Université de Bordeaux

HDR

Développements bioinformatiques pour la modélisation des réseaux biologiques : « Des réseaux de régulation aux réseaux métaboliques »

Patricia Thébault

Rapport scientifique présenté en vue de l'obtention de

L'Habilitation à Diriger des Recherches

Composition du jury

Alain Denise	Professeur LRI, Paris	Rapporteur
Claudine Médigue	Directeur de recherche CNRS, UMR8030 Genoscope	Rapporteur
Mohamed Elati	Maitre de conférences, ISSB, Paris	Rapporteur
Christine Gaspin	Directrice de recherche, INRA Toulouse	Examinateur
Jean Christophe Taveau	Professeur Université, Bordeaux	Examinateur
Rodolphe Thiébaut	PUPH, INSERM, CHU de Bordeaux	Examinateur
Guillaume Blin	Professeur Université de Bordeaux	Examinateur

TABLES DES MATIERES

PREAMBULE

Issue d'une formation de biochimiste, j'ai obtenu un DEA en 1996. C'est au cours de mon stage de recherche que j'ai découvert la bioinformatique en tant qu'utilisatrice biochimiste. Mon engouement pour cette discipline m'a motivée à poursuivre mon cursus par un DESS double compétence d'informatique appliqué aux sciences de la Vie et de la Terre, obtenu en 1996 et un doctorat soutenu en juillet 2004.

J'ai effectué mes premiers pas en bioinformatique suite au DESS en tant qu'ingénieur dans le cadre d'un CDD de deux ans dans l'équipe de Des Higgins à l'University College Cork en Irlande. Mes contributions ont consisté à spécifier et à co-développer un serveur de calcul et les interfaces clientes en java, dédiés à l'alignement de séquences ribosomales. J'ai ensuite rejoint pour deux ans l'équipe de Daniel Kahn du Laboratoire des Interactions Plantes-Micro-organismes (LIPM) de l'INRA de Toulouse. Mon projet consistait en l'accompagnement bioinformatique du projet international de séquençage de la bactérie symbiotique *S. meliloti* et du projet Génoscope-INRA/CNRS de séquençage de la bactérie pathogène *R. solanecearum*. En appui à J. Gouzy (LIPM, IR, INRA, Toulouse) j'ai développé l'environnement d'annotation semi-automatisé iANT, dont l'originalité dans les années 2000, a été de proposer un outil informatique permettant à la fois le traitement automatique des séquences et l'évaluation/édition des résultats par un annotateur expert. Cette période a été essentielle dans mon parcours pour étendre et renforcer ma culture bioinformatique sur l'ensemble des outils d'analyse de séquences et pour mettre à profit ma double compétence pour interagir étroitement avec des biologistes afin de répondre au mieux à leur demande.

Mon intérêt pour la compréhension et le développement de méthodes informatiques et mathématiques m'a ensuite amenée à joindre à ma formation un doctorat que j'ai souhaité réaliser dans un laboratoire d'informatique et de statistique, l'unité Biométrie et Intelligence Artificielle de l'INRA de Toulouse. Mon doctorat a été co-dirigé par Christine Gaspin et Thomas Schiex. Mon sujet de thèse a porté sur le développement d'une approche permettant la recherche de motifs structurés sous contraintes d'interactions dans les séquences génomiques.

Mes travaux de recherche se sont centrés sur la recherche de motifs structurés de type ARN avec l'objectif de proposer et d'implémenter, tout en restant efficace d'un point de vue algorithmique, une nouvelle modélisation intégrant la spécification des interactions intermoléculaires de type ARN/ARN. Je me suis ainsi intéressée à la modélisation et à la résolution de cette problématique en combinant le formalisme des réseaux de contraintes, thématique de recherche de mes encadrants, avec des algorithmes de pattern-matching/arbre des suffixes.

J'ai également investi le domaine de l'algorithmique du texte pour choisir les méthodes les plus appropriées à mes problématiques et j'ai développé en C++ un logiciel généraliste de recherche de motifs structurés. Afin de le rendre utilisable par la communauté biologiste, j'ai également développé un langage utilisateur pour décrire les molécules à rechercher et j'ai développé une interface web.

A l'issu de mon doctorat, j'ai rejoins l''équipe Baobab au sein du laboratoire Biométrie et Biologie Evolutive à Lyon pour y réaliser un post-doctorat de deux ans. Mes travaux de recherche se sont orientés vers la modélisation des réseaux métaboliques et m'ont permis de m'initier à des approches issues de la théorie des graphes et d'élargir l'ensemble de mes compétences.

Recrutée depuis le 1er septembre 2007 par l'Université de Bordeaux, j'occupe un poste de MCU en bioinformatique au sein de l'UF Biologie du collège ST de Bordeaux. Je suis rattachée à l'équipe MABioVis du LaBRI (responsable : Guillaume Blin) et j'exerce mon activité de recherche plus spécifiquement dans le thème "Exploration Visuelle et Analytique de Données Massives, EVADoMe" (responsable : Romain Bourqui). Mes travaux de recherche se focalisent sur le développement de nouvelles approches bioinformatiques, en collaboration avec des biologistes et informaticiens. Plus concrètement, je suis particulièrement intéressée pour contribuer à l'amélioration d'algorithmes existants afin de les rendre plus pertinents pour répondre à des questions biologiques mais aussi par le développement de nouveaux algorithmes et à leur mise en œuvre dans un contexte biologique. L'aspect finalisé de mes travaux est un point essentiel de mon activité de recherche en bioinformatique qui se focalise sur deux thèmes principaux incluant la mise en œuvre d'approches comparatives en biologie et le développement de nouveaux modèles pour la biologie des systèmes.

J'ai choisi de focaliser ce manuscrit sur mes travaux en biologie des systèmes, thématique que j'investis depuis maintenant plusieurs années à travers de nombreux projets et collaborations. Mes travaux, dans le contexte de la biologie des systèmes, font l'objet de nombreuses collaborations avec Romain Bourqui. Ils ont eu pour objectifs d'utiliser les méthodes d'optimisation en visualisation de graphes pour faciliter l'intégration et l'interprétation de grandes quantités de données biologiques. En me situant en tant qu'utilisatrice dans un premier temps, mes contributions dans le processus du *design* de systèmes informatiques ont porté sur l'identification de questions biologiques nécessitant des compétences pointues dans la mise à disposition de méthodes et outils pour la visualisation de grandes quantités de données. Je me suis aussi appuyée sur le savoir-faire et la dynamique de l'équipe pour acquérir de nouvelles compétences en visualisation des données dans le contexte de la biologie des systèmes.

CHAPITRE 1- CONTEXTE SCIENTIFIQUE: DES DONNEES BIOLOGIQUES A LA BIOINFORMATIQUE

Mes travaux de recherche se focalisent sur le développement de nouvelles approches bioinformatiques en étroite connexion avec les données biologiques, et impliquent de nombreuses collaborations entre des biologistes, bioinformaticiens et informaticiens. Plus concrètement, je suis particulièrement intéressée par l'application de résultats algorithmiques dans un contexte biologique pour, d'une part, proposer de nouvelles approches bioinformatiques et, d'autre part, mieux appréhender des questions scientifiques issues de la Biologie. L'aspect applicatif de nos travaux est un point essentiel et implique de débuter ce rapport par quelques éléments de contexte pour mettre en relation les données en Biologie, où la volumétrie et la complexité ne cessent d'augmenter, avec les développements informatiques nécessaires à leur étude. Dans ce cadre, je me suis particulièrement intéressée aux approches de visualisation en biologie des systèmes.

1.1 L'ARRIVEE DES DONNEES MASSIVES EN BIOLOGIE

Big Data in biology?
"Big Data" has surpassed "systems biology" and "omics" as the hottest buzzword in the biological sciences, but is there any substance behind the hype? Certainly, we have learned about various aspects of cell and molecular biology from the many individual high-throughput data sets that have been published in the past 15–20 years. These data, although useful as individual data sets, can provide much more knowledge when interrogated with Big Data approaches, such as applying integrative methods that leverage the heterogeneous data compendia in their entirety.
.
Issue de « Implications of Big Data for cell biology », *Mol. Biol. Cell*, vol. 26, n° 14, p. 2575-2578, juill. 2015.

En l'espace d'une vingtaine d'années, la révolution des technologies de production de données biologiques à haut débit, en diminuant les coûts et temps de production, a donné lieu à une augmentation sans précédent du nombre de données et de leur hétérogénéité [1]. Ces avancées technologiques impactent les approches expérimentales et continuent de révolutionner la biologie et l'ensemble de ses applications.

Impliquée dans la recherche en biologie depuis maintenant 17 ans, je peux témoigner de cette révolution et de son impact sur la nature des projets de recherche, ne serait-ce qu'au travers de mes différentes implications. Avant le début des années 2000, un projet ou une thèse en Biologie se focalisait essentiellement sur l'étude d'un gène ou d'une petite famille de gènes. J'ai eu la chance d'assister et de prendre part, dès 2001, à un premier changement d'échelle avec l'arrivée des nouvelles approches de séquençage. Dès lors, la nature des projets a été rapidement modifiée en se focalisant davantage sur l'analyse d'un génome entier. Dans ce contexte, j'ai été impliquée dans le développement d'un nouveau système d'annotation semi-automatique qui fut

en parallèle exploité par plusieurs projets sur le séquençage de bactéries modèles (*Sinorhizobium meliloti* et *Ralstonia solenecarum*). En 2005, avec l'avènement des nouvelles technologies de séquençage (NGS), un nouveau virage a été pris ajoutant une nouvelle dimension à ces projets. Il devenait possible d'étudier, en parallèle, plusieurs souches d'un même organisme sur la base, par exemple, de phénotypes d'intérêts (en analysant par approche de génomique comparative des souches pathogènes *versus* des souches vaccinales).
En 2008, j'ai ainsi été impliquée dans le projet ANR EVOLMYCO «Etude à grande échelle des génomes des mycoplasmes de ruminants : évolution et adaptation de bactéries minimales à des hôtes complexes » en prenant la responsabilité de la partie bioinformatique. Au cours de ce projet, nous avons assisté à l'évolution des différentes méthodes expérimentales de production de données et leur impact sur l'amélioration de la qualité des résultats obtenus.

Aujourd'hui, cette révolution continue de modifier les Sciences du Vivant. Il est aujourd'hui bien établi que la production de données massives permet d'accomplir à très grande échelle des analyses qui ne pouvaient pas l'être auparavant. Ainsi, les objectifs actuels de la communauté en Bioinformatique est de développer des méthodes et outils pour pouvoir prendre en compte simultanément l'ensemble de ces grandes masses de données afin d'en exploiter la valeur ajoutée.

1.1.1 UN ETAT DES LIEUX NON EXHAUSTIF

Un premier focus sur les données génomiques avec le génome humain

L'exemple le plus classique pour illustrer cette nouvelle ère de données massives en Biologie est celui des technologies de séquençage apparues en 2005. On peut facilement prendre conscience du changement d'échelle amené par les NGS en regardant les évolutions récentes des projets en Biologie.

Si l'on considère les travaux en génomique autour de l'humain, très rapidement après le projet « Human Genome Project » dans les années 2000, la taille et complexité des projets qui ont suivi démontrent parfaitement ce changement d'échelle. Ainsi, les projets ENCODE [2], HapMap [3] ou encore 1000 Genomes [4] comptent parmi les projets les plus importants et ont rendu nécessaire la génération de catalogues d'information. ENCODE fondé en 1999 était au départ centré sur l'humain. Aujourd'hui, il supporte plus de 80 espèces de vertébrés et propose des ressources très diverses telles que des ensembles de gènes, des alignements de génomes entiers, des répertoires de gènes homologues, des annotations sur les variants ou encore la détection de régions de régulations. Le projet HapMap, quant à lui, a eu pour objectif de développer une carte d'haplotypes sur le génome humain. A l'issu de ce projet, une base de données de plus de 3 millions de variants d'ADN a vu le jour, laquelle a permis de réaliser les premières études d'association du génome (GWAS) visant à localiser plus de 600 facteurs de risques génétiques pour différentes maladies (diabète, cancer du sein…). Après ce projet, 1000 Genomes a pris le relai en 2008 en exploitant et améliorant les applications à partir des technologies de séquençage. Ce projet est aujourd'hui dans sa troisième phase et rend disponible à la communauté scientifique une base de données résultant du séquençage de 2504 individus issus de 26 populations. Grâce à lui, le nombre de variants disponibles a dépassé les 80 millions. En se basant sur ces données, L'EBI, dans son rapport annuel de 2016, prédit que de 2 à 40 exabytes (1 exa = 10^9 gigabytes) de données génomiques humaines devraient être produites d'ici 2025 (voir **Figure 1**).

D'après [5] (voir également **Figure 1**), les estimations sur les 10 prochaines années montrent une multiplication du volume par deux tous les 7 à 12 mois, avec chaque année une augmentation de l'ordre de l'exabase ($1*10^{18}$) de séquences dans les bases de données. En

suivant ces estimations, et en considérant seulement les séquences génomiques, on devrait atteindre l'échelle du zettabase (1*10[21] bytes) en 2025 avec 2,5 millions d'espèces séquencées.
Il est également important de considérer ce bouleversement de production de volume de données en regardant les différents types de données omiques tels que :

- **les données de séquences** disponibles à l'ENA[1] représentaient **798,2 millions de séquences** au 16/01/2017. Les nouvelles technologies de séquençage ont rendu possible une diversification des applications portant sur le génome, l'épigénome, le transcriptome [6] et le dégradome. Comme illustrées dans la Figure 2, les méthodes-SEQ se diversifient selon les différentes questions biologiques les motivant. Par exemple, **les données métagénomiques**. ont connu également une croissance vertigineuse en 2015 notamment grâce à l'addition des ensembles de données provenant de l'expédition Tara Oceans dans L'EBI Metagenomics.
- **les données de métabolomiques** avec MetaboLights[2], un exemple parmi d'autres, représentaient **9,018 terabytes de données** au 24/11/2016. Ces données ont, à leur tour, connu la croissance la plus importante en 2016 [7] (pour une revue des productions des plateformes à haut débit en métabolomique voir [8]).
- **les données protéomiques** [9] stockées à UNIPROT et Trembl représentaient **73 millions de protéines** au 17/01/2017. Malgré la mise en place d'une procédure pour réduire la redondance en 2015 qui a permis de diminuer le nombre de protéines de 90 à 46 millions, leur volumétrie ne cesse d'augmenter et devrait prochainement ré-atteindre les 90 millions.

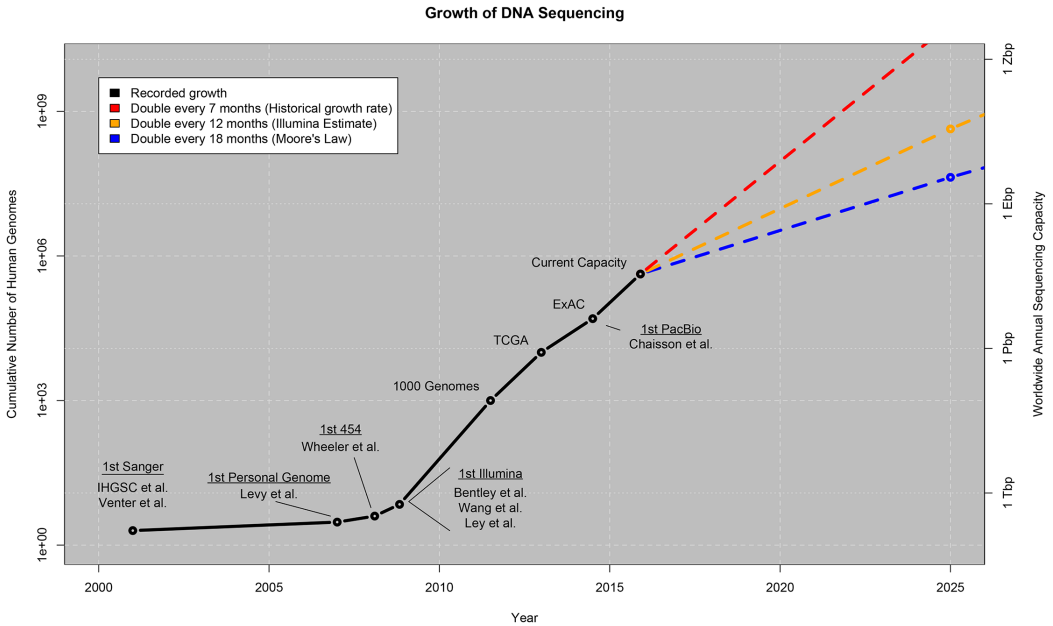


Figure 1 : La production de grandes masses de données impactent l'ensemble des approches omiques [5].

D'autres domaines sont également impactés par la production à haut débit. Par exemple, si on s'intéresse à l'imagerie, la production de données de structure vient récemment de

[1] ENA : European Nucleotide Archive < EMBL-EBI : www.ebi.ac.uk/ena

[2] MetaboLights (http://www.ebi.ac.uk/metabolights/) est une base de données d'expériences en métabolomiques.

connaître une envolée grâce à la microscopie électronique en 3D (3D EM) et a été nommée « *Technology of the Year* » par le journal Nature Methods en 2015 [7].

Cet état des lieux n'est bien sûr qu'une image partielle de la réalité mais il illustre déjà le changement d'échelle auquel la Biologie est en train de faire face. Parallèlement à cette production de données, le terme « Big Data », désignant les capacités technologiques à traiter de très grandes masses de données avec des infrastructures numériques, est également de plus en plus utilisé au sein de notre communauté.

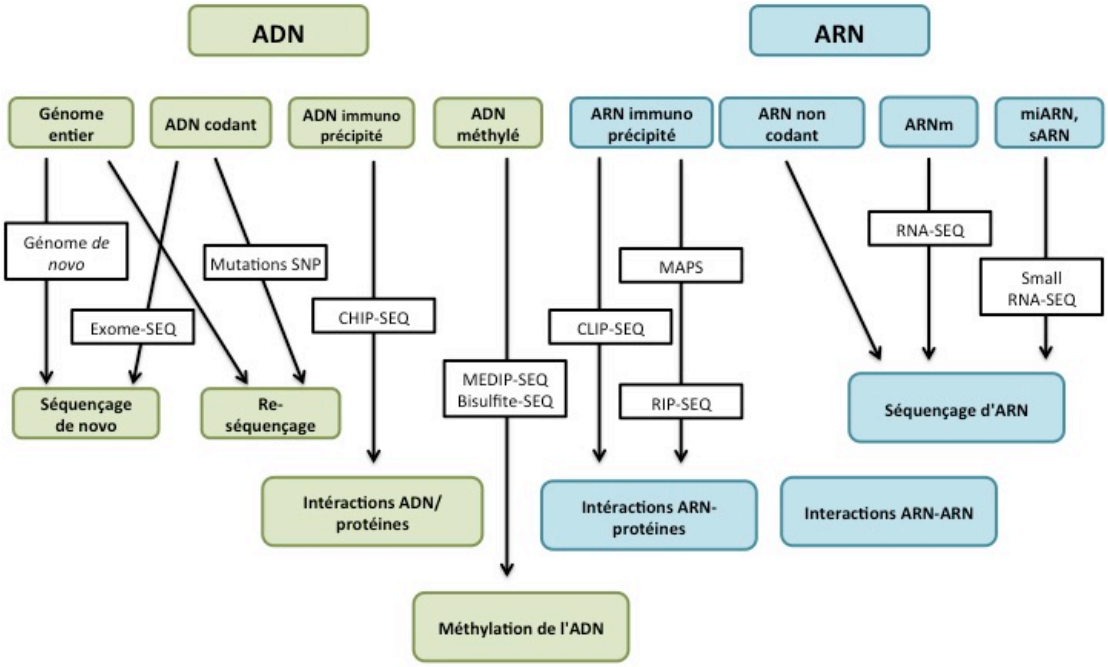


Figure 2: Liste non exhaustive d'applications issues du NGS.

Quelles sont les caractéristiques qui définissent les données massives en Biologie ?

Pour aborder cette question, nous nous sommes, dans un premier temps, tournés vers la littérature scientifique. Baro *et al* [10] ont ainsi réalisé une analyse statistique en 2015 à partir du corpus des publications de *pubmed*. Ayant pour objectif de proposer une définition du terme « Big Data » en Santé, ils ont collecté l'ensemble des articles contenant le terme Big Data dans le titre ou résumé des articles. Après une expertise manuelle, pour ne conserver que ceux qui étaient associés à un jeu de données expérimentales, ils ont calculé systématiquement deux valeurs pour chaque publication :

- ***n*** pour le nombre d'individus (ne se réfère pas forcément au nombre d'organismes mais peut également correspondre à un nombre de séquences) et
- ***p*** pour le nombre de variables analysées pour décrire un jeu de données expérimentales (p peut correspondre à des données qualitatives ou quantitatives (par exemple, les caractéristiques physicopchimiques des acides aminés)).

A partir de ces deux valeurs, ils ont observé une augmentation constante de $Log(n^* p)$ en fonction du temps entre 2009 et 2014 [10].

Il est cependant important de souligner qu'en Biologie, la production de données massives a largement précédé l'usage du terme Big Data, et pour aller plus loin, il serait pertinent de dénombrer pour chaque publication le nombre d'organismes étudiés (de même s'il s'agit de plusieurs souches d'une même bactérie ou plusieurs espèces), des volumes de données

brutes/valorisées obtenus, les différents types de données (s'il s'agit d'analyse combinant par exemple des données de protéomiques, génomiques, métabolomiques…).

En effet, ce type d'analyse permettrait certainement de confirmer que grâce aux nouvelles technologies haut débit, la tendance actuelle est de produire plusieurs types de données pour un même organisme, et/ou pour plusieurs organismes. Ainsi, les données et leurs productions sont devenues des produits de la recherche scientifique comme le démontre l'apparition de journaux dédiés à leur diffusion (par exemple, *Genome Annoucem*ent ou encore *Scientific Data* proposé par la revue Nature). La communauté bioinformatique est directement impactée par cet état de fait à plusieurs niveaux, allant du stockage de ces données à la nécessiter de les exploiter, dans des cadres globaux et intégratifs. Bien que les aspects liés au stockage des données soient un point crucial et nécessitent une attention très importante de la part de la communauté scientifique internationale (par exemple, en 2011, face à l'augmentation sans précédent du nombre de données en génomique, le NCBI a annoncé ne plus pouvoir prendre en charge le stockage et la mise à disposition des données non produites à partir de quelques organismes modèles [11]), ce mémoire s'intéresse davantage à leur accessibilité, leur exploitation et leur analyse par la communauté bioinformatique.

1.1.2 LES CONSEQUENCES DE L'AUGMENTATION DES DONNEES BIOLOGIQUES

Les conséquences en termes d'entrepôts et de bases de données

Une première conséquence directe de cette production de données massives est directement observable par la taille qu'elles occupent dans les bases de données nécessaires à leur stockage. En effet, l'EBI illustre parfaitement le changement d'échelle lié à cette production massive de données en devenant en 2013 l'un des entrepôts les plus importants de données biologiques avec un volume de 20 petabytes ($20*10^{15}$ bytes) de données [7]. En se focalisant sur les données biologiques, dont le volume atteignait 2 petabytes en 2014, leur quantité continue d'augmenter en doublant de volume chaque année.

Dispensant depuis plus de 10 ans des enseignements sur les bases de données en Biologie, j'ai eu à maintes reprises l'occasion de vérifier cette progression exponentielle, avec un changement d'échelle encore plus marqué dans les 3 dernières années. Bien entendu, il faut prendre en compte les avantages et inconvénients. Il est, par exemple, de plus en plus facile d'avoir accès aux séquences d'organismes encore très peu étudiés et de pouvoir ainsi améliorer la reconstruction d'un arbre phylogénétique pour l'étude d'une famille de gènes. Cependant, il n'est pas toujours évident de pouvoir expliquer et rassurer les étudiants sur l'expertise manuelle à effectuer lorsque, pour un même gène d'un organisme d'intérêt, on obtient un nombre astronomique de réponses à des requêtes ciblées. La redondance et la qualité des données que l'on trouve dans ces bases doivent être prises en compte et demandent une certaine expérience pour faire les bons choix.

On observe également, dans de nombreux domaines (par exemple les petits ARNs non codants, les domaines de protéines…), le développement de plusieurs ressources pour un même type de données. Il est souvent difficile de s'y retrouver pour un étudiant ou jeune chercheur. Pour illustrer cette complexité, il suffit de s'intéresser aux petits ARNs non codant dont le nombre a considérablement augmenté ces dernières années grâce au séquençage et pour lesquels plusieurs bases de données spécialisées existent. Dans ce contexte, pour aider l'accessibilité à ces différentes sources de données dispersées, *RNAcentral* [12] propose une ressource intégrée qui regroupe actuellement 23 bases de données expertes dédiées aux ARNs fonctionnels et offre ainsi une large palette de données en lien (englobant des micros ARNs, snoARNs, structure d'ARN etc.) avec l'intégration d'informations issus d'autres sources plus

généralistes (bases de données sur des organismes tels que *Arabidopsis thaliana*, la souris, l'humain…).

Il devient ainsi nécessaire d'avoir recours à des portails spécialisés, points d'entrée pour faciliter l'accès et l'utilisation de ces ressources.

En effet, si quelques années en arrière, il était possible, pour certaines communautés travaillant sur des organismes modèles, d'exploiter conjointement plusieurs sources d'informations, leur diversité et multiplicité actuelles rendent aujourd'hui cette intégration plus difficile à mettre en œuvre.

A partir de ce constat, Gomez-Cabrero *et al* [13] identifient deux points essentiels à prendre en compte par les producteurs de nouveaux types de données:

- mettre à profit les expériences passées pour réfléchir, très en amont, à la définition de standards permettant de rendre les données interopérables et ainsi de faciliter leur exploitation,
- prendre en compte la redondance des données pour proposer des solutions de synthèse.

Un besoin important pour la définition de standards pour décrire les données

Très rapidement, avec la diversité et les volumes de données produites, on a pu observer dans la communauté scientifique la mise en place de consortiums ayant pour objectif de proposer la définition de standards dédiés à différents types de données et offrant des solutions pour rendre les données interopérables, mais aussi pour stocker, partager et analyser l'ensemble des données [14].

Un exemple de développement de standard en Biologie des systèmes

En biologie des systèmes, on trouve actuellement un grand nombre de standards plus ou moins étendus dans leur façon de prendre en compte certaines particularités des données. Par exemple, *SYSTEMS BIOLOGY GRAPHICAL NOTATION* (*SBGN*, http://www.sbgn.org) vise à définir un standard de représentation graphique de diagrammes prenant en compte la connaissance experte des interactions et régulations moléculaires. *BIOLOGICAL PATHWAYS EXCHANGE* (*BIOPAX*, http://www.biopax.org), quant à lui, exploite une ontologie pour décrire formellement les voies métaboliques, les interactions moléculaires, les voies de transduction du signal, la régulation de l'expression des gènes et les interactions génétiques. Le plus largement utilisé par les logiciels de visualisation est *SYSTEMS BIOLOGY MARKUP LANGUAGE* (*SBML*, http://sbml.org/Main_Page) qui propose une description standard des modèles mathématiques de systèmes biochimiques [15]. Exploitant *SBML*, *SIMULATION EXPERIMENT DESCRIPTION MARKUP LANGUAGE* (*SED*-ML : http://sed-ml.org) propose un standard de description afin d'analyser et de réutiliser les résultats d'analyses de simulations.

Comme le montre cette diversité de standards dans un domaine très spécifique comme la biologie des systèmes, la variété importante des types et sources de données entraîne un besoin important de définition et surtout d'utilisation de ces standards [14].

En 2011, nous nous sommes confrontés à ce problème dans le cadre du travail de Bryan Hernandez (effectuant son stage de master du MIT, USA) et en collaboration avec Matt DeJongh (University Hope College, USA). Initialement motivés par la comparaison de deux méthodes de reconstruction de réseaux métaboliques de bactéries avec les approches de BIOCYC[3] et

[3] The Biocyc Database Collection : https://biocyc.or

RAST/SEED[4], nous pensions pouvoir définir un cadre de comparaison en nous appuyant sur le format SBML. Cependant, malgré l'existence de ce standard, les disparités d'usage des deux systèmes ont rendu la tâche ardue, voire impossible. Le projet a alors naturellement été recentré sur la comparaison et la visualisation de réseaux métaboliques entre deux bactéries en utilisant le logiciel Systryp [16], et en développant l'outil Rast2Systryp (non publié). Bien que le format SBML soit pris en charge par Systryp ou un autre logiciel de visualisation tel que Cytoscape, il a été nécessaire de développer un plugin à façon pour permettre la simple connexion entre un outil de prédiction de réseau et un outil de visualisation.

Depuis quelques années, l'accent semble être mis sur le développement d'initiatives visant à proposer de nouveaux outils et approches en amont de la recherche d'information à partir des données brutes (en intégrant par exemple une couche sémantique) afin de faciliter l'utilisation et l'interopérabilité de ces connaissances. Dans ce sens, l'initiative COMBINE (COmputational Modeling in BIology NEtwork) qui organise régulièrement des workshops (COMBINE 2016: 7th Computational Modeling in Biology Network Workshop) vise à améliorer l'interopérabilité entre ces standards et favoriser l'ensemble des efforts pour définir les nouveaux besoins.

En Europe pour l'ensemble des sciences du vivant

L'initiative *BIOSHARING* [5] est un point d'entrée pour accéder à une collection d'entrepôts dans les sciences du Vivant. Elle recense 671 standards ainsi que 833 entrepôts de données dans les domaines de la biologie, de la santé et de l'environnement.. Ce projet est un des composants du nœud ELIXIR UK, et s'intègre dans le projet *ELIXIR EXCELERATE*. Une autre conséquence de cette ère des Big Data est également de favoriser le foisonnement d'initiatives pour l'intégration de données et de services à l'instar de celles lancées par *ELIXIR* pour l'Europe.

La communauté scientifique des Sciences du Vivant a rapidement pris conscience des changements d'échelle induits par la production massive de données. Le caractère traditionnel du partage des données en biologie a très vite nécessité le besoin de développer de nouvelles bases de données, puis des standards pour faciliter et uniformiser leurs échanges au sein des communautés scientifiques, et enfin des entrepôts de bases de données pour faciliter l'exploitation de ces données dont le nombre et la dispersion ne cessent de croître.

1.1.3 QUALIFIER LES « BIG DATA »

L'explosion récente des masses de données numérisées avec Internet (connue sous l'appellation : *evolution of Internet of Things (IoT)*) n'a pas seulement impacté les Sciences du Vivant, même si, dans ce domaine, on observe une des plus grandes complexités résultant de la forte hétérogénéité des données biologiques. Sont également concernés de nombreux domaines avec des applications très diverses telles que les villes intelligentes [17] (par exemple avec le développement des stations de vélos libres), la cyber-sécurité [18] ou encore le développement des objets connectés.

En s'appuyant sur la littérature scientifique en informatique, où le terme Big Data est largement utilisé aujourd'hui, les principales caractéristiques citées sont la volumétrie importante et la complexité apportée par leur diversité. Il est cependant important de préciser que l'utilisation du terme Big Data fait référence aux données massives pluri-dimensionnelles ou

[4] http://www.theseed.org/wiki/Home_of_the_SEED

[5] BioSharing.org. Oxford University, 2009. Disp. sur : https://www.biosharing.org/biodbcore/

non structurées dont la production et le stockage se font aujourd'hui en continu avec une croissance exponentielle de leur volume.

