rs# identifier). Each line is parsed and a variant entry is integrated into MyGeneFriends as a couple of genomic position and allele, allowing precise definition of the relationships between diseases and mutations.

Data flow management

The data flow management involves the integration of data from diverse sources into the MyGeneFriends database and comprises several steps. The first step is to retrieve the data from each source (database, ftp server, local laboratory files), with parsing and cleaning of the data if necessary.

The second step is extraction of additional information. For example, to extract keywords from textual data linked to genes and diseases, the python nltk (Bird 2006) library is used to tokenize the text into phrases and words, stem words in order to retrieve a canonical form, filter words from the NCBI stopwords file and in-house filters for word size, numbers and special characters. Then, we take advantage of the gensim (Řehůřek and Sojka 2011) library to calculate the Inverse Document Frequency (IDF) of the keywords, and the TF*IDF (Term Frequency * Inverse Document Frequency). The IDF is used as a specificity score and the TF*IDF is used to weight the relationship between a keyword and a gene or disease. VEP (McLaren et al. 2010) is used to link variations retrieved from ClinVar to Ensembl transcripts and to estimate their effect. The effects are then automatically classified in more general categories using the Sequence Ontology (Eilbeck et al. 2005) data.

The third step involves the comparison of remote and local data and the generation of news events. We use one or more fields from an item as a unique identifier. If a remote item has an identifier not present in our local database, it is considered to be a NEW event. If a local item has an identifier not present in the remote source, it is considered to be a DELETE event. If the identifier is present in both remote and local sources, the items are compared field by field to generate UPDATE events. Once these events are generated, the local database is synchronized to the remote source. Choosing a suitable identifier is a crucial problem to detect whether an item has changed, or whether it is a different new item. For example, the name of a disease is an unsuitable identifier, as it can and will change over time. The identification code of a disease combined with the data source (e.g. OMIM:209900) is a better identifier, but may also cause problems: for instance since the beginning of 2015, the OMIM 209900 code does not identify the

Bardet Biedl family, but only the Bardet Biedl 1 syndrome. For each data source, we have selected the field or the combination of fields that was the most stable over time.

Finally, comparison data is generated if necessary. Our goal is to represent the news in a way that is the most suitable and natural given its biological context. When an actor is linked to a new publication, the publication is downloaded and made accessible directly from the news. When a sequence is updated, a sequence alignment is generated using ClustalW (Thompson et al. 1994). When a textual information changes, such as the description of a disease, the google-diff (Fraser 2012) python library is used to compare both versions of the text and highlight what was removed and what was added. The advantage of this library is fast comparison of texts and detection of the smallest number of events to transform one text into the other.

Data display as Word Clouds

The information content (IC) metric (Reyes-Palomares et al. 2013) is used to estimate the specificity of a feature (localization, phenotype) describing an actor. The IC is defined as

$$IC(c) = -logP(c)$$

The specificity is then defined as the IC normalized in the range [0-1], where 0 corresponds to the minimal font size and 1 to the maximal font size during word cloud rendering.

Friendship suggestions and affinity score

Candidate actors (a_c) are scored relative to the active Topic, based on the sum of scores (S) between a candidate actor and all actors of the same type in the active Topic (a_t). To score genes, we use the global STRING score. To score diseases, we use the number of common phenotypes between two diseases.

$$\mathrm{score}(a_c) = \sum_{t=0}^{N} S(a_t, a_c)$$

The top ten candidates are then suggested as new friends.

The affinity score (aaff) is a metric reflecting the proximity between an actor and the content of user Topic and thus the relevance of befriending this actor. It is shown on gene and disease profile pages when applicable. We calculate it as:

$$a_{aff} = \frac{a_c}{\max a_c} \times 100$$

Publications suggestions

To suggest pertinent publications given the content of the active Topic, we query our elasticsearch server with a set of weighted keywords. The weight for each keyword given the content of the Topic (k_t) equals the sum of scores describing the relationship between this

keyword and all actors from the Topic. If the human has explicitly added this keyword to the Topic, a boost of (k_h) is applied. For elasticsearch, weights between 0 et 1 reduce relevance of a term, and weights greater than 1 increase it. To ensure the weight of the keyword increases relevance, we add 1.

$$weight(k_t) = 1 + \sum_{t=0}^{N} ka_t + k_h$$

ACKNOWLEDGEMENTS

This work was funded by ANR Investissement d'Avenir Bioinformatique BIP:BIP (ANR10-BINF03-05).

FIGURE LEGENDS

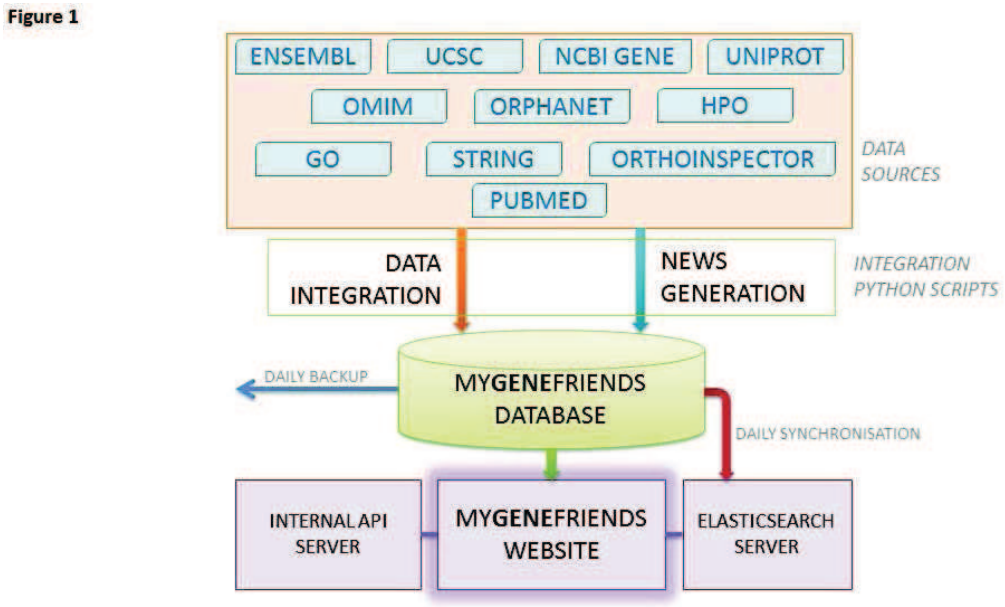


Figure 1. The architecture of the MyGeneFriends infrastructure allows simultaneous data integration and news generation from various asynchronous data sources and the integration of MyGeneFriends in a web of local services.

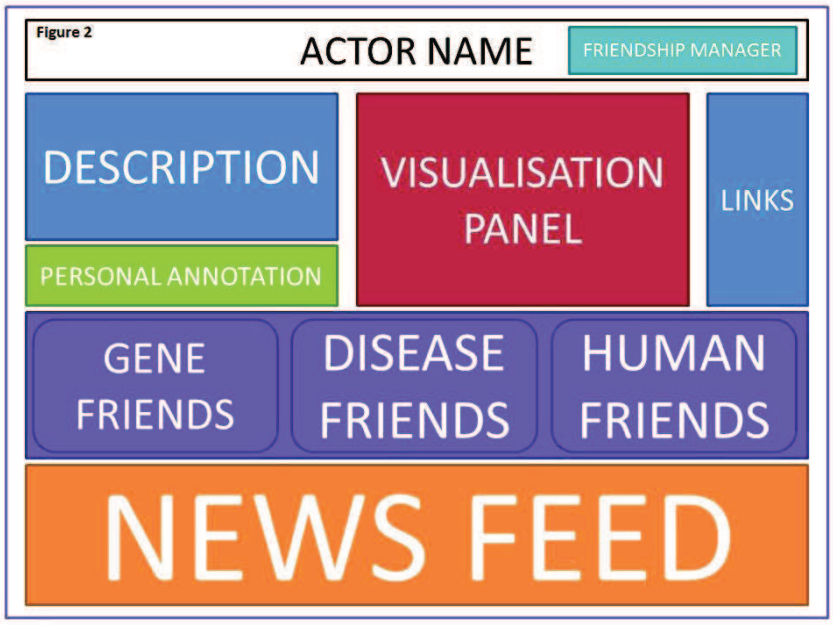


Figure 2. Schematic profile page for actors (genes, diseases and humans), using conventions from popular social networks, such as presentation, description, friendships and news feed for standardized and easy interaction with information.

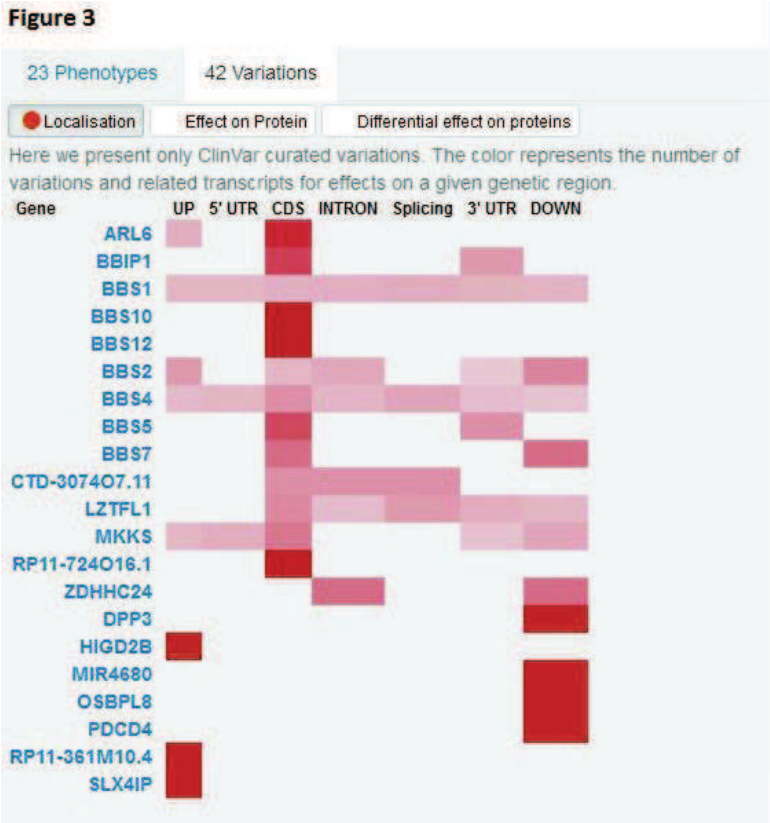


Figure 3. Visualization of variations associated with Bardet-Biedl Syndrome (BBS) based on their density in structural elements of transcripts from different genes. High density is shown in red and low density in light pink. High density of variations is observed in the CDS of BBS10 and ARL6 genes, while a more homogeneous repartition in seen for BBS1 and BBS4 for example. Genes with variations only in the neighboring regions, like HIGD2B and SLX4IP, are listed at the end.

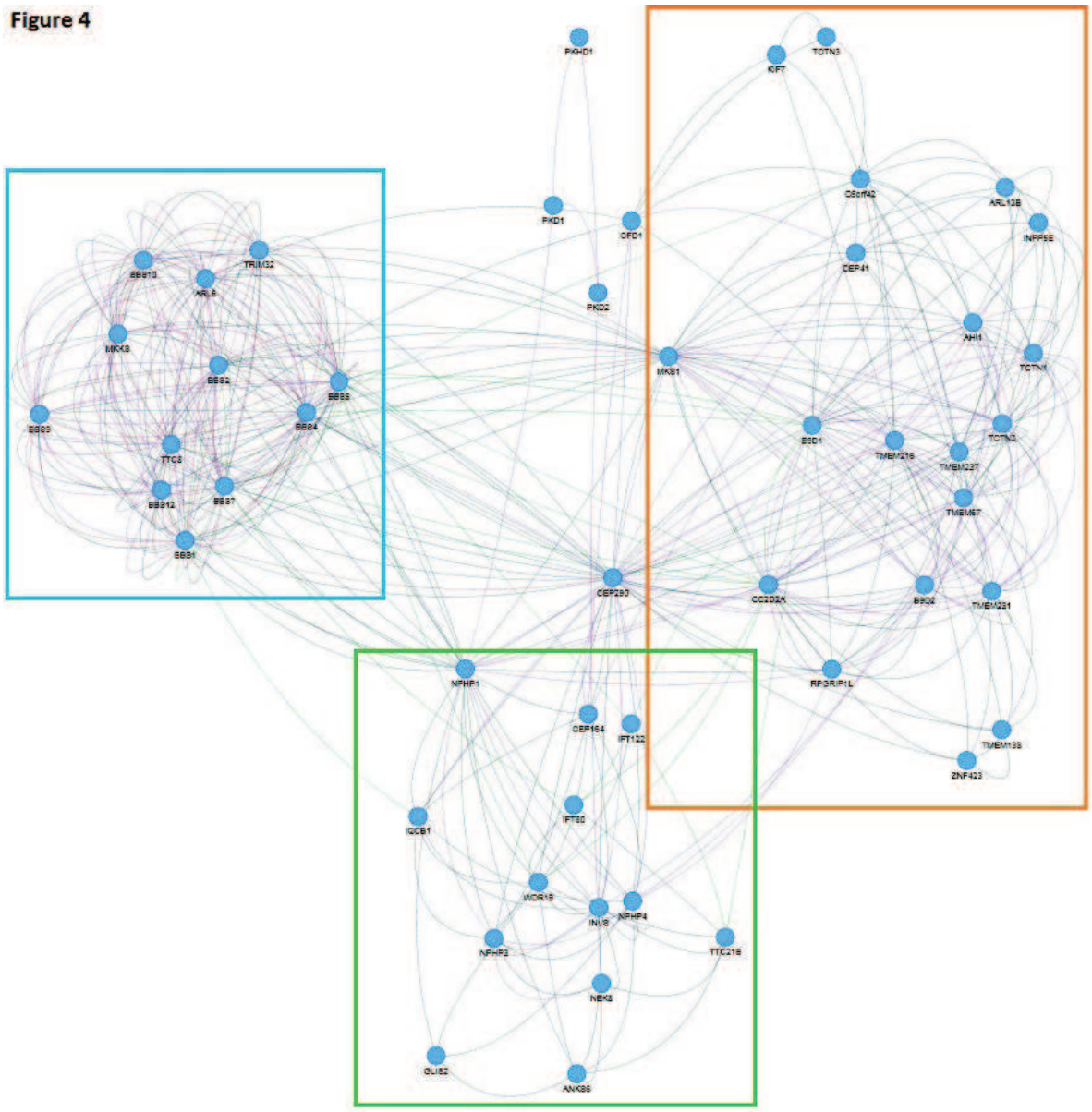


Figure 4. Network of 52 genes related to the publication "Congenital Hepatic Fibrosis Overview" (Gunay-Aygun et al. 1993). Red links represent shared public human friends, grey links represent shared diseases, violet links represent STRING relationships, and green links represent a similar evolutionary profile. Highly connected genes are automatically clustered to form subgroups. The blue rectangle highlights the genes associated with Bardet-Biedl Syndrome, the orange rectangle the genes associated with Joubert and Meckel syndromes and the green rectangle, genes associated with the Senior-Loken syndrome and nephronophthisis.

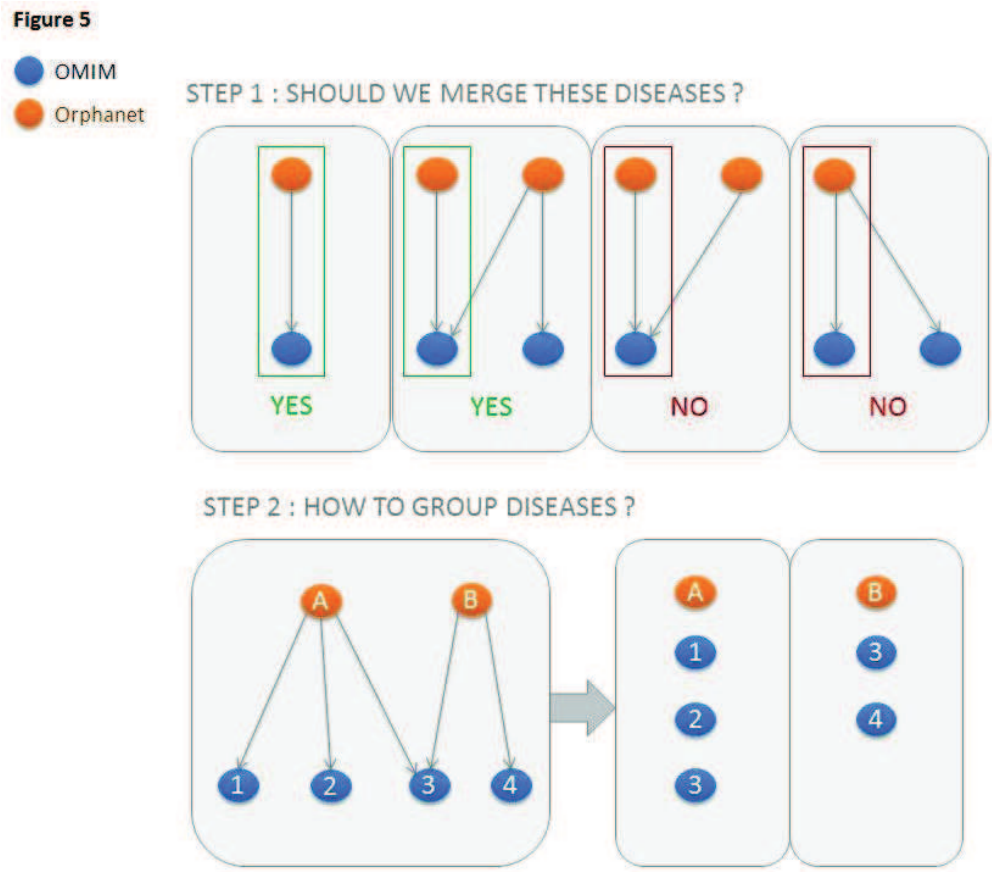


Figure 5. Diseases from different, contradictory data sources are merged using two rules. First, disease A is merged with disease B only if B has a single link to disease A, and no other diseases have a single link to A. Second, if a disease has links to several other diseases, a disease group called Metadisease is created.

REFERENCES

2015. Gene Ontology Consortium: going forward. *Nucleic Acids Res* **43**: D1049-1056.

Amberger JS, Bocchini CA, Schiettecatte F, Scott AF, Hamosh A. 2015. OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. *Nucleic Acids Res* **43**: D789-798.

Andrew IS, Tim W, Serge B, Hilmar L, Keith AC, David B, Jie Z, Richard S, Mimi H, Gabriel K et al. 2004. A gene atlas of the mouse and human protein-encoding transcriptomes. *Proc Natl Acad Sci U S A* **101**: 6062-6067.

Apweiler R, Bairoch A, Wu CH, Barker WC, Boeckmann B, Ferro S, Gasteiger E, Huang H, Lopez R, Magrane M et al. 2004. UniProt: the Universal Protein knowledgebase. *Nucleic Acids Res* **32**: D115-119.

Ayme S. 2003. [Orphanet, an information site on rare diseases]. *Soins*: 46-47.

Barrett T, Wilhite SE, Ledoux P, Evangelista C, Kim IF, Tomashevsky M, Marshall KA, Phillippy KH, Sherman PM, Holko M et al. 2013. NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**: D991-995.

Bellazzi R. 2014. Big data and biomedical informatics: a challenging opportunity. *Yearbook of medical informatics*.

Benz D, Eisterlehner F, Hotho A, Jaschke R, Krause B, Stumme G. 2009. Managing Publications and Bookmarks with BibSonomy. *20th Acm Conference on Hypertext and Hypermedia (Hypertext 2009)*: 323-324.

Bird S. 2006. NLTK: the natural language toolkit. Vol %6, pp. 69--72 %&. Association for Computational Linguistics, Stroudsburg, PA, USA.

Chen H, Sharp BM. 2004. Content-rich biological network constructed by mining PubMed abstracts. *BMC Bioinformatics* **5**: 147.

Consortium E. 2014. Europe PMC: a full-text literature database for the life sciences and platform for innovation. *Nucleic Acids Res*.

Cunningham F, Amode MR, Barrell D, Beal K, Billis K, Brent S, Carvalho-Silva D, Clapham P, Coates G, Fitzgerald S et al. 2015. Ensembl 2015. *Nucleic Acids Res* **43**: D662-669.

Dennis G, Jr., Sherman BT, Hosack DA, Yang J, Gao W, Lane HC, Lempicki RA. 2003. DAVID: Database for Annotation, Visualization, and Integrated Discovery. *Genome Biol* **4**: P3.

Doms A, Schroeder M. 2005. GoPubMed: exploring PubMed with the Gene Ontology. *Nucleic Acids Res* **33**: W783-786.

Eilbeck K, Lewis SE, Mungall CJ, Yandell M, Stein L, Durbin R, Ashburner M. 2005. The Sequence Ontology: a tool for the unification of genome annotations. *Genome Biol* **6**.

Fraser N. 2012. google-diff-match-patch-Diff, Match and Patch libraries for Plain Text.

Giglia E. 2011. PubMed in progress: latest changes in MeSH and MyNCBI. *Eur J Phys Rehabil Med* **47**: 525-528.

Giglia E, Spinelli O. 2009. PubMed reloaded: new interface, enhanced discovery. *Eur J Phys Rehabil Med* **45**: 631-636.

Gonzalez-Alcaide G, Castello-Cogollos L, Castellano-Gomez M, Agullo-Calatayud V, Aleixandre-Benavent R, Alvarez FJ, Valderrama-Zurian JC. 2013. Scientific publications and research groups on alcohol

consumption and related problems worldwide: authorship analysis of papers indexed in PubMed and Scopus databases (2005 to 2009). *Alcohol Clin Exp Res* **37 Suppl 1**: E381-393.

Gormley C, Tong Z. 2015. *Elasticsearch: The Definitive Guide*. " O'Reilly Media, Inc.".

Gunay-Aygun M, Gahl WA, Heller T. 1993. Congenital Hepatic Fibrosis Overview. In *GeneReviews(R)*, (ed. RA Pagon, et al.), Seattle (WA).

Hamosh A, Scott AF, Amberger J, Valle D, McKusick VA. 2000. Online Mendelian Inheritance in Man (OMIM). *Hum Mutat* **15**: 57-61.

Hodis E, Prilusky J, Sussman JL. 2010. Proteopedia: A collaborative, virtual 3D web-resource for protein and biomolecule structure and function. *Biochem Mol Biol Educ* **38**: 341-342.

Hoffmann R. 2008. A wiki for the life sciences where authorship matters. *Nat Genet* **40**: 1047-1051.

Hoffmann R, Valencia A. 2004. A gene network for navigating the literature. *Nat Genet* **36**: 664.

Hofker MH, Fu J, Wijmenga C. 2014. The genome revolution and its role in understanding complex diseases. *Biochim Biophys Acta* **1842**: 1889-1895.

Jee K, Kim GH. 2013. Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. *Healthc Inform Res* **19**: 79-85.

Kamphans T, Krawitz PM. 2012. GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. *Bioinformatics* **28**: 2515-2516.

Khalid Z, Max F, Harold R, Jason M, Christian Tannus L, Gary DB, Quaid M. 2013. GeneMANIA Prediction Server 2013 Update. *Nucleic Acids Res*.

Kitchin R. 2014. *Big Data, new epistemologies and paradigm shifts*.

Lally A, Fodor P. 2011. Natural language processing with prolog in the IBM watson system. *Retrieved June*.

Landrum MJ, Lee JM, Riley GR, Jang W, Rubinstein WS, Church DM, Maglott DR. 2014. ClinVar: public archive of relationships among sequence variation and human phenotype. *Nucleic Acids Res* **42**: D980-985.

Lee E, Helt GA, Reese JT, Munoz-Torres MC, Childers CP, Buels RM, Stein L, Holmes IH, Elsik CG, Lewis SE. 2013. Web Apollo: a web-based genomic annotation editing platform. *Genome Biol* **14**: R93.

Linard B, Allot A, Schneider R, Morel C, Ripp R, Bigler M, Thompson JD, Poch O, Lecompte O. 2015. OrthoInspector 2.0: Software and database updates. *Bioinformatics* **31**: 447-448.

Mandloi S, Chakrabarti S. 2015. PALM-IST: Pathway Assembly from Literature Mining - an Information Search Tool. *Sci Rep* **5**: 10021.

McLaren W, Pritchard B, Rios D, Chen Y, Flicek P, Cunningham F. 2010. Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. *Bioinformatics* **26**: 2069-2070.

Nielsen M. 2009. The Fourth Paradigm: Data-Intensive Scientific Discovery. *Nature* **462**: 722-723.

Papanikolaou N, Pavlopoulos GA, Pafilis E. 2014. BioTextQuest+: a knowledge integration platform for literature mining and concept discovery. ….

Prosdocimi F, Linard B, Pontarotti P, Poch O, Thompson JD. 2012. Controversies in modern evolutionary biology: the imperative for error detection and quality control. *BMC Genomics* **13**: 5.

Pruitt KD, Brown GR, Hiatt SM, Thibaud-Nissen F, Astashyn A, Ermolaeva O, Farrell CM, Hart J, Landrum MJ, McGarvey KM et al. 2014. RefSeq: an update on mammalian reference sequences. *Nucleic Acids Res* **42**: D756-763.

Řehůřek R, Sojka P. 2011. Gensim—Statistical Semantics in Python. *Gensim—Statistical Semantics in Python*.

Reyes-Palomares A, Rodriguez-Lopez R, Ranea JA, Sanchez Jimenez F, Medina MA. 2013. Global analysis of the human pathophenotypic similarity gene network merges disease module components. *PLoS One* **8**: e56653.

Sherry ST, Ward MH, Kholodov M, Baker J, Phan L, Smigielski EM, Sirotkin K. 2001. dbSNP: the NCBI database of genetic variation. *Nucleic Acids Res* **29**: 308-311.

Stelzer G, Inger A, Olender T, Iny-Stein T, Dalah I, Harel A, Safran M, Lancet D. 2009. GeneDecks: paralog hunting and gene-set distillation with GeneCards annotation. *Omics* **13**: 477-487.

Szklarczyk D, Franceschini A, Wyder S, Forslund K, Heller D, Huerta-Cepas J, Simonovic M, Roth A, Santos A, Tsafou KP et al. 2015. STRING v10: protein-protein interaction networks, integrated over the tree of life. *Nucleic Acids Res* **43**: D447-D452.

Thompson JD, Higgins DG, Gibson TJ. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res* **22**: 4673-4680.

Tsuruoka Y, Tsujii J, Ananiadou S. 2008. FACTA: a text search engine for finding associated biomedical concepts. *Bioinformatics*.

Van Landeghem S, Bjorne J, Wei CH, Hakala K, Pyysalo S, Ananiadou S, Kao HY, Lu Z, Salakoski T, Van de Peer Y et al. 2013. Large-scale event extraction from literature with multi-level gene normalization. *PLoS One* **8**: e55814.

Wei CH, Kao HY, Lu Z. 2013. PubTator: a web-based text mining tool for assisting biocuration. *Nucleic Acids Res* **41**: W518-522.

3.3 Faire face à la complexité des données biologiques et l'hétérogénéité des ressources bio-informatiques

Pour intégrer ces différents acteurs dans le réseau MyGeneFriends, il m'a fallu faire face d'une part, à la complexité des données biologiques traversée par les aspects de volume, vélocité, variété, variabilité, interconnexion mais également, à l'écosystème hétérogène et parfois contradictoire, des ressources bio-informatiques.

3.3.1 Gene

Lors de l'intégration des acteurs gènes et des informations les décrivant, j'ai été confronté à des défis de différents types : pratiques, avec les problèmes d'identifiants uniques et de correspondance robuste entre identifiants au travers des différentes bases de données utilisées ; biologiques, avec les notions de transcrit canonique qui fait toujours débat à l'ère de la biologie à haut débit et de la multiplication des transcrits alternatifs ; ergoniques, notamment pour décrire de façon synthétique les liens entre mutations et transcrits.

3.3.1.1 Faire correspondre les id des différentes sources de données

Afin de réduire les problèmes inhérents à l'établissement de correspondance entre identifiants de gènes venant de plusieurs sources de données (par exemple entre Ensembl et NCBI), j'utilise un processus d'intégration des correspondances de plusieurs sources.

Un échange par email m'a appris qu'Ensembl établit la correspondance entre ses identifiants et les identifiants externes en se basant sur la similarité de séquences (recherche par BLAST en considérant tous les hits avec plus de 80% de similarité). Ceci génère, dans un certain nombre de cas, des correspondances multiples (un-à-plusieurs ou plusieurs-à-plusieurs). Les correspondances d'identifiants effectuées par le NCBI se heurtent également à ce problème.

La première problématique fut donc de choisir entre spécificité (n'utiliser que les correspondances qui se retrouvent dans toutes les sources) et sensibilité (conserver et complémenter les différentes correspondances). J'ai choisi de privilégier la sensibilité afin d'annoter le plus de gènes.

Le protocole de *remapping* utilisé par MyGeneFriends priorise les sources de correspondances d'identifiants. Après une série de tests, nous avons retenu l'ordre de priorité suivant : « MappingEnsembl », « MappingParSymbole » puis « MappingNCBI ». Les différents fichiers de correspondance sont filtrés pour ne conserver que les relations de type un-à-un (Tableau 17), puis appliqués dans l'ordre aux différents identifiants Ensembl. On peut voir que, même si la majorité des correspondances se retrouvent dans tous les *mappings*, chacun apporte des relations complémentaires des autres (Figure 32).

Tableau 17: Le filtrage pour ne conserver que les relations de type un-à-un enlève une partie des relations pour chaque source.

	Ensembl	NCBI	Symbol
Total correspondances	26320	22067	35564
Correspondances OneToOne	23763	21960	29692

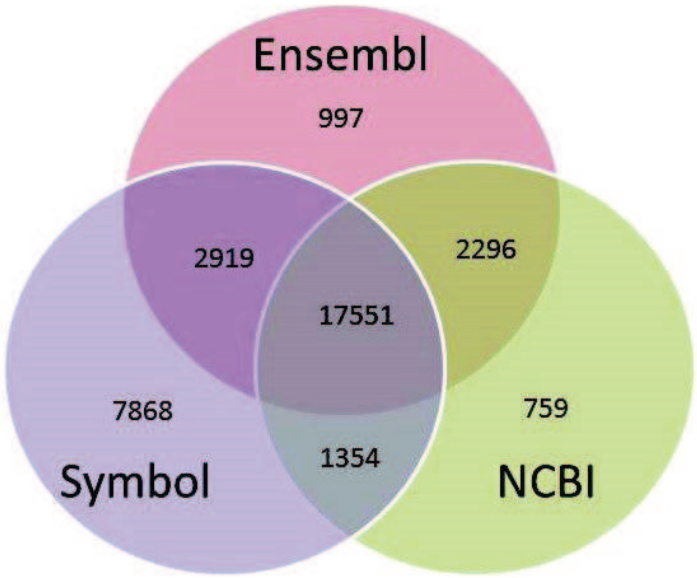


Figure 32: La combinaison entre différentes sources de mise en correspondance des identifiants Ensembl et NCBI permet d'améliorer l'attribution d'une correspondance unique vers des identifiants NCBI aux gènes de MyGeneFriends issus de Ensembl.

3.3.1.2 Choisir un transcrit canonique

La notion de transcrit canonique permet d'associer à un gène un transcrit principal. Bien que la définition théorique et officielle désigne comme transcrit canonique le transcrit le plus exprimé, le manque de données sur l'expression de transcrits poussent les ressources bio-informatiques à adopter des définitions plus pratiques. Ainsi la colonne *canonical_transcript* dans la base de données Ensembl désigne les transcrits codant pour la plus longue protéine [préférablement avec la plus longue séquence codante consensus (CCDS)] car ils sont utilisés pour des analyses de génomique comparative. Pour le classement des transcrits sur le site d'Ensembl, les labels comme TSL (*Transcript Support Level*, une méthode permettant de juger de la qualité d'identification d'un transcrit), APPRIS (Rodriguez et al., 2013) ou la présence de CCDS sont utilisés en complément d'autres données. Mais les informations sur le classement global ne sont pas

disponibles dans la base de données de Ensembl, ce qui m'a obligé à classer les transcrits moi-même pour MyGeneFriends.

