Développement de méthodes bioinformatiques dédiées à la prédiction et l'analyse des réseaux métaboliques et des ARN non codants

Amine Ghozlane

N° d'ordre : 4617

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX 1

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par Amine GHOZLANE

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : *INFORMATIQUE*

Développement de méthodes bioinformatiques dédiées à la prédiction et à l'analyse des réseaux métaboliques et des ARN non codants.

Directrice de recherche : Maylis DELEST
Co-Directrice : Isabelle DUTOUR
Co-Directrice : Patricia THÉBAULT

Soutenue le : 20 Novembre 2012

Devant la commission d'examen formée de :

M. DENISE, Alain	Professeur	Université Paris-Sud	Rapporteur
Mme GASPIN, Christine	Directrice de recherche	INRA	Rapportrice
M. JOURDAN, Fabien	Chargé de recherche	INRA	Examinateur
M. MAZAT, Jean-Pierre	Professeur	Université Bordeaux 2	Examinateur

Développement de méthodes bioinformatiques dédiées à la prédiction et l'analyse des réseaux métaboliques et des ARN non codants.

Résumé :

L'identification des interactions survenant au niveau moléculaire joue un rôle crucial pour la compréhension du vivant. L'objectif de ce travail a consisté à développer des méthodes permettant de modéliser et de prédire ces interactions pour le métabolisme et la régulation de la transcription. Nous nous sommes basés pour cela sur la modélisation de ces systèmes sous la forme de graphes et d'automates.

Nous avons dans un premier temps développé une méthode permettant de tester et de prédire la distribution du flux au sein d'un réseau métabolique en permettant la formulation d'une à plusieurs contraintes. Nous montrons que la prise en compte des données biologiques par cette méthode permet de mieux reproduire certains phénotypes observés *in vivo* pour notre modèle d'étude du métabolisme énergétique du parasite *Trypanosoma brucei*. Les résultats obtenus ont ainsi permis de fournir des éléments d'explication pour comprendre la flexibilité du flux de ce métabolisme, qui étaient cohérentes avec les données expérimentales. Dans un second temps, nous nous sommes intéressés à une catégorie particulière d'ARN non codants appelés sRNAs, qui sont impliqués dans la régulation de la réponse cellulaire aux variations environnementales. Nous avons développé une approche permettant de mieux prédire les interactions qu'ils effectuent avec d'autres ARN en nous basant sur une prédiction des interactions, une analyse par enrichissement du contexte biologique de ces cibles, et en développant un système de visualisation spécialement adapté à la manipulation de ces données. Nous avons appliqué notre méthode pour l'étude des sRNAs de la bactérie *Escherichia coli*. Les prédictions réalisées sont apparues être en accord avec les données expérimentales disponibles, et ont permis de proposer plusieurs nouvelles cibles candidates.

Mots-clefs : Bioinformatique, Graphe, réseau de Petri, métabolisme, Flux Balance Analysis, ARN non codant, prédiction des cibles des sRNAs.

Discipline : Informatique

Development of bioinformatic methods dedicated to the prediction and the analysis of metabolic networks and non-coding RNA.

Abstract :

The identification of the interactions occurring at the molecular level is crucial to understand the life process. The aim of this work was to develop methods to model and to predict these interactions for the metabolism and the regulation of transcription. We modeled these systems by graphs and automata.

Firstly, we developed a method to test and to predict the flux distribution in a metabolic network, which consider the formulation of several constraints. We showed that this method can better mimic the *in vivo* phenotype of the energy metabolism of the parasite *Trypanosoma brucei*. The results enabled to provide a good explanation of the metabolic flux flexibility, which were consistent with the experimental data. Secondly, we have considered a particular class of non-coding RNAs called sRNAs, which are involved in the regulation of the cellular response to environmental changes. We developed an approach to better predict their interactions with other RNAs based on the interaction prediction, an enrichment analysis, and by developing a visualization system adapted to the manipulation of these data. We applied our method to the study of the sRNAs interactions within the bacteria *Escherichia coli*. The predictions were in agreement with the available experimental data, and helped to propose several new target candidates.

Key words : Bioinformatic, Graph, Petri nets, metabolism, Flux Balance Analysis, Non-coding RNA, sRNA target prediction.
Field : Computer science

Laboratoire Bordelais de Recherche en Informatique (LaBRI)
Université Bordeaux 1,
351, cours de la libération
33405 Talence (FRANCE)

Table des matières

Introduction

Le fonctionnement de la cellule peut se modéliser par un système complexe impliquant l'interaction d'entités biologiques de natures variées. À l'échelle de la cellule, ce système aboutit à la réalisation de phénotypes complexes. Pour mieux comprendre le vivant, il est ainsi nécessaire d'étudier les molécules qui composent ces systèmes, de même que leurs interactions. Grâce aux récentes avancées technologiques dans le domaine de la biologie, il est à présent plus facile de détecter et de caractériser ces différentes molécules. Nous disposons ainsi de plus en plus d'informations sur ces entités. Cependant l'identification de leurs interactions, et ainsi de leurs rôles, reste une question difficile à aborder, de par le nombre d'interactions en jeu et la complexité des mécanismes sous-jacents.

Des efforts importants ont donc été réalisés en Bioinformatique pour intégrer et donner du sens à ces nouvelles données grâce à la définition de modèles. Un modèle permet de décrire de manière simplifiée des connaissances. Il peut être construit à partir d'une hypothèse sur les données ou le résultat d'une prédiction, puis être comparé au phénotype observé en simulant par exemple son fonctionnement (sur la base des principes physiques et chimiques qui régissent le vivant). Une application intéressante des modèles en biologie peut être réalisée sous la forme de graphes. Cette approche permet de représenter de manière intuitive les relations entre les différents éléments, telles que les interactions de métabolites et de protéines sous la forme de réseaux métaboliques ou l'interaction de molécules d'ARN entre elles sous la forme de réseaux de régulation. Elle peut également permettre d'intégrer plus facilement des mesures biologiques dans le modèle et de mieux visualiser ces mécanismes, pour améliorer la prédiction de nouveaux états ou de nouvelles interactions du système.

Durant cette thèse, nous avons ainsi abordé les thématiques de la prédiction du fonctionnement d'un réseau métabolique et de la prédiction des interactions des ARN non codants par la modélisation de ces systèmes sous la forme de réseaux. Nous avons considéré ces deux types de modélisations sous différentes échelles, en nous basant pour l'analyse des réseaux métaboliques sur une approche dite ascendante (*Bottom-up*). Nous avons initié notre étude à partir des différents éléments identifiés expérimentalement, avant de prédire à l'aide de la modélisation et de la simulation des phénotypes globaux. À l'inverse, nous avons réalisé pour les ARN non codants une approche dite descendante (*Top-down*). Nous avons recherché au sein de l'ensemble des interactions prédites pour une catégorie d'ARN non codant appelés sRNAs, les éléments qui pouvaient être impliqués dans certains mécanismes moléculaires bien spécifiques.

Prédiction de la distribution des flux au sein d'un réseau métabolique

Lors de la reconstruction du réseau métabolique d'un organisme, on cherche à identifier l'ensemble des réactions qui peuvent survenir. Ceci peut être réalisé de deux manières. Une première approche consiste à comparer le génome de l'espèce étudiée à celui d'une ou plusieurs

autres espèces proches d'un point de vue génétique. Il est ainsi possible d'inférer, par la similarité entre ces génomes, les réactions de l'organisme étudié. Cependant lorsque le mécanisme considéré est unique, il est nécessaire de procéder à la détection des différentes réactions par expérimentation. La question qui se pose alors porte sur la fonctionnalité et le phénotype du modèle formulé.
Afin d'étudier cette problématique, nous nous sommes donc intéressés à la prédiction de la distribution des flux au sein d'un réseau métabolique. Un flux correspond à la proportion de métabolites empruntant une réaction. Il peut être mesuré expérimentalement, et comparé au résultat obtenu lors de la simulation du modèle reconstruit, ce qui peut apporter de nombreuses informations sur la fonctionnalité ou non du modèle. Pour cette étude, il est aussi nécessaire de considérer la disponibilité des paramètres du réseau, leur nature, ainsi que le niveau de réalisme de la simulation qui peut être obtenu. Ces problèmes de vérification du fonctionnement et d'intégration des données d'un réseau métabolique ont ainsi été abordés par différentes méthodes. Cependant, la définition de multiples objectifs pour contraindre un réseau métabolique n'avait jusque-là été abordée que de façon théorique, sans proposer d'outil bioinformatique. Pour répondre à ces questions, nous avons donc développé et implémenté une nouvelle approche au sein du framework Metaboflux. Cette méthode se base sur la prédiction de la distribution des flux au sein d'un réseau métabolique modélisé sous la forme d'un *Flux Petri net*, soumis à plusieurs contraintes formulées à partir des connaissances biologiques.

Prédiction des cibles des sRNAs

D'autre part, nous avons considéré la question de la prédiction des interactions d'une catégorie particulière d'ARN non codants appelés sRNAs. Ces ARN ont un rôle important pour coordonner la réponse de l'organisme lors d'un changement des conditions environnementales, en régulant l'expression et la traduction d'autres ARN. Grâce aux nouvelles technologies de séquençage, il est aujourd'hui possible d'identifier bien plus rapidement qu'auparavant ces éléments. Il se pose alors la question de leur rôle et donc des interactions qu'ils effectuent.
L'approche bioinformatique peut identifier des cibles potentielles de ces sRNAs et donc simplifier le travail de recherche du biologiste, en réduisant le nombre de conditions et de cibles à étudier à quelques candidats. Différentes études ont également montré l'apport de l'analyse d'enrichissement ou de la visualisation pour la prédiction de cibles. L'intégration de l'ensemble de ces analyses en un seul système pour les sRNAs n'avait cependant pas été réalisée jusque-là. De plus, ces méthodes ne permettent pas en l'état actuel de rechercher simultanément les cibles de plusieurs sRNAs, ce qui est pourtant nécessaire pour identifier des motifs de régulation impliquant plusieurs de ces ARN. Nous avons donc développé le pipeline iRNA qui permet de prendre en compte la prédiction des interactions et le contexte des connaissances biologiques pour ces éléments, avec l'utilisation d'un système de visualisation adapté à ces données.

Nos apports ont ainsi consisté pour ces deux thématiques au développement de méthodes permettant d'améliorer l'intégration des données biologiques. Nous nous sommes également intéressés dans chaque cas à la visualisation de ces données sous la forme de graphes. Nous avons pu aborder par cette approche les problèmes biologiques sous un angle différent, en considérant par exemple la topologie (grâce à la théorie des graphes) ou la dynamique de ces réseaux (grâce à la théorie des automates) et observer un réalisme accru de nos prédictions. Les résultats que nous avons obtenus pour le métabolisme ont ainsi fait l'objet d'une publication [Ghozlane *et al.* 2012]. Quant au travail sur les ARN, nous publierons prochainement les

travaux effectués sur le plan des nouvelles visualisations développées.

Plan du mémoire

Le manuscrit est organisé en cinq chapitres et deux parties. Le Chapitre 1 présente les notions de théorie des graphes et d'automates qui constituent les principaux outils théoriques utilisés pour nos études. Dans la première partie, nous nous intéressons au travail relatif au métabolisme. Ainsi, nous présentons dans le Chapitre 2 ce qu'est le métabolisme et faisons l'état de l'art des méthodes employées pour le reconstruire et le modéliser. Cette étude nous a amené à proposer notre propre méthode d'analyse des réseaux métaboliques, qui est décrite dans le chapitre suivant. Le Chapitre 3 détaille le fonctionnement de la méthode développée et présente les résultats obtenus pour notre modèle d'étude *Trypanosoma brucei*. Dans la seconde partie, nous présentons le travail réalisé pour l'étude des ARN non codants et la prédiction des cibles des sRNAs. Dans le Chapitre 4, nous établissons un état de l'art sur les connaissances dont on dispose sur ces ARN et sur les méthodes permettant de les détecter et de prédire leurs interactions. Dans le Chapitre 5, nous présentons l'approche que nous avons développée pour améliorer la recherche des cibles des sRNAs et les résultats préliminaires obtenus pour notre modèle d'étude *Escherichia coli*.

Chapitre 1

Intégration des données en bioinformatique

Sommaire

1.1 Graphes

Les graphes constituent un outil majeur en bioinformatique pour l'intégration des données biologiques et la modélisation des systèmes biologiques. Nous présentons dans ce chapitre les principales notions et notations de la théorie des graphes nécessaires à la compréhension de ce manuscrit (Section 1.1.2), puis nous aborderons leur utilisation dynamique au travers des automates (Section 1.2). Pour une application du concept des graphes et automates sous forme de réseau métabolique, se référer à la Section 3.3.2 et pour une application aux ARN non codants, se référer à la Section 5.2.2.

1.1.1 Qu'est-ce qu'un graphe ?

La théorie des graphes est un outil de modélisation utilisé dans un grand nombre de disciplines en bioinformatique, allant de la modélisation moléculaire à la génétique des populations. Les graphes offrent une abstraction mathématique qui permet de décrire les relations entre un ensemble d'objets (par exemple : protéines, ARN, métabolites...). Un graphe consiste en un ensemble de sommets (ou nœuds) et un ensemble d'arêtes qui connectent les sommets. Les sommets servent à représenter les entités, et les arêtes à représenter les relations entre ces entités. La relation modélisée par les arêtes peut être dichotomique : présence ou pas d'une arête modélisant l'interaction ou non de deux entités, ou plus générale, dans ce cas, c'est le poids discret ou continu de l'arête entre les deux objets qui pourra être considéré. Par

exemple, pour un graphe complet de gènes modélisant leurs co-citations dans la littérature, la distance mesurant l'homologie entre ces gènes sera considérée pour l'interprétation de ce graphe et non la dichotomie des arêtes (Fig. 1.1).

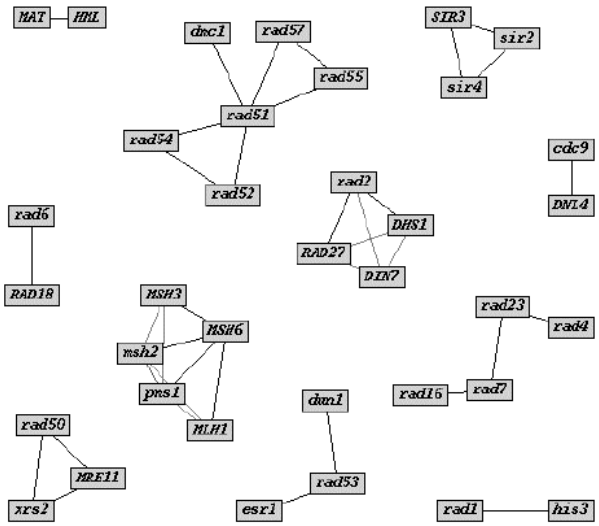


Figure 1.1 – Graphe généré par la recherche des gènes co-cité dans une publication avec le nom "DNA repair". Tous les gènes sont possiblement connectés, seule la limitation du seuil de distance est considérée pour réduire le nombre d'arêtes entre les gènes [Stapley et Benoit 2000].

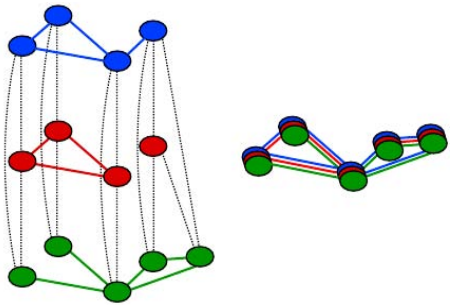
Deux formes généralisées des graphes présentent un intérêt tout particulier en biologie : les multigraphes et les hypergraphes. Les sections suivantes se proposent de les décrire brièvement, nous présenterons ensuite une application des hypergraphes et des multigraphes respectivement au travers des réseaux métaboliques (Section 3) et des réseaux d'interaction d'ARN (Section 5).

Multigraphe

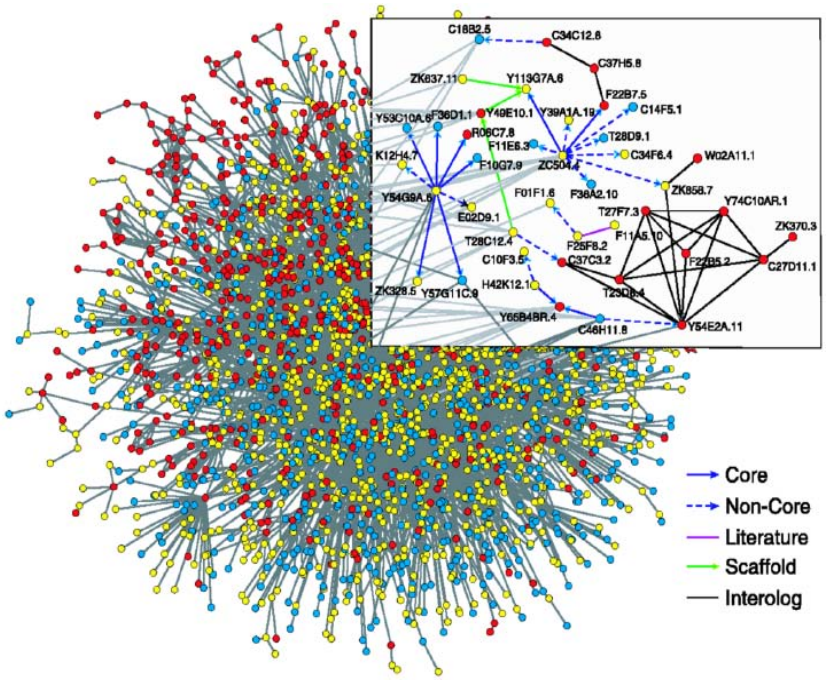
Les multigraphes sont une généralisation des graphes. Ils permettent d'avoir plusieurs ensembles d'arêtes incidentes à un même nœud. Étant donné n réseaux primaires $G_i = (V_i, E_i)$ (voir Def. 12) et une correspondance entre les nœuds de ces réseaux, l'empilement des V_i en se basant sur les ensembles E_i permet de construire le multigraphe correspondant. Dans l'exemple (Fig. 1.2a), un multigraphe est construit par l'alignement de trois ensembles d'arêtes. Les multigraphes sont très utilisés en biologie pour représenter les relations entre des entités selon différentes conditions. Chaque ensemble d'arêtes est alors employé pour modéliser les interactions provenant d'une source de données définie (littérature, banques de données, organismes...). Deux sommets peuvent ainsi être connectés par une ou plusieurs arêtes mettant en avant par ce biais, la condition dans laquelle la relation a été identifiée. On retrouve plusieurs exemples d'utilisation des multigraphes pour l'analyse comparative de deux organismes [Sharan et Ideker 2006] ou la mise en évidence de relation dans un interactome [Li *et al.* 2004] (Fig. 1.2b). Dans ce cas, la redondance de l'arête dans les différentes sources de données (littérature, expérience d'hybridation et de prédiction bioinformatique) constitue un argument en faveur de l'existence de la relation.

Hypergraphe

Les hypergraphes généralisent les graphes par la multiplication de l'incidence des arêtes. Les arêtes ne relient plus un ou deux sommets, mais un nombre quelconque de sommets, compris entre un et le nombre de sommets de l'hypergraphe (Fig. 1.3), pour une définition formelle voir (Def. 13). Les hypergraphes trouvent de très nombreuses applications en biologie où ils sont représentés sous leur forme spécialisée telle que les graphes d'interaction pour la modélisation des interactions protéine-protéine et des voies de régulation ou sous la forme de graphe biparti pour les réseaux métaboliques et interactions ARN-ARN. Pour une revue concernant l'application des hypergraphes en biologie, se référer à [Klamt *et al.* 2009].



(a) Un multigraphe



(b) Interactome de *Caenorhabditis elegans*

Figure 1.2 – Un exemple de multigraphe et un exemple d'utilisation en biologie.
(a) À gauche, trois réseaux (bleu, rouge et vert) sont représentés avec leur relation de correspondance représentée par des pointillé. À droite, le multigraphe d'alignement correspondant. Il correspond à l'empilement des trois réseaux [Denielou 2010]. (b) Un exemple de multigraphe issu de l'interactome de Caenorhabditis elegans [Li et al. *2004]. Les sommets correspondent à des protéines et les arêtes correspondent à leur interaction selon différentes sources de données.*

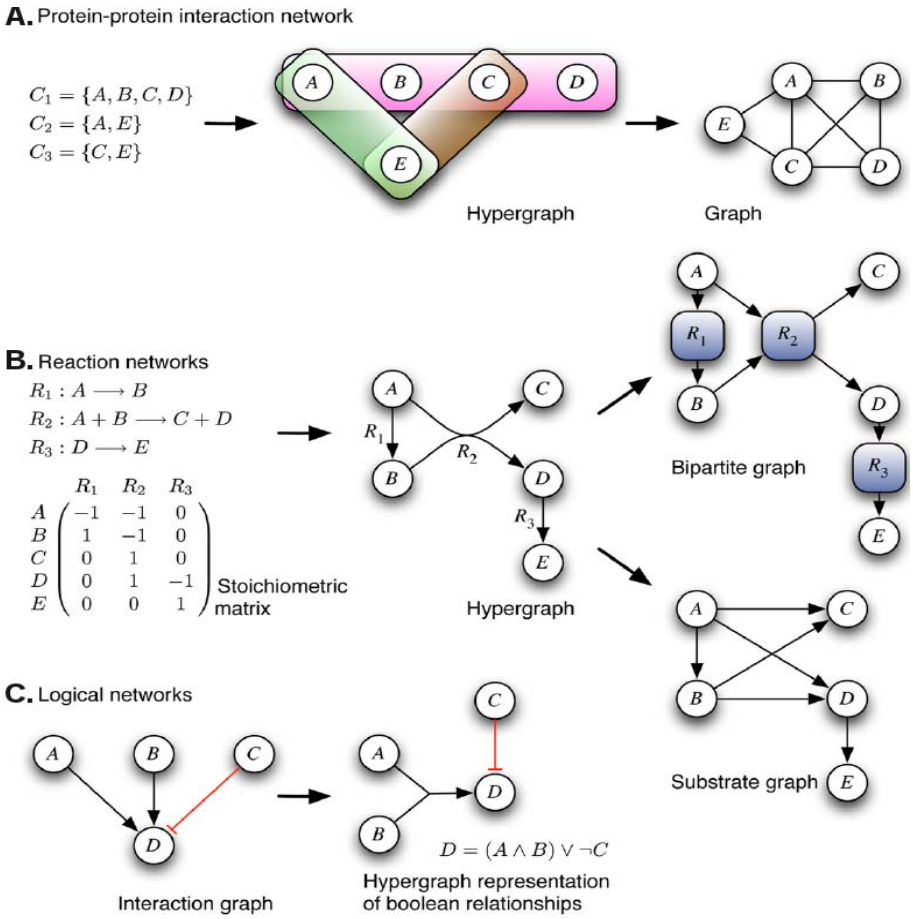


FIGURE 1.3 – Exemples d'utilisation des hypergraphes et des hypergraphes orientés en biologie [Klamt *et al.* 2009].
(A) Hypergraphe modélisant des interactions entre protéines. (B) Hypergraphes modélisant des réseaux métaboliques. (C) Hypergraphes modélisant des interactions.