Les notions classiquement utilisées pour qualifier les « Big data » sont les trois "Vs": volume, variété et vélocité. Ces trois « V » expriment les difficultés liées au traitement en raison de la combinaison de leur taille (**volume**), de la fréquence des mises à jour (**vélocité**) ou de leur diversité (**variété**). La **véracité** est un quatrième "V" parfois ajouté pour décrire les problèmes de fiabilité liés notamment à l'explosion du nombre de sources de collecte etc.. Certains articles mentionnent également un cinquième «V» pour la **valorisation**.

La projection de cette définition sur les Sciences du Vivant met facilement en relief les différents V impactés par cette ère de données massives. Volume, vélocité et variété définissent les données biologiques et justifient le besoin (très sensible) de réarranger les données entre elles. Par exemple, les séquenceurs haut débit génèrent de très grands volumes de données qu'il est d'abord nécessaire d'assembler avec des outils efficaces avant d'être en mesure d'analyser un génome.

1.1.4 Nouveaux challenges pour la communauté bioinformatique

Les nouveaux challenges pour une équipe de petite taille se situent actuellement dans la maîtrise des méthodes et des outils (mise en œuvre de ceux qui sont disponibles, développement de nouveaux pour répondre à des questions spécifiques) pour donner de la valeur aux données (d'expression, de séquences, protéomiques, imagerie…). Les solutions à développer doivent être pensées de manière à aider à la proposition de nouvelles hypothèses et à tendre vers une compréhension intégrative plus réaliste du Vivant.

Nos travaux actuels se positionnent dans ce contexte avec, comme principaux objectifs, de proposer (i) de nouvelles méthodes bioinformatiques pour le traitement de données hétérogènes et (ii) de les mettre en œuvre pour produire de nouvelles connaissances.

Data integration

Le besoin de connecter différentes sources de connaissances entre elles, avec l'objectif d'augmenter la plus-value de ces données en facilitant leur utilisation conjointe, n'est pas nouveau. Dès l'apparition des bases de données en Biologie, les différents systèmes utilisés par la communauté s'intéressaient déjà à cette question. La Figure 3 illustre un type de requête complexe qui peut être réalisée dans un système d'interrogation tel que SRS[6] en exploitant les liens croisés entre différentes bases de données [19].

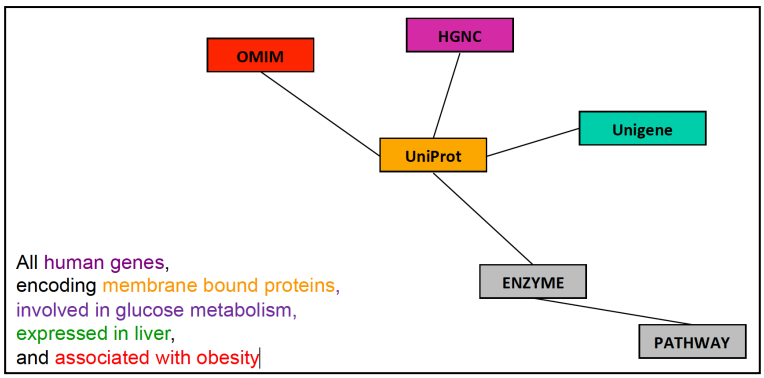


Figure 3 : Exemple de requête complexe exploitant les liens croisés entre différentes bases de données.

[6] Le service EMBL-EBI Sequence Retrieval System (SRS) a été remplacé par l'ENA le 19/12/2013.

Cependant, l'intégration ici consiste essentiellement à fournir une accessibilité aux données reliées entre elles (par exemple accéder à l'ensemble des fiches de protéines ou structure 3D reliées à la fiche d'un gène). Un niveau plus avancé serait d'être en mesure d'exploiter conjointement ces données en les intégrant à des approches de fouilles de données. Ce dernier point est devenu le verrou actuel depuis l'explosion de la quantité des données et l'augmentation de leur complexité. C'est dans ce contexte que le terme *analytics* (largement utilisé aujourd'hui par les médias) est apparu, lequel recouvre les nouvelles approches d'analyse de fouilles de données adaptées aux données massives.

Ainsi, l'intégration des données a donné lieu à de nouveaux challenges en fouille et analyse de données. Un exemple peut être donné avec les approches dites "trans-omics » [20] qui visent à connecter différents niveaux d'information afin d'améliorer les qualités prédictives de modèles biologiques *in silico* (voir Figure 4).

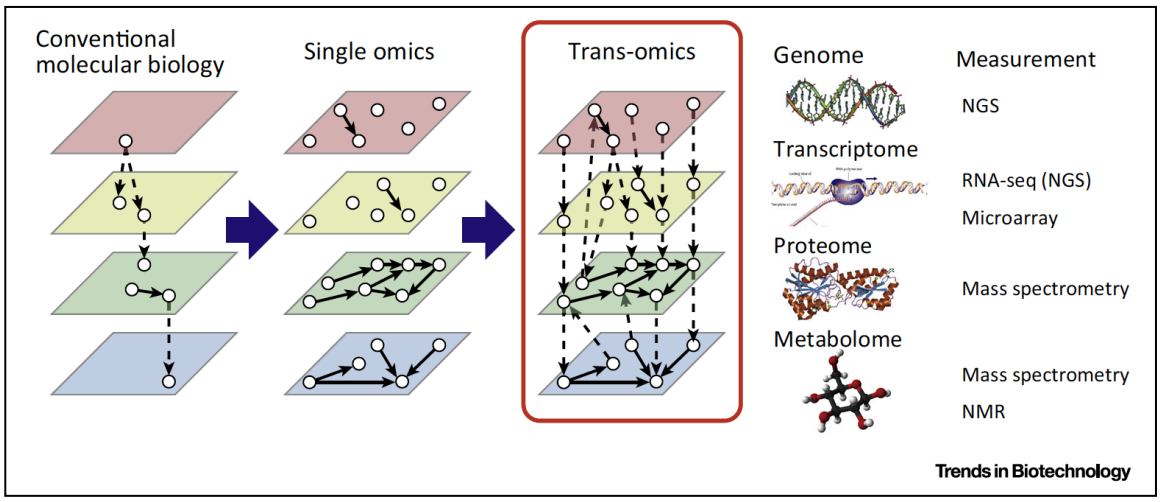


Figure 4 : Illustration du terme "trans omics" intégrant différents niveaux d'information issues des données omiques [20].

Plus concrètement, l'intégration se focalise le plus souvent sur deux types d'analyses :

- **Les analyses horizontales** (*horizontal integrative meta-analysis*) ont pour objectifs de combiner plusieurs études génomiques de même type (par exemple, des données d'expression, des analyses de méthylations…) afin d'améliorer la puissance et la fiabilité des analyses (pour une illustration sur les avantages de la combinaison de données d'expression issues de différents laboratoires voir [21])
- **Les analyses d'intégration verticales** (*vertical integrative analysis*) visent à combiner plusieurs types de données omiques (par exemple, données d'expression avec des données de méthylation …) issues d'un même organisme pour mieux appréhender son fonctionnement dans sa globalité et tendre vers une réalité biologique plus vraisemblable.

Actuellement, ce sont les analyses d'intégration verticale qui connaissent le plus grand essor. En effet, l'information extraite de la combinaison de différents types de données est aujourd'hui atteignable et « incontournable » pour réaliser des analyses *in-silico* de bien meilleure qualité [22].

Par le passé, les approches d'intégration se focalisaient principalement sur la prise en compte de deux types de données omiques au maximum, le plus souvent en raison de la

difficulté (coût et temps) de produire différents types de données pour les mêmes conditions expérimentales. Aujourd'hui, étendre ce type d'analyse ne nécessite pas toujours la production de nouvelles données grâce à l'intégration de connaissances déjà disponibles. Par exemple, une solution pourrait consister à exploiter la version 2.0 de la base de données COLOMBOS [23] qui donne un accès à des collections de données d'expression génétique spécifiques à un organisme, BEO et *micro-array*, et combinent des résultats et métadonnées d'expériences pour plusieurs organismes.

Dans la même veine, Ecomics [24], entièrement dédié à *Escherichia coli*, fournit un accès à une collection de données divisée en deux parties, afin de combiner les profils multi-omiques obtenus à l'échelle du génome avec les méta-données expérimentales (voir Figure 5).

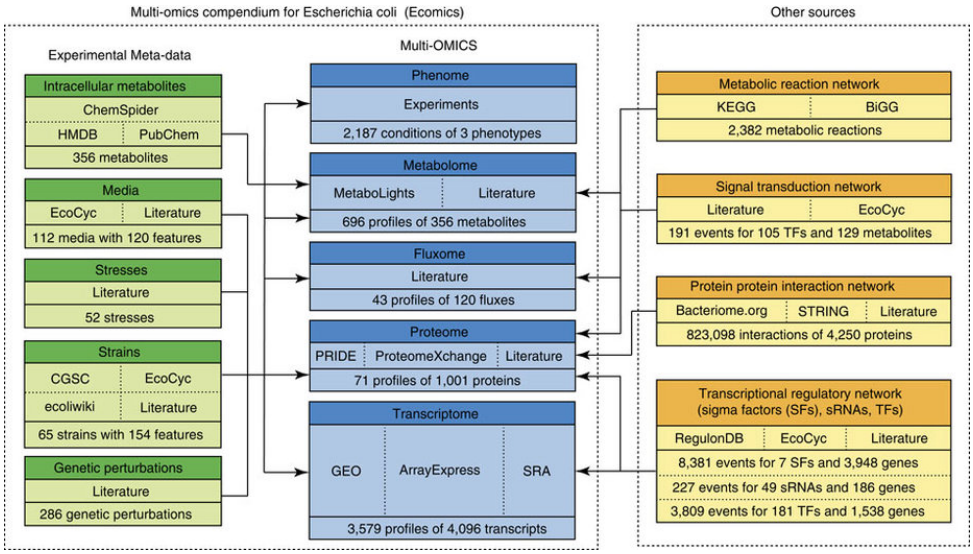


Figure 5: Schéma représentant la collection de données *Ecomics [24]*

Un autre exemple peut être donné avec la **médecine de précision**. Jusqu'à récemment, les données cliniques et les données « omiques » étaient gérées et exploitées de manière indépendante. Ainsi, la prise en charge des patients se faisait exclusivement à l'aide des données cliniques tandis que l'interprétation fonctionnelle des gènes se basait principalement sur des données et connaissances « omiques ». Mais depuis peu, avec le développement de la médecine de précision et des thérapies ciblées, les données génomiques et leur interprétation constituent un élément indispensable pour la prise de décision thérapeutique au même titre que les données cliniques du patient [25]. Parallèlement, être capable de connaître le rôle des gènes dans les maladies ou dans la réponse d'un patient à un médicament est déterminant.

Une solution, pour répondre à ces besoins d'intégration d'information disparates, peut s'orienter vers la proposition de nouvelles mesures de similarité sémantique entre des sources d'information hétérogènes. C'est dans ce contexte, que nous démarrons le projet IBOK (*Integration of Biomedical and Omics Knowledge for a synthetic functional annotation of gene sets*) qui vise à développer de nouvelles approches informatique exploitant plusieurs sources de connaissances. Concrètement, l'objectif est de déterminer l'ensemble restreint des annotations les plus pertinentes permettant d'expliquer la fonction biologique d'un groupe de gènes (Voir la **section Perspectives**).

Objectifs et méthodes pour l'intégration des données

Les objectifs des méthodes développées pour intégrer des données peuvent être résumés comme suit (voir Figure 6 extrait de [26]): (i) la découverte de mécanismes moléculaires, (ii) le groupement d'individus ou échantillons, (iii) la prédiction d'un résultat. La plupart des méthodes bioinformatiques se sont, pour l'instant, davantage focalisées sur les deux premiers aspects [26].

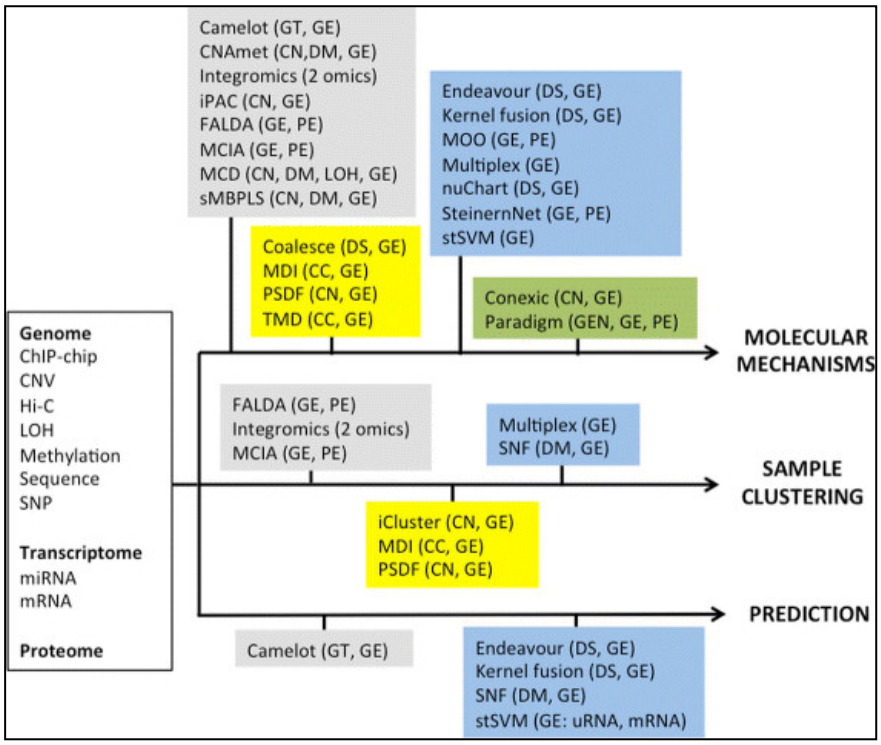


Figure 6 : Illustration du développement de nouvelles approches d'analyse en fonction des 3 objectifs cités par [26].

1.2.5 CONCLUSION

L'analyse des données, et non leur production, est devenue le goulot d'étranglement de la recherche en bioinformatique. Le développement de nouvelles méthodes d'exploration et de visualisation de données, d'algorithmes de recherche plus efficaces, de nouvelles stratégies informatiques d'intégration et d'interprétation d'informations permettant d'exploiter la diversité des données font partie des défis de la recherche en bioinformatique et informatique.

Pour accéder à cette nouvelle complexité du Vivant, il est nécessaire de re-visiter certains algorithmes et d'en concevoir de nouveaux afin de fournir aux biologistes les outils logiciels permettant de réaliser des prédictions, d'explorer les données, d'effectuer des analyses statistiques en grandes dimensions etc.. De part la diversité et la spécificité des données, l'expert biologiste a certainement un rôle encore plus important à jouer devant cet amas de connaissances. Des outils de visualisation adaptés doivent être développés pour rendre l'interprétation des données plus facile et favoriser la génération de nouvelles hypothèses dans le processus d'analyse [27]. La conception de nouvelles approches, permettant de naviguer dans les données afin d'en extraire l'information pertinente, constitue une voie de recherche à laquelle nous nous intéressons pour aider à interpréter cette masse de données.

1.2 A MODELISATION EN BIOINFORMATIQUE

> ***What is Systems Biology ?***
> *The question "What is Systems Biology?" is occasionally discussed, although I find the question "Why model?" more important. I shall thus focus on the arguments for mathematical modeling of cellular systems, beginning with my favorite definition of systems biology: Systems biology is the science that studies how biological function emerges from the interactions between the components of living systems and how these emergent properties enable and constrain the behavior of those components.*
> *Mathematical models are formal representations of natural systems that can help answer questions about the complex system they represent.*
> Extrait de : « Why model? « , Wolkenhauer Olaf, Frontiers in Physiology, 2014.

Face à l'augmentation de la diversité et de la quantité des données de biologie depuis ces dernières années, la modélisation des processus complexes de la cellule connaît également un essor très marqué. Cet axe de recherche est au centre d'une nouvelle discipline qu'est la biologie des systèmes dont l'objectif est de fournir des outils mathématiques et informatiques afin de comprendre les relations entre le génotype et le phénotype. Pour répondre à cette question, il est essentiel d'identifier les différents objets biologiques dans toute leur diversité ainsi que leurs connexions afin d'être en mesure d'appréhender la complexité d'un système biologique. La définition d'un système complexe est une question récurrente pour différentes disciplines, telles que l'informatique ou la biologie. Dans ce cadre, cette section a pour objectif d'introduire les concepts essentiels à prendre en compte pour proposer de nouvelles approches (globales et intégratives) de modélisation en biologie des systèmes.

1.2.1 DE LA CELLULE A UN SYSTEME COMPLEXE

En biologie, l'existence de différents niveaux d'organisation de la cellule justifie aisément la définition d'un système complexe [28]) (voir Figure 7). Les différents constituants moléculaires des cellules (gènes, ARN , protéines, métabolites) composent un premier niveau d'organisation. Viennent ensuite les interactions simples entre ces différents constituants pour définir les briques élémentaires du système. À leur tour, les interactions entre ces différents éléments peuvent être intégrées pour former des modules en charge des fonctions cellulaires. Enfin, ces modules sont imbriqués de façon hiérarchique pour définir l'organisation fonctionnelle à grande échelle de la cellule.

Plus généralement, nous pouvons nous appuyer sur la définition donnée par Annick Lesnes [29] qui s'est attachée à identifier les caractéristiques essentielles des systèmes complexes en se basant sur des travaux issus de la physique et de la biologie. La caractérisation d'un système complexe implique généralement la prise en compte des critères suivants :

- un nombre important de composants élémentaires et hétérogènes (ex : ADN, ARN, protéines, métabolites),
- un nombre important d'interactions de différents types et non linéaires entre ces composants,
- des contraintes biologiques extérieures au système non négligeables.

La simple addition des propriétés de ses composants élémentaires (ARN, protéines, ...) est insuffisante pour étudier leur fonctionnement global. En effet, il est impératif de considérer les propriétés induites par leurs nombreuses interactions ainsi que l'organisation multi-échelle

de la cellule. Le recours à une approche systémique peut être une solution pour appréhender les différents objets biologiques élémentaires dans leur complexité. Elle implique la prise en compte des états individuels, de l'impact des interactions ou rétroactions et des modifications extérieures. Cette approche peut prendre forme dès le processus de modélisation en s'appuyant sur des représentations graphiques où on peut matérialiser des objets avec leurs interactions.

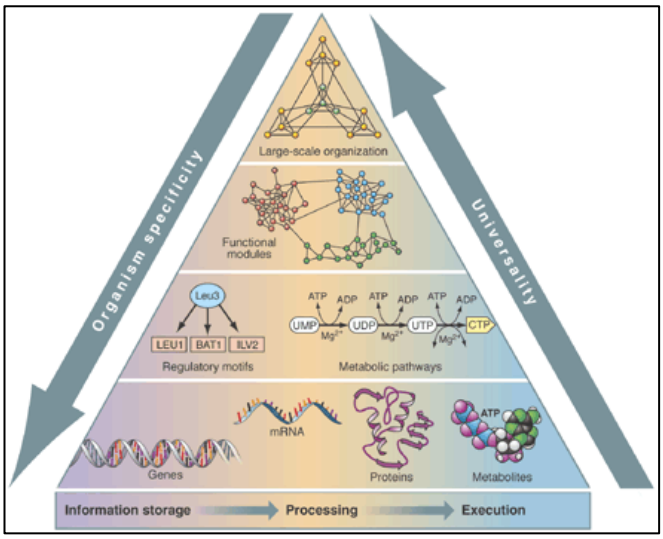


Figure 7 : Représentation des différents niveaux d'organisation d'une cellule d'après [30].

1.2.2 APPORT DE LA THEORIE DES GRAPHES

En bioinformatique, la modélisation de tels systèmes fait assez naturellement et classiquement appel à la théorie des graphes en tant qu'outil de représentation et de modélisation. Les graphes proposent une modélisation intuitive des relations inter-moléculaires du Vivant, tout en fournissant une large boîte à outils de méthodes d'analyse.

Plus formellement, un simple graphe G se définit par la paire **(V,E)**, où :

- **V** est un ensemble fini d'éléments appelés sommets ou noeuds (exemple gène X ou Y)
- **E** est une relation binaire définie sur V (par exemple sous ensemble de VxV) et où les éléments de E sont appelés des arcs ou arêtes.

Dans le cas de réseaux génétiques, les gènes ou leurs produits peuvent se formaliser par les sommets du graphe et leurs interactions (en spécifiant ou non leur nature) sont représentées par des arcs reliant ces sommets. Dans le cadre du métabolisme, un simple graphe peut modéliser le métabolisme (les sommets pour les enzymes et les arêtes pour les métabolites) afin d'en étudier les caractéristiques topologiques. A l'inverse un modèle s'appuyant sur des réseaux de Petri ou des équations différentielles permettra de prendre en compte la dynamique de la cellule.

Plus globalement, l'éventail des modèles utilisés en bioinformatique s'étend des modèles logiques/qualitatifs aux modèles quantitatifs en s'appuyant sur différents cadres formels mathématiques. C'est essentiellement la question biologique et les données que l'on souhaite prendre en compte qui orienteront le choix du modèle.

Le processus nécessaire à la construction du modèle est généralement décrit selon un cycle itératif nécessaire à la définition de l'ensemble des variables nécessaires. Similairement

aux critères requis pour définir un modèle complexe [29], la définition d'un modèle mathématique nécessite les trois ingrédients suivants [31] :

- des variables pour représenter les objets biologiques,
- des relations ou connexions mathématiques entre ces variables / objets biologiques,
- des contraintes pour prendre en compte le contexte de l'analyse.

Comme mentionné précédemment, la complexité du modèle doit se faire au regard des questions biologiques et des données dont on dispose. En d'autres termes, la définition simple des graphes proposée ci-dessus peut être étendue selon ce que l'on souhaite modéliser. Dans ce contexte, Lenovere [32] a identifié trois principales catégories de graphes :

- les graphes dirigés où une direction est imposée à l'interaction en X et Y (par exemple, X régule Y),
- les graphes séquentiels où un chemin peut être défini pour relier les sommets en entrée et sortie (par exemple, X synthétise Y, lequel synthétise Z …),
- un graphe dit mécanistique pour représenter des informations décrivant les différents états de ces sommets (par exemple la cinétique associée à un enzyme).

Ainsi, pour modéliser **un réseau d'interactions**, un simple graphe peut être suffisant pour représenter les interactions symétriques entre des molécules. Il n'est pas nécessaire de prendre en compte d'autres informations sur l'effet de leur interaction (X interagit avec Y et réciproquement, un exemple est donné avec un graphe de co-expression où deux sommets/gènes sont reliés s'ils sont co-exprimés avec r >+0.75).

Modéliser **le flux d'activités** est une tâche plus difficile et nécessite de prendre en compte de manière précise des informations sur la nature de l'interaction (exemple X régule Y) , sans pour autant contraindre la spécification des mécanismes sous-jacents à l'interaction. Les arcs sont dirigés mais la valeur des sommets ne varie pas. Ce dernier point est pris en compte par le niveau suivant.

La description d'un Processus permet de décrire des transferts de masse et de faire varier des valeurs associées aux sommets. Les labels ou valeurs associées aux sommets peuvent dans ce cas évoluer au cours du temps.

En général, le modélisateur choisira le niveau le plus simple permettant l'intégration des données dont il dispose. En effet, ce choix influencera les développements méthodologiques à réaliser par la suite. A chacun de ces modèles correspond un cadre mathématique formel auquel une boîte à outils de méthodes d'analyse peut être reliée. Par exemple, si on souhaite identifier dans un réseau biologique des sous-graphes fortement connectés en adaptant des méthodes de *clustering*, la difficulté des approches à implémenter ne sera pas identique pour un simple graphe ou un réseau de Petri.

1. 2.3 CONCLUSION

L'identification et la compréhension des règles gouvernant un système complexe est un des problèmes auxquels s'intéresse la communauté scientifique en bioinformatique. Elles nécessitent avant tout l'identification des caractéristiques topologiques et biologiques pour, évaluer la robustesse du système (capacité à réagir aux perturbations extérieures), le comparer à un autre système (un autre organisme), étudier sa dynamique en effectuant des simulations etc.. Ces différentes questions biologiques peuvent se traduire formellement par des questions informatiques classiques auxquelles je m'intéresse et qui seront développées dans les chapitres suivants:

- *clustering* de graphes ayant pour objectif de décomposer un système complexe en modules **(*cf.* section 2 sur la modélisation de réseau de régulation)**,
- inférence de motif **(*cf.* section 4 sur les perspectives)**
- recherche de motif (signature topologique) dans un graphe, **(*cf.* section 2 sur la modélisation de réseau de régulation)**
- alignement de graphes **(*cf.* section 4 sur les perspectives)**
- analyse du flux **(cf. section 3 sur le métabolisme)**

Afin de prendre en compte le caractère massif des données que nous avons à traiter et faciliter l'expertise des analyses issues de nos développements, je me suis également intéressée à l'apport des approches de visualisation. Le contexte général de ces approches est introduit dans la section suivante.

1.3 APPORT DE LA VISUALISATION EN BIOINFORMATIQUE

> ***Why do people visualize data?***
> *People visualize data either to consume or produce information relevant to a domain specific problem or interest. Visualization design and evaluation involves a mapping between domain problems or interests and appropriate visual encoding and interaction design choices. This mapping translates a domain-specific situation into abstract visualization tasks, which allows for succinct descriptions of tasks and task sequences in terms of why data is visualized, what dependencies a task might have in terms of input and output, and how the task is supported in terms of visual encoding and interaction design choices. Describing tasks in this way facilitates the comparison and cross-pollination of visualization design choices across application domains; the mapping also applies in reverse, whenever visualization researchers aim to contextualize novel visualization techniques.*
> Extrait de : Matthew Brehmer, PhD Dissertation. "Why Visualization? Task Abstraction for Analysis and Design [29]". April, 2016.

Comme décrit dans la première section de cette introduction, la complexité et la taille des données à modéliser dans les réseaux biologiques sont considérables. Dans la seconde section ont été évoqués les avantages à intégrer diverses sources de données hétérogènes pour améliorer significativement la qualité des modélisations *in silico*.

Une conséquence directe de ces grandes masses de données et de leur intégration est de rendre leur analyse et leur manipulation difficiles pour les biologistes. Le passage à l'échelle nécessité par le caractère massif des données biologiques, est devenu depuis ces dernières années, un problème incontournable en bioinformatique. Pour y répondre, la communauté s'intéresse de plus en plus à proposer des systèmes de visualisation dédiés à l'analyse des données biologiques. En effet, une stratégie classique exploitée par les approches de visualisation repose sur le découpage des données afin de les étudier par itération au moyen de propositions de représentations graphiques pertinentes. Le principal avantage est de tirer partie de capacités d'analyses visuelles de l'expert en le plaçant au cœur du processus d'analyse. Les méthodes développées en visualisation d'information en biologie ont ainsi pour objectifs de faciliter l'extraction de connaissances afin de proposer de nouvelles hypothèses en vue d'une validation expérimentale.
Dans ce contexte, nous proposons de présenter dans cette section les principes et méthodes issues de la communauté scientifique qui s'intéresse aux approches méthodologiques en visualisation.

1.3.1 OBJECTIFS DES APPROCHES EN VISUALISATION

Les approches en visualisation de données ont pour objectifs de proposer des représentations graphiques d'informations à partir de données, relations ou concepts souvent abstraits et à très grande volumétrie. Grâce à la proposition des structures visuelles pertinentes pour un contexte donné, elles ont vocations à faciliter une lecture plus rapide ainsi qu'une meilleure mémorisation de l'information en permettant de densifier/synthétiser l'information au moyen de différentes variables visuelles (objets graphiques, couleurs, formes, position).

En science, le recours à ces approches est fréquent pour **communiquer** des résultats où une représentation graphique sera considérée plus efficace si elle permet d'améliorer la compréhension en exploitant le caractère explicatif. En outre, le choix de la représentation graphique est souvent guidé par la solution la plus adaptée pour mettre en avant une information, augmenter son attractivité tout en prenant en compte de le caractère massif des données.

Un autre objectif de ces approches est également de permettre **l'exploration et l'analyse des informations pour faciliter la découverte de connaissances** en autorisant la confrontation de différents points de vue avec des niveaux de détails à choisir. La comparaison de représentations visuelles dans le cadre d'analyses d'information complexes et multiples est également un atout de ces différentes approches. Nos travaux se positionnant dans ce périmètre, nous nous attachons par la suite à décrire le cadre formel apporté par les approches de visualisation pour la définition et la mise en œuvre de solutions informatiques pour l'exploration et l'analyse de grandes masses de données en biologie.

1.3.2 DEFINIR UN SYSTEME DE VISUALISATION

En pratique, la visualisation peut se décrire comme un processus centré sur les données dont l'état va varier au cours de l'analyse et permettre plusieurs niveaux d'exploration (voir Figure 8).

Au premier niveau, se trouvent les données brutes, dont le format peut être très variable. Elles ont vocation à être transformées par l'application de méthodes bioinformatiques. Par exemple, une analyse classique effectuée à partir de données d'expression issues d'une analyse de type RNAseq consiste à comparer, grâce à une approche de *clustering*, les différents niveaux d'expression des gènes en fonction de différentes conditions expérimentales. L'identification de groupes de gènes ayant une expression corrélée pour plusieurs conditions permet ainsi d'identifier des profils d'expression d'intérêt.

Dans le second niveau, les données transformées sont associées à un ensemble de structures visuelles grâce à une étape de ***mapping* visuel**. L'objectif lors de cette étape est de faire correspondre les éléments (variables et objets) de la table de données à des objets graphiques. Dans le cas de l'exemple cité précédemment, on pourra par exemple associer les gènes à des cercles, les facteurs de transcriptions à des triangles et les relations de co-expression à des arêtes dont l'épaisseur peut varier en fonction de la corrélation d'expression des deux gènes.

Le troisième niveau nécessite la proposition de vues extraites à partir des structures visuelles. Toujours en se basant sur l'exemple, une vue pertinente pour l'analyse des co-expressions pourrait consister à extraire les sous-graphes fortement connectés grâce à l'application d'un algorithme de *clustering* de graphes (par exemple avec un algorithme de détection de communauté tel que Glay [33]). La vue obtenue met ainsi en relief les sous-graphes correspondant à des groupes de gènes fortement co-exprimés pour différentes conditions. Une

autre possibilité pourrait être de demander l'extraction des facteurs de transcription avec leur voisinage pour visualiser les régulateurs dont l'expression est corrélée à des groupes de gènes.

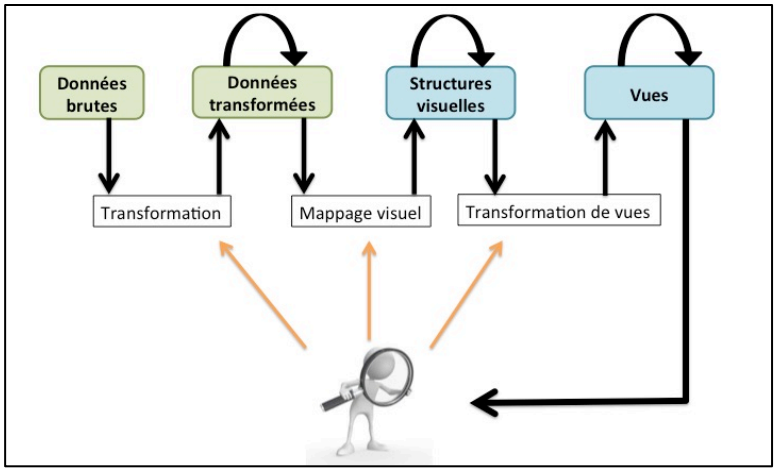


Figure 8 : Processus d'analyse reposant sur l'utilisation d'un système de visualisation. Les flèches noires représentent les flux de données et les flèches orange les interactions de l'utilisateur avec le système.