Pour déterminer un transcrit canonique, j'associe des scores aux différents transcrits (S_t) en m'appuyant sur des critères auxquels sont associés des scores pondérés (S_c), avec par ordre décroissant d'importance le fait que le transcrit considéré:

- soit le seul transcrit,
- soit le seul transcrit protéique,
- possède la correspondance d'une séquence canonique UniProt (UniProt considère comme canoniques les séquences protéiques les plus fréquentes et les plus semblables aux séquences orthologues d'autres espèces),
- possède la correspondance d'un ou plusieurs transcrits de la banque RefSeq,
- possède une CCDS.

Si plusieurs transcrits possèdent le score le plus élevé, ils sont alors classés par taille et le plus long est considéré comme transcrit canonique. Le score d'un transcrit est donné par la formule :

$$S_t = \sum_{c=0}^{n} S_c \; ssi \; critère \; s'applique \; \text{à} \; t$$

3.3.1.3 Simplifier les données sur l'expression des gènes et les rendre visualisables

Les données des expériences RNAseq disponibles gratuitement, bien qu'ayant un certain nombre d'avantages dus à cette technique (notamment l'identification des transcrits), ne sont pour l'instant pas comparables en terme de quantité et richesse de tissus associés aux données issues des expériences *microarrays*. J'ai donc dans un premier temps intégré l'expérience HGA (*Human Genome Atlas*) *microarray* qui décrit l'expression des gènes dans 79 tissus. Cela a posé un certain nombre de problèmes car, pour simplifier au maximum l'accès aux données, je ne voulais pas obliger le chercheur à spécifier à chaque fois un *probeset* donné ou un transcrit donné pour observer l'expression.

Je définis donc la relation entre un gène et un tissu pour une expérience donnée par l'intensité de signal maximale associée à un *probeset* lié au transcrit canonique de ce gène. Sans être parfaite, c'est une bonne approximation. Cette intensité de signal est transformée en couleur (HSL, *Hue, Saturation, Lightness*) par un traitement décrit dans la partie Matériels et Méthodes.

Cela a également nécessité la schématisation de 41 tissus humains (présentés dans le Tableau 18, plusieurs tissus sont extraits des fichiers SVG de Wikipédia) afin de pouvoir analyser visuellement l'expression d'un gène (exemple présenté sur la Figure 40). Le système de visualisation d'expression est suffisamment souple pour permettre l'ajout de tissus supplémentaires et la visualisation de l'expression dans d'autres organismes que l'Homme.

Tableau 18: Tissus schématisés dans MyGeneFriends afin d'afficher l'expression des gènes

Corps entier	Cerveau	Fœtus
Whole_brain	**Brain_thalamus**	**Placenta**
Trachea	**Cerebellum**	**Fetal_brain**
Heart	**Cerebellum_peduncles**	**Fetal_liver**
Bone	**Hypothalamus**	**Fetal_lung**
Lung	**Medulla_oblongata**	**Fetal_thyroid**
Liver	**Occipital_lobe**	
Psoas	**Olfactory_bulb**	
Kidney	**Parietal_lobe**	
Adrenal_gland	**Pituitary**	
Adrenal_cortex	**Pons**	
Thymus	**Prefrontal_cortex**	
Pancreas	**Spinal_cord**	
Whole_blood	**Temporal_lobe**	
Prostate		
Testis		
Ovary		
Uterus		
Skin		
Thyroid		
Tongue		
Alivary_gland		
Tonsil		
Appendix		

3.3.2 Maladie

Nous avons vu dans la partie Introduction que définir les maladies comme des entités distinctes était un problème ardu. Cette problématique ressurgit lorsque l'on compare les listes de maladies proposées par différentes sources de données, qui possèdent de nombreuses contradictions. C'est un défi à relever pour intégrer et fusionner des maladies de différentes sources de données. Mais c'est également un outil pour créer des groupes de maladies similaires, que nous avons réunies sous le terme de 'Méta-Maladie'.

3.3.2.1 Agglomérer les sources de données contradictoires pour mieux définir la maladie

Afin de créer une liste de maladies aussi exhaustive et objective que possible, MyGeneFriends possède un système d'intégration de maladies de sources différentes (pour l'instant OMIM et Orphanet).

a) Chaque maladie, peu importe la source d'information choisie, est décrite par un objet *BasicDisease*, avec des informations comme nom, description, synonymes, identifiants, et références extérieures. A ces informations vont s'ajouter deux listes. La première

contiendra toutes les maladies fusionnées à celle-ci. La deuxième contiendra toutes les maladies absorbées par celle-ci.

b) Ces objets *BasicDisease* sont intégrés dans un pool de maladies.
c) On traverse une première fois ce pool pour identifier toutes les relations *un à un* entre maladies et les fusionner pour constituer une seule maladie. Pour chaque source, on détermine laquelle possède le plus d'informations sur la maladie. La source la plus informative donnera son nom à la maladie résultante.
d) On parcourt une deuxième fois ce pool pour créer des Méta-Maladies à partir de relations *un à plusieurs* et *plusieurs à plusieurs*, en permettant à des maladies d'en absorber d'autres.

Les objets *BasicDisease* toujours valides (qui n'ont pas été fusionnés à d'autres) seront intégrés dans MyGeneFriends. Les maladies présentes dans MyGeneFriends sont donc pour une part, spécifiques à des sources de données et pour une autre, fusionnées à partir de sources différentes (Tableau 19).

Tableau 19 : Répartition par source des maladies dans MyGeneFriends

Pur OMIM	Pur Orphanet	Maladies fusionnées
4965	**6337**	**3159**

3.3.2.2 Regrouper des maladies semblables

Les Méta-Maladies permettent à MyGeneFriends de regrouper des maladies semblables. Elles constituent des groupes plus ou moins hétérogènes qui peuvent être visualisés sous forme de réseau. MyGeneFriends possède à l'heure actuelle plus de 700 méta-maladies qui regroupent près de 3500 maladies (voir Tableau 20 pour plus de détails). Les relations entre maladies et Méta-Maladies sont de type plusieurs-à-plusieurs, ce qui signifie qu'une maladie peut faire partie de plusieurs Méta-maladies. La maladie qui possède des liens vers plusieurs autres maladies et qui sert donc de base pour créer une méta-maladie est également présente dans MyGeneFriends sous forme de maladie.

Tableau 20: Regroupement des maladies en méta-maladies

Caractéristique	
Total des maladies regroupées en méta-maladies	3431
Nombre maximum de maladies contenues dans une méta-maladie	77
Nombre minimum de maladies contenues dans une méta-maladie	2
Nombre moyen de maladies dans une méta-maladie	5,3
Nombre moyen de méta-maladies auxquelles appartient une maladie	1,13
Nombre maximal de méta-maladies auxquelles appartient une maladie	8

L'étude des réseaux des différentes méta-maladies permet de différencier plusieurs profils. Un premier profil illustré par la Figure 33 correspond par exemple à la méta-maladie « Familial isolated dilated cardiomyopathy». On remarque que les maladies sont très connectées entre elles et forment un ensemble compact.

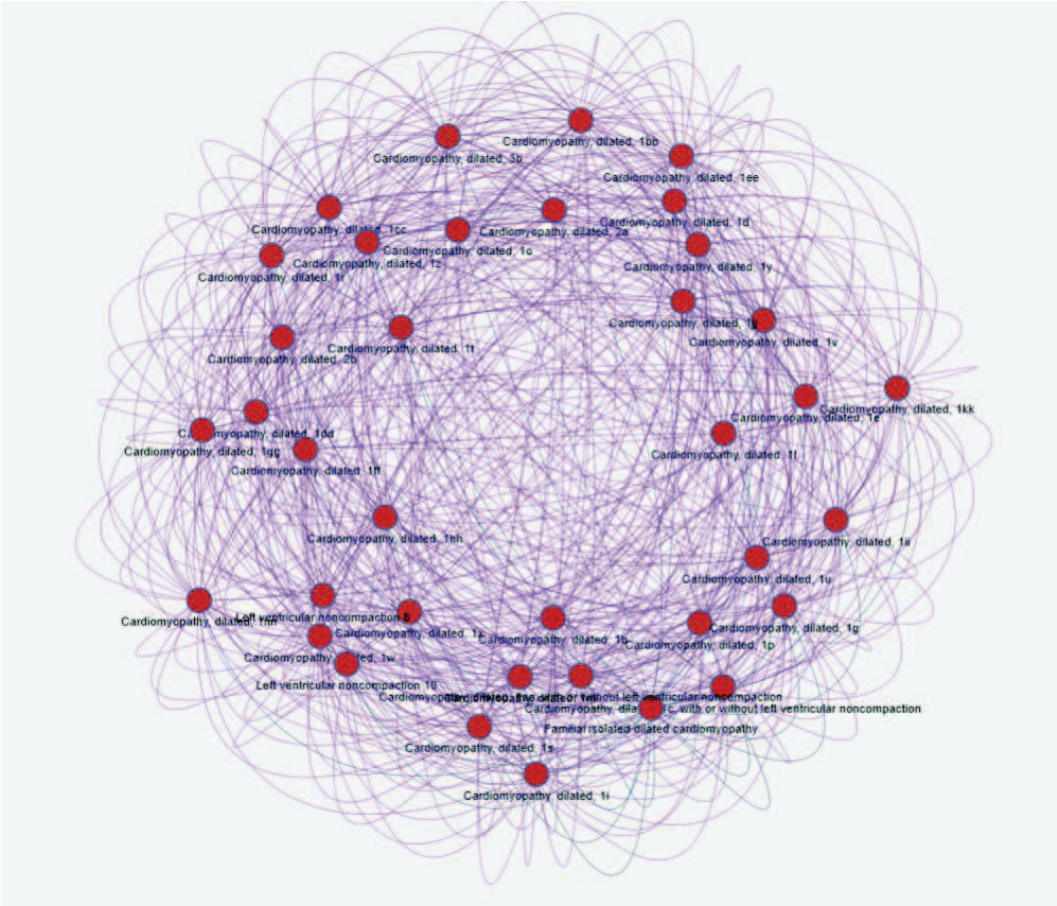


Figure 33: Réseau des maladies appartenant à la méta-maladie « Familial isolated dilated cardiomyopathy ». Les maladies sont reliées par les gènes et phénotypes qu'elles ont en commun. Les trois maladies n'ayant pas de gènes ou phénotypes en commun avec les autres maladies ne sont pas présentes sur la figure.

Un autre profil de méta-maladie (*Figure 34*) montre trois grands groupes de maladies, chacun possédant des maladies ayant un vaste réseau de phénotypes en commun. Les trois groupes sont reliés par des liens de gènes communs à la maladie « *Autosomal recessive non-syndromic sensorineural deafness type DFNB* » dont les informations ont servi de base à la création de cette méta-maladie. Après analyse des trois groupes de maladies, on peut remarquer qu'elles se distinguent principalement par la précision de leur annotation phénotypique. Ainsi le groupe A est principalement caractérisé par le phénotype « *Sensorineural hearing impairement* », le groupe B

est principalement relié par le phénotype *« Hearing impairement »*, et le groupe C par le phénotype *"Prelingual sensorineural hearing impairment"*. Une analyse plus poussée permettrait de savoir si ces trois groupes pourraient être candidats à former des maladies distinctes.

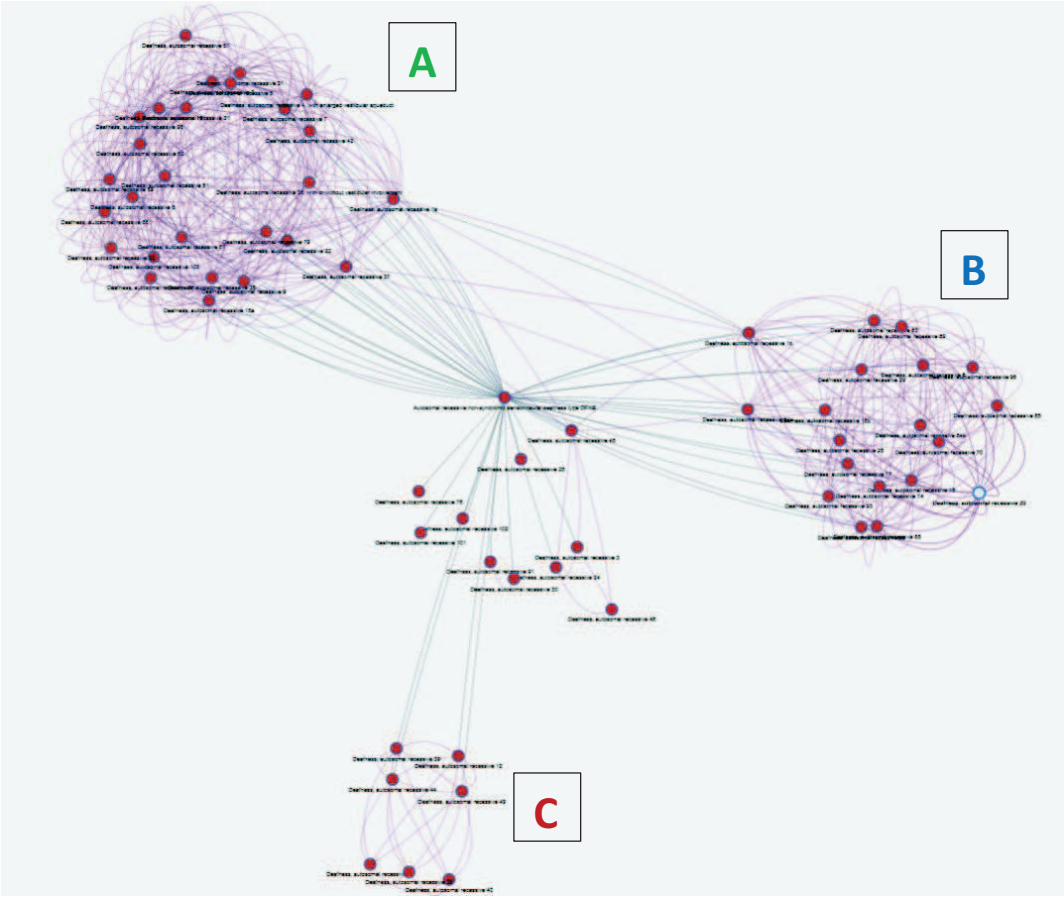


Figure 34: Les maladies appartenant à la méta-maladie "Autosomal recessive non-syndromic sensorineural deafness type DFNB" forment trois grands groupes distincts, A, B et C.

3.4 Décrire et formaliser les acteurs grâce aux mots clefs

Après avoir intégré les acteurs et leurs données, il était essentiel d'en extraire les informations qui allaient permettre à MyGeneFriends de comprendre les grandes catégories auxquelles ces acteurs appartenaient afin de faire fonctionner ses mécanismes de personnalisation de l'information et ainsi :

- Faire ressortir les passages les plus importants pour l'acteur humain des différents textes (publications, descriptions de gènes et maladies, etc…),
- Caractériser automatiquement les utilisateurs humains suivant leurs intérêts publics,
- Suggérer des publications pertinentes par rapport aux intérêts courants de l'utilisateur.

J'ai choisi d'utiliser à cette fin un pool de mots clefs, construit en analysant les données textuelles relatives aux différents acteurs. Les descriptions, noms complets et termes GO pour les gènes, et les descriptions et phénotypes pour les maladies ont ainsi été « digérées » pour caractériser les acteurs de manière générale, sans être trop distinctif.

J'utilise la très populaire librairie python de traitement du langage naturel, nltk, pour diviser le texte en phrases, puis en mots, et enfin extraire la forme racine du mot. La librairie propose plusieurs racinisateurs (*stemmer* en anglais) comme les racinisateurs Porter ou Snowball (Porter, 1980) ou le racinisateur de Lancaster (Paice, 1990). J'ai choisi d'utiliser le racinisateur de Porter car il offrait la meilleure qualité de racinisation (Jivani, 2011).

Un mot clef MyGeneFriends est composé de :

- La forme racinisée qui sert à le comparer aux autres, le rechercher dans les textes et construire une requête pour le moteur de recherche ElasticSearch (voir chapitre 2.2.2),
- L'ensemble des mots correspondant à la forme racinisée qui ont été retrouvés dans les textes,
- Une forme canonique utilisée pour l'affichage du mot clef sur les profils des acteurs.

Chaque mot clef est caractérisé par la valeur IDF (*Inverse Document Frequency*) qui définit sa rareté dans le corpus de textes considérés. Je calcule cette valeur grâce à la librairie python gensim.

Le flux de mots clefs initial passe dans un pipeline de filtrage qui permet d'écarter des mots de faible intérêt :

- Les mots apparaissant dans le fichier *stopwords* du NCBI répertoriant les mots à faible contenu informationnel (*a*, *to*, *from*, etc…) ainsi que certains mots biologiques (gène, protéine),
- Les mots dont la composition en caractères non alphabétiques est supérieure à la composition en caractères alphabétiques.

La relation entre un mot clef et un document est définie par la valeur du $tfidf = f_t * idf_t$ qui favorise les mots apparaissant fréquemment dans un document donné et peu dans les autres documents (voir chapitre 2.3.2).

Les mots clefs et leurs scores constituent un modèle représentant l'intérêt courant de l'acteur humain, utilisé par les mécanismes de personnalisation de l'information décrits dans le chapitre 3.2). Le score qualifiant la relation entre un acteur et un mot clef, est égal au score *tfidf* maximal du mot clef obtenu dans les documents relatifs à cet acteur. A l'usage, cela semble être une bonne approximation.

L'ensemble des mots clefs extraits ainsi que les relations entre mots clefs et acteurs sont intégrés à la base de données MyGeneFriends.

Lorsque l'utilisateur ajoute des mots clefs et des acteurs à un *Topic* (qui représente un projet scientifique de l'acteur humain), MyGeneFriends recalcule les scores des différents mots clefs par rapport à ce *Topic*, le score entre un mot clef et un Topic (t_k) étant égal à la somme des scores a_k de ce mot clef avec tous les acteurs du *Topic*. Un poids supplémentaire (h_k) est ajouté si ce mot clef a été rajouté explicitement au *Topic* par l'utilisateur :

$$t_k = 1 + h_k + \sum_{a}^{N} a_k$$

Ce système possède des imperfections et peut être amélioré dans le futur sur plusieurs points.

La distinction entre un mot clef pertinent et un mot clef peu informatif est floue pour la machine, et souvent même pour l'humain. Cela rend difficile le filtrage des mots clef ou l'établissement des seuils. L'utilisation de fichiers de *stopwords* existants n'a pas suffi à éliminer une partie du bruit dans les mots clef hautement scorés (par exemple *ratio, review, locus, patient*, …), il m'a donc fallu filtrer manuellement une partie des mots clef pour les ajouter à un nouveau fichier *stopwords* pour MyGeneFriends.

L'utilisation de la mesure de l'entropie (présentée dans l'introduction de la thèse)(Lochbaum and Streeter, 1989) de chaque mot et la sélection des mots à l'entropie la plus élevée permettrait également d'augmenter la qualité du set de mots clefs. L'entropie est une meilleure métrique que l'idf pour capturer l'importance d'un mot dans la collection de documents car elle tient compte de l'importance du mot dans les documents où il est présent.

Les mots clef composés d'un seul mot sont limités pour transmettre un sens plus complexe. Pour faire face à cela on pourrait envisager de détecter et utiliser des groupes nominaux en complément des mots clef uniques.

Enfin, des erreurs de racinisation peuvent réunir des mots dont les sens sont très différents comme « organic », « organisms », « organ », ou « irritable » et « irritant ». La racinisation doit donc être complétée par une contextualisation des mots.

D'autre part, afin d'annoter plus efficacement les acteurs humains, il faudrait classifier les mots clefs qui leur sont reliés en des catégories, soit en utilisant une ontologie soit en utilisant un classificateur. Cela nécessite de créer des groupes de mots clef et de donner à chaque groupe un nom adéquat et informatif. Plusieurs techniques existent pour cela, on peut citer par exemple LSA (*Latent Semantic Analysis*) et LDA (*Latent Dirichlet Allocation*), toutes deux implémentées dans la librairie *gensim*. LSA se base sur la cooccurrence des mots dans les mêmes documents, et permet par exemple de regrouper des mots au sens semblable (même profil) (Dumais, 2004) et capturer une partie de la similarité du sens.

En plus de décrire un acteur humain par des mots clef ou catégories de mots clef, MyGeneFriends

devra également décrire les *Topics* par les catégories de mots clef liées aux acteurs qu'ils contiennent. Cela permettra, en plus de la visualisation des acteurs sous forme de réseau, de déterminer si le *Topic* est homogène (tous les acteurs peuvent être décrits par une ou deux grandes catégories) ou contient des sous-groupes implicites d'acteurs (beaucoup de catégories identifiées).

En parallèle à la mise en place d'une solution simple de caractérisation de liens entre acteurs et mots clef par approche tf*idf, j'ai exploré des pistes plus complexes, aux résultats prometteurs pour l'avenir.

L'utilisation du modèle Word2Vec (radimrehurek.com/gensim/models/word2vec.html) est une direction très intéressante. Dans le traitement de la phrase, cette approche vectorielle associe à chaque mot son importance dans le document donné (Hotho et al., 2005).

Word2vect permet de manipuler les vecteurs de n dimensions en les additionnant ou soustrayant par exemple. Ainsi KING – MAN + WOMAN correspondra à un vecteur dont le vecteur le plus similaire sera QUEEN.

Parmi les utilisations que cette approche permet figurent :

- Le regroupement des mots semblables,
- Le calcul de la distance entre deux mots,
- La capacité à trouver l'intrus dans une liste de mots (le mot qui a le moins de points communs avec les autres).

Afin de tester cette approche sur des termes biologiques, j'ai téléchargé et chargé en mémoire un modèle word2vect de Pubmed et PMC créé en 2013 et disponible à l'adresse (http://evexdb.org/pmresources/vec-space-models/).

Les tests n'ont pas été concluants et le modèle ne s'est pas révélé capable de discriminer les termes appartenant à des problématiques biologiques différentes. Je pense cependant que ce genre d'approche peut s'avérer très utile, si elle est maitrisée, pour découvrir les véritables nouveautés dans les publications, à savoir des idées nouvelles ou révolutionnaires. On pourrait pour cela, « rechercher l'intrus » parmi les termes les plus importants de la publication, pour détecter si l'un d'entre eux est hautement improbable dans ce contexte, par rapport au modèle des publications existantes.

J'ai également testé la possibilité de représenter les liens entre mots clef, et entre mots clef et acteurs, sous forme de graphe. Pour cela je me suis servi de la base de données graphique Neo4j car la puissance de son moteur de requêtes Cypher m'avait beaucoup intéressé. Je voulais l'utiliser pour identifier dans le réseau tous les nœuds relatifs au *Topic* de l'utilisateur (acteurs, mots clef) pour ensuite pouvoir attribuer un score à tous les mots clef du réseau (Figure 35), et les

utiliser pour évaluer des phrases et paragraphes de publications, afin de détecter les plus intéressants pour l'utilisateur. Cependant, l'intégration de Neo4j avec la base de données principale de MyGeneFriends était problématique et les premiers tests de requêtes se sont révélés trop lents. J'ai donc opté dans un premier temps pour la gestion de ce réseau par la base de données relationnelle principale, sans utiliser une base de données en graphe.

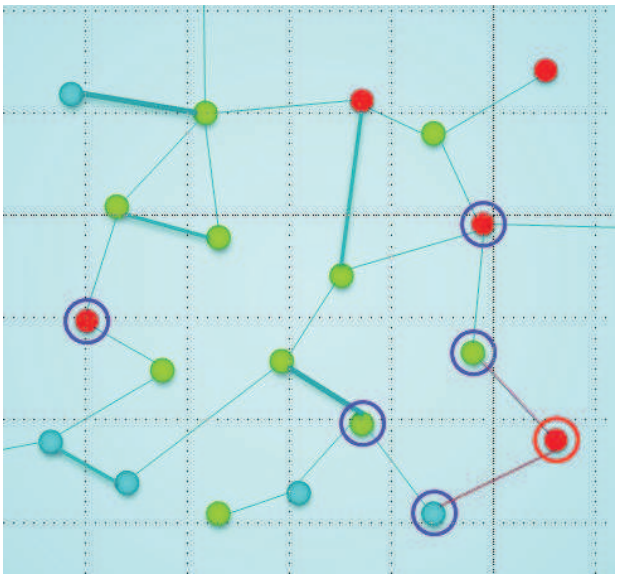


Figure 35 : Schéma de l'organisation d'acteurs et de mots clefs sous forme de réseau hétérogène comportant des gènes (points bleus), de maladies (points rouges) et des mots clef (points verts). Les arêtes représentent un score de similarité entre deux nœuds. L'idée était de marquer dans ce réseau tous les nœuds (gènes, maladies, mots clef) relatifs au Topic courant (nœuds entourés par des cercles violets), ce qui permettrait de donner un score à n'importe quel autre nœud par rapport à ce Topic (par exemple la maladie entourée par le cercle rouge).

3.5 Architecture

La base de données principale MyGeneFriends utilise un serveur Postgres et contient 49 tables de taille très variable, allant de quelques KiloOctets à plusieurs GigaOctets de données. On peut séparer ces tables en trois familles :

- Les tables qui contiennent des données biologiques,
- Les tables qui contiennent des données sur les utilisateurs,
- Les tables qui lient les deux types de données.

Assurer un bon fonctionnement de cette base de données bio-sociale et de la plate-forme web MyGeneFriends a nécessité un certain nombre de choix techniques, architecturaux et philosophiques. Ces choix sont en grande partie le résultat de l'expérience acquise durant la création de Parsec et concernent aussi bien le choix des librairies et des technologies de bases de données, l'architecture et l'organisation des données, mais aussi le choix d'un modèle centralisé ou décentralisé pour MyGeneFriends.

3.5.1 Choix techniques

Faire de bons choix techniques est essentiel pour développer rapidement une application, être flexible au changement de contraintes et aux demandes de nouvelles fonctionnalités, corriger facilement et rapidement les bugs, mettre à jour le site web sans qu'il soit indisponible plus de quelques secondes, gérer efficacement et rapidement de grandes quantités de données, et enfin pouvoir gérer une charge croissante de trafic et d'utilisateurs.

3.5.1.1 Barre de recherche : pour une recherche parallèle et décentralisée

Afin de simplifier au maximum la recherche d'informations sur MyGeneFriends, j'ai choisi de m'inspirer de Facebook pour créer une seule barre de recherche permettant d'atteindre tout type d'informations sur le site web. J'ai légèrement modifié la fonctionnalité de complétion automatique de la librairie JavaScript jqueryUI afin de lancer chaque recherche de ressource en parallèle (Figure 36).

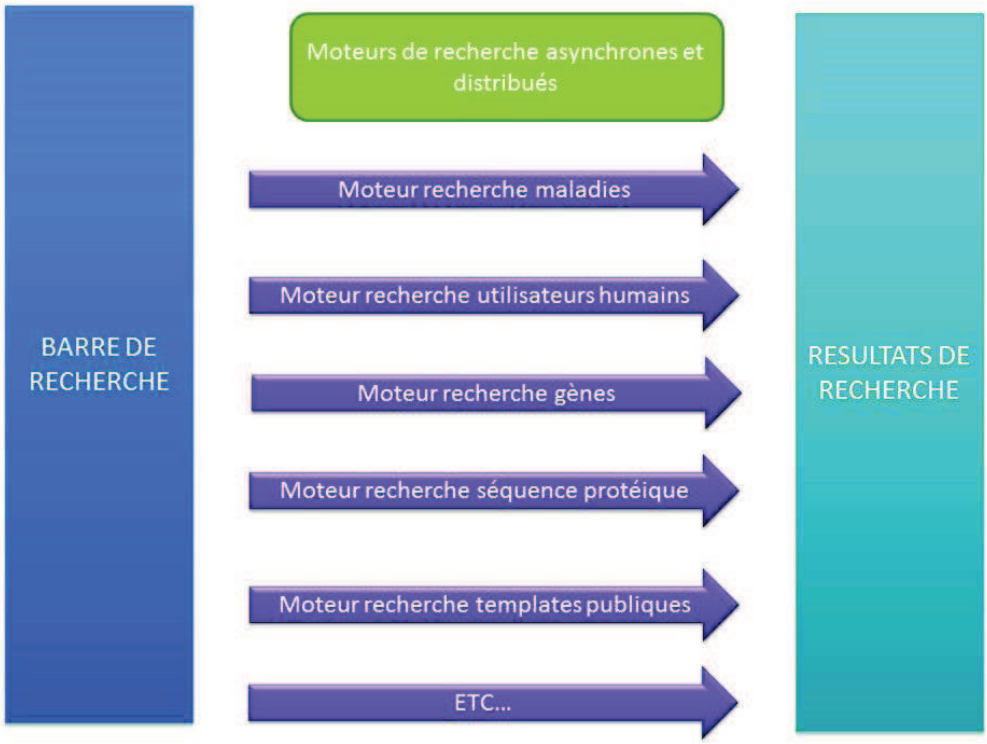


Figure 36: Schéma du fonctionnement parallélisé de la barre de recherche

Chaque moteur de recherche indépendant est accédé par une URL, reçoit la chaine cherchée, et s'engage à retourner un JSON contenant des informations spécifiques (Tableau 21).

Tableau 21: Caractéristiques des entrées d'un résultat de recherche renvoyé par un moteur de recherche

Champ JSON	Description
id	Un identifiant unique à l'ensemble des résultats de recherche, composé de l'identifiant du moteur et de l'index du résultat, par exemple : g1
label	Le nom principal du résultat renvoyé, par exemple « RARA » ou « Alexis »
alternativeLabel	Le nom complémentaire, par exemple le stable_id d'un gène ou l'adresse mail d'un humain
category	Le nom de moteur de recherche utilisé, ex : Disease, Humans, TopicTemplates, etc…
url	L'URL dont les informations seront chargées lorsque l'utilisateur cliquera sur le résultat de la recherche.
searchResultAction	Définit l'action à effectuer lorsque l'utilisateur clique sur le résultat de la recherche. Pour l'instant deux modes sont disponibles : '*show-in-modal*' et '*show-in-webpage*'. '*show-in-modal*' affiche le contenu de l'url dans un modal, tandis que '*show-in-webpage*' ouvre l'url dans une nouvelle page web.

A mesure que les résultats de recherche des différents moteurs arrivent, ils s'affichent dans les résultats de recherche. Ce couplage faible entre barre de recherche et les moteurs effectuant une recherche précise permet de distribuer la recherche, qui peut en partie être effectuée par des serveurs distants (Figure 36).

3.5.1.2 SQL ou NoSQL

La base de données est un élément essentiel des nombreux services web modernes, garantissant la pérennité des données, leur mise à jour, leur accès rapide, et permettant même d'effectuer des traitements. Le choix de la technologie de la base de données est donc important. Deux grandes familles sont disponibles aujourd'hui : les solutions SQL et NoSQL (Tableau 22).

Après une longue période de domination sans partage des bases de données relationnelles (PostgreSQL, MySQL, Oracle, SQLserver, etc…), on a vu récemment non seulement l'émergence de solutions NoSQL (que l'on traduit par *not only SQL,* et non comme un « non » au SQL). MongoDB, ElasticSearch, les bases clef-valeur comme Redis, ou les bases de données graphiques comme Neo4j constituent le fer de lance de ces nouvelles solutions. . Beaucoup pensent que la manipulation des mégadonnées implique forcément l'utilisation de technologies NoSQL sensées être plus performantes et que le NoSQL va rapidement remplacer les bases de données relationnelles traditionnelles.