1.1.2 Définition et notations

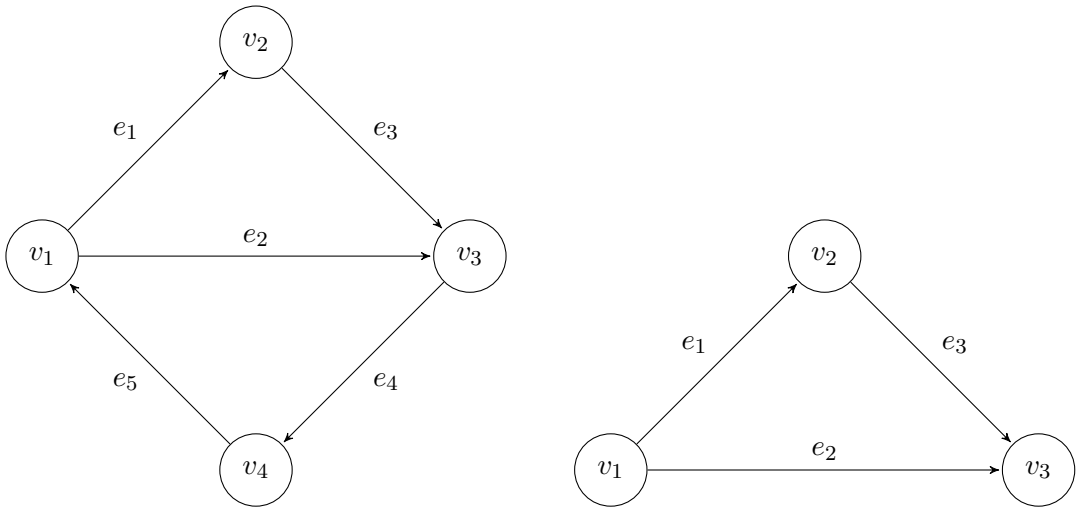
Les définitions suivantes sont issues de [Gross et Yellen 2004]. Nous commençons par une définition formelle des termes graphe, sommet et arête.

Définition 1 (**Graphe non orienté**). *Un graphe $G = (V, E)$ est composé d'un ensemble de sommets, également appelé nœuds, noté V et d'un ensemble d'arêtes, noté E. Soit $e = (u, v)$, une arête, on appelle les sommets u et v, les extrémités de l'arête e.*

Définition 2 (**Graphe orienté**). *Un graphe $G = (V, E)$ est composé d'un ensemble de sommets, également appelé nœuds, noté V et d'un ensemble d'arcs, noté E. Soit $a = (u, v)$, un arc, on appelle respectivement les sommets u et v, la source et la destination de l'arc a.*

Définition 3 (**Sous-graphe**). *Soit un graphe $G' = (V', E')$, G' est un sous-graphe de G si et seulement si G' est un graphe, $V' \subseteq V$ et $E' \subseteq E$.*

Un graphe et un de ses sous-graphes sont donnés en exemple (FIG. 1.4).



(a) Un graphe orienté G (b) Un sous-graphe de G

FIGURE 1.4 – Un graphe orienté et un de ses sous-graphes.
(a) Graphe orienté $G = (V, E)$ avec $V = v_1, v_2, v_3, v_4$ et $E = e_1, e_2, e_3, e_4, e_5$ (b) Sous-graphe G' de G avec $V' = v_1, v_2, v_3$ et $E' = e_1, e_2, e_3$.

Définition 4 (**Adjacence de sommet**). *Un sommet u est adjacent au sommet v s'ils sont liés par une arête.*

Définition 5 (**Incidence**). *Si le sommet v est associé à l'arête e, alors v est incident à e et e est incident à v.*

Définition 6 (**Voisinage d'un sommet**). *Soit un graphe $G = (V, E)$, $u \in V$. On appelle voisinage de u dans G, noté $N_G(u)$, l'ensemble $\{v | \{u, v\} \subseteq E\}$.*

Définition 7 (**Degré**). *Le degré d'un sommet v dans un graphe G, noté $deg(v)$, est le nombre d'arête incident à v.*

L'ajout d'attribut aux graphes permet d'aller plus loin que la simple étude de sommets et d'arêtes.

Définition 8 (**Attribut de sommet**). *Un attribut de sommet est une fonction de l'ensemble sommet à un ensemble de valeurs possibles d'attribut.*

Définition 9 (**Attribut d'arête**). *Un attribut d'arête est une fonction de l'ensemble arête à un ensemble de valeurs possibles d'attribut.*

Définition 10 (**Chemin dans un graphe**). *Un chemin dans un graphe $G = (V, E)$ est défini par une suite finie de sommets et d'arcs issus de V et E consécutifs, telle que pour tout arc e_i appartenant au chemin, v_i et v_{i+1} sont des extrémités de e_i. La longueur du chemin dépend du nombre d'arcs qui le composent.*

Définition 11 (**Graphe connexe**). *Un graphe est connexe si quels que soient les sommets u et v du graphe G, il existe un chemin de u vers v ou de v vers u.*

Définition 12 (**Multigraphe**). *Un multigraphe (ou pseudographe) $\mathcal{M} = (V, E, f)$ est composé d'un ensemble de sommets notés V, de plusieurs ensembles d'arêtes, notés E et une fonction $f \in E \to \mathbb{P}_2(V)$, f indiquant ici le sommet auquel est connecté chaque type d'arête.*

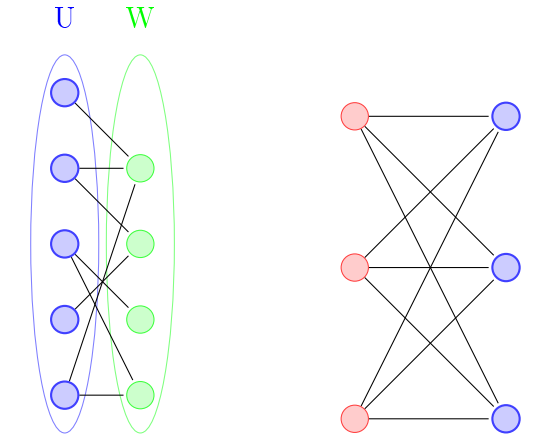
Dans ce mémoire, nous nous sommes plus particulièrement intéressé aux hypergraphes et à leur modélisation sous la forme de graphes bipartis.

Définition 13 (**Hypergraphe**). *Un hypergraphe $\mathcal{H} = (V, \mathcal{F})$ a un ensemble de sommets notés V et un ensemble d'arêtes notés $\mathcal{F}$ qui consiste en plusieurs sous-ensembles de E. On suppose que $\mathcal{F} \neq \emptyset$ et que $|F| \geq 2$ pour tout $F \in \mathcal{F}$.*

Définition 14 (**Graphe biparti**). *Un graphe ou hypergraphe $G = (V, E)$ est un graphe biparti, si les nœuds du graphe peuvent être divisés en deux sous-ensembles U et W tels que chaque arête ait une extrémité dans U et l'autre W.*

Définition 15 (**Graphe biparti complet**). *Un graphe biparti complet est un graphe biparti dans lequel tous les sommets d'une partition sont adjacents à tous les sommets de l'autre partition.*

Les deux graphes (Fig. 1.5) sont des graphes bipartis.



(a) Graphe biparti

(b) Graphe biparti complet

Figure 1.5 – Deux graphes bipartis.
(a) Un graphe biparti à 10 sommets. (b) Un graphe biparti complet à 6 sommets.

1.2 Automates

Une utilisation courante des graphes en biologie est réalisée via les automates. Ils sont notamment utilisés pour l'analyse lexicale des séquences et l'analyse de la dynamique des interactions des éléments biologiques (comportement des différents réseaux de la cellule : réseau de gènes, réseau métaboliques...). Nous abordons dans cette section le fonctionnement des automates (Section 1.2.1) puis l'application qui en a été réalisée durant cette thèse sous la forme de réseau de Petri (Section 1.2.2).
Les définitions ici présentes sont issues de [Sakarovitch 2003]. Nous utiliserons le formalisme et des exemples issues de la théorie des langages et automates pour décrire de manière formelle les automates.

1.2.1 Qu'est-ce qu'un automate ?

Les automates sont des objets mathématiques, qui permettent de modéliser des systèmes. Un automate est constitué d'un ensemble d'états du système, reliés entre eux par des tran-

sitions représentant les évènements. Un exemple d'automate simple est celui des automates finis :

Définition 16 (**Automate fini**). *Un automate fini $\mathcal{A}$ est un quintuplé $(Q, \Sigma, \delta, I, T)$ où :*
- *Q est l'ensemble fini d'états de $\mathcal{A}$*
- *Σ est l'alphabet de $\mathcal{A}$*
- *$\Delta : Q \times \Sigma \to Q$ est la fonction de transition*
- *$I \subseteq Q$ est l'ensemble des états initiaux de $\mathcal{A}$*
- *$T \subseteq Q$ est l'ensemble des états finaux de $\mathcal{A}$*

On représente classiquement un automate par un graphe orienté et étiqueté par les lettres de l'alphabet Σ (Fig. 1.6). Plus précisément, les sommets du graphe sont les états de l'automate et les arcs représentent les transitions de l'automate. Les états initiaux sont désignés par une flèche entrante et les états finaux par un double cercle. Un triplet (q_1, b, q_2) appartenant à l'automate $\mathcal{A}$ sera noté $q_1 \xrightarrow{b} q_2$.

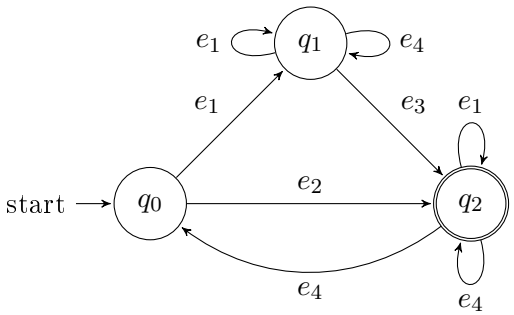


Figure 1.6 – Un automate fini $\mathcal{A}$. On a pour l'automate $\mathcal{A}$: $Q = \{q_0, q_1, q_2\}$, Σ=$\{e_1, e_2, e_3, e_4\}$, I=$\{q_0\}$ et T=$\{q_2\}$

Un automate ainsi décrit peut être utilisé comme analyseur syntaxique qui accepte ou rejette une chaine de caractères donnée selon les règles décrites ci-dessous.

Définition 17 (**Chemin dans un automate fini**). *Un chemin π dans un automate fini est une suite de transitions $\delta_1, \delta_2, ..., \delta_k \in \Delta$ consécutives. Ainsi, pour tout $i \in [1, n-1]$, l'état final de la transition e_i est l'état initial de la transition e_{i+1}.*

Définition 18 (**Reconnaissance d'un mot par un automate**). *Un mot m sur l'alphabet Σ est reconnu par un automate $\mathcal{A} = (Q, \Sigma, \delta, I, T)$, s'il existe un chemin de I vers T tel que la séquence des étiquettes des arcs du chemin forme le mot. On dit également que l'automate accepte le mot m. L'ensemble des mots reconnus par un automate forme le langage reconnu par cet automate.*

Par exemple, le mot 'abaa' est reconnu par l'automate $\mathcal{B}$ (Fig. 1.7) car il existe un chemin dans le graphe, partant de l'état initial et aboutissant à l'état d'acceptation tel que la concaténation des symboles étiquetant les arcs du chemin ABABB est égale au mot.

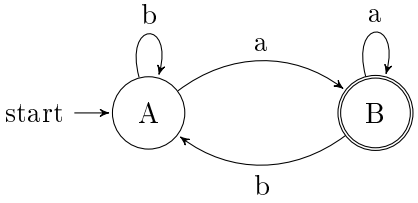


FIGURE 1.7 – Automate fini $\mathcal{B}$.

Les différentes classes d'automates

Les automates finis sont répartis en deux classes, les automates finis déterministes (AFD) et les automates finis non-déterministes (AFN). Les automates finis déterministes se différencient des automates non-déterministes par la fonction de transition. Dans le cas d'un automate déterministe, la fonction de transition associe un état à un couple composé d'un état et d'un symbole $\delta : Q \times \Sigma \to Q$ alors que pour un automate non déterministe, cette fonction associe un ensemble d'états à un couple composé d'un état et d'un symbole. L'automate peut donc transiter à partir d'un état q_0 et sur un symbole a vers plusieurs états. La fonction de transition est par conséquent définie de la façon suivante : $\delta : Q \times \Sigma \to \mathcal{P}(\mathcal{Q})$ où $\mathcal{P}(\mathcal{Q})$ est l'ensemble des parties de Q.
Un automate fini déterministe peut donc être décrit de la manière suivante :

Définition 19 (**Automate fini déterministe**). *Un automate fini déterministe $\mathcal{A}$ est un automate fini où :*
- *L'ensemble I n'est composé que d'un seul état initial*
- *Il n'existe pas deux transitions de même étiquette issues d'un même état*

Ainsi, le déterminisme d'un automate se traduit, dans la représentation graphique, par le fait qu'il ne peut y avoir plus d'un arc possédant la même étiquette émanant d'un même état. Pour une configuration de l'AFD, il n'existera donc au plus qu'un seul mouvement possible, tandis que l'AFN présentera des configurations pour lesquelles plus d'un mouvement est possible. Nous pouvons utiliser pour illustrer cela, la reconnaissance du suffixe ab d'un mot par un automate.
L'automate fini $\mathcal{C}$ décrit ci-dessous reconnait les mots construits sur l'alphabet $\Sigma = \{a, b\}$ avec le suffixe ab, tel que :
- $\mathcal{Q} = \{q_0, q_1, q_2\}$;
- $\Sigma = \{a, b\}$;
- $\delta_1(q_0, a) = \{q_0, q_1\}, \delta_2(q_0, b) = \{q_0\}, \delta_3(q_1, b) = \{q_2\}$;
- $I = q_0$;
- $T = \{q_2\}$.

La figure 1.8 représente le graphe de cet automate. L'existence des transitions (q_0, a, q_0) et (q_0, a, q_1) indique qu'il est non déterministe.

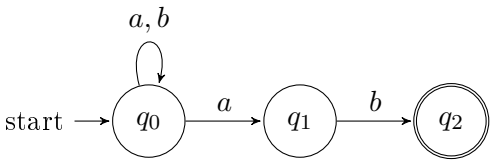


FIGURE 1.8 – Automate fini $\mathcal{C}$ reconnaissant le langage $\{a, b\}$ finissant par ab.

La reconnaissance du mot *aaabab* par cet automate est représentée dans la figure 1.9. Partant de la configuration $(q_0, aaabab)$, l'automate effectue en parallèle deux mouvements qui le mènent vers les configurations $(q_0, aabab)$ et $(q_1, aabab)$. La configuration $(q_1, aabab)$ ne permet aucun mouvement, ce processus s'arrête. La configuration $(q_0, aabab)$ mène à $(q_0, abab)$ et $(q_1, abab)$...

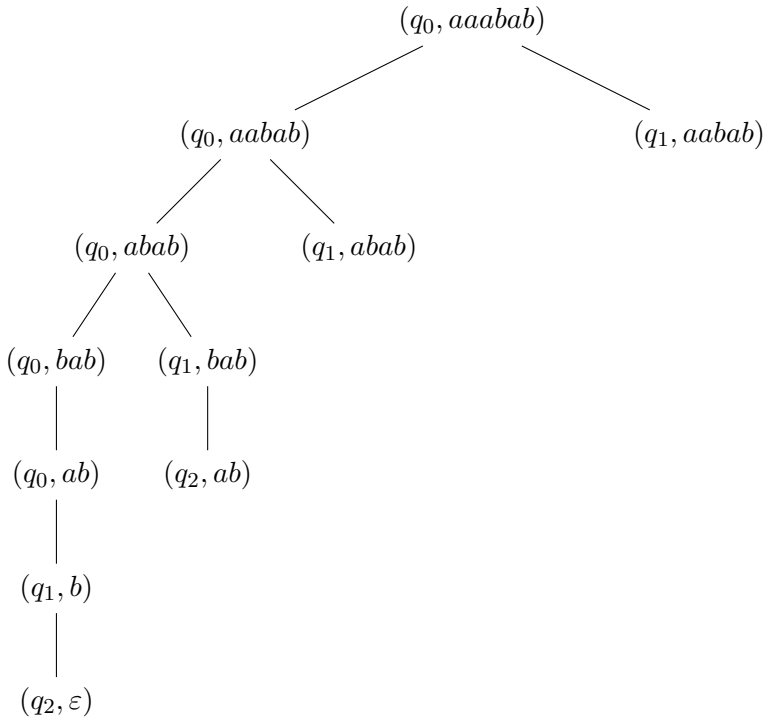


FIGURE 1.9 – Reconnaissance non déterministe du mot *aaabab* par l'automate $\mathcal{C}$

Comme observé ci-dessus, l'utilisation d'un automate non déterministe est compliquée car il est nécessaire de dédoubler le processus de lecture de façon à poursuivre tous les chemins possibles. Dans ces conditions, il est souvent préférable de définir un AFD équivalent, qui reconnait comme précédemment les mots suffixés par *ab* (FIG. 1.10). Soit l'automate $\mathcal{C}'$:

- $\mathcal{Q} = \{q_0, q_1, q_2\}$;
- $\Sigma = \{a, b\}$;
- $\delta_1(q_0, a) = \{q_1\}$, $\delta_2(q_0, b) = \{q_0\}$, $\delta_3(q_1, a) = \{q_1\}$, $\delta_4(q_1, b) = \{q_2\}$, $\delta_5(q_2, a) = \{q_1\}$, $\delta_6(q_2, b) = \{q_0\}$;
- $I = q_0$;
- $T = \{q_2\}$.

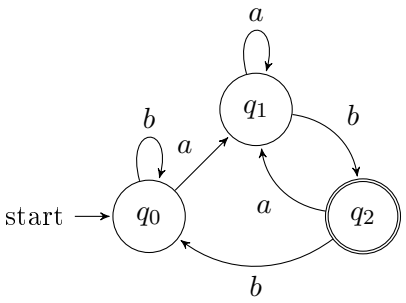
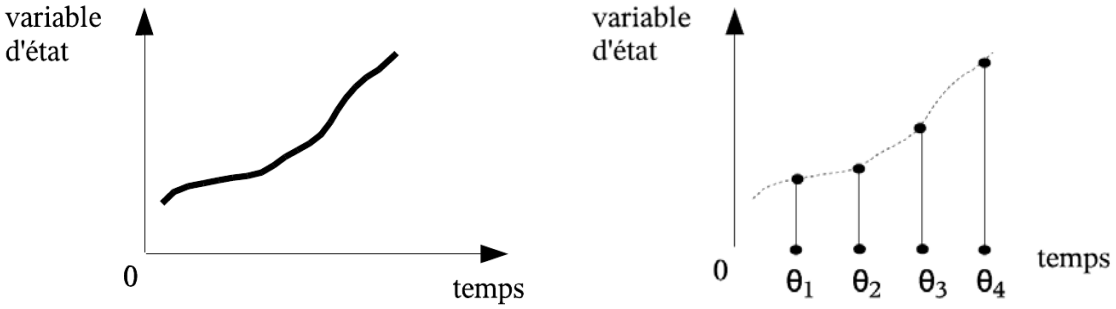


FIGURE 1.10 – Automate fini $\mathcal{C}'$ reconnaissant le langage $\{a, b\}$ finissant par ab.

Dynamique des automates

Les automates peuvent être utilisés pour modéliser un système dynamique. Ces systèmes sont composés d'entités homogènes ou hétérogènes dont les interactions évoluent constamment au cours du temps. Pour les modéliser, il est nécessaire de revoir la fonction de transition de l'automate qui va alors exprimer l'ensemble des évènements susceptibles de le faire changer d'état. Cinq types de systèmes peuvent être identifiés selon le type de variables modélisées et selon leur évolution au cours du temps.
Les systèmes continus et échantillonnés sont utilisés pour modéliser des environnements où les variables d'état sont continues (FIG. 1.11). Les variables d'état continues sont souvent employées en biologie pour quantifier des grandeurs physiques d'éléments (concentration d'un métabolite, poids de cellule...). Ces variables prennent leur valeur dans le domaine des réels $\mathbb{R}$ et leur évolution est souvent représentée par un système d'équations différentielles. Les systèmes continus permettent de modéliser les cas où toutes les variables du modèle sont continues, y-compris le temps, qui est représenté par une variable continue (temps dense) (FIG. 1.11a), tandis que les systèmes échantillonnés vont représenter les variables d'états dans un système où le temps est représenté par une suite d'instant $\theta_1, \theta_2, ..., \theta_n$ (FIG. 1.11b). Ces modèles sont notamment utilisés pour la modélisation d'évènements à l'échelle cellulaire.

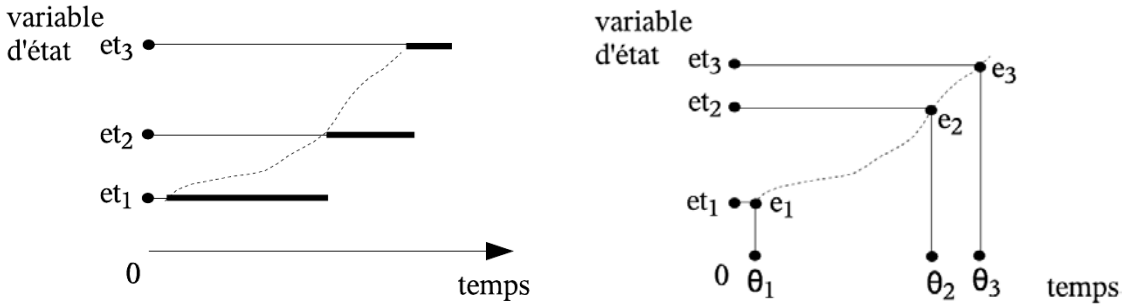


(a) Système continu (b) Système échantillonné

FIGURE 1.11 – Deux systèmes modélisant des variables d'état continues [Valette 2007].
(a) Un système continu représentant une variable d'état et un temps continus. (b) Un système échantillonné représentant une variable d'état continue avec un temps échantillonné à une suite d'instants $\theta_1, \theta_2, \theta_3, \theta_4$.