Ainsi, pour définir un système de visualisation, il est essentiel de regarder trois critères : les données, les tâches que l'on souhaite proposer et bien sûr, l'utilisateur, un biologiste dans notre cas. Dans le cadre de nos contributions (*cf.* Chapitre 2), nous nous sommes appuyés sur le modèle de Munzner (Munzner's nested model) (voir Figure 9) qui définit plusieurs étapes dans son processus de *design* et d'évaluation de processus comme illustré dans l'exemple suivant :

1. le **contexte** : à partir d'une question biologique identifiée comme l'inférence de motifs récurrents dans un réseau biologique,
2. une **étape d'abstraction des données et des tâches** est nécessaire. Les motifs correspondent par exemple à un enchaînement spécifique de sommets/arêtes et leur identification implique de rechercher des groupes de sommets ayant un voisinage similaire,
3. **l'encodage visuel** définit les **algorithmes** à mettre en œuvre. Dans le cas présent, cela revient à combiner un algorithme de recherche de plus court chemin avec un outil d'interaction pour visualiser les sous-graphes calculés.

Cette structuration est exploitée dans le chapitre 2 où nous présentons nos contributions dans rNAV.

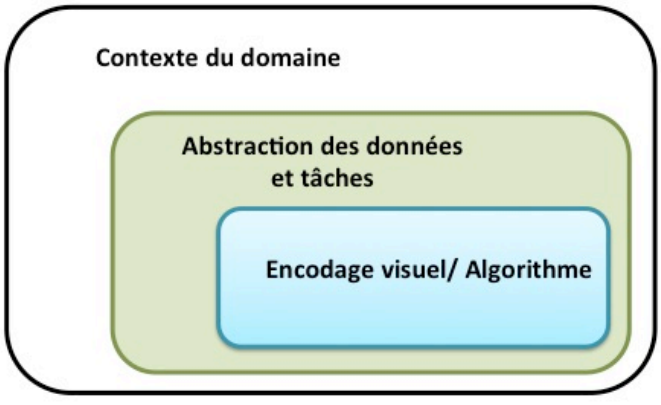


Figure 9 : Modèle de MUNZER pour la conception de système de visualisation [34].

1.3.3 APPROCHES METHODOLOGIQUES EN VISUALISATION : UNE BOITE A OUTILS

En s'appuyant sur nos capacités visuelles innées et un certain apprentissage, les vues synthétiques facilitent et accélèrent la compréhension de grands de jeux de données. Elles permettent d'acquérir plus facilement des informations complexes tout en mettant en œuvre un raisonnement analytique au moyen de vues multiples sur les données.
En conséquence, la proposition d'un système de visualisation pertinent nécessite de prendre en compte différents critères présentés ici et développés par la suite:

- **Premièrement**, **la représentation de l'information** est un critère à prendre en compte pour définir la structure visuelle la plus adaptée au type de données qualitatives ou quantitatives (exemple une représentation sous la forme de réseaux ou hiérarchies pour des données très structurées ou de manière séquentielle pour des données temporelles).
- **Ensuite, le choix des tâches et interactions** que l'on souhaite proposer dans un système de visualisation. **Les techniques d'interaction** (exemple : sélection, filtrage, zoom) permettent de décupler le pouvoir expressif des données et de mieux les appréhender lors du passage à l'échelle. Elles sont essentielles pour permettre à l'utilisateur de changer de point de vue au fur et à mesure de son analyse grâce à l'utilisation de différents encodages visuels de vues.
- **Enfin**, les méthodes de visualisation sont également efficaces pour **articuler la connaissance implicite** afin de stimuler de nouvelles façons de penser. Ainsi différentes méthodes seront plus adaptées pour aider à réduire la complexité des données lors d'une analyse ou, à l'inverse, faciliter l'identification de plusieurs solutions à une question ou un problème.

CHOISIR UNE STRUCTURE VISUELLE ADAPTEE AUX DONNEES

Le nombre de formes visuelles proposées par la communauté de visualisation est important. Une illustration de cette très grande variété et richesse est donnée avec la classification[7] proposée par [35] (voir Figure 10). Cette classification est inspirée de la table périodique des éléments chimiques pour classer les formes visuelles en fonction de leur complexité graphiques (équivalent à la notion de période en chimie), et de leurs applications (équivalent à la notion de groupes dans le tableau).

[7] Une version dynamique et illustrée de la table est donnée à cette URL: http://www.visual-literacy.org/periodic_table/periodic_table.html#

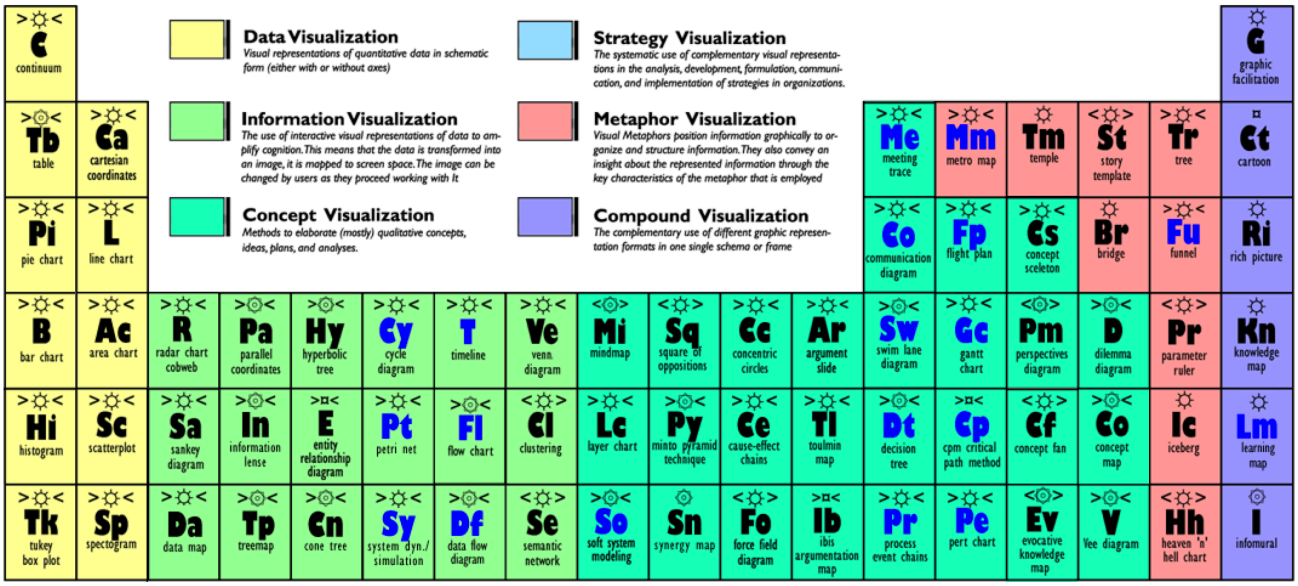


Figure 10: Tableau périodique pour les différentes méthodes de visualisation d'après [35]

Dans nos travaux, nous nous sommes focalisés sur les approches de **visualisation d'information** appartenant à la seconde catégorie de formes visuelles coloriées en vert sur la Figure 10.

Pour illustrer, nous proposons de discuter le choix de structures dans trois domaines d'applications auxquels nous nous intéressons.

1) Visualisation des prédictions de structures secondaires d'ARN :

Plusieurs représentations visuelles sont classiquement utilisées pour visualiser la structure secondaire des ARNs [36], laquelle correspond à une vue simplifiée du repliement réel et fonctionnel de la molécule.
La première image de la Figure 11 montre un diagramme en arcs et permet de visualiser au moyen d'arcs les régions en interaction.
La seconde image de la Figure 11, la plus usuelle, est un graphe où les bases sont symbolisées par les sommets, les arêtes représentant les liaisons covalentes formant la séquence de la molécule. Le dessin du graphe est calculé dans ce type de structure de manière à rapprocher les segments en interaction formant les hélices qui sont ensuite matérialisées par des rectangles colorés.
La troisième image de la Figure 11 est une représentation en montagne qui repose sur un graphe en deux dimensions où l'axe des abscisses correspond aux bases et l'axe des ordonnées au nombre de bases pour joindre les deux bases aux extrémités.
Ces trois structures visuelles sont particulièrement adaptées à la représentation et à la comparaison des structures secondaires définies par le sous ensemble des interactions pouvant être visualisées dans un plan sans croisement des arcs, d'après la première image. Elles sont moins adaptées à la prise en compte des structures ternaires (qui impliquent des croisements entre les arcs dans le plan) et pour les visualiser, il est fréquent d'avoir recours à une visualisation sous forme de dot plot telle que présentée dans **la quatrième image de la** Figure 11**.** Cette dernière structure visuelle a l'avantage également de faciliter la visualisation de plusieurs structures correspondant à l'ensemble des structures sous-optimales secondaires pouvant être prédites par une approche bioinformatique telle que *mfold* [37]. En effet, lorsque

la véritable structure d'un ARN n'a pas été validée expérimentalement, il sera plus pertinent pour un biologiste de visualiser, en plus de la structure présentant l'énergie la plus faible, l'ensemble des structures alternatives. Pour cela, le dot plot donné par la matrice de contact où une case colorée permet de visualiser un appariement potentiel (les bases des deux séquences correspondant pour l'une aux lignes et pour la seconde aux colonnes), autorise la superposition de plusieurs hélices. La couleur peut également être corrélée à la valeur de l'énergie libre correspondant aux appariements.

En 2015, **Biovis**[8] (conférence : *5th Symposium on Biological Data Visualization*) a ainsi proposé un *contest* en *design* pour encourager la proposition de nouvelles structures visuelles adaptées à la représentation de cet ensemble de structures sous-optimales. Nous y avons proposé une structure visuelle (voir Figure 12) basée sur la représentation en graphe de la structure secondaire de plus faible énergie et servant ensuite de squelette à la superposition d'arcs pour montrer les structures alternatives.

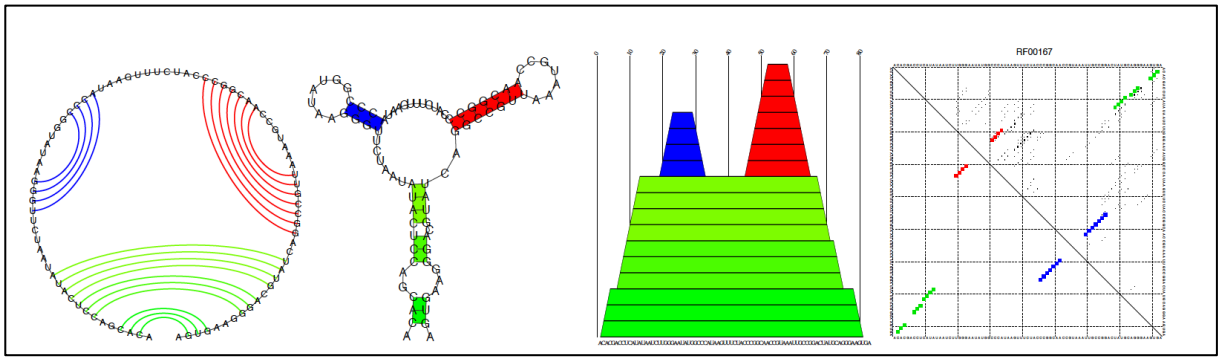


Figure 11 : Différentes structures visuelles pour la représentation de la structure secondaire des ARNs (issues de [36])

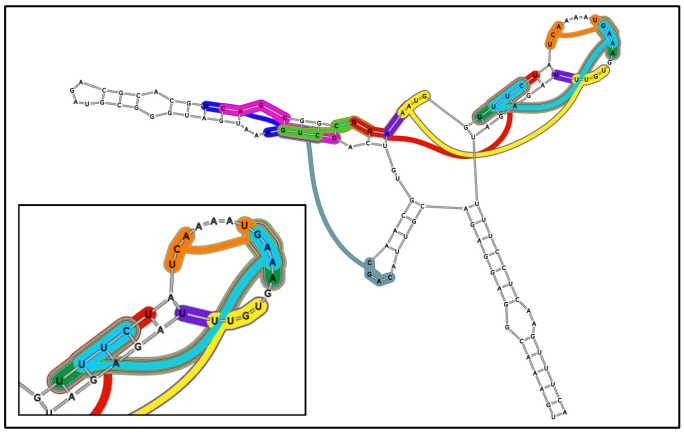


Figure 12 : Représentation de l'ensemble des structures sous-optimales calculées par *mfold* [37]. Le graphe représente la structure secondaire de plus faible énergie et les hélices formées par les structures alternatives sont visualisées au moyen d'enveloppes liant deux régions et selon un code de couleur [38].

2) Les données d'expression de gènes:

Pour la visualisation de résultats de données d'expression, la communauté a souvent recours aux ***heatmaps*** (voir **a** dans Figure 13) également nommées matrices de corrélation [39].

[8] http://biovis.net/year/2015/info/overview-0.html

Les colonnes correspondent aux différentes conditions expérimentales et les lignes aux gènes dont l'expression est étudiée. Il est ainsi facile de rapidement visualiser les blocs de gènes dont l'expression est corrélée pour un sous ensemble de conditions expérimentales, grâce à un encodage de couleur (généralement un dégradé allant du vert au rouge correspond aux valeurs d'expression des gènes). Dans ce type de structure visuelle, l'ordonnancement des lignes (des gènes) est un critère important et est calculé par un algorithme de *clustering* dont l'objectif est de regrouper les gènes en fonction de leur profil d'expression et non de les regrouper en fonction des conditions expérimentales.

Pour prendre en compte les deux aspects, une analyse en composantes principales (ACP) peut être pertinente afin de transformer les données grâce à l'application d'une réduction. Le but de ces méthodes est de réduire l'ensemble des données en identifiant les relations entre les variables (dans notre cas, les conditions expérimentales) et d'y projeter ensuite les individus de l'analyse (ici, les gènes). Ainsi, il devient possible d'identifier (i) les groupes de gènes corrélés mais également (ii) les composantes principales données par l'ACP et correspondant aux combinaisons des différentes conditions expérimentales expliquant la variabilité entre les gènes. Les ***scatter plots*** sont les structures visuelles associées à ce type.

Avec les mêmes objectifs, et sans pré-traitement (tels que le *clustering* ou l'ACP) la représentation **en coordonnées parallèles** (voir **b** dans Figure 13) a l'avantage de proposer à l'utilisateur une exploration interactive de ces données en considérant à la fois les gènes et les conditions. Cette structure visuelle est très adaptée à l'exploration pour identifier de petites différences entre les échantillons. Elle a également l'avantage de pouvoir être interactive pour modifier intuitivement l'ordonnancement des lignes/gènes ou des colonnes/conditions expérimentales. Les coordonnées parallèles sont aussi plus adaptée à l'analyse de données temporelles [39] où on souhaite évaluer les fluctuations des quantités de manière précise. L'utilisation des variations de la couleur des *heatmaps* n'est pas un critère facilement interprétable par l'œil humain pour identifier une variation au cours du temps. Au contraire, l'utilisation de l'encodage spatial par une courbe de niveau permet non seulement une lecture plus précise des valeurs absolues, mais également de mieux comprendre les tendances complexes au moyen d'un graphique. Ainsi, la comparaison de différentes analyses d'expression sera plus aisée à partir de la comparaison de ces structures visuelles.

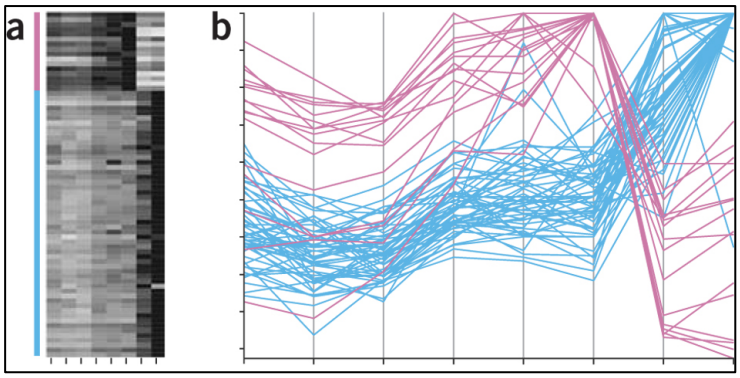


Figure 13 : Structures visuelles pour la représentation des données d'expression : (a) correspond à une *heatmap* et (b) aux coordonnées parallèles, issue de [39].

3) La visualisation des résultats d'enrichissement d'annotation pour un groupe de gènes.

Pour interpréter biologiquement un groupe de gènes, une approche classique en bioinformatique consiste à avoir recours à des méthodes d'enrichissement (pour détail voir la Section *Enrichissement des annotations Fonctionnelles* à la page 61). Elles permettent d'associer les gènes à un groupe de termes d'annotations issues, en général, de la Gene Ontology. Ces données d'annotation des gènes sont des informations particulièrement difficiles à interpréter et le recours à des structures visuelles est très souvent associé aux outils implémentant des approches d'enrichissement. Les termes d'annotation de la Gene Ontology étant hiérarchisés (où un terme ancêtre décrivant une fonction biologique de haut niveau sera relié à des termes enfants décrivant des fonctions plus précises), le graphe est une représentation assez usuelle (voir Figure 14). Il est ainsi par exemple possible d'y ajouter le score statistique associé à chaque annotation (et calculé par la méthode d'enrichissement) grâce à la densité de la coloration et à la taille du sommet. Cependant, ce type de visualisation sera difficilement interprétable si le nombre de termes à représenter est important. De plus, les relations d'hiérarchie sont difficilement visibles et rendent la manipulation d'un tel graphe difficile. Une autre solution, plus adaptée pour les données reliées sémantiquement, est d'avoir recours à des *treemaps* (voir Figure 14) où les termes (rectangles) partageant un ancêtre commun sont regroupés selon un code de couleur spécifique. Dans ce type de structure, il est également possible d'ajouter d'autres informations, telles que le score donné par l'enrichissement, en exploitant la taille des rectangles. Egalement, des interacteurs peuvent être associés à ces structures pour prendre en compte la visualisation de la hiérarchie des termes en fonction de différents points de vue.

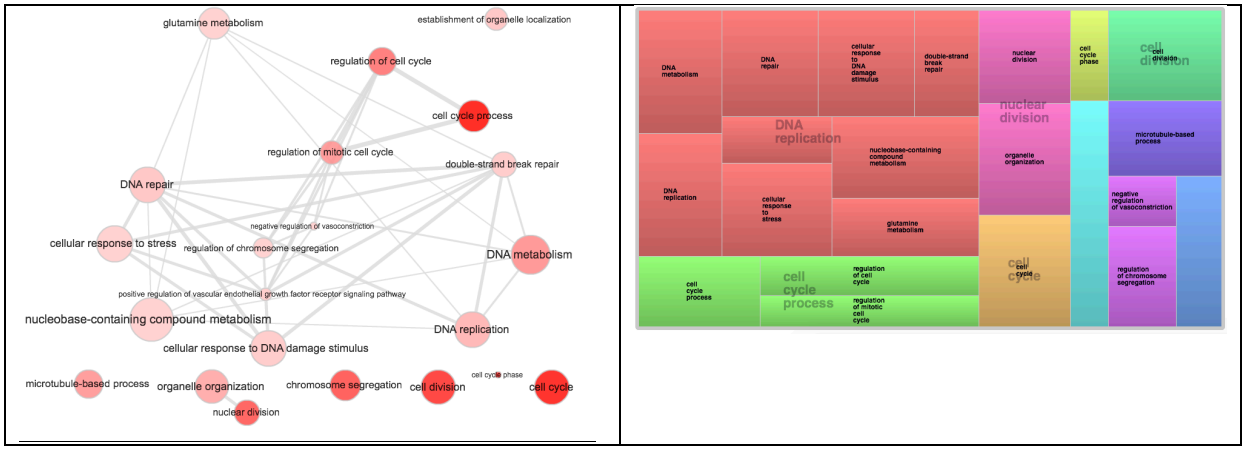


Figure 14 : Graphe et *Treemap* pour visualiser les annotations des gènes sur-exprimés à l'issue d'une analyse d'enrichissement. Les termes descendant d'un même terme ancêtre, indiqué dans un rectangle en gris, sont colorés avec un code couleur identique (illustration générée avec REVIGO [40])

LES TECHNIQUES D'INTERACTION

Dans le cadre de nos travaux, nous nous sommes particulièrement intéressés à ces techniques pour proposer aux biologistes la définition des différentes tâches exploratoires à effectuer pour l'analyse de leurs données. C'est d'ailleurs à ce niveau que la combinaison des approches en visualisation et en bioinformatique est la plus cruciale. En effet, l'usage d'interacteurs présente avant tout l'avantage de filtrer/zoomer ces données et ainsi de mettre en œuvre un raisonnement analytique par itération et décomposition.

Pour proposer de tels outils, nous nous sommes appuyés sur les recommandations ***« Overview first, zoom and filter, then details-on-demand » de Schneiderman***. Ces recommandations sont aujourd'hui reconnues comme une base de la recherche d'information et ont été largement appliquées aux réseaux biologiques, dont la taille et la complexité des données rendent difficiles les représentations manuelles. Aussi, il est important de pouvoir visualiser dans leur globalité les données puis de les explorer en identifiant au fil des analyses des sous-graphes d'intérêt qui seront obtenus grâce à l'application d'algorithmes de bioinformatique pertinents.

Le filtrage dynamique facilite l'interaction avec l'utilisateur. Il exploite ses capacités à percevoir très rapidement au cours de ses actions sur la vue globale, les différences et similitudes des différentes représentations. Les différentes vues globales sont ainsi remises à jour immédiatement (apparition ou disparition des sommets et ou arcs) en fonction des requêtes exprimées au moyen des éléments d'interaction (curseurs ou menus de sélection). Il devient alors possible de détecter les critères affectant les données filtrées et analyser plus en détail certaines régions du graphe global. En appliquant ce type d'analyse avec différents paramètres sur des vues dupliquées du graphe, l'analyse comparative des différents filtres est améliorée grâce à des systèmes qui permettent de suivre la sélection de certains nœuds sur tous les sous-graphes extraits. Un tel environnement est utile pour développer une analyse multi-échelles en suivant précisément certains sommets/ molécules d'intérêt (des exemples sont donnés avec le logiciel Osprey [41]).

Certaines plateformes de visualisation offrent la possibilité d'effectuer plusieurs fois cette opération avec différents paramètres et d'observer les résultats obtenus à partir de plusieurs graphes visualisables dans une même fenêtre. Egalement, elles peuvent permettre d'extraire des sous-graphes sur la base d'une propriété (ex : la topologie) commune.

1.3.4 VISUALISATION DE RESEAUX BIOLOGIQUES

En biologie des systèmes, lorsque les données omiques peuvent être modélisées par des graphes, les méthodes de visualisation fournissent une véritable boîte à outils pour construire des systèmes de visualisation dédiés à l'analyse. Depuis quelques années, le nombre de propositions d'outils a fortement augmenté en réponse aux nouveaux besoins résultant des données massives et de leur intégration. Des preuves de cet état de fait sont observables dans la littérature scientifique. Par exemple, en 2010, le journal *nature methods* (number 3, Volume 7, S1-S28, 90 pages), a consacré un numéro avec cinq revues entièrement dédié à l'illustration de l'étendue des applications et problématiques auxquelles la visualisation de données biologiques peut contribuer.

1) <u>Avantage de la combinaison d'approches de visualisation et de méthodes bioinformatiques.</u>

Dans le cadre d'une revue publiée dans *Briefings in Bioinformatics* en 2015, nous nous sommes intéressés aux avantages liés à la combinaison d'approches issues des deux communautés pour l'analyse de réseaux biologiques.

D'une manière générale, la combinaison de méthodes bioinformatiques avec des approches de visualisation facilite l'exploration intuitive des données selon le processus

analytique illustré dans la **Figure 15**. A partir de données biologiques massives et complexes, la visualisation de l'information produit des représentations visuelles pour aider l'expert à construire des hypothèses et pouvoir ensuite les tester à l'aide d'algorithmes de bioinformatique. Pour accroître le niveau de confiance de l'utilisateur, les algorithmes de bioinformatique peuvent soit réduire la portée de l'étude, soit intégrer d'autres connaissances biologiques. Une telle boucle analytique produit de nouvelles données qui peuvent potentiellement être utilisées pour motiver de nouvelles études expérimentales.

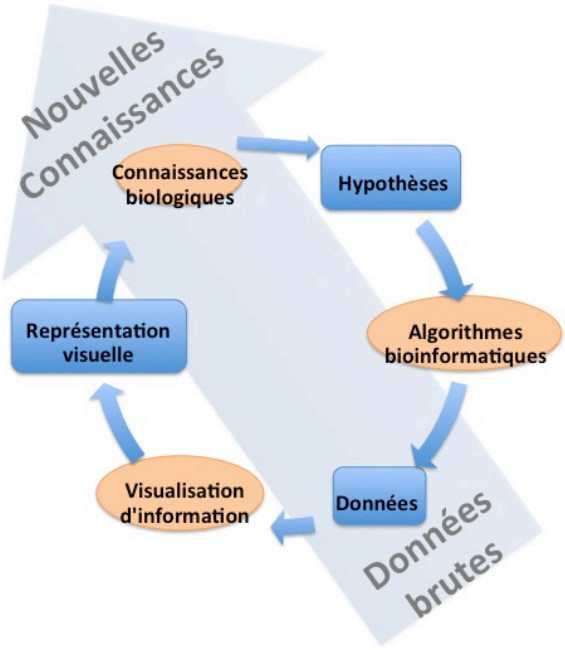


Figure 15 : Le processus itératif d'analyse exploitant la visualisation et la bioinformatique pour inférer de nouvelles informations/connaissances à partir des données. Les ovales orange représentent les domaines scientifiques impliqués et les carrés bleus les résultats intermédiaires de l'analyse. Extrait de Thebault *et al* [42].

2) Illustration : Analyse topologique des réseaux, d'une analyse globale à locale

L'architecture des réseaux biologiques est connue pour être dépendante du type des différentes données *omiques*. En conséquence, les mesures classiques topologiques issues de la théorie des graphe ont été largement étudiées et comparées pour déduire des règles biologiques [43],[28]. Pour mettre en œuvre de telles analyses, des plateformes de visualisation telles que Cytoscape [44] ou Tulip (développé par David Auber dans le thème EVADoMe du LaBRI [45]) proposent une large gamme d'outils de calcul de métriques topologiques (étendue respectivement dans des plugins/Cytoscape ou perspectives/Tulip). D'une manière générale, les algorithmes exploités par ces approches de visualisation vont du dessin de graphe à des algorithmes d'analyse dédiés, en passant par le calcul de mesures topologiques.

Plus généralement, **les stratégies classiques *Bottom-up*** sont basées sur une décomposition itérative du graphe avec pour objectif principal l'identification de motifs ou modules présentant une structure topologique particulière. L'intégration pour chacun de ces modules d'informations supplémentaires (exemple des informations d'annotation ou de localisation génomique si les sommets sont des gènes) aide à sélectionner les modules les plus pertinents au regard de la question biologique posée par l'expert. Par exemple, Modi *et al* [46]

ont construit, à partir de données transcriptomiques, le réseau de régulation centré sur les petits ARNs non codant de *E. coli*. Ils ont proposé une représentation de leurs données sous la forme d'un graphe biparti et ont intégré à cette représentation des informations d'annotation en colorant les gènes selon des catégories fonctionnelles (en prenant en compte la sur représentation de termes d'annotation). Le graphe résultant donne une visualisation globale de la régulation des petits ARNs étudiés. Il devient aisé d'identifier les gènes régulés par plusieurs ARNnc, ainsi que les petits ARNs hyper connectés en charge de la régulation coordonnée à plusieurs conditions de stress.

Les analyses *Down-Top* sont efficaces, quant à elle, pour initier l'exploration à partir d'un sous-graphe d'intérêt, en préambule à l'exploration de son voisinage. Un exemple est donné dans [42] où à partir des gènes régulés par un ARNnc (dont les interactions ont été validées expérimentalement), de nouveaux gènes cibles présentant les mêmes caractéristiques sont proposés. Dans ce cas également, outre la structure du graphe (un gène et un ARNnc sont connectés par un arc si une interaction est prédite), des informations supplémentaires sont rajoutées (annotation des gènes et zone d'interaction entre les deux molécules). Ces stratégies exploitent l'hypothèse selon laquelle une partie non négligeable du graphe est inconnue (dû soit à un manque de données ou la capacité des outils de prédiction) et peut être considérée comme cachée. En se basant sur une sous-partie du graphe bien connue, les caractéristiques topologiques et/ou hétérogènes sont apprises pour être ensuite étendues à l'ensemble du réseau pour le filtrer [47].

Avant de développer de telles stratégies et mettre à profit les avantages apportés par les approches de visualisation, il est essentiel de prendre en compte le nombre important de faux positifs générés par les outils de bioinformatique et de proposer des systèmes d'expertise pour améliorer la qualité des réseaux reconstruits. Une solution évoquée brièvement dans les deux exemples consiste à prendre en compte des données supplémentaires et souvent hétérogènes (pour une revue des différentes stratégies, voir [43]).

1.3.5 CONCLUSION

Nos travaux se positionnent dans ce contexte, où grâce à des collaborations interdisciplinaires, j'ai eu pour objectif de proposer de nouveaux outils et méthodes pour l'analyse de réseaux biologiques. Ce manuscrit se divise en quatre chapitres.

Ce premier chapitre introductif a présenté le contexte scientifique de mes thèmes de recherche à l'interface de la bioinformatique et de la biologie, et en se focalisant sur les enjeux des analyses de données massives et les apports de la visualisation pour l'analyse des données et/ou l'interprétation des résultats d'analyse.

Le second chapitre décrit nos travaux et contributions sur les réseaux de régulation centrés sur les petits ARNs. Dans ce cadre, en nous appuyant sur le modèle de Munzer (voir Figure 9), nous présentons un système de visualisation combinant des approches bioinformatiques et de visualisation.

Le chapitre 3 présente nos développements méthodologiques dédiés à l'analyse des réseaux métaboliques ainsi que le cas d'application qu'ils ont fourni pour le système de visualisation Systryp.

Enfin, **le chapitre 4** évoque les deux axes de perspectives à court et moyen terme, avec respectivement l'objectif de proposer une nouvelle méthode pour l'annotation des groupes de gènes, et celui de combiner les réseaux de régulation aux réseaux métaboliques.

CHAPITRE 2- LES RESEAUX DE REGULATION BACTERIEN CENTRES SUR LES PETITS ARNS NON CODANTS

Depuis 2011, une part importante de notre travail porte sur la construction et analyse de la topologie des réseaux de régulation centrés sur les petits ARNs. Sur cette thématique, sont présentés dans cette section une partie de la thèse d'Amine Ghozlane (co-encadrement avec Isabelle Dutour), du stage de master de recherche d'Amina Bedrat et du contrat (CDD IE, 12 mois) de William Benchimol (co-encadrement avec Romain Bourqui & Isabelle Dutour).