Tableau 22: Comparaison des solutions SQL et NoSQL

Concept	SQL	NoSQL
Représentation des données	Tables	Documents, clef-valeur, graphe, etc…
Schémas	Prédéfinis	Flexibles, « sans schéma »
Préférence de scalabilité	Verticale (plus de puissance)	Horizontale (plus de machines)
Langage	SQL	Dépend de la base de données, par exemple Cypher pour Neo4j
Exemples	Postgres, MySQL, DB2, Oracle, Sqlite	MongoDB, Redis, Neo4j

L'un des points forts mis en avant par les solutions NoSQL est leur nature *shema-free*, en opposition au fait que les bases de données relationnelles possèdent des tables dont le schéma est clairement défini (un nombre donné de colonnes, chacune de type défini). L'avantage d'une solution sans schéma est donc la possibilité d'y intégrer des données non structurées. Le piège du *schema-free* est, en contrepartie, que la logique des données, ou le schéma, est simplement transféré de la base de données au code qui y accède, sans la possibilité pour la base de données, de garantir l'intégrité structurale des données qu'elle contient.

Les bases NoSQL revendiquent également leur facilité de mise à l'échelle horizontale pour s'adapter graduellement à l'augmentation de la charge, mais cela passe par une *eventual consistency* : l'information délivrée à un client peut ne pas être à jour sur un des nœuds. Ce manque de cohérence n'est pas très préjudiciable s'il s'agit du nombre de *likes* sur Twitter, mais bien plus grave s'il s'agit d'informations biologiques.

Face à cet apparent dilemme, la réponse ne passe pas nécessairement par le choix d'une solution au dépend de l'autre, comme nous allons le voir. Ainsi, lorsqu'on s'intéresse à l'architecture des grands du web (Google, Facebook, Twitter, Instagram), on se rend compte que les solutions SQL sont à la base de ces services comme garants des données, et complémentées par des solutions NoSQL pour gérer efficacement le nombre d'accès. Cela est illustré par le Tableau 23. Plutôt que de privilégier une technologie aux dépens de l'autre, la meilleure solution est de réunir SQL et NoSQL dans ce que l'on appelle le *Polyglot Persistence* (Miner, 2012) qui combine l'utilisation de plusieurs systèmes de gestion de bases de données pour profiter des avantages de chacune.

De plus, les solutions SQL possèdent des réponses aux problèmes de mégadonnées, comme le *sharding* qui consiste à diviser une table en plusieurs tables pour optimiser la taille des index et améliorer les temps d'accès. C'est une fonctionnalité massivement exploitée par Facebook par exemple.

La réplication permet de faire face au crash d'un serveur de base de données, mais aussi d'avoir une mise à l'échelle horizontale pour supporter une plus grande charge de requêtes. Elle est implémentée dans les bases de données relationnelles (Postgres, Mysql, etc…) par la synchronisation entre un serveur maître et des serveurs esclaves.

L'intégration du type de données JSON dans les dernières versions de MySQL et Postgres les rapproche encore plus en termes de fonctionnalités des solutions NoSQL. Ce type de colonne assure la validation des données (seulement un JSON valide accepté) et permet d'accéder aux champs du JSON, de les indexer, de les inclure dans les requêtes, etc…

Dans MyGeneFriends, j'ai également choisi de combiner des technologies SQL et NoSQL afin de tirer parti des avantages de chacune.

Tableau 23 : Quelques exemples de l'utilisation de la Polyglot Persistance par des grands acteurs du web

Service	Partie SQL	Partie NoSQL
Instagram	PostgreSQL	Redis
Facebook	MySQL	Memcached, flashcache
Twitter	MySQL	Redis, Flock
LinkedIn	Oracle et MySQL	Voldemort

3.5.1.3 Vitesse vs Norme

Il existe un ensemble de règles de bonne conduite lorsqu'il s'agit d'architecturer les tables d'une base de données. Certaines de ces règles peuvent cependant rentrer en conflit avec les exigences d'efficacité d'accès aux données. Ainsi, diviser une table en plusieurs tables de schéma identique (*sharding*) permet de diviser la taille des tables et des index et d'accélérer les requêtes. Imaginons que l'on veuille stocker les relations entre deux gènes, ces relations provenant de sources différentes. Une table avec les colonnes « gene1, gene2, score, source » serait la plus normalisée, mais aussi la plus lente puisqu'elle contiendra un nombre très important de lignes. Créer les tables *string_related_gene* « gene1, gene2, score » et *orthology_related_gene* « gene1, gene2, score » permet de nettement accélérer la récupération de relations d'une source donnée.

Certains attributs textuels répétitifs comme le biotype d'un transcrit sont, selon les règles, stockés dans une table spécifique, « biotypes » par exemple, et leur identifiant est stocké dans la table d'origine. Cependant, cette approche entraine un grand nombre de jointures supplémentaires réduisant les performances. J'ai donc choisi de stocker à chaque fois ces données textuelles dans la table d'origine.

3.5.1.4 ORM ou non ORM

L'utilisation d'un ORM (*Object Relational Mapping*) qui simplifie les accès à la base de données possède un ensemble d'avantages et d'inconvénients.

Les requêtes sont simples et rapides à écrire, et en réponse, on récupère des objets directement utilisables. De plus, l'utilisation d'un ORM en couple avec une librairie de gestion de JSON comme Jackson (github.com/FasterXML/jackson) permet en quelques lignes de code d'ajouter des fonctionnalités API à la plateforme (requête, récupération des objets, transformation en JSON, envoi au client).

En théorie, un ORM permet de se concentrer sur la logique métier et d'ignorer la syntaxe spécifique de la base de données utilisée. En pratique, dès que la requête devient complexe, il est nécessaire d'utiliser du pur SQL, qui pourra être complété par des conditions plus triviales faites avec la machinerie de l'ORM.

Un des grands soucis avec les ORM est l'impossibilité, pour l'instant du moins, de faire des parcours d'index seuls (*index only scans*). Ces requêtes sont les plus rapides car au lieu d'utiliser l'index pour localiser l'information dans la table, l'information est directement récupérée de l'index. Ainsi un index sur les colonnes (A,B) pourra satisfaire seul la requête « select B from T where A = 'a' ». Les ORM identifient les objets par la colonne *id* de la table, cette colonne est donc toujours chargée par l'ORM empêchant l'utilisation de l'index décrite précédemment.

Malgré ces désavantages, après avoir fait le choix d'une approche sans ORM pour Parsec, le gain en productivité permis par l'ORM EBean m'a paru considérable sur MyGeneFriends et justifie son utilisation.

3.5.1.5 AJAX : un peu, beaucoup, à la folie ?

L'utilité d'un chargement parallélisé de l'information en utilisant la technologie AJAX est indiscutable. Pourtant, la question du nombre d'éléments de la page web à charger par AJAX peut légitimement se poser. Intuitivement, on pensera à charger par AJAX uniquement les fonctionnalités les plus massives ou lentes à calculer. Certains sites adoptent l'autre extrême en assurant toute la navigation sur leur site par AJAX, sans aucun rechargement de page. AJAX a un autre grand avantage mis à part la réduction du temps d'attente du chargement d'une page, c'est de pouvoir gérer le contenu de la page (chargement correct ou erreur) au cas par cas.

J'ai donc fait le choix de charger toutes les parties critiques du site web de MyGeneFriends indépendamment par AJAX, de sorte qu'en cas de problème avec l'une de ces parties un simple cadre informatif apparaitra sur le site web informant l'utilisateur que la fonctionnalité en question n'est pas disponible.

3.5.1.6 API : savoir exposer

L'utilisation d'APIs est indispensable pour améliorer la tolérance à la panne ou au bug en disséminant les fonctionnalités sur plusieurs serveurs (architecture orientée service). Elles permettent également à MyGeneFriends d'exposer ses données (publiques) et fonctionnalités aux autres programmes du laboratoire. C'est pour cela que j'ai tenu à rendre disponibles par API les fonctionnalités les plus importantes de MyGeneFriends.

Ces API sont disponibles sous forme d'un serveur spécialisé permettant d'exécuter des programmes sur la machine. Les programmes actuellement accessibles sont :

- VEP, l'API permet de caractériser une variation sur une position donnée,
- Clustalw, l'API permet de réaliser un alignement entre deux séquences

Un autre serveur permet d'accéder rapidement à des entrées OMIM (téléchargées et traitées automatiquement), par identifiants ou mots clefs. Les entrées entières ou les sections sont alors récupérables sous forme de JSON. L'intérêt de ce serveur par rapport au serveur API OMIM officiel est que son installation locale permet une charge et une rapidité de réponse bien supérieures.

Le serveur principal de MyGeneFriends expose également un certain nombre d'API, dont les requêtes permettant d'accéder au *Social Graph* ou au *News Stream*. Pour des raisons de sécurité, l'accès à la plupart de ces API nécessite d'avoir un compte. Chaque requête renvoie une réponse sous forme de JSON.

Tableau 24: Quelques exemples d'accès API aux services de MyGeneFriends

URL	Description
/api/genes/search/[search_term]	Renvoie les gènes dont le symbole ou synonyme commence par la chaine de caractères soumise.
/api/interests/templates/search[search_term]	Renvoie les *Topics* publics dont le nom commence par la chaine de caractères soumise
/api/genes/network/data/for_publication/[pmid]	Renvoie une liste de gènes liés à cette publication et une liste de toutes les amitiés disponibles sur MyGeneFriends entre ces gènes
/api/genes/network/data/custom/[gene_identifiers]	Renvoie une liste de toutes les amitiés disponibles entre la liste d'identifiants de gènes soumis. Ces identifiants peuvent être des symboles, stable_id ou synonymes.
/api/genes/by_tag/orthology/[tag]	Renvoie la liste de gènes appartenant à un profil évolutionnaire spécifique.

/api/go/list/gene/[gene_id]	Renvoie la liste des termes GO associés à un gène spécifié par son identifiant, avec les valeurs de spécificité associées.
/api/diseases/network/for_metadisease/[metadisease_id]	Renvoie une liste de maladies appartenant à la méta-maladie donnée et une liste d'amitiés disponibles sur MyGeneFriends entre ces maladies
/api/publications/[pmid]	Renvoie la publication associée avec des informations produites par MyGeneFriends comme le nombre total de *Likes* et de *Dislikes* que la publication a reçus.
/api/expression/[gene_id]/[experience_id]	Renvoie l'expression du gène donné pour l'expérience donnée dans tous les tissus de l'expérience avec l'intensité de signal et l'intensité de signal transformée par MyGeneFriends
/api/news/stream/[page_nb]	Renvoie les *news* liées aux acteurs du *Topic* actif. Ces *news* sont accessibles page par page, avec dix *news* par page.

3.5.1.7 Mise à jour automatique

Afin d'exécuter les différents scripts de mise à jour de MyGeneFriends, j'ai choisi d'utiliser le serveur Jenkins installé dans notre laboratoire. D'une part, cette solution me permet de paramétrer les tâches avec des options à chaque exécution. Ainsi chaque tâche peut être exécutée en mode *Dev* (avec les options de développement et sur la base de données de développement) ou *Prod* (avec les options de production et sur la base de données de production). Une option supplémentaire permet de stipuler si la mise à jour doit générer des *news* (inutile lors de la première exécution). Jenkins permet également de relier les tâches entre elles pour que l'exécution sans erreurs d'une tâche lance une tâche fille, grâce à un graphe de dépendances (Figure 37).

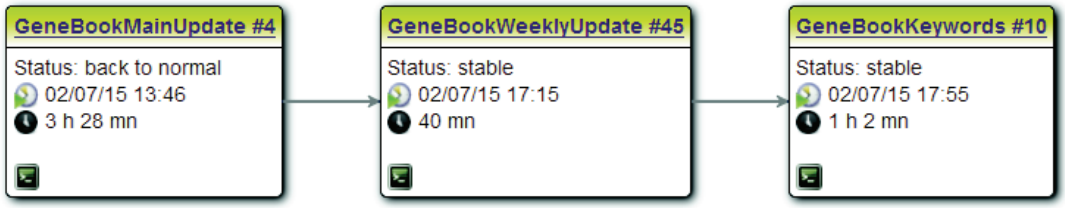


Figure 37: Un exemple de dépendance de tâches gérée par Jenkins.

Etant donné que certaines sources de données sont mises à jour plus souvent que d'autres, et que la mise à jour de certaines données doit obligatoirement entrainer la mise à jour d'autres, MyGeneFriends utilise des tâches distinctes avec une fréquence de lancement différente (Tableau 25).

Tableau 25: Principales tâches exécutées par Jenkins pour MyGeneFriends

Tâche	Description
GeneBookCheckDatasources	Vérification de la disponibilité des sources de données listées dans le fichier de configuration
GeneBookDeployIntegrator	Tests sur le code et copie du code du dossier développement vers le dossier production, contenant les scripts qui seront lancés par Jenkins
GeneBookKeywords	Extraction de mots clef des données textuelles liées aux gènes et aux maladies
GeneBookMainUpdate	Mise à jour de gènes, de transcrits, des termes GO et des données d'expression des gènes
GeneBookPublications	Mise à jour des publications et des liens entre publications et gènes
GeneBookWeeklyUpdate	Maladies et Variations

3.5.1.8 Robustesse et pérennité

Les aspects de robustesse et pérennité concernent le bon fonctionnement de MyGeneFriends, mais également sa maintenance et la facilité de son évolution.

Le bon fonctionnement de MyGeneFriends est assuré d'une part par la sauvegarde journalière de la base de données sur un serveur indépendant et d'autre part, afin d'assurer une *intégration en continue* des scripts d'intégration de données, chaque nuit Jenkins exécute un ensemble de tests sur le code dans le dossier *dev* qui sera alors recopié dans le dossier *prod* si tous les tests ont réussi. C'est à partir de ce dossier *prod* que Jenkins va lancer les tâches d'intégration de données et de génération de *news*. Un autre script lancé par Jenkins vérifie que tous les URL et chemins de fichiers spécifiés dans le fichier de configuration sont valides et qu'aucune ressource n'est brusquement devenue indisponible.

Un système de réponse rapide aux imprévus permet de garantir le bon fonctionnement du site web de MyGeneFriends. Lorsqu'un incident se produit sur le site web, une notification est immédiatement envoyée par une API vers le serveur YouTrack (système de gestion de tickets). Un ticket de l'accident est créé et un mail est envoyé à l'administrateur courant de MyGeneFriends pour le prévenir. Afin d'assurer des corrections de bugs et des mises à jour du site web très rapides, j'ai écrit un script permettant de remplacer l'ancienne version du site web par la nouvelle en quelques secondes.

Enfin, un bouton présent sur le site de MyGeneFriends permet également à tout utilisateur de créer très facilement un ticket pour signaler un bug ou demander une nouvelle fonctionnalité.

3.5.1.9 *Configuration spécialisée pour une intégration efficace des données*

J'ai étendu la librairie python de gestion de fichiers de configuration *argparse* afin de permettre de gérer au mieux les différentes ressources dont j'intègre les données. Ainsi, la configuration est séparée en plusieurs parties. La partie [BasicConf] contient les sources de données par défaut. Les sources définies dans les parties [Dev] et [Prod] disposent d'informations propres à chaque mode (les informations de connexion à une base de données par exemple) et écrasent les sources définies dans [BasicConf]. Enfin, chaque espèce dispose de sources de données propres, et peut écraser les informations contenues dans les parties précédentes. Les informations d'accès aux données spécifiques du génome humain sont par exemple présentes dans la partie [Human].

Les noms des paramètres présents dans le fichier de configuration sont également strictement définis. Chaque nom de paramètre présente des informations séparées par des points permettant au système de le tester correctement. Le premier mot du paramètre doit forcément faire partie de la liste des termes autorisés décrits dans le Tableau 26.

Tableau 26 : Termes autorisés en début des noms des paramètres de configuration

Premier mot du nom du paramètre	Description
Database	Paramètres de connexion à la base de données
Url	URL pour accéder à une ressource HTTP
Ftp	Paramètres d'accès à un serveur FTP
Path	Chemin vers un fichier sur un disque du laboratoire
Species	Informations décrivant une espèce
Parameter	Paramètre utilisé à l'intérieur du programme.

En identifiant la nature de chaque paramètre, un script exécuté chaque jour peut vérifier que les fichiers définis par Path existent toujours et sont lisibles, que les URL pointent vers des pages web disponibles (code 200), que les bases de données sont accessibles, etc…

3.5.2 Choix conceptuels

Les choix conceptuels qui seront présentés ici concernent les décisions qui régissent la « philosophie » d'une ressource web et les bases de l'expérience utilisateur. Tout comme les choix techniques vont améliorer les performances ou la maintenance de l'infrastructure, les choix conceptuels permettront de définir des objectifs à long terme.

3.5.2.1 Egaux mais pas trop ?

L'une des principales questions soulevées par le choix del'architecture de la base de données est à quel point les acteurs présentés comme égaux sur le site web doivent l'être au niveau de l'implémentation. Bien sûr, l'égalité a de gros avantages du point de vue de la réutilisation du code et de la maintenance. Les acteurs gène et maladie partagent les mêmes modèles pour les pages de profil, et l'interface i_Acteur expose un grand nombre de méthodes permettant leur traitement indifférencié par une grande partie du système. L'acteur humain se singularise par plusieurs points: des mécanismes d'authentification lui sont associés, il a son mot à dire sur l'acceptation de demandes d'amitié…

La question de savoir à quel point les acteurs non humains doivent être actifs est une autre problématique, car tout un ensemble d'informations peut être présenté par le système à l'utilisateur sous forme passive ou active. Prenons les suggestions d'amitié (présentés plus en détail dans le chapitre 3.2). Le système détermine quels sont les gènes et maladies qui peuvent vous intéresser. Il y a deux manières de présenter cette information à l'utilisateur. La manière active est de faire en sorte que les gènes et les maladies les mieux notés par le système de suggestion envoient des demandes d'amitié à l'utilisateur de la même manière que d'autres humains lui envoient des demandes d'amitié. La manière passive est d'afficher un encadré avec une suggestion de gènes et de maladies avec lesquelles l'utilisateur peut vouloir devenir ami. La première manière de faire donne l'impression à l'utilisateur que les gènes et maladies sont des acteurs actifs et interagissent avec lui, mais cela peut vite devenir lassant de recevoir de nombreuses notifications. J'ai donc choisi l'approche passive qui est plus discrète. Un compromis peut toutefois être trouvé, par exemple, seul le meilleur candidat envoie à l'utilisateur une demande d'amitié, et les autres se retrouvent dans la liste de suggestions.

3.5.2.2 Modèle centralisé ou décentralisé

Pendant le développement de MyGeneFriends, la question s'est posée de savoir s'il fallait concevoir un système téléchargeable et installable par chaque laboratoire intéressé (modèle décentralisé) ou s'il ne devait y avoir qu'un seul serveur MyGeneFriends hébergé au LBGI et accessible par tous les laboratoires (modèle centralisé). Chacun de ses modèles a, là aussi, ses avantages et ses inconvénients.

Le modèle décentralisé permet une très bonne confidentialité des données, étant donné qu'elles ne sortent pas du laboratoire des utilisateurs. Ce modèle nécessite que le laboratoire possède un agent capable d'installer et de maintenir le système.

Le modèle centralisé, inspiré de l'approche choisie par la plupart des réseaux sociaux, permet à n'importe qui de se connecter très facilement à MyGeneFriends sans aucun aspect technique à gérer. Il peut en revanche soulever des réticences de la part des utilisateurs à voir leurs données de recherche hébergées en dehors de leur laboratoire.

Un troisième critère fut décisif pour le choix du modèle : l'aspect réseau. La grande force des réseaux sociaux modernes est la taille de leur *Social Graph* et des mécanismes de personnalisation, prédiction, suggestion, partage, communication qu'il offre. On ne saurait imaginer un Facebook dans chaque collège ou centre de recherche. Pour cette raison, nous nous sommes orientés vers un modèle centralisé de MyGeneFriends, bien qu'il entraine le défi de gagner et conserver la confiance des utilisateurs afin qu'ils acceptent de partager leurs données avec notre réseau.

3.6 Tourné vers l'expert

MyGeneFriends est en premier lieu tourné vers l'expert en lui proposant une interface reprenant un ensemble de codes popularisés par les réseaux sociaux. D'une part la navigation dans son réseau d'amitiés et l'identification d'informations pertinentes. D'autre part, la suggestion d'informations et d'acteurs d'intérêt. Puis enfin, une personnalisation du flux d'informations biologiques sous forme d'un flux de *news* qui lui donnera accès à une information filtrée, personnalisée et présentée avec une visualisation adaptée.

3.6.1 Réseautage par visualisation interactive

Une des forces de MyGeneFriends est de considérer l'expert humain comme partie intégrante du réseau biologique, l'humain est donc à la fois une partie de l'information contenue dans le réseau et l'utilisateur de ce réseau. Le réseautage est une partie importante de l'utilisation de MyGeneFriends puisqu'il permet de découvrir de nouveaux amis intéressants mais également, d'identifier des *hubs* ou propriétés particulières de certains réseaux.

J'utilise la librairie JavaScript vis.js pour présenter sous forme de réseau les liens d'amitié entre différents acteurs. Parmi les différentes simulations de physique disponibles, j'ai choisi d'utiliser le modèle Barnes Hut qui permet le mieux former des groupes d'acteurs hautement interconnectés. On peut voir dans le manuscrit de la publication sur MyGeneFriends, la capacité du système à grouper les gènes associés à une publication en des groupes distincts biologiquement pertinents. La visualisation en réseau permet également d'évaluer la cohérence d'un *Topic* et de son sujet de recherche. La Figure 38 montre par exemple le réseau d'amitiés entre gènes ajoutés à un *Topic* expérimental, mêlant des gènes relatifs au syndrome du BBS et des

gènes relatifs aux fibres musculaires (myosine, actine, etc…). Pour construire ce *Topic,* j'utilise simplement la fonctionnalité *Copy* me permettant de construire un *Topic* à partir de gènes présents dans d'autres *Topics*. Si cette distribution de gènes se trouvait dans un vrai *Topic*, ce serait le signe qu'il devrait être divisé en deux *Topic* distincts.

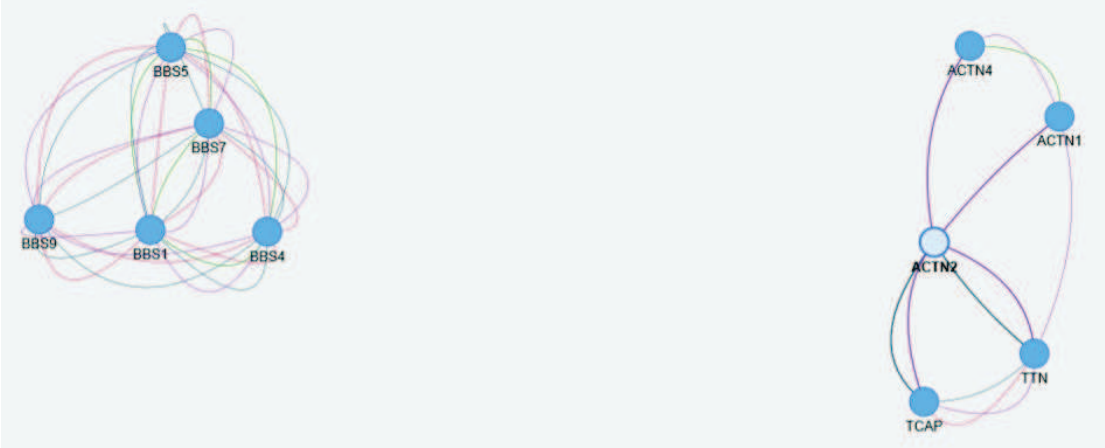


Figure 38: Réseau de gènes appartenant à un Topic mêlant deux thèmes distincts, des gènes liés au Bardet-Biedl et des gènes de fibres musculaires. Les liens violets représentent les amitiés par STRING (Szklarczyk et al., 2015), les liens gris les amitiés par amies maladies communes, les liens verts les amitiés dues au profil d'évolution similaire, et les liens rouges les amitiés par amis humains publics communs.

3.6.2 Suggestions

Comme nous l'avons vu dans l'introduction, les mécanismes de suggestion sont très fréquents sur des sites de commerce, de streaming musical ou vidéo, et les réseaux sociaux. Le principe des suggestions est de se baser sur des informations que l'on possède sur l'utilisateur afin de trouver des informations qui s'y rapprochent le plus, et qui sont donc susceptibles de l'intéresser.

MyGeneFriends utilise les suggestions à la fois pour aider l'expert à étendre son réseau, et pour identifier des informations d'intérêt. Pour cela, j'ai implémenté des mécanismes de suggestion d'amis gènes et maladies (décris dans le manuscrit MyGeneFriends) afin d'aider l'utilisateur à étendre son *Topic*. De plus, les suggestions de publications peuvent apporter de nouvelles informations sur les acteurs déjà présents, mais aussi présenter d'autres acteurs liés à ces publications. Lorsque l'expert s'intéresse à des profils d'expression tissus-spécifiques des gènes, MyGeneFriends lui suggérera les gènes qui correspondent au mieux au profil qu'il a construit interactivement. Enfin, la suggestion d'auteurs de publications permet à l'utilisateur de se faire une idée sur les principaux auteurs du champ de recherche qui l'intéresse.

3.6.2.1 Suggestions de publications

Les publications sont une source d'informations non structurée très utilisée par les scientifiques, il était donc évident que MyGeneFriends devait être capable de proposer aux experts les publications qui seraient les plus à même de les intéresser. Plusieurs mécanismes de proposition de publications peuvent être envisagés :

- Proposer les publications qui correspondent aux mots clefs que l'utilisateur saisit. C'est le mécanisme utilisé par PubMed ou n'importe quel gestionnaire de publications. Cette solution peut être intéressante si l'utilisateur sait exactement ce qu'il veut mais l'enferme dans ses connaissances.
- Partir des mots clefs fournis par l'utilisateur et utiliser des ontologies pour étendre ces mots clefs. Cette approche est par exemple utilisée par GoPubmed (Doms and Schroeder, 2005).
- Proposer des publications ressemblant aux publications que l'utilisateur possède déjà ou a aimé. Cette méthode est également très utilisée par les gestionnaires de publications comme, par exemple, ReadCube (www.readcube.com). Elle nécessite que l'utilisateur ait déjà choisi des publications.
- Proposer des publications mentionnant le nom des gènes et maladies avec lesquelles l'utilisateur est ami. Cette méthode ne permettrait pas de trouver des publications qui traitent du sujet qui intéresse l'utilisateur sans citer explicitement les gènes ou maladies avec lesquelles il est ami.
- Constituer automatiquement un set de mots clefs avec un score et les utiliser pour rechercher des publications. C'est la méthode que j'ai choisie car elle n'oblige pas l'utilisateur à sélectionner des mots clefs, et ne l'enferme pas dans les publications directement liées aux acteurs qu'il a sélectionnés. Cette approche m'a intéressé car elle permet d'explorer le champ d'intérêt courant de l'utilisateur, mais également de définir l'utilisateur en fonction de ses intérêts.

Postgres possède plusieurs fonctionnalités intéressantes pour indexer des champs textuels et les interroger. Il est cependant loin d'être aussi puissant que les solutions dédiées, et en particulier un des acteurs majeurs dans ce domaine : ElasticSearch. Afin d'attribuer un score aux publications par rapport aux mots clefs extraits automatiquement par MyGeneFriends, j'ai installé un serveur ElasticSearch synchronisé quotidiennement avec la table des publications de la base de données MyGeneFriends. ElasticSearch permet de « booster » certains mots clef dans la requête en augmentant ou diminuant leur pertinence, ne rechercher que des parties d'un mot, etc…

J'utilise donc les scores d'associations entre mots clefs et les acteurs du *Topic* actif comme *boost* pour les mots clefs lors de l'interrogation d'ElasticSearch. De plus, j'interroge ElasticSearch avec la racine de chaque mot clef afin de prendre en compte toutes les variantes.

Un exemple de chaine de caractères construite par MyGeneFriends pour être envoyée à ElasticSearch pourrait être :

spermatocyt*^2.6 vasodil*^1.6 self-mutil*^1.2

Cela signifie qu'ElasticSearch recherchera les publications contenant les mots dont la racine est spermatocyt (spermatocyte, spermatocytes, spermatocytic, etc…), vasodil (vasodilators, vasodilation, vasodilating, vasodilator, etc…) et self-mutil (self-mutilating, self-mutilation) en scorant le mieux les documents qui les possèdent tous les trois, et en scorant mieux les publications qui possèdent les mots dont la racine est spermatocyt que celles qui possèdent les mots dont la racine est « self-mutil ».

3.6.2.1.1 Analyse d'un cas

Afin de montrer l'intérêt de cette approche, on peut analyser un exemple de publications suggérées à partir du contenu d'un *Topic*. Nous pourrons voir si l'objectif de l'approche choisie a été respecté, à savoir obtenir des publications pertinentes par rapport au sujet du *Topic* sans être cantonné aux gènes et maladies présents dans le *Topic*.

Je commence par construire un *Topic* orienté sur plusieurs récepteurs nucléaires :

Symbole	Description
NR1H4 (FXR)	Récepteur à l'acide biliaire
PPARA	Récepteur activé par la prolifération des peroxysomes
RARA	Récepteur à l'acide rétinoïque
VDR	Récepteur à la vitamine D

Je récupère grâce à une API (/api/keywords/active) le profil que MyGeneFriends a construit pour représenter ce qui peut m'intéresser. Parmi les vingt mots clefs qui le constituent, on retrouve dans les mieux notés, ceux qui décrivent l'ensemble de ces gènes, sous forme racinisée : ('regul', 'receptor', 'transcript', 'bind', 'promot', 'nuclear', 'ligand-depend', 'sequence-specif'), quelques mots clefs décrivent plus précisément certains des gènes : ('peroxisom', 'bile', 'vitamin', 'acid', 'dihydroxyvitamin')

Je m'attends donc à ce que MyGeneFriends me recommande des publications liées à mon sujet d'intérêt, et non à des gènes précis. Les dix publications les mieux évaluées recommandées par MyGeneFriends sont présentées dans le Tableau 27.

Tableau 27: Publications recommandées par MyGeneFriends en se basant sur le contenu du Topic que j'ai construit.

PMID	Titre
18768987	Transactivation of the hepatitis B virus core promoter by the **nuclear receptor FXRalpha**.
19666701	**Vitamin D3 and its nuclear receptor** increase the expression and activity of the human proton-coupled folate transporter.
15890193	The human peroxisome proliferator-activated receptor delta gene is a primary target of **1alpha,25-dihydroxyvitamin D3 and its nuclear receptor**.
16357103	An essential role of the CAAT/**enhancer** binding protein-alpha in the **vitamin D-induced** expression of the human steroid/bile acid-sulfotransferase (SULT2A1).
15521013	Evidence for a new human CYP1A1 regulation pathway involving **PPAR-alpha and 2 PPRE sites**.
23703729	Epigenetic regulation of MicroRNA-122 by peroxisome proliferator activated receptor-gamma and hepatitis b virus X protein in hepatocellular carcinoma cells.
15878955	**General receptor** for phosphoinositides 1, a novel **repressor** of thyroid hormone receptor action that prevents deoxyribonucleic acid binding.
9171239	Analysis of the functional role of **steroid receptor** coactivator-1 in ligand-induced transactivation by thyroid hormone receptor.
12562861	Regulation of human delta-6 desaturase gene transcription: identification of a **functional direct repeat-1 element**.
12914524	Functional interference between **estrogen-related receptor alpha** and **peroxisome proliferator-activated receptor alph**a/9-cis-retinoic acid receptor alpha heterodimer complex in the nuclear receptor response element-1 of the medium chain acyl-coenzyme A dehydrogenase gene.

Ces publications concernent comme attendu un vaste domaine, tout en se concentrant sur le sujet qui m'intéresse, à savoir la régulation de transcription. Certaines des publications traitent des gènes présents dans le *Topic,* d'autres, de gènes au rôle semblable.