Les systèmes discrets ou à évènements discrets permettent quant à eux de modéliser des variables d'état discrètes. Ces variables prennent leur valeur dans le domaine des entiers na-

turels $\mathbb{N}$. Ces systèmes représentent respectivement un temps continu associé à une représentation discrète des états (Fig. 1.12a) et un temps représenté par une suite d'évènements (Fig. 1.12b). Les systèmes à évènement discret ont une importance cruciale notamment pour la recherche [Navarro et Raffinot 2002], l'analyse statistique [Lladser *et al.* 2008], et la découverte de motif [Tompa *et al.* 2005] (pour revue voir [Marschall 2011]).



(a) Système discret (b) Système à évènement discret

Figure 1.12 – Deux systèmes modélisant des variables d'état discrètes [Valette 2007].
(a) Un système discret modélisant une variable d'état discrète $et1, et2, et3$ *dans un temps continu. (b) Un système discret modélisant une variable d'état discrète* $et1, et2, et3$ *dans un temps échantillonné* $\theta_1, \theta_2, \theta_3$.

Enfin les systèmes hybrides comprennent à la fois des variables d'état continues et des variables d'état discrètes et permettent de manipuler à la fois du temps continu et du temps discret sous la forme d'un ensemble d'évènements. Ils sont notamment utilisés pour représenter des évènements se produisant à différentes échelles dans la cellule, tel que l'état discret d'une cellule combiné aux flux continu des réactions (pour revue [Casagrande *et al.* 2005]).

1.2.2 Réseaux de Petri

L'utilisation des automates pour modéliser le fonctionnement dynamique des réseaux biologiques doit faire face, de nos jours, à la nature complexe de certains de ces systèmes (pour revue voir [Fromm 2004]). Les réseaux métaboliques sont composés de nombreuses entités interagissant les unes avec les autres, créant ainsi un grand ensemble d'interconnexions. Cet ensemble d'interactions génère sans aucune application d'un principe global, des phénomènes globaux difficiles à prédire. La modélisation de ces systèmes en utilisant des automates classiques pose de nombreux de problèmes liés à leur représentation limitée. En effet lorsque l'on considère un réseau métabolique à l'échelle du métabolisme cellulaire (Fig. 1.13), il est difficile avec ces derniers d'identifier des sous-systèmes mais aussi de gérer l'explosion combinatoire du nombre d'états. Deux évolutions des automates ont été réalisées pour représenter des réseaux de manière décomposée : les automates cellulaires et les réseaux de Petri. Nous nous intéresserons plus particulièrement dans cette thèse à cette dernière modélisation très utilisée en biologie.

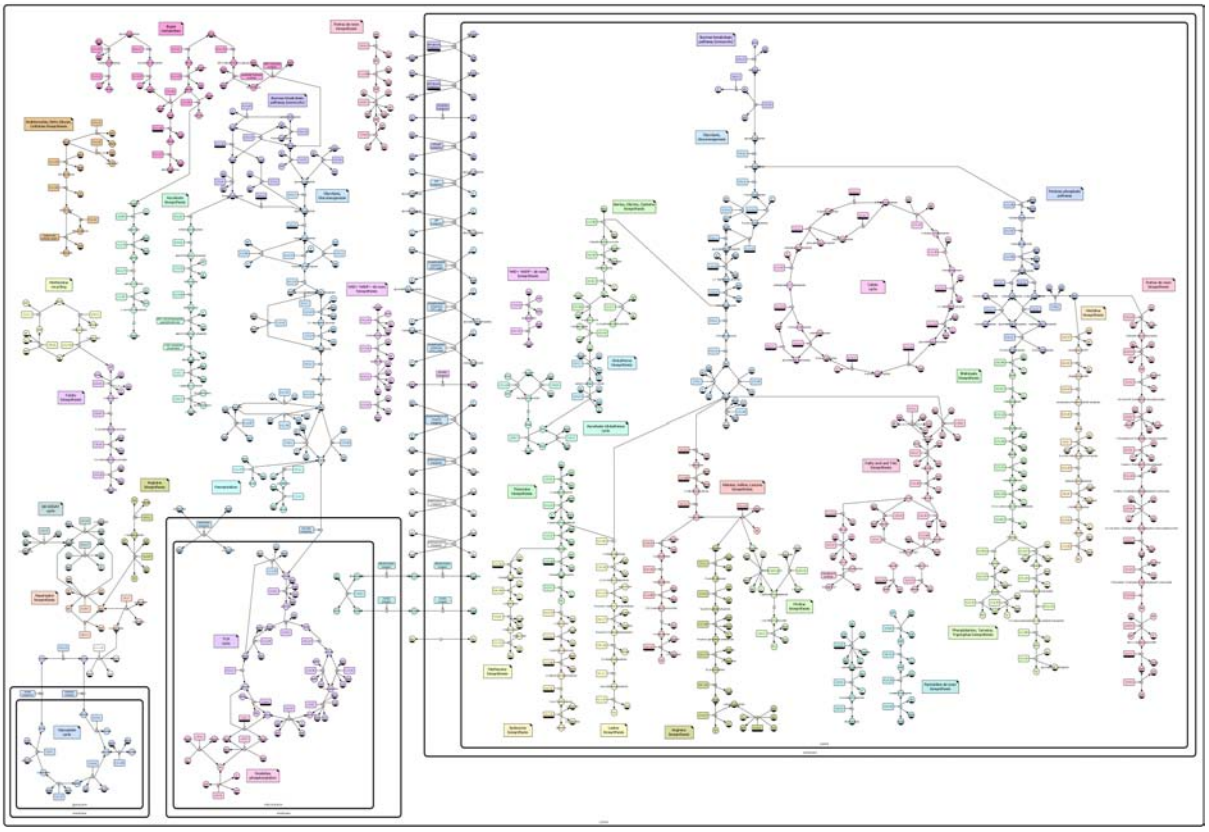


FIGURE 1.13 – Représentation du métabolisme central des plantes (Illustration issue de : SBGN competition 2010). La représentation de ce système par un modèle d'automate classique présenterait une explosion combinatoire du nombre d'évènements possibles.

Principe

Les définitions présentées ici sont issues de [Heiner *et al.* 2008] et [Koch *et al.* 2011] ; pour avoir plus d'informations sur les éléments que nous modélisons avec les réseaux de Petri, se référer à la Section 2.1.
Les réseaux de Petri (PN) ont été formulés par [Petri 1962] avec l'objectif de définir un formalisme mathématique pour représenter et analyser des systèmes causaux avec concurrence. Le formalisme proposé se présente sous la forme d'un graphe biparti pondéré et orienté qui comporte deux types de nœuds : des places et des transitions. Les places, représentées graphiquement par des cercles, sont associées à des états du système et les transitions, représentées par des rectangles, décrivent les changements d'état. Des jetons sont aussi présents au titre d'objets discrets servant à représenter la dynamique du système (FIG. 1.14). Leur répartition sur le réseau correspond au marquage du réseau de Petri. On peut donc définir les réseaux de Petri de la manière suivante :

Définition 20 (**Réseau de Petri**). *Un réseau de Petri est un 4-uplet* $\mathcal{N} = (P, T, f, m_0)$ *où :*
- P *et* T *sont des ensembles finis, non vides et disjoints où* P *est l'ensemble des places et* T *est l'ensemble des transitions.*
- $f : ((P \times T) \cup (T \times P)) \to \mathbb{N}$ *définit l'ensemble des arcs avec leurs multiplicités.*
- $m_0 : P \to \mathbb{N}$ *est le marquage initial du réseau de Petri où pour chaque place* $p \in P$*, il y a* $n \in \mathbb{N}$ *jetons.*

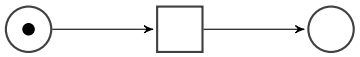


FIGURE 1.14 – Réseau de Petri comportant deux places, une transition et un jeton.

La première utilisation des réseaux de Petri en biologie a été réalisée par [Reddy *et al.* 1993], qui proposa de les utiliser pour représenter les voies métaboliques. L'applicabilité de ces réseaux pour la modélisation des réseaux métaboliques a depuis été démontrée et étendue par différents auteurs [Koch *et al.* 2005; Heiner *et al.* 2008].

Dans ce formalisme, les places servent à modéliser, de manière générale, tous les différents composants de la cellule, tels que les métabolites, les composants chimiques, ainsi que tout élément servant de précurseur ou de produit à une réaction. Les transitions jouent un rôle actif dans les réseaux de Petri puisqu'elles modélisent toutes les actions se produisant dans la cellule, telles que les réactions chimiques, les associations protéine-ligand ou encore les transports entre les compartiments cellulaires. La quantité des éléments associés aux transitions (enzyme, transporteur...) est considérée, par défaut, comme étant en quantité suffisante pour la transition. En conséquence, pour limiter cette quantité, il est nécessaire d'ajouter des places modélisant ces éléments. Les arcs du réseau de Petri sont orientés depuis les places représentant les précurseurs vers les réactions, puis des réactions aux produits. Ainsi les "pré"-places, notées $\bullet t$, vont correspondre aux précurseurs et les "post"-places, notées $t\bullet$, vont correspondre aux produits des transitions. Le poids des arcs, aussi appelé multiplicité de l'arc, correspond à la stœchiométrie de la réaction. Un arc avec un poids de 1 correspond à la valeur par défaut des arcs, bien que non indiqué. Enfin, chaque place contient un nombre donné de jetons. Ces jetons représentent une quantité du composant exprimée en nombre de molécules ou dans une unité mesurant la quantité de matière de l'élément (mole, gramme...). Les jetons sont représentés par des points noirs ou des chiffres. Ainsi la réaction de phosphorylation du glucose Gluc + ATP $\rightarrow$ G6P + ADP est décrite par un réseau de Petri de la manière suivante :

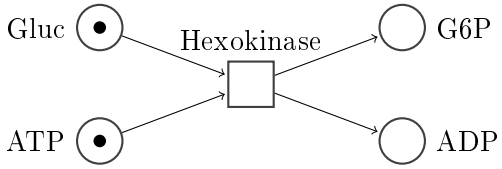


FIGURE 1.15 – Réaction de phosphorylation du glucose. On a pour ce réseau de Petri P={Gluc, ATP, G6P, ADP}, T={Hexokinase}, m_0(Gluc)=1 et m_0(ATP)=1.

Dynamique d'exécution

Les réseaux de Petri sont des automates finis non-déterministes, dont la dynamique repose sur l'évolution du marquage du réseau. Ce processus consiste en la réalisation du franchissement de la transition (aussi appelé tir) selon deux étapes : l'évaluation de la précondition et le tir lui-même. Nous pouvons décrire ce processus de la manière suivante :

Définition 21 (Franchissement d'une transition). *Soit un réseau de Petri* $\mathcal{N} = (P, T, f, m_0)$ *:*

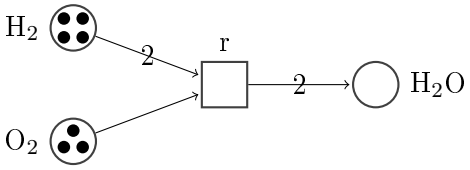
- *Une transition t de* $\mathcal{N}$ *est activée dans un marquage m, si* $\forall p \in \bullet t : m(p) \geq f(p,t)$*, sinon la transition reste inactive.*
  f(p,t) décrit le nombre de jetons retirés des places p lors d'une occurrence de t. Le nombre de jetons à retirer correspond à la multiplicité de l'arc. Inversement, f(t,p) décrit le nombre de jetons ajoutés aux places p .

- *Une transition t activée peut être tirée.*
- *Quand t de m est tirée, un nouveau marquage m' est atteint avec $\forall p \in P : m'(p) = m(p) - f(p,t) + f(t,p)$.*
- *Le tir se produit de manière automatique et atomique, sans consommer de temps.*

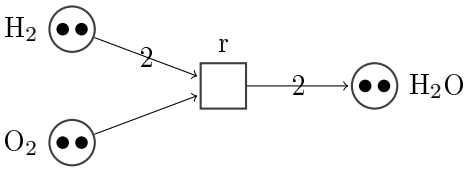
Nous pouvons illustrer ce processus par l'exemple fourni par [Murata 1989] pour modéliser la synthèse de l'eau :

$$2H_2 + 0_2 \rightarrow 2H_2O$$

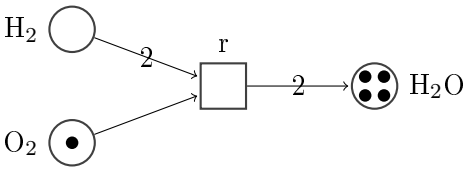
Cette réaction requiert 2 moles de dihydrogène et une mole de dioxygène pour produire 2 moles d'eau. Nous pouvons représenter cette réaction sous la forme d'un réseau de Petri de la manière suivante :



Ce réseau présente une réaction qui peut être réalisée, son occurrence nous permet d'atteindre ce premier état intermédiaire :



Puis, lors d'une seconde occurrence de la transition, on obtient l'état suivant :



Les jetons ont donc été déplacés entre les places connectées à r suivant le sens des arcs. La dynamique se poursuit jusqu'à ce que plus aucune transition ne soit activable. Ce dernier état correspond à l'état final du réseau.

Activités parallèles

Les réactions parallèles (ou concurrentes) peuvent être exprimées par les réseaux de Petri. Par exemple, les réactions de la phosphoglucose isomérase (t_1) de la glycolyse et de la glucose-6-phosphate déshydrogénase (t_2) de la voie des pentoses phosphates se produisent en même temps dans la cellule et consomment toutes deux du glucose-6-phosphate (G6P), tel que :

$$\text{G6P} \xrightarrow{t_1} \text{F6P}$$

et

$$\text{G6P} + 2\text{NADP}^+ \xrightarrow{t_2} \text{Ru5P} + 2\text{NADPH}$$

La composition de ses deux réactions peut être exprimée sous la forme du réseau de Petri. On trouve ci-après un exemple :

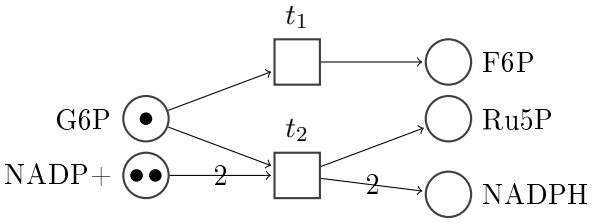


FIGURE 1.16 – Le réseau de Petri modélisant la réaction de la Glucose-6-phosphate isomérase et Glucose-6-phosphate déshydrogénase.

Dans cet état (FIG. 1.16), le jeton de G6P peut être consommé par la réaction t_1 ou la réaction t_2. Le réseau de Petri n'explicite pas l'évolution de la dynamique à tenir, il est nécessaire de faire un choix pour savoir quelle réaction tirer. Cette structure de réseau présente donc une concurrence entre t_1 et t_2 pour laquelle un choix non-déterministe doit être effectué.

Éléments remarquables des réseaux de Petri

Différentes particularités topologiques peuvent être identifiés dans les réseaux de Petri. Nous définissons ici quelques notions spécifiques aux réseaux de Petri que nous devrons considérer dans la définition de notre simulateur de réseau de Petri (voir Section 3.3.2). Soit un réseau de Petri $\mathcal{N}$:

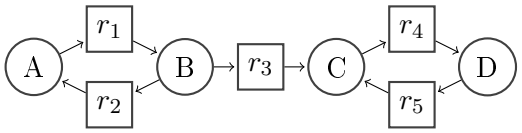


FIGURE 1.17 – Réseau de Petri $\mathcal{N}$ [Nabli 2011].

Piège (*Trap*) Un piège désigne un sous-ensemble $\mathcal{Q}$ des places d'un réseau de Petri $\mathcal{N}$ où chaque transition, ayant une place de $\mathcal{Q}$ en entrée, n'a que des places de $\mathcal{Q}$ en sortie. Par exemple, les places C et D de $\mathcal{N}$ (FIG. 1.17) forment un piège. Les transitions r_4 et r_5 qui enlèvent un jeton de C ou D, ajoutent aussi un jeton à C ou D. Cette régularité structurale implique une préservation des jetons de $\mathcal{Q}$. Ainsi si une place de $\mathcal{Q}$ est marquée, $\mathcal{Q}$ restera marqué pour tous les marquages suivants de $\mathcal{N}$ et ne perdra jamais ses jetons.

Siphon Un siphon désigne un sous-ensemble $\mathcal{S}$ des places d'un réseau de Petri $\mathcal{N}$, où chaque transition, ayant une entrée dans $\mathcal{S}$, a au moins une sortie dans $\mathcal{S}$. Par exemple, les places A et B de $\mathcal{N}$ (FIG. 1.17) forment un siphon. Les transitions r_1 et r_2 qui enlèvent un jeton de A ou B, peuvent ajouter un jeton à A ou B. Cependant, une fois vides de jetons, les places de $\mathcal{S}$ ne regagneront jamais de jetons.

Impasse (*Deadlock*) Un réseau de Petri $\mathcal{N}$, ayant un marquage initial m_0, est une impasse si plus aucune transition ne peut être activée à un marquage m' issu de m_0.

Vivacité (*Liveness*) Une transition t est en vie si pour chaque marquage m' issu de m_0, il y a une série de transitions permettant de l'activer. Un réseau de Petri $\mathcal{N}$ est vivant, si toutes les transitions de $\mathcal{N}$ sont en vie. À l'inverse, le réseau de Petri $\mathcal{N}$ atteint son état final si aucune transition de $\mathcal{N}$ n'est en vie.

Infinité Il est possible d'indiquer dans un réseau de Petri, des sources et des puits capables respectivement de produire ou consommer des jetons en nombre infini. Ces sources sont représentées par des transitions ou parfois par des places (Fig. 1.18).

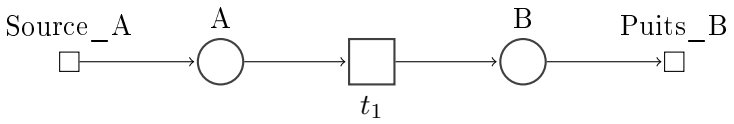


Figure 1.18 – Réseau de Petri présentant une structure infinie.

1.2.3 Extensions des réseaux de Petri

La dynamique des réseaux de Petri, comme vu précédemment, permet de simuler de manière intuitive le flux métabolique des molécules dans la cellule. Cependant, cette première définition est insuffisante pour intégrer dans la modélisation les principes physiques qui dirigent le métabolisme (voir Section 2.1.2). Aussi, différentes extensions aux réseaux de Petri ont été réalisées pour mieux modéliser ce processus. Nous détaillons ici les deux principales extensions existantes pour la modélisation continue et la modélisation stochastique. Enfin, nous présentons en Section 3.3.2 le système de modélisation que nous avons développé à partir des réseaux de Petri stochastiques généralisés.

Modélisation continue

La modélisation continue a pour objectif d'expliquer et de prédire la quantité des métabolites d'un système dynamique. Ces modèles s'attachent principalement à l'intégration des données cinétiques des réactions chimiques modélisées (voir Section 2.3.1). Les réseaux de Petri continus [David et Alla 1987] représentent la principale implémentation de ce type de modélisation dans les réseaux de Petri.

Réseaux de Petri continus Les réseaux de Petri continus sont définis par un système d'équations différentielles ordinaires où le système de marquage et de transitions des réseaux de Petri sont modifiés. Le marquage des places ne correspond plus à un entier mais correspond à un réel positif, appelé valeur du jeton (*token value*), qui est interprétée comme la concentration de la molécule modélisée par la place. Les transitions modifient cette valeur selon un taux de tir (*reaction rate*) correspondant à la vitesse de la réaction et sont tirées de manière instantanée pour simuler un flot continu. Ceci peut être formalisé par la définition suivante issue de [Heiner *et al.* 2008] :

Définition 22 (**Réseau de Petri continu**). *Un réseau de Petri continu est un quintuplé* (P, T, f, v, m_0) *où :*

- P *et* T *sont des ensembles finis, non vides et disjoints.* P *est l'ensemble des places et* T *est l'ensemble des transitions.*
- $f : ((P \times T) \cup (T \times P)) \to \mathbb{R}^+$ *définit l'ensemble des arcs.*

- $v : T \to H$ *est une fonction qui assigne une fonction de taux de tir* h_t *à chaque transition* t*, tel que :*
  $H = \cup_{t \in T} \{h_t | h_t : \mathbb{R}^{|\bullet t|} \to \mathbb{R}\}$
  l'ensemble de toutes les fonctions de tir, et $v(t) = h_t$ *pour toutes les transitions* $t \in T$
- $m_0 : P \to \mathbb{R}^+$ *est le marquage initial du réseau de Petri.*

La fonction décrivant le taux de tir des réactions, h_t, correspond généralement à une cinétique de *mass-action* relative à une équation de Michaelis-Menten ou de Hill (voir Section 2.1.2). La variation du marquage d'une place au cours du temps est donnée, quant à elle, par la fonction (Eq. 1.1) :

$$\frac{dM}{dt} = \sum_{t \in \bullet p} f(t,p)v(t) - \sum_{t \in \bullet p} f(p,t)v(t) = U \times v(t), \ avec \ M = m(p). \tag{1.1}$$

où le marquage de chaque place m dépend de la stœchiométrie ($f(t,p)$ et $f(p,t)$) et du taux de tir $v(t)$ des transitions t produisant ou consommant p au cours du temps. On peut ainsi exprimer le marquage par la fonction (Eq. 1.2) :

$$m(t) = m_0 + \int_0^t U \times v(u) \times du \tag{1.2}$$

où le marquage du réseau de Petri $M(t)$ ne représente plus de cette manière un état discret mais un changement continu au cours du temps depuis l'état initial M_0.
Les réseau de Petri continus intègrent ainsi l'ensemble des principes mathématiques décrivant les réactions chimiques. Leur utilisation est cependant aujourd'hui limitée par le manque de données cinétiques sur les réseaux métaboliques, seules quelques voies métaboliques sont suffisamment détaillées pour effectuer ce type de modélisation. Un second type de modélisation s'est donc imposé pour pallier ces contraintes : la modélisation stochastique.