Ces travaux ont été le fruit d'une collaboration interdisciplinaire avec Romain Bourqui, chercheur dans le domaine de la visualisation, et se sont concrétisés par le développement du logiciel rNAV (a,e) dédié à la construction et la représentation des réseaux de régulation. Les représentations graphiques issues de rNAV tirent partie des capacités visuelles de l'expert biologiste en le mettant au cœur du processus d'analyse pour explorer et analyser le graphe par l'application de filtres et d'algorithmes de fouille de données. Ces travaux nous ont permis de valoriser nos contributions en biologie (d), en bioinformatique (a,c) et en visualisation de données (b,e).

En m'appuyant sur mes compétences en Biologie et en bioinformatique et en exploitant mes connaissances sur les systèmes de régulation impliquant les petits ARNs bactériens, mes contributions sur ce thème ont porté sur la définition du *design* de rNAV et le choix des méthodes et outils bioinformatiques pertinents associés aux analyses.

a) Bourqui R, Dutour I, Dubois J, Benchimol W, Thébault P. rNAV 2.0: A visualization tool for bacterial sRNA-mediated regulatory networks mining,. 2017, accepté dans BMC Bioinformatics
b) Joris Sansen, Patricia Thebault, Isabelle Dutour, Romain Bourqui. Visualization of sRNA-mRNA Interaction Predictions. Information Visualisation (IV), 2016 20th International Conference, 342-347.
c) Patricia Thébault, Romain Bourqui, William Benchimol, Christine Gaspin, Pascal Sirand- Pugnet, Raluca Uricaru and Isabelle Dutour. Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks. Briefings in Bioinformatics, in press, 2015. URL
d) William Benchimol, Patricia Thébault, Thomas Bandres, Jonathan Dubois, Isabelle Dutour, Romain Bourqui. rNAV, a new software mixing bioinformatics and visualization approaches for analysing bacterial sRNA-mediated regulatory networks. 6th Bordeaux RNA Club Workshop, June 26-27, 2014.
e) Dubois J, Ghozlane A, Thébault P, Dutour I, Bourqui R. Genome-wide detection of sRNA targets with rNAV. in Proceedings in 3rd IEEE Symposium on Biological Data Visualization, 13-14 October 2013, Atlanta, USA. 2013; 81-88.

2.1 CONTEXTE BIOLOGIQUE –LA REGULATION ET LES ARNS

La régulation différentielle est au cœur de la diversité des phénotypes et de l'adaptabilité du vivant. Elle est essentielle pour moduler les quantités de protéines produites par la cellule, dans le cadre de son fonctionnement normal (mécanisme général de régulation) ou en réponse à des conditions de stress extérieures (mécanisme opportuniste).

Chez les bactéries, au niveau transcriptionnel elle repose « essentiellement » sur les facteurs de transcription (FT). Au niveau post-transcriptionnel, le rôle des petits ARNs non codants (ARNnc) a été mis en évidence plus récemment (pour revue voir [48]). Depuis, l'intérêt porté à ces molécules ainsi qu'à leurs fonctions régulatrices n'a cessé de susciter l'intérêt de la communauté scientifique comme le démontrent de nombreuses publications [49],[50]. Le recours aux petits ARNs non codants pour réguler l'expression des ARNm, permet à la cellule

d'ajuster l'expression de ses gènes dans de nombreux processus biologiques, tels que la modulation de la transcription, la traduction, la stabilité de l'ARNm [51],[52]… Le rôle déterminant des ARNnc dans l'établissement de la virulence a été également décrit chez plusieurs pathogènes bactériens comme *Vibrio cholerae* , *Staphylococcus aureus* [53] ou encore *Listeria monocytogenes* [54],[55] .

Quels sont les avantages au maintien de ces régulateurs?

Une explication du maintien de ce type de régulation au cours de l'évolution pourrait être liée au faible coût énergétique nécessaire à la synthèse d'un ARNnc *versus* une protéine. Initialement suggérée par Gottesman *et al* [56], cette hypothèse a été ensuite démontrée chez des bactéries pathogènes pour la régulation de facteurs de virulence [57],[58].

Plus spécifiquement, la régulation opérée par un ARNnc nécessite qu'il **interagisse** avec les régions 5' UTR des ARNm pour former un complexe ARNnc/ARNm. Cette interaction induit ensuite des changements conformationnels qui empêchent ou favorisent la traduction de l'ARNm en protéine. Par ailleurs, le complexe ARNnc/ARNm peut être dégradé, provoquant une disparition complète de l'ARNm. Contrairement à l'ADN, l'ARN est en général une molécule simple brin, ce qui donne aux bases le constituant la possibilité de s'apparier avec d'autres bases grâce à des liaisons hydrogènes [59]. Une molécule d'ARN a ainsi la capacité de se replier sur elle-même et/ou d'interagir avec d'autres molécules d'ARN. Une interaction entre deux molécules d'ARN va impliquer les bases des deux régions pour former des paires dites canoniques, ou encore, paires Watson-Crick grâce à des liaisons hydrogènes (une adénine avec une uracile ou une guanine avec une cytosine). On peut également trouver dans l'ARN des paires de bases non canoniques, ou non Watson-Crick. Dans ce cas-là, les bases peuvent s'associer entre elles, en fonction des atomes accepteurs d'un côté et donneurs de liaison hydrogène de l'autre. Pour qu'une interaction entre deux régions d'ARN puisse être stable, il est nécessaire qu'elle implique un nombre suffisant d'appariements successifs. Cependant, une interaction peut aussi contenir un nombre limité d'insertions/délétions de bases et/ou de mis-appariements. Ces « erreurs » fragilisent l'interaction en ayant une incidence sur l'énergie libre de cette dernière, et de fait, elles restent limitées dans leur quantité.

La modélisation de l'ensemble de ces régulations peut s'effectuer à partir de réseaux de régulation. Pour construire de tels réseaux, il est nécessaire d'identifier l'ensemble des acteurs potentiels (ARNnc et 5'UTR ARNm) ainsi que leurs interactions. Actuellement, les données expérimentales dont nous disposons sont encore peu nombreuses. A titre d'exemple, **sRNAtarBASE 3.0** contient des informations pour 475 interactions validées à partir de 201 petits ARNs. Parmi ces 475 interactions, 90 impliquent différentes souches de salmonella et 214 de *E. coli*. Cette surreprésentation des entérobactéries est certainement liée au fait que la plupart des travaux actuels impliquent des GRAM– et montre qu'il est nécessaire, dans les années à venir d'étendre nos connaissances en étudiant une plus grande diversité de bactéries. L'identification de ces régulations implique cependant de développer des protocoles expérimentaux onéreux et chronophages et il est souvent nécessaire d'avoir recours à des méthodes de prédiction afin de prioriser les candidats.

Notre travail se positionne dans ce contexte pour proposer de nouvelles approches combinant les méthodes et outils de la bioinformatique et de la visualisation de données pour analyser les réseaux de régulations où les petits ARNs non codants sont impliqués en tant qu'acteurs majeurs de la régulation.

2.2 PREDICTION DE CIBLES DES PETITS ARNS NON CODANTS

2.2.1 LES METHODES DE PREDICTION DE CIBLES

Les logiciels existants (pour revue voir [60],[61],[42]) se classent dans différentes catégories en fonction de la méthode implémentée pour la prédiction des interactions. Une première classe exploite exclusivement la structure primaire des séquences. Elle comprend les logiciels classiques de recherche de similarité (blastn, fasta, ssearch). La recherche d'une interaction entre deux ARNs se fait grâce à l'utilisation de matrices de substitutions modifiées pour prendre en compte les appariements de type Watson & Crick et Wobble (à la place de la similarité). Une seconde classe de logiciels plus sophistiqués exploite les informations de la structure secondaire en se basant sur des approches d'énergie minimum (MFE). Certains logiciels, au départ conçus pour calculer le repliement optimal (en termes d'énergie libre) de la structure secondaire d'un simple ARN (*the RNA folding problem*), ont ainsi été adaptés à la recherche d'interaction(s) entre deux ARNs. Pour cela, certains exploitent la concaténation des deux molécules et recherchent des repliements inter et intra molécules en les mettant en compétition (RNAfold [62] et targetRNA [63]). D'autres s'appuient sur la prédiction de régions accessibles en exploitant au préalable la structure calculée des deux ARNs. Dans ce dernier cas de figure, les interactions entre deux régions non impliquées dans des hélices sont alors favorisées (RNAup [64] et IntaRNA [65]).

2.2.2 BIAIS DES DONNEES D'APPRENTISSAGE

Le choix d'un outil de prédiction n'est pas évident en général et s'appuie souvent sur les analyses comparatives proposées dans les revues. Cependant, ces tests sont très dépendants du jeu de données qui a été utilisé pour les réaliser. Comme évoqué précédemment, 64% des paires (complexe ARNnc/ARNm) validées expérimentalement, sont essentiellement issues de deux bactéries très proches l'une de l'autre sur le plan évolutif (*E. coli* et salmonella). Parmi ces paires, on observe également une surreprésentation de quelques petits ARNs, plus largement étudiés (par exemple, 44 paires pour **GcvB** ou 33 paires pour **RyhB**). Ces données parcellaires et biaisées impactent le développement des approches de prédiction qui sont certainement plus adaptées à ces bactéries.

En effet, ces deux entérobactéries partagent, entre autre, plusieurs caractéristiques telles que leur % moyen en GC ainsi que le recours à une protéine chaperonne HFQ pour la formation des complexes ARNnc/ARNm. Cette protéine, impliquée dans la modulation de la stabilité des ARNm et l'efficacité de leur traduction [66], se lie de façon préférentielle à des régions simple brin riches en A/U et proches d'une tige boucle. Cependant, bien que des homologues de cette protéine aient été identifiés chez plusieurs bactéries avec des variations importantes pour leurs expressions, beaucoup n'en possèdent pas [67]. Ces caractéristiques sont assez spécifiques des entérobactéries et ne seront pas forcément applicables à l'ensemble des bactéries, notamment celles dépourvues de la chaperonne HFQ et avec un %GC moyen faible (exemple : les mycoplasmes).

2.2.3 ANALYSES COMPARATIVES

Pour guider le choix d'un biologiste, plusieurs études sont proposées dans la littérature scientifique (pour un jeu de données spécifique) afin d'évaluer « l'efficacité » des différents logiciels sur la base du nombre de faux positifs (prédiction d'interactions non réelles) et faux négatifs (vraies interactions non prédites). Le meilleur outil est souvent présenté comme celui

permettant d'obtenir le meilleur compromis entre deux valeurs spécificité et sensibilité grâce par exemple à l'analyse d'une courbe ROC (voir première image de **Figure 16**). Un outil sera considéré comme ayant une bonne spécificité et sensibilité s'il permet respectivement de minimiser le nombre de faux positifs et faux négatifs.

Pour aller plus loin, Pain *et al* [61] ont proposé de quantifier le nombre moyen d'expériences à réaliser pour trouver une interaction, à partir de données expérimentales issues exclusivement *de E. coli*. Pour cela, ils ont évalué les performances de plusieurs logiciels en analysant le classement des cibles en fonction de leur score de prédiction. Ils ont ainsi démontré que le rang médian des vraies interactions était inférieur à 5 pour trois logiciels appartenant aux méthodes les plus sophistiquées de prédiction (intègrent des informations de structure des deux molécules d'ARN et/ou des informations sur la conservation des interactions). Pour les autres outils, ce nombre moyen peut atteindre des valeurs très élevées.

Nous nous sommes également intéressés à l'évaluation de ces outils en nous focalisant sur la signification de leur prédiction [42]. Pour cela, nous avons utilisé de vraies séquences mais également des séquences ayant subit au préalable un mélange aléatoire des bases. Notre analyse s'est concentrée sur 3 outils choisis sur la base des meilleurs résultats donnés par une courbe ROC classique (voir première image de **Figure 16**). Deux de ces outils sont également dans la top liste proposée par Pain *et al* [61]. Le troisième outil, ssearch, qui n'a pas été utilisé dans leur test, est une implémentation de l'algorithme de Smith & Waterman [68] et exploite exclusivement la structure primaire de la séquence.

Pour effectuer cette analyse, notre choix d'étude s'est porté sur l'ARNnc **GcvB**, pour lequel plus de 30 cibles ont été validées expérimentalement (chez salmonella et/ou *E. coli* [69]) et dont l'implication importante dans la régulation des gènes de salmonella a été suggérée à partir d'analyses à haut débit du transcriptome [69]. Nous avons comparé les résultats obtenus en utilisant la séquence de **GcvB** ainsi que des séquences contenant le même pourcentage de bases mais dont l'ordre a été modifié aléatoirement. De manière surprenante, la distribution des scores donnés par les trois outils analysés (IntaRNA, RNAup et ssearch) n'est pas significativement différente lorsque la véritable séquence de **GcvB** est utilisée (voir première image de **Figure 16**). Considérant la taille des interactions recherchées ainsi que la taille de l'alphabet de l'ADN, ces observations suggèrent que l'utilisation d'informations exclusivement basées sur la séquence est insuffisante pour discriminer les vraies interactions par rapport au bruit de la prédiction.

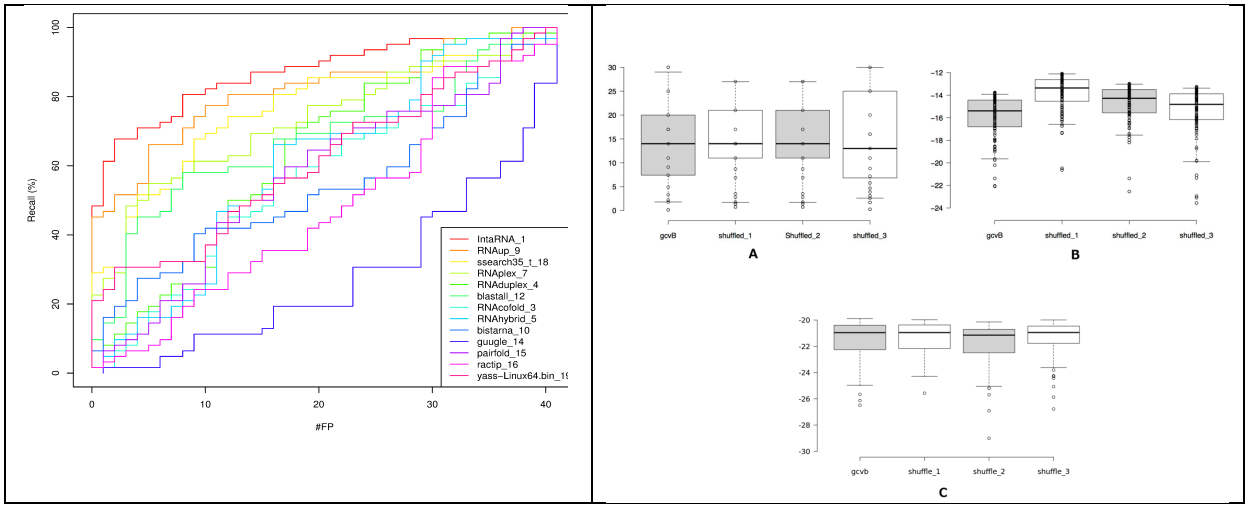


Figure 16 : (1) : Courbes ROC et (2) : distribution des scores de similarité obtenus avec ssearch (A) et les valeurs du calcul de l'énergie données par IntaRNA (B) et RNAup (C). Quatre séquences ARNs ont été

utilisées (GcvB et séquence GcvB dont les bases ont été mélangées au hasard) pour calculer les scores des 100 meilleures prédictions pour chaque logiciel.

En effet, comme énoncé précédemment, l'ensemble des règles biochimiques régissant ces interactions est encore méconnu et rend nécessaire la prise en compte d'autres caractéristiques biologiques. De plus, le nombre de prédictions obtenues à partir des outils peut facilement être problématique pour la mise en œuvre d'une analyse experte et manuelle. Pour pallier à ce problème, une solution souvent choisie implique l'application d'un filtre sur le score en fonction du nombre de prédictions obtenues. En d'autres termes, la spécificité est souvent choisie au détriment de la sensibilité.

Pour améliorer l'efficacité de ces outils, une autre alternative peut consister à se focaliser dans un premier temps sur leur sensibilité, en diminuant le seuil de sélection des scores, puis dans un second temps sur la spécificité grâce à l'intégration de connaissances biologiques hétérogènes. Ce postulat a été le point de départ de notre réflexion pour proposer un nouvel environnement d'analyse dédié à l'analyse de réseau de régulation centré sur les petits ARNs non codants. Dans ce contexte, nous avons développé le logiciel rNAV dont un des objectifs est de fournir un système informatique de visualisation combinant les algorithmes classiques de bioinformatique pour la prédiction de cibles à des approches de visualisation dédiées aux grands graphes.

2.3 DEFINITION D'UN SYSTEME DE VISUALISATION

Le logiciel rNAV [9] a été implémenté sous licence LGPL [70]. Il est une émanation (appelée perspective par la communauté Tulip) de la plateforme Tulip [71] qui nous a fourni un environnement générique de visualisation de données dédié à l'analyse et à la visualisation de des relations entre les données.

Pour concevoir ce système, combinant des approches de visualisation et bioinformatique, nous nous sommes appuyés sur le cadre formel proposé par Menzer *et al* [72] (décrit page 22) qui a l'avantage de centrer sa conception sur le contexte d'étude. Nous avons ainsi procédé de la manière suivante : (i) identifier en premier lieu les questions biologiques pour en déduire ensuite (ii) les données et tâches qui devront être représentées et enfin (iii) les modéliser et identifier les problématiques informatiques liées afin de choisir les algorithmes et « interacteurs » visuels pertinents. Les sections suivantes illustrent quelques fonctionnalités de rNAV en décrivant nos développements selon cette structuration.

2.3.1 MODELISATION D'UN RESEAU DE REGULATION CENTRE SUR LES ARNNC

Questions biologiques identifiées	**Donnée/tâche abstraction**	**Implémentation/Algorithme/Interaction visuelle**
Analyser le réseau dans son intégralité	Proposer une visualisation du réseau	Représenter le réseau par un diagramme nœud-lien

[9] http://rnav.labri.fr

1) Les données

Les données d'entrée du logiciel rNAV sont une liste de petits ARNs et une liste de régions 5'UTR d'ARNm. L'extraction des 5'UTRS peut également être réalisée à partir d'un fichier *genbank* (contient les positions des gènes) en fonction des positions relatives au premier codon du gène. La prédiction d'interactions est ensuite effectuée selon le logiciel choisi, IntaRNA ou ssearch.

2) Implémentation des tâches en bioinformatique et visualisation

Comme illustré dans l'introduction, plusieurs structures visuelles peuvent être considérées pour représenter les réseaux. Parmi celles-ci, les plus intéressantes sont les diagrammes basés sur des matrices (exemple matrice d'adjacence) et les diagrammes basés sur des nœud-liens (graphes avec des sommets et des arcs). Alors que les diagrammes basés sur les matrices sont des modes de représentation efficaces pour l'exploration visuelle de graphes denses en facilitant la perception de densités locales, ceux basés sur les nœud-liens sont plus adaptés à la représentation de réseaux biologiques où les objets étudiés nécessitent de prendre en compte des informations supplémentaires et hétérogènes. Ces informations peuvent alors être associées aux sommets ou arcs en tant qu'attributs, et utilisées pour modifier la représentation de sommets ou arêtes en fonction de critères (exemple : deux types de sommets à différencier).

Dans rNAV, nous avons choisi de modéliser le réseau par un graphe biparti (voir **Figure 17**) avec (i) deux types de sommets, un type pour chaque molécule ARNnc et ARNm et (ii) des arcs correspondants aux interactions prédites entre un ARNnc et un ARNm. Les interactions ou arcs correspondent aux résultats du logiciel de prédiction choisi et contiennent des attributs (annotation fonctionnelle du gène correspondant au 5'UTR de l'ARNm, score de prédiction, région impliquée dans l'interaction …).

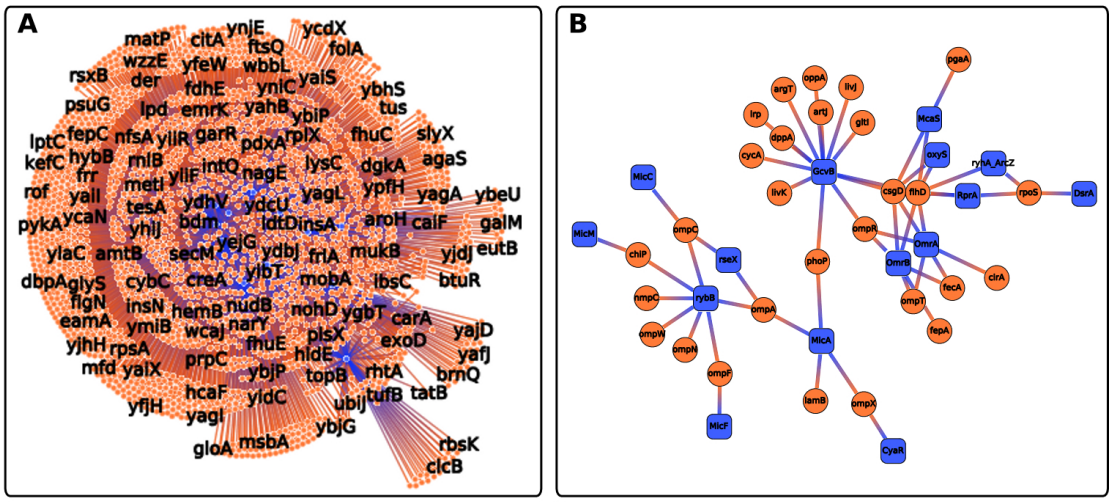


Figure 17: Réseau de régulation centré sur la régulation du biofilm chez *E. coli* (extrait de [42]). Les ovales oranges et carrés bleus représentent respectivement les 5'UTR d'ARNm (ou gènes) et petits ARNS régulateurs.

2.3.2 INTEGRATION DES INFORMATIONS BIOLOGIQUES DES ARNNC REGULATEURS

Questions biologiques identifiées	Donnée/tâche abstraction	Implémentation/Algorithme/Interaction visuelle
Détection de motifs spécifiques de régulation selon Beisel *et al* [73]	Visualiser le voisinage des nœuds : détection de voisinages similaires	Fournir la visualisation du voisinage en exploitant l'outil d'interaction dédié Bring & CO. La solution choisie : algorithme de calcul du plus court chemin puis extraction du sous réseau d'intérêt.
Identifier les régions dont le contenu en bases peut être contraint par l'évolution pour maintenir un grand nombre d'interactions avec différents gènes.	Identifier des nœuds sur la base d'information de leurs arêtes adjacentes.	Fournir des algorithmes/filtres pour identifier les ARNs en fonction de leurs régions d'interaction.

1) Contexte biologique : diversité des types d'appariements

L'intégration d'information topologique et/ou biologique a pour objectif d'aider l'expert à rassembler un faisceau d'arguments pour proposer de nouveaux candidats. Par exemple, la détection de motifs spécifiques de régulation ainsi que la conservation d'une région d'interaction sont autant d'informations à prendre en compte pour l'expertise des cibles. D'après la classification proposée par Beisel *et al* [73] (**Figure 18**), deux motifs nous ont particulièrement intéressés :

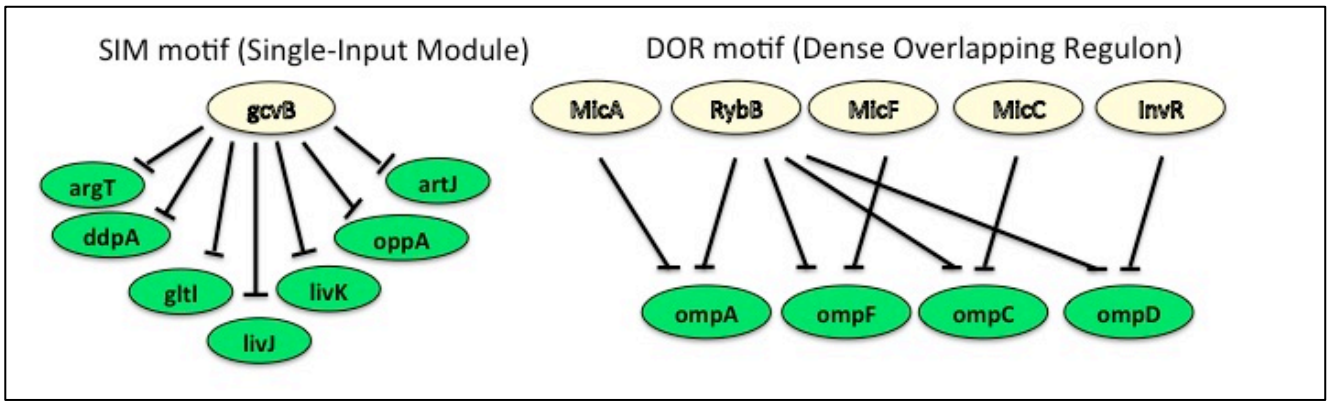


Figure 18 : Motifs de régulation SIM et DOR d'après Beisel *et al* [73].

- **le motif SIM (*Single-Input Module*)** est centré sur un ARNnc interagissant avec un groupe de cibles pouvant être impliquées dans un même processus biologique (ou une voie métabolique commune). *De facto*, l'utilisation singulière d'un unique ARNnc confère à la cellule l'avantage de pouvoir très rapidement s'adapter en modifiant l'expression simultanée de plusieurs gènes.
- **Le motif DOR (*Dense overlapping regulon*)** implique quant à lui plusieurs ARNnc interagissant avec plusieurs groupes de cibles. Ces ARNnc peuvent réguler la même cible et *vice versa*. Grâce à ces différentes combinaisons d'interactions, la cellule a la capacité de coordonner l'expression de plusieurs gènes impliqués dans plusieurs processus biologiques, et cela en réponse à plusieurs *stimuli* extérieurs.

Skippington *et al* [74] ont exploité ce formalisme pour analyser les relations et comportements de ARNnc orthologues chez 27 espèces de *Escherichia coli* ou *shigella*. L'analyse

croisée de l'absence/présence d'orthologues des petits ARNs identifiés expérimentalement (chez salmonella et/ou *E. coli K12*) ainsi que du type de motif de régulation formé, ont servi de base à la proposition de règles régissant les petits ARNs. Par exemple, les petits ARNs formant des motifs SIM ont tendance à être plus conservés parmi les bactéries étudiées. Cette observation est en accord avec l'hypothèse d'une pression évolutive sur certaines régions de l'ARNnc afin de préserver les interactions avec plusieurs ARNm [75],[76].

En d'autres termes, pour un ARNnc seront priorisées les cibles:

- dont les interactions impliquent une région commune du ARNnc,
- qui codent pour des protéines participant à une même fonction biologique.

De même, pour chaque cible, on priorisera celles dont la région impliquée dans l'interaction :

- correspond au site de fixation du RBS,
- et/ou est impliquée dans une interaction avec d'autres ARNnc.

La recherche et analyse combinée de ces informations à partir du graphe initial est une tâche difficile et surtout chronophage. En effet, la taille du graphe peut être très importante : le réseau initial de régulation prédit avec IntaRNA à partir de 15 ARNnc (impliqués dans la formation du biofilm chez *E. coli)* donne **6705 arêtes et 2913 sommets**. Pour faciliter ce type d'exploration, nous avons proposé des outils de manipulation du graphe permettant d'explorer ce dernier à la demande, en fonction du niveau d'analyse souhaité et de l'expertise de l'utilisateur (en mettant par exemple le focus sur un sommet/ARN et son voisinage).

2) Implémentation des tâches en bioinformatique et visualisation

L'intérêt que l'expert porte sur certaines représentations du graphe évolue au cours de son exploration et implique de se focaliser à différents moments sur des conformations précises. Pour cela, nous avons choisi de proposer **un filtrage dynamique** basé sur des techniques de ***fisheye*** **déformant** afin de faire disparaître ou apparaître des sommets selon leur degré d'intérêt (*DOI*). Le DOI se définit selon l'information à représenter et permet de mesurer l'importance des nœuds à visualiser mais également leur distance par rapport à la sélection de l'utilisateur. Ce sont ensuite les opérateurs implémentant ces techniques qui appliquent des transformations aux nœuds pour les rendre visible ou invisible en fonction de l'action de l'utilisateur et en conciliant les détails et le contexte pour une interprétation graphique et esthétique des données.

Deux méthodes sont classiquement utilisées pour faire apparaitre des informations sur le graphe: *« **Link Sliding** »* et *« **Bring et Go** »*. La première, *« Link Sliding »*, fait apparaitre des informations variables (en fonction du DOI) selon le déplacement du curseur le long d'une arête. Pour la seconde technique, *« Bring & Go »*, la déformation de l'image a pour finalité de permettre le rapprochement des nœuds adjacents et de celui qui a été sélectionné [77]. Dans notre cas, à partir d'un nœud sélectionné, il était essentiel de conserver une visualisation sous forme de graphe pour représenter l'ensemble des sommets interagissant avec le sommet sélectionné. En nous inspirant de [78], nous avons donc choisi la méthode « *Bring & GO »* pour développer un outil d'interaction dédié aux ARNs avec deux principales fonctionnalités (**Figure 19**). Pour un ARN d'intérêt (la sélection), (i) le sous-graphe des plus proches voisins est redessiné (effet

zoom) et (ii) une visualisation globale des régions d'interactions pour l'ensemble des cibles reliées est proposée (un petit relié à plusieurs ARNm ou *vice versa*). Cette seconde fonctionnalité est fournie grâce à la superposition du dessin d'un rectangle contenant une courbe représentant le nombre d'interactions impliquant chaque base de l'ARN (la taille du rectangle mime la taille de l'ARN). Ces informations peuvent ensuite être exploitées pour, à la demande, sélectionner une zone d'intérêt et faire apparaître ou disparaître les sommets selon leur interaction avec cette région. Cet interacteur a également été combiné à un algorithme de clustering pour représenter les régions partageant un groupe de cibles (voir B_1 et B_2 de la **Figure 19**).

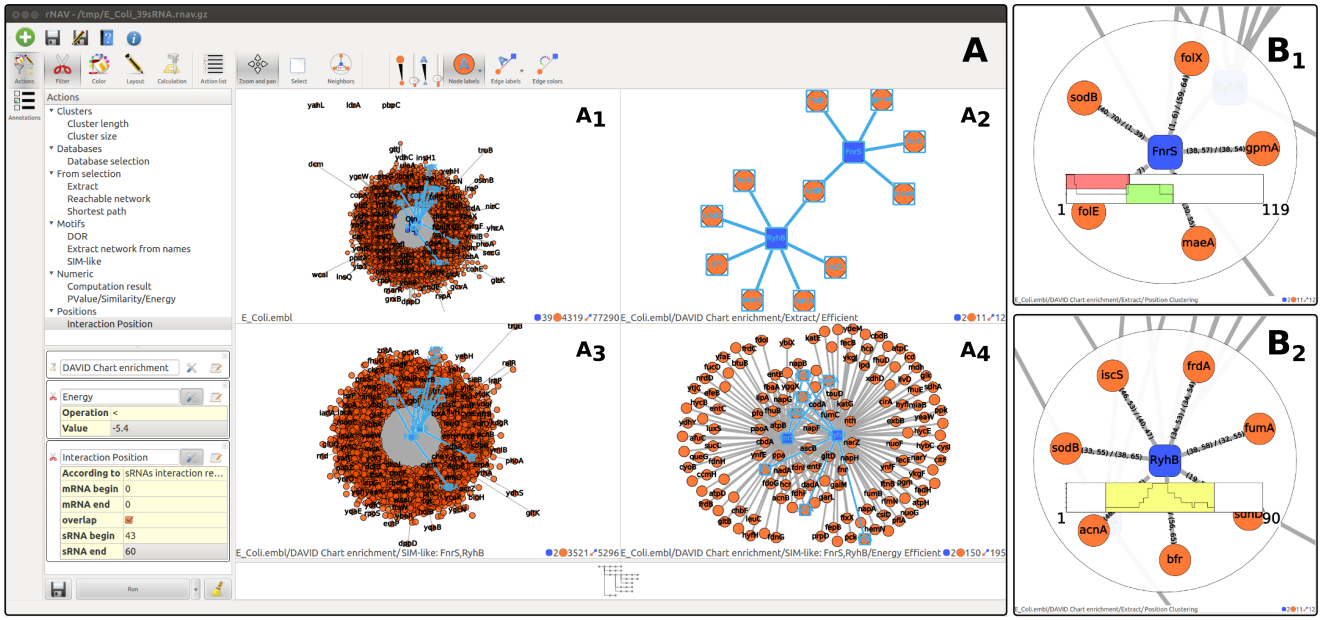


Figure 19 : Stratégie d'analyse dans rNAV pour explorer les sous réseaux de régulation.