La suggestion de publications dans son implémentation actuelle est entièrement basée sur les mots clef et la correspondance entre les mots clef et les publications dans la base de données de MyGeneFriends. Cela pose trois problèmes :

- Si la qualité de racinisation de certains termes biologiques essentiels est mauvaise, cela peut empêcher de bien évaluer les publications les plus pertinentes.
- Des publications anciennes peuvent se retrouver bien notées alors qu'elles sont la plupart du temps moins pertinentes que des publications récentes. La date de la publication doit donc intervenir dans l'évaluation du score et du classement des publications pour la suggestion. ElasticSearch permet d'améliorer le score des publications datant de moins d'un mois ou d'un an par exemple, mais il s'agit de trouver le bon équilibre entre pertinence et récence.

- MyGeneFriends ne télécharge les publications que lorsqu'un lien entre gène et publication apparait dans le fichier gene2pubmed du NCBI. La quantité de publications dans la base de données de MyGeneFriends (un demi-million à l'heure actuelle) n'est donc qu'un sous-ensemble des publications disponibles sur Pubmed, ce qui réduit la quantité de publications qui peuvent être recommandées.

3.6.2.2 Suggestion d'auteurs de publications

L'objectif de MyGeneFriends est également d'aider l'expert à identifier les personnes les plus influentes dans le domaine qui l'intéresse. Pour cela, MyGeneFriends identifie l'auteur qui a contribué au plus de publications relatives aux acteurs du *Topic* actif *via* un traitement en plusieurs étapes :

1. Récupérer toutes les publications liées aux gènes du *Topic* actif.
2. Si la publication a plus de trois auteurs, identifier uniquement le premier, deuxième et dernier auteur de la publication (dans le cas contraire, on considère tous les auteurs).
3. Calculer le nombre de publications par auteur.
4. Classer les auteurs par nombre de publications et afficher le ou les auteurs les plus productifs.

Si l'on poursuit l'étude du cas décrit dans le paragraphe précédent, on peut voir que le système de recommandation suggère également l'auteur qui pourrait m'intéresser. Ce système n'est pas basé sur les mots clefs, mais sur le lien établi par le NCBI entre gènes et publications. Pour ce *Topic* MyGeneFriends me recommande Staels Bart avec un score de 15 (il est premier, deuxième ou dernier auteur de quinze publications liées à des gènes de mon *Topic*). Les « Récepteurs nucléaires dans le syndrome métabolique » sont en effet un des principaux thèmes de recherche du professeur Staels Bart.

3.6.2.3 Suggestion de gènes en fonction d'un profil d'expression

MyGeneFriends possède également un outil permettant à l'expert de décrire un profil d'expression en sélectionnant sur une vue schématique d'un corps humain, un ensemble de tissus dans lequel un gène est ou n'est pas exprimé (Figure 39). MyGeneFriends va alors lui suggérer les gènes qui correspondent le mieux au profil défini.

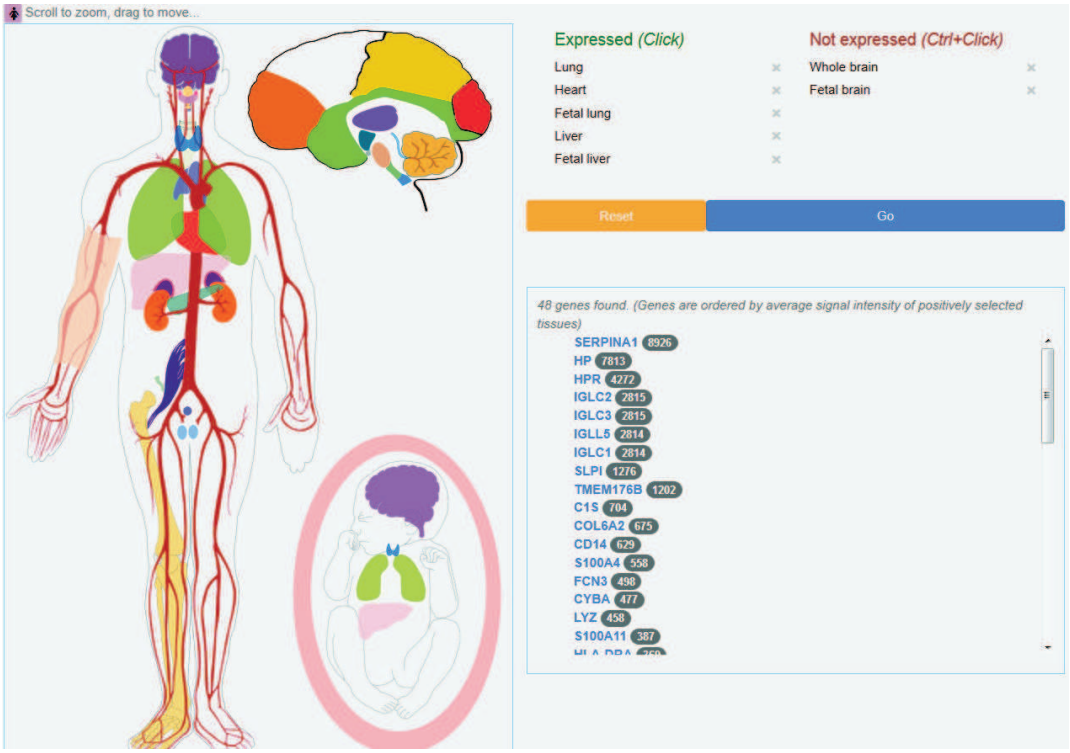


Figure 39: Filtre d'expression permettant à l'utilisateur de construire un profil d'expression qui permettra à MyGeneFriends de suggérer les gènes qui le respectent le mieux. Les gènes sont classés par l'intensité moyenne de signal dans les tissus sélectionnés positivement.

Le système sélectionne dans un premier temps, tous les gènes qui ne sont exprimés dans aucun tissu sélectionné dans la catégorie *Not expressed* et qui s'expriment dans tous les tissus de la catégorie *Expressed*. Je me base sur le seuil d'intensité de signal de 19.02 défini dans notre laboratoire pour cette expérience comme limite d'expression ou non d'un gène. Le classement des gènes se fait ensuite sur la base de l'intensité moyenne du signal des tissus sélectionnés dans la catégorie « *Expressed* ». La Figure 40 montre que le gène SERPINA1, le mieux noté par le système, est effectivement très exprimé dans des organes comme les poumons et le foie, et n'est pas exprimé dans le cerveau.

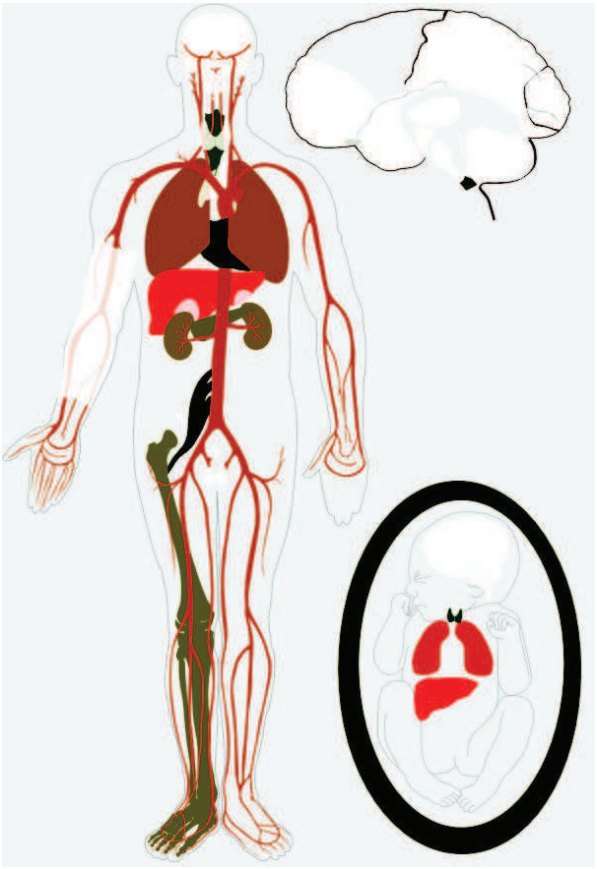


Figure 40: Expression du gène SERPINA1 dans différents tissus, visualisée sous forme de heatmap sur le schéma du corps humain adulte, du cerveau et du fœtus. Les tissus dans lesquels le gène est fortement exprimé sont de couleur qui tend vers le rouge, les tissus dans lesquels le gène est peu exprimé tendent vers le bleu. Les tissus dans lesquels le gène n'est pas exprimé sont en blanc.

Par la suite, ce classement pourra être amélioré en tenant compte de l'ordre des gènes les plus exprimés dans un tissu, et non de la valeur de l'intensité d'expression. Cela permettra de nuancer les tissus dont l'intensité de signal maximale est beaucoup plus forte que celle d'autres tissus.

3.6.3 Détection des pics d'intérêt grâce à la Timeline

Une visualisation chronologique (*Timeline)* permet d'évaluer l'évolution de l'intérêt de la communauté de chercheurs pour les gènes d'un *Topic* sous la forme du nombre de publications relatives aux gènes du *Topic* par année. Lorsque le *Topic* contient peu de gènes, la popularité de chaque gène est affichée et peut être comparée. Lorsque le Topic contient beaucoup de gènes, une tendance globale basée sur la somme totale des publications est visualisée. Le *Topic* « Ciliopathies » créé dans notre laboratoire contient 56 gènes liés aux ciliopathies. La visualisation sous forme de *Timeline* (Figure 41) de la popularité de ces gènes permet de détecter deux pics. Le premier intervient au début des années deux mille, et le second vers l'année 2010.

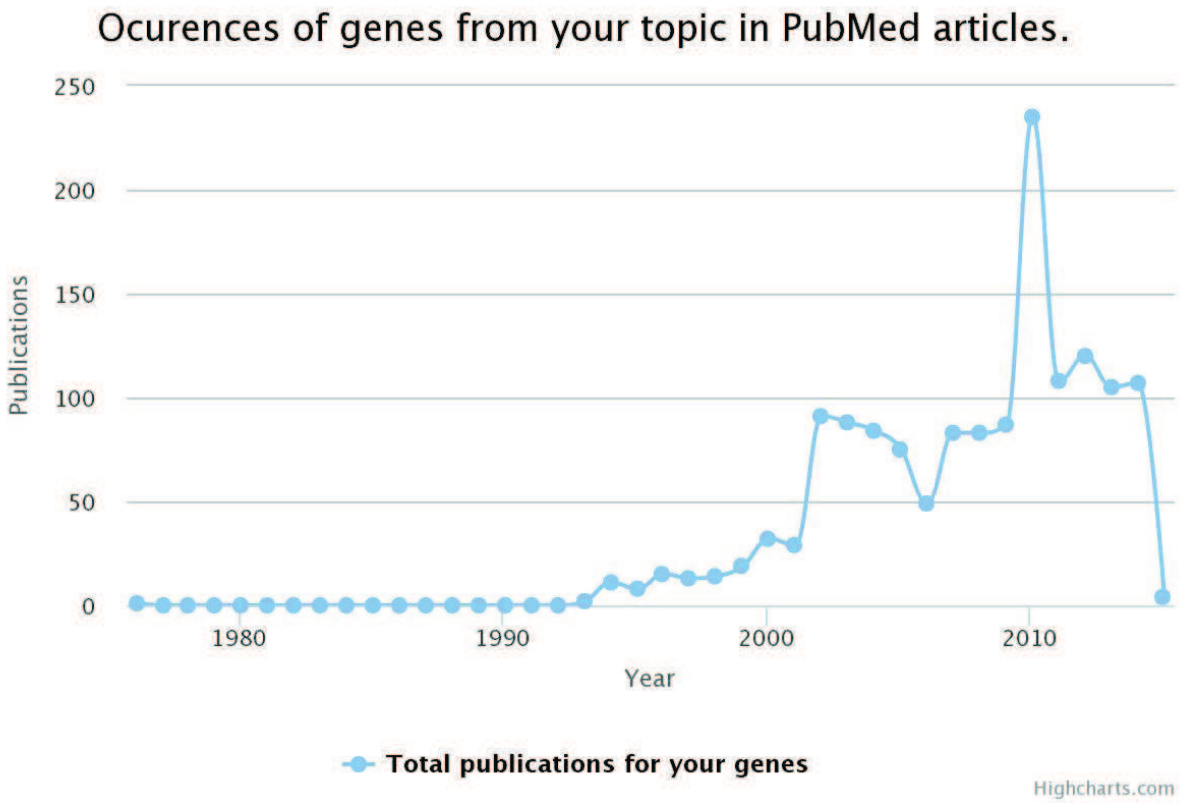


Figure 41: Représentation en Timeline produite par MyGeneFriends afin de suivre l'évolution du nombre de publications liés à des gènes liés aux ciliopathies selon les années.

Après analyse de la littérature, le premier pic semble lié à la découverte du rôle essentiel du cil dans de nombreuses fonctions de l'organisme, en particulier pour le fonctionnement des reins (Barr and Sternberg, 1999; Hildebrandt et al., 2011).

Le deuxième pic pourrait être dû à l'utilisation massive de méthodes à haut débit (séquençage d'exomes de patients par exemple) et des données et découvertes qui en ont résulté.

3.6.4 Les News : des acteurs qui vous tiennent informés sur leur vie

La génération de *news* se produit au moment de l'intégration de nouvelles données dans la base de données de MyGeneFriends. Des adaptateurs spécifiques se chargent de se connecter à une source particulière de données et d'en traiter le contenu afin de le transformer en une liste d'entrées intégrables dans la table *Diff* de la base de données MyGeneFriends (Tableau 28). La table *Diff* permet de stocker tous les évènements NEW, UPDATE et DELETE survenus pendant la mise à jour.

Tableau 28: Informations relatives à une entrée de la table Diff, décrivant une news

Champ	Description
catégorie	Définit le type de *news* : base de données ou publication
source	Définit la source de données associée à cette actualité
entity	Associée au nom de la table dans laquelle les données sont intégrées, elle décrit l'entité précise concernée par la nouveauté : gène, maladie, lien gène à gène, etc…
entity_event	Décrit le type de *news* : NEW, UPDATE ou DELETE
entity_id	Permet de retrouver la ligne précise de la table qui a été mise à jour
entity_name	Permet de nommer une entité non acteur (par exemple un transcrit) dans la nouveauté
feature	Correspond à une colonne précise, et donc un détail précis sur l'entité : symbole du gène, description de la maladie, etc…
old_value	Contient l'ancienne valeur, dans le cas d'un UPDATE ou DELETE
new_value	Contient la nouvelle valeur dans le cas d'un NEW ou UPDATE
diff_value	Calculée au moment de la génération de *news* pour des comparaisons complexes entre *old_value* et *new_value*, comme une comparaison de textes ou un alignement

L'intégration en elle-même se passe en plusieurs étapes :

1. Définition d'une clef comme identifiant unique d'une entrée,
2. Construction d'un cache des entrées locales pour cette entité,
3. Construction de caches pour les entités parentes de cette entité.

Pour chaque entrée distante, j'utilise la clef pour la comparer à l'entrée locale. Si la clef n'est pas présente dans la base de données locale, on considère cela comme un évènement NEW, l'entrée est ajoutée à la table correspondante et une entrée est créée dans la table *Diff* pour enregistrer cet évènement.

Lorsque la clef est trouvée dans la table locale, pour chaque colonne de l'entrée je compare les valeurs locale et distante, et lorsqu'elles sont différentes, un évènement UPDATE est ajouté à la table *Diff*. Comme le montre la figure Figure 42, de nombreux acteurs sont concernés par la mise à jour de leurs propriétés.

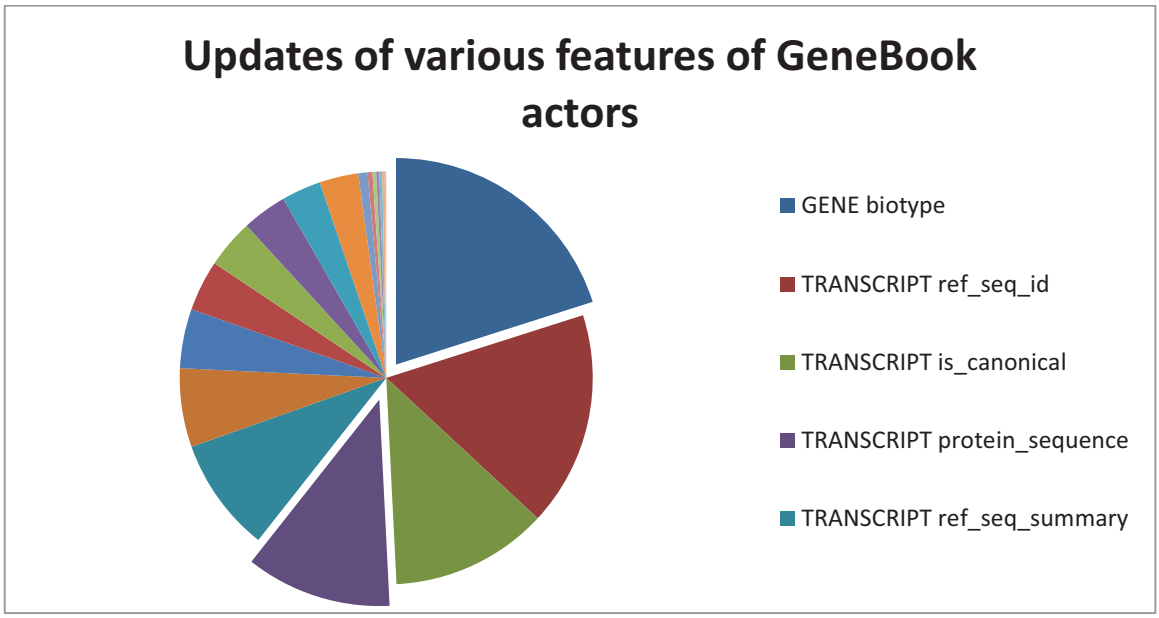


Figure 42: Proportion d'évènements de type mise à jour liés aux différentes propriétés (biotype, description, etc...) des acteurs de MyGeneFriends ou des objets (ex: Transcrit) qui leur sont associés.

La génération de *news* sur les gènes apparaissant dans les publications devra être améliorée. Ainsi, afin de détecter le lien entre gène et publication, je me suis basé sur le fichier *gene2pubmed*, mis à jour quotidiennement au NCBI. C'est un fichier comprenant tous les liens entre publications et gènes de beaucoup d'espèces dont les données sont utilisées par le NCBI pour afficher les publications liées à un gène et les gènes liés à une publication. MyGeneFriends a dans sa démarche de génération de *news* collecté les statistiques d'évolution de ce fichier dans le temps (création de nouvelles relations et suppression de relations existantes).

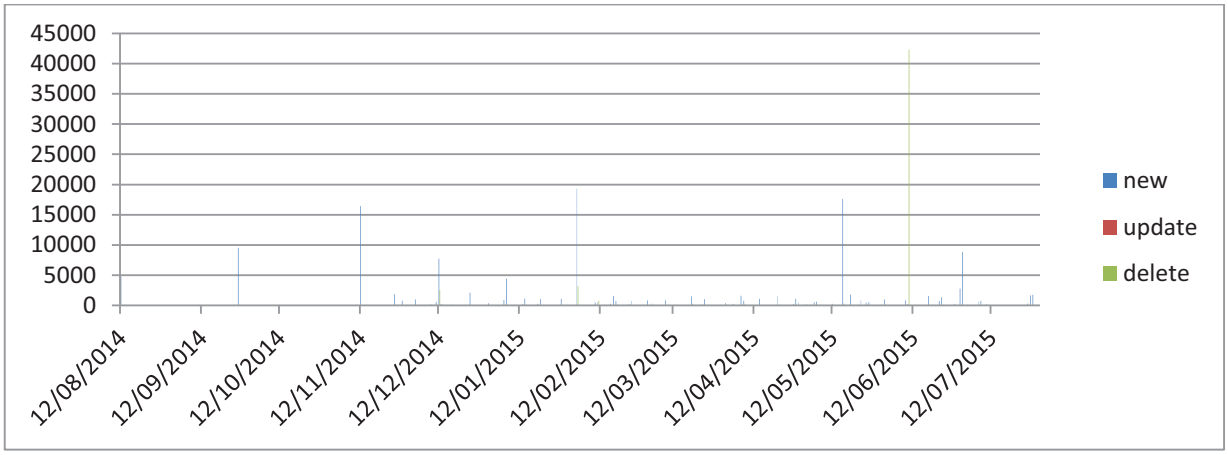


Figure 43: Principaux évènements détectés dans le fichier gene2pubmed par le système de génération de news de MyGeneFriends durant une année

Plusieurs points remettent en cause l'utilisation future de ce fichier pour générer des *news*. D'une part, comme le montre la Figure 43, il est très instable et bien que de nombreuses relations sont ajoutées quotidiennement, de nombreuses relations ont été supprimées. L'exemple le plus extrême est la suppression de 42000 relations en juin 2015. D'autre part, les algorithmes du NCBI semblent retravailler l'ensemble des publications à chaque fois puisque lors des mises à jour du fichier une partie de nouvelles relations créées concernent de vielles publications. Enfin, leur méthode ne détecte pas tous les liens entre gènes et publications.

Deux alternatives sont possibles. Pubtator propose un fichier *gene2pubtator* reprenant le même format que *gene2pubmed* et les annotations de *gene2pubmed* mais en les agrémentant d'annotations réalisées par GNormPlus (Wei et al., 2015). Ce fichier est mis à jour une fois par mois mais offre un ensemble de relations entre gènes et maladies plus complet et serait une alternative à court terme. A long terme, on peut envisager le traitement par MyGeneFriends de toutes les nouvelles publications journalières pour y détecter les gènes grâce à une instance locale de GNormPlus.

3.7 Comprendre l'évolution de l'information biologique grâce à notre archive de News

Depuis presque un an de fonctionnement, le pipeline automatique de génération de *news* de MyGeneFriends a produit plus d'un million de *news*. Mis à part le flux de *news* que propose MyGeneFriends à ses utilisateurs, cette archive de *news* permet d'analyser l'évolution de l'information dans les banques bio-informatiques et d'y détecter des tendances qui peuvent nous éclairer sur la manière dont évolue notre connaissance des acteurs biologiques. Autant pour les gènes que pour les maladies, ces *news* permettent de faire ressortir quelques tendances intéressantes.

On pourrait se poser de nombreuses questions et tenter d'y répondre avec les informations disponibles dans notre archive, mais je ne présenterai ici que deux exemples d'analyse, dans lesquels nous nous intéresserons à l'évolution des types des gènes et à l'évolution des noms de maladies.

3.7.1 Evolution des biotypes

On peut voir sur la Figure 42 que le biotype de nombreux gènes évolue en fonction des mises à jour (plus de 14000 updates). Une analyse plus approfondie (Figure 44) révèle que ces changements concernent surtout l'ajout de précision de l'annotation des pseudogènes (s'ils sont transcrits ou processés par exemple) et le reclassement de gènes codant pour des protéines en lincRNA (*long non-coding RNA*).

Schématiquement, les pseudogènes sont souvent des fragments de gènes autrefois fonctionnels que des modifications ont réduit au silence (Goodhead and Darby, 2015). L'intérêt croissant pour les pseudogènes et une précision accrue de leur annotation peut être liée à la découverte récente

(permise par le séquençage à haut débit et les avancées de la recherche sur les ARN non codants) de leur rôle dans la régulation transcriptionelle et post-transcriptionelle ainsi que leur implication dans de multiples cancers (Xiao-Jie et al., 2015).

Les lincRNA sont de longs transcrits (plus de 200 nucléotides) qui ne codent pas pour des protéines (Khorkova et al., 2015). Le reclassement de gènes précédemment considérés comme protéiques ou comme transcrits processés, en lincRNA, peut s'expliquer par la découverte du rôle majeur des lincRNA dans de nombreux mécanismes biologiques, liés aux maladies respiratoires et cardiovasculaires (Liu et al., 2015), les maladies auto-immunes (Sigdel et al., 2015), les cancers (Blume et al., 2015), et bien d'autres.

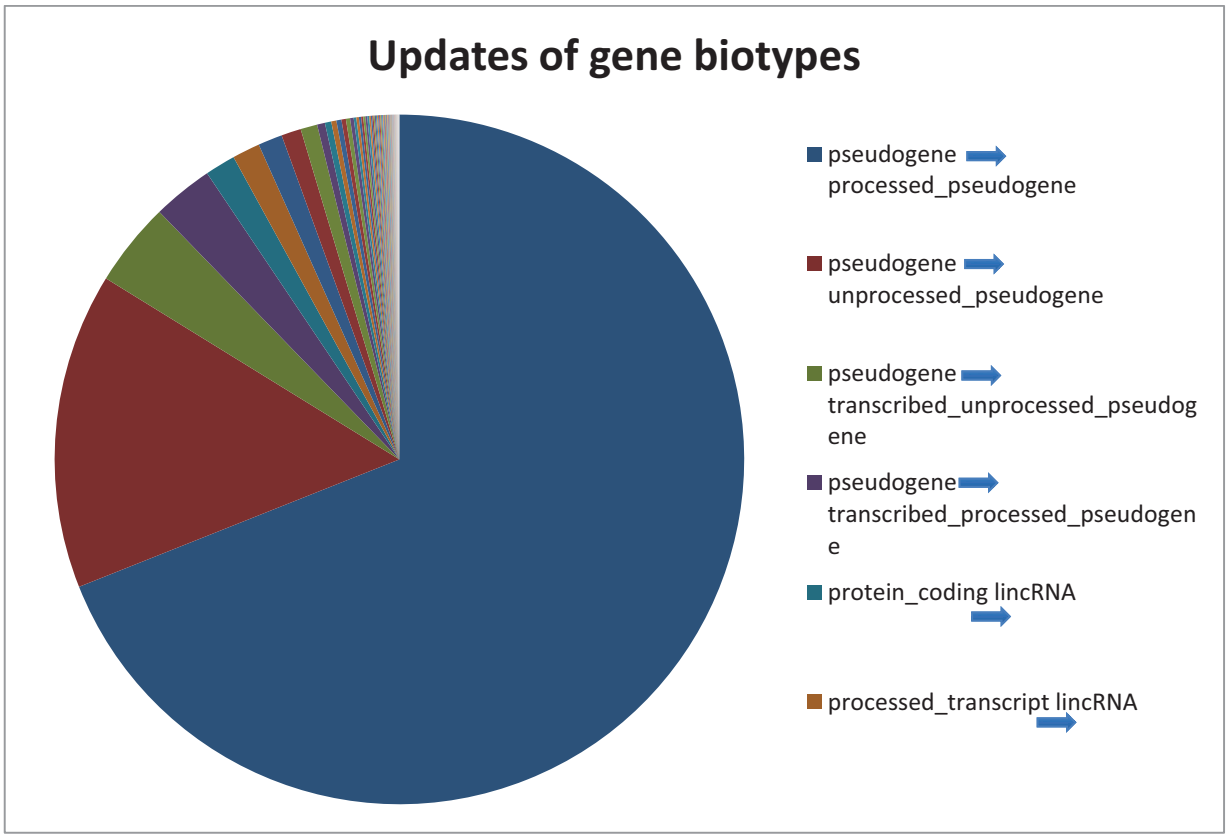


Figure 44: Principales mises à jour concernant la modification du type de gène Ensembl, les types pseudogène et lincRNA sont parmi les plus touchés. Pour chaque couleur, la légende montre d'abord le terme avant la mise à jour (ex : pseudogene) puis le terme après la mise à jour (ex : processed_pseudogene)

3.7.2 Evolution des noms de maladies

Les maladies sont également un acteur très actif et en pleine évolution. La description et le nom des maladies sont les informations qui évoluent le plus (Figure 45). Une analyse manuelle de ces changements a permis d'identifier plusieurs tendances (Tableau 29).

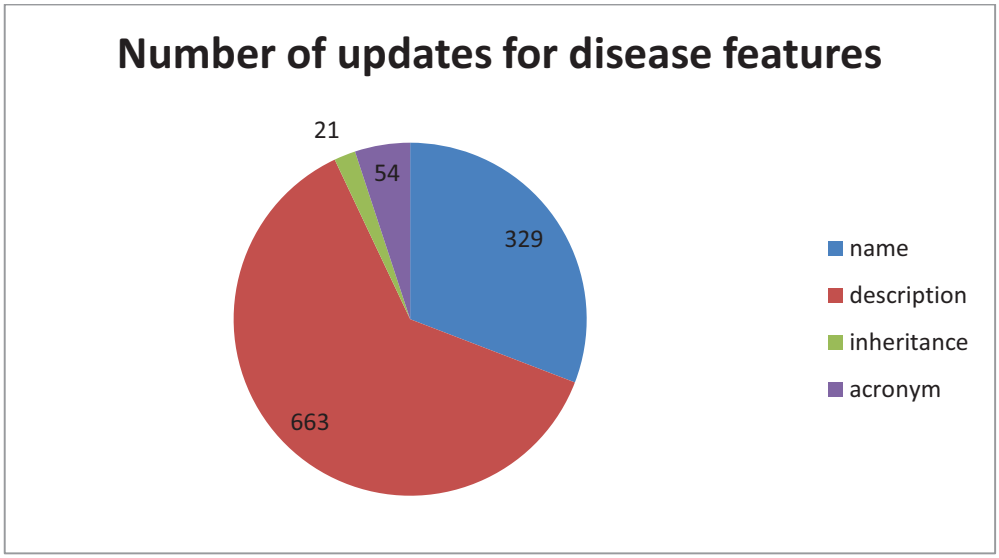


Figure 45: Les mises à jour des différentes informations liées aux maladies. L'ajout ou le retrait d'informations ne sont pas pris en compte, uniquement leur modification.

Tableau 29: Quelques tendances majoritaires dans l'évolution du nom des maladies

Tendance	Exemple
Modification de l'ordre des mots	« *Unilateral renal agenesis* » est devenu « *Renal agenesis, unilateral* »
Corrections d'orthographe	« *Jumping frenchman of maine*" est devenu "*Jumping frenchmen of maine* »
Un adjectif devient un nom	« *Pancreatic adenoma* » est devenu « *Adenoma of pancreas* »
Remplacement de phénotypes supplémentaires par le numéro d'un sous type d'une maladie	« *Myasthenic syndrome, congenital, with pre- and postsynaptic defects* » est devenu « *Myasthenic syndrome, congenital, 8* »

Cette base de noms alternatifs nous permettra de détecter encore mieux les maladies de bases de données différentes qui devraient être fusionnées. Il nous suffira pour cela d'entrainer un classificateur bayésien afin d'évaluer la probabilité que deux noms différents décrivent la même maladie.

Conclusions et perspectives

Dans une vision synthétique, on peut dire que ce travail de thèse a été traversé par deux centres d'intérêt qui me touchent tout particulièrement, à savoir : le rôle des mégadonnées, en l'occurrence, en biologie, et la place de l'humain face aux machines et aux mégadonnées.

Les mégadonnées révolutionnent notre monde !

Filles de la Physique et de l'Astronomie, leur irruption sur la Toile, en particulier dans les réseaux sociaux, a fait basculer notre perception des mégadonnées. Ainsi, leur volume démesuré est passé de la notion d'entrave, de collecte, gestion, visualisation et analyse, à celle d'opportunité de connaissance, personnalisation et adaptation fine aux problèmes.

Dans ce contexte, leur émergence récente en Biologie offre, à mes yeux, un atout unique de bénéficier de solutions éprouvées de valorisation des mégadonnées au profit de l'étude de la fonction et de l'évolution des gènes et génomes et de leurs rôles dans les maladies humaines.

Les machines remplacent les hommes !