Modélisation stochastique

La modélisation stochastique est basée sur la méthode de simulation des réactions chimiques décrites par [Gillespie 1976, 1977]. Elle propose de simuler de manière simplifiée l'évolution dans le temps d'un système chimique réactionnel grâce à trois hypothèses fondamentales. Premièrement, le système est considéré comme étant à l'équilibre thermique, ce qui signifie que la température est identique en tout point du milieu pour toutes les réactions. Deuxièmement, la concentration des métabolites est supposée homogène dans le milieu réactionnel. Enfin, le temps d'occurrence d'une réaction est considéré comme dépendant principalement du temps nécessaire à deux réactifs pour se rencontrer, alors que le temps que la réaction se produise une fois les conditions réunies, est considéré comme négligeable. Grâce à ces trois hypothèses, il est possible de décrire l'évolution du système au cours du temps par une équation chimique principale (CME) [Lei 2010], qui correspond à un processus de markovien continu au cours du temps (CTMC) (Fig. 1.19). Ce formalisme a été intégré au sein des réseaux de Petri stochastiques [Molloy 1982] et de plusieurs autres variants tels que : les réseaux de Petri stochastiques généralisés (GSPNs) [Marsan *et al.* 1984], *Stochastic Activity Networks* [Meyer *et al.* 1985], *Stochastic Reward Nets* [Muppala *et al.* 1994], *Stochastic Well-Formed Colored Nets* [Chiola *et al.* 1993]. Compte-tenu des travaux réalisés dans cette thèse, nous nous limiterons à la description du formalisme principal des réseaux de Petri stochastiques et à leur extension sous la forme de GSPNs.

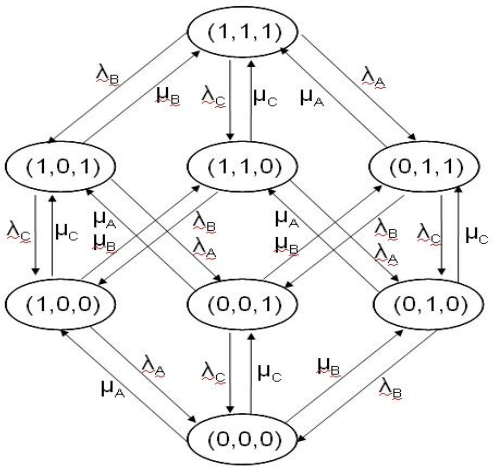


FIGURE 1.19 – Chaque ensemble correspond à l'état des molécules chimiques du système, tandis que les transitions correspondent à l'occurrence des réactions.

Réseaux de Petri stochastiques (SPNs) Le formalisme des SPNs, initialement proposé par [Molloy 1982], se présente, comme pour les réseaux de Petri classiques, avec un nombre de jetons discret aux places et des transitions qui dépendent du marquage m des places, de la stœchiométrie et d'une variable supplémentaire correspondant au taux de tirs X_t (équivalent à un temps d'attente), associée à chaque transition t. X_t est une variable aléatoire comprise entre $[0, +\infty[$, qui suit une probabilité de distribution exponentielle. En conséquence, on peut définir un SPN selon la définition suivante [Heiner *et al.* 2008] :

Définition 23 (**Réseau de Petri stochastique**). *Un réseau de Petri stochastique ($\mathcal{SPN}$) est un quintuplé (P, T, f, v, m_0) où :*

- *P et T sont des ensembles finis, non vides et disjoints. P est l'ensemble des places et T est l'ensemble des transitions.*
- *$f : ((P \times T) \cup (T \times P)) \to \mathbb{N}$ définit l'ensemble des arcs.*
- *$v : T \to H$ est une fonction qui assigne une fonction stochastique h_t à chaque transition t, tel que :*
  $H = \cup_{t \in T} \{h_t | h_t : \mathbb{N}^{|\bullet t|} \to \mathbb{R}^+\}$
  est l'ensemble des fonctions stochastiques, et $v(t) = h_t$ pour toutes les transitions $t \in T$
- *$m_0 : P \to \mathbb{N}$ est le marquage initial du réseau de Petri.*

Soit une transition activée (dont les pré-places sont suffisamment marquées), le temps d'attente pour le tir de cette transition dépend de la fonction de densité de probabilité associée à la variable aléatoire X_t, tel que :

$$f_{X_t}(\tau) = \lambda_t(m) e^{(-\lambda_t(m) * \tau)}, \tau \geq 0 \tag{1.3}$$

où τ est la variable aléatoire et $\lambda_t(m)$ est le taux de transition associé au marquage m défini par :

$$\lambda_t(m) = h_t = k_t \times N \times \prod_{p \in \bullet t} \left(\frac{m(p)}{N} \right) \tag{1.4}$$

où k_t est le taux de distribution de la transition t déduit à partir de la cinétique de la réaction selon le procédé décrit par [Wilkinson 2006] et N le nombre de niveaux de l'échelle de concentration employée (FIG. 1.20).

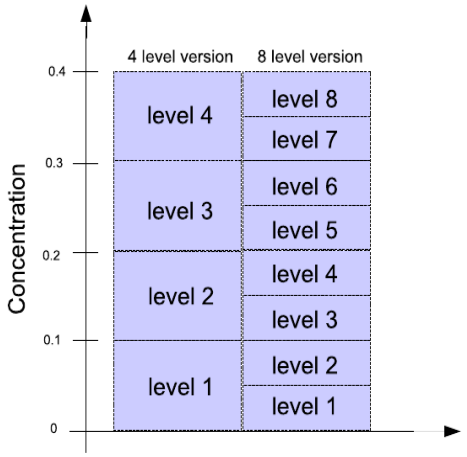


FIGURE 1.20 – Partitionnement de l'échelle de concentration en des niveaux discrets [Heiner *et al.* 2008]

L'intégration d'un tel processus de modélisation s'est avérée aussi efficace qu'un système continue pour modéliser l'évolution de la concentration de certains composés dans un système biologique (FIG. 1.21). Les SPNs ont montré qu'en considérant un nombre de paramètres cinétiques plus restreint que celui employé par les systèmes continus, il est possible d'approximer efficacement l'évolution de la concentration des métabolites au cours du temps dans un réseau métabolique.

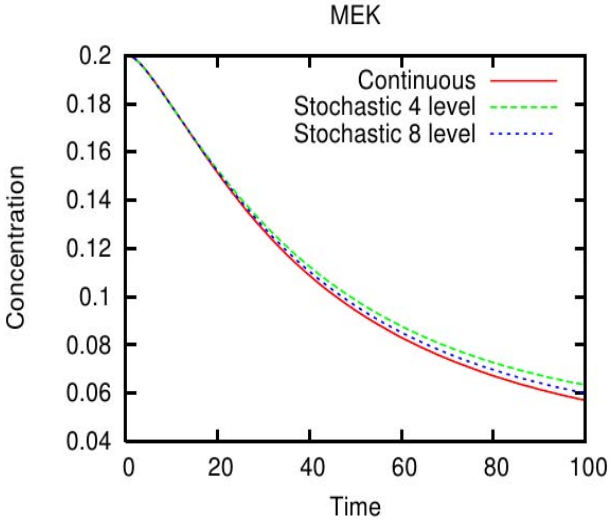


FIGURE 1.21 – Simulation au cours du temps de la consommation de MAPK (*mitogen-activated protein kinase*)/ ERK (*extracellular signal-regulated kinases*) kinase (MEK) selon la modélisation continue et stochastique [Heiner *et al.* 2008].

Réseaux de Petri stochastiques généralisés (GSPNs) Les réseaux de Petri stochastiques généralisés (GSPNs), formulés par [Marsan *et al.* 1984], introduisent deux paramètres supplémentaires par rapport aux SPNs, pour l'occurrence de transitions immédiates. Les transitions immédiates sont des transitions qui sont tirées de manière prioritaire par rapport aux transitions temporelles, et qui ne sont également pas soumise à un temps d'attente tir après leur sélection. Ces transitions sont particulièrement intéressantes pour modéliser des réactions chimiques car elles permettent de s'abstraire totalement des paramètres cinétiques lorsqu'ils ne sont pas disponibles, tout en considérant une notion de répartition de différentielle des métabolites entre les réactions. Ainsi, lorsque plusieurs transitions immédiates sont activées, une sélection de la transition à tirer est déterminée par le calcul d'une probabilité de tir. On peut donc définir les GSPNs par la définition suivante [Lamprecht *et al.* 2011] :

Définition 24 (**Réseau de Petri stochastique généralisé**). *Un réseau de Petri stochastique généralisé ($\mathcal{GSPN}$) est un 7-tuple $(P, T, f, e, w, pri, m_0)$ où :*
- *P et T sont des ensembles finis, non vides et disjoints. P est l'ensemble des places et T est l'ensemble des transitions.*
- *$f : ((P \times T) \cup (T \times P)) \to \mathbb{N}$ définit l'ensemble des arcs.*
- *$e : T \times \mathbb{N}^P \to \mathbb{B}$ est l'activation de chaque transition.*
- *$w : T \times \mathbb{N}^P$ est le poids de chaque transition.*
- *$pri : T \to \mathbb{N}^P$ est la priorité de chaque transition.*
- *$m_0 : P \to \mathbb{N}$ est le marquage initial du réseau de Petri.*

Soit l'état courant d'un marquage donné par $m \in \mathbb{N}^P$, une transition $t \in T$ d'un GSPN est activée si $e(t, m)$ est vraie et s'il n'y a pas de transition $t' \in T$ avec $pri(t') > pri(t)$ et dont $e(t', m)$ est aussi vraie. Les transitions dont la priorité est supérieure à 0 correspondent ainsi aux transitions immédiates, tandis que celles dont la priorité est égale à 0 correspondent aux transitions temporelles. Lorsque plusieurs transitions immédiates sont activées, la détermination de la transition à tirer ne dépend plus de la priorité, mais de la probabilité liée au poids de la transition $w(t)$. Soit $E(m) \subseteq T$ correspondant à l'ensemble des transitions immédiates actives, le poids de la transition $w(t)$ est défini par :

$$w_t(m) = \frac{w(t)}{\sum_{t' \in E(m)} w(t')} \tag{1.5}$$

Ces nouveaux paramètres apportent ainsi une flexibilité qui a permis de les appliquer notamment pour la modélisation de complexe de signalisation au Ca^{2+} [Lamprecht *et al.* 2011] et également de motiver l'application de ce formalisme pour l'étude du flux au sein des réseaux métaboliques (voir Section 3.3.2).

1.3 Travail de thèse

Le travail réalisé dans cette thèse se place dans ce contexte et ces problématiques d'intégration de données biologiques. L'essentiel des principes et des formalisations des graphes et des automates ayant été défini, nous verrons comment ont été adaptés leurs définitions et leurs usages pour l'analyse de données dont nous disposions.
Nous proposons dans cette thèse un nouveau formalisme de réseau de Petri s'appliquant à la prédiction des flux dans un réseau métabolique (voir Section 3), ainsi qu'une démarche intégrative reposant sur les multigraphes pour la prédiction des cibles des sRNAs (voir Section 5). La visualisation de ces réseaux et des informations qu'ils contiennent a aussi été l'objet

d'un intérêt majeur dans ce travail. Parmi les nombreuses solutions logicielles permettant de visualiser des graphes (Cytoscape [Smoot *et al.* 2011] pour les réseaux métaboliques et génomiques, Snoopy [Rohr *et al.* 2010] pour les réseaux de Petri, ...), nous nous sommes plus particulièrement intéressé au logiciel Tulip [Auber 2003] développé au sein de l'équipe MaBioVis au LaBRI. Ce logiciel permet de dessiner et d'interagir avec de très grands graphes. Il constitue également une boite à outil importante, en mettant à disposition sous la forme de plugins de très nombreux algorithmes de visualisation et de mesures sur les graphes.
Au cours de ce travail, nous avons exploité Tulip de deux manières. Avec les réseaux métaboliques, nous avons considéré la visualisation dynamique des graphes. Nous avons ainsi mis à contribution le dessin pour étudier la variation d'un paramètre du graphe selon la condition du modèle. Avec les réseaux d'interaction des ARN, nous avons plus particulièrement employé la fouille interactive des données que permet Tulip. Malgré la taille très importante de nos graphes, nous avons pu les analyser simplement, en les considérant de manière analytique selon un critère spécifique et en sélectionnant au fur et à mesure les filtres les plus pertinents. Ces éléments ont ensuite été redessinés en un sous-graphe avant de répéter cette analyse.
Ce travail nous a, dans chaque cas, permis de mettre en évidence des éléments d'intérêt sous-jacent et parfois dissimulés dans nos données. Nous avons pu, grâce à Tulip, sélectionner les outils et développer les plugins qui répondaient plus particulièrement à nos problématiques biologiques. Ainsi, nous avons pu participer au développement de deux logiciels basés sur Tulip, Systrip [Dubois *et al.* 2012] qui permet de visualiser les réseaux métaboliques (voir Section 3.4.2) et iRNA_visu qui permet de visualiser les interactions des sRNAs (voir Section 5.2.2).

Première partie

Prédiction de la distribution des flux au sein d'un réseau métabolique

Chapitre 2

Le métabolisme

Sommaire

Nous présentons dans ce chapitre les concepts biologiques et mathématiques nécessaires à la compréhension des développements réalisés en ce sens dans cette thèse. Dans la première partie, nous introduirons les notions utiles à la compréhension du métabolisme, puis nous aborderons dans une seconde partie les différentes méthodes d'exploration du métabolisme. Enfin nous verrons les différentes méthodes de modélisation du métabolisme dans lesquels s'intègre l'approche développée dans cette thèse. Les notions ici présentées sont issues de [Koch *et al.* 2011] et [Chang 2000].

2.1 Qu'est-ce que le métabolisme ?

Le métabolisme comprend l'ensemble des réactions pour la production et la dégradation des composés organiques qui se produisent dans les organismes vivants. Le métabolisme permet aux organismes de transformer des substrats en produits, leur permettant ainsi de croître, de se reproduire, de maintenir leur structure et d'interagir avec leur environnement. Le métabolisme cellulaire présente une organisation très structurée (voir Fig.1.13) qui met en œuvre, en grande majorité, des réactions de nature chimique impliquant une grande variété de molécules.

Le métabolisme est classiquement divisé en deux catégories : le catabolisme et l'anabolisme. Le catabolisme désigne l'ensemble des réactions permettant de dégrader de la matière organique. Il regroupe les réactions permettant de récolter de l'énergie comme la glycolyse ou le cycle de Krebs, et de produire les substrats nécessaires aux réactions anaboliques. L'énergie chimique produite par le catabolisme est stockée sous la forme d'ATP (Adénosine

Tri-Phosphate) et/ou de cofacteurs réduits tels que le NADH (Nicotinamide Adénine Dinucléotide) ou le NADPH (Nicotinamide Adénine Dinucléotide Phosphate). L'anabolisme est le processus de synthèse organique qui utilise l'énergie issue du catabolisme ou de l'environnement (photosynthèse). Il permet de synthétiser à partir de molécules simples (oses, acides aminés, acides gras) des macromolécules (protéines, lipides, acides nucléiques).

2.1.1 Les différents acteurs du métabolisme

Le métabolisme est organisé en voies métaboliques, qui comprennent un substrat principal qui est transformé par une série de réactions. On représente ces réactions par leur équation bilan (Fig. 2.1) dans laquelle sont spécifiés (i) les éléments participant à la réaction tels que les réactifs : l'ATP, le D-glucose, les produits : ADP, G6P et l'enzyme : hexokinase et (ii) la stœchiométrie dans laquelle se produit la réaction, c'est à dire la proportion des différents éléments qui sera consommée ou produite.

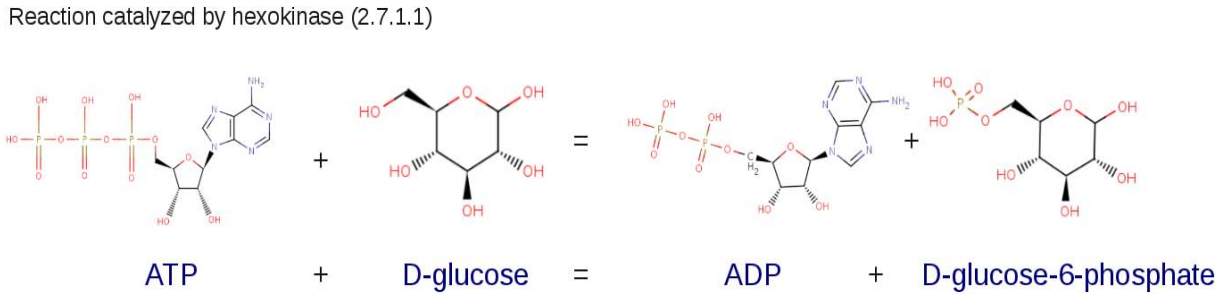


Figure 2.1 – Équation bilan de la réaction de phosphorylation catalysée par l'hexokinase (Illustration issue de : Brenda Enzyme Database)

Les réactions comportent un à plusieurs composés chimiques transformés, avec parfois l'intervention d'une ou plusieurs enzymes. Les voies métaboliques, que forment ces réactions, peuvent être linéaires avec un composé spécifique en entrée et en sortie, ou cycliques, avec un produit final qui peut être réutilisé pour un nouveau cycle de réactions.

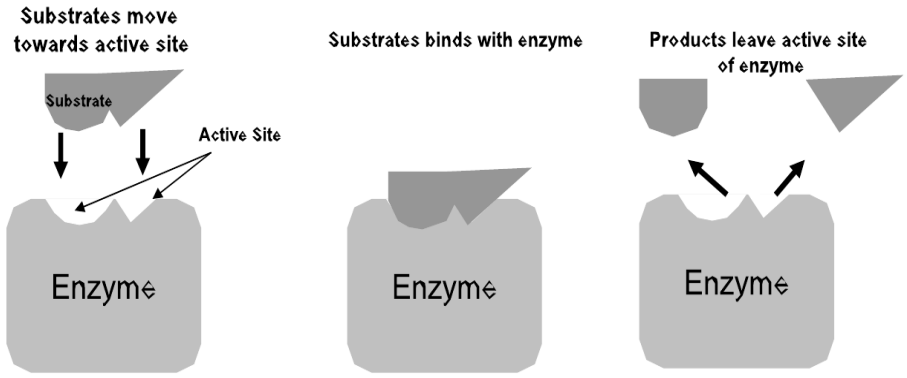
Les métabolites

Les métabolites sont à la base du métabolisme, ce sont des molécules organiques produites et consommées dans la cellule par le métabolisme. Ces molécules sont de faible masse moléculaire (par exemple le glucose) par rapport aux macromolécules (tel que le glycogène) qui constituent des polymères de masse moléculaire élevée. On identifie les métabolites selon deux catégories : les métabolites primaires et secondaires. Les métabolites primaires correspondent à l'ensemble des métabolites essentiels à la croissance, au développement et/ou à la reproduction de l'organisme. Ces métabolites sont issus du métabolisme des glucides, des acides gras, des protéines et des acides nucléiques. À cette première catégorie sont naturellement opposés les métabolites secondaires qui désignent l'ensemble des métabolites non essentiels à la survie de l'organisme tels que les antibiotiques, les pigments et les hormones.

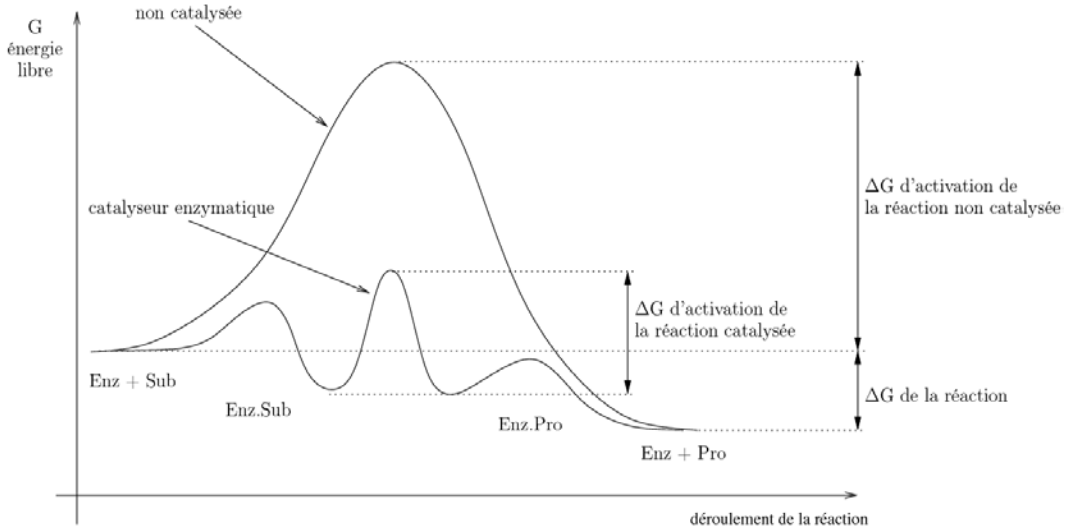
Les enzymes

Les enzymes jouent aussi un rôle crucial dans le métabolisme. En effet, ces protéines sont capables de catalyser des réactions (avec une augmentation du ratio de la réaction de l'ordre de

10^6 à 10^{18}) par un processus se déroulant basiquement en 3 étapes (Fig. 2.2a). Premièrement, les substrats se lient à une région de l'enzyme appelée le site actif. Ce site présente de manière générale une forme permettant la formation d'interactions faibles et réversibles entre l'enzyme et les substrats. Le complexe enzyme-substrat ainsi formé stabilise l'état de transition de la réaction [Koshland 1958] et abaisse le seuil d'énergie nécessaire à l'activation de la réaction ΔG (Fig. 2.2b) (voir Section 2.1.2). L'énergie, plus faible, qui doit être absorbée par les réactifs pour atteindre l'état de transition augmente donc la vitesse de la réaction. Enfin, les produits obtenus se détachent de l'enzyme.



(a) Processus de fonctionnement de la catalyse enzymatique.



(b) Évolution de l'énergie de Gibbs selon la progression de la réaction, pour une réaction non catalysée et une réaction catalysée.

Figure 2.2 – Processus de catalyse enzymatique [Boyer 2004].
(a) Description schématique d'une réaction transformant un produit en deux substrats. Le substrat se lie premièrement à l'enzyme sur son site actif par des liaisons faibles, puis l'enzyme catalyse la réaction qui le transforme en produit et libère le produit [Koch et al. *2011]. (b) Comparaison de l'énergie nécessaire à la réaction selon la progression de la réaction. Enz, enzyme ; Sub, substrat ; Pro, produit ; ΔG, énergie d'activation.*

Sans ce processus, le temps de réalisation de certaines réactions du métabolisme serait incompatible avec la vie. Bien que les métabolites puissent réagir entre eux de diverses manières,

les enzymes sont très spécifiques et ne peuvent catalyser la réaction que d'un nombre restreint de réactifs partageant une même structure. L'ensemble des réactions que peut faire une cellule est ainsi globalement déterminé par les enzymes qu'elle est capable de produire, celles-ci favorisant un certain nombre de réactions par rapport aux autres possibles. Certaines voies métaboliques vont donc être absentes chez certains organismes, lorsqu'ils seront incapables de produire les enzymes catalysant les réactions de la voie.
Les enzymes peuvent dans certains cas agir seule ou avec l'aide de cofacteurs (ions métalliques, cofacteurs énergétiques), et leur vitesse de réaction peut être affectée par plusieurs facteurs tels que la concentration des substrats et produits ou celle de facteurs de régulation, appelés inhibiteurs ou activateurs.