2.3.3 L'ANNOTATION DES ARNNC

Questions biologiques identifiées	Donnée/tâche abstraction	Implémentation/Algorithme/Interaction visuelle
Exploiter la connaissance experte du biologiste pour explorer le graphe de régulation (par exemple, le nom des gènes ou d'un processus biologique)	Filtrer le graphe en exploitant les attributs.	Fournir des algorithmes de filtrage/sélection pour réduire le nombre d'éléments à représenter en fonction de leur nom, annotation…

1) Contexte biologique : l'annotation d'un ARNnc

Comme évoqué précédemment, la régulation de plusieurs cibles par un même ARN pourrait être un moyen économique et rapide pour la cellule de modifier son fonctionnement [56]. En effet, un ARNnc peut réguler plusieurs cibles impliquées dans un même processus biologique et permettre ainsi à la cellule de rapidement modifier son fonctionnement en réponse à un stress environnemental [79]. La prise en compte de cette caractéristique, grâce à l'annotation des gènes, est pertinente pour réduire le nombre de candidats et se focaliser sur ceux impliqués dans un même processus biologique. Modi *et al* [79], en se basant sur les termes d'annotation utilisés par les cibles, ont proposé une annotation fonctionnelle des régulateurs.

Par exemple, **RyhB** est impliqué dans des processus de régulation de l'homéostasie au fer, **GcvB** dans le métabolisme des aminos acides, **micF** dans la mobilité cellulaire etc.. Ces annotations ont été déduites à partir de méthodes d'enrichissement d'annotation. La stratégie mise en œuvre par ces méthodes s'appuie sur des approches statistiques pour identifier, au sein du groupe de gènes considéré, les sous-groupes possédant une annotation sur-représentée par rapport à une référence (exemple : le génome complet). Un score statistique est ensuite calculé en fonction de la présence du terme (pour une revue, voir Huang et al. [80]).

2) Implémentation des tâches en bioinformatique et visualisation

En s'inspirant des travaux de Modi *et al* [79], nous avons implémenté dans rNAV des opérateurs pour calculer des enrichissements d'annotation de gènes à partir de la sélection d'un sous-graphe d'intérêt. Cette fonctionnalité est un très bon exemple des avantages résultant de la combinaison de méthodes bioinformatiques à des approches de visualisation. En effet, l'annotation fonctionnelle de groupes de cibles est effectuée grâce à l'utilisation de web-services de DAVID [81]. Notre choix s'est porté sur ce logiciel car une grande majorité des organismes séquencés y sont référencés et il exploite également plusieurs bases de connaissances biologiques (exemple : Gene Ontology, KEGG Pathways, UniProt, Sequence Features). Le calcul d'enrichissement s'effectue en sélectionnant un opérateur à appliquer sur un sous-graphe sélectionné par l'utilisateur au cours de l'exploration du réseau. Ainsi, selon le focus auquel il s'intéresse, plusieurs calculs d'enrichissement peuvent être effectués simultanément. Le score obtenu par DAVID est ensuite exploité comme attribut par rNAV et des interacteurs dynamiques (méthode *fisheye*) sont proposés pour afficher ou rendre invisibles les ARNs en fonction de leur résultat d'annotation.

2.3.4 REPRODUIRE ET SAUVEGARDER LES ANALYSES

Questions biologiques identifiées	**Donnée/tâche abstraction**	**Implémentation/Algorithme/Interaction visuelle**
Reproduire une analyse tout en modifiant des paramètres.	Sauvegarder les enchaînements d'analyses effectuées.	Sauvegarder ou réutiliser des enchainements d'analyses.
Analyser/comparer des sous graphes issus de l'application en parallèle de plusieurs filtres.	Appliquer des enchaînements de filtres/actions sur le même sous réseau afin de détecter des éléments communs.	Fournir une vue à multi-niveaux.
Autoriser de longues analyses et les reproduire sur de nouvelles données.	Autoriser des sessions multiples et garder une sauvegarde des analyses.	Importer/mettre à jour/sauvegarder la vue de l'arbre d'exploration en autorisant les retours arrières pour visualiser un état du réseau intermédiaire dans l'enchaînement des analyses.

1) Contexte biologique : trouver les paramètres les plus pertinents d'un logiciel de prédiction

Classiquement, les analyses de bioinformatique impliquent de développer et répéter plusieurs calculs, en faisant varier certains paramètres afin d'identifier les plus pertinents, sur l'ensemble des données. Ces opérations impliquent de lancer des algorithmes coûteux en temps de calcul et d'analyser un ensemble important de résultats. Pour ces différentes raisons, il est crucial de pouvoir étudier à la volée une partie des données sélectionnées en fonction de critères ou caractéristiques biologiques d'intérêt. Grâce à des systèmes de visualisation dédiés, les

algorithmes bioinformatiques peuvent alors être spécifiquement et rapidement appliqués à cette sélection. Un autre avantage est que l'on peut multiplier le nombre d'analyses à appliquer à cette extraction en vue d'identifier l'analyse la plus pertinente. Il nous a semblé pertinent d'offrir la possibilité à l'utilisateur de mettre en œuvre par lui-même une stratégie d'analyse et d'exploration de ses données et de pouvoir la sauvegarder.

2) Implémentation des tâches en bioinformatique et visualisation

Les différentes opérations effectuées dans rNAV sont appliquées séquentiellement à partir d'un graphe de départ, et correspondent au développement du pipeline d'analyse mis en œuvre par l'utilisateur. Pour visualiser et sauvegarder les résultats issus de ces analyses, rNAV propose un système de visualisation d'arbre d'opérations. Un nouveau nœud est créé lorsque l'utilisateur applique un algorithme de visualisation ou de bioinformatique sur le graphe en cours d'analyse. Le nœud racine correspond au graphe initial, et l'état du graphe résultant de l'application d'un algorithme est ensuite associé au nœud enfant. Lorsque qu'un nouveau pipeline d'analyse est effectué à partir d'un nœud intermédiaire de ce graphe, une nouvelle branche d'analyse est alors visualisée. Cet arbre d'exploration est fourni avec des fonctionnalités de sauvegarde et peut être aussi ré-appliqué à un nouveau graphe (voir **Figure 20**).

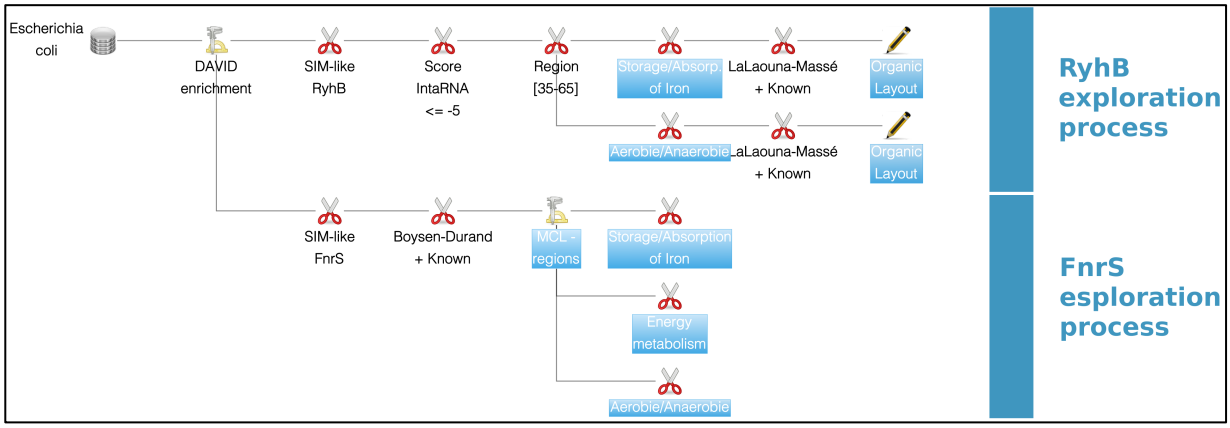


Figure 20 : Arbre d'exploration pour sauvegarder les différentes étapes d'analyse. La sélection d'un nœud de cet arbre affiche le graphe intermédiaire correspondant.

2.4 CAS D'ETUDE

Grâce à rNAV, nous avons pu rapidement mettre en œuvre différentes analyses et réaliser une expertise des données qui semblaient au départ assez complexe :

- chez *Escherichia coli,* nous avons étudié un réseau de régulation centré sur (i) l'ARNnc **GcvB** [70], une quinzaine de petits ARNs impliqués dans la régulation du biofilm de la bactérie [42] et la régulation orchestrée par **FnrS** et **Ryhb** (cas d'étude présenté dans un article récemment accepté dans BMC Bioinformatics).
- Nous avons également réalisé des analyses visant à établir le réseau de régulation de *Mycoplasma pneumoniae* et *Mycoplasma capricolum,* où les ARNnc joueraient un rôle majeur. Ce travail a été soutenu dans le cadre de deux projets Myco-RNA PEPS CNRS/Idex Université de Bordeaux 2013-20154 et Myco-BioBrick AP 2014 de l'Action Thématique « Synthetic Biology Bordeaux (SB2)»)

2.4.1 Exemple d'analyse avec l'ARNnc FnrS d'E. coli

La compilation des données expérimentales disponibles chez *Escherichia coli* et/ou *salmonella* (en s'appuyant sur la base de données sRNAtarBASE [82], EcoCyc [83] et les travaux de Wright *et al* [84]) nous a permis d'identifier **13 cibles connues régulées par FnrS**.

FnrS, est un petit ARN de 122 paires de bases. Il est produit dans des conditions d'anaérobie sous le contrôle des protéines FNR et ArcA. Le nombre de cibles de ce petit ARN est estimé à plus de 30 gènes principalement impliqués dans le métabolisme énergétique [85]. Par exemple, l'implication de **FnrS** dans la régulation d'enzymes intervenant lors du shift de l'état aérobie à anaérobie en condition de stress oxydatif a été démontré dans [86]. L'ensemble des cibles validées par des approches expérimentales a permis d'identifier deux amorces de régions d'interactions couvrant respectivement les bases 3 à 6 et 50 à 60 (expérimentalement démontrées par mutagénèse dirigée). Les deux régions d'interactions correspondantes ont été retrouvées en utilisant l'outil de *clustering* des cibles de rNAV où chaque cluster (déterminant une région d'interaction partagée par un groupe de cibles) est représentée au moyen d'un rectangle coloré (voir Figure 21).

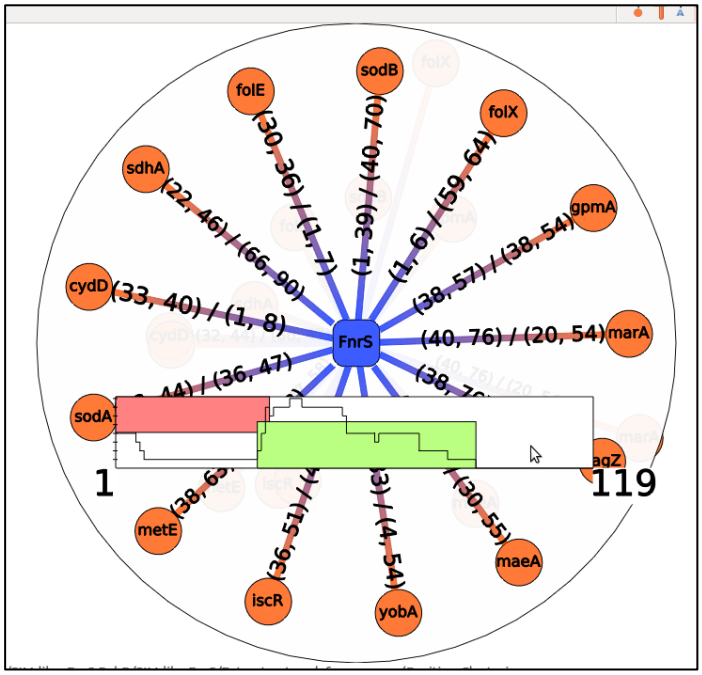


Figure 21 : Représentation visuelle, selon rNAV, du motif SIM de régulation de **FnrS** à partir de données expérimentales. Pour chaque base du petit ARN (le carré bleu correspond ici à un petit ARN et les ovales oranges à des gènes cibles putatifs), une courbe est dessinée (dans un rectangle proportionnel au nombre de bases de l'ARN) pour représenter le nombre d'interaction impliquant chaque base. Un algorithme de clustering a été appliqué sur les régions d'interactions, et deux régions distinctes ont été proposées (elles correspondent aux positions démontrées par mutagénèse dirigée).

Nous avons utilisé des données de RNAseq ([86] et [87]), à partir desquelles nous avons identifié des groupes de gènes présentant un pattern d'expression inversé par rapport à l'expression d'un petit ARN. Entre un ARNnc et un ARNm, pour différencier une relation directe ou d'une relation indirecte, une expérience de RNAseq est insuffisante. En effet, il est nécessaire

de développer un protocole expérimental pour chaque paire (en mutant certaines positions de l'interaction pour identifier une perte d'interaction). Aussi, avec l'objectif de mettre en évidence les meilleurs arguments pour prioriser les cibles, nous nous sommes focalisés sur le sous réseau des cibles validées et celui des cibles hypothétiques (celles pour lesquelles une interaction directe avec le petit ARN n'a pas encore été démontrée). Nous avons ensuite appliqué différents filtres déduits des connaissances actuelles disponibles pour les 13 cibles validées pour prioriser de nouveaux candidats à partir des données de RNAseq. Ces différentes étapes sont représentées sous forme d'un arbre dans la **Figure 20** et la sélection des différents nœuds de l'arbre permet de redessiner automatiquement le sous graphe correspondant.

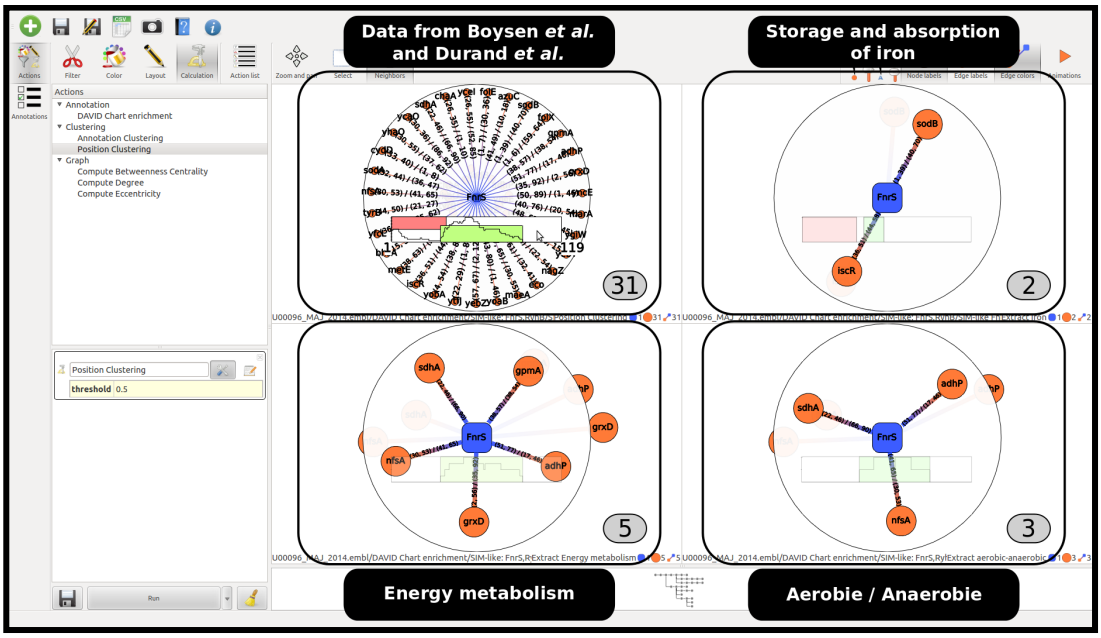


Figure 22 : Stratégie d'analyse avec rNAV pour explorer le sous-réseau SIM-*like* du motif FnrS.

Dans un premier temps, nous avons filtré le réseau de régulation pour extraire le motif SIM-*like* du réseau avec les 32 cibles issues de l'analyse des données de RNAseq. Nous avons ensuite appliqué l'algorithme de clustering MCL pour grouper les cibles sur la base de leurs positions d'interaction impactant **FnrS**. A partir de ce sous-réseau et en exploitant la position d'interaction, nous avons ensuite filtré les cibles présentant un enrichissement d'annotation relié (i) au métabolisme énergétique et (ii) aux processus biologiques spécifiques des états d'aérobie/d'anaérobie. L'analyse des résultats (voir **Figure 22**) nous a ainsi permis d'identifier 3 nouvelles cibles (adhP, grdX et nfsA) sur la base de contraintes biologiques similaires aux cibles déjà connues. Les interactions prédites pour ces 3 cibles, ainsi que pour deux autres cibles validées expérimentalement, sont représentées dans la Figure 23. D'intérêt particulier, nous avons ainsi pu observer des interactions impactant des bases validées grâce à l'interaction avec les deux cibles maeA et gpmA.

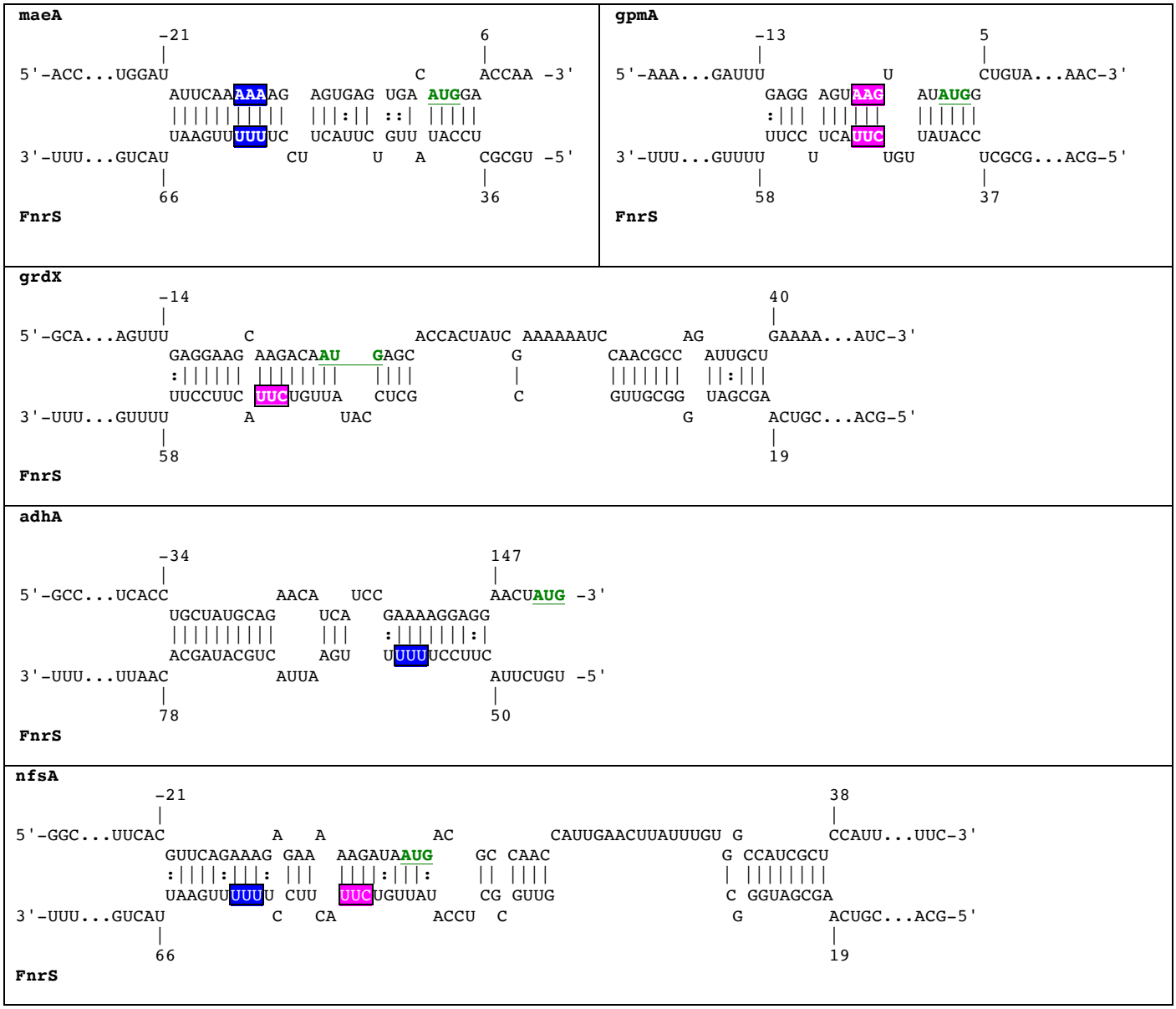


Figure 23 : Visualisation des régions d'interaction. Les deux premières images correspondent à deux cibles où les interactions ont été validées par mutation. Les bases correspondantes sont représentées par une boîte colorée en bleu pour la cible maeA et en rose pour la cible gpmA. Le codon start des ARNm est indiqué par des bases coloriées en vert et soulignées. Les trois images suivantes représentent les interactions prédites par IntaRNA pour les 3 nouvelles cibles proposées et, à titre indicatif, les positions correspondantes aux positions validées pour les deux cibles maeA et gpmA sont représentées.

2.4.2 Exemple d'analyse avec les ARNnc Mcs2 et Mcs4b de Mycoplasma capricolum

Dans le cadre du projet MycoRNA (collaboration avec l'équipe SaAB, INRA de Toulouse et l'équipe Mollicutes, INRA de Bordeaux) nous nous sommes intéressés à développer une stratégie de prédiction des cibles d'ARNnc, spécifiquement adaptée à l'exploitation simultanée des données génomiques disponibles pour un phylum de bactéries pathogènes (voir Figure 24).

Nous sommes partis du modèle de *Mycoplasma capricolum* où l'existence de 2 ARNnc (**Mcs2** et **Mcs4b**) a été confirmée expérimentalement [89] et avons étudié les réseaux de régulation impliquant ces deux ARNnc au sein du groupe des spiroplasmes (sous groupe des mycoplasmes). Une approche de génomique comparative a aussi été mise en oeuvre pour

intégrer les ARNnc orthologues d'autres bactéries appartenant au groupe des mycoplasmes. Ce travail a permis de proposer 9 nouveaux petits ARN putatifs.

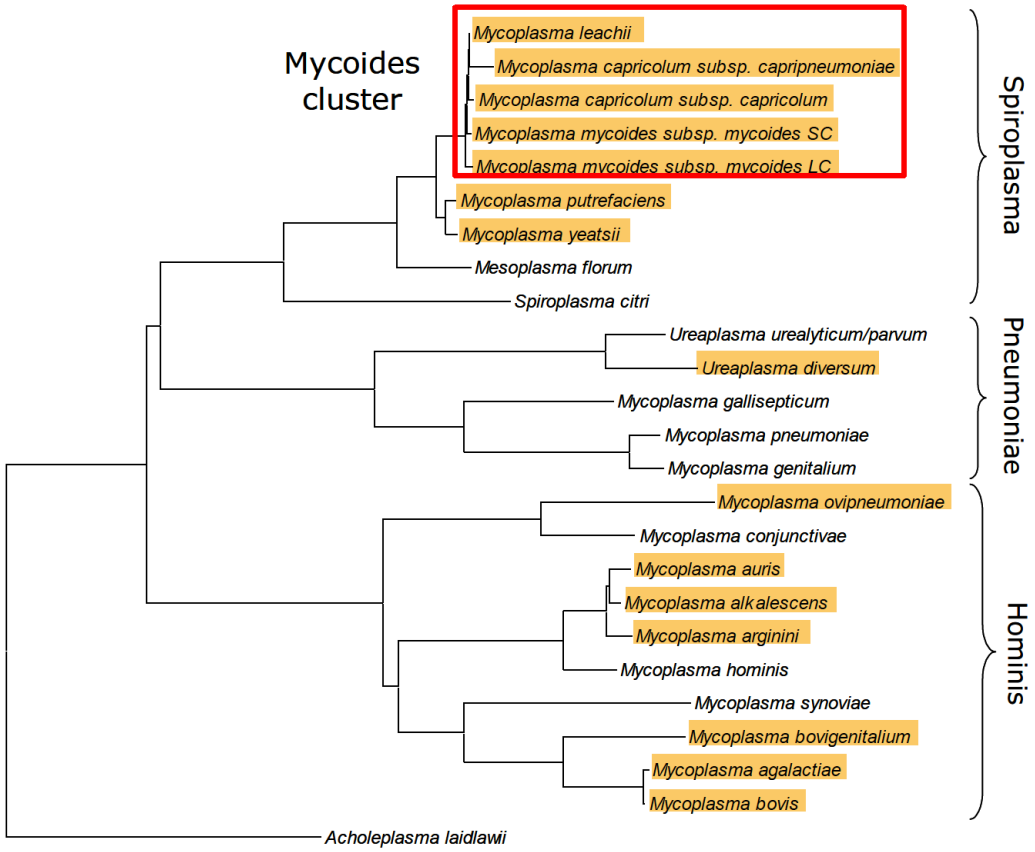


Figure 24 : Phylogénie des mollicutes. Les espèces encadrées en orange correspondent aux bactéries pour lesquelles plusieurs souches ont été séquencées dans le cadre du projet EVOLMYCO.

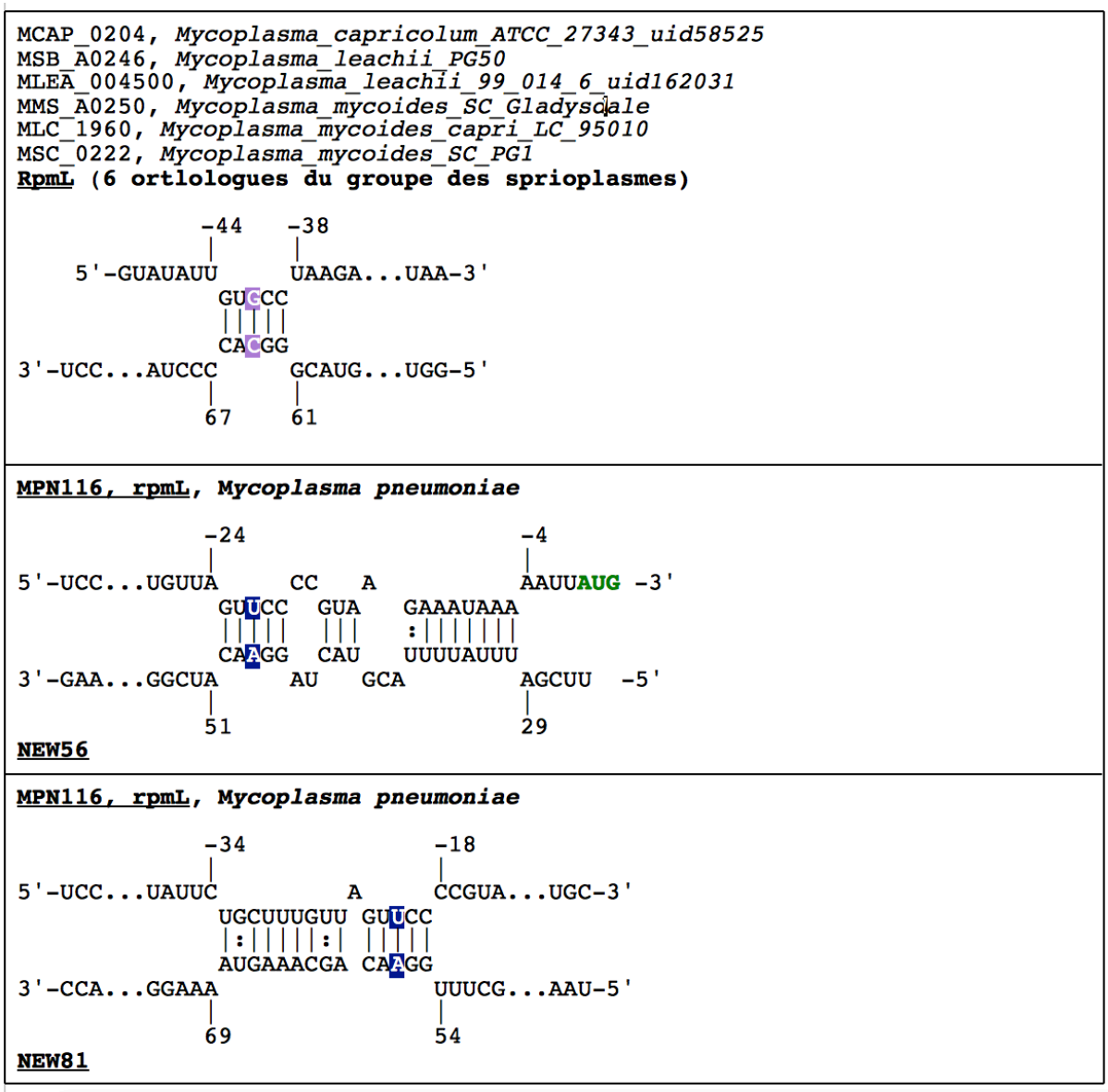
Nous avons alors engagé une recherche de cibles pour l'ensemble des petits ARN et avons également intégré en parallèle les ARNnc proposés par Guell et *al* [88] pour *mycoplasma pneumoniae* (groupe des pneumoniae), pour rechercher de possibles interactions avec les ARNm de cette bactérie. Il est important de noter qu'au moment de notre analyse, les données de *mycoplasma pneumoniae* étaient les seules disponibles pour l'ensemble des mollicutes. La prise en compte d'une espèce relativement éloignée d'un point de vue phylogénétique présente l'intérêt de pouvoir identifier des patterns de régulation conservés entre des mollicutes éloignés. En effet, mettre en lumière la possibilité d'un matintien d'une régulation pour des cibles spécifiques est également une information permettant de prioriser les candidats.

Pour illustrer cette analyse, la Figure 25 synthétise les résultats obtenus pour 3 gènes cibles impliqués dans la traduction et dont le phénotype pourrait être corrélé à des variations de croissance cellulaire. Ce choix a été motivé par [90] démontrant une large représentation de ce type de régulation chez les gamma-protéobactéries.