Cette crainte, qui remonte aux débuts de la mécanisation, n'est pas infondée à l'heure où le remplacement de la main-d'œuvre par des machines s'étend aux emplois de plus en plus qualifiés. Cependant, là encore, les mégadonnées et leur complexité inhérente nous obligent à nuancer cette affirmation et à comprendre que les connaissances procédant des mégadonnées sont multiples, allant de savoirs explicites faciles à identifier et à extraire rapidement avec la puissance de calcul dont nous disposons (*i.e.* définitions, fréquences, liens simples entre évènements…) jusqu'à des connaissances et fonctionnalités non incluses qu'il nous revient d'imaginer et d'inventer (i.e. blogs et donneurs d'alerte, réalité augmentée, internet des objets…). Dans ce face à face hommes-machines, les machines et l'informatique ont remporté de grands succès et, à cet égard, les réalisations d'IBM (DeepBlue, Watson…) sont emblématiques remisant l'homme au second plan aux échecs, à Jeopardy et bientôt peut-être, dans le diagnostique de maladies complexes. Cependant, elles connaissent également des échecs cuisants pour l'humanité (effet d'emballement des bourses, analyse des risques terroristes du 11 Septembre… météorologie !) qui nous montrent que l'étude de systèmes complexes nécessite encore une interaction profonde entre l'expert humain et les machines ; interaction qui passera, sans doute, par la création de nouveaux rapports entre homme et machine.

De nouveau, dans ce contexte, la biologie ouvre de réelles perspectives car elle constitue, à mes yeux, le domaine de prédilection pour inventer ce nouveau rapport entre chercheur et information. En effet, en tant que système particulièrement complexe, constitué d'éléments hétérogènes, transitoires, hautement interdépendants et devant s'adapter à des environnements incroyablement disparates, la biologie questionne constamment les limites et l'efficacité des

algorithmes les plus élaborés. De plus, au regard de l'importance des enjeux en termes de devenir de l'humanité, de sa santé, de son environnement, la biologie peut (doit ?) sensibiliser et mobiliser toutes les forces de la société.

Dans ce cadre, ma thèse a été axée sur le développement de ressources visant à faciliter l'accès personnalisé et efficace aux grandes quantités d'informations, la visualisation de ces données et leur interconnexion en réseaux et l'exploitation de l'information biologique. J'ai pu développer de nouvelles approches et outils en tirant partie des évolutions en matière de visualisation des données, d'interaction en temps réel avec les données, et des concepts inventés ou popularisés par les réseaux sociaux. Dans l'absolu, l'ensemble de ces travaux et réalisations a abouti, ou a contribué, aux développements de trois outils : Parsec, OrthoInspector et MyGeneFriends. Cependant, on peut distinguer plusieurs jalons et enseignements. Ainsi, Parsec m'a poussé à rechercher la rapidité et l'instantanéité dans les traitements des données afin de répondre en temps réel aux requêtes des utilisateurs. J'ai été pour la première fois, confronté à l'hétérogénéité des données biologiques et aux difficultés de leur intégration. OrthoInspector m'a permis de saisir l'importance d'une visualisation et interaction efficace des données, et la force des réseaux et ce, dans le cadre d'un logiciel déjà établi, impliquant mon adaptation aux langages et choix de mes prédécesseurs. Ces deux expériences m'ont permis de définir à quoi devait ressembler le projet phare de ma thèse, MyGeneFriends, qui cristallisait ma volonté de créer une ressource web qui serait orientée vers une large communauté de chercheurs à laquelle je voulais proposer une relation plus simple, étroite et amusante avec les mégadonnées.

Ce travail n'est qu'un premier pas sur le chemin que j'aimerais parcourir. Une grande partie de la thèse a consisté à mettre en place l'architecture de MyGeneFriends. D'une part, créer les scripts d'analyse et d'intégration de données, de génération de *news*. D'autre part, tester et implémenter de multiples concepts d'interaction homme-information permettant une expérience d'utilisation fluide et cohérente.

De nombreux aspects de MyGeneFriends ont nécessité une longue réflexion, des essais et erreurs, et parfois de longs débats. Ainsi le premier concept de MyGeneFriends était axé uniquement autour des gènes (d'où son premier nom, GeneBook), mais nous nous sommes rapidement rendus compte que les maladies devaient rejoindre le réseau de par leur importance dans la recherche biomédicale. Nous avons longtemps discuté sur la manière la plus juste d'intégrer des maladies de sources différentes, et notre désir d'offrir une visualisation des variations différente de celles qui existent et la plus utile au personnel médical, nous a poussés à organiser une réunion avec nos collègues praticiens.

Bien qu'ayant passé beaucoup de temps à travailler l'architecture et l'intégration des données, j'ai surtout pu mettre en place des fonctionnalités clef, reflétant l'esprit de MyGeneFriends :

- Un flux d'actualités personnalisé
- Des suggestions d'acteurs et de publications, ainsi que des scores d'affinité pour juger à quel point un profil d'acteur peut nous intéresser
- Un accès simplifié et visuel aux données biologiques relatives aux gènes et maladies
- Création de profils décrivant les intérêts scientifiques publics de chaque utilisateur sous forme de vecteur de mots clefs scorés.
- Aide à la lecture, en attirant le regard du lecteur vers les passages de textes qui risquent de l'intéresser
- Réseautage, et visualisation du « social graph » sous forme d'un réseau avec groupement naturel d'acteurs hautement connectés

L'avenir est encore plus excitant !

A court terme, les perspectives concernent la finalisation de fonctionnalités de MyGeneFriends dont le concept a été testé, mais que je n'ai pas eu le temps d'optimiser. Cela inclut la fouille de textes, les suggestions d'amitié, les scores d'affinité et le classement des gènes par niveau d'expression dans différents tissus. Certaines fonctionnalités très intéressantes que je n'ai pas eu le temps d'ajouter devront être implémentées comme la visualisation de réseaux mixtes de phénotypes-maladies ; la possibilité de recherche de sites génomiques avec l'intégration d'un accès à Parsec *via* la barre de recherche de MyGeneFriends ; l'affichage d'acteurs gènes proximaux. Les utilisateurs pourront également « approuver » des termes GO ou HPO des acteurs, comme ils le font pour les compétences de leurs collègues sur LinkedIn. L'intégration des mots clefs représente une base pour une fonctionnalité qui m'intéresse énormément : l'attribution de scores aux paragraphes de publications entières, en fonction des intérêts courants de l'acteur humain, lui permettant de ne lire que les passages qui sont susceptibles de l'intéresser !

A moyen terme, les perspectives concernent l'extension de MyGeneFriends aux concepts modernes de communication que ce soit sur le terrain de facilités de développement ou sur celui des échanges entre membres de MyGeneFriends.

Ainsi, un des aspects les plus intéressants de Facebook, auquel je n'ai malheureusement pas eu le temps de m'atteler lors de la conception de MyGeneFriends, est la possibilité pour des développeurs tiers d'ajouter leurs propres applications. Cette approche est bien plus intéressante qu'une infrastructure simplement modulaire, car elle permet à chaque application de profiter immédiatement du graphe social complet du réseau.

Une autre fonctionnalité sur laquelle j'aimerais travailler concerne le '*GeneDating*', qui permettrait de favoriser les rencontres et collaborations entre chercheurs partageant les mêmes centres d'intérêts, dans les congrès scientifiques par exemple. MyGeneFriends construit déjà des profils décrivant les intérêts scientifiques publics de chaque utilisateur sous forme de vecteur. On peut donc calculer la compatibilité scientifique entre deux chercheurs par une formule comme la *cosine similarity* par exemple. Une extension de MyGeneFriends au monde réel serait donc une application qui ferait vibrer le smartphone d'un acteur humain dans un congrès lorsqu'il se trouve à coté de quelqu'un possédant un vecteur semblable (et s'intéressant donc publiquement aux mêmes choses).

Enfin, les micro-publications seraient un moyen structuré de communiquer les résultats de son travail sous forme de *micro-blogging* semblable à celui de Facebook ou Twitter. Chaque micro-publication concernerait un sujet très précis, devrait contenir moins de 200 mots et être composée uniquement de mots appartenant à un vocabulaire contrôlé.

Enfin, je pense que l'avenir réside dans une transformation encore plus profonde des communications entre mégadonnées et chercheurs, transformation qui impliquera sans doute, une autonomie et une capacité de décision grandissantes des données vis-à-vis de l'humain. Sans conteste, les développements autour de l'internet des objets préfigurent ces évolutions qui autorisent nos frigos à nous réveiller en pleine nuit quand le lait vient à manquer ou notre aspirateur à nous harceler si son sac est plein ! Dans le cadre de MyGeneFriends, cette transformation pourrait impliquer une place plus importante donnée aux deux acteurs non-humains, à savoir les gènes et les maladies. Ainsi, une des multiples avancées potentielles pourrait permettre aux gènes et maladies de 'juger' de façon autonome (dans une version simplifiée '*like*' ou '*dislike*') d'autres acteurs ou publications, fournissant ainsi, le 'point de vue' d'un gène ou d'une maladie sur une publication où ils apparaissent. De même, on peut aisément concevoir que, dans le cadre des développements de l'internet des objets, nos acteurs de la biotechnologie moderne (séquenceurs, spectromètres, scanners,…) deviennent des acteurs de MyGeneFriends et communiquent, de façon autonome, avec certains gènes de MyGeneFriends, pour les 'prévenir', en temps réel, de leurs présences dans des résultats expérimentaux inédits.

Annexes

1 Annexe 1

OrthoInspector 2.0: Software and database updates Bioinformatics, 2015

BIOINFORMATICS APPLICATIONS NOTE Vol. 31 no. 3 2015, pages 447–448
doi:10.1093/bioinformatics/btu642

Databases and ontologies

Advance Access publication October 1, 2014

OrthoInspector 2.0: Software and database updates

Benjamin Linard[1,2], Alexis Allot[1], Raphaël Schneider[1], Can Morel[1], Raymond Ripp[1], Marc Bigler[1], Julie D. Thompson[1], Olivier Poch[1] and Odile Lecompte[1,*]

[1]LBGI, Computer Science Department, ICube, UMR 7357, University of Strasbourg, CNRS, Fédération de médecine translationnelle, 4 rue Kirschleger 67085 Strasbourg, France and [2]Department of Life Sciences, Natural History Museum, Cromwell Road, London SW7 5BD

Associate Editor: Jonathan Wren

ABSTRACT

Summary: We previously developed OrthoInspector, a package incorporating an original algorithm for the detection of orthology and inparalogy relations between different species. We have added new functionalities to the package. While its original algorithm was not modified, performing similar orthology predictions, we facilitated the prediction of very large databases (thousands of proteomes), refurbished its graphical interface, added new visualization tools for comparative genomics/protein family analysis and facilitated its deployment in a network environment. Finally, we have released three online databases of precomputed orthology relationships.

Availability: Package and databases are freely available at http://lbgi.fr/orthoinspector with all major browsers supported.

Contact: odile.lecompte@unistra.fr

Supplementary information: Supplementary data are available at *Bioinformatics* online.

Received on July 1, 2014; revised on September 4, 2014; accepted on September 21, 2014

1 INTRODUCTION

High throughput comparative analyses, functional annotations, or evolutionary studies involve massive transfers of information between organisms using orthology inference. As defined by Fitch (1970), orthologs are homologous genes that diverged from an ancestral speciation event, while paralogs emerged from a duplication event. Today, it is widely accepted that orthologs generally share similar functions, whereas paralogs can potentially evolve new functions. Numerous algorithms based on the results of Blast searches were developed to infer orthology relations (see Kristensen *et al*., 2011; Altenhoff and Dessimoz, 2012 for reviews). We previously developed an orthology inference algorithm also based on Blast and implemented it in the OrthoInspector (OI) package (Linard et al., 2011). Our focus was to maintain a balance between sensitivity and specificity (Linard *et al*., 2011; Dalquen *et al*., 2013). Contrary to most other packages, OI is not limited to predictions and provides tools for comprehensive mining of large orthology databases, nonspecialist use through a desktop graphical interface and can easily be deployed in a network environment. Here, we describe the main improvements implemented in the second version of the OrthoInspector package.

*To whom correspondence should be addressed.

2 DISTINCTIVE FEATURES

2.1 Requirements

OI 2.0 requires the NCBI blast+ toolkit (Camacho *et al*., 2009), a Java compatible operating system and a local or remote access to a SQL database. Any SQL engine is compatible as long as it allows Java connections via a JDBC driver. OI provides an extended support for Postgresql and MySQL engines, for which more operations are automated. To predict new orthologs, OI requires as inputs one proteome in FASTA format per species of interest and a blast all-against-all constructed from the same proteomes.

2.2 Implementation and network exploitation

The package is based on database/client interactions and separated into three clients (Fig. 1A): a command-line for initial orthology predictions and database installation, a query command-line to retrieve precomputed predictions and a graphical interface designed for desktop querying and data visualization. Limited computational resources are required as the clients delegate many operations to the SQL engines and their optimized dataset manipulation capabilities. Management of large databases (thousands of species) is, however, facilitated by several options from the installation command-line. The components of the graphical interface make easier the mining and the visualization of complex orthology relationships for nonspecialists. A small 6six species database dump can be downloaded from the website to rapidly test these tools. The three clients can be used on a single desktop computer but can also be deployed in a network. Orthology databases can then be stored in one server while several users/pipelines exploit the clients for various purposes (Fig. 1B). Then, OI responsiveness will mainly depend on server and network speed.

2.3 Eukaryote and prokaryote databases

We have constructed three orthology databases with OI. The first database, named "Prokaryotes," contains orthologs between 120 Archaea and 1568 Bacteria proteomes. The "Eukaryotes" dataset contains 259 complete proteomes and covers all main eukaryotic phyla, from unicellular organisms to plants, fungi, and metazoan. The last dataset, "Quest For Orthologs" (QFO), combines bacteria, archaea and eukaryote proteomes and corresponds to the latest version of the orthology benchmark released by the QFO Consortium (Dessimoz *et al*., 2012).

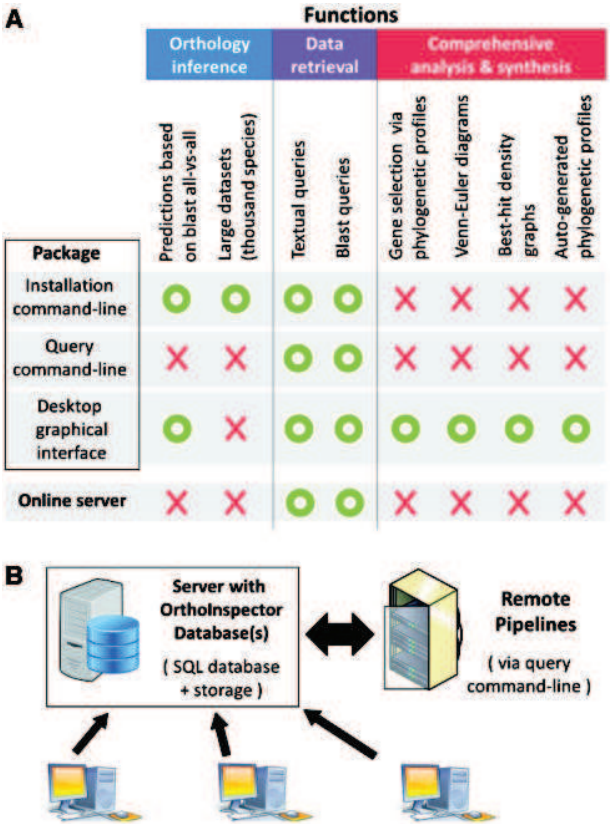


Fig. 1. Package organization and main functionalities. (A) Command-lines are used for initial orthology inference, database querying and to handle large datasets. The graphical interface is used for all other tasks. (B) Typical deployment of the package on a network

Supplementary File S1 lists all the species included in these databases and their taxonomy.

3 MAIN ADDITIONS

3.1 Large-scale phylogenetic profiles

Several analyses are now supported by an interactive tree of life in the graphical interface, facilitating in particular the establishment of "phylogenetic profile" queries. A selection of presence/absence criteria at different levels in the tree, allows the extraction of large-scale sets of genes that respect the profile through the orthology criteria. For instance, one can retrieve all Microsporidia sequences with orthologs in Basidiomycota but not in Ascomycota (Supplementary Fig. S2).

3.2 Best-hit density graph and Euler diagrams

Two new visualization tools are now part of the graphical interface. First, the "best-hit density graph" is designed to analyze the orthologous relationships linking genes in a particular family and to reveal potential subfamilies. Through a dynamic graph representation of BLAST best hits linking a protein family, the user can explore conservation patterns within the set by modifying the BLAST score or *E*-value thresholds on the fly (Supplementary File S3). This tool can be used to adapt the delineation of subfamilies to the evolutionary rate of the family under consideration or to a given phylogenetic scope. Second, Venn diagrams (3 organisms), but also more complex Euler diagrams (>3 organisms), can be generated. When more than 3 organisms are considered, diagram overlaps are based on the VennEuler library (Wilkinson, 2012), which provides a statistical framework to estimate the best possible circle-based representation.

3.3 Web server

Our precomputed datasets (Eukaryote, Prokaryote, and QFO) can be accessed via a web server allowing ortholog retrieval by textual or Blastp searches. A list of organisms can be selected with an interactive species tree. Orthology relationships corresponding to the query sequence are compiled in a table format with phylum color codes to facilitate the analysis of phylum specific orthology distributions and produce a user-friendly and intuitive overview of the revealed evolutionary history (see Supplementary File S4 for a case study). All datasets can be downloaded in CSV and OrthoXML formats (Schmitt *et al.*, 2011).

4 CONCLUSION

OI is a package dedicated to the efficient calculation and analysis of orthology data and allows a rapid and intuitive analysis of relationships associated with large clades. The OI web server allows the retrieval of precomputed orthology data from thousands of eukaryote and prokaryote proteomes.

ACKNOWLEDGEMENTS

Funding: This work was supported by the ANR [grant ANR-10-INSB-05-01 FRISBI, grant ANR-10-BINF-03-02 BIPBIP] and Institute funds from the CNRS, the Faculté de Médecine de Strasbourg and the Université de Strasbourg.

Conflict of interest: none declared.

REFERENCES

Altenhoff,A.M. and Dessimoz,C. (2012) Inferring orthology and paralogy. *Methods Mol. Biol.*, **855**, 259–279.

Camacho,C. *et al.* (2009) BLAST+: architecture and applications. *BMC Bioinformatics*, **10**, 421.

Dalquen,D.A. *et al.* (2013) The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. *PLoS One*, **8**, e56925.

Dessimoz,C. *et al.* (2012) Toward community standards in the quest for orthologs. *Bioinformatics*, **28**, 900–904.

Fitch,W.M. (1970) Distinguishing homologous from analogous proteins. *Syst. Zool.*, **19**, 99–113.

Kristensen,D.M. *et al.* (2011) Computational methods for Gene Orthology inference. *Brief Bioinform.*, **12**, 379–391.

Linard,B. *et al.* (2011) OrthoInspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, **12**, 11.

Schmitt,T. *et al.* (2011) Letter to the editor: SeqXML and OrthoXML: standards for sequence and orthology information. *Brief Bioinform.*, **12**, 485–488.

Wilkinson,L. (2012) Exact and approximate area-proportional circular Venn and Euler diagrams. *IEEE Trans. Vis. Comput. Graph.*, **18**, 321–331.

2 Annexe 2

Retinoic acid receptor subtype-specific transcriptotypes in the early zebrafish embryo Mol Endocrinol, 2014

Retinoic Acid Receptor Subtype-Specific Transcriptotypes in the Early Zebrafish Embryo

Eric Samarut,* Cyril Gaudin,* Sandrine Hughes, Benjamin Gillet, Simon de Bernard, Pierre-Emmanuel Jouve, Laurent Buffat, Alexis Allot, Odile Lecompte, Liubov Berekelya, Cécile Rochette-Egly, and Vincent Laudet

Institut de Génomique Fonctionnelle de Lyon (E.S., C.G., S.H., B.G., L.B., V.L.), Université de Lyon, Université Lyon 1, Centre National de la Recherche Scientifique (CNRS), Ecole Normale Supérieure de Lyon, 69364 Lyon Cedex 07, France; Institut de Génétique et de Biologie Moléculaire et Cellulaire (E.S., A.A., O.L., C.R.-E.), Institut National de la Sante et de la Recherche Medicale, U596, CNRS, UMR7104, Université de Strasbourg, BP 10142, 67404 Illkirch Cedex, France.; and AltraBio SAS (S.B., P.-E.J., L.B.), Lyon, France

Retinoic acid (RA) controls many aspects of embryonic development by binding to specific receptors (retinoic acid receptors [RARs]) that regulate complex transcriptional networks. Three different RAR subtypes are present in vertebrates and play both common and specific roles in transducing RA signaling. Specific activities of each receptor subtype can be correlated with its exclusive expression pattern, whereas shared activities between different subtypes are generally assimilated to functional redundancy. However, the question remains whether some subtype-specific activity still exists in regions or organs coexpressing multiple RAR subtypes. We tackled this issue at the transcriptional level using early zebrafish embryo as a model. Using morpholino knockdown, we specifically invalidated the zebrafish endogenous RAR subtypes in an in vivo context. After building up a list of RA-responsive genes in the zebrafish gastrula through a whole-transcriptome analysis, we compared this panel of genes with those that still respond to RA in embryos lacking one or another RAR subtype. Our work reveals that RAR subtypes do not have fully redundant functions at the transcriptional level but can transduce RA signal in a subtype-specific fashion. As a result, we define RAR subtype-specific transcriptotypes that correspond to repertoires of genes activated by different RAR subtypes. Finally, we found genes of the RA pathway (*cyp26a1*, *raraa*) the regulation of which by RA is highly robust and can even resist the knockdown of all RARs. This suggests that RA-responsive genes are differentially sensitive to alterations in the RA pathway and, in particular, *cyp26a1* and *raraa* are under a high pressure to maintain signaling integrity. **(*Molecular Endocrinology* 28: 260–272, 2014)**

Retinoic acid (RA) is the main active metabolite of vitamin A and has highly pleiotropic effects, being involved in major cellular processes such as cell proliferation and differentiation (1–3). It acts during development as a crucial morphogen for axes patterning and organogenesis during development and regulates adult homeostasis (4–6). Its wide action relies on the fact that it controls a complex gene-regulatory network and cross talk with several other important signaling pathways (eg, fibroblast growth factor, Hox etc.) controlling physiology and embryonic development. At the molecular level, RA acts through binding to nuclear retinoic acid receptors (RARs) that harbor a modular structure encompassing two main structured domains, a DNA-binding domain (DBD) and a ligand-binding domain, and work as ligand-dependent transcription factors (7–10). Basically, RARs work as heterodimers with retinoid X receptors, another nuclear receptor, following a classical ON/OFF binary

ISSN Print 0888-8809 ISSN Online 1944-9917
Printed in U.S.A.

Received November 5, 2013. Accepted December 12, 2013.
First Published Online January 8, 2014

* E.S. and C.G. contributed equally to this work.

Abbreviations: ATRA, all-*trans*-retinoic acid; DBD, DNA-binding domain; DR, direct repeat; hpf, hours postfertilization; NR, nuclear receptor; NTD, N-terminal domain; PPAR, peroxisome proliferator-activated receptor; qPCR, quantitative PCR; RA, retinoic acid; RAR, RA receptor; RARE, retinoic acid response element;

model of action (11). In the absence of ligand (apo state), RAR/ retinoid X receptor heterodimers are thought to be bound to specific sequences located in the regulatory sequences of target genes, named RAREs for retinoic acid response elements, where they interact with large corepressor protein complexes that keep the chromatin in a silent state, thus preventing transcription (OFF) (12). Upon binding of RA (holo state), drastic changes in the conformation of the RAR ligand-binding domain allow the dissociation of corepressors and favor the binding of large coactivator complexes endowing enzymatic activities that modify the surrounding chromatin to an active state (ON) (13, 14). As a result, transcriptional machinery can be recruited in a sequential and coordinated manner to activate transcription of the associated genes.

There are 3 RAR subtypes in mammals: α (NR1B1), β (NR1B2), and γ (NR1B3) that are encoded by distinct genes (15). The specific functions of each of them have been deeply investigated in mutant mice lacking RARα, RARβ, and/or RARγ (16, 17). Although RAR single-knockout mutants displayed congenital and postnatal abnormalities, they were viable and did not recapitulate the retinoic acid-deprived phenotypes defining the vitamin A deficiency syndromes (18, 19). Recapitulation of vitamin A deficiency phenotypes was only observed in double-knockout mutants in which at least 2 retinoid receptors were invalidated leading to the conclusion that RARs are, at least partially, functionally redundant.

As a result of a third round of whole-genome duplication in the teleost lineage, teleost fishes doubled their *rar* genes. However, in the zebrafish genome only 4 *rar* genes are present, *raraa*, *rarab*, *rarga*, *rargb*, because of the secondary loss of the β-subtypes, specifically in zebrafish (20–22). Although different isoforms (generated by alternative promoter usage and/or alternative splicing) for each subtypes are known in the mouse (23, 24), the existence of such isoforms for all zebrafish RARs has never been described. Knockdown of RARs using morpholino-oligonucleotides in zebrafish also revealed some shared and specific functions for each RAR subtype. Indeed, RARα-B is necessary for pectoral fin development whereas RARγ-A and RARγ-B are required for correct pharyngeal arches patterning (25). In addition, it is worth noting that during zebrafish embryogenesis, RARs can depict both exclusive and overlapping expression patterns (20, 21, 25). As an example, between 35 and 48 hours postfertilization (hpf), retina expresses only the RARα-B subtype, the craniofacial mesoderm expresses specifically RARγ-A subtype, whereas all 4 RARs are expressed in branchial arches. Thus, the specific function of one RAR subtype can therefore be correlated to its particular expression pattern. However, in regions where different RAR subtypes are coexpressed, it has never been addressed whether their functions are fully redundant or whether one subtype with specific intrinsic properties drives specific responses. For example, some embryonic regions, such as the pectoral fin bud, express several RAR subtypes (ie, RARα-B, RARγ-A, RARγ-B), but only one is required for the development of the organ suggesting that, even expressed in the same region, each RAR subtype could have specific functions.

Shedding light on such functional subtype specificity would be of crucial interest for the understanding of the mechanisms controlling RAR activity in vivo. Here, we tackled this issue in vivo at the transcriptomic level using knockdown strategy in the early zebrafish embryo. Zebrafish gastrula presents many advantages as being an easily accessible in vivo model, but with no highly differentiated structures or tissue. Moreover, it allows easy knockdown of a protein of interest by morpholino injection in an in vivo context, targeting endogenous proteins. We show that the different RAR subtypes regulate different repertoires of genes, providing evidence for RAR subtype-specific transcriptional activity in vivo. Furthermore, we observed that the expression of some RA-target genes is highly robust and could participate in the maintenance of the integrity of the RA pathway in vivo.

Materials and Methods

Fish stocks

AB-TU zebrafish strains were reared at 28.5°C and kept under a 10-hour dark, 14-hour light cycle and staged as described elsewhere (26).

Treatment of zebrafish embryos

All-*trans*-retinoic acid (ATRA, Sigma) was diluted in 100% ethanol at a stock concentration of 10^{-2} M. Wild-type embryos were treated from the stock solution to a final concentration of 10^{-7} M, diluted in embryo medium. Negative control embryos were treated with 100% ethanol as vehicle. The treatment occurred from the sphere stage (4 hpf) until the 75%-epiboly stage (8 hpf). Embryos were fixed in RNAlater (Ambion) and kept overnight at 4°C. They were processed the day after or frozen at −20°C.

Morpholino injection

Morpholino oligonucleotides were purchased from GeneTools and resuspended in ultrapure water at 2 mM stock solution. Of each translation-blocking morpholino oligonucleotides 3.5 ng were injected into one-cell stage embryos using a Femto-Jet Eppendorf microinjector. Sequences of morpholinos targeting zebrafish RARs are provided in Figure 2C. α-morphants used for transcriptome analysis are injected with ATG-blocker morpholinos against *raraa*_001, *raraa*_201, *rarab*_001, and *rarab*_201, whereas γ-morphants are injected with ATG-blocker

morpholinos against *rarga*_001, *rarga*_201, *rargb*_001, *rargb*_201. Full morphants are injected with a combination of ATG-blocker morpholinos against all RARs. Morpholinos against other nuclear receptors are listed here: peroxisome proliferator-activated receptor (PPAR)βa, AGCTGCCGCCGCTGCCCTCATCACT; PPARβb, GGTTCTC CGGTCTGACTGCTAAATC; nuclear receptor (NR)2F1A, AGACGCTAACT ACCATTGCCATATC; NR2F1B, CATGGCTGGAATCTTTT ACGCGAAT; NR2F2, AGCCTCTCCACACTACCA TTGCCAT; NR2F5, CACTGATTTACTACCAT TGCCATGC; RARα-A_Ex2-Int2, TAAATCTAGTGTCTTACCTTGCAGC; RARα-B_Ex2-Int2, AGTTGGAAGCGTGGCCTT ACCTTAC. Splice-blocker morpholino efficiency was checked by RT-PCR from total RNA extracts (primer sequences available upon request).

Total RNA extraction and SOLiD library preparation

Two independent clutches were injected corresponding to experimental duplicates. Total RNA from RNAlater-fixed embryos was extracted using RNAsolv reagent (Omega Biotek) following the manufacturer's standard protocol. About 10 μg of total RNA (RNA Integrity Number > 9) per sample (30–40 embryos) was purified onto Dynabeads Oligo$(dT)_{25}$ (Life Technologies) yielding in between 100 and 500 ng of polyA-purified RNA. Multiplexed libraries were prepared following the SOLiD total RNA-Seq kit protocol provided by Life Technologies, including a conversion step to adapt to our 5500Wildfire SOLiD sequencer. RNA quantitation and quality assessment was performed on a 2200 TapeStation (Agilent Technologies).

Sequencing and data analysis

Sequencing has been performed on a SOLiD 5500 sequencer upgraded with the Wildfire technology. Preanalysis of the raw data and mapping have been achieved with Lifescope 2_5.1. Ensembl zebrafish genome Zv9 release 71 from April 2013 has been used as a reference genome. Only genes with more than one count per million in at least 2 samples were kept for the statistical analysis. Raw library size normalization factors were computed using the trimmed mean of M-values normalization method (27) in the R package edgeR. Function voom from the R package limma (28) was then applied to the count data to yield log2 counts per million and associated observational-level weights. A linear model was fit to each gene by using function lmFit from limma to compute a clutch-corrected average expression level for each treatment condition. Statistical contrasts were then computed for comparisons of interest. The empirical Bayes method was used to compute moderated *P* values that were then corrected for multiple comparisons using the Benjamini and Hochberg's false discovery rate controlling procedure (29). The hierarchical agglomerative clusterings of genes (associated with given heat maps) were performed with the following parameters: 1) metric: euclidean distances between genes (genes are represented by their log [fold-changes] values); and 2) linkage criterion: complete linkage .

In situ hybridization and probe cloning

Specific probes for each RAR (2 isoforms for each 4 zebrafish RAR subtypes) corresponding to their 5′-untranslated region and first exon (sequences upon request) were cloned within the pCS2+ vector using the TOPO TA cloning kit (Invitrogen). After sequencing, antisense probes were in vitro transcribed using SP6 or T7 RNA polymerases. Whole-mount in situ hybridization of zebrafish embryos was performed as described by Thisse and Thisse (30). Stained embryos were kept in 80% glycerol and photographed.