2.1.2 Règles régissant le métabolisme

Thermodynamique

Le métabolisme fonctionne, d'un point de vue thermodynamique, sous la forme d'un système ouvert irréversible, qui échange continuellement avec son environnement de l'énergie et de la matière. Ce système n'est jamais à l'équilibre mais se trouve toujours dans un état dit "stationnaire dynamique", obtenu lorsque la vitesse d'apparition d'un composé est compensée par la vitesse de sa dégradation. Le fonctionnement de ce type de système est décrit par l'enthalpie. L'enthalpie est une mesure de l'énergie totale (H) d'un système thermodynamique. Elle est exprimée en joule ou en calorie et elle peut être décrite de la manière suivante :

$$H = G + T \times S \tag{2.1}$$

où G, l'enthalpie libre (ou énergie libre) de Gibbs, est l'énergie qu'il faut fournir pour effectuer un travail et TS est l'énergie entropique (aussi appelée énergie du désordre). En biochimie, la variation de l'enthalpie libre est la principale énergie qui va nous intéresser car elle nous renseigne sur le sens d'évolution du système réactionnel. Elle est exprimée de la manière suivante :

$$\Delta G = \Delta H - T \times \Delta S \tag{2.2}$$

Prenons, pour exemple, un système composé des métabolites A, B, C et D tel que :

$$aA + bB \rightleftarrows cC + dD \tag{2.3}$$

avec la constante de réaction K (*constante de Gibbs*) définie par :

$$K = \frac{[C]^c[D]^d}{[A]^a[B]^b} \tag{2.4}$$

où [A] et [B] sont les concentrations des substrats, [C] et [D] sont les concentrations des produits et a,b,c,d leurs coefficients stœchiométriques. Il est possible de décrire la variation de l'enthalpie libre de ce système réactionnel (notée $\Delta_r G$) en fonction de K, grâce à la *relation de Gibbs*, tel que :

$$\Delta_r G = \Delta_r G^\circ + R \times T \times \ln K \tag{2.5}$$

où $\Delta_r G$ dépend de la variation de l'enthalpie standard de formation $\Delta_r G^\circ$; de R, la constante des gaz parfaits ; de T, la température en Kelvin et de K, la constante de Gibbs.
Ainsi, il est possible de décrire avec $\Delta_r G$ le fonctionnement énergétique du système (2.3).

- Si $\Delta_r G < 0$ avec $\Delta G_{CD} < \Delta G_{AB}$, la réaction est dite exergonique car elle fournit de l'énergie au système.
- Si $\Delta_r G > 0$ avec $\Delta G_{AB} < \Delta G_C D$, la réaction est dite endergonique car elle consomme de l'énergie.
- Si $\Delta_r G = 0$, la réaction ne consomme pas d'énergie.

À température et pression constante, les réactions s'opèrent dans le sens de la diminution de l'enthalpie libre. Le métabolisme doit donc pour fonctionner, s'assurer que les enthalpies des réactions sont bien négatives pour transformer les métabolites. Pour cela, le métabolisme peut, soit diminuer la valeur du coefficient de Gibbs K par une proportion plus importante des substrats par rapport aux produits, soit coupler aux réactions endergoniques, une réaction apportant de l'énergie. En pratique, le métabolisme réalise ces deux opérations en même temps. Les réactions du catabolisme ont lieu en flux tendu. Le métabolisme opère à une consommation immédiate des produits de dégradation et stocke l'énergie récoltée sous forme d'ATP, en ajoutant une liaison phosphodiester à une molécule d'ADP. Tandis que les réactions anaboliques procèdent à l'hydrolyse de l'ATP en ADP, pour récolter l'énergie dont elles ont besoin pour fonctionner. Cette dernière réaction est souvent combinée au sein de l'enzyme catalysant la réaction.

Cinétique des réactions

La vitesse d'une réaction chimique correspond à la vitesse du processus transformant le substrat en produit. Cette notion sera distinguée dans cette thèse de celle du flux métabolique, qui correspond à la proportion de molécules empruntant une voie métabolique indépendamment de la vitesse de la réaction. Nous nous intéresserons ici à la cinétique des réactions catalysées par des enzymes. On peut décrire ce processus, tel que vu précédemment (Fig. 2.2a), de manière simplifiée par l'équation suivante :

$$E + S \underset{k_{-1}}{\overset{k_1}{\rightleftarrows}} ES \xrightarrow{k_2} E + P \tag{2.6}$$

où E et S représentent respectivement l'enzyme et le substrat, ES est le complexe enzyme-substrat, P est le produit et k_1, k_{-1} et k_2 correspondent aux constantes d'équilibre de cette réaction, tel que k_1 est la constante d'association de [E][S], k_{-1} et k_2 sont les constantes de dissociation de [ES].
La vitesse de cette réaction s'exprime en $mol.L^{-1}.s^{-1}$. Cette unité décrit la quantité molaire de substrat transformée par unité de volume de la solution et unité de temps. Lorsque la réaction se déroule à l'intérieur d'une cellule, elle est plutôt exprimée en $mmol.h^{-1}.(gDW)^{-}1$ qui décrit la quantité de substrat transformée par heure et par quantité de cellule, DW représentant la masse sèche des cellules. La vitesse de la réaction (Eq. 2.6) dépend de nombreux facteurs, tels que la concentration du substrat, du produit et de l'enzyme, la température, le pH, etc... On peut cependant simplifier cette cinétique en ne considérant que quelques paramètres. [Menten et Michaelis 1913] ont ainsi proposé de décrire ce mécanisme en fonction de la concentration. La vitesse v est alors égale à :

$$v = -\frac{d[S]}{dt} = \frac{d[P]}{dt} = k_2[ES] \tag{2.7}$$

où v représente la vitesse de la réaction et $\frac{d[S]}{dt}$ et $\frac{d[P]}{dt}$ représentent le taux de consommation de S et de production de P au cours du temps. L'évolution de la production de P suit alors la dynamique observée (Fig. 2.3).

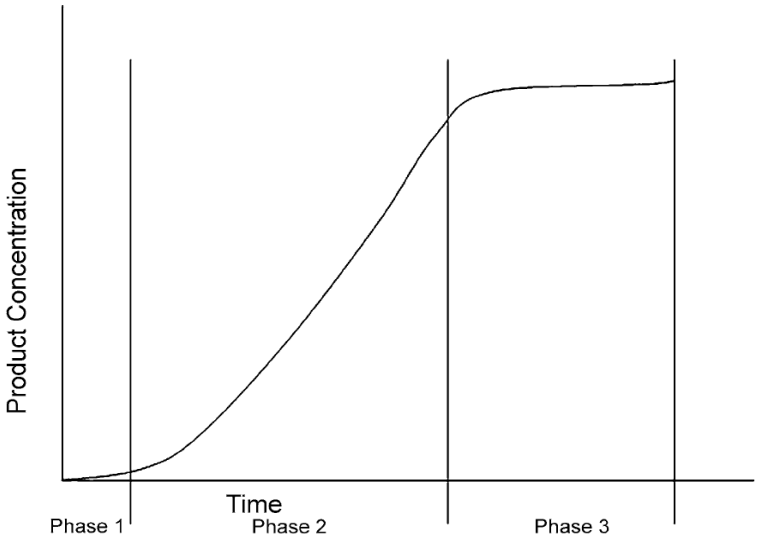


FIGURE 2.3 – Courbe représentant la concentration d'un produit en fonction du temps [Koch *et al.* 2011].
La formation du produit au cours du temps peut être divisée en trois phases. Au début de la réaction, la concentration de produit est relativement faible (phase 1), puis la concentration de produit augmente linéairement avec le temps (phase 2), enfin l'enzyme a transformé tout le substrat en produit, on atteint la phase stationnaire (phase 3).

La vitesse de production de ES est la différence entre les vitesses de deux réactions élémentaires décrivant sa formation et sa disparition :

$$\frac{d[ES]}{dt} = k_1[E][S] - k_{-1}[ES] - k_2[ES] \tag{2.8}$$

Cette équation différentielle ne peut être intégrée de façon explicite sans approximation qui la simplifie. Il est nécessaire de dériver cette expression vers la variable la plus facile à mesurer : la concentration du substrat ($[S]$).

Équilibre rapide Michaelis et Menten ont émis l'hypothèse que pour la première étape de la réaction, le substrat était présent en excès dans le milieu par rapport à la quantité d'enzyme, tel que $k_{-1} >> k_2$. La formation de ES est donc un état d'équilibre qui est rapidement atteint. La constante de dissociation K_s est ainsi donnée par l'équation :

$$K_s = \frac{k_{-1}}{k_1} = \frac{[E][S]}{[ES]} \tag{2.9}$$

Sachant que la concentration totale de l'enzyme est égale à $[E]_0 = [E] + [ES]$, on obtient :

$$K_s = \frac{([E]_0 - [ES])[S]}{[ES]} \tag{2.10}$$

On a donc :

$$[ES] = \frac{[E]_0[S]}{K_s + [S]} \tag{2.11}$$

Qui nous donne :

$$v = \frac{d[P]}{dt} = \frac{k_2[E]_0[S]}{K_s + [S]} \tag{2.12}$$

Ainsi lorsque la concentration de substrat est faible ($[S] << K_s$), on a :

$$v = \frac{k_2[E]_0[S]}{K_s} \tag{2.13}$$

et lorsque la concentration de substrat est élevée ($[S] >> K_s$), l'équation (2.12) est égale à :

$$v = \frac{d[P]}{dt} = k_2[E]_0 = V_{max} \tag{2.14}$$

où V_{max} est la vitesse maximale de la réaction, qui est obtenue lorsque toutes les enzymes sont complexées à un substrat.

État stationnaire L'approximation de l'équilibre rapide n'est cependant pas appropriée pour les systèmes biochimiques car les réactions sont consécutives dans le métabolisme et le produit devient rapidement le substrat de la réaction subséquente. Il n'y a pas d'équilibre possible. Au lieu d'un équilibre rapide, [Briggs et Haldane 1925] ont donc proposé l'hypothèse que peu de temps après la mise en contact de l'enzyme et du substrat, la concentration du complexe enzyme substrat tende vers une valeur constante (FIG. 2.4) correspondant à l'état stationnaire (*steady state* en anglais) tel que :

$$\frac{d[ES]}{dt} = 0 = k_1[E][S] - k_{-1}[ES] - k_2[ES] \tag{2.15}$$

On obtient donc après développement :

$$[ES] = \frac{k_1[E]_0[S]}{k_1[S] + k_{-1} + k_2} \tag{2.16}$$

Ainsi la vitesse de la réaction est égale à :

$$v = \frac{d[P]}{dt} = \frac{V_{max}[S]}{K_M + [S]} \tag{2.17}$$

où v dépend de $[S]$, de V_{max} et de K_M, *la constante de Michaelis*, qui est l'inverse de la mesure d'affinité entre l'enzyme et le substrat, tel que :

$$K_M = \frac{k_{-1} + k_2}{k_1} \tag{2.18}$$

Ainsi, plus K_M est petit, plus l'affinité entre le substrat et l'enzyme est forte. La mesure de K_m est obtenue à la concentration à laquelle la réaction a atteint la moitié de sa vitesse maximale ($\frac{V_{max}}{2}$) (FIG. 2.5).

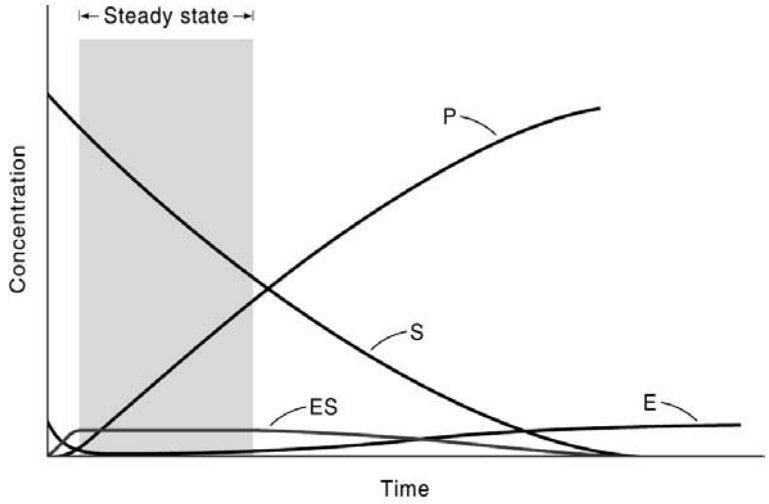


FIGURE 2.4 – Concentration des différents éléments au cours du temps d'une réaction catalysée par une enzyme [Chang 2000].
La concentration initiale du substrat est largement supérieure à la concentration de l'enzyme. Les paramètres k_1, k_{-1} et k_2 sont constants. L'état stationnaire de la réaction est désigné par un carré gris, il englobe la période de temps où [ES] est constante.

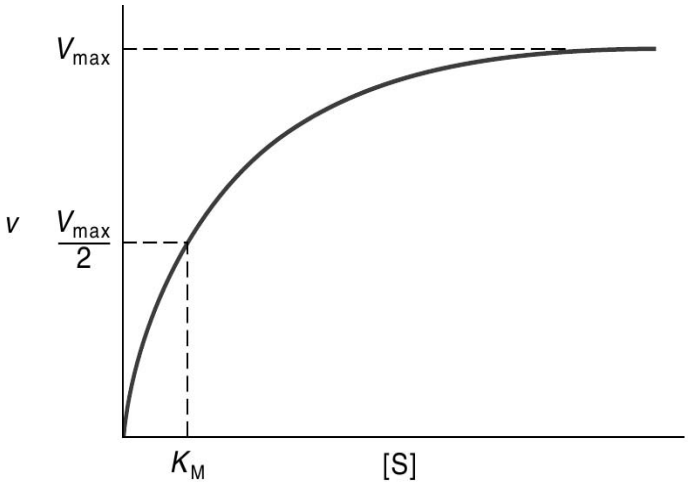


FIGURE 2.5 – Courbe décrivant la vitesse de la réaction en fonction de la concentration du substrat $[S]$ [Chang 2000].
Aux concentrations élevées de substrat ($[S] >> K_M$), la vitesse de la réaction tend vers V_{max}. L'enzyme est saturée et la vitesse de la réaction dépend linéairement de sa quantité. Aux concentrations faibles de substrat ($[S] << K_M$), la vitesse de la réaction tend vers $\frac{V_{max}}{K_M}$.

Contrôle du métabolisme

La cinétique de Michaelis-Menten, vue précédemment, permet de modéliser des transformations simples opérées par des enzymes. Cependant les réactions du métabolisme suivent souvent des cinétiques plus complexes, à cause des différents types de régulations mises en œuvre par la cellule. En effet, le métabolisme est un système dont la production et la consommation est contrôlée. Bien que ceci ne soit pas le sujet de ce manuscrit, nous souhaitons

rappeler au lecteur la multiplicité des régulations existantes, celles-ci allant d'un contrôle simple de la quantité d'enzyme et de son activité, à des mécanismes de contrôle bien plus complexes, tels que l'inhibition compétitive, la régulation allostérique, la régulation redox et la rétro-inhibition, ces régulations fonctionnant de pair avec des systèmes de contrôle du stress cellulaire ou de l'expression des protéines.

2.2 Reconstruction des réseaux métaboliques

Le métabolisme d'un organisme vivant se compose d'un grand nombre de réactions. Par exemple, pour *Escherichia coli* qui est un des organismes les plus étudiés, plus de 1400 réactions différentes ont été identifiées [Keseler *et al.* 2011]. Un enjeu majeur de ses dernières années est de rechercher et de caractériser toutes les voies métaboliques qui composent les différents organismes (Fig. 2.6).

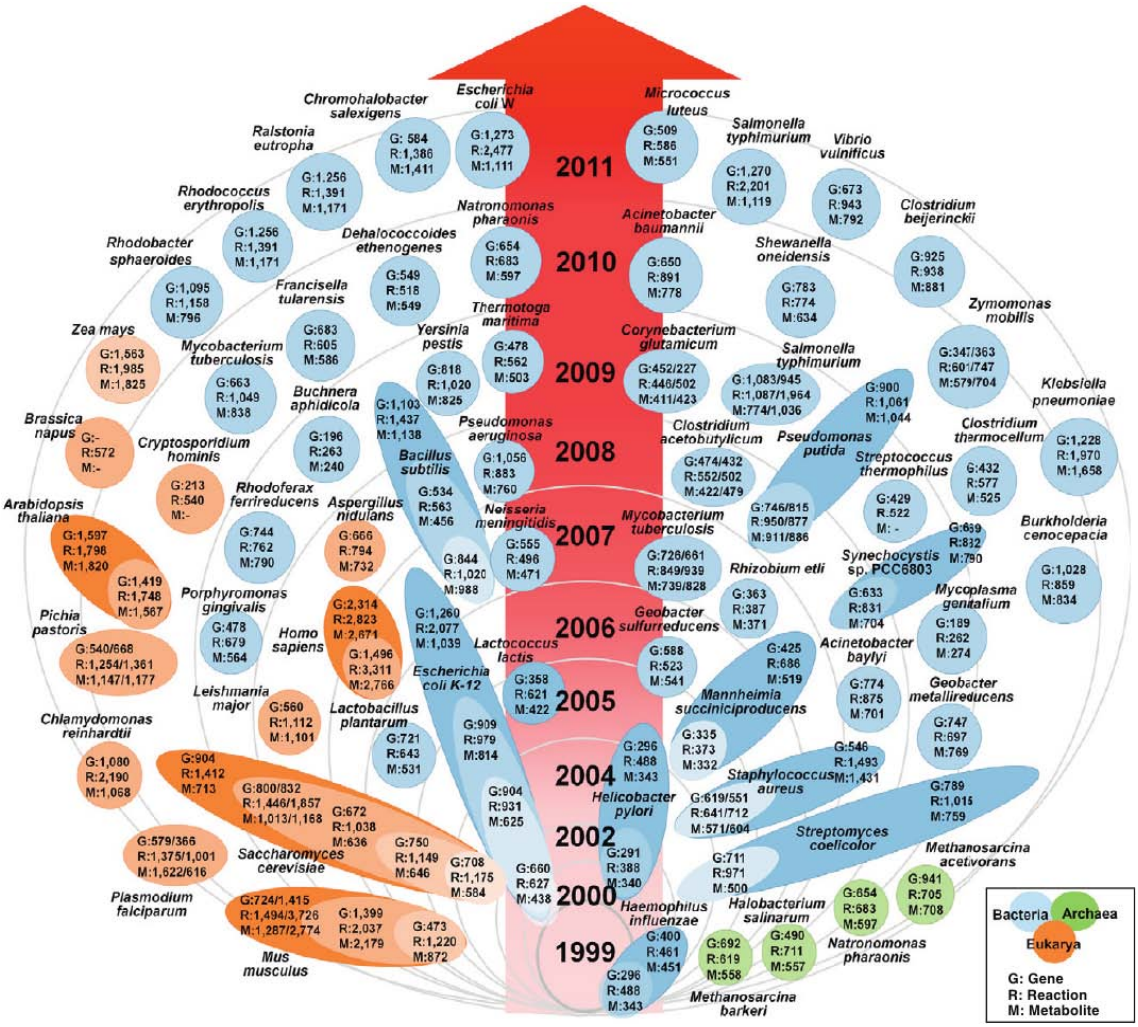


Figure 2.6 – Frise chronologique montrant les organismes dont le métabolisme a été reconstruit depuis 1999 [Kim *et al.* 2011].
Les ellipsoïdes orange, bleue et verte représentent respectivement les eucaryotes, les procaryotes et les archées. Les lettres G,R et M représentent le nombre de gènes, de réactions et de métabolites identifiés dans ces modèles.

La diversité biologique des organismes pourrait suggérer qu'ils adoptent une infinité de systèmes différents, cependant ce n'est pas le cas. Ils conservent généralement un mode de fonctionnement similaire de leurs voies métaboliques, avec par exemple le même enchainement d'enzymes pour catalyser une série de réactions. La reconstruction du réseau métabolique d'un organisme consiste donc bien souvent à inférer des relations entre les gènes, les protéines et les réactions (*GPR association*), sur la base des relations déjà connues. Différentes approches ont été développées à cet effet, nous détaillerons les approches expérimentales et bioinformatiques.

2.2.1 Approches expérimentales

L'approche expérimentale est historiquement la première approche employée pour la reconstruction des réseaux métaboliques. Dès la fin du 20^e siècle, elle permit de découvrir pour quelques organismes modèles tels que *Escherichia coli*, une grande partie des réactions et des voies métaboliques de son métabolisme [Gross *et al.* 1996], permettant déjà d'appréhender le métabolisme de cette espèce dans sa globalité.
Cette approche bénéficie à présent d'importantes avancées grâce à de nouvelles techniques d'investigation telles que la spectrométrie de masse, la chromatographie à haute performance ou la résonance magnétique nucléaire (RMN). Ces nouvelles technologies permettent d'étudier le métabolisme de manière bien plus poussée qu'auparavant, rendant possible l'identification de nouvelles voies métaboliques pour des organismes connus, comme celles identifiées pour le métabolisme du glucose [Fischer et Sauer 2003] ou de la pyrimidine [Loh *et al.* 2006] d'*Escherichia coli*. Enfin, l'approche expérimentale joue un rôle essentiel avec l'approche bioinformatique pour vérifier, valider et compléter les prédictions effectuées (voir Section 2.2.2), 30-40% des activités métaboliques de la classification de l'*EC*[1] (*Enzyme Commission*) ne correspondant pas encore à la séquence de gènes connus [Lespinet et Labedan 2005].

2.2.2 Approches bioinformatiques

Plusieurs types de données peuvent être utilisés pour reconstruire un réseau métabolique par une approche bioinformatique. On distinguera à cette occasion l'approche descendante dite *top-down*, utilisant en premier lieu l'information brute du génome pour reconstruire le squelette du réseau (*genome-scale network reconstruction : GENRE*) avant de s'intéresser au détail plus fin du réseau, et l'approche ascendante dite *bottom-up*, qui part d'une connaissance spécifique du réseau, basée sur des données de métabolomique et de protéomique, pour en arriver par assemblage à reconstruire le réseau principal.

Méthode *Top-Down*

Quatre étapes sont nécessaires pour reconstruire le réseau métabolique à partir du génome : (1) l'annotation du génome, (2) la reconstruction automatique du réseau métabolique, (3) le raffinement du réseau et (4) l'expérimentation *in vitro* [Thiele et Palsson 2010] (Fig. 2.7).