- **La cible S10, rpsJ** : 3 orthologues de Mcs2, 2 orthologues de Mcs4b ainsi que 5 nouveaux candidats ciblent potentiellement l'ARNm de ce gène en interagissant sur une même région. D'intérêt particulier, 3 orthologues de Mcs2 présents dans 3 souches de mycoplasmes plus éloignées de *Mycoplasma capriculum* n'ont pas donné de résultats de prédiction d'interactions et pourrait résulter d'un non maintien de l'interaction dans

l'évolution des mycoplasmes. En regardant plus spécifiquement les régions d'interaction concernées, on observe une différence de 2 bases entre le groupe des 3 Mcs2 où une interaction avec rpsJ a été prédite et le groupe des 3 Mcs2 où aucune interaction n'a pu être prédite.

- **La cible L35, rpmL**: Les 6 orthologues de Mcs2 étudiés ainsi que les 2 orthologues de Msc4b ont des régions d'interaction prédites dans deux zones différentes. Le pattern de régulation de cette cible est également prédit chez *mycoplasma pneumoniae* grâce à deux ARNnc : NEW56 et NEW81. Une analyse plus approfondie des séquences montre la présence d'une double mutation entre les séquences des ARNm et entre les ARNnc pouvant correspondre à la présence de covariations pour maintenir l'interaction entre le groupe des spiroplasmes et celui des pneumoniae.



- **La cible S13, rpsM** : une première région est ciblée par les 6 orthologues de Mcs2 et comme indiqué sur la Figure 25, la région impliquant les ARNnc est localisée juste en aval d'une région variable entre les 6 orthologues. On peut également observer qu'une seconde région peut impliquer une interaction avec 2 des 9 nouveaux ARNnc candidats.

Ces résultats ont permis de motiver une validation par les biologistes en s'appuyant sur une approche par mutagénèse. Afin de construire des mutants, il a été nécessaire de supprimer les gènes des sRNAs dans le génome de *M. capricolum* cloné dans la levure. Ces génomes modifiés ont ensuite été transplantés dans une cellule receveuse, *M. capricolum*. Ces premiers résultats très encourageants ont motivé de nouvelles expériences (en cours), qui visent à comparer l'expression de protéines par spectrométrie de masse et leur transcriptome séquencé par RNAseq.

2.5 CONCLUSION

Par le développement de rNAV, nous nous sommes intéressés aux avantages apportés par la combinaison d'approches de visualisation avec des méthodes bioinformatiques. Une des qualités principales de ce système est d'offrir à l'utilisateur la possibilité d'explorer ses données à la volée en testant différentes hypothèses. De plus, **l'intégration *a posteriori* de caractéristiques biologiques**, telles que la conservation de la région d'interaction de l'ARNnc pour maintenir l'interaction avec plusieurs cibles, l'implication des multi cibles dans un même processus biologique …), facilite le classement des candidats pour identifier les plus pertinents.

L'intégration *a priori* des caractéristiques biologiques est également une piste que nous commençons à explorer en développant une nouvelle approche de prédiction en collaboration avec Julien Allali (thématique de recherche : algorithmique du texte). Dans ce contexte, le projet **RnaPredict** vise à étendre l'algorithme de Smith & Waterman pour prendre en compte dans son calcul de score l'énergie libre résultante de l'interaction de deux ARNs. Une originalité de l'approche est de ne pas considérer exclusivement la meilleure solution mais l'ensemble des meilleures solutions pour une paire ARNnc/ARNm. En effet, cette étape nous a semblé essentielle pour améliorer la sensibilité. Un algorithme de *clustering* est ensuite appliqué pour regrouper les prédictions de chaque ARNnc (prenant en compte plusieurs solutions pour un ARNm) *versus* tous les ARNm en fonction de la région d'interaction sur l'ARNnc. *In fine*, pour chaque ARNnc seront proposées des cibles regroupées en fonction de leur région d'interaction sur l'ARN. Chacun de ces groupes sera ensuite exploité pour proposer une annotation unifiée de l'ensemble des cibles (voir Perspectives). Grâce à la succession de ces différentes étapes, le choix d'un candidat pour une paire d'ARNnc/ARNm sera guidé par le choix d'une région d'interaction jugée plus pertinente car susceptible d'interagir avec d'autres cibles impliquées dans une même fonction biologique.

Enfin, l'ensemble de ces résultats de prédiction sera proposé aux biologistes dans un environnement de visualisation dédié où les régions d'interaction seront représentées visuellement au regard de la structure secondaire de l'ARNnc. Une exploration interactive sera proposée pour analyser ces résultats en fonction de la région d'interaction et de l'annotation associée au groupe de gènes associés.

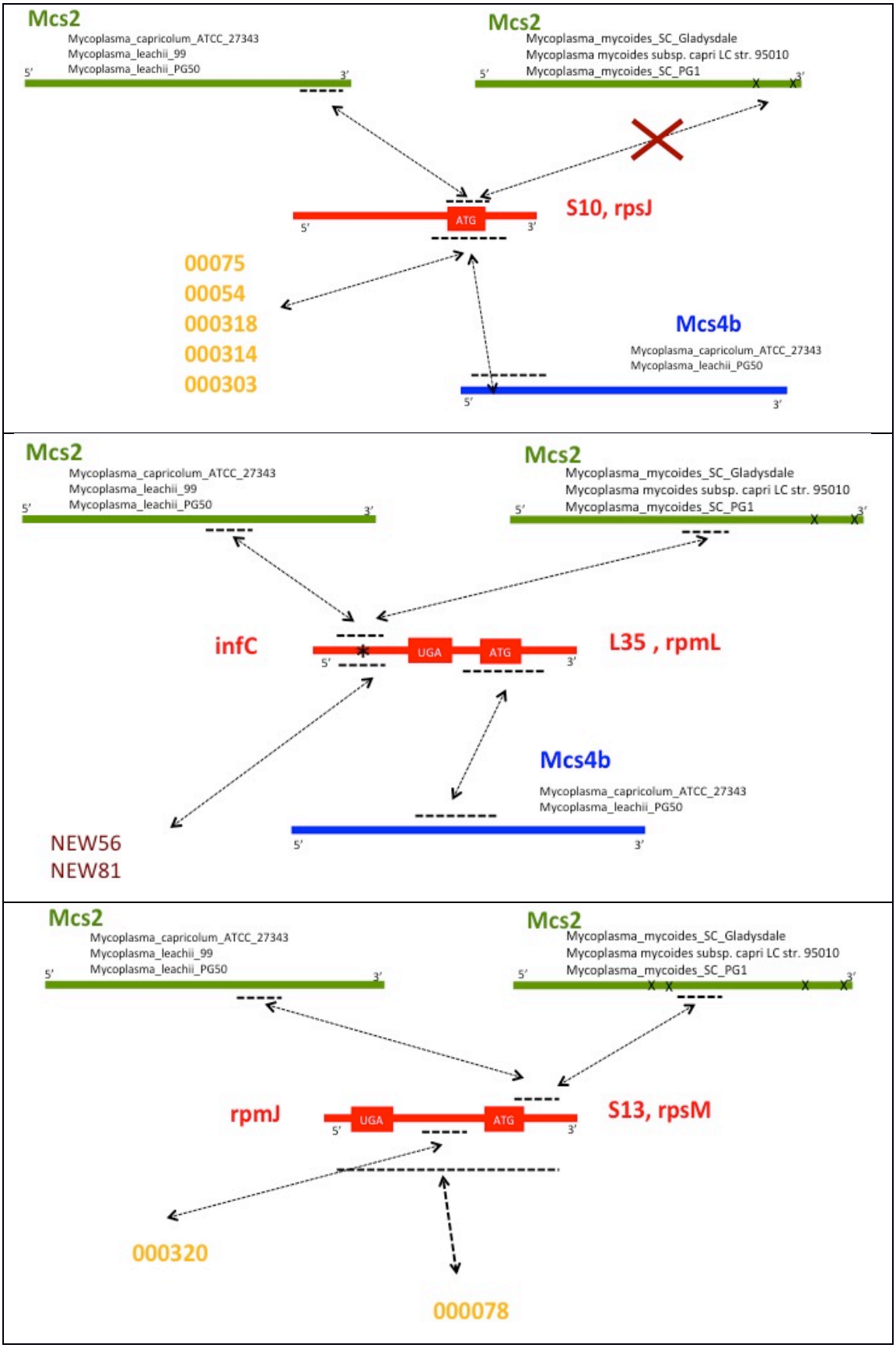


Figure 25 : Synthèse des résultats d'analyse obtenus pour 3 cibles (resp. rpsJ, rpmL et rpsM) impliquées dans la régulation de la traduction. Les 3 séquences sont représentées en rouge, les ARNnc des spiroplasmes en vert. Les nouveaux ARNnc candidats pour les spiroplasmes sont indiqués en orange et les ARNnc identifiés par Guell *et al* [88] chez *mycoplasma pneumoniae* sont nommés NEW56 et NEW81..

CHAPITRE 3- LES RESEAUX METABOLIQUES

Nos travaux présentés dans ce chapitre s'inscrivent dans le champ disciplinaire de la biologie des systèmes avec un intérêt particulier pour l'analyse de la dynamique des interactions entre les objets biologiques étudiés, à mettre en perspective avec l'intégration de données omiques hétérogènes. Sur cette thématique, sont décrits les travaux d'Amine Ghozlane (co-encadrement avec Isabelle Dutour) réalisés dans le cadre de son stage de master de recherche et de sa thèse, en collaboration avec Frédéric Bringaud (biologiste expert du métabolisme des trypanosomes). Plus spécifiquement, mes contributions ont porté sur la définition et le choix des méthodes dédiées à l'analyse du métabolisme d'un protozoaire.

Une approche complémentaire à l'analyse topologique (illustrée dans le chapitre précédent) consiste à analyser la dynamique des réseaux biologiques grâce à l'intégration de données (semi)-quantitatives. La modélisation abstraite sous forme de graphe peut représenter un système biologique par un réseau contenant des informations hétérogènes mécanistiques. Le niveau local sert à définir les règles propres du système, et le niveau global permet de décrire le comportement de ce système dans un environnement intégré. Pour améliorer la qualité prédictive de ces modèles simulés, il est également essentiel de prendre en compte l'ensemble des données (qualitatives et/ou semi-quantitatives) expérimentales disponibles. Cependant, leurs combinaisons peuvent engendrer plusieurs objectifs conflictuels à satisfaire. Pour ces raisons, nous avons proposé une nouvelle approche basée « flux », combinant un modèle simplifié de réseaux de Petri stochastiques à une heuristique d'optimisation combinatoire (d) d'une fonction multi-objective intégrant de multiples contraintes biologiques hétérogènes. Cette étude a fourni un cas d'application qui a permis de valoriser le logiciel de visualisation Systryp développé par Jonathan Dubois et Romain Bourqui (b) et dédié à l'analyse des séries de données temporelles dans le contexte des réseaux métaboliques.

a) Patricia Thebault. A multi-objective heuristic integrating biological data for in metabolic networks, Open Source Software for Systems, Pathways, Interactions and Networks (SPIN-OSS) retreat being held at the Wellcome Trust Genome Campus, EBI, England, 14-16 November 2012.

b) Jonathan Dubois, Ludovic Cottret, Patricia Thebault, Frédéric Bringaud, Amine Ghozlane, Fabien Jourdan, David Auber and Romain Bourqui. *Systryp: a visual environment for the investigation of time-series data in the context of metabolic networks.* IV 2012, 16th IEEE international conference on Information Visualisation. (2012) 8-13th July 2012, Montpellier.

c) Vincent Lacroix, Ludovic Cottret, Patricia Thébault, Marie-France Sagot: An Introduction to Metabolic Networks and Their Structural Analysis. IEEE/ACM Trans. Comput. Biology Bioinform. 5(4): 594-617 (2008)

d) Amine Ghozlane, Frédéric Bringaud, Hayssam Souedan, Isabelle Dutour1, Fabien Jourdan and Patricia Thébault. Flux analysis of the Trypanosoma brucei glycolysis based on a multi-objective criteria bioinformatic approach. Advances in Bioinformatics, 2012, :159423. doi: 10.1155/2012/159423.

e) Amine Ghozlane, Frédéric Bringaud, Fabien Jourdan and PatriciaThébault. Metaboflux : a method to analyse flux distributions in metabolic networks. (2010). Metabolomics conference 2010 . Amsterdam june 27 - july 1

3.1 CONTEXTE BIOLOGIQUE – LA MODELISATION DU METABOLISME

3.1.1 OBJECTIFS

Le métabolisme d'une cellule détermine ses propriétés biochimiques et physiologiques. Il se décrit par l'ensemble des processus métaboliques nécessaires à sa mise en œuvre, lesquels sont le résultat des différentes interactions entre les éléments qui le composent, tels que les enzymes et métabolites. Pour le modéliser et étudier son fonctionnement, on a classiquement recours à la représentation abstraite et simplifiée d'un réseau métabolique. Pour le construire, la

connaissance du génome est une des premières étapes nécessaires à l'identification des enzymes. Des données de transcriptomique et de protéomique peuvent ensuite enrichir ce modèle en apportant des informations supplémentaires sur l'activité de ces enzymes ou réactions biochimiques lors de la transformation des métabolites. Les données de fluxome ou métabolomique nous renseignent quant à elles sur les concentrations de ces métabolites en tant que produits finaux et/ou intermédiaires. La combinaison et l'intégration de l'ensemble de ces données sont essentielles pour améliorer la modélisation du métabolisme et en appréhender le fonctionnement de manière plus réaliste.

Comme évoqué dans l'introduction (cf. section 1.2 page 18), le choix du modèle doit être effectué au regard des questions biologiques le motivant et des données dont on dispose. Bien que **la complexité des réseaux** implique de prendre en compte leur structure, l'étude de **leur dynamique** nécessite de disposer de données quantitatives ou semi-quantitatives telles que la concentration des métabolites ou la quantité d'enzymes produite par la cellule. Entre autres, le développement d'un modèle métabolique dynamique peut aider à identifier le fonctionnement métabolique optimal d'un organisme pour la production de métabolites. Egalement, l'analyse de **la distribution des flux métaboliques** fournit des informations pertinentes afin d'identifier les voies empruntées en fonction de différentes conditions externes. En se basant sur cette connaissance, il est alors possible de proposer une optimisation de la production de métabolites d'intérêt, de comprendre la robustesse/plasticité du métabolisme, d'analyser le caractère essentiel de métabolites ou le rôle déterminant de certaines réactions biochimiques pour un phénotype donné…

Un flux est défini ici par le nombre de molécules, par unité de temps, traversant chaque réaction enzymatique. Pour un réseau métabolique modélisé sous forme de graphe, chaque chemin correspond à un flux possible véhiculant des molécules anabolisées ou catabolisées par les réactions qu'elles traversent. Ainsi, la production d'un métabolite donné résulte d'un flux, de même le niveau de croissance cellulaire mesurable pour une biomasse donnée est un flux. La combinatoire des différents flux pose des défis d'analyse pour la bioinformatique et la prise en compte d'informations supplémentaires telles que le phénotype ou la production de certains métabolites devient cruciale.

Nos travaux se positionnent dans ce cadre et ont eu pour objectifs de proposer un nouveau modèle intégrant des données semi-quantitatives pour prédire la répartition optimale des flux dans un réseau métabolique.

3.1.2 Analyse de Flux

Pour étudier la distribution des flux d'un réseau métabolique, une approche classique en bioinformatique s'appuie sur une modélisation exploitant la définition de contraintes. Par exemple, le FBA (Flux Balance Analyses) [89][90] est une méthode qui cherche à minimiser l'écart entre les données expérimentales et les données calculées en faisant varier les valeurs de flux. Les valeurs calculées sont comparées aux données expérimentales, généralement en exploitant une approche des moindres carrés. En fonction de la différence obtenue, une variation systématique est imposée aux flux non contraints grâce à l'application d'un algorithme de minimisation dont la nature (simplex, stratégie évolutionnaire, recuit simulé, etc.) dépend du logiciel utilisé.

Plus formellement, le cadre mathématique sur lequel repose le FBA s'appuie sur des matrices stœchiométriques décrivant le réseau de réactions. La répartition des flux est ensuite calculée à partir de méthodes de programmation linéaire basées sur les contraintes. Ces

contraintes permettent d'injecter des informations sous la forme d'une fonction objectif (par exemple la croissance de la cellule ou la production d'ATP du système) afin d'optimiser la prédiction des flux dans l'espace solution (décrivant tous les comportements possibles des flux).

Bien qu'il soit communément admis d'utiliser la maximisation de la production de la biomasse pour les bactéries modélisée par une simple fonction objectif, il est nécessaire pour définir l'état de stabilité d'un système modélisant une cellule eucaryote de prendre en compte la combinaison de plusieurs objectifs (par exemple, l'équilibre entre différentes concentrations de métabolites ou encore de maintenir ces concentrations). Plusieurs travaux théoriques ont été proposés pour combiner plusieurs fonctions objectifs qu'elles soient conflictuelles ou non. Par exemple, [91] et, [92] ont étendu le cadre du FBA pour calculer la répartition des flux en contraignant un ou plusieurs flux optimaux. La première étape de ces approches consiste à fixer les bornes de l'espace des solutions. La fonction d'optimisation est ensuite définie comme une distance multiparamétrique qui peut traiter des données de différents types (métabolite de la biomasse, flux , etc.). La distance est ensuite optimisée par des méthodes non linéaires ou linéaires, adaptées à des données multidimensionnelles. La résolution du problème revient à trouver une configuration du système aussi proche que possible de l'état optimal pour lequel il n'existe pas d'alternative. En d'autres termes, l'amélioration d'un objectif n'est plus possible sans produire un impact négatif sur un autre objectif (défini comme une solution optimale de Pareto). Plusieurs analyses ont été effectuées en combinant ainsi de 2 à 4 objectifs, illustrant également l'avantage issu de l'addition des données expérimentales pour améliorer la réalité prédictive du modèle [92]. Outre le grand intérêt de ces articles, ils n'ont pas donné lieu à l'implémentation de logiciels bioinformatiques disponibles publiquement.

Dans ce contexte, nous avons choisi de proposer une heuristique pour le calcul de la répartition de flux optimale en cherchant à satisfaire les contraintes identifiées à partir des données expérimentales disponibles au moyen d'une fonction multi-objectifs. Plus précisément, notre objectif a consisté à développer une approche pour prédire l'ensemble des paramètres du modèle pour lesquels le comportement simulé du réseau s'approche le plus de la réalité mimée par les données semi-quantitatives disponibles.

3.2 DEVELOPPEMENT DE METABOFLUX

La méthodologie choisie a été implémentée dans le logiciel **Metaboflux**. Cette approche combine un simulateur de réseau métabolique et une méta-heuristique probabiliste pour optimiser plusieurs fonctions objectifs prenant en compte différentes contraintes biologiques.

3.2.1 DEFINITION DES FLUX PETRI NET

Le formalisme des réseaux de Petri (PN) a été le point de départ de nos travaux. Il a l'avantage d'offrir un cadre mathématique pour modéliser et simuler les réseaux biologiques [93] et a été largement exploité pour décrire la dynamique des réseaux biologiques (pour revue voir [94]. Il s'appuie sur un graphe biparti dirigé, composé de deux types de sommets (nommés Places et Transitions), et permet de décrire la dynamique au travers d'une suite d'événements discrets. Il se définit dans le cas des réseaux métaboliques par le triplet **{P,T,A}** :

- Un ensemble fini de *places* $\mathbf{P = \{P_1, P_2, \ldots, P_n\}}$: métabolites
- Un ensemble fini de *transitions* $\mathbf{T = \{T_1, T_2, \ldots, T_n\}}$: réactions enzymatiques

- Un ensemble fini d'arcs orientés et unidirectionnels **A = {A_1, A_2, ..., A_n}** reliant exclusivement une transition à une place.

Chaque place contient un ensemble de *tokens* qui vont représenter symboliquement la quantité du métabolite correspondant. A un instant donné, le nombre de *tokens* contenus dans chaque place d'un **PN** définit le marquage du PN et symbolise son état courant. La dynamique est simulée par les déplacements de *tokens* qui passent d'une place à une autre (si le PN l'autorise). Le franchissement d'une transition validée consiste à enlever un *token* de toutes les places d'entrée de la transition et à en déposer un dans toutes les places de sortie de cette même transition. Les **SPN** (stochastiques PN) ajoutent de l'indéterminisme en associant une probabilité au déclenchement des transitions. Ils ont été étendus aux **GPSN** (Generalized Stochastic Petri Nets) par Marsan *et al* [95] pour prendre en compte des transitions qui se font immédiatement. Dans les GSPN, il devient possible de définir des transitions immédiates grâce au "poids" qui leur est associé et, lorsque plusieurs transitions immédiates sont possibles, ces poids sont utilisés pour déterminer quelle transition doit être tirée.

Combinant le cadre donné par le **FBA** et les **GSPN**, nous avons étendu le formalisme pour prendre en compte le poids des flux modélisé par des probabilités en définissant les FPN. Un FPN (Flux Petri Net) est un PN stochastique auquel est associé un poids modélisant les flux au niveau des transitions. Pendant la simulation, lorsque plusieurs transitions peuvent être franchies, ces poids servent à calculer la distribution des probabilités de toutes les transitions possibles. En d'autres termes, plus la valeur de la probabilité est élevée, plus le nombre de métabolites traversant la réaction/transition est élevé (voir **Figure 26**).

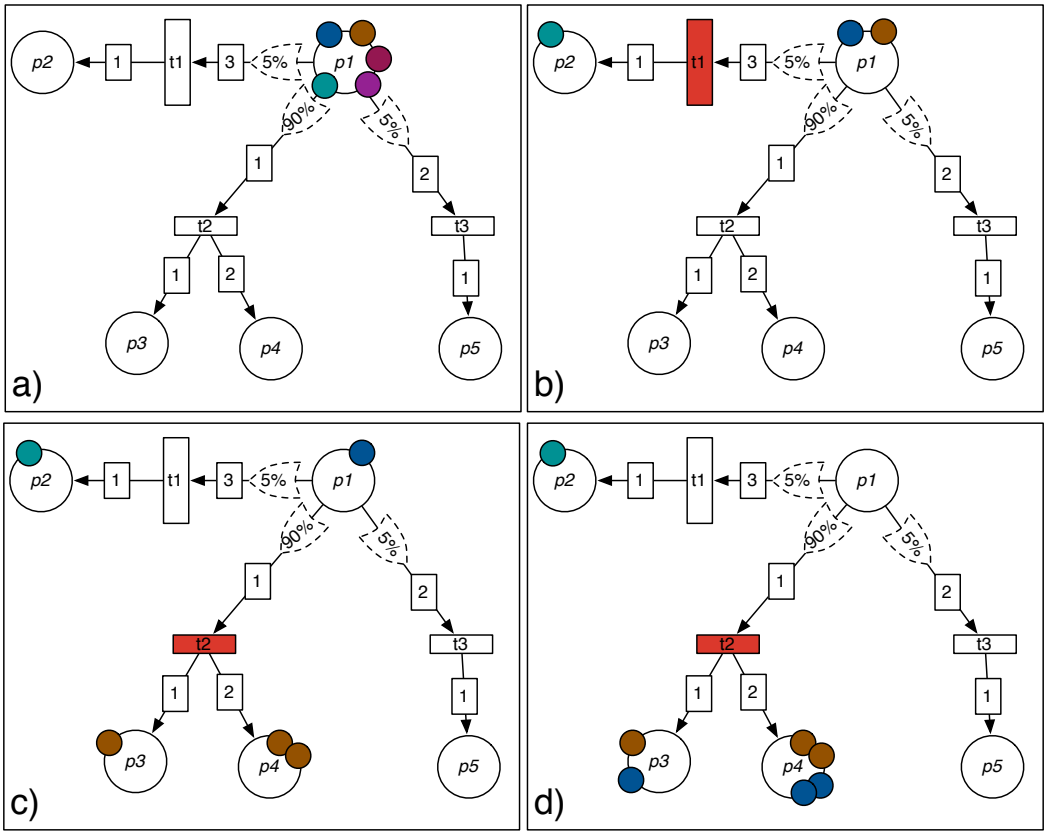


Figure 26 : Illustration de la dynamique des Flux Petri Nets (FPN). Les rectangles symbolisent les transitions (réactions), les cercles représentent les places (métabolites) et les tokens le flux. La succession des images illustre le déplacement des *tokens*.

3.2.2 FONCTION MULTI-OBJECTIFS

Les données expérimentales que nous souhaitons prendre en compte peuvent être de différentes natures et sont formulées sous la forme de contraintes à satisfaire. Le calcul d'une distance euclidienne entre le marquage final et attendu (par exemple, concentration d'un métabolite observée dans le modèle *versus* la concentration déterminée par expérimentation) nous sert ensuite à mesurer la différence entre les valeurs de concentration ou flux calculées par le modèle et les données expérimentales.

Il est possible d'indiquer par exemple (voir **Figure 27**):

- le maintien d'un équilibre pour l'ATP,
- la quantité attendue de produits finaux devant être produits par le réseau métabolique (un exemple est donné avec le succinate),
- le flux relatif attendu pour plusieurs enzymes.

$$dis\tan ce = \sqrt{\left(\frac{ATP_{obs} - ATP_{\exp}}{ATP_{\exp}}\right)^2 + \left(\frac{succinate_{obs} - succinate_{\exp}}{succinate_{\exp}}\right)^2 + \left(\frac{Flux_B + Flux_C - Flux_A}{Flux_B}\right)^2}$$

Figure 27 : Fonction de *distance* pour intégrer les données expérimentales.

Les données expérimentales à prendre en compte sont définies par des intervalles de valeurs représentant la quantité d'un métabolite ou par la définition de relations entre les flux. La qualité du simulateur sera fonction de sa capacité à minimiser la valeur de la *distance* pour un marquage final donné. Plusieurs itérations, testant différents marquages, sont effectuées jusqu'à obtention d'une valeur optimale de *distance* (mesure la satisfaction des contraintes entre données attendues et observées).

3.2.3 ALGORITHME HEURISTIQUE D'OPTIMISATION

L'objectif est d'obtenir un FPN avec un marquage qui respecte les équilibres, les proportions des produits finaux et les cofacteurs observés. L'obtention de ce comportement attendu revient à trouver **un flux F** de telle sorte que les marquages finaux des transitions soient en accord avec les données expérimentales. Cette étape consiste à rechercher les valeurs des flux minimisant la fonction de *distance*. La recherche de ce minimum est un problème d'optimisation combinatoire, et peut s'appréhender avec une méthode de recuit simulé ou un algorithme génétique. En nous basant sur une comparaison de leurs performances (voir [96]), notre choix s'est porté sur le recuit simulé dont les performances sont supérieures aux algorithmes génétiques.

Concrètement, le simulateur implémenté dans Métaboflux effectue ses calculs selon deux boucles d'itérations illustrées par la Figure 28 **:**

1. **Une première boucle simule le FPN** pour tester différents marquages, avec des valeurs aléatoires, jusqu'à ce que le nombre d'itérations maximum (paramètre à définir) soit atteint ou qu'il ne soit plus possible de franchir de transitions (tous les *tokens* étant déplacés). La probabilité qu'une solution de flux soit acceptée dépend de la valeur de *distance* obtenue. Pour éviter les minimums locaux, cette méthode d'exploration de l'espace des solutions est combinée à une recherche locale basée sur la méthode de Nelder-Mead/downhill-simplex [97].
2. **La seconde boucle** permet de répéter l'étape précédente jusqu'à ce que la procédure d'optimisation converge vers une solution optimale.

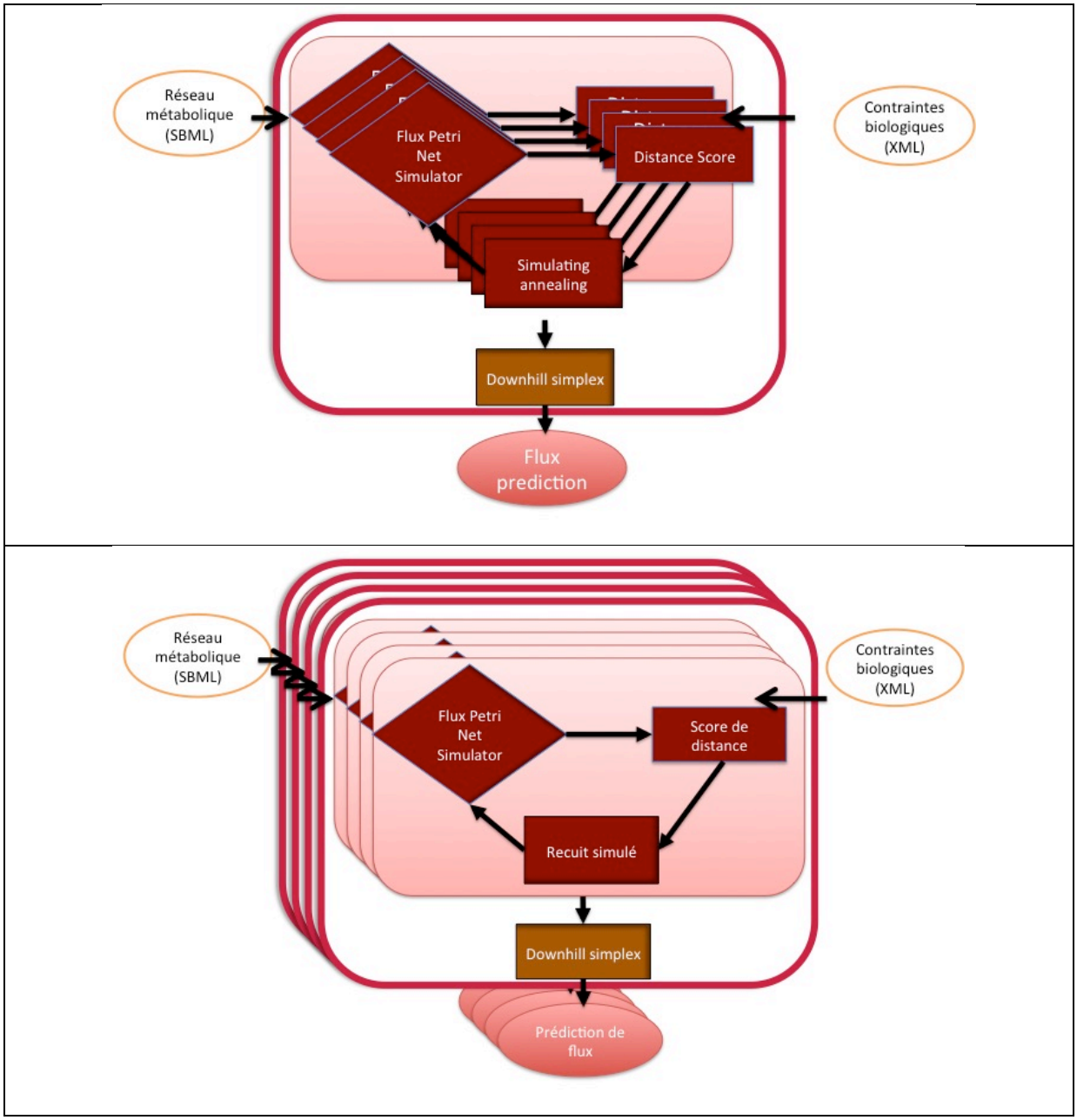


Figure 28 : Fonctionnement itératif de Metaboflux : (1) différents marquages du FPN sont testés pour un nombre d'itérations fixé, (2) l'étape précédente est répétée jusqu'à convergence d'une solution minimale.