RT-quantitative PCR (qPCR)

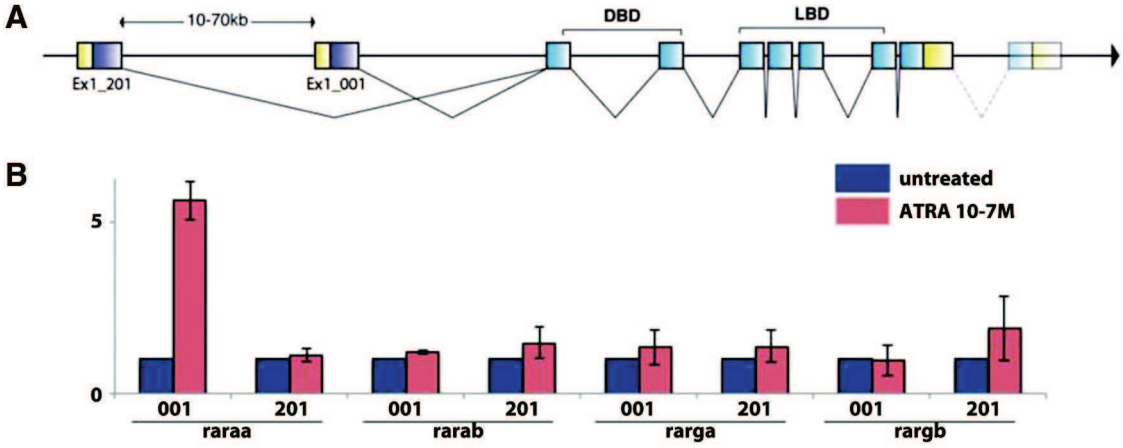
qPCR primers for each RAR isoforms were designed using the Universal Probe Library tool from Roche. Reverse transcription was performed from 1 μg total RNA using a mix of oligo$(dT)_{15}$ and random primers (Roche) and the Transcriptor Reverse Transcriptase (Roche). Quantitative PCR was performed on 3 μL of 1:10-diluted cDNA using QuantiTect SYBR green (QIAGEN) on a 96-well plate on a MX3000P Stratagene system. *Polr2d* gene was used as a referenced gene for quantification.

Results

Zebrafish RAR genes encode 2 isoforms, including *raraa*

The Ensembl database predicts 2 variants for each subtype (except for *raraa*), all following the same splicing pattern (Figure 1A), which consists of the alternative use of the first exon (31). First we checked and validated the expression of the different RARs (α-A, α-B, γ-A and γ-B) in early zebrafish embryos (8 hpf) by RT-qPCR and whole-mount in situ hybridization (Figure 1B and Supplemental Figure 1 published on The Endocrine Society's Journals Online web site at http://mend.endojournals.org). Moreover, using transcriptomic data from our laboratory, we described a second isoform for *raraa*, which follows the same genomic organization as for the other *rar* genes (Supplemental Figure 2). We cloned the full-length sequence of this newly identified isoform from total RNA extracts and submitted the sequence to GenBank (accession no. KF148618). As a result, we confirmed the existence of 2 isoforms for each zebrafish RAR subtype (named *_001* and *_201* following the Ensembl nomenclature), which differ in their first exon encoding for the N-terminal domain of the receptor (Figure 1C).

We checked their expression pattern in the early zebrafish embryo, and we found that all RAR isoforms were expressed almost ubiquitously at 8 hpf (Supplemental Figure 1). We also checked by RT-qPCR to determine whether some of the *rar* isoforms are responsive to RA in the early zebrafish gastrula (Figure 1B) and found that only *raraa_001* is activated by ATRA exposure from 4–8 hpf, which is consistent with previous observations (25). In our conditions at the early stage examined, the other isoforms are virtually insensitive to RA. The alignment of



C

gene	chr	splice variant	alternative exon	ENSEMBL ID	NCBI ID	protein size (aa)	ATG MO sequence
raraa	chr 12	raraa_001	exon 1	ENSDART00000079336	NM_131406.2	444	GGTTCACATCCACACTCTCATACAT
		raraa_201	exon 1	N/A	KF148618	447	CACCTCTACAGTTCGCCCACCGTCA
rarab	chr 3	rarab_001	exon 1	ENSDART00000044677	NM_131399.1	458	CCACAACGTCCACGCTCTCGTACAT
		rarab_201	exon 1	ENSDART00000126022	N/A	460	GTTTCCAGACATGGCACGCTCAGAC
rarga	chr 23	rarga_001	exon 1	ENSDART00000049551	NM_131339.1	499	CCAGAGCCTCCATACAGTCGAACAT
		rarga_201	exon 1	ENSDART00000103158	N/A	507	GGCCAGTTGTGTGCTTCAACAGCCT
rargb	chr 11	rargb_001	exon 1 and 8	N/A	NM_001083310.1	489	CGTCTCTCATACCAAGTGGGCTGTT
		rargb_201	exon 1 and 8	ENSDART00000052546	N/A	482	GATGATGATGGCCCTTCCCCAGCGT

Figure 1. Zebrafish RAR subtypes and isoforms. A, Splicing pattern of zebrafish RARs. Each *rar* subtype gene generates at least two different isoforms differing in their first exon encoding the N-terminal part of the receptor by alternative promoter usage. The more distal alternative exon 1 belongs to transcripts _201 and the more proximal to transcripts _001 (Ensembl nomenclature). The DBD of the receptor is encoded by exons 2 and 3, and the LBD is encoded by the fourth to seventh exon. Note that for *rargb* gene only, the deposited sequences show the last eighth exon also alternatively spliced between _001 and _201 isoforms leading to different C-terminal F domains at the protein level. B, RT-qPCR detection of each RAR isoform from whole mRNA extracts of untreated 8 hpf embryos (blue) or after treatment with ATRA at 10^{-7}M from 4–8 hpf (red). The error bars correspond to SD between at least 2 replicates from independent clutch. C, Table recapitulating for each RAR subtype, its bearing chromosome, the name and accession number of each splice variants, the corresponding size of the protein product, and the ATG-blocker morpholino sequence used for knockdown. Ex, exon; LBD, ligand-binding domain.

the 8 RAR protein sequences is provided in Supplemental Figure 3.

Designing new morpholinos and knockdown strategy

Previous studies describing the effects of RAR knockdown in zebrafish used ATG blocker morpholinos targeting only one isoform (*_001*s) of each subtype (25). In order to perform a knockdown of all RAR subtypes, we designed new ATG blocker morpholinos (that avoid mRNA translation) specific to the *_201* isoforms (Figure 1C), and splice-blocking morpholinos targeting a common exon (that disrupt correct mRNA splicing). We also investigated whether some RA-responsive genes in the early zebrafish embryo are regulated in a RAR subtype-specific fashion, by coinjecting morpholinos targeting 1) both α-A and α-B receptors; 2) both γ-A and γ-B receptors, and 3) all four subtypes, in one-cell stage embryos (Figure 2A). We treated control (ie, uninjected) and injected embryos with ATRA at 10^{-7} M from 4 hpf (sphere stage) to 8 hpf (75% epiboly stage) and fixed them for RNA extraction. The knockdown strategy was verified by RT-qPCR on canonical target genes (eg, *hox* genes) in order to define the amount of morpholino injection and to compare the splice-blocker morpholinos vs ATG-blocker morpholinos (Figure 2, B and C). We observed that splice-blocker morpholinos are less efficient than ATG morpholinos for the abolition of RA-induced target gene expression in the early embryo (Figure 2C). This is consistent with the fact that maternal RNAs that are not targeted by splice-blocker morpholinos might still be present in the early blastula, still allowing a slight activation of target gene expression (32). We also checked the expression of ventral (*eve1*) and dorsal (*shha*) markers of the zebrafish gastrula in these morphants (33). No change in the expression pattern of these markers was observed (Figure 2D), confirming that the gastrula patterning is not affected at this early stage by the knockdown of RAR. As a result, the changes in gene expression observed should not be assimilated to changes in developmental territories induced by the lack of RARs but rather in transcriptional differences. Once all the experimental parameters were set up, we widened our analysis to the whole transcriptome. We injected 2 independent spawnings with the various morpholino combinations and prepared libraries

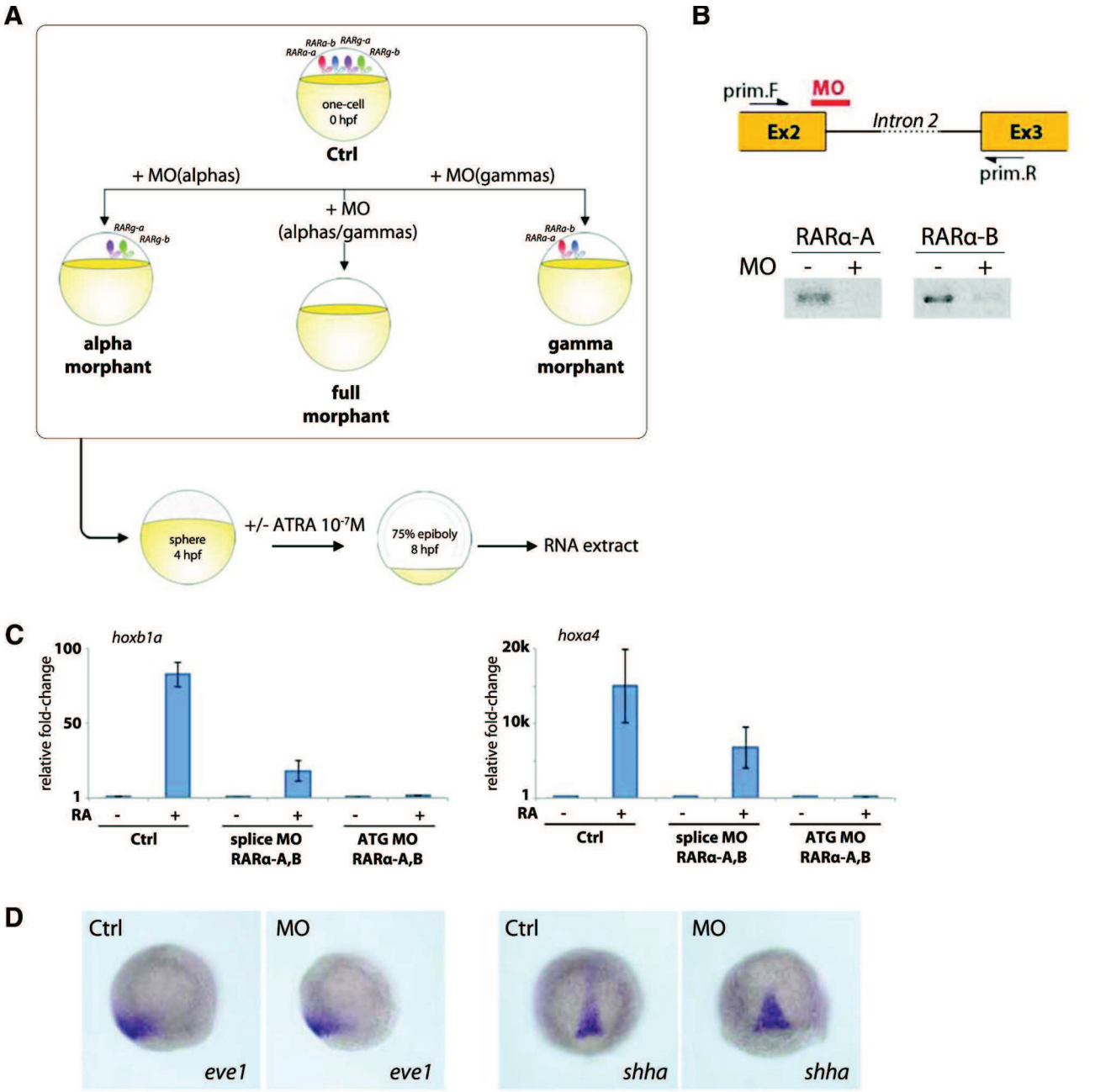


Figure 2. RAR knockdown strategy. A, Differential knockdown induced by morpholino injection. Combination of morpholinos are injected at the one-cell stage to invalidate both α-A and α-B subtypes (α-morphants), both γ-A and γ-B subtypes (γ-morphants), or all 4 RARs (full morphants). B, For RARα-A and RARα-B, splicing blocker morpholinos spanning the junction between exon 2 and intron 2 were designed, and their efficiency was checked by RT-PCR using primers in exon 2 and exon 3. The amplicon was barely detectable after morpholino injection, validating the impaired splicing of the mRNA targeted. C, Comparison of the effects of splice-blocker vs ATG-blocker morpholinos targeting RARα-A and RARα-B on the expression of *hoxb1a* and *hoxa4* by RT-qPCR. The values correspond to the relative fold-change compared with untreated controls (lane 1), and the error bars correspond to SD between at least 2 replicates from independent clutch. D, Whole-mount in situ hybridization showing the expression of *eve1* and *shha* in uninjected embryos (Ctrl) or injected with ATG-blocker morpholinos against RARα-A and RARα-B (MO). Ctrl, control; Ex, exon; MO, morpholino oligonucleotide.

from extracted RNA for transcriptome analysis using SOLiD next generation sequencer.

RA-responsive genes in the zebrafish gastrula

We compared pooled control (EtOH treated) embryos to embryos exposed to ATRA for 4 hours from 4–8 hpf, and we established a list of 311 differentially expressed genes, selected on their *P* value (adjusted *P* value $< 0{,}05$) and their fold activation (logFC > 1) (Figure 3A and Supplemental Table 1). This list contains 311 genes differentially expressed, comparing pooled control (EtOH treated) and RA-treated embryos, selected on their *P*

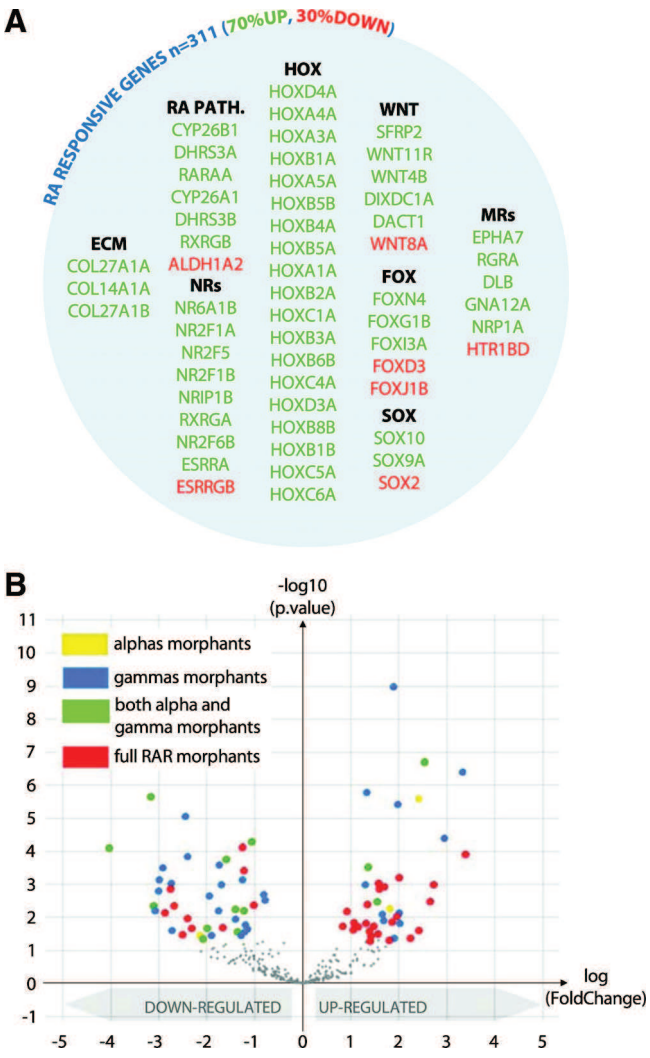


Figure 3. RA-responsive genes and their RAR-subtype dependency at the basal level. A, List of the main genes differentially expressed after a treatment of embryos from 4–8 hpf with 10^{-7}M ATRA. A total of 311 genes was identified after filtration on their *P* value $< .05$ and logFC > 1. Among them, 70% are up-regulated (in green) and 30% are down-regulated (in red) after ATRA exposure. The full list of genes with their corresponding fold-change and *P* value is provided in Supplemental Table 1. ECM, extracellular matrix; MR, membrane receptor. B, The expression of the 311 RA-responsive genes identified was checked in the different RAR morphants with regard to uninjected embryos at the basal level (ie, without ATRA exposure). Each gene colored in the volcano plot is differentially expressed (DE) in the full morphants, and they are plotted according to its fold-change (x-axis) and its *P* value (y-axis). Yellow genes are specifically DE in α-morphants (also in full morphant but not in γ-morphants) whereas blue genes are specifically DE in γ-morphants (also in full morphant but not in α-morphants). Green genes correspond to genes that are colored both in yellow and blue (DE both in α- or γ-morphants) and red genes are DE only in the full morphant (ie, lacking all RARs). The full list of genes and their fold-change (FC) and *P* value in each condition is provided in Supplemental Table 2.

value (adjusted *P* value $< 0{,}05$) and their fold activation (logFC > 1). Among these genes, 70% were up-regulated upon RA treatment (ie, 219 genes), and 30% (ie, 92) had their overall expression down-regulated. As expected, this list contained canonical RA target genes that validate our approach. Among these genes, we found genes implicated in the RA metabolism pathway (eg, *cyp26a/b1*, *dhrs3a/b*, *aldh1a2),* genes from the WNT pathway, Hox genes, and other homeobox transcription factors (eg, *sox*, *fox, meis*) (34–36). We also found some new RA targets belonging to nuclear receptors (estrogen-related receptors: *esrra*, *esrrgb*), membrane receptors (ephrin receptor 7, retinal G protein, serotonin receptor), and genes of the extracellular matrix (collagen type XIV, type XXVII) (Figure 3A and Supplemental Table 1). This list of RA-responsive genes is of prime interest for a wide scientific community because no study has ever assayed RA-transcriptional response in vivo, in the early zebrafish embryo. In the present work, this list is used as a reference for further comparison with RAR subtype-specific morphants.

RAR subtype-specific basal regulation of RA-target genes

We first investigated whether among the above 311 RA-responsive genes, some are also under the control of RARs in the early embryo, in the absence of exogenous ligand. In other words, we analyzed the consequences of our RAR subtype-specific knockdown on the basal expression of these genes (ie, without RA exposure) (Figure 3B and Supplemental Table 2). We therefore tested whether these genes, regulated after exogenous RA exposure, could indeed be under the control of endogenous RAR in the early embryo. Only 27% of these RA-responsive genes (n = 85/311) exhibit a significant differential expression in the full-RAR morphants, suggesting that many of them are not regulated by RARs at the basal level. Within these genes, some are only affected in α-morphants (yellow genes in Figure 3A), others are affected only in γ-morphants (blue genes in Figure 3B), some are affected in both situations (green genes in Figure 3, A and B), and lastly some genes are only affected in the full-RAR morphants (red genes in Figure 3, A and B). Such results indicate that the basal expression of RA-target genes can be regulated specifically by one RAR subtype (either α- or γ-subtypes) or can rely on both for their normal basal expression (eg, requires all 4 RARs). Of note, only 20 genes are differentially expressed in the α-morphants (count yellow and green genes in Figure 3B) whereas 58 genes are affected in the γ-morphants (count blue and green genes in Figure 3B), suggesting that there are more genes the basal regulation of which is dependent on the γ-subtypes. Interestingly, 85% of the genes that are differentially expressed in the α- or γ-morphants are also found in the full-RAR morphants, therefore emphasizing the validity of the effects observed in the various situa-

tions. Besides, nearly half of the affected genes (37 of 85) are only affected in the full-RAR morphant (red genes in Figure 3B), meaning that, in some cases, the different RAR subtypes can be redundant. Interestingly, at the basal level, we observed as much down-regulated as up-regulated genes in our RAR morphants. In the absence of ligand, RARs are thought to actively repress the transcription of their target (1, 37). The fact that we observe many up-regulated genes after RAR knockdown suggests that RARs do have a basal repressive activity on some genes in the early zebrafish embryo in contrast to what has been recently suggested (37).

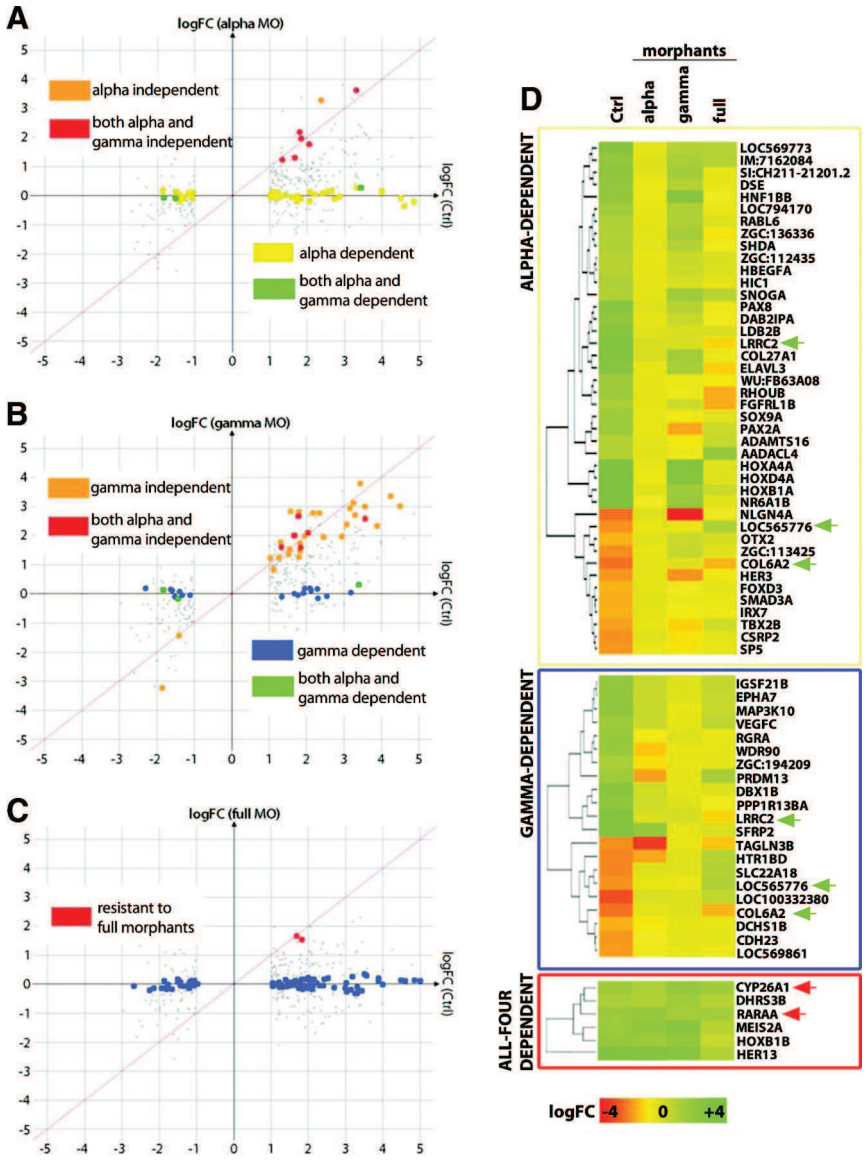


Figure 4. Comparison of the effect of RA on RA-responsive genes in control vs morphant embryos. Comparison of differentially expressed (DE) genes after ATRA exposure in control vs α-morphants (panel A), γ-morphants (panel B), or full morphants (panel C). Each RA-responsive gene is plotted against its fold-change (FC) after ATRA exposure in control embryos (x-axis) and its FC in morphants (y-axis). If a gene has a same significant FC in control and morphant embryos, it is colored and follows the red diagonal. If the FC of a gene is abolished in morphants, it drops near the x-axis and is colored if its FC in morphants is less than 10% of the FC in controls. In each situation the genes are differentially colored depending on their behavior (subtype dependent or independent) in the different morphants. Green genes correspond to genes that are colored both in yellow and blue (merge). The full list of genes and their FC and *P* value in each condition is provided in Supplemental Table 3. D, α-Dependent (yellow box), γ-dependent (blue box), and all four dependent (red box) genes are depicted on heat maps. The color code is function of the logFC. Note that the *P* values are not taken into account in the heat map color coding. All the *P* values are available in Supplemental Table 3. Green arrows indicate the genes found both α- and γ-dependent. Red arrows indicate the genes that are still DE in the full morphants. MO, morpholino oligonucleotide.

Differential regulation of RA-responsive genes by α- or γ-RAR subtypes in vivo

Next we analyzed the consequences of our RAR-subtype specific knockdown on the expression of our list of 311 genes after exogenous RA exposure. After comparing the list of RA-responsive genes in wild-type vs RAR-morphant embryos, we considered genes that are still significantly induced in the morphants as RA-target genes the expression of which is independent of the knocked down receptors (orange and red spots in Figure 4, A and B). Conversely, genes the expression of which in RAR morphants was drastically decreased (<10% of the fold activation observed in controls) were considered as completely abolished and therefore as fully dependent of the targeted receptors (yellow and blue spots in Figure 4, A and B). In the α-morphants, 42 genes were completely abolished although they are significantly responding to RA in controls, (yellow spots, Figure 4A; and Supplemental Table 3) meaning that these genes are strictly requiring RARα subtypes (either RARα-A and/or RARα-B) for their RA-dependent regulation. Similarly, we counted 21 genes that strictly require the RARγ subtypes (either RARγ-A and/or RARγ-B) for being regulated by RA (blue spots in Figure 4B and Supplemental Table 3). It is worth noting, that 3 of these genes were completely abolished both in α and γ-morphants (green spots in Figure 4, A and B), meaning that they strictly require all 4 RAR sub-

types to be regulated by RA. In these subtype-specific morphants, we also observed some genes that are not affected by the morpholino-induced RAR knockdown and that continue to be regulated in the same way as in controls (orange spots in Figure 4, A and B). Among these nonaffected genes, 6 continue to be normally regulated by RA both in α- and γ-morphants (red spots in Figure 4, A and B), suggesting that for these genes, the activity of the different RAR subtypes is redundant. In the full-RAR morphants situation (Figure 4C), 4 of these 6 genes (ie, *dhrs3b, meis2a, hoxb1b, her13*) are no longer responding to RA. This observation indicates that these genes can be regulated either by α- or γ-subtypes indifferently and are therefore regulated by RAR in a truly redundant manner. Altogether, these results indicate that in the early zebrafish, some RA-responsive genes are specifically regulated by α- or γ-subtypes independently (yellow and blue genes, respectively) whereas others (green genes) require all 4 subtypes to be RA responsive. Of note, almost twice as many genes appear sensitive to α-morpholino knockdown (yellow and green genes in Figure 4, A and B) compared with γ-morphants (blue and green genes in Figure 4, A and B). This suggests that the α-RAR subtypes are the main subtypes involved in the regulation of RA-induced gene expression in the early zebrafish gastrula whereas, as observed above (Figure 3B), γ-RAR subtypes appear to control the basal RA regulation of more genes.

The extreme cases of strict subtype specificity that we just described concern only a small fraction of all RA-responsive genes identified, and most the genes (in gray in Figure 4, A–C) exhibit an intermediate specificity. Indeed, there is a continuum of transcriptional behaviors in response to RA in the different morphants, which we confirmed for some genes in an independent RT-qPCR assay (Supplemental Figure 4), suggesting that for many genes, proper RA regulation requires the availability of all RARs but does not rely on a full dependency on one specific RAR subtype but rather on a partial functional redundancy among subtypes.

cyp26a1 and *raraa* are highly robust RA-responsive genes

During this analysis, we observed only 2 genes that are still significantly RA regulated in the full-RAR morphants as they are in controls (red spots, Figure 4C). Interestingly, these 2 genes, *cyp26a1* and *raraa*, are members of the RA-signaling pathway and are known to be canonical RA target genes in zebrafish (38, 39). As several other NRs such as chicken ovalbumin upstream promoter-transcription factor II, thyroid hormone receptor 4, retinoid orphan receptor-β, and PPARβ have been recently proposed as being activated by RA and we therefore wonder if the apparent RAR-independant RA-regulation of *cyp26a1* and *raraa* could be mediated by one of these receptors (40–43). Members of the chicken ovalbumin upstream promoter-transcription factor II group (NR2F1A, NR2F1B, NR2F2 and NR2F5) as well as the 2 PPARβ (PPARβ-A and PPARβ-B) are expressed in the zebrafish gastrula (20), we therefore knocked down these receptors and checked whether *cyp26a1* and *raraa* induction by RA was altered. We found that the level of RA regulation was virtually unaltered after single or combined knockdown of these receptors (Figure 5A and data not shown), suggesting that these resistant genes are not under the control of these NRs. Therefore, to check whether these genes are regulated by RA through a RAR-dependent mechanism, we monitored their expression by RT-qPCR after treatment with a pan-RAR inverse agonist BMS493 (Figure 5B). Interestingly, the antagonist treatment completely abolishes the RA activation of these genes, thus confirming that their RA regulation is mediated through RARs. Thus, their surprising resistance to full-RAR morphants can be explained by: 1) the fact that other undescribed isoforms of RARs could exist and are not targeted by our morpholinos or 2) the fact that it is known that morpholino injection only leads to incomplete knockdown (44). It is important to note in that respect that, these 2 genes represent only 0.6% of the 311 responsive genes, validating that the knockdown induced by our morpholino is highly efficient. The fact that *cyp26a1* and *raraa* continue to be activated in these morphants as in controls suggest that their RA-induced activation is kept under a high robustness pressure in the early embryo. In addition, this would also suggest that different levels of robustness of RA regulation exist among RA-target genes and that some genes, such as *cyp26a1* and *raraa*, could be considered as the most robust genes that can be regulated by RA even in the case of a drastic reduction of RARs availability.