1. http ://www.chem.qmul.ac.uk/iubmb/

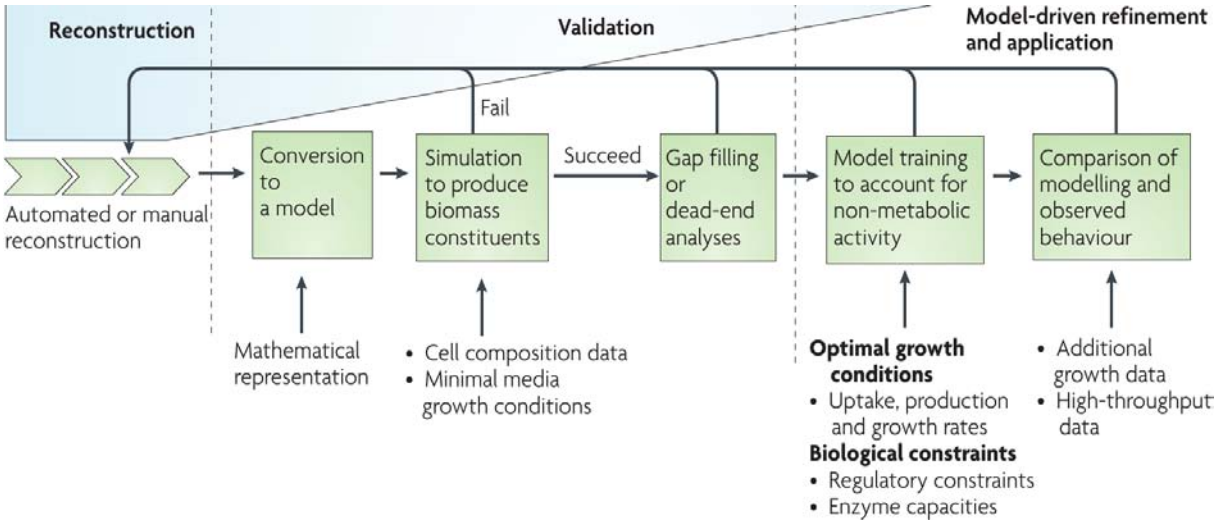
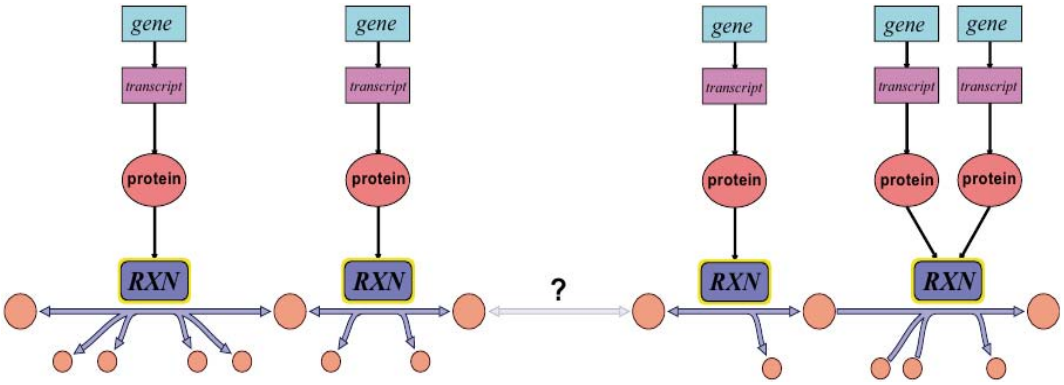


FIGURE 2.7 – Diagramme illustrant une méthodologie classiquement employée pour la reconstruction d'un réseau métabolique [Feist *et al.* 2009].

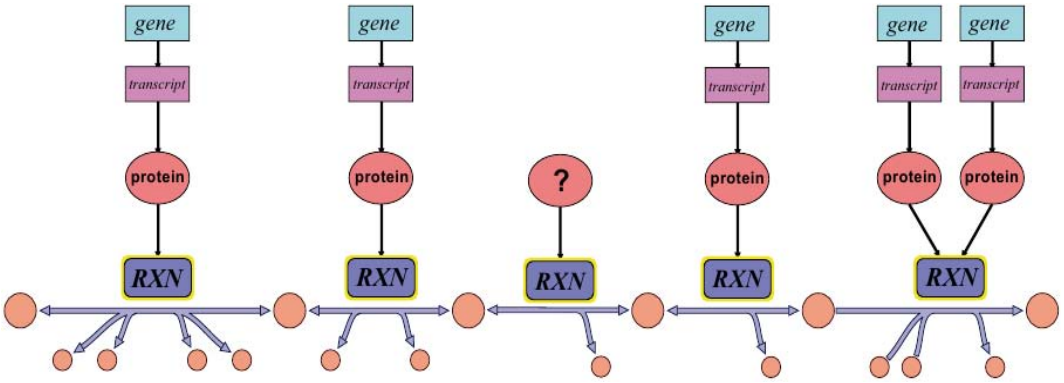
Première étape - Annotation du génome Depuis le premier génome séquencé [Jou *et al.* 1972], les avancées technologiques dans le domaine du séquençage ADN ont permis de réduire le coût et le temps nécessaire pour obtenir la séquence complète d'un génome [Kircher et Kelso 2010]. Ainsi, plus de 3000 génomes, séquencés et annotés, sont à présent disponibles[2]. Ceci constitue une base de connaissances majeure, qui est utilisée pour annoter par homologie d'autres organismes.
Sans vouloir exposer un état de l'art concernant le principe de l'annotation fonctionnelle des génomes qui n'est pas l'objet de cette thèse, nous souhaitons indiquer au lecteur que ces différentes connaissances des génomes sont disponibles au travers de différentes bases de données, soit spécifiques pour certains organismes : EcoCyc pour *E. coli* [Keseler *et al.* 2011], SGD pour *Saccharomyces* [Cherry *et al.* 1998], soit généralistes : EntrezGene [Maglott *et al.* 2011], CMR [Davidsen *et al.* 2010], Genome Reviews [Kulikova *et al.* 2007] ou IMG [Markowitz *et al.* 2006]. Il convient ensuite de les exploiter à l'aide d'outils d'alignements tels que BLAST [Altschul *et al.* 1990] pour identifier les gènes homologues ou directement par des outils d'annotation automatique tels que ERGO [Overbeek *et al.* 2003] ou RAST [Aziz *et al.* 2008].

Deuxième étape - Reconstruction automatique du réseau métabolique Grâce aux enzymes identifiées lors de l'annotation fonctionnelle, il est possible d'effectuer une première reconstruction du réseau métabolique. Cette étape est réalisée à l'aide d'outils tels que SEED [DeJongh *et al.* 2007] ou Pathway Tools [Karp *et al.* 2010], qui sont capables d'exploiter les bases de données dédiées aux protéines et au métabolisme : KEGG [Kanehisa et Goto 2000], BRENDA [Scheer *et al.* 2011], Metacyc [Caspi *et al.* 2006], pour fournir une définition initiale du réseau. À cette première étape, le réseau comporte encore beaucoup de gaps (réactions inconnues qui consomment ou produisent un métabolite) et de réactions orphelines (réactions connues comme existantes mais dont le gène n'a pu être identifié) (FIG. 2.8). Ils correspondent à une importante fraction de gènes (31 à 80%) qui n'ont pu se voir assigner une fonction [Iliopoulos *et al.* 2001; Suthers *et al.* 2009].

2. http ://www.genomesonline.org, http ://www.ebi.ac.uk/integr8



(a) Exemple de gap.



(b) Exemple de réaction orpheline.

FIGURE 2.8 – Diagramme illustrant la manière dont les (a) gaps et (b) les réactions orphelines surviennent lors de la reconstruction d'un réseau [Orth et Palsson 2010].
Lors du processus d'identification gène-transcrit-protéine-réaction, des gaps et des réactions orphelines peuvent être notées dans le réseau reconstruit. Les gaps correspondent à des réactions inconnues nécessaires pour produire ou consommer un métabolite. Les réactions orphelines correspondent à des réactions connues pour lesquelles aucun gène encodant pour l'enzyme nécessaire n'a été identifié. Dans cette figure, les métabolites sont représentés par des cercles oranges et les réactions par des flèches bleues. Les associations gène-protéine correspondent des relations portant sur la transcription et la traduction de chacune des entités.

Troisième étape - Raffinement du réseau Le raffinement du réseau métabolique est une étape indispensable de la reconstruction. Il permet de corriger trois principaux défauts des réseaux obtenus à l'issue de l'annotation. Le premier aspect concerne la cohérence du réseau métabolique avec les principes généraux du métabolisme. Il est en effet nécessaire de vérifier la stœchiométrie des réactions par curation manuelle car celles-ci peuvent être mal indiquées dans les bases de données. Il faut aussi vérifier la faisabilité thermodynamique du modèle en considérant la direction des réactions et les ratios des réactions par prédiction de l'énergie de Gibbs des voies métaboliques [Henry *et al.* 2009]. Enfin, il faut déterminer si le modèle généré respecte les contraintes systémiques et physicochimiques de la cellule. Pour cela, un modèle métabolique (*Genome-scale model* - GEM) est généré à partir du réseau et simulé sous contrainte à l'aide de logiciels tel que la COBRA toolbox [Schellenberger *et al.* 2011]. Cette étape nécessite de définir des objectifs métaboliques correspondant à la composition de la biomasse. La biomasse désigne les composants clés de l'organisme nécessaires à son maintien. Sa compo-

sition est déterminée expérimentalement à partir de la teneur en carbone et en acides aminés de différentes populations cellulaires de l'espèce donnée [Gonzalez *et al.* 2010]. Le nombre de facteurs inclus dépend des caractéristiques métaboliques de l'espèce, 61 facteurs ont été par exemple inclus pour *Mycoplasma pneumoniae* [Suthers *et al.* 2009]. Le second aspect concerne les gaps et les réactions orphelines pour lesquels différents outils ont été réalisés afin d'identifier les éléments manquants tels que GapFill [Kumar *et al.* 2007] ou BNICE [Hatzimanikatis *et al.* 2005] pour les gaps (*gap-filling*) et PHFiller-GC [Green et Karp 2007], SEED [Osterman 2006] ou ADOMETA [Chen et Vitkup 2006] pour les réactions orphelines (*orphan-filling*). Le dernier aspect concerne l'intégration des spécificités métaboliques de l'organisme considéré (compartimentalisation de certaines réactions, voies métaboliques spécifiques) qui nécessite la considération de la littérature liée à l'espèce [DeJongh *et al.* 2007; Karp *et al.* 2010].

Quatrième étape - Expérimentation *in vitro* L'expérimentation *in vitro* du modèle permet enfin de confronter le réseau métabolique reconstruit aux données expérimentales (Fig. 2.9). Cette étape permet de vérifier par exemple l'utilisation des métabolites intermédiaires [Van Noorden 2010; Niittylae *et al.* 2009] ou de comparer l'essentialité de certains gènes par mutagenèse en comparaison du modèle GENRE [Cameron *et al.* 2008; Gallagher *et al.* 2007].

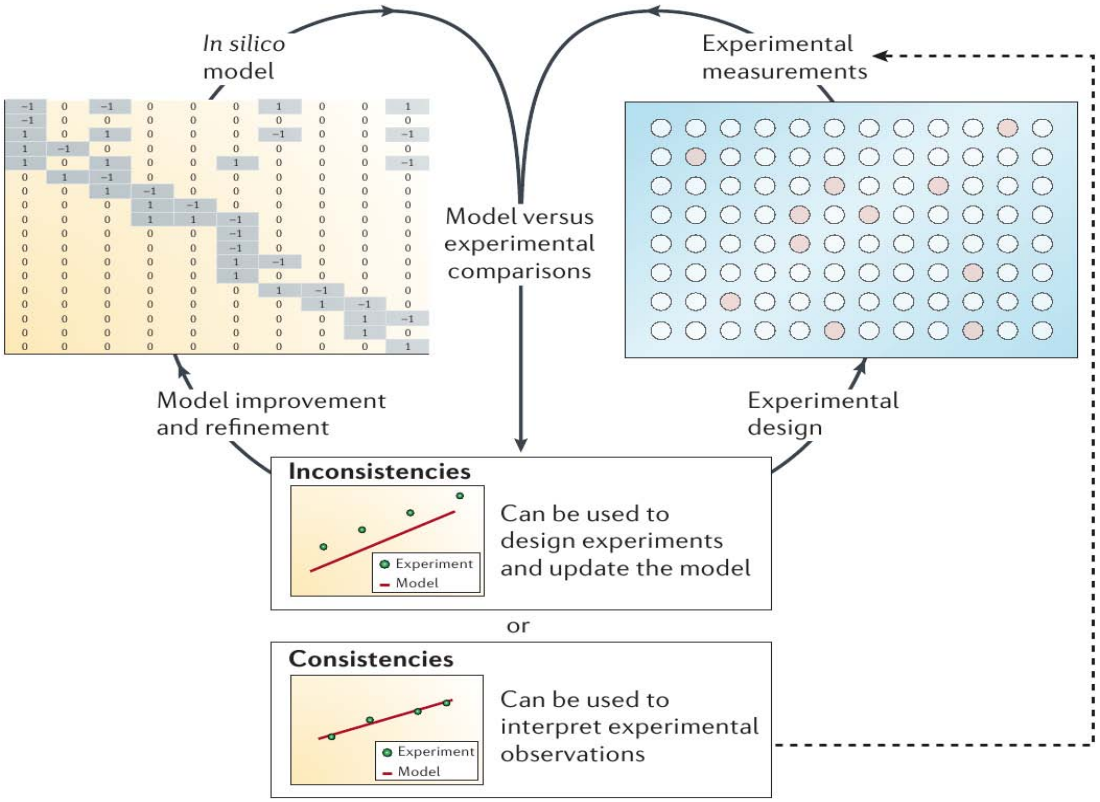


Figure 2.9 – Schéma de la procédure d'expansion du réseau [Reed *et al.* 2006].
La comparaison du modèle avec les données expérimentales permet d'identifier les éléments consistants et inconsistants. Les éléments consistants vont aider à l'interprétation des données biologiques, tandis que les éléments inconsistants génèrent des hypothèses sur l'organisme. Une procédure d'identification ou d'élimination de ses composants et/ou de leurs interactions est alors initiée par comparaison avec les données expérimentales pour améliorer la formulation du modèle.

Méthodes *Bottom-up*

Bien que très employée, l'approche *top-down* reste limitée par les connaissances actuelles en biologie et nécessite, comme nous l'avons vu précédemment, la répétition d'un processus de curation pour limiter les gaps, les réactions orphelines et les propagations de mauvaises annotations. Une seconde approche s'est donc développée en parallèle : l'approche *bottom-up*. Cette approche se base sur le principe de curation manuelle du réseau, composant par composant. Elle consiste à une intégration progressive des réactions et des voies métaboliques décrites par la littérature, ou lorsqu'elles ne sont pas connues, par une reconstruction progressive du réseau [Bringaud *et al.* 2006]. Une base de données BiGG (*Biochemical Genetic and Genomic knowledgebase*) [Schellenberger *et al.* 2010] permet enfin de stocker ces différents réseaux, 6 organismes sont à présent disponibles *Homo sapiens Recon 1* [Duarte *et al.* 2007], *Escherichia coli* K-12 *i*JR 904 [Reed *et al.* 2003] et K-12 MG1655 *i*JR204 [Feist *et al.* 2007], *Saccharomyces cerevisiae i*ND750 [Duarte *et al.* 2004], *Staphyloccocus aureus i*SB619 [Becker et Palsson 2005], *Methanosarcina barkeri i*AF692 [Feist *et al.* 2006] et *Helicobacter pylori i*lT341 [Thiele *et al.* 2005].

2.3 Modélisation du métabolisme

Lors de la reconstruction d'un réseau métabolique, plusieurs étapes de modélisation sont réalisées suivant les connaissances acquises (voir Section 2.2). Différentes approches ont ainsi été développées pour représenter, analyser et surtout tester ces réseaux. Dans cette partie, nous ferons une revue des différentes types de modélisation permettant de tester la consistance des réseaux métaboliques, avant de nous concentrer sur les évolutions apportées par la modélisation dynamique sous contraintes auxquelles s'intègre l'approche développée dans cette thèse (voir Section 3).

2.3.1 Les différents formalismes de modélisation

Il existe trois principaux types de modélisation du métabolisme selon le niveau d'analyse des réseaux considérés d'après [Stelling 2004] (Fig. 2.10).

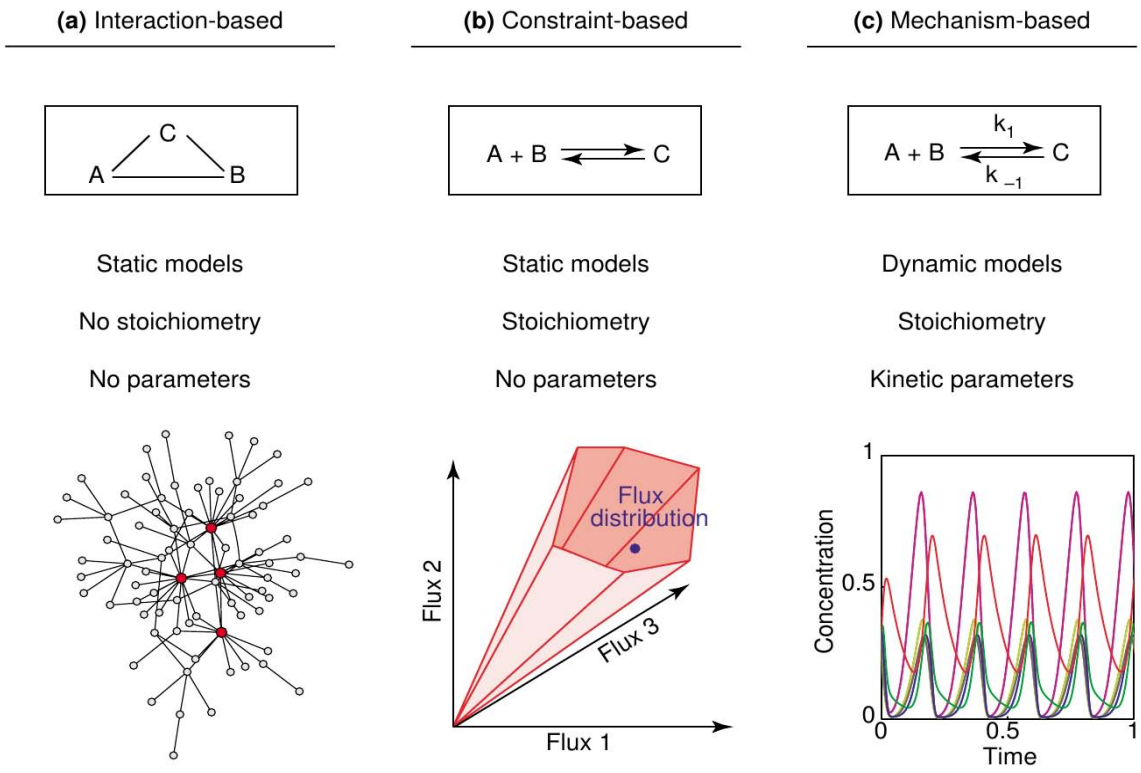


FIGURE 2.10 – **Les différents formalismes de modélisation des réseaux métaboliques** (a) Modélisation basée uniquement sur les interactions, (b) Modélisation à base de contraintes, (c) Modélisation cinétique tenant compte de l'évolution au cours du temps des quantités d'enzyme et de métabolites [Stelling 2004].

Modélisation structurelle Le premier niveau de cette classification se rapporte à la modélisation structurelle des réseaux (FIG. 2.10a). Ces modèles statiques s'adressent principalement à l'analyse et la vérification structurelle des réseaux reconstruits. Ils se basent sur les interactions entre métabolites et effecteurs (enzymes, transporteurs,...) pour créer un graphe d'interaction. Ils simplifient fortement la définition du métabolisme, car ils sont statiques et ne tiennent pas compte du type d'interaction (catalyse, régulation...). Les réseaux d'interaction sont très utilisés pour étudier l'organisation structurelle du métabolisme, et à ce titre, la théorie des graphes fournit beaucoup de méthodes pour analyser la connectivité des métabolites ou des effecteurs [Bergmann *et al.* 2004], les chemins du graphe utilisés pour analyser leurs longueurs [Ma et Zeng 2003] ou leurs modes élémentaires [Schuster *et al.* 1999], ou encore pour identifier des motifs particuliers tels que les *choke points* (enzymes consommant ou produisant un unique métabolite) utilisés comme cible de médicaments [Yeh *et al.* 2004; Rahman et Schomburg 2006].

On oppose à cette première approche les modèles dynamiques, qui sont utilisés pour analyser le fonctionnement des réseaux métaboliques. Ces modèles servent, principalement, à faire le lien entre la connaissance de la structure du réseau, que l'on obtient lors de la reconstruction, et le phénotype observé chez la cellule. Ils correspondent aux modèles "basés sur les contraintes" (*constraint-based modeling*) (FIG. 2.10b) et aux modèles cinétiques (*mechanism-based modeling*) (FIG. 2.10c). Ces approches comportent, toutes deux, une représentation structurale (et mathématique) du réseau qui est combinée à un ensemble de règles issues du métabolisme (voir Section 2.1) et, si elles sont disponibles, des mesures quantitatives de

paramètres permettant la simulation du mécanisme chimique [Papin *et al.* 2003]. En intégrant plus de connaissances, ces réseaux sont aussi plus réalistes, plus complexes et plus petits, à mesure que le niveau de détail intégré est plus important (FIG. 2.11).

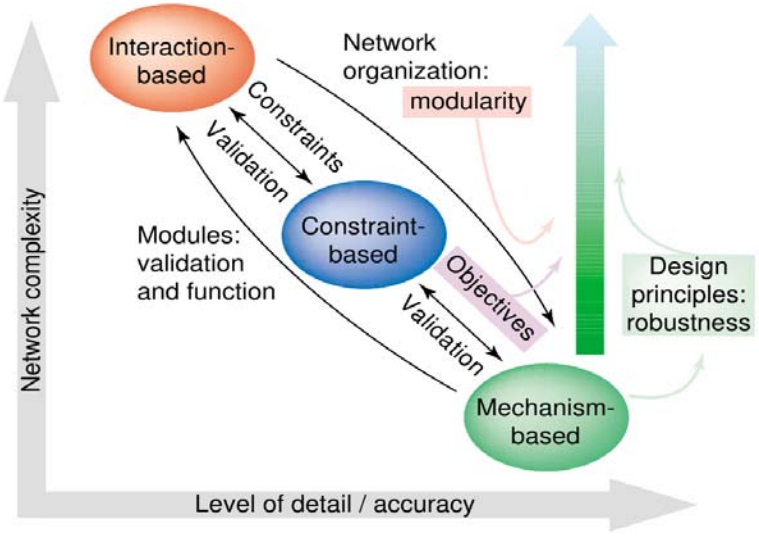


FIGURE 2.11 – Diagramme représentant les trois classes de modélisations (structurelle, basées sur les contraintes et cinétique) selon leur degré de précision de l'analyse et la taille des réseaux qu'elles peuvent prendre en compte. Les flèches noires se réfèrent aux interactions entre ces méthodes. La flèche verte indique la robustesse de l'analyse suivant le niveau considéré [Stelling 2004].