3.3 CAS D'ETUDE

3.3.1 APPLICATION AU METABOLISME ENERGETIQUE DE TRYPANOSOMA BRUCEI

Le cas d'étude exploité par Metaboflux a porté sur l'analyse de *T. brucei*. Ce parasite alterne son cycle de vie entre l'hôte mammifère où il prend **une *forme sangui*ne (BSF)** et l'insecte vecteur où il existe sous ***la forme procy*clique (PF)**. Son métabolisme du glucose est très divergent pour chacune de ces formes. De plus, ce parasite a la particularité d'effectuer les six premières étapes de la glycolyse dans les glycosomes, qui sont des organelles similaires au peroxysome. Cette stratégie métabolique unique implique une utilisation strictement contrôlée des cofacteurs métaboliques (ATP et NAD +) à l'intérieur de l'organite où le non maintien des équilibres est léthal pour le parasite.

A partir des données publiées sur le réseau métabolique de *T. Brucei* [98][99], nous avons construit deux modèles correspondant à chacune des formes.

- **Le modèle BSF** a été utilisé pour valider notre approche (voir [100]) en intégrant deux contraintes : le maintien de l'équilibre des co-facteurs glycosomiques NAD^+/NADH et ATP/ADP (conditions vitales du protozoaire) ainsi qu'une valeur maximum d'ATP. Pour les deux conditions testées et simulées avec ou sans oxygène, les proportions attendues de pyruvate et ATP ont été obtenues.
- Dans le cas du **modèle PF** (voir Figure 29), la distribution du flux entre les différentes branches du réseau n'ayant encore jamais été analysée, nous avons exploité Metaboflux pour simuler son métabolisme énergétique. Les contraintes biologiques formulées pour ce modèle sont :
    1. maintien des ratios entre les co-facteurs glycosomiques NAD^+/NADH et ATP/ADP,
    2. un taux d'acétate excrété en rapport avec la quantité de succinate produit,
    3. un taux attendu de succinate dans le glycosome et dans la mitochondrie,
    4. un flux minimal pour la réaction PEP → OAA (étape 19).
    5. un flux attendu pour la réaction MAL → PYR (étapes 25, 26).

Nos résultats nous ont permis de démontrer qu'une proportion d'acétate/succinate comprise entre 26-50% autorise la satisfaction des contraintes du modèle. Ce modèle confirme que le choix des voies métaboliques en charge de la production d'acétate *versus* le succinate est très flexible et cette zone de plasticité est en accord avec plusieurs résultats expérimentaux effectués dans différentes conditions [101],[102].

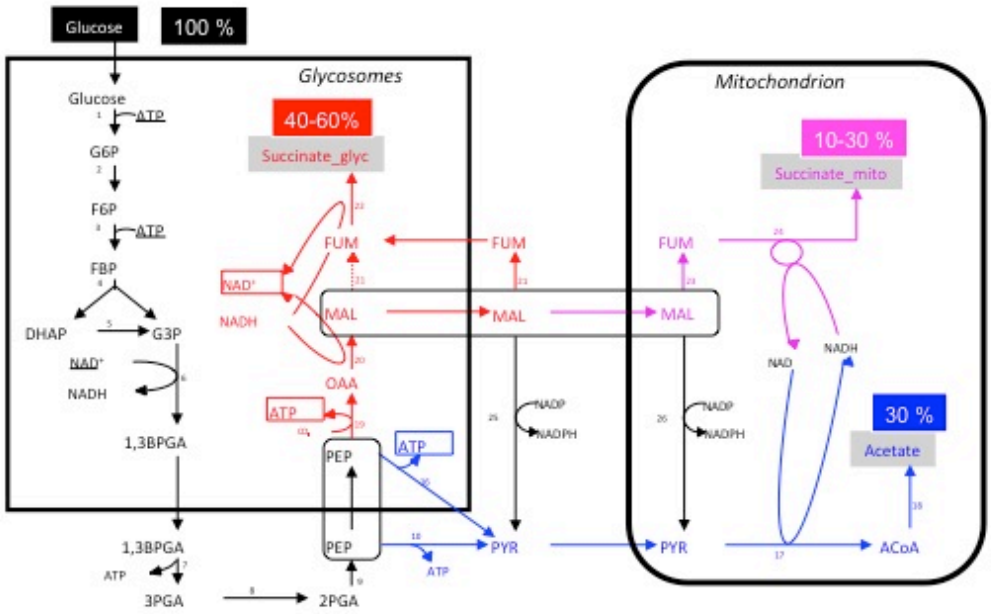


Figure 29 : Représentation du réseau métabolique énergétique des formes PF de T. brucei.

3.3.2 Cas d'application pour la visualisation

Pour comprendre les raisons de la flexibilité des concentrations d'acétate et de succinate produites, nous avons exploité le logiciel Systryp [16] dans le cadre d'une collaboration avec Romain Bourqui. Ce système de visualisation dédié à l'analyse des séries de données temporelles dans le contexte des réseaux métaboliques a l'avantage de proposer une visualisation dynamique des données temporelles.

Les différents états possibles du réseau ont été calculés par **Metaboflux** en faisant varier la quantité d'acétate excrétée *versus* la quantité de succinate dans un intervalle de valeurs variant de 1 à 99%, avec un pas de 5%. Les résultats ont ensuite été intégrés à **Systryp** pour visualiser la dynamique des flux et des concentrations des métabolites (voir Figure 30). Le logiciel **Systryp** a été paramétré pour représenter graphiquement les différentes valeurs des arêtes et sommets (*resp.* flux et concentrations de métabolites) proportionnellement aux valeurs obtenues pour chaque simulation de Metaboflux. La Figure 30 illustre quatre captures d'écran d'étapes intermédiaires de l'animation et démontre le rôle central des enzymes maliques dans le processus de flexibilité. Cette analyse est en accord avec d'autres expérimentations effectuées avec Metaboflux [100], lesquelles ont démontré également une corrélation entre la production d'acétate et le flux métabolique impliquant les enzymes maliques.

En conclusion, Metaboflux et Systryp nous ont permis d'expliquer et de visualiser l'importante flexibilité du trypanosome dans la production de succinate *versus* acétate, par l'intermédiaire des étapes impliquant l'enzyme malique (étapes 24 et 25).

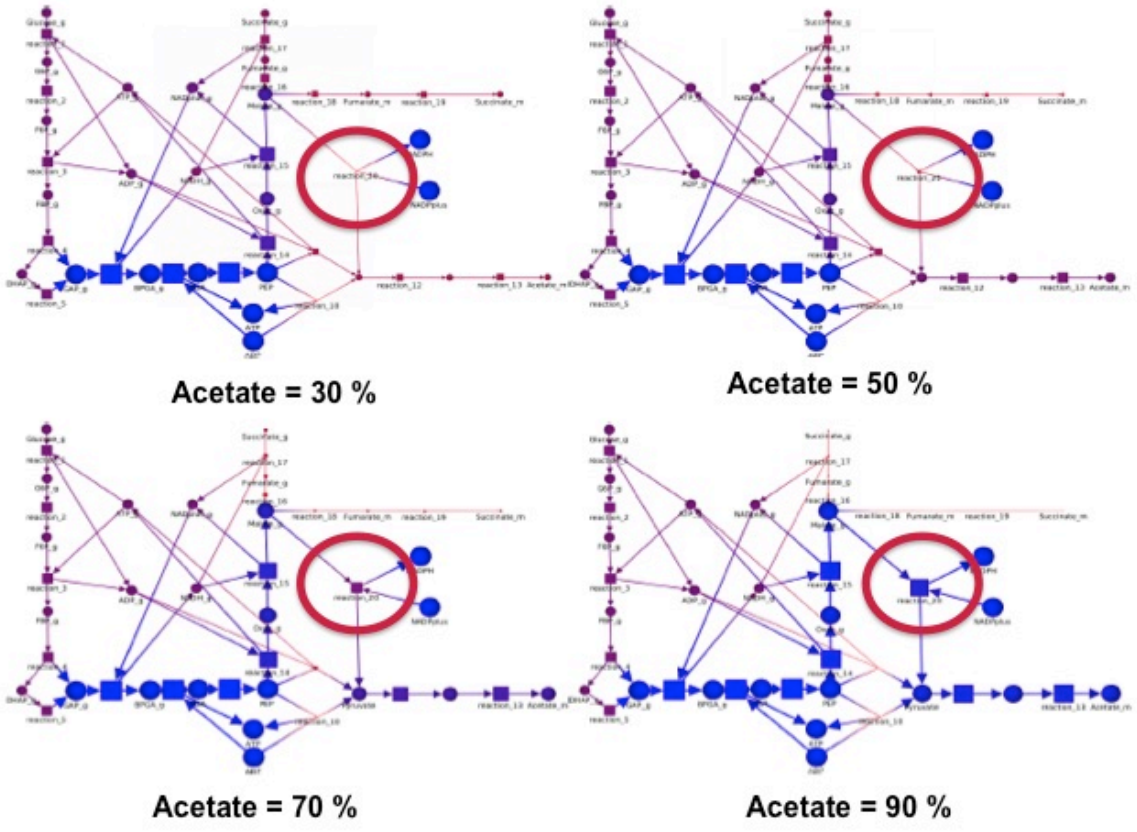


Figure 30 : Captures d'écran de Systryp correspondant à l'analyse de 4 résultats de Metaboflux obtenus pour différents ratios d'acétate/succinate. Un usage variable de l'enzyme malique est indiqué par un cercle rouge.

3.4 CONCLUSION

Grâce au bénéfice issu des grandes masses de données aujourd'hui disponibles, l'intégration de données hétérogènes est devenue essentielle pour affiner le caractère prédictif de nos modèles. En effet, cela revient à combiner différents points de vue du fonctionnement de la cellule et améliore nettement la réalité des modèles. Si l'analyse de la structure des réseaux biologiques reste l'approche initiale nécessaire pour appréhender la complexité de systèmes biologiques, l'étude de la dynamique contribue à obtenir une vision plus intégrée du fonctionnement cellulaire.

La compréhension de la régulation est également une étape importante dans l'étude du métabolisme en intégrant les différents niveaux d'organisation de la cellule. Le "phénotype métabolique" est une propriété dynamique du système biologique et ne peut être prédit sur la base de ses composantes statiques uniquement sans prendre en compte leurs relations avec des régulateurs. Ainsi, il sera intéressant d'identifier les gènes de "haut niveau" dans le réseau de régulation ayant une connexion directe ou indirecte avec un ensemble d'enzymes impliqués dans un même processus biologique pour caractériser les "interrupteurs" responsables d'un changement de métabolisme, ce qui pourrait donner lieu à des applications en biologie de synthèse. Dans ce cadre, nous envisageons différentes pistes de développement qui seront évoquées dans la partie **Perspectives**.

CHAPITRE 4 – TRAVAUX EN COURS ET PERSPECTIVES

Depuis les années 2000, l'essor des méthodes de comparaison de séquences biologiques et des systèmes les intégrant avec l'ensemble des données dont on dispose (par exemple Microscope dédié aux bactéries [103]) a été essentiel pour la découverte de nouvelles connaissances à l'échelle du génome telle que la compréhension de l'histoire évolutive de différentes espèces, famille de gènes ou des relations entre phénotype/génotype… Egalement, dans le monde des bactéries, les analyses des « core » et « pan » génome, qui exploitent la comparaison des répertoires de gènes de différents organismes, sont aujourd'hui un outil classique pour l'analyse du caractère pathogène des bactéries [104].

Aujourd'hui, la diversité, la quantité et la disponibilité des données nous permettent d'accéder à un niveau supérieur d'échelle où nous sommes en capacité d'étudier les objets biologiques ainsi que leurs nombreuses interactions. Ce niveau d'accès aux données offre également de nombreuses perspectives pour les approches comparatives des réseaux biologiques et va certainement améliorer notre compréhension du fonctionnement des organismes vivants dans leur globalité. Les applications résultantes de ces analyses sont nombreuses et nous fournissent une vision plus réaliste des ressemblances et différences de plusieurs organismes. Par exemple, la comparaison de réseaux biologiques permet d'identifier des structures conservées entre différents réseaux, lesquelles peuvent se définir comme des modules conservés au cours de l'évolution. Leur identification, d'un point de vue fonctionnel et intégrant des informations de régulation, devrait permettre d'ouvrir de nombreuses voies tant en biologie des systèmes qu'en bio-ingénierie avec la biologie de synthèse.
C'est dans ce contexte que se positionnent mes perspectives de recherche, avec un intérêt particulier pour proposer de nouvelles méthodes en bioinformatique pour la comparaison de différents organismes basée sur :

- l'interprétation fonctionnelle des groupes de gènes (peut correspondre à un sous graphe d'un réseau de régulation, par exemple l'ensemble des cibles d'un ARNnc),
- la prise en compte de manière concertée des réseaux de régulation et des réseaux métaboliques.

En d'autres termes, nos travaux visent à (1) développer de nouvelles approches informatiques pour **l'intégration unifiée de données hétérogènes multi-organismes** et à moyen-terme (2) étendre nos modèles basés sur les graphes pour **représenter et analyser de manière concertée la régulation et le métabolisme**. Plus spécifiquement, ces deux volets nous permettront d'aborder le traitement efficace des graphes en adaptant des approches pour la comparaison de graphes.

D'un point de vue applicatif, ces développements relèvent d'une question biologique générale centrée sur les relations entre l'évolution des génomes et l'évolution des réseaux biologiques. L'étude de ces relations, si on l'intègre dans une approche comparative, permettrait d'apporter des éléments supplémentaires pour aider à la compréhension des mécanismes d'évolution, et de fournir des pistes pour l'interprétation des réseaux en termes de contraintes

physiologiques. Cette étude pourrait aussi aider à tracer l'évolution des systèmes biologiques, et à comprendre comment ces systèmes complexes se sont mis en place au cours du temps.

4.1 DEVELOPPEMENT DE METHODES D'INTEGRATION DE SOURCES DE CONNAISSANCES HETEROGENES POUR UNE ANNOTATION UNIFIEE DE GROUPES DE GENES.

Ce travail, initié en 2016, résulte d'une collaboration avec Fleur Mougin, MCU de l'équipe ERIAS U1219 (thématique de recherche : représentation des connaissances et informatique médicale), et du co-encadrement d'Aaron Ayllón Benitez depuis 2015 (stage de master de recherche et un contrat CDD IE de 10 mois) qui poursuit actuellement ses travaux dans le cadre d'un contrat doctoral depuis septembre 2016. Ses travaux ont pour objectifs de développer de nouvelles approches informatiques visant à intégrer des informations de bases de connaissances pour une annotation fonctionnelle d'un ensemble de gènes regroupés sur la base d'un critère biologique.

4.1.1 CONTEXTE BIOLOGIQUE

L'expertise et la compréhension des réseaux biologiques auquel nous nous intéressons (un exemple est donné dans le chapitre 2 avec les réseaux de régulation) nécessitent de les interpréter biologiquement. Concrètement, cette compréhension implique d'identifier des annotations pertinentes afin de refléter la fonction biologique d'un groupe de gènes connectés dans le graphe. Pour mettre à profit les masses de données dont nous disposons, il est nécessaire d'exploiter conjointement plusieurs sources d'information dont la quantité et la diversité posent des défis d'intégration à plusieurs niveaux. En effet, ces informations issues de différentes bases de données sont à la fois de types très hétérogènes et reliées implicitement (par exemple, Gene Ontology, KeGG, Uniprot).

Pour illustrer, la Gene Ontology (GO) (ontologie la plus utilisée pour annoter les gènes) [105], regroupe plus de 30 000 termes pour décrire les rôles des gènes. Ces termes sont reliés hiérarchiquement et GO Annotation permet d'assigner un ou plusieurs termes de GO à un gène donné [106]. En moyenne, environ 10 termes sont associés à chaque gène dans le génome humain. Ceci implique que lorsque l'on annote un groupe de 100 gènes (par exemple, les 100 meilleures cibles prédites pour un microARN), 1000 termes GO (certains annotant plusieurs gènes) doivent être exploités pour interpréter biologiquement ce groupe. Si aucun traitement n'est appliqué pour déterminer les termes les plus pertinents, un tel nombre a un intérêt limité pour identifier une ou plusieurs fonctions communes à des sous-ensembles de gènes du groupe d'origine.

Pour palier à ces difficultés, nos travaux visent à concevoir de nouvelles approches informatiques pour le développement de procédures d'annotation intégrée avec l'objectif d'éliminer explicitement les redondances d'information et de pondérer les termes « ressemblants » proposés par de multiples sources. Dans ce contexte, nous développons de nouvelles approches de fouille de données pour déterminer l'ensemble restreint des annotations les plus pertinentes à associer à un groupe de gènes.

4.1.2 ENRICHISSEMENT DES ANNOTATIONS FONCTIONNELLES

Pour identifier les fonctions les plus représentatives d'un groupe de gènes, une approche classique en bioinformatique repose sur les méthodes d'enrichissement. L'idée sous-jacente de ces méthodes vise à comparer sur la base de leurs annotations:

- des groupes de gènes (exemple : gènes co-exprimés ou co-régulés pour une pathologie) et
- un catalogue de gènes de référence (exemple : l'ensemble complet des gènes de l'organisme considéré).

La stratégie mise en œuvre par les méthodes d'enrichissement s'appuie sur des approches statistiques pour identifier, au sein du groupe de gènes considérés, les sous-groupes possédant une annotation sur-représentée par rapport aux annotations utilisées par la référence (exemple : le génome complet). Un score statistique est ensuite calculé en fonction de la représentation du terme.

Dans ce cadre, plusieurs logiciels ont été proposés (voir pour revue Huang *et al.* [107]). Le choix de l'un de ces logiciels n'est pas une tâche facile et nécessite de prendre en considération plusieurs points essentiels dont nous avons proposé une liste dans [42]:

- Un premier verrou rencontré par ces différentes approches concerne la nomenclature des gènes cités. En effet, il est primordial que cette dernière soit identique au sein des bases de données et de la liste à analyser. Plusieurs identifiants de gènes peuvent être utilisés par différentes bases, et leur mise à jour pose également des problèmes dans le maintien de ces outils et a contraint pour plusieurs d'entre eux à restreindre leur application à des organismes modèles.
- Ces différentes méthodes fournissent une information redondante en sélectionnant des termes qui ont des liens de parenté au sein de la hiérarchie de GO, et rendent difficile, et parfois biaisée, l'interprétation des résultats [108],[109]. Un exemple est donné pour le gène atpD de *Escherichia coli* pour lequel 9 termes d'annotations dans la catégorie **Biological Process** sont proposés et dont la liste contient : transport, proton transport, ion transport.
- Les différentes mesures statistiques classiquement utilisées par les approches d'enrichissement peuvent avoir un impact variable sur la robustesse des analyses. La qualité, le type de données analysées ou encore les méthodes utilisées pour la construction des groupes [110] peuvent ainsi influencer les résultats selon le test utilisé.

Considérant tous ces critères, DAVID-WS [81] est actuellement le logiciel le plus largement utilisé dans la communauté scientifique et a été intégré au développement de rNAV [70]. Cette plateforme propose un serveur d'analyse pour de nombreuses espèces et intègre une large liste de bases de données d'annotation allant de bases de données généralistes (exemple : UniProt) à des bases plus spécialisées exploitant des informations telles que la conservation de séquences ou de domaines protéiques. Un autre avantage de DAVID-WS est d'exploiter plusieurs bases de données. L'intégration des résultats est ensuite réalisée a posteriori en regroupant les différents termes sur-représentés en fonction de leur co-utilisation par les gènes. Au final, cette étape de *clustering* naïve permet de regrouper les termes redondants de la GO tout en intégrant les annotations d'autres bases de données. *In fine*, bien qu'un ensemble de clusters de termes soit proposé au biologiste, une expertise manuelle reste nécessaire pour déterminer le processus biologique pertinent.

Pour pallier aux limites posées par les méthodes d'enrichissement et fournir une méthode automatique, nous avons choisi de privilégier les relations sémantiques décrites dans GO entre les termes d'annotation afin d'identifier les termes les plus représentatifs de la fonction d'un groupe de gènes.

4.1.3 OBJECTIFS

Notre approche exploitera indépendamment les termes d'annotation puis les gènes les utilisant. En effet, la proposition d'une annotation unifiée pour un groupe de gènes implique de différencier la redondance générée par la présence de termes similaires, de la complémentarité apportée par la combinaison de termes différents pour annoter un gène.

Dans ce cadre, trois volets sont envisagés :

- un premier volet vise à synthétiser l'ensemble des termes d'annotation utilisés par les gènes d'un jeu de données, sans tenir compte du nombre de gènes du groupe associé,
- un second volet ré-exploite ces résultats pour proposer une interprétation fonctionnelle d'un groupe en comparant les annotations des gènes le composant,
- un troisième volet s'appuiera sur les développements précédents pour : (i) étendre les méthodes en intégrant de nouvelles bases de connaissances pour l'annotation des gènes et (ii) mettre en place de nouvelles stratégies de comparaison des annotations de groupes de gènes intra ou inter-organismes.

Nos développements exploitent, de prime abord, la Gene Ontology pour la mise en œuvre de cette approche. Le choix de cette base de connaissances est d'une part motivée par sa large utilisation dans les projets d'annotation (afin d'appliquer et de valoriser rapidement nos travaux) et d'autre part, par le potentiel que représentent ses relations sémantiques. Sa structure et ses termes guideront l'intégration et la mise en relation avec de nouvelles bases de connaissances en identifiant les relations sémantiques pertinentes entre les termes de GO et ceux d'une nouvelle base.

4.1.4 SYNTHETISER LES TERMES D'ANNOTATION

1) Comparaison des termes d'annotations

La **similarité sémantique** est une métrique variant de 0 à 1 qui mesure la ressemblance entre deux entités (exemple : les termes de GO) d'une ontologie. Cette mesure peut être relativement sophistiquée afin de ne pas comparer deux entités exclusivement en fonction de leur libellé mais plutôt en fonction de leur contexte sémantique (par exemple, l'ensemble des entités reliées à l'entité d'intérêt). De nombreux travaux se sont intéressés à cette question et, en particulier, Pesquita et *al.* [111] ont effectué une revue d'articles proposant des méthodes de similarité sémantique entre termes de GO ou entre gènes annotés par un terme de GO. Les différentes méthodes se regroupent en trois catégories :

- les méthodes se basant sur les relations entre termes de GO (*edge based*) où par exemple une mesure de la similarité va correspondre au plus court chemin entre deux termes,
- les méthodes se basant sur les termes eux-mêmes (*node based*) où par exemple le nombre de gènes annotés par un terme est exploité, et
- les méthodes hybrides exploitant les termes et leurs relations.

Pour identifier les caractéristiques et avantages de ces différentes mesures, nous développons actuellement une approche comparative en considérant plusieurs jeux de données constitués par l'ensemble des termes de GO annotant des groupes de gènes humain et d'*E. coli*. Pour cela, la similarité a été calculée pour chaque paire de termes et été reportée dans une matrice de distance. Un algorithme de *clustering* hiérarchique ascendant a été ensuite appliqué pour répartir de manière itérative les termes dans un nombre de classes non défini *a priori* et permettre leur représentation en dendrogramme. Nous avons ensuite exploité deux méthodes de coupure classiques (méthode elbow & silhouette [112]) pour générer des clusters agrégeant les termes les plus similaires.

2) Identification des termes les plus pertinents dans un cluster de termes

A partir des résultats de *clustering*, l'étape suivante consiste à identifier les termes les plus représentatifs de chaque cluster. L'objectif principal sera de trouver le meilleur compromis entre identifier un nombre de termes restreint et le niveau de précision apporté par ces termes. La résolution de ce problème revient, à partir du parcours du graphe de la GO (où les nœuds correspondant aux termes du cluster sont marqués), à identifier le « meilleur » sous-ensemble de nœuds parents représentatifs des termes d'annotations regroupés. Pour cette étape, nous travaillons avec Julien Allali à la définition d'un algorithme de parcours d'arbre adapté à la GO. Le problème est difficile surtout en raison de l'absence de contrainte d'exclusion claire. Aussi, une solution consistera à fixer *a priori* la valeur du nombre de termes représentatifs du cluster.

4.1.5 COMPARAISON DES NOUVELLES ANNOTATIONS DES GENES

Les nouveaux termes représentatifs des groupes d'annotations "similaires" seront ensuite réaffectés aux gènes. A ce niveau, nous souhaitons à nouveau appliquer une réduction sur le nombre de termes associés au groupe de gène. Pour cela, pour le jeu de données considéré, il est nécessaire de différencier (i) les termes qui co-annotent souvent les mêmes gènes (information redondante) et (ii) les termes qui sont utilisés de manière complémentaire. En nous appuyant sur une étude prospective [113] exploitant le cadre formel de la recherche des itemsets fréquents et la GO, nous souhaitons développer une nouvelle méthode de comparaison d'annotation de gènes.

Plus précisément, la recherche d'itemsets fréquents est une branche des méthodologies de fouille de données visant à inférer des règles d'association entre des objets sur la base des items qui leur sont rattachés (pour une revue de leurs applications en bioinformatique, voir [114]). Dans le cas présent, les objets représentent les gènes tandis que les items sont les termes les annotant. Une transaction modélise la relation entre chaque gène et la liste de termes qui lui est associée. **La recherche des itemsets fréquents** consiste à rechercher les associations de termes qui co-annotent des gènes pour un nombre de termes donné. Plus formellement, le support d'un itemset comptabilise le nombre de gènes que les termes annotent simultanément.

L'adaptation du cadre formel donné par les itemsets fréquents nous permettra également de définir les différentes propriétés des règles d'association entre les termes d'annotation synthétiques. En nous appuyant sur la formalisation de différentes catégories d'itemset (voir Figure 31), nous en déduirons les relations significatives entre gènes. Par exemple :

- les termes **d'un itemset fréquent fermé** (définit un itemset dont le support est supérieur à celui de tous les itemsets le contenant) reflètent un niveau de redondance puisqu'ils co-occurrent très souvent.
- les termes **d'un itemset fréquent maximal** (définit un itemset fréquent dont aucun des itemsets le contenant n'est fréquent) reflètent une complémentarité des termes d'annotation.

Notre stratégie d'extraction de règles d'association comportera les deux étapes suivantes : (i) recherche exhaustive de l'ensemble des itemsets fréquents d'une taille donnée et (ii) caractérisation des propriétés des différentes classes d'itemset. Un point essentiel dans le choix des algorithmes d'extraction à adapter portera sur leurs performances pour un passage à l'échelle afin de prendre en compte de grands volumes de données. Dans ce sens, des travaux récents de prototypage ont été menés (dans le cadre d'un encadrement d'étudiants vietnamiens en master d'informatique) et nous ont fourni une implémentation en java de l'algorithme MAFIA [115]. Grâce à une représentation des données par des index bitmap, cet algorithme montre de très bonnes performances pour la prise en compte de grands jeux de données et servira de point de départ à nos travaux.

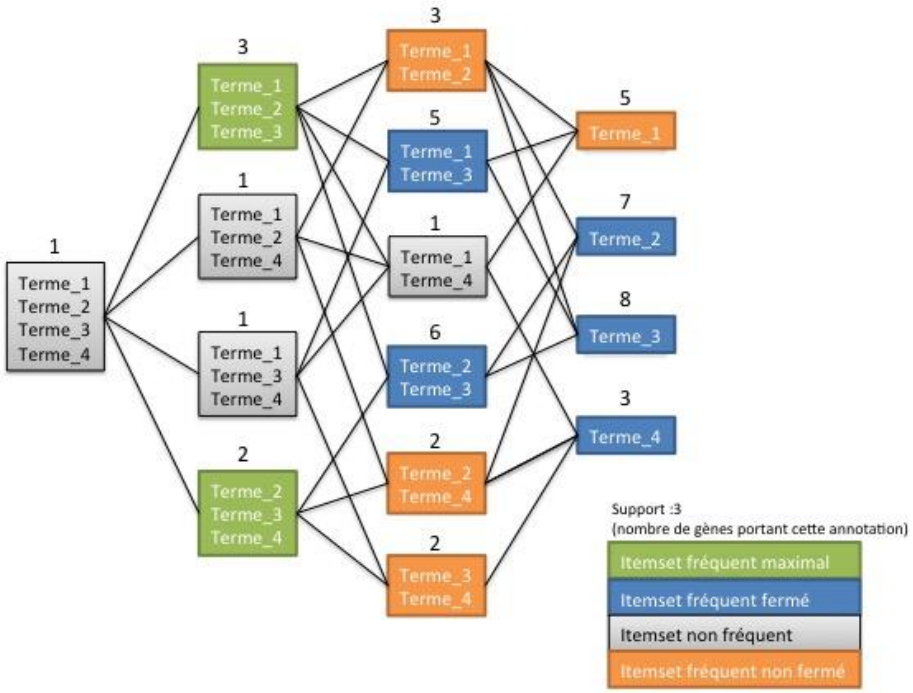


Figure 31 : Représentation des différentes classes d'itemsets fréquents.

4.1.6 INTEGRATION MULTI-ECHELLES (MULTI SOURCES ET MULTI ORGANISMES)

La suite de nos travaux s'intéressera à l'intégration d'autres sources de connaissances pour annoter les gènes. En effet, la première partie de ce projet se focalise principalement sur GO pour générer l'annotation unifiée. Cependant, d'autres sources de connaissances peuvent apporter des annotations complémentaires pour interpréter plus précisément les fonctions d'un groupe de gènes. On citera en particulier des sources de connaissances biologiques telles que KEGG (décrit les réseaux métaboliques), EC (décrit les fonctions enzymatiques) ou encore INTERPRO (décrit les domaines protéiques conservés).

L'intégration des termes d'annotation issus de ces nouvelles bases de connaissances est classiquement réalisée *a posteriori* pour comparer les gènes (exemple le *clustering* effectué par DAVID). Avec l'objectif de proposer une nouvelle annotation synthétisant les termes issus de multiples sources de connaissance, nous souhaitons (compétences en représentation des connaissances de Fleur Mougin) intégrer ces termes en amont afin:

- d'effectuer cette intégration une seule fois, indépendamment des gènes à comparer (diminution du post-traitement) et,
- de pondérer un terme d'annotation en fonction du nombre de sources le spécifiant.

Dans ce cadre, les technologies du Web sémantique et de l'initiative *Linked Open Data* seront particulièrement utiles. Dans le domaine biomédical en particulier, des travaux se sont attachés à rendre disponibles certaines sources de connaissances dans des formats exploitables automatiquement (RDF, OWL) avec la volonté de les interconnecter entre elles [116][117][118]. Cependant, les liens décrits entre les différentes sources de connaissances sont au niveau des données; les références croisées entre identifiants (de gènes) permettent de récupérer des informations issues de différentes sources de connaissances, mais il n'y a pas d'intégration au niveau des termes d'annotation eux-mêmes. En revanche, Callahan et *al.* ont exploité une ontologie comme pivot pour faire correspondre les connaissances de plusieurs sources intégrées dans Bio2RDF [119]. Ce travail, particulièrement intéressant, n'intègre cependant pas toutes les sources de connaissances permettant d'annoter les groupes de gènes. La description de certaines connaissances n'est pas assez précise. Nos développements s'inspireront de ce travail pour les étendre et les adapter à la spécificité de nos données.