The transcriptional activity of RARα-A and RARα-B is not redundant

In order to decipher in more details the RAR subtype-specific transcriptional activity, we decided to knock down the RARα subtypes (ie, RARα-A and RARα-B) separately. Therefore, we specifically knocked down RARα-A or RARα-B by morpholino injection (Figure 5C), and we checked by RT-qPCR the expression of genes that appeared to be sensitive to RARαs knockdown in our whole-transcriptome assay (Supplemental Table 3). For the 4 genes tested, each single knockdown of RARα-A or

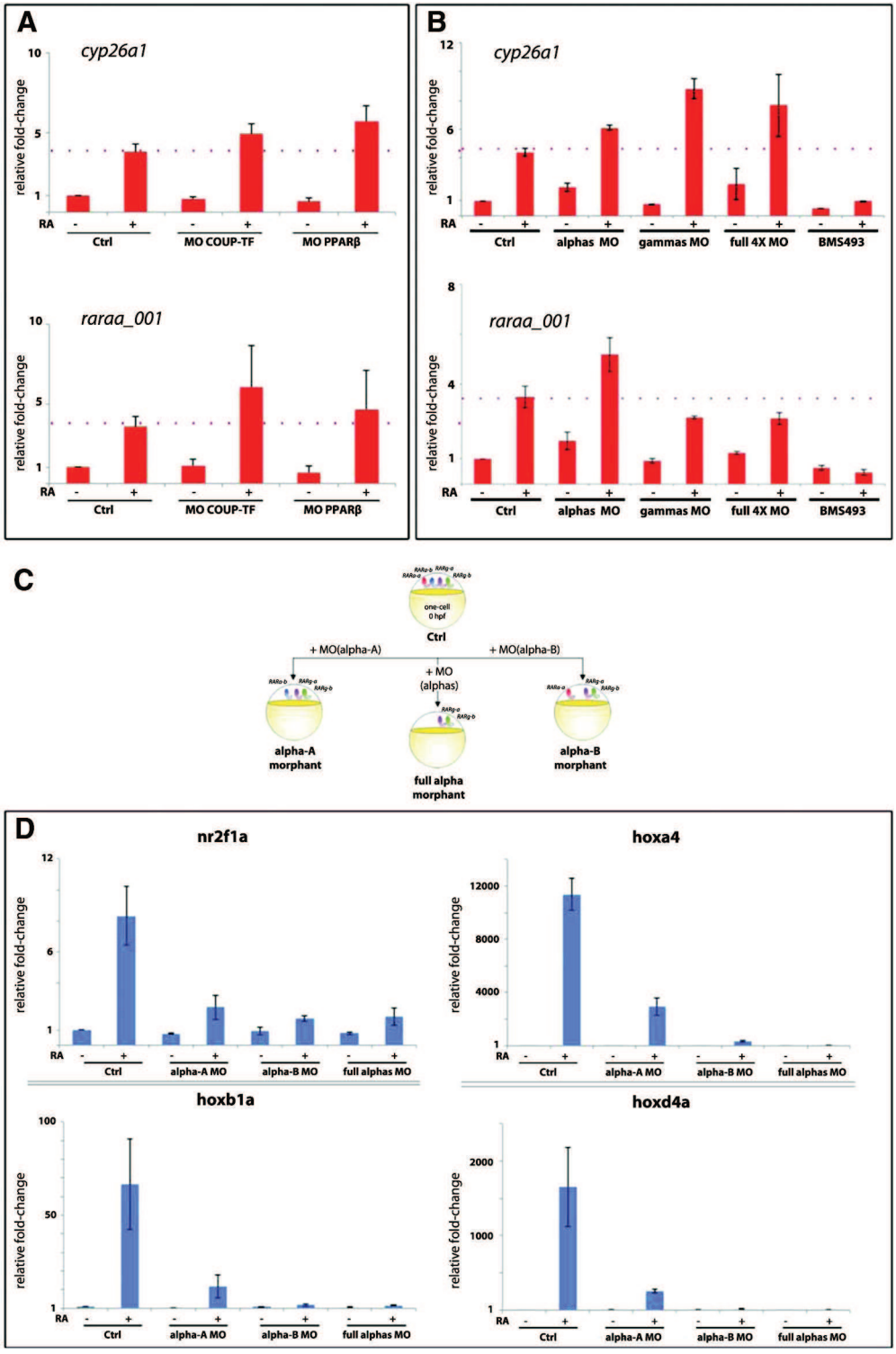


Figure 5. RAR antagonist treatment and knockdown of other NRs and of single RARα-A and RARα-B. A, The expression of *cyp26a1* and *raraa_001* checked by RT-qPCR is still activated by RA after knockdown of COUP-TFII receptors (NR2F1A, NR2F1B, NR2F2, and NR2F5) or PPARβ-A/β-B. B, Validation of the activation of *cyp26a1* and *raraa_001* by RA in the different RAR morphants. The activation is abolished by pan-RAR inverse agonist (BMS493) treatment 1 hour prior to ATRA exposure. C and D, The expression of α-dependent genes (ie, *nr2f1a*, *hoxa4*, *hoxb1a*, *hoxd4a*) after ATRA exposure (10^{-7} M, 4–8 hpf) was checked by RT-qPCR in uninjected embryos (Ctrl) or after single knockdown of RARα-A or RARα-B compared with full α-morphant (ie, both RARα-A and RARα-B) induced by ATG-blocker morpholino injection. The values correspond to the relative fold-change compared with untreated controls (lane 1), and the error bars correspond to SD between at least 2 replicates from independent clutch. Ctrl, control; MO, morpholino oligonucleotide.

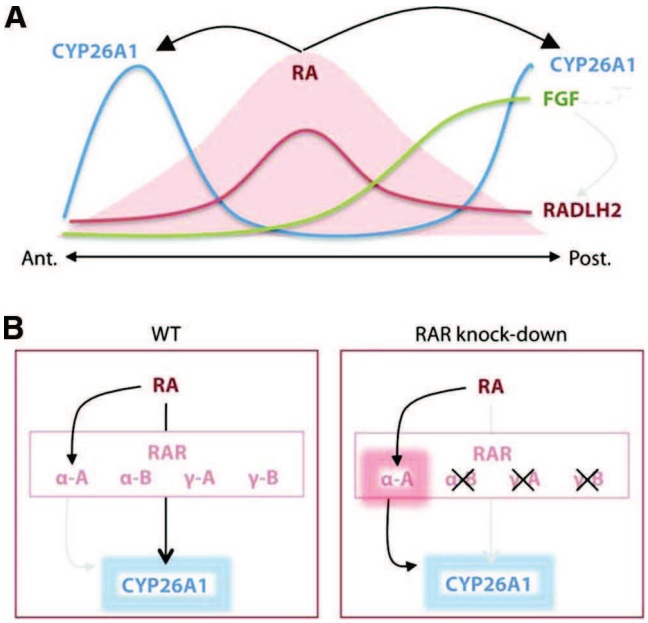


Figure 6. Robust *cyp26a1* and *raraa* RA regulation could participate in maintaining RA signaling in vivo. A, The proposed two-tailed gradient along the antero-posterior axis of the embryo is controlled by local degradation of RA by CYP26A1 the expression of which is activated by RA and inhibited by FGF pathway (52). RALDH2 synthetizes RA and its expression is activated by FGF. B, In WT embryos, *cyp26a1* is activated by RA by any of the 4 RAR subtypes. However, RAR-αA is the only subtype to be strongly activated by RA. In the case of an impaired availability of RARs (eg, knockdown), the robust activation of *raraa* by RA ensures the subsequent activation of *cyp26a1*, necessary for the control of RA level in vivo. Ant, anterior; FGF, fibroblast growth factor; Post, posterior; WT, wild type.

RARα-B strongly reduces their RA-induced expression (Figure 5D). This suggests that both RARα-A and RARα-B are required for proper RA regulation of these genes. Although the expression level in the single α-morphants is not as reduced as in the full α-morphants, these results suggest that, in contrast to what is commonly thought, both RARα subtypes have independent transcriptional activities and are not fully redundant at the transcriptional level, at least at this stage. Of note is that the relative abundance of *raraa* and *rarab* transcripts in the early zebrafish embryo is the same (36); therefore, the changes observed are not due to differences of expression level between both subtypes.

Discussion

Many studies successfully deciphered the specific function of each RAR subtype in mice and zebrafish using standard forward genetics approaches examining developmental phenotypes that can be associated with the lack of one specific RAR subtype (19, 25). However, it is still not understood why one developmental process could require only one specific RAR subtype although others are expressed in the same region of interest. This is expected to rely on differences in the recruitment and/or transcriptional activity of the different RAR subtypes. In this work, we provide evidences that the different RAR subtypes, even coexpressed in vivo, can drive RA transcriptional response in a subtype-specific fashion. Indeed, we were able to discriminate genes that 1) are specifically regulated by one or another RAR subtype; 2) strictly require all subtypes for their correct RA regulation; and 3) can be regulated either by one or another subtype indifferently. Drawing the parallel with RAR subtype-specific phenotypes that have been characterized in zebrafish (25, 45), our results revealed what we call RAR subtype-specific "transcriptotypes" that correspond to a repertoire of genes activated in a subtype-specific fashion in response to RA. Moreover, we observed that the γ-subtypes of RARs appear to be more involved in the basal regulation of RA-responsive gene expression, whereas α-subtypes are most implicated in the transcriptional RA response in the early zebrafish embryo. These differences are unlikely to be due to different expression levels because in a 12-hpf embryo, the relative mRNA level of α (ie, α-A and α-B) and γ (ie, γ-A ang γ-B) subtypes are equivalent (36).

More than simply providing exhaustive lists of genes regulated by distinct RARs in vivo, we also examined their expression pattern in the ZFIN database at 8 hpf in the gastrula and later in development (Supplemental Figure 5). We did not observe any obvious correlation between the subtype dependency of a gene and its specific expression pattern in the gastrula. In fact, many genes harbor a ubiquitous expression pattern at 8 hpf whereas others present more specific, but variable, expression patterns (Supplemental Figure 5). As a result, the subtype-specific transcriptotypes we described are more likely to rely on differences of RAR transcriptional activity at the chromatin level instead of differences of their activity in specific developmental territories. We also checked the expression of these genes later in development to determine whether their subtype dependency at 8 hpf correlates with their expression pattern later in development. Interestingly, the α-dependent genes at 8 hpf are mainly expressed in the central nervous system later in development, whereas our γ-dependent genes are predominantly expressed in the eye (retina or lens), heart, and somites of the zebrafish larvae (Supplemental Figure 5). Although all 4 RARs are expressed in the central nervous system, only RARα-A and RARα-B are expressed in the eye (20). Altogether, these observations suggest that the specific subtype dependency of a gene is not necessarily the same at 8 hpf and later in development. We therefore propose that

this dependency does not rely on fixed features but rather on dynamic mechanisms.

How can such specificity be driven at the transcriptional level? Because the DBD of the different RAR subtypes are highly conserved (92%), especially the P box (100% identity), which determines the specific interaction between RAR and DNA (46), we do not expect differences at the receptor level that could explain why one RAR subtype is specifically involved in the regulation of specific gene as we observed here. However, one can speculate that the specificity relies on specific features of the RAREs that would promote the recruitment of one specific RAR subtype. We exhaustively looked for canonical and nonclassical RAREs within zebrafish genome using the PARSEC (PAtteRn Search and Contextualization) platform, which has been recently developed to identify a specific sequence pattern within the whole genome (47). We look for direct-repeats (DRs) of the consensus RARE nucleotide sequence RGKTSA (48) separated by 1–10 nucleotides (DR1–DR10) near the coding region of our RA-responsive genes (+/− 10 kb). Unfortunately, neither specific consensus sequence nor special DR space enrichment was found to explain the specific recruitment of one specific RAR subtype. However, recent studies have revealed that the repertoire of RAR-binding sites in vivo is highly diverse and that the classical RARE consensus does not recapitulate all true binding sites (49). As a result, an in vivo analysis of RAR subtype-specific binding sites by sequential chromatin immunoprecipation sequencing using antibodies recognizing specifically each subtype would be necessary to have a better view in what could drive this RAR-subtype specificity at the DNA level, such as the presence of binding sites for other transcription factors near RAREs. Lastly, the unstructured N-terminal domain (NTD) of RARs is a good candidate for being involved in this specificity. In fact, thanks to its proximity with the DBD, proteins interacting with the NTD could regulate RAR recruitment onto DNA as recently hypothesized and/or the recruitment of specific coactivator complexes (50). Because NTDs are the most divergent parts of RARs between subtypes (Supplemental Figure 3B), one can speculate that differences in the NTDs can participate in driving RAR subtype-specific recruitment onto chromatin through specific protein interactions. This would require further work aiming at identifying RAR subtype-specific NTD-interacting proteins in vivo.

Finally, our work also provides evidences that RA-responsive genes are not all sensitive to the same level of alterations in the RA pathway. In fact, we found genes of the RA pathway itself the expression of which is not affected at all, even after a potent knockdown of all RARs. This suggests that the regulation of these genes by RA is extremely robust, suggesting that keeping their RA regulation is subjected to a strong pressure that can participate to maintain the integrity of the pathway in vivo. This is consistent with many studies that highlighted the robustness of the RA-signaling pathway in the developing embryo, thanks to complex cross talks and many feedback regulations controlling *cyp26a1* expression (51, 52) (Figure 6A). Furthermore, our results show that *raraa* expression is also robustly induced by RA, and this could be considered as a security path that can compensate any alteration of RAR availability to finally maintain the RA regulation of *cyp26a1* (Figure 6). The high robustness of *cyp26a1* expression in response to RA might be associated with the fact that 3 canonical RAREs have been described within its promoter (38, 39). Furthermore, using the PARSEC platform, we found 4 more canonical RAREs in the promoter region of *cyp26a1* and two DR5s near the *raraa* gene (data not show). We can wonder whether the presence of multiple canonical RAREs in the *cis*-regulatory regions of these genes could be correlated to their robustness.

Our work therefore provides new insights into RAR activity at the transcriptional level in an in vivo context examining endogenous proteins and opening new prospects for further research to decipher the mechanisms regulating their activity in vivo.

Acknowledgments

We thank Laure Bernard for technical help and comprehensive support. We thank Yann Gibert for fruitful discussion. We benefited from the zebrafish in-breeding facilities developed by the Pateau de Recherche Experimentale de Criblage In vivo technical platform at the Structure Fédérative de Recherche Biosciences Gerland and from the sequencing platform of the IGFL.

Address all correspondence and requests for reprints to: Vincent Laudet, IGFL-ENS de Lyon, 32–34 Avenue Tony Garnier, 69007 Lyon, France. E-mail: vincent.laudet@ens-lyon.fr.

This work was supported by La ligue Contre le Cancer and the French Ministry of Research, the Centre National de la Recherche Scientifique, and the Ecole Normale Supérieure de Lyon. Of note, it was not funded by Agence Nationale pour la Recherche, who rejected the project. We also thank E.G and L.S for support.

Disclosure Summary: The authors have nothing to disclose.

References

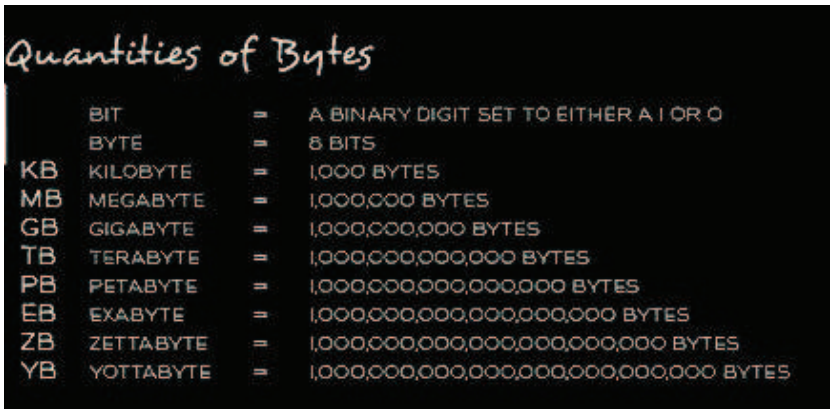
1. **Samarut E, Rochette-Egly C.** Nuclear retinoic acid receptors: conductors of the retinoic acid symphony during development. *Mol Cell Endocrinol*. 2012;348(2):348–360.

2. **Clagett-Dame M, DeLuca HF.** The role of vitamin A in mammalian reproduction and embryonic development. *Annu Rev Nutr*. 2002; 22:347–381.
3. **Zile MH.** Function of vitamin A in vertebrate embryonic development. *J Nutr*. 2001;131(3):705–708.
4. **Begemann G, Marx M, Mebus K, Meyer A, Bastmeyer M.** Beyond the neckless phenotype: influence of reduced retinoic acid signaling on motor neuron development in the zebrafish hindbrain. *Dev Biol*. 2004;271(1):119–129.
5. **Maves L, Kimmel CB.** Dynamic and sequential patterning of the zebrafish posterior hindbrain by retinoic acid. *Dev Biol*. 2005; 285(2):593–605.
6. **Aulehla A, Pourquie O.** Signaling gradients during paraxial mesoderm development. *Cold Spring Harb Perspect Biol*. 2010;2(2): a000869.
7. **Chambon P.** A decade of molecular biology ot retinoic acid receptors. *FASEBJ*. 1996;10(9):940–954.
8. **Germain P, Altucci L, Bourguet W, Rochette-Egly C, Gronemeyer H.** Nuclear receptor superfamily: principles of signaling. In: Miyamoto J, Burger J, eds. *Pure Appl Chem*. Vol 75. 2003:11–12, 1619–1664.
9. **Germain P, Staels B, Dacquet C, Spedding M, Laudet V.** Overview of nomenclature of nuclear receptors. *Pharmacol Rev*. 2006;58(4): 685–704.
10. **Gronemeyer H, Gustafsson JA, Laudet V.** Principles for modulation of the nuclear receptor superfamily. *Nat Rev Drug Discov*. 2004; 3(11):950–964.
11. **Rochette-Egly C, Germain P.** Dynamic and combinatorial control of gene expression by nuclear retinoic acid receptors (RARs). *Nucl Recept Signal*. 2009;7:e005.
12. **Chakravarti D.** *Regulatory Mechanisms in Transcriptional Signaling*. 1st ed. Vol 87. New York: Elsevier; 2009.
13. **Moras D, Gronemeyer H.** The nuclear receptor ligand-binding domain: structure and function. *Curr Opin Cell Biol*. 1998;10(3): 384–391.
14. **Dilworth FJ, Chambon P.** Nuclear receptors coordinate the activities of chromatin remodeling complexes and coactivators to facilitate initiation of transcription. *Oncogene*. 2001;20(24):3047–3054.
15. **Germain P, Chambon P, Eichele G, et al.** International Union of Pharmacology. LX. Retinoic acid receptors. *Pharmacol Rev*. 2006; 58(4):712–725.
16. **Dolle P.** Developmental expression of retinoic acid receptors (RARs). *Nucl Recept Signal*. 2009;7:e006.
17. **Niederreither K, Dollé P.** Retinoic acid in development: towards an integrated view. *Nat Rev Genet*. 2008;9(7):541–553.
18. **Mark M, Ghyselinck NB, Chambon P.** Function of retinoid nuclear receptors: lessons from genetic and pharmacological dissections of the retinoic acid signaling pathway during mouse embryogenesis. *Annu Rev Pharmacol Toxicol*. 2006;46:451–480.
19. **Mark M, Ghyselinck NB, Chambon P.** Function of retinoic acid receptors during embryonic development. *Nucl Recept Signal*. 2009;7:e002.
20. **Bertrand S, Thisse B, Tavares R, et al.** Unexpected novel relational links uncovered by extensive developmental profiling of nuclear receptor expression. *PLoS Genet*. 2007;3(11):e188.
21. **Hale LA, Tallafuss A, Yan YL, Dudley L, Eisen JS, Postlethwait JH.** Characterization of the retinoic acid receptor genes raraa, rarab and rarg during zebrafish development. *Gene Expr Patterns*. 2006;6(5): 546–555.
22. **Waxman JS, Yelon D.** Comparison of the expression patterns of newly identified zebrafish retinoic acid and retinoid X receptors. *Dev Dyn*. 2007;236(2):587–595.
23. **Zelent A, Mendelsohn C, Kastner P, et al.** Differentially expressed isoforms of the mouse retinoic acid receptor β generated by usage of two promoters and alternative splicing. *EMBO J*. 1991;10(1):71–81.
24. **Mollard R, Viville S, Ward SJ, Decimo D, Chambon P, Dolle P.** Tissue-specific expression of retinoic acid receptor isoform transcripts in the mouse embryo. *Mech Dev*. 2000;94(1–2):223–232.
25. **Linville A, Radtke K, Waxman JS, Yelon D, Schilling TF.** Combinatorial roles for zebrafish retinoic acid receptors in the hindbrain, limbs and pharyngeal arches. *Dev Biol*. 2009;325(1):60–70.
26. **Kimmel CB, Ballard WW, Kimmel SR, Ullmann B, Schilling TF.** Stages of embryonic development of the zebrafish. *Dev Dyn*. 1995; 203(3):253–310.
27. **Robinson MD, Oshlack A.** A scaling normalization method for differential expression analysis of RNA-seq data. *Genome Biol*. 2010;11(3):R25.
28. **Smyth GK.** Linear models and empirical bayes methods for assessing differential expression in microarray experiments. *Stat Appl Genet Mol Biol*. 2004;3:Article3.
29. **Benjamini Y, Hochberg Y.** Controlling the false discovery rate: a practical and powerful approach to multiple testing. *J Royal Statistical Soc Series B (Methodological)*. 1995;57(1):289–300.
30. **Thisse C, Thisse B.** High-resolution in situ hybridization to whole-mount zebrafish embryos. *Nat Protocols*. 2008;3(1):59–69.
31. **Birney E, Andrews TD, Bevan P, et al.** An overview of Ensembl. *Genome Res*. 2004;14(5):925–928.
32. **Tadros W, Lipshitz HD.** The maternal-to-zygotic transition: a play in two acts. *Development*. 2009;136(18):3033–3042.
33. **Hammerschmidt M, Pelegri F, Mullins MC, et al.** dino and mercedes, two genes regulating dorsal development in the zebrafish embryo. *Development*. 1996;123:95–102.
34. **Balmer JE, Blomhoff R.** Gene expression regulation by retinoic acid. *J Lipid Res*. 2002;43(11):1773–1808.
35. **Feng L, Hernandez RE, Waxman JS, Yelon D, Moens C.** Dhrs3a regulates retinoic acid biosynthesis through a feedback inhibition mechanism. *Dev Biol*. 2010;338(1):1–14.
36. **Oliveira E, Casado M, Raldua D, Soares A, Barata C, Pina B.** Retinoic acid receptors' expression and function during zebrafish early development. *J Steroid Biochem Mol Biol*. 2013;138:143–151.
37. **Waxman JS, Yelon D.** Zebrafish retinoic acid receptors function as context-dependent transcriptional activators. *Dev Biol*. 2011; 352(1):128–140.
38. **Hu P, Tian M, Bao J, et al.** Retinoid regulation of the zebrafish cyp26a1 promoter. *Dev Dyn*. 2008;237(12):3798–3808.
39. **Li J, Hu P, Li K, Zhao Q.** Identification and characterization of a novel retinoic acid response element in zebrafish cyp26a1 promoter. *Anat Rec (Hoboken)*. 2012;295(2):268–277.
40. **Zhou XE, Suino-Powell KM, Xu Y, et al.** The Orphan Nuclear Receptor TR4 Is a Vitamin A-activated Nuclear Receptor. *J Biol Chem*. 2011;286(4):2877–2885.
41. **Stehlin-Gaon C, Willmann D, Zeyer D, et al.** All-trans retinoic acid is a ligand for the orphan nuclear receptor ROR β. *Nat Struct Biol*. 2003;10(10):820–825.
42. **Kruse SW, Suino-Powell K, Zhou XE, et al.** Identification of COUP-TFII orphan nuclear receptor as a retinoic acid-activated receptor. *PLoS Biol*. 2008;6(9):e227.
43. **Shaw N, Elholm M, Noy N.** Retinoic acid is a high affinity selective ligand for the peroxisome proliferator-activated receptor β/δ. *J Biol Chem*. 2003;278(43):41589–41592.
44. **Bill BR, Petzold AM, Clark KJ, Schimmenti LA, Ekker SC.** A primer for morpholino use in zebrafish. *Zebrafish*. 2009;6(1):68–77.
45. **D'Aniello E, Rydeen AB, Anderson JL, Mandal A, Waxman JS.** Depletion of retinoic acid receptors initiates a novel positive feedback mechanism that promotes teratogenic increases in retinoic acid. *Plos Genetics*. 2013;9(8).
46. **Lee MS, Kliewer SA, Provencal J, Wright PE, Evans RM.** Structure of the retinoid X receptor α DNA binding domain: a helix required for homodimeric DNA binding. *Science*. 1993;260(5111):1117–1121.
47. **Allot A, Anno YN, Poidevin L, Ripp R, Poch O, Lecompte O.**

PARSEC: PAtteRn SEarch and Contextualization. *Bioinformatics*. 2013;29(20):2643–2644.

48. **Lalevée S, Anno YN, Chatagnon A, et al.** Genome-wide in silico identification of new conserved and functional retinoic acid receptor response elements (direct repeats separated by 5 bp). *J Biol Chem*. 2011;286(38):33322–33334.
49. **Delacroix L, Moutier E, Altobelli G, et al.** Cell-specific interaction of retinoic acid receptors with target genes in mouse embryonic fibroblasts and embryonic stem cells. *Mol Cell Biol*. 2010;30(1):231–244.
50. **Lalevée S, Bour G, Quinternet M, et al.** Vinexinβ, an atypical "sensor" of retinoic acid receptor γ signaling: union and sequestration, separation, and phosphorylation. *FASEB J*. 2010;24(11):4523–4534.
51. **White RJ, Nie Q, Lander AD, Schilling TF.** Complex regulation of cyp26a1 creates a robust retinoic acid gradient in the zebrafish embryo. *PLoS Biol*. 2007;5(11):e304.
52. **Shimozono S, Iimura T, Kitaguchi T, Higashijima S, Miyawaki A.** Visualization of an endogenous retinoic acid gradient across embryonic development. *Nature*. 2013;496(7445):363–366.

3 Annexe 3 : Taille des données



(https://www.flickr.com/photos/lafabriquedeblogs/5595668903/)

Références bibliographiques

Aagaard, M.M., Siersbaek, R., and Mandrup, S. (2011). Molecular basis for gene-specific transactivation by nuclear receptors. Biochim Biophys Acta *1812*, 824-835.

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.

Allison, M. (2008). Is personalized medicine finally arriving? Nat Biotechnol.

Allot, A., Anno, Y.N., Poidevin, L., Ripp, R., Poch, O., and Lecompte, O. (2013). PARSEC: PAtteRn SEarch and Contextualization. Bioinformatics *29*, 2643-2644.

Amar, R., Eagan, J., and Stasko, J. (2005). Low-level components of analytic activity in information visualization. INFOVIS 05: IEEE Symposium on Information Visualization, Proceedings, 111-117.

Amberger, J.S., Bocchini, C.A., Schiettecatte, F., Scott, A.F., and Hamosh, A. (2015). OMIM.org: Online Mendelian Inheritance in Man (OMIM(R)), an online catalog of human genes and genetic disorders. Nucleic Acids Res *43*, D789-798.

Aral, S., and Walker, D. (2012). Identifying influential and susceptible members of social networks. Science *337*, 337-341.

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.

Augenlicht, L.H., Wahrman, M.Z., Halsey, H., Anderson, L., Taylor, J., and Lipkin, M. (1987). Expression of cloned sequences in biopsies of human colonic tissue and in colonic carcinoma cells induced to differentiate in vitro. Cancer Res *47*, 6017-6021.

Ayme, S. (2003). [Orphanet, an information site on rare diseases]. Soins, 46-47.

Bailey, T., Krajewski, P., Ladunga, I., Lefebvre, C., Li, Q., Liu, T., Madrigal, P., Taslim, C., and Zhang, J. (2013). Practical guidelines for the comprehensive analysis of ChIP-seq data. PLoS Comput Biol *9*, e1003326.

Bailey, T.L., Boden, M., Buske, F.A., Frith, M., Grant, C.E., Clementi, L., Ren, J., Li, W.W., and Noble, W.S. (2009). MEME SUITE: tools for motif discovery and searching. Nucleic Acids Res *37*, W202-208.

Bairoch, A., and Apweiler, R. (2000). The SWISS-PROT protein sequence database and its supplement TrEMBL in 2000. Nucleic Acids Res *28*, 45-48.

Bakshy, E., Rosenn, I., Marlow, C., and Adamic, L. (2012). The role of social networks in information diffusion. Proceedings of the 21st International Conference on World Wide Web.

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.

Barabasi, A.L. (2007). Network medicine--from obesity to the "diseasome". N Engl J Med *357*, 404-407.

Bardet, G. (1995). On congenital obesity syndrome with polydactyly and retinitis pigmentosa (a contribution to the study of clinical forms of hypophyseal obesity). 1920. Obes Res *3*, 387-399.

Barr, M.M., and Sternberg, P.W. (1999). A polycystic kidney-disease gene homologue required for male mating behaviour in C. elegans. Nature *401*, 386-389.

Barrett, T., Wilhite, S.E., Ledoux, P., Evangelista, C., Kim, I.F., Tomashevsky, M., Marshall, K.A., Phillippy, K.H., Sherman, P.M., Holko, M.*, et al.* (2013). NCBI GEO: archive for functional genomics data sets--update. Nucleic Acids Res *41*, D991-995.

Bestetti, R.B., Restini, C.B., and Couto, L.B. (2014). Development of anatomophysiologic knowledge regarding the cardiovascular system: from Egyptians to Harvey. Arq Bras Cardiol *103*, 538-545.
Biedl, A. (1995). A pair of siblings with adiposo-genital dystrophy. 1922. Obes Res *3*, 404.
Bird, S. (2006). NLTK: the natural language toolkit (Stroudsburg, PA, USA: Association for Computational Linguistics).
Blake, J. (2004). Bio-ontologies-fast and furious. Nat Biotechnol *22*, 773-774.
Blanchette, M., Kent, W.J., Riemer, C., Elnitski, L., Smit, A.F., Roskin, K.M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E.D*., et al.* (2004). Aligning multiple genomic sequences with the threaded blockset aligner. Genome Res *14*, 708-715.
Blume, C.J., Hotz-Wagenblatt, A., Hullein, J., Sellner, L., Jethwa, A., Stolz, T., Slabicki, M., Lee, K., Sharathchandra, A., Benner, A*., et al.* (2015). p53-dependent non-coding RNA networks in chronic lymphocytic leukemia. Leukemia *29*, 2015-2023.
Bollen, J., Mao, H.N., and Zeng, X.J. (2011). Twitter mood predicts the stock market. Journal of Computational Science *2*, 1-8.
Bonabeau, E., Dorigo, M., and Theraulaz, G. (2000). Inspiration for optimization from social insect behaviour. Nature *406*, 39-42.
Boyd, D.M., and Ellison, N.B. (2007). Social network sites: Definition, history, and scholarship. J Comput-Mediat Comm *13*, 210-230.
Brown, T.A. (2002). Genomes. Wiley-Liss.
Burgess, J., and Bruns, A. (2012). Twitter archives and the challenges of" Big Social Data" for media and communication research. M/C Journal.
Campbell, M., Hoane, A.J., and Hsu, F.H. (2002). Deep blue. Artificial Intelligence *134*, 57-83.
Cano, I., Lluch-Ariet, M., Gomez-Cabrero, D., Maier, D., Kalko, S., Cascante, M., Tegner, J., Miralles, F., Herrera, D., Roca, J*., et al.* (2014). Biomedical research in a Digital Health Framework. J Transl Med *12 Suppl 2*, S10.
Carneiro, H.A., and Mylonakis, E. (2009). Google trends: a web-based tool for real-time surveillance of disease outbreaks. Clin Infect Dis *49*, 1557-1564.
Cha, M., Mislove, A., and Gummadi, K.P. (2009). A measurement-driven analysis of information propagation in the flickr social network. Proceedings of the 18th International Conference on World Wide Web, 721--730.
Chen, C.L.P., and Zhang, C.Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences *275*, 314-347.
Chen, H., and Sharp, B.M. (2004). Content-rich biological network constructed by mining PubMed abstracts. BMC Bioinformatics *5*, 147.
Chen, M., Mao, S., and Liu, Y. (2014). Big data: A survey. Mobile Networks and Applications.
Choudhury, R. (2003). Application of Evolutionary Algorithms for Multiple Sequence Alignment. In Application of Evolutionary Algorithms for Multiple Sequence Alignment (Computational Molecular Biology. Biochemistry).
Clarke, L., Zheng-Bradley, X., Smith, R., Kulesha, E., Xiao, C., Toneva, I., Vaughan, B., Preuss, D., Leinonen, R., Shumway, M*., et al.* (2012). The 1000 Genomes Project: data management and community access. Nat Methods *9*, 459-462.
Consortium, E.P., Birney, E., Stamatoyannopoulos, J.A., Dutta, A., Guigo, R., Gingeras, T.R., Margulies, E.H., Weng, Z., Snyder, M., Dermitzakis, E.T*., et al.* (2007). Identification and analysis of functional elements in 1% of the human genome by the ENCODE pilot project. Nature *447*, 799-816.
Cook, K.A., and Thomas, J.J. (2005). Illuminating the path: The research and development agenda for visual analytics. Illuminating the path: The research and development agenda for visual analytics.
Crespo, I., Doucey, M.A., and Xenarios, I. (2015). Can social networks help to infer causality in the tumor microenvironment? BC2.

Crooks, G.E., Hon, G., Chandonia, J.M., and Brenner, S.E. (2004). WebLogo: a sequence logo generator. Genome Res *14*, 1188-1190.
Cunningham, F., Amode, M.R., Barrell, D., Beal, K., Billis, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates, G., Fitzgerald, S.*, et al.* (2015). Ensembl 2015. Nucleic Acids Res *43*, D662-669.
Cunningham, H., Tablan, V., Roberts, A., and Bontcheva, K. (2013). Getting more out of biomedical documents with GATE's full lifecycle open source text analytics. PLoS Comput Biol *9*, e1002854.
Cunningham, T.J., and Duester, G. (2015). Mechanisms of retinoic acid signalling and its roles in organ and limb development. Nat Rev Mol Cell Biol *16*, 110-123.
Danecek, P., Auton, A., Abecasis, G., Albers, C.A., Banks, E., DePristo, M.A., Handsaker, R.E., Lunter, G., Marth, G.T., Sherry, S.T.*, et al.* (2011). The variant call format and VCFtools. Bioinformatics *27*, 2156-2158.
Dardel, F., and Bensoussan, P. (1988). DNAid: a Macintosh full screen editor featuring a built-in regular expression interpreter for the search of specific patterns in biological sequences using finite state automata. Comput Appl Biosci *4*, 483-486.
Dayhoff, M.O. (1965). Atlas of protein sequence and structure. Atlas of protein sequence and structure.
Deans, A.R., Lewis, S.E., Huala, E., Anzaldo, S.S., Ashburner, M., Balhoff, J.P., Blackburn, D.C., Blake, J.A., Burleigh, J.G., Chanet, B.*, et al.* (2015). Finding our way through phenotypes. PLoS Biol *13*, e1002033.
Devi, G.R. (2006). siRNA-based approaches in cancer therapy. Cancer Gene Ther *13*, 819-829.
Doms, A., and Schroeder, M. (2005). GoPubMed: exploring PubMed with the Gene Ontology. Nucleic Acids Res *33*, W783-786.
Donath, J., and Boyd, D. (2004). Public displays of connection. Bt Technology Journal *22*, 71-82.
Dumais, S.T. (2004). Latent semantic analysis. Annu Rev Inform Sci *38*, 189-230.
Eddy, S.R. (1998). Profile hidden Markov models. Bioinformatics *14*, 755-763.
Eddy, S.R. (2012). The C-value paradox, junk DNA and ENCODE. Curr Biol *22*, R898-899.
Ehrenberg, M., Elf, J., and Hohmann, S. (2009). Systems Biology: Nobel Symposium 146. FEBS Lett *583*, 3881.
Esteller, M. (2011). Non-coding RNAs in human disease. Nat Rev Genet *12*, 861-874.
Feigelson, E.D., and Babu, G.J. (2012). Big data in astronomy. Significance *9*, 22-25.
Fernald, G.H., Capriotti, E., Daneshjou, R., Karczewski, K.J., and Altman, R.B. (2011). Bioinformatics challenges for personalized medicine. Bioinformatics *27*, 1741-1748.
Ferrucci, D., Levas, A., Bagchi, S., Gondek, D., and Mueller, E.T. (2013). Watson: Beyond Jeopardy! Artificial Intelligence *199*, 93-105.
Finn, R.D., Clements, J., and Eddy, S.R. (2011). HMMER web server: interactive sequence similarity searching. Nucleic Acids Res *39*, W29-37.
Fitch, W.M. (1970). Distinguishing homologous from analogous proteins. Syst Zool *19*, 99-113.
Fjeldberg, H.C. (2008). Polyglot Programming. Polyglot Programming.
Gehlenborg, N., O'Donoghue, S.I., Baliga, N.S., Goesmann, A., Hibbs, M.A., Kitano, H., Kohlbacher, O., Neuweger, H., Schneider, R., Tenenbaum, D.*, et al.* (2010). Visualization of omics data for systems biology. Nat Methods *7*, S56-68.
Gene Ontology, C. (2015). Gene Ontology Consortium: going forward. Nucleic Acids Res *43*, D1049-1056.
Genomes Project, C., Abecasis, G.R., Altshuler, D., Auton, A., Brooks, L.D., Durbin, R.M., Gibbs, R.A., Hurles, M.E., and McVean, G.A. (2010). A map of human genome variation from population-scale sequencing. Nature *467*, 1061-1073.
Gheldof, N., Smith, E.M., Tabuchi, T.M., Koch, C.M., Dunham, I., Stamatoyannopoulos, J.A., and Dekker, J. (2010). Cell-type-specific long-range looping interactions identify distant regulatory elements of the CFTR gene. Nucleic Acids Res *38*, 4325-4336.
Giglia, E., and Spinelli, O. (2009). PubMed reloaded: new interface, enhanced discovery. Eur J Phys Rehabil Med *45*, 631-636.

Goh, K.I., and Choi, I.G. (2012). Exploring the human diseasome: the human disease network. Brief Funct Genomics *11*, 533-542.
Goodhead, I., and Darby, A.C. (2015). Taking the pseudo out of pseudogenes. Curr Opin Microbiol *23*, 102-109.
Greene, J.A., Choudhry, N.K., Kilabuk, E., and Shrank, W.H. (2011). Online social networking by patients with diabetes: a qualitative evaluation of communication with Facebook. J Gen Intern Med *26*, 287-292.
Grimmelmann, J. (2008). Saving facebook. Saving facebook.
Hawn, C. (2009). Take Two Aspirin And Tweet Me In The Morning: How Twitter, Facebook, And Other Social Media Are Reshaping Health Care. Health Affairs *28*, 361-368.
Heer, J., Card, S.K., and Landay, J.A. (2005). Prefuse: a toolkit for interactive information visualization. Paper presented at: Prefuse: a toolkit for interactive information visualization (ACM).
Hendlisz, A. (2015). Of art and science: is personalized medicine getting personal enough? Curr Opin Oncol *27*, 349-350.
Henn, B.M., Botigue, L.R., Bustamante, C.D., Clark, A.G., and Gravel, S. (2015). Estimating the mutation load in human genomes. Nat Rev Genet *16*, 333-343.
Hey, T. (2012). The Fourth Paradigm - Data-Intensive Scientific Discovery. Comm Com Inf Sc *317*, 1-1.
Hildebrandt, F., Benzing, T., and Katsanis, N. (2011). Ciliopathies. N Engl J Med *364*, 1533-1543.
Hodis, E., Prilusky, J., and Sussman, J.L. (2010). Proteopedia: A collaborative, virtual 3D web-resource for protein and biomolecule structure and function. Biochem Mol Biol Educ *38*, 341-342.
Hoffmann, R. (2008). A wiki for the life sciences where authorship matters. Nat Genet *40*, 1047-1051.
Hotho, A., Nürnberger, A., and Paaß, G. (2005). A Brief Survey of Text Mining. Ldv Forum.
Huang, D.W., Sherman, B.T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M.W., Lane, H.C*., et al.* (2007). DAVID Bioinformatics Resources: expanded annotation database and novel algorithms to better extract biology from large gene lists. Nucleic Acids Res *35*, W169-175.
Iseli, C., Ambrosini, G., Bucher, P., and Jongeneel, C.V. (2007). Indexing strategies for rapid searches of short words in genome sequences. PloS one *2*, e579.
Jacobson, D., Woods, D., and Brail, G. (2011). APIs: A strategy guide. APIs: A strategy guide.
Jamali, S., and Rangwala, H. (2009). Digging Digg: Comment Mining, Popularity Prediction, and Social Network Analysis. Wism: 2009 International Conference on Web Information Systems and Mining, Proceedings, 32-38.
Jee, K., and Kim, G.H. (2013). Potentiality of big data in the medical sector: focus on how to reshape the healthcare system. Healthc Inform Res *19*, 79-85.
Jiménez-Dıaz, G., and Menéndez, H.D. (2011). Predicting performance in team games. II for Systems.
Jivani, A.G. (2011). A comparative study of stemming algorithms. Int J Comp Tech Appl.
Kamphans, T., and Krawitz, P.M. (2012). GeneTalk: an expert exchange platform for assessing rare sequence variants in personal genomes. Bioinformatics *28*, 2515-2516.
Katz, L. (1953). A new status index derived from sociometric analysis. Psychometrika *18*, 39-43.
Kent, W.J. (2002). BLAT--the BLAST-like alignment tool. Genome Res *12*, 656-664.
Khorkova, O., Hsiao, J., and Wahlestedt, C. (2015). Basic biology and therapeutic implications of lncRNA. Adv Drug Deliv Rev *87*, 15-24.
King, S.P., Burgener, C., Paretti, C.T., and Davis, M.E. (2010). System and method for contextual advertising based on status messages.
Klein, R.J., Zeiss, C., Chew, E.Y., Tsai, J.Y., Sackler, R.S., Haynes, C., Henning, A.K., SanGiovanni, J.P., Mane, S.M., Mayne, S.T*., et al.* (2005). Complement factor H polymorphism in age-related macular degeneration. Science *308*, 385-389.
Kohl, M., Wiese, S., and Warscheid, B. (2011). Cytoscape: software for visualization and analysis of biological networks. Methods Mol Biol *696*, 291-303.

Kohler, S., Doelken, S.C., Mungall, C.J., Bauer, S., Firth, H.V., Bailleul-Forestier, I., Black, G.C., Brown, D.L., Brudno, M., Campbell, J.*, et al.* (2014). The Human Phenotype Ontology project: linking molecular biology and disease through phenotype data. Nucleic Acids Res *42*, D966-974.
Kohler, S., Schulz, M.H., Krawitz, P., Bauer, S., Dolken, S., Ott, C.E., Mundlos, C., Horn, D., Mundlos, S., and Robinson, P.N. (2009). Clinical diagnostics in human genetics with semantic similarity searches in ontologies. Am J Hum Genet *85*, 457-464.
Korb, O., Monecke, P., Hessler, G., Stutzle, T., and Exner, T.E. (2010). pharmACOphore: multiple flexible ligand alignment based on ant colony optimization. J Chem Inf Model *50*, 1669-1681.
Lalevee, S., Anno, Y.N., Chatagnon, A., Samarut, E., Poch, O., Laudet, V., Benoit, G., Lecompte, O., and Rochette-Egly, C. (2011). Genome-wide in silico identification of new conserved and functional retinoic acid receptor response elements (direct repeats separated by 5 bp). J Biol Chem *286*, 33322-33334.
Lally, A., and Fodor, P. (2011). Natural language processing with prolog in the IBM watson system. Retrieved June.
Lampe, C., Ellison, N., and Steinfield, C. (2007). A Familiar Face(book): Profile Elements as Signals in an Online Social Network. Conference on Human Factors in Computing Systems, Vols 1 and 2, 435-444.
Lander, E.S., Linton, L.M., Birren, B., Nusbaum, C., Zody, M.C., Baldwin, J., Devon, K., Dewar, K., Doyle, M., FitzHugh, W.*, et al.* (2001). Initial sequencing and analysis of the human genome. Nature *409*, 860-921.
Landrum, M.J., Lee, J.M., Riley, G.R., Jang, W., Rubinstein, W.S., Church, D.M., and Maglott, D.R. (2014). ClinVar: public archive of relationships among sequence variation and human phenotype. Nucleic Acids Res *42*, D980-985.
Langmead, B., and Salzberg, S.L. (2012). Fast gapped-read alignment with Bowtie 2. Nat Methods *9*, 357-359.
Lappalainen, I., Lopez, J., Skipper, L., Hefferon, T., Spalding, J.D., Garner, J., Chen, C., Maguire, M., Corbett, M., Zhou, G.*, et al.* (2013). DbVar and DGVa: public archives for genomic structural variation. Nucleic Acids Res *41*, D936-941.
Lee, E., Helt, G.A., Reese, J.T., Munoz-Torres, M.C., Childers, C.P., Buels, R.M., Stein, L., Holmes, I.H., Elsik, C.G., and Lewis, S.E. (2013). Web Apollo: a web-based genomic annotation editing platform. Genome Biol *14*, R93.
Li, J.B., Gerdes, J.M., Haycraft, C.J., Fan, Y.L., Teslovich, T.M., May-Simera, H., Li, H.T., Blacque, O.E., Li, L.Y., Leitch, C.C.*, et al.* (2004). Comparative and basal genomics identifies a flagellar and basal body proteome that includes the BBS5 human disease gene. Cell *117*, 541-552.
Linard, B., Allot, A., Schneider, R., Morel, C., Ripp, R., Bigler, M., Thompson, J.D., Poch, O., and Lecompte, O. (2015). OrthoInspector 2.0: Software and database updates. Bioinformatics *31*, 447-448.
Linard, B., Thompson, J.D., Poch, O., and Lecompte, O. (2011). OrthoInspector: comprehensive orthology analysis and visual exploration. BMC Bioinformatics *12*, 11.
Linville, A., Gumusaneli, E., Chandraratna, R.A., and Schilling, T.F. (2004). Independent roles for retinoic acid in segmentation and neuronal differentiation in the zebrafish hindbrain. Dev Biol *270*, 186-199.
Liu, W., Pellegrini, M., and Wang, X. (2014). Detecting communities based on network topology. Sci Rep *4*, 5739.
Liu, Y., Zhang, R., and Ying, K. (2015). Long noncoding RNAs: novel links in respiratory diseases (review). Mol Med Rep *11*, 4025-4031.
Lochbaum, K.E., and Streeter, L.A. (1989). Comparing and Combining the Effectiveness of Latent Semantic Indexing and the Ordinary Vector-Space Model for Information-Retrieval. Information Processing & Management *25*, 665-676.
Lotia, S., Montojo, J., Dong, Y., Bader, G.D., and Pico, A.R. (2013). Cytoscape app store. Bioinformatics *29*, 1350-1351.

Luo, H., Sun, Y., Wei, G., Luo, J., Yang, X., Liu, W., Guo, M., and Chen, R. (2015). Functional Characterization of Long Noncoding RNA Lnc_bc060912 in Human Lung Carcinoma Cells. Biochemistry *54*, 2895-2902.
Ma, H., Zhou, D., Liu, C., Lyu, M.R., and King, I. (2011). Recommender systems with social regularization. Proceedings of the Fourth ACM International Conference on Web Search and Data Mining, 287-296.
MacCaw, A. (2012). The Little Book on CoffeeScript (" O'Reilly Media, Inc.").
Magrane, M., and Consortium, U. (2011). UniProt Knowledgebase: a hub of integrated protein data. Database (Oxford) *2011*, bar009.
Mahela, O.P., Shaik, A.G., and Gupta, N. (2015). A critical review of detection and classification of power quality events. Renew Sust Energ Rev *41*, 495-505.
Marcotte, E.M., Xenarios, I., and Eisenberg, D. (2001). Mining literature for protein-protein interactions. Bioinformatics *17*, 359-363.
McLaren, W., Pritchard, B., Rios, D., Chen, Y., Flicek, P., and Cunningham, F. (2010). Deriving the consequences of genomic variants with the Ensembl API and SNP Effect Predictor. Bioinformatics *26*, 2069-2070.
McLean, C.Y., Bristor, D., Hiller, M., Clarke, S.L., Schaar, B.T., Lowe, C.B., Wenger, A.M., and Bejerano, G. (2010). GREAT improves functional interpretation of cis-regulatory regions. Nature biotechnology *28*, 495-501.
McPherson, M., Smith-Lovin, L., and Cook, J.M. (2001). Birds of a feather: Homophily in social networks. Annual Review of Sociology *27*, 415-444.
Medina-Rivera, A., Defrance, M., Sand, O., Herrmann, C., Castro-Mondragon, J.A., Delerce, J., Jaeger, S., Blanchet, C., Vincens, P., Caron, C.*, et al.* (2015). RSAT 2015: Regulatory Sequence Analysis Tools. Nucleic Acids Res *43*, W50-56.
Miner, S. (2012). Using NoSQL. Using NoSQL.
Moynihan, R. (2003). The making of a disease: female sexual dysfunction. Bmj *326*, 45-47.
Nabi, Z. (2013). Big Data: The Next Frontier. Big Data: The Next Frontier.
Nadkarni, A., and Hofmann, S.G. (2012). Why Do People Use Facebook? Pers Individ Dif *52*, 243-249.
Nagy, A., and Patthy, L. (2014). FixPred: a resource for correction of erroneous protein sequences. Database (Oxford) *2014*, bau032.
Ng, P.C., and Henikoff, S. (2003). SIFT: predicting amino acid changes that affect protein function. Nucleic Acids Res *31*, 3812-3814.
Norman, G. (2011). Editorial: Medicine man meets machine. Adv Health Sci Educ Theory Pract *16*, 147-150.
Novina, C.D., and Sharp, P.A. (2004). The RNAi revolution. Nature *430*, 161-164.
Okuda, S., Yamada, T., Hamajima, M., Itoh, M., Katayama, T., Bork, P., Goto, S., and Kanehisa, M. (2008). KEGG Atlas mapping for global analysis of metabolic pathways. Nucleic Acids Res *36*, W423-426.
Onisko, A., Druzdzel, M.J., and Wasyluk, H. (2001). Learning Bayesian network parameters from small data sets: application of Noisy-OR gates. Int J Approx Reason *27*, 165-182.
Paice, C.D. (1990). Another stemmer. ACM SIGIR Forum *24*, 56-61.
Park, P.J. (2009). ChIP-seq: advantages and challenges of a maturing technology. Nat Rev Genet *10*, 669-680.
Pereira, N.L., Sargent, D.J., Farkouh, M.E., and Rihal, C.S. (2015). Genotype-based clinical trials in cardiovascular disease. Nat Rev Cardiol *12*, 475-487.
Piasecki, B.P., Burghoorn, J., and Swoboda, P. (2010). Regulatory Factor X (RFX)-mediated transcriptional rewiring of ciliary genes in animals. Proc Natl Acad Sci U S A *107*, 12969-12974.
Pollard, K.S., Hubisz, M.J., Rosenbloom, K.R., and Siepel, A. (2010). Detection of nonneutral substitution rates on mammalian phylogenies. Genome Res *20*, 110-121.

Pop, F. (2014). High Performance Numerical Computing for High Energy Physics: A New Challenge for Big Data Science. Advances in High Energy Physics.
Porter, M.F. (1980). An Algorithm for Suffix Stripping. Program-Autom Libr *14*, 130-137.
Prosdocimi, F., Linard, B., Pontarotti, P., Poch, O., and Thompson, J.D. (2012). Controversies in modern evolutionary biology: the imperative for error detection and quality control. BMC Genomics *13*, 5.
Rasheed, M.M., and Ghazali, O. (2010). Server Worm Detection by Using Intelligent Failure Connection Algorithm. Computer Science & Telecommunications *27*.
Řehůřek, R., and Sojka, P. (2011). Gensim—Statistical Semantics in Python. Gensim—Statistical Semantics in Python.
Reyes-Palomares, A., Rodriguez-Lopez, R., Ranea, J.A., Sanchez Jimenez, F., and Medina, M.A. (2013). Global analysis of the human pathophenotypic similarity gene network merges disease module components. PLoS One *8*, e56653.
Rochel, N., Ciesielski, F., Godet, J., Moman, E., Roessle, M., Peluso-Iltis, C., Moulin, M., Haertlein, M., Callow, P., Mely, Y.*, et al.* (2011). Common architecture of nuclear receptor heterodimers on DNA direct repeat elements with different spacings. Nat Struct Mol Biol *18*, 564-570.
Rochette-Egly, C., and Germain, P. (2009). Dynamic and combinatorial control of gene expression by nuclear retinoic acid receptors (RARs). Nucl Recept Signal *7*, e005.
Rodriguez, J.M., Maietta, P., Ezkurdia, I., Pietrelli, A., Wesselink, J.J., Lopez, G., Valencia, A., and Tress, M.L. (2013). APPRIS: annotation of principal and alternative splice isoforms. Nucleic Acids Res *41*, D110-117.
Rosenbloom, K.R., Armstrong, J., Barber, G.P., Casper, J., Clawson, H., Diekhans, M., Dreszer, T.R., Fujita, P.A., Guruvadoo, L., Haeussler, M.*, et al.* (2015). The UCSC Genome Browser database: 2015 update. Nucleic Acids Res *43*, D670-681.
Roure, D.D., Goble, C., and Bhagat, J. (2008). myExperiment: Defining the social virtual research environment. eScience, 182-189.
Sadakane, K. (2007). Compressed suffix trees with full functionality. Theory of Computing Systems *41*, 589-607.
Samarut, E., and Rochette-Egly, C. (2012). Nuclear retinoic acid receptors: conductors of the retinoic acid symphony during development. Mol Cell Endocrinol *348*, 348-360.
Sandelin, A., Alkema, W., Engstrom, P., Wasserman, W.W., and Lenhard, B. (2004). JASPAR: an open-access database for eukaryotic transcription factor binding profiles. Nucleic Acids Res *32*, D91-94.
Sarkar, C., and Maitra, A. (2008). Deciphering the cis-regulatory elements of co-expressed genes in PCOS by in silico analysis. Gene *408*, 72-84.
Scheidecker, S., Etard, C., Pierce, N.W., Geoffroy, V., Schaefer, E., Muller, J., Chennen, K., Flori, E., Pelletier, V., Poch, O.*, et al.* (2014). Exome sequencing of Bardet-Biedl syndrome patient identifies a null mutation in the BBSome subunit BBIP1 (BBS18). J Med Genet *51*, 132-136.
Schwartz, S., Kent, W.J., Smit, A., Zhang, Z., Baertsch, R., Hardison, R.C., Haussler, D., and Miller, W. (2003). Human-mouse alignments with BLASTZ. Genome Res *13*, 103-107.
Scully, J.L. (2004). What is a disease? EMBO Rep *5*, 650-653.
Serrano-Pozo, A., Qian, J., Monsell, S.E., Betensky, R.A., and Hyman, B.T. (2015). APOEepsilon2 is associated with milder clinical and pathological Alzheimer disease. Ann Neurol *77*, 917-929.
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.
Sidiroglou, S., Locasto, M.E., Boyd, S.W., and Keromytis, A.D. (2005). Building a reactive immune system for software services. USENIX Association Proceedings of the General Track: 2005 UNENIX Annual Technical Conference, 149-161.

Siepel, A., Bejerano, G., Pedersen, J.S., Hinrichs, A.S., Hou, M., Rosenbloom, K., Clawson, H., Spieth, J., Hillier, L.W., Richards, S.*, et al.* (2005). Evolutionarily conserved elements in vertebrate, insect, worm, and yeast genomes. Genome Res *15*, 1034-1050.
Sigdel, K.R., Cheng, A., Wang, Y., Duan, L., and Zhang, Y. (2015). The Emerging Functions of Long Noncoding RNA in Immune Cells: Autoimmune Diseases. J Immunol Res *2015*, 848790.
Signorini, A., Segre, A.M., and Polgreen, P.M. (2011). The use of Twitter to track levels of disease activity and public concern in the U.S. during the influenza A H1N1 pandemic. PLoS One *6*, e19467.
Silver, N. (2012). The weatherman is not a moron. New York Times.
Singh-Blom, U.M., Natarajan, N., Tewari, A., Woods, J.O., Dhillon, I.S., and Marcotte, E.M. (2013). Prediction and validation of gene-disease associations using methods inspired by social network analyses. PLoS One *8*, e58977.
Smart, J. (2011). Jenkins: the definitive guide. Jenkins: the definitive guide.
Southern, E., Mir, K., and Shchepinov, M. (1999). Molecular interactions on microarrays. Nat Genet *21*, 5-9.
Steen, R.G. (2011). Retractions in the scientific literature: do authors deliberately commit research fraud? J Med Ethics *37*, 113-117.
Su, A.I., Wiltshire, T., Batalov, S., Lapp, H., Ching, K.A., Block, D., Zhang, J., Soden, R., Hayakawa, M., Kreiman, G.*, et al.* (2004). A gene atlas of the mouse and human protein-encoding transcriptomes. Proc Natl Acad Sci U S A *101*, 6062-6067.
Sun, J., Ding, W., Zhi, J., and Chen, W. (2015). MiR-200 suppresses metastases of colorectal cancer through ZEB1. Tumour Biol.
Szklarczyk, D., Franceschini, A., Wyder, S., Forslund, K., Heller, D., Huerta-Cepas, J., Simonovic, M., Roth, A., Santos, A., Tsafou, K.P.*, et al.* (2015). STRING v10: protein-protein interaction networks, integrated over the tree of life. Nucleic Acids Res *43*, D447-452.
Tellier, I., Duchier, D., Eshkol, I., Courmet, A., and Martinet, M. (2012). Apprentissage automatique d'un chunker pour le français. Actes de TALN'12, papier court (poster).
Tempel, S. (2012). Using and understanding RepeatMasker. Methods Mol Biol *859*, 29-51.
Tenenbaum, J.D., Sansone, S.A., and Haendel, M. (2014). A sea of standards for omics data: sink or swim? J Am Med Inform Assoc *21*, 200-203.
Theissen, G. (2002). Secret life of genes. Nature *415*, 741.
Thompson, J.D., Gibson, T.J., and Higgins, D.G. (2002). Multiple sequence alignment using ClustalW and ClustalX. Curr Protoc Bioinformatics *Chapter 2*, Unit 2 3.
Tian, P. (2011). Convergence: Where West meets East. Nature *480*, S84-86.
Tikkinen, K.A., Leinonen, J.S., Guyatt, G.H., Ebrahim, S., and Jarvinen, T.L. (2012). What is a disease? Perspectives of the public, health professionals and legislators. BMJ Open *2*.
Topol, E.J. (2014). Individualized medicine from prewomb to tomb. Cell *157*, 241-253.
Tumasjan, A., Sprenger, T.O., Sandner, P.G., and Welpe, I.M. (2010). Predicting Elections with Twitter: What 140 Characters Reveal about Political Sentiment. ICWSM.
Turatsinze, J.V., Thomas-Chollier, M., Defrance, M., and van Helden, J. (2008). Using RSAT to scan genome sequences for transcription factor binding sites and cis-regulatory modules. Nat Protoc *3*, 1578-1588.
Ugander, J., Karrer, B., Backstrom, L., and Marlow, C. (2011). The anatomy of the facebook social graph. arXiv preprint arXiv:11114503.
Ukkonen, E. (1995). On-line construction of suffix trees. Algorithmica *14*, 249-260.
Välimäki, N., Mäkinen, V., Gerlach, W., and Dixit, K. (2007). Engineering a compressed suffix tree implementation. Experimental Algorithms, 217-228.
van Helden, J., Andre, B., and Collado-Vides, J. (2000). A web site for the computational analysis of yeast regulatory sequences. Yeast *16*, 177-187.

Van Landeghem, S., Bjorne, J., Wei, C.H., Hakala, K., Pyysalo, S., Ananiadou, S., Kao, H.Y., Lu, Z., Salakoski, T., Van de Peer, Y*., et al.* (2013). Large-scale event extraction from literature with multi-level gene normalization. PLoS One *8*, e55814.
Van Noorden, R. (2014). Online collaboration: Scientists and the social network. Nature *512*, 126-129.
Verburg, F.P., Goedegebure, M.C., and Tienen, V.F. (2014). Peer Matching Tool. Peer Matching Tool.
Vidal, M. (2009). A unifying view of 21st century systems biology. FEBS Lett *583*, 3891-3894.
Wang, M., Lamers, R.J., Korthout, H.A., van Nesselrooij, J.H., Witkamp, R.F., van der Heijden, R., Voshol, P.J., Havekes, L.M., Verpoorte, R., and van der Greef, J. (2005). Metabolomics in the context of systems biology: bridging traditional Chinese medicine and molecular pharmacology. Phytother Res *19*, 173-182.
Wasserman, W.W., and Sandelin, A. (2004). Applied bioinformatics for the identification of regulatory elements. Nat Rev Genet *5*, 276-287.
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.
Weaver, J., and Tarjan, P. (2013). Facebook Linked Data via the Graph API. Semantic Web *4*, 245-250.
Wei, C.H., Kao, H.Y., and Lu, Z. (2015). GNormPlus: An Integrative Approach for Tagging Genes, Gene Families, and Protein Domains. Biomed Res Int *2015*, 918710.
Weinreich, S.S., Mangon, R., Sikkens, J.J., Teeuw, M.E., and Cornel, M.C. (2008). [Orphanet: a European database for rare diseases]. Ned Tijdschr Geneeskd *152*, 518-519.
Wilkinson, L. (2012). Exact and approximate area-proportional circular Venn and Euler diagrams. IEEE Trans Vis Comput Graph *18*, 321-331.
Wingender, E. (2008). The TRANSFAC project as an example of framework technology that supports the analysis of genomic regulation. Brief Bioinform *9*, 326-332.
Wong, P.C., Shen, H.W., Johnson, C.R., Chen, C., and Ross, R.B. (2012). The top 10 challenges in extreme-scale visual analytics. IEEE Comput Graph Appl *32*, 63-67.
Wu, C., Mark, A., and Su, A.I. (2014). MyGene. info: gene annotation query as a service. bioRxiv.
Xiao-Jie, L., Ai-Mei, G., Li-Juan, J., and Jiang, X. (2015). Pseudogene in cancer: real functions and promising signature. J Med Genet *52*, 17-24.
Yachdav, G., Goldberg, T., Wilzbach, S., Dao, D., Shih, I., Choudhary, S., Crouch, S., Franz, M., Garcia, A., Garcia, L.J*., et al.* (2015). Anatomy of BioJS, an open source community for the life sciences. Elife *4*.
Zarifia, M.H., Ghalehjogh, N.K., and Baradaran-nia, M. (2015). A new evolutionary approach for neural spike detection based on genetic algorithm. Expert Syst Appl *42*, 462-467.
Zeeberg, B.R., Feng, W., Wang, G., Wang, M.D., Fojo, A.T., Sunshine, M., Narasimhan, S., Kane, D.W., Reinhold, W.C., Lababidi, S*., et al.* (2003). GoMiner: a resource for biological interpretation of genomic and proteomic data. Genome Biol *4*, R28.
Zhang, Z., and Gerstein, M. (2003). Of mice and men: phylogenetic footprinting aids the discovery of regulatory elements. J Biol *2*, 11.
Zhao, S., Zhong, L., and Wickramasuriya, J. (2011). Human as real-time sensors of social and physical events: A case study of twitter and sports games. arXiv preprint arXiv:11064300.
Zittrain, J. (2009). Law and Technology The End of the Generative Internet. Commun Acm *52*, 18-20.
Zuberi, K., Franz, M., Rodriguez, H., Montojo, J., Lopes, C.T., Bader, G.D., and Morris, Q. (2013). GeneMANIA prediction server 2013 update. Nucleic Acids Res *41*, W115-122.

Alexis ALLOT

MyGeneFriends :
vers un nouveau rapport entre chercheurs et mégadonnées

Résumé

Ces dernières années, la biologie a subi une profonde mutation, impulsée notamment par les technologies à haut débit et la montée de la génomique personnalisée. L'augmentation massive et constante de l'information biologique qui en résulte offre de nouvelles opportunités pour comprendre la fonction et l'évolution des gènes et génomes à différentes échelles et leurs rôles dans les maladies humaines. Ma thèse s'est articulée autour de la relation entre chercheurs et information biologique, et j'ai contribué à (OrthoInspector) ou créé (Parsec, MyGeneFriends) des systèmes permettant aux chercheurs d'accéder, analyser, visualiser, filtrer et annoter en temps réel l'énorme quantité de données disponibles à l'ère post génomique. MyGeneFriends est un premier pas dans une direction passionnante, faire en sorte que ce ne soient plus les chercheurs qui aillent vers l'information, mais que l'information pertinente aille vers les chercheurs sous une forme adaptée, permettant l'accès personnalisé et efficace aux grandes quantités d'informations, la visualisation de ces informations et leur interconnexion en réseaux.

Mots clefs : génomique, réseaux sociaux, maladies, personnalisation, infrastructure web, mégadonnées

Summary

In recent years, biology has undergone a profound evolution, mainly due to high throughput technologies and the rise of personal genomics. The resulting constant and massive increase of biological data offers unprecedented opportunities to decipher the function and evolution of genes and genomes at different scales and their roles in human diseases. My thesis addressed the relationship between researchers and biological information, and I contributed to (OrthoInspector) or created (Parsec, MyGeneFriends) systems allowing researchers to access, analyze, visualize, filter and annotate in real time the enormous quantity of data available in the post genomic era. MyGeneFriends is a first step in an exciting new direction: where researchers no longer search for information, but instead pertinent information is brought to researchers in a suitable form, allowing personalized and efficient access to large amounts of information, visualization of this information, and their integration in networks.

Keywords : genomics, social networks, diseases, personalisation, web framework, bigdata