Modélisation "basée sur les contraintes" La modélisation "basée sur les contraintes" correspond ainsi au niveau intermédiaire d'intégration des connaissances biologiques. Elle repose sur l'hypothèse centrale que les cellules ont évolué de manière à ce que leur réseau métabolique se comporte de manière optimale face aux contraintes du métabolisme [Price *et al.* 2003; Kauffman *et al.* 2003]. La modélisation par contrainte est notamment utilisée pour prédire la distribution du flux au sein d'un métabolisme (voir Section 2.1.2).
Selon l'hypothèse centrale de la modélisation "basées sur les contraintes", le flux des métabolites emprunte un chemin optimal dans le réseau pour produire le plus possible de constituants nécessaires à la vie. Deux approches principales ont été développées, à partir de ce postulat, pour prédire l'état probable d'un réseau métabolique : le FBA (*Flux Balance Analysis*) [Fell et Small 1986; Orth *et al.* 2010], qui est capable de prédire une distribution des flux compte-tenu d'un objectif à maximiser, et le MFA (*Metabolic Flux Analysis*) [Lee *et al.* 1999], qui prédit une distribution du flux par combinaison des contraintes du métabolisme et d'une mesure expérimentale de certains flux au sein du réseau. Ces approches s'abstraient des données cinétiques, qui sont rarement disponibles [Palsson 2006; Llaneras et Picó 2008], en considérant les flux à un état *pseudo*-stationnaire, où le flux correspond à un rendement net de la réaction. Elles ne vont donc requérir qu'une connaissance de la structure du réseau, du sens des réactions, de la stœchiométrie et du rendement en métabolites souhaités pour la fonction objective du FBA ou de données de flux de certaines voies obtenues par exemple lors d'expériences de marquage ^{13}C pour le MFA [Sauer 2006; Wiechert 2001]. Ainsi, ces méthodes vont être particulièrement importantes pour conjecturer du fonctionnement d'un réseau (voir Section 2.2.2) selon sa capacité à satisfaire pleinement ou en partie les contraintes (d'après les prédictions

de ses méthodes) et selon la correspondance entre les flux prédits et les observations expérimentales. Compte-tenu des données expérimentales disponibles pour le modèle d'étude et de l'approche développée durant cette thèse, nous nous focaliserons sur une revue du FBA et de ses évolutions en Section 2.3.2.

Modélisation cinétique Enfin, le dernier niveau de cette classification correspond aux modèles cinétiques qui prennent en compte le plus de données biologiques. Ces modèles intègrent des mesures détaillées sur le réseau, telles que la concentration des métabolites et les paramètres cinétiques des réactions, pour simuler et tester avec exactitude un réseau métabolique. L'utilisation de tels modèles permet souvent de formuler des hypothèses expérimentalement testables de phénomènes parfois complexes [Vilar *et al.* 2003]. Cependant ils ne sont que peu ou pas utilisés lors de la reconstruction des réseaux métaboliques car la complexité des systèmes biologiques rend difficile l'acquisition de telles données et implique souvent des mécanismes inconnus, particulièrement lorsque l'étude se porte à l'échelle d'un génome [Kitano 2002].

2.3.2 Flux Balance Analysis (FBA)

Principe

Une analyse par FBA consiste en 2 étapes : la définition des contraintes et l'optimisation d'un objectif métabolique (Fig. 2.12).

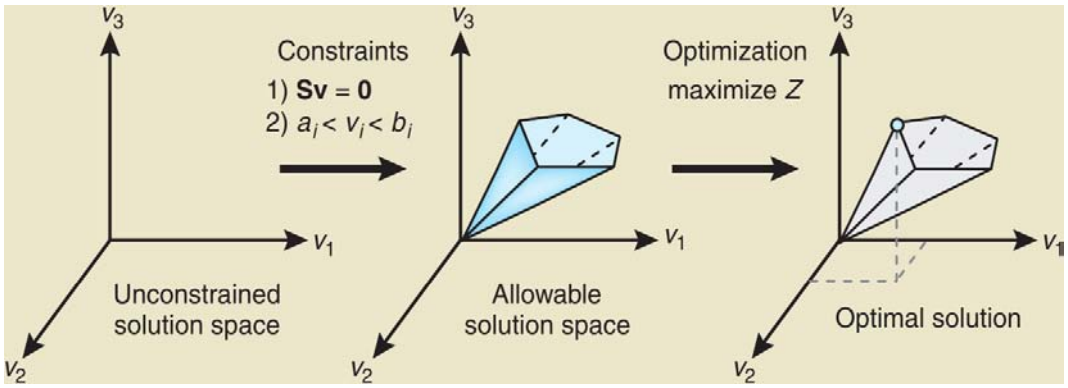


Figure 2.12 – Schéma décrivant le concept de l'analyse réalisée par le FBA [Orth *et al.* 2010]. *Deux étapes sont considérées dans la recherche de la distribution du flux par FBA. La première étape consiste à définir des contraintes d'après 1) les règles physicochimiques du métabolisme et 2) les limites de capacité imposées au flux par l'utilisateur. La seconde étape consiste à définir à l'objectif métabolique qui est optimisé pour identifier la distribution optimale du flux d'après les contraintes précédemment définies.*

La principale contrainte d'un modèle FBA repose sur la topologie du réseau métabolique et ses contraintes physico-chimiques (voir Section 2.1.2). L'intégration de cette contrainte permet de restreindre l'ensemble des solutions à des distributions de flux cohérentes d'un point de vue biologique. Cette première étape est concrètement réalisée par la définition du formalisme élémentaire des réactions. Un exemple de système réactionnel est présenté Figure 2.13 pour illustrer ce processus.

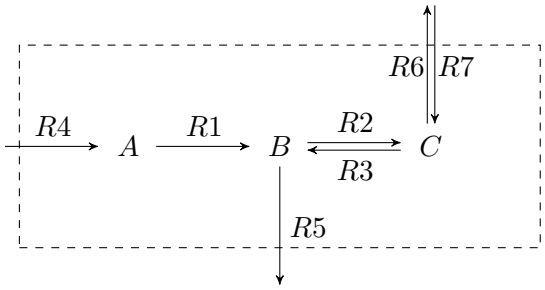


FIGURE 2.13 – Un exemple de système réactionnel [Lee *et al.* 2006]. Ce réseau comporte 7 réactions comprises dans le système intracellulaire.

Dans ce système, la variation de la concentration des métabolites A et B peut être exprimée par les relations suivantes (E_Q. 2.19) :

$$\begin{aligned} \frac{d[A]}{dt} &= -v_1 + v_4 \\ \frac{d[B]}{dt} &= v_1 - v_5 \end{aligned} \tag{2.19}$$

où $\frac{d[A]}{dt}$ et $\frac{d[B]}{dt}$ correspondent à la variation de la concentration des métabolites A et B respectivement obtenue par la différence entre les ratios (v_1, v_4 et v_5) auxquels les métabolites sont produits et consommés (voir Section 2.1.2). On peut remarquer que cette définition suit le principe de conservation des masses avec aucune création ou perte de matière. Ce système, dans le cas des métabolites A et B, peut être écrit sous forme matricielle de la manière suivante :

$$\begin{vmatrix} \frac{d[A]}{dt} \\ \frac{d[B]}{dt} \end{vmatrix} = \begin{vmatrix} -1 & 1 & 0 \\ 1 & 0 & -1 \\ 0 & 0 & 0 \end{vmatrix} \times \begin{vmatrix} v_1 \\ v_4 \\ v_5 \end{vmatrix} \tag{2.20}$$

Appliqué à l'ensemble du système, on obtient la relation de l'équilibre des masses (*mass balance*) caractérisant tout système réactionnel biologique (EQ. 2.21) :

$$\mathcal{V} = S \times v \tag{2.21}$$

où $\mathcal{V}$ est le vecteur de vitesse des réactions, S est la matrice des coefficients stœchiométriques du système réactionnel et v est le vecteur de flux des réactions. La matrice stœchiométrique S est ici de taille $m \times n$, chaque entrée de cette matrice représente la stœchiométrie d'un unique composé m (en ligne), pour une unique réaction n (en colonne). Ce coefficient est négatif pour les composants consommés et positif pour les métabolites produits. Un coefficient zéro est appliqué si le composant ne participe pas à la réaction. La matrice S constitue ainsi une matrice creuse car les réactions ne vont généralement consommer qu'un ou deux métabolites en entrée. Concernant le vecteur de flux, les paramètres des réactions ne sont généralement pas connus. Sa valeur est donc souvent exprimée à l'aide d'une mesure relative, ou parfois par la mesure du poids de cellule (quantité de substrat transformée par heure et par quantité de cellule) correspondant à la contribution de la réaction au flux dans la cellule [Smith et Robinson 2011].

À l'état stationnaire, la variation de vitesse des réactions $\mathcal{V}$ est égale à 0 pour toutes les réactions intracellulaires [Varma et Palsson 1994] (voir aussi Section 2.1.2), tel que :

$$S \times v = 0 \tag{2.22}$$

Cette considération de l'équilibre des masses à l'état stationnaire permet de simplifier le système, puisqu'il ne dépend plus que de la stœchiométrie et du vecteur de flux. On obtient ainsi la contrainte d'équilibre des masses qui constitue le principal support du modèle FBA. Il est ensuite possible d'associer à cette contrainte des limites sur les flux. En effet lorsque l'orientation thermodynamique de certaines réactions est connue, on peut orienter le flux de manière à respecter cette contrainte. Pour cela, le FBA permet de limiter la capacité de v entre des limites haute et basse (*upper/lower bound limits*), où le flux ne pourra se produire que dans un sens et être soit nul, soit très faible dans le sens opposé.

La second étape consiste à définir l'objectif métabolique qui correspond au phénotype cellulaire. Cet objectif est formulé à l'aide d'une fonction objective correspondant généralement à une réaction du réseau dont le flux est maximisé ou minimisé. Cet objectif peut être illustré dans notre modèle de la manière suivante :

$$\begin{aligned} &Objectif : max\ Z = v_5 \\ &Constraintes : \\ &\qquad \begin{array}{c} \\ A \\ B \\ C \end{array} \begin{array}{c} \begin{array}{ccccccc} R_1 & R_2 & R_3 & R_4 & R_5 & R_6 & R_7 \end{array} \\ \begin{bmatrix} -1 & 0 & 0 & 1 & 0 & 0 & 0 \\ 1 & -1 & 1 & 0 & -1 & 0 & 0 \\ 0 & 1 & -1 & 0 & 0 & -1 & 1 \end{bmatrix} \end{array} \begin{bmatrix} v_1 \\ \vdots \\ v_7 \end{bmatrix} = 0 \\ &\qquad 0 \leq v_1, ..., v_7 \leq 10 \end{aligned} \tag{2.23}$$

où l'objectif est constitué par le flux de la réaction Z que l'on va tenter de maximiser sous la contrainte de la matrice stœchiométrique et d'une contrainte de valeur sur les flux v. Ce problème constitue un problème de programmation linéaire qui est résolu par un optimiseur linéaire. On obtient ainsi pour notre modèle initial la répartition du flux suivante (Fig. 2.14) :

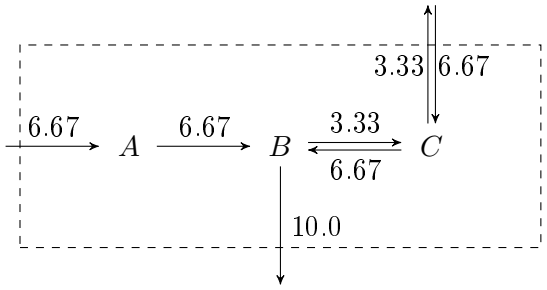


Figure 2.14 – $Z = 10, v = [6.67\ 3.33 6.67\ 6.67\ 10.0\ 3.33\ 6.67]^T$, [Lee *et al.* 2006]. Les flux ici prédits correspondent à la contribution relative des réactions. Comme on peut l'observer, le flux de la réaction Z a été maximisé grâce à orientation du flux des précédentes réactions pour que Z soit maximal. Les contraintes sur le flux v empêchent ici d'avoir un flux négatif sur une réaction.

La fonction objectif ainsi employée permet de formuler différents types de contraintes correspondant par exemple à la croissance cellulaire, la synthèse d'un produit, la minimisation

de la consommation d'ATP, la consommation de nutriment [Segrè *et al.* 2002]... Pour cela, une réaction artificielle est généralement ajoutée au réseau. Par exemple dans le cas d'un objectif de production de biomasse (constituants nécessaires à la croissance de l'organisme), cette réaction (appelée réaction de biomasse) consomme des métabolites précurseurs à une stœchiométrie qui simule la production réelle de ses constituants [Orth *et al.* 2010]. Cette stratégie a ainsi permis d'optimiser efficacement le flux de grands réseaux métaboliques tels que celui d'*Escherichia coli* [Edwards et Palsson 2000; Covert et Palsson 2002] et *Helicobacter pylori* [Schilling *et al.* 2002], pour lesquels les prédictions de maximisation de la biomass sont apparus consistantes avec les mesures expérimentales [Kauffman *et al.* 2003]. D'autres applications de cette méthode ont depuis été réalisées notamment pour la formulation d'hypothèse de fonctionnement métabolique dans le cas de dysfonctionnements métaboliques [Smith et Robinson 2011].

Évolutions

La méthode FBA a attiré par ses succès l'attention de la communauté qui souhaite à présent en améliorer la signifiance biologique. Plusieurs autres méthodes ont ainsi été développées dans l'objectif de diversifier et d'intégrer plus de connaissances biologiques dans les analyses effectuées par FBA. Ces méthodes peuvent être regroupées dans deux catégories selon qu'elles modifient le principe de fonctionnement de FBA (*FBA modification*) ou qu'elles intègrent FBA dans une analyse plus large (*FBA intégration*) (Fig. 2.15).

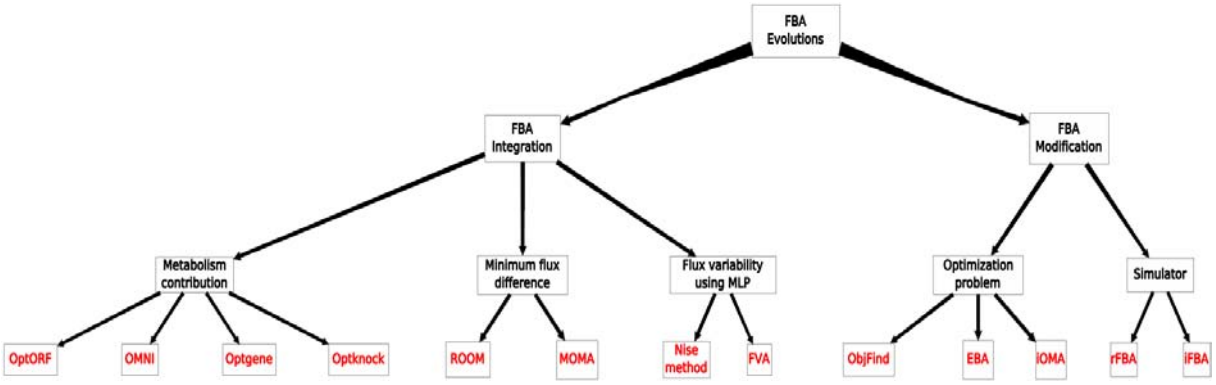


Figure 2.15 – Les différentes évolutions du FBA.

La première catégorie, désignée sous le nom de *FBA modification*, s'attache à modifier le problème résolu par FBA en permettant d'intégrer des données de thermodynamique et de cinétique lorsqu'elles sont disponibles. Ces changements ont pour objectif d'améliorer le réalisme de la contrainte du réseau par rapport à une distribution de flux prédite par un FBA classique. Dans cette catégorie, on retrouve par exemple rFBA [Covert et Palsson 2002] et iFBA [Covert *et al.* 2004] qui ajoutent respectivement un modèle booléen pour modéliser la régulation transcriptionnelle et un modèle cinétique pour intégrer les paramètres de certaines réactions connues. D'autres méthodes telles que EBA (*Energy Balance Analysis*) [Beard *et al.* 2002] suppriment des solutions thermodynamiquement infaisables proposées par FBA. Enfin, des solutions telles que Objfind [Burgard et Maranas 2003] et iOMA [Yizhak *et al.* 2010] permettent d'imposer des valeurs de flux issues de l'expérimentation à certaines des réactions.

La seconde catégorie de méthodes ne touche pas au fonctionnement de FBA mais l'intègre dans une analyse plus large du réseau métabolique. Une première classe d'outils réalise ainsi

de multiples optimisations par FBA pour tester et intégrer différents objectifs métaboliques dans la prédiction du flux. Par exemple, FVA (*Flux Variability Analysis*) [Gudmundsson et Thiele 2010] réalise deux optimisations d'un même modèle par FBA. Chacune de ses optimisations permet de prédire le flux nécessaire pour maximiser ou minimiser la fonction objective. Ces prédictions sont ensuite fusionnées pour prédire les zones de flexibilité du flux au sein du réseau où la différence entre le flux maximum et minimum est importante. Dans cette même classe, la méthode Nise (*Noninferior set estimation*) [Oh *et al.* 2009] est capable de déterminer une distribution de flux consensus entre trois objectifs métaboliques (maximisation de la production d'un métabolite, maximisation du flux d'une réaction et maximisation de la production de biomasse dans l'exemple présenté) pour un même réseau. Le réseau est en fait optimisé trois fois pour chacun de ces objectifs et un consensus des distributions de flux est calculé entre ces derniers pour déterminer le meilleur compromis. Une seconde classe d'outils s'attache à prédire la distribution du flux d'une espèce mutante en minimisant la différence par rapport au modèle sauvage. Ces méthodes sont particulièrement utilisées pour tester l'effet de perturbations sur un réseau métabolique lorsque des mutations du réseau sauvage existent ou que l'on souhaite tester la cohérence et l'exhaustivité du réseau sauvage. Par exemple, MOMA (*Minimization Of Metabolic Adjustment*) [Segrè *et al.* 2002] calcule la distribution du flux du mutant en estimant la distance minimum relative à la solution métabolique de l'espèce sauvage. Le logiciel ROOM (*Regulatory on/off minimization*) [Shlomi *et al.* 2005], dans cette même idée, détermine le nombre minimum de changements au niveau du flux entre l'espèce sauvage et l'espèce mutante pour prédire le flux de ce dernier. Enfin, une dernière classe de méthodes cherche les différentes stratégies de production pour permettre la surproduction d'un métabolite ou la reproduction d'un phénotype sauvage (certains métabolites excrétés). Ces méthodes telles que Optknock [Burgard *et al.* 2003], Optgene [Patil *et al.* 2005], OMNI [Herrgård *et al.* 2006] et OptORF [Kim et Reed 2010] se basent sur FBA pour tester la distribution du flux selon une certaine formulation du réseau.
Deux éléments peuvent être notés pour ces différentes évolutions. Premièrement, avec cette augmentation du nombre d'objectifs à intégrer, les auteurs de ces méthodes ont dû employer des méthodes d'optimisation plus évoluées à mesure que le problème à résoudre se voyait complexifier. Ainsi, depuis l'optimisation linéaire employée par FBA pour satisfaire une fonction objective, ces méthodes ont successivement employé des méthodes telles que des optimisations quadratiques (MOMA, iOMA), des optimisations à deux niveaux (Objfind), des solveurs linéaires de type MILP (*Mixed integer linéar programming*) et jusqu'à des solveurs non-linéaires pour les problèmes les plus complexes (EBA) (pour revue sur ce sujet voir [Banga 2008]). Deuxièmement, face à la multitude de méthodes proposées, différents logiciels tels que OptFlux [Rocha *et al.* 2010] ou Fasimu [Hoppe *et al.* 2011] tentent à présent de regrouper l'ensemble de ces méthodes d'analyse dans des frameworks logiciels d'analyse de réseau métabolique.

2.4 Conclusion

La reconstruction des réseaux métaboliques est de nos jours, une thématique qui prend de l'ampleur pour intégrer et donner du sens aux données obtenues grâce aux nouvelles technologies de séquençage. Les méthodes d'analyse bioinformatiques jouent, comme nous l'avons vu, un rôle très important dans ce processus. La méthode FBA représente à l'heure actuelle l'une des principales méthodes pour la vérification des réseaux reconstruits, en étant à présent implémentée dans la plupart des pipelines dédiés à la reconstruction des réseaux *top-down*.
La procédure de reconstruction *bottom-up* ne dispose pas à notre connaissance d'un éventail de

solutions pour la vérification de la fonctionnalité de ces réseaux. Il apparaît également, comme nous le verrons Section 3.4.3, que les méthodes FBA ne conviennent pas parfaitement à ce contexte. Nous avons donc souhaité développer une méthode de simulation et de vérification de ces réseaux, qui s'adapte mieux à ce type de modèle en intégrant d'avantage l'ensemble des données de protéomique et de métabolomique disponibles pour ces réseaux.

Chapitre 3

Metaboflux : théorie et applications

Sommaire

Nous présentons dans ce chapitre l'approche qui a été développée au cours de cette thèse et qui a été implémentée dans le framework Metaboflux [Ghozlane *et al.* 2012]. Metaboflux permet de tester et de prédire la distribution du flux au sein d'un réseau métabolique. Nous verrons en premier lieu les questions qui se posaient pour notre modèle d'étude : *Trypanosoma brucei* qui ont motivé ce développement, puis nous présenterons le principe de la méthode et les résultats obtenus. Nous discuterons enfin d'une comparaison de notre méthode avec la méthode FBA et de l'apport que peut apporter le logiciel Systrip pour la visualisation des prédictions de Metaboflux.

3.1 Contexte biologique

Les trypanosomes sont des protistes unicellulaires jouant le rôle de parasite ubiquitaire des eucaryotes supérieurs, incluant les insectes, les plantes et les mammifères. Parmi les nombreuses espèces appartenant à la famille des trypanosomatides, *Trypanosoma brucei*, *Trypanosoma cruzi* et *Leishmania* spp sont responsables de maladies chez l'Homme. La plupart de ces parasites vivent chez plus d'un hôte au cours de leur cycle de vie et vont rencontrer ainsi des environnements très différents. Ces différentes formes parasitaires ont donc développé des morphologies et des métabolismes très distincts. Nous nous sommes particulièrement intéressés au cours de cette étude à *T. brucei* qui est un parasite des vertébrés et d'un insecte hématophage (Glossina palpalis) en Afrique. Son cycle de vie digénétique est classiquement subdivisé en quatre phases sur la base de critères morphologiques (Fig. 3.1). Les formes procyclique et épimastigote correspondent respectivement aux phases de multiplication du

parasite dans l'intestin puis dans les glandes salivaires de la glossine. Cette multiplication s'effectue par scissiparité (division cellulaire) de l'organisme. *T. brucei* se transforme enfin en Trypomastigote métacyclique, forme infectieuse des vertébrés, injectée lors de la piqûre de la glossine. Chez les vertébrés, les trypomastigotes circulants se multiplient dans le sang, la lymphe et le liquide céphalo-rachidien de l'hôte, engendrant alors la maladie du sommeil (chez l'Homme) et le Nagana (chez les animaux).

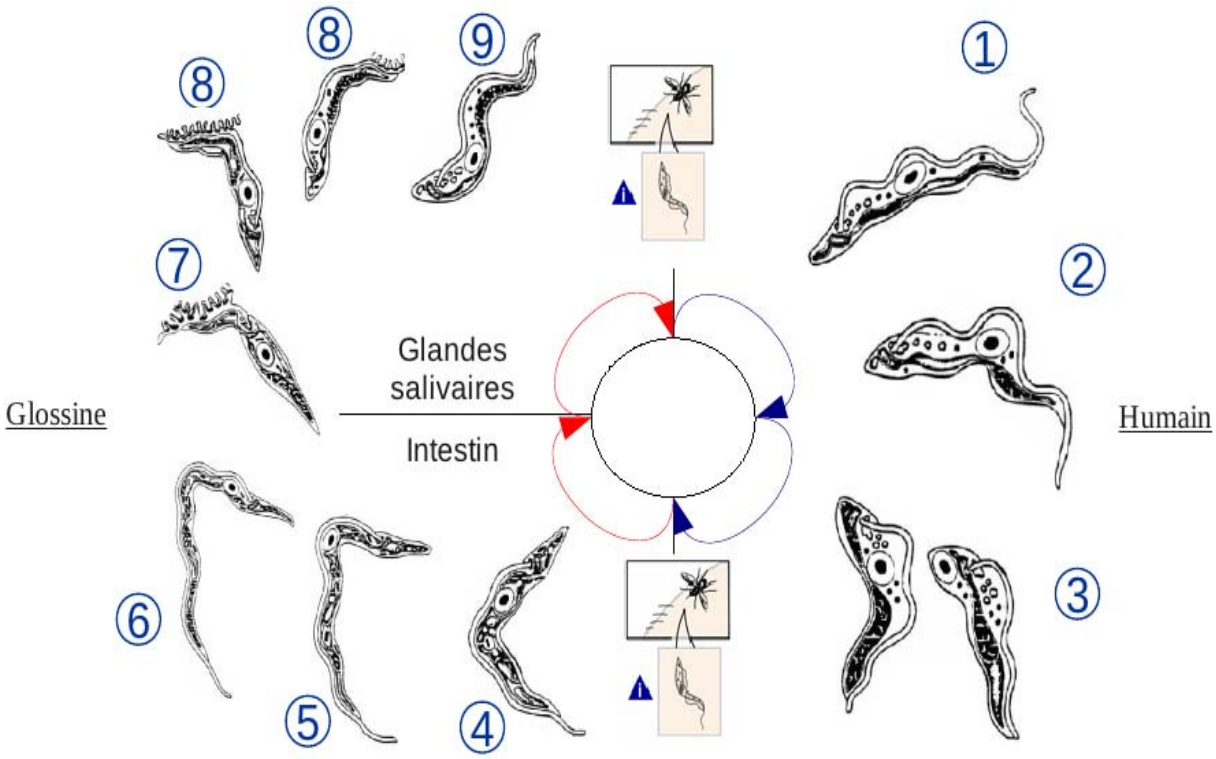


FIGURE 3.1 – Cycle de vie de *Trypanosoma brucei*.
1-2-3. Trypomastigote circulant, 4-5-6. Trypomastigote procyclique, 7-8. Épimastigote, 9. Trypomastigote métacyclique.

Le métabolisme énergétique nécessaire à la vie et au développement (mobilité, multiplication, etc...) du parasite constitue un élément clé pour la compréhension de son fonctionnement. En conditions physiologiques, le métabolisme énergétique du trypanosome varie selon son stade de développement et l'hôte dans lesquels il évolue. *T. brucei* utilise principalement comme source énergétique le glucose disponible dans les fluides des vertébrés et la L-proline disponible chez l'insecte. Trois compartiments cellulaires vont constituer le support de ce métabolisme : le cytosol, la mitochondrie et le glycosome (FIG. 3.2).
Les glycosomes sont des organelles (apparentées aux peroxysomes) spécifiques à l'ordre des Kinétoplastidae où vont se dérouler les 6 ou 7 premières étapes de la glycolyse [Opperdoes et Borst 1977]. Les enzymes glycolytiques y représentent jusqu'à 90% du contenu protéique [Michels *et al.* 2006] et laissent supposer que le principal rôle des glycosomes est d'assurer le métabolisme énergétique. Les tenants et aboutissants de cette compartimentation unique du métabolisme glucidique sont toujours mal connus chez la forme procyclique de *T. brucei*. Cependant, elle semble jouer un rôle essentiel pour la régulation des premières étapes de la glycolyse puisque la ré-allocation de ces enzymes dans le cytosol (par blocage de l'adressage

des protéines glycosomales dans l'organelle) entraine un "effet turbo" létal correspondant à la consommation complète de l'ATP de la cellule par l'activité (non régulée) de l'hexokinase et de la phosphofructokinase [Teusink *et al.* 1998], [Bakker *et al.* 2000]. L'équilibre des rapports ATP/ADP et NAD^+/NADH à l'intérieur des glycosomes est ainsi certainement indispensable à son fonctionnement car aucun transporteur ATP/ADP ou NAD^+/NADH n'a été mis en évidence dans la membrane du glycosome. Ceci implique par voie de conséquence que chaque molécule de NAD^+ ou d'ATP consommée durant les premières étapes de la glycolyse doit être régénérée au sein de l'organelle (Fig. 3.3).

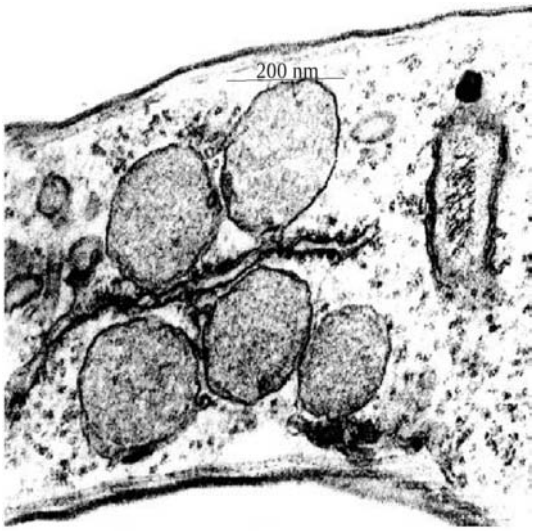


Figure 3.2 – Glycosomes - Vue par microscopie électronique

Chez leur hôte mammifère, les formes sanguines de *T. brucei* (BSF) développent un métabolisme énergétique très simple et bien connu. Ce métabolisme se base sur la conversion du glucose en pyruvate qui est le seul produit excrété de la glycolyse en présence d'oxygène (Fig. 3.3A). En anaérobiose, des proportions équimolaires de pyruvate et de glycérol sont excrétées à partir du glucose (Fig. 3.3B). Dans les deux conditions, l'ATP requis pour le développement du parasite est produit par la pyruvate kinase cytosolique (étape 10 Fig. 3.3). En revanche, la forme procyclique de *T. brucei* (PF), qui évolue dans l'intestin de l'insecte, développe un métabolisme énergétique plus complexe. Lors de la croissance dans un milieu riche, la PF utilise en priorité le glucose comme source de carbone et d'ATP. Au cours de cette glycolyse, le phospho*enol*pyruvate (PEP) est produit dans le cytosol et constitue un point de branchement remarquable (Fig. 3.3C). Il peut être converti en pyruvate, qui est ensuite transformé en acétate dans la mitochondrie [Rivière *et al.* 2004; Millerioux *et al.* 2012]. Le PEP peut aussi ré-entrer dans les glycosomes pour être converti en succinate au sein des organelles ou dans la mitochondrie [Besteiro *et al.* 2005; Coustou *et al.* 2005].
Une topologie du réseau métabolique a été proposée pour la forme procyclique [Bringaud *et al.* 2006, 2010]. La proportion des métabolites finaux excrétés par ce métabolisme a été déterminée par RMN [Coustou *et al.* 2005]. Pour une quantité de glucose donnée, environ la moitié est transformée en succinate dans le glycosome, 20% en succinate dans la mitochondrie, et 30% est transformée en acétate dans la mitochondrie. Nous disposions également de données préliminaires (non publiées) de l'équipe de Frédéric Bringaud, iMET, Bordeaux (en collaboration avec l'équipe de Jean Charles Portais, ENSA, Toulouse) suggérant une possible

équivalence entre le flux passant par les enzymes maliques et la somme des flux résultant de la pyruvate kinase et de la pyruvate phosphate dikinase (étapes 10,16 et 25,26 Fig.3.3). Toutefois, nous ne disposons pas actuellement d'une vue globale et validée expérimentalement des mécanismes moléculaires de ce métabolisme énergétique. L'utilisation de la modélisation est donc nécessaire pour tester si le modèle courant est compatible avec les contraintes de ce métabolisme et dans un second temps, identifier une distribution du flux entre les différentes branches identifiées.
L'objectif principal de ce travail a ainsi été de proposer une analyse permettant de tester le réseau métabolique du trypanosome en intégrant les différentes données expérimentales acquises sur ce réseau : (i) la topologie publiée du réseau [Bringaud *et al.* 2006, 2010], (ii) le maintien des rapports NAD^+/NADH et ATP/ADP à l'équilibre dans les glycosomes et (iii) les données expérimentales.

3.2 Motivations

Plusieurs approches bioinformatiques peuvent être employées pour adresser ce problème de distribution des flux, au premier rang duquel se trouve la méthode FBA et ses différentes évolutions (voir Section 2.3.2). Ces méthodes permettent, comme nous l'avons vu précédemment, de trouver une distribution du flux optimale en considérant la stœchiométrie des réactions comme support du réseau et une fonction objective modélisant par exemple la croissance, la biomasse ou la production d'ATP. Lors d'une reconstruction *top-down* d'un réseau métabolique, une recherche de la bonne formulation de cette fonction est nécessaire de manière à identifier l'objectif dont l'optimisation permet d'obtenir une distribution du flux équivalente à celle mesurée expérimentalement [Feist et Palsson 2010] (voir Section 2.3.2). Des simulations ont également été réalisées pour tester séparément la production de métabolites et déterminer si la formulation du réseau permettait d'observer les proportions observées expérimentalement [Smith et Robinson 2011]. Le réseau métabolique est ainsi considéré comme acceptable s'il est capable de satisfaire toutes ces considérations.
Cependant il serait intéressant de pouvoir procéder à une approche intégrative des contraintes et des mesures expérimentales sans recherche à priori de l'objectif cellulaire. Les modèles testés devraient ainsi répondre favorablement à l'union des contraintes auxquelles ils sont soumis *in vivo*. De plus, la définition d'autres types d'objectifs métaboliques s'avère nécessaire pour prédire une distribution du flux appropriée en cas de considération de réseaux partiels (lors de reconstruction *bottom-up*) et/ou présentant des particularités métaboliques tels que les glycosomes chez *T. brucei*. Ces systèmes peuvent être en effet dirigés par d'autres contraintes, tel que l'homéostasie cellulaire avec aucune consommation ou production nette d'un métabolite/cofacteur clé ($NAD(P)^+$, NAD(P)H, ATP, ADP, etc...) ou le maintien d'une certaine concentration de métabolites dans le système. La prédiction de la distribution des flux devrait prendre en compte ces propriétés, en plus des autres données métaboliques concernant les ratios excrétés de certains produits, car la qualité des prédictions pourrait être favorablement affectée.
La considération combinée de différentes contraintes avec la fonction objective est une tâche difficile. L'augmentation du nombre de contraintes accroit la complexité du problème, où il peut apparaître que certaines combinaisons soient en contradiction. Pour ces raisons, des études récentes ont investi le problème de la distribution du flux au travers de la définition d'une fonction multi-objectif [Oh *et al.* 2009; Nagrath *et al.* 2010]. Ces approches proposent un algorithme pour inférer une distribution du flux pour chacun des objectifs, puis dans un second temps de trouver une distribution du flux consensus par l'estimation d'une distance multi-

paramétrique. Cette distance est ensuite minimisée par une méthode non-linéaire ou linéaire adaptée aux données multi-dimensionnelles. L'objectif est ainsi d'identifier une configuration du système aussi proche que possible de l'état optimal, où l'amélioration de la distribution du flux pour un objectif n'impacte pas les autres objectifs (appelée solution pareto optimale). Ces analyses ont montré que la combinaison de 2, 3 ou 4 objectifs issus des données expérimentales bénéficiait au modèle, permettant ainsi de mieux mimer les conditions présentes dans la cellule [Nagrath *et al.* 2010]. Cependant, aucun outil disponible publiquement n'est lié à ces articles. Enfin, aucune de ces méthodes n'offre, à notre connaissance, d'approche qualitative ou semi-quantitative permettant de prendre en compte de nouveaux types de contraintes.
Nous avons donc implémenté une méthode capable d'intégrer les diverses connaissances biologiques dont on disposait sur ces réseaux. Nous nous sommes particulièrement intéressés à développer une méthode qui ne requière pas de données cinétiques (souvent difficile à obtenir), mais permettant d'intégrer plusieurs contraintes à partir des propriétés connues d'un réseau métabolique. Pour atteindre cet objectif, nous avons opté pour un formalisme différent de FBA, en nous basant sur une méthodologie combinant un simulateur de réseaux métaboliques (basé sur les réseaux de Petri) couplé à un algorithme d'optimisation heuristique, qui est capable de calculer une distribution optimale des flux compte-tenu d'une fonction multi-objectif.

3.3 Principe

3.3.1 Données du modèle biologique

Le métabolisme du glucose des BSF et PF des trypanosomes diffèrent considérablement. Nous verrons en premier lieu le modèle simple et bien connu de la BSF (Fig. 3.3A-B). Pour ce modèle, la distribution du flux entre les deux branches principales a été mesurée expérimentalement. Le second modèle (Fig. 3.3C) décrit le métabolisme du glucose plus élaboré de la PF. Trois branches s'interconnectent dans ce modèle, mais aucune distribution du flux entre ces branches n'a été dressée dans la littérature. Notre approche sera validée au travers du modèle BSF pour lequel nous disposons de données issues de la littérature [Bakker *et al.* 1997, 1999], avant d'analyser la distribution du flux entre les branches principales de la PF.

Trypanosoma brucei forme sanguine (BSF)

Lors de sa croissance en présence d'oxygène, les BSF convertissent le glucose en pyruvate avec intervention de trois compartiments cellulaires (les glycosomes, le cytosol et la mitochondrie) [Bringaud *et al.* 2006]. Ce produit est ensuite excrété de la cellule. Pour de multiples raisons de simplification, les compartiments du cytosol et de la mitochondrie sont fusionnés dans le modèle de la Figure 3.3A. Les sept premières étapes glycolytiques ont lieu dans les glycosomes (étapes 1-7), tandis que les trois étapes suivantes aboutissent au pyruvate dans le cytosol (étapes 8-10). Il est à noter que le glucose (qui est un hexose) est transformé en deux molécules de triose phosphate : le dihydroxyacétone phosphate (DHAP) et du glycéraldéhyde 3-phosphate (G3P). Dans les glycosomes, les molécules d'ATP consommées aux étapes 1 et 3 sont régénérées par l'étape 7 et le NAD^+ consommé à l'étape 6 est régénéré dans l'organelle par la conversion du DHAP en glycerol 3-phosphate (G3P) (étape 11). Ce dernier est ensuite à nouveau transformé en DHAP dans la mitochondrie (étape 12). Ils forment ainsi un cycle DHAP/G3P qui transfère des électrons du DHAP au dioxygène pour produire de l'eau (étapes 12-14). Deux molécules de pyruvate sont ainsi produites à partir d'une molécule de glucose consommée, avec une production nette de deux molécules d'ATP dans le cytosol par la pyru-

vate kinase (étape 10). En anaérobiose les électrons ne pouvant être transférés au dioxygène, ils sont éliminés sous forme de glycérol qui est produit à partir du G3P. Dans ces conditions, le bilan est donc d'une molécule de pyruvate et d'une molécule de glycérol produites pour chaque molécule de glucose consommée, avec une production nette d'une seule molécule d'ATP dans le cytosol (Fig. 3.3B).

Trypanosoma brucei forme procyclique (PF)

Lors de sa croissance, la conversion du glucose par PF en produits excrétés, succinate et acétate, implique des étapes enzymatiques glycosomales, cytosoliques et mitochondriales [Bringaud *et al.* 2006]. Les 6 premières étapes glycolytiques ont lieu dans les glycosomes et consomment 2 molécules d'ATP et 2 molécules de NAD^+ par molécule de glucose consommée (étapes 1-6 Fig. 3.3C), tandis que les trois étapes suivantes, qui mènent à la production de phospho*énol*pyruvate (PEP), ont lieu dans le cytosol (étapes 7-9). Le PEP constitue un point d'embranchement clé dans ce réseau, (i) une branche mène à la production d'acétate dans la mitochondrie (étapes 10, 17-18) [Rivière *et al.* 2004; Millerioux *et al.* 2012] et (ii) les deux autres branches amènent à la production de succinate dans les glycosomes (étapes 19-22) et dans la mitochondrie (étapes 23-24) [Bochud-Allemann et Schneider 2002; Coustou *et al.* 2003]. La branche glycosomale du succinate est critique pour la glycolyse car elle permet de régénérer une molécule d'ATP (étape 19) et deux molécules de NAD^+ (étapes 20, 22) par molécule de succinate produite. Le modèle comporte aussi les réactions cytosoliques et mitochondriales des enzymes maliques (ME, étapes 25 et 26 respectivement) [Coustou *et al.* 2008] qui constituent un lien entre les branches succinate et acétate. Il est aussi important de mentionner que : (i) le cycle de l'acide tricarboxylique (aussi appelé cycle de Krebs) n'est pas utilisé par la PF en tant que cycle, car l'acétyl-CoA n'est pas (ou très peu) converti en citrate, il est essentiellement converti en acétate [Van Weelden *et al.* 2003], (ii) l'essentiel de l'ATP est produit par des phosphorylations réalisées au niveau des substrats issus du glucose, avec une contribution non essentielle de la synthase $F_0/F_1 - ATP$ mitochondrial [Bochud-Allemann et Schneider 2002; Lamour *et al.* 2005; Coustou *et al.* 2008; Millerioux *et al.* 2012], (iii) les molécules de NADH produites dans la mitochondrie par le complexe de la pyruvate deshydrogénase (étape 17) sont régénérées par la fumarate réductase mitochondrial (étape 24), impliquant que l'activité de la chaine respiratoire ne soit pas requise pour maintenir l'équilibre redox mitochondriale associée au métabolisme du glucose [Coustou *et al.* 2005]. Ceci a pour conséquence que le cycle de Krebs, la chaine respiratoire et la synthase $F_0/F_1 - ATP$ ne sont pas inclus dans notre modèle.

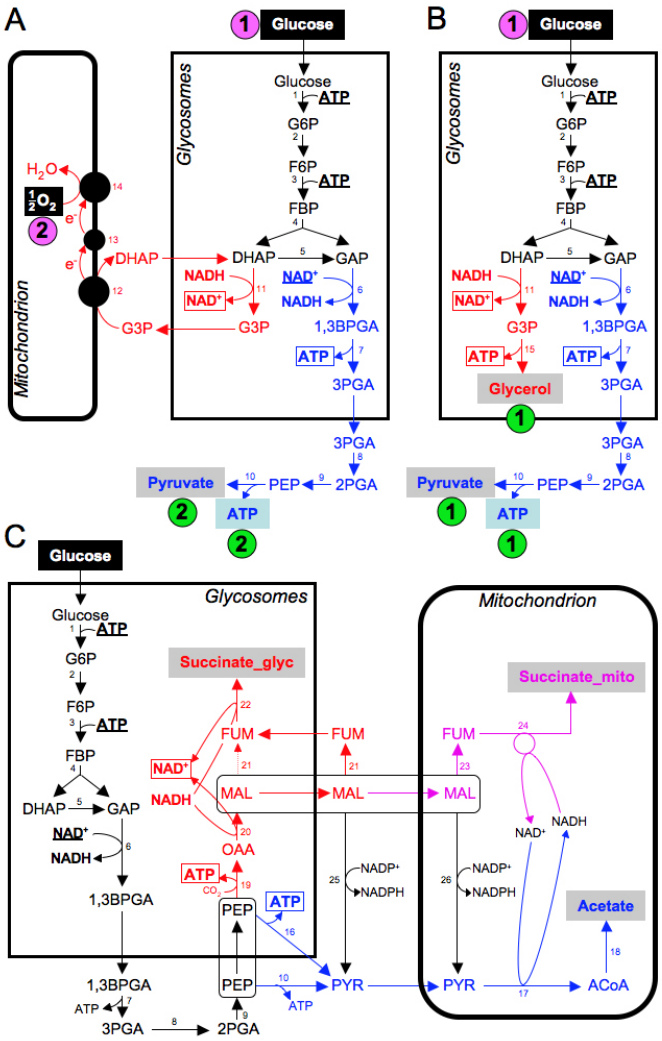


FIGURE 3.3 – Réseau métabolique du métabolisme glucidique des formes sanguines et procycliques de *T. brucei*

Les schémas A et B correspondent au modèle métabolique des formes sanguines de T. brucei (BSF) en condition d'aérobiose et d'anaérobiose, respectivement. Le schéma C représente le modèle métabolique de la forme procyclique (PF) dans un milieu riche en glucose. Les produits excrétés à l'issue de ce métabolisme sont écrits en rouge, vert ou violet. Dans les schémas A et B, les branches métaboliques qui consomment et régénèrent du NAD^+ sont en bleu et rouge respectivement, tandis que le code couleur du schéma C est respectivement bleu, rouge et violet pour les branches qui produisent de l'acétate, du succinate dans les glycosomes et du succinate