Nous envisageons également d'apporter des contributions en génomique comparative en étendant nos travaux pour prendre en compte les relations d'homologie/orthologie entre gènes. L'intégration des relations pourrait s'effectuer à deux niveaux : (i) au niveau des termes d'annotations pour prendre en compte les relations d'orthologie et (ii) au niveau des gènes en définissant de nouveaux types de supports correspondant à deux gènes orthologues et permettant de favoriser la recherche des itemsets fréquents partagés par deux organismes donnés.

4-2 COMPARAISON DE RESEAUX BIOLOGIQUES

Les travaux que j'ai menés dans le cadre des réseaux de régulation et des réseaux métaboliques ont, jusqu'à présent, été menés indépendamment. Je souhaite m'intéresser à leur modélisation et analyse conjointe dans le cadre de développement d'approches bioinformatiques de **comparaison de réseaux métaboliques**. Dans ce contexte, le projet ModulNet (en collaboration avec Raluca Uricaru) vise à développer des techniques d'**alignement de graphes à partir de *graines*** (*e.g.,* sous graphes d'intérêt biologique).

4.2.1 CONTEXTE BIOLOGIQUE

La comparaison de réseaux métaboliques de deux organismes est un problème fondamental en Bioinformatique qui présente plusieurs applications allant de l'identification de fonctions qui leur sont communes/spécifiques à la compréhension de différences essentielles dans le phénotype. Comme beaucoup d'objets biologiques, les réseaux métaboliques font de plus apparaître de nombreux motifs/modules répétés. La détection de ces répétitions dans les

réseaux doit permettre une meilleure compréhension de leur fonctionnement. Cependant, ces motifs ne sont généralement pas répétés de manière exacte mais subissent de légères modifications, qui rendent leurs détections difficiles et dépendantes de méthodes de comparaison. Il est crucial de pouvoir identifier l'absence/présence de certains nœuds et motifs dans les réseaux, et de relier ces informations structurelles au phénotype. En effet, le "phénotype métabolique" est une propriété émergeante d'un réseau ne pouvant être prédite seulement sur la base de ses composantes statiques uniquement. Pour aborder cette question, nous avons choisi d'exploiter des données de régulation pour identifier les sous-graphes composés d'enzymes co-régulés et susceptibles d'être conservés au cours de l'évolution.

4.2.2 DEFINITION D'UN MODULE DANS UN RESEAU BIOLOGIQUE

A partir d'un réseau métabolique, une approche orientée structure définit les modules par une sous-partie connexe du graphe répondant à une propriété topologique particulière. Une seconde approche, complémentaire de la première, consiste à inférer les modules après avoir identifié des groupes de composants biologiques corrélés d'un point de vue fonctionnel. Enfin, certaines définitions de modularité intègrent des notions d'évolution entre espèces en se basant sur l'apparition ou la disparition de groupes d'éléments au cours du temps. Idéalement, une définition de la modularité devrait permettre la définition d'un réseau en unités élémentaires significatives d'un point de vue biologique.

Pour aborder cette question, nous souhaitons mettre à profit la **comparaison inter-organismes** ainsi que la réalité biologique pour aborder les différentes méthodes de décomposition de réseaux en modules. Typiquement, on s'intéressera aux comportements physiologiques atypiques à mettre en perspective avec une décomposition en modules.

Dans un premier temps, à partir d'un réseau métabolique G_M d'une bactérie, nous évaluerons plusieurs méthodes d'inférence de motifs/modules (par exemple pour inférer des sous-graphes de taille k en exploitant [120]). Nous développerons ensuite une méthode qui exploitera le graphe de régulation G_R pour **isoler parmi ces sous-graphes métaboliques, ceux qui sont co-régulés**. Les informations de régulation seront essentielles pour filtrer l'ensemble de modules d'intérêt. Des approches se basant sur des métriques exploitant des signatures de type *graphlets* [121] seront également évaluées dans le même cadre d'analyse. Les graphlets (sous-graphes induits non isomorphes calculés sur le réseau global) sont des mesures caractéristiques prenant en compte les similarités structurales et définissant des signatures sur la topologie des graphes.

4.2.2 ALIGNEMENT ET VISUALISATION DE RESEAUX BIOLOGIQUES

Les résultats de la première partie nous serviront à amorcer la seconde étape visant à proposer une nouvelle méthode pour la comparaison de réseaux métaboliques de plusieurs organismes. Une piste envisagée consiste à développer une heuristique d'alignement global guidée par l'utilisation de points d'ancrage à partir de sous-graphes présentant des propriétés spécifiques (par exemple des gènes hautement connectés ou des relations conservées chez plusieurs espèces).

L'ensemble de nos travaux pourra être appliqué à la comparaison du métabolisme des bactéries. En effet, ces travaux nous permettront de mener des analyses de comparaison sur le métabolisme de bactéries vaccinales et pathogènes, ou encore résistantes aux antibiotiques. L'objectif est d'identifier les meilleurs gènes candidats pour expliquer un comportement physiologique atypique.

4.2.3 VISUALISATION DE LA COMPARAISON DE DEUX RESEAUX BIOLOGIQUES

En collaboration avec Romain Bourqui, un volet consistera également à initier des approches de visualisation pour la représentation des résultats de comparaison de deux réseaux métaboliques. La prise en compte des modules inférés et conservés nous permettra de construire des représentations facilitant leur identification visuelle tout en représentant de manière différenciée les parties non alignées des deux réseaux biologiques. De plus, nous nous efforcerons d'intégrer en un seul graphe les différents types de réseaux manipulés. Alors que la visualisation simultanée de ces réseaux a été faite en général en utilisant plusieurs vues liées (une vue par type de réseau [115), nous nous intéresserons à la définition de nouvelles représentations dédiées aux réseaux multiplexes biologiques (permettent de représenter des multicouches intriquées entre elles lorsque les données et interactions se complexifient), c'est à dire des réseaux contenant des relations d'ordre différents (par exemple, "régule" ou encore "produit").

BIBLIOGRAPHIE

[1] K. Dolinski et O. G. Troyanskaya, « Implications of Big Data for cell biology », *Mol. Biol. Cell*, vol. 26, nº 14, p. 2575, juill. 2015.

[2] « An Integrated Encyclopedia of DNA Elements in the Human Genome », *Nature*, vol. 489, nº 7414, p. 57-74, sept. 2012.

[3] International HapMap 3 Consortium *et al.*, « Integrating common and rare genetic variation in diverse human populations », *Nature*, vol. 467, nº 7311, p. 52-58, sept. 2010.

[4] The 1000 Genomes Project Consortium, « A global reference for human genetic variation », *Nature*, vol. 526, nº 7571, p. 68-74, oct. 2015.

[5] Z. D. Stephens *et al.*, « Big Data: Astronomical or Genomical? », *PLOS Biol.*, vol. 13, nº 7, p. e1002195, juil 2015.

[6] S. Goodwin, J. D. McPherson, et W. R. McCombie, « Coming of age: ten years of next-generation sequencing technologies », *Nat. Rev. Genet.*, vol. 17, nº 6, p. 333-351, juin 2016.

[7] EMBL-European Bioinformatics Institute, « EMBL-EBI Annual Scientific Report 2016 ». .

[8] M. de Raad, C. R. Fischer, et T. R. Northen, « High-throughput platforms for metabolomics », *Curr. Opin. Chem. Biol.*, vol. 30, p. 7-13, févr. 2016.

[9] Z. Zhang, S. Wu, D. L. Stenoien, et L. Paša-Tolić, « High-throughput proteomics », *Annu. Rev. Anal. Chem. Palo Alto Calif*, vol. 7, p. 427-454, 2014.

[10] E. Baro, S. Degoul, R. Beuscart, et E. Chazard, « Toward a Literature-Driven Definition of Big Data in Healthcare », *BioMed Res. Int.*, vol. 2015, 2015.

[11] « Closure of the NCBI SRA and implications for the long-term future of genomics data storage », *Genome Biol.*, vol. 12, nº 3, p. 402, 2011.

[12] T. RNAcentral Consortium, « RNAcentral: a comprehensive database of non-coding RNA sequences », *Nucleic Acids Res.*, vol. 45, nº D1, p. D128-D134, avr. 2017.

[13] D. Gomez-Cabrero *et al.*, « Data integration in the era of omics: current and future challenges », *BMC Syst. Biol.*, vol. 8, nº Suppl 2, p. I1, mars 2014.

[14] S. A. Chervitz *et al.*, « Data standards for Omics data: the basis of data sharing and reuse », *Methods Mol. Biol. Clifton NJ*, vol. 719, p. 31-69, 2011.

[15] M. Hucka *et al.*, « Systems Biology Markup Language (SBML) Level 2 Version 5: Structures and Facilities for Model Definitions », *J. Integr. Bioinforma.*, vol. 12, nº 2, p. 271, sept. 2015.

[16] B. R. Dubois J Cottret L, Ghozlane A, Auber D, Bringaud F, Thébault P, Jourdan F., « Systrip: a visual environment for the investigation of time-series data in the context of metabolic networks », *Proc 16th Int. Conf. Inf. Vis. IV12*, p. pp 204-213, 2012.

[17] M. M. Rathore, A. Ahmad, A. Paul, et S. Rho, « Urban planning and building smart cities based on the Internet of Things using Big Data analytics », *Comput. Netw.*, vol. 101, p. 63-80, juin 2016.

[18] « Big boost in cyber-security spending », *Netw. Secur.*, vol. 2011, nº 12, p. 20, déc. 2011.

[19] C. Town, Éd., *Functional Genomics*. Dordrecht: Springer Netherlands, 2002.

[20] K. Yugi, H. Kubota, A. Hatano, et S. Kuroda, « Trans-Omics: How To Reconstruct Biochemical Networks Across Multiple 'Omic' Layers », *Trends Biotechnol.*, vol. 34, nº 4, p. 276-290, avr. 2016.

[21] V. Detours, J. E. Dumont, H. Bersini, et C. Maenhaut, « Integration and cross-validation of high-throughput gene expression data: comparing heterogeneous data sets », *FEBS Lett.*, vol. 546, nº 1, p. 98-102, juill. 2003.

[22] D. N. Macklin, N. A. Ruggero, et M. W. Covert, « The Future of Whole-Cell Modeling », *Curr. Opin. Biotechnol.*, vol. 28, p. 111-115, août 2014.

[23] M. Moretto *et al.*, « COLOMBOS v3.0: leveraging gene expression compendia for cross-species analyses », *Nucleic Acids Res.*, vol. 44, nº Database issue, p. D620-D623, janv. 2016.

[24] M. Kim, N. Rai, V. Zorraquino, et I. Tagkopoulos, « Multi-omics integration accurately predicts cellular state in unexplored conditions for Escherichia coli », *Nat. Commun.*, vol. 7, p.

13090, oct. 2016.
[25] C. S. Kruse, R. Goswamy, Y. Raval, et S. Marawi, « Challenges and Opportunities of Big Data in Health Care: A Systematic Review », *JMIR Med. Inform.*, vol. 4, nº 4, p. e38, nov. 2016.
[26] M. Bersanelli *et al.*, « Methods for the integration of multi-omics data: mathematical aspects », *BMC Bioinformatics*, vol. 17, nº 2, p. S15, 2016.
[27] V. Satagopam *et al.*, « Integration and Visualization of Translational Medicine Data for Better Understanding of Human Diseases », *Big Data*, vol. 4, nº 2, p. 97-108, juin 2016.
[28] O. Z. Barabasi AL, « Network biology: understanding the cell's functional organization. », *Nat Rev Genet*, vol. 5, nº 2, p. 101–113, févr. 2004.
[29] A. Lesne, « Complex Networks: from Graph Theory to Biology », *Lett. Math. Phys.*, vol. 78, nº 3, p. 235-262, nov. 2006.
[30] Z. N. Oltvai et A.-L. Barabási, « Systems biology. Life's complexity pyramid », *Science*, vol. 298, nº 5594, p. 763-764, oct. 2002.
[31] A. Lesne, « Robustness: confronting lessons from physics and biology », *Biol. Rev. Camb. Philos. Soc.*, vol. 83, nº 4, p. 509-532, nov. 2008.
[32] N. Le Novère, « Quantitative and logic modelling of molecular and gene networks », *Nat. Rev. Genet.*, vol. 16, nº 3, p. 146-158, mars 2015.
[33] G. Su, A. Kuchinsky, J. H. Morris, D. J. States, et F. Meng, « GLay: community structure analysis of biological networks », *Bioinformatics*, vol. 26, nº 24, p. 3135-3137, déc. 2010.
[34] T. Munzner, « A Nested Process Model for Visualization Design and Validation », *IEEE TVCG*, vol. 15, nº 6, p. 921-928, 2009.
[35] M. J. Eppler et R. Lengler, « Towards a periodic table of visualization methods », in *Proceedings of the 2007 IASTED Conference on Graphics and Visualization in Engineering*, Florida, 2007, p. 1.
[36] T. Lengauer, « Bioinformatics — From Genomes to Therapies », in *Bioinformatics - From Genomes to Therapies*, Wiley-VCH Verlag GmbH, 2008, p. 1-24.
[37] M. Zuker, « Mfold web server for nucleic acid folding and hybridization prediction », *Nucleic Acids Res.*, vol. 31, nº 13, p. 3406-3415, juill. 2003.
[38] « 6_Joris_Sansen.pdf ». .
[39] N. Gehlenborg et B. Wong, « Points of view: Heat maps », *Nat. Methods*, vol. 9, nº 3, p. 213-213, mars 2012.
[40] F. Supek, M. Bošnjak, N. Škunca, et T. Šmuc, « REVIGO Summarizes and Visualizes Long Lists of Gene Ontology Terms », *PLOS ONE*, vol. 6, nº 7, p. e21800, juil 2011.
[41] B.-J. Breitkreutz, C. Stark, et M. Tyers, « Osprey: a network visualization system », *Genome Biol.*, vol. 4, nº 3, p. R22, 2003.
[42] P Thebault, R Bourqui, C Gaspin, P Sirand-Pugnet, R Uricaru, I Dutour., « Advantages of mixing bioinformatics and visualization approaches for analyzing sRNA-mediated regulatory bacterial networks », *Brief. Bioinform.*, vol. 16, nº 5, p. 795-805, 2014.
[43] T. Aittokallio et B. Schwikowski, « Graph-based methods for analysing networks in cell biology », *Brief. Bioinform.*, vol. 7, nº 3, p. 243-255, janv. 2006.
[44] P. Shannon *et al.*, « Cytoscape: a software environment for integrated models of biomolecular interaction networks. », *Genome Res*, vol. 13, nº 11, p. 2498–2504, nov. 2003.
[45] Auber D, « Tulip- A Huge Graph Visualization Framework. », *P Mutzel M Jυunger Ed. Graph Draw. Softw. Math. Vis. Springer-Verl. 2003*, p. pp 105?126, 2003.
[46] S. R. Modi, D. M. Camacho, M. A. Kohanski, G. C. Walker, et J. J. Collins, « Functional characterization of bacterial sRNAs using a network biology approach », *Proc. Natl. Acad. Sci.*, vol. 108, nº 37, p. 15522-15527, 2011.
[47] L. S. Ahn YY Bagrow JP, « Link communities reveal multiscale complexity in networks. », *Nature*, vol. 466, nº 7307, p. 761–764, août 2010.
[48] J. S. Mattick, « RNA regulation: a new genetics? », *Nat. Rev. Genet.*, vol. 5, nº 4, p. 316-323, avr. 2004.
[49] W. K. Storz G Vogel J., « Regulation by Small RNAs in Bacteria: Expanding Frontiers. », *Mol Cell*, vol. 43, nº 6, p. 880-891, sept. 2011.

[50] M. E. Desnoyers G Bouchard MP, « New insights into small RNA-dependent translational regulation in prokaryotes. », *Trends Genet*, vol. 29, nº 2, p. 92–98, févr. 2013.
[51] P. Romby et E. Charpentier, « An overview of RNAs with regulatory functions in gram-positive bacteria. », *Cell Mol Life Sci*, vol. 67, nº 2, p. 217–237, janv. 2010.
[52] S. Brantl et R. Bruckner, « Small regulatory RNAs from low-GC Gram-positive bacteria. », *RNA Biol*, vol. 11, nº 5, févr. 2014.
[53] A. Toledo-Arana, F. Repoila, et P. Cossart, « Small noncoding RNAs controlling pathogenesis. », *Curr Opin Microbiol*, vol. 10, nº 2, p. 182–188, avr. 2007.
[54] A. Toledo-Arana *et al.*, « The Listeria transcriptional landscape from saprophytism to virulence. », *Nature*, vol. 459, nº 7249, p. 950–956, juin 2009.
[55] C. P. Mandin P Repoila F, Vergassola M, Geissmann T., « Identification of new noncoding RNAs in Listeria monocytogenes and prediction of mRNA targets. », *Nucleic Acids Res*, vol. 35, nº 3, p. 962-974, 2007.
[56] S. G. Gottesman S, « Bacterial small RNA regulators: versatile roles and rapidly evolving variations. », *Cold Spring Harb Perspect Biol*, vol. 3, nº 12, déc. 2011.
[57] S. Khandige, T. Kronborg, B. E. Uhlin, et J. Møller-Jensen, « sRNA-Mediated Regulation of P-Fimbriae Phase Variation in Uropathogenic Escherichia coli », *PLoS Pathog.*, vol. 11, nº 8, août 2015.
[58] A. E. Rentschler, S. D. Lovrich, R. Fitton, J. Enos-Berlage, et W. R. Schwan, « OmpR regulation of the uropathogenic Escherichia coli fimB gene in an acidic/high osmolality environment », *Microbiology*, vol. 159, nº Pt 2, p. 316-327, févr. 2013.
[59] N. B. Leontis et E. Westhof, « Analysis of RNA motifs. », *Curr Opin Struct Biol*, vol. 13, nº 3, p. 300–308, juin 2003.
[60] W. E. Vogel J, « Target identification of small noncoding RNAs in bacteria. », *Curr Opin Microbiol*, vol. 10, nº 3, p. 262–270, juin 2007.
[61] A. Pain, A. Ott, H. Amine, T. Rochat, P. Bouloc, et D. Gautheret, « An assessment of bacterial small RNA target prediction programs », *RNA Biol.*, vol. 12, nº 5, p. 509-513, 2015.
[62] M. Andronescu, Z. C. Zhang, et A. Condon, « Secondary structure prediction of interacting RNA molecules. », *J Mol Biol*, vol. 345, nº 5, p. 987–1001, févr. 2005.
[63] B. Tjaden, « Computational identification of sRNA targets. », *Methods Mol Biol*, vol. 905, p. 227–234, 2012.
[64] U. Muckstein, H. Tafer, J. Hackermuller, S. H. Bernhart, P. F. Stadler, et I. L. Hofacker, « Thermodynamics of RNA-RNA binding. », *Bioinformatics*, vol. 22, nº 10, p. 1177–1182, mai 2006.
[65] B. R. Busch A Richter AS, « IntaRNA: efficient prediction of bacterial sRNA targets incorporating target site accessibility and seed regions. », *Bioinformatics*, vol. 24, nº 24, p. 2849–2856, déc. 2008.
[66] L. S. Waters et G. Storz, « Regulatory RNAs in bacteria. », *Cell*, vol. 136, nº 4, p. 615–628, févr. 2009.
[67] A. Jousselin, L. Metzinger, et B. Felden, « On the facultative requirement of the bacterial RNA chaperone, Hfq. », *Trends Microbiol*, vol. 17, nº 9, p. 399–405, sept. 2009.
[68] W. R. Pearson, « Searching protein sequence libraries: comparison of the sensitivity and selectivity of the Smith-Waterman and FASTA algorithms. », *Genomics*, vol. 11, nº 3, p. 635–650, nov. 1991.
[69] J. Wang *et al.*, « sRNATarBase 3.0: an updated database for sRNA-target interactions in bacteria », *Nucleic Acids Res.*, vol. 44, nº D1, p. D248-253, janv. 2016.
[70] J. Dubois, A. Ghozlane, P. Thebault, I. Dutour, et B. Romain, « Genome-wide detection of sRNA targets with rNAV », in *Proc. of the 3rd Symposium on Biological Data Visualization*, 2013, p. 81–88.
[71] D. Auber, « Graph Drawing Software », P. Mutzel et M. Junger, Éd. Springer-Verlag, 2003.
[72] T. Munzner, « A Nested Model for Visualization Design and Validation », *IEEE Trans. Vis. Comput. Graph.*, vol. 15, nº 6, p. 921-928, nov. 2009.
[73] C. L. Beisel, T. B. Updegrove, B. J. Janson, et G. Storz, « Multiple factors dictate target selection by Hfq-binding small RNAs. », *EMBO J*, vol. 31, nº 8, p. 1961–1974, avr. 2012.

[74] R. M. Skippington E, « Evolutionary dynamics of small RNAs in 27 Escherichia coli and Shigella genomes. », *Genome Biol Evol*, vol. 4, nº 3, p. 330–345, 2012.
[75] M. H. Peer A, « Accessibility and Evolutionary Conservation Mark Bacterial Small-RNA Target-Binding Regions. », *J Bacteriol*, vol. 193, nº 7, p. 1690-1701, avr. 2011.
[76] H. R. Mika F, « Small Regulatory RNAs in the Control of Motility and Biofilm Formation in E. coli and Salmonella », *Int. J. Mol. Sci.*, vol. 14, nº 3, p. 4560-4579, 2013.
[77] C. Muelder, L. Gou, K.-L. Ma, et M. X. Zhou, « Multivariate Social Network Visual Analytics », p. 37-59, 2014.
[78] T. Moscovich, F. Chevalier, N. Henry, E. Pietriga, et J.-D. Fekete, « Topology-Aware Navigation in Large Networks », in *SIGCHI conference on Human Factors in computing systems*, Boston, États-Unis, 2009, p. 2319–2328.
[79] C. J. Modi SR Camacho DM, Kohanski MA, Walker GC, « Functional characterization of bacterial sRNAs using a network biology approach. », *Proc Natl Acad Sci U A*, vol. 108, nº 37, p. 15522-15527, sept. 2011.
[80] L. R. Huang DW Sherman BT, « Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists. », *Nucleic Acids Res*, vol. 37, nº 1, p. 1–13, janv. 2009.
[81] X. Jiao *et al.*, « DAVID-WS: A Stateful Web Service to Facilitate Gene/Protein List Analysis. », *Bioinformatics*, avr. 2012.
[82] Y. Cao *et al.*, « sRNATarBase: a comprehensive database of bacterial sRNA targets verified by experiments. », *RNA*, vol. 16, nº 11, p. 2051–2057, nov. 2010.
[83] I. M. Keseler *et al.*, « EcoCyc: fusing model organism databases with systems biology. », *Nucleic Acids Res*, vol. 41, nº Database issue, p. D605–D612, janv. 2013.
[84] P. R. Wright *et al.*, « CopraRNA and IntaRNA: predicting small RNA targets, networks and interaction domains. », *Nucleic Acids Res*, vol. 42, nº Web Server issue, p. W119–W123, juill. 2014.
[85] C. Michaux, N. Verneuil, A. Hartke, et J.-C. Giard, « Physiological roles of small RNA molecules. », *Microbiology*, vol. 160, nº Pt 6, p. 1007–1019, juin 2014.
[86] S. Durand et G. Storz, « Reprogramming of anaerobic metabolism by the FnrS small RNA », *Mol Microbiol*, vol. 75, p. 1215–1231, 2010.
[87] A. Boysen, J. Moller-Jensen, B. Kallipolitis, P. Valentin-Hansen, et M. Overgaard, « Translational regulation of gene expression by an anaerobically induced small non-coding RNA in Escherichia coli. », *J Biol Chem*, vol. 285, nº 14, p. 10690–10702, avr. 2010.
[88] M. Guell *et al.*, « Transcriptome complexity in a genome-reduced bacterium. », *Science*, vol. 326, nº 5957, p. 1268–1271, nov. 2009.
[89] D. A. Fell et J. R. Small, « Fat synthesis in adipose tissue. An examination of stoichiometric constraints. », *Biochem J*, vol. 238, nº 3, p. 781–786, sept. 1986.
[90] J. D. Orth, I. Thiele, et B. Ø. Palsson, « What is flux balance analysis? », *Nat Biotechnol*, vol. 28, nº 3, p. 245–248, mars 2010.
[91] Y.-G. Oh, D.-Y. Lee, S. Y. Lee, et S. Park, « Multiobjective flux balancing using the NISE method for metabolic network analysis. », *Biotechnol Prog*, vol. 25, nº 4, p. 999–1008, 2009.
[92] D. Nagrath, M. Avila-Elchiver, F. Berthiaume, A. W. Tilles, A. Messac, et M. L. Yarmush, « Integrated energy and flux balance based multiobjective framework for large-scale metabolic networks. », *Ann Biomed Eng*, vol. 35, nº 6, p. 863–885, juin 2007.
[93] I. Koch, B. H. Junker, et M. Heiner, « Application of Petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. », *Bioinformatics*, vol. 21, nº 7, p. 1219–1226, avr. 2005.
[94] C. Chaouiya, « Petri net modelling of biological networks. », *Brief Bioinform*, vol. 8, nº 4, p. 210–219, juill. 2007.
[95] M. Ajmone Marsan, G. Conte, et G. Balbo, « A class of generalized stochastic Petri nets for the performance evaluation of multiprocessor systems », *ACM Trans. Comput. Syst. TOCS*, vol. 2, nº 2, p. 93–122, 1984.
[96] S. Bandyopadhyay, S. Saha, U. Maulik, et K. Deb, « A Simulated Annealing-Based Multiobjective Optimization Algorithm: AMOSA », *IEEE Trans. Evol. Comput.*, vol. 12, nº 3, p. 269–

283, 2008.
[97] J. A. Nelder et R. Mead, « A Simplex Method for Function Minimization », *Comput. J.*, vol. 7, nº 4, p. 308-313, janv. 1965.
[98] F. Bringaud, L. Rivière, et V. Coustou, « Energy metabolism of trypanosomatids: adaptation to available carbon sources. », *Mol Biochem Parasitol*, vol. 149, nº 1, p. 1–9, sept. 2006.
[99] F. Bringaud, C. Ebikeme, et M. Boshart, « Acetate and succinate production in amoebae, helminths, diplomonads, trichomonads and trypanosomatids: common and diverse metabolic strategies used by parasitic lower eukaryotes. », *Parasitology*, vol. 137, nº 9, p. 1315–1331, août 2010.
[100] A. Ghozlane, F. Bringaud, H. Soueidan, I. Dutour, F. Jourdan, et P. Thébault, « Flux Analysis of the Trypanosoma brucei Glycolysis Based on a Multiobjective-Criteria Bioinformatic Approach », *Adv. Bioinforma.*, vol. 2012, p. 159423, 2012.
[101] V. Coustou *et al.*, « Fumarate is an essential intermediary metabolite produced by the procyclic Trypanosoma brucei. », *J Biol Chem*, vol. 281, nº 37, p. 26832–26846, sept. 2006.
[102] S. W. H. van Weelden *et al.*, « Procyclic Trypanosoma brucei do not use Krebs cycle activity for energy generation. », *J Biol Chem*, vol. 278, nº 15, p. 12854–12863, avr. 2003.
[103] D. Vallenet *et al.*, « MicroScope in 2017: an expanding and evolving integrated resource for community expertise of microbial genomes », *Nucleic Acids Res.*, vol. 45, nº D1, p. D517-D528, janv. 2017.
[104] L. Rouli, V. Merhej, P.-E. Fournier, et D. Raoult, « The bacterial pangenome as a new tool for analysing pathogenic bacteria », *New Microbes New Infect.*, vol. 7, p. 72-85, juin 2015.
[105] M. Ashburner *et al.*, « Gene ontology: tool for the unification of biology. The Gene Ontology Consortium », *Nat. Genet.*, vol. 25, nº 1, p. 25-29, mai 2000.
[106] E. Camon *et al.*, « The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro », *Genome Res.*, vol. 13, nº 4, p. 662-672, avr. 2003.
[107] D. W. Huang, B. T. Sherman, et R. A. Lempicki, « Bioinformatics enrichment tools: paths toward the comprehensive functional analysis of large gene lists », *Nucleic Acids Res.*, vol. 37, nº 1, p. 1-13, janv. 2009.
[108] G. Yu, F. Li, Y. Qin, X. Bo, Y. Wu, et S. Wang, « GOSemSim: an R package for measuring semantic similarity among GO terms and gene products », *Bioinforma. Oxf. Engl.*, vol. 26, nº 7, p. 976-978, avr. 2010.
[109] J. Wang *et al.*, « GO-function: deriving biologically relevant functions from statistically significant functions », *Brief. Bioinform.*, vol. 13, nº 2, p. 216-227, mars 2012.
[110] T. Bleazard, J. A. Lamb, et S. Griffiths-Jones, « Bias in microRNA functional enrichment analysis », *Bioinforma. Oxf. Engl.*, vol. 31, nº 10, p. 1592-1598, mai 2015.
[111] C. Pesquita, D. Faria, A. O. Falcão, P. Lord, et F. M. Couto, « Semantic similarity in biomedical ontologies », *PLoS Comput. Biol.*, vol. 5, nº 7, p. e1000443, juill. 2009.
[112] « NbClust: An R Package for Determining the Relevant Number of Clusters in a Data Set | Charrad | Journal of Statistical Software ». [En ligne]. Disponible sur: https://www.jstatsoft.org/article/view/v061i06. [Consulté le: 18-juill-2016].
[113] P. H. Guzzi, M. Milano, et M. Cannataro, « Mining Association Rules from Gene Ontology and Protein Networks: Promises and Challenges. », *Procedia Comput. Sci.*, vol. 29, nº 0, p. 1970-1980, 2014.
[114] S. Naulaerts *et al.*, « A primer to frequent itemset mining for bioinformatics », *Brief. Bioinform.*, vol. 16, nº 2, p. 216-231, mars 2015.
[115] D. Burdick, M. Calimlim, et J. Gehrke, « MAFIA: a maximal frequent itemset algorithm for transactional databases », in *17th International Conference on Data Engineering, 2001. Proceedings*, 2001, p. 443-452.
[116] M. Samwald *et al.*, « Linked open drug data for pharmaceutical research and development », *J. Cheminformatics*, vol. 3, nº 1, p. 19, 2011.
[117] F. Belleau, M.-A. Nolin, N. Tourigny, P. Rigault, et J. Morissette, « Bio2RDF: towards a mashup to build bioinformatics knowledge systems », *J. Biomed. Inform.*, vol. 41, nº 5, p. 706-716, oct. 2008.

[118] M. J. García Godoy, E. López-Camacho, I. Navas-Delgado, et J. F. Aldana-Montes, « Sharing and executing linked data queries in a collaborative environment », *Bioinforma. Oxf. Engl.*, vol. 29, nº 13, p. 1663-1670, juill. 2013.
[119] A. Callahan, J. Cruz-Toledo, et M. Dumontier, « Ontology-Based Querying with Bio2RDF's Linked Open Data », *J. Biomed. Semant.*, vol. 4 Suppl 1, p. S1, avr. 2013.
[120] V. Lacroix, G. Fernandes C., et M. Sagot, « Reaction motifs in metabolic networks », in *Proceedings of 5th Workshop on Algorithms for BioInformatics (WABI'05), Lecture Notes in BioInformatic*, 2005, vol. 3692, p. 178-191.
[121] T. Milenković et N. Przulj, « Uncovering biological network function via graphlet degree signatures », *Cancer Inform.*, vol. 6, p. 257-273, 2008.
[122] W. M. Bourqui R, « Visualizing temporal dynamics at the genomic and metabolic level », *Procof 13th Int. Conf. Inf. Vis.*, p. 317?322, 2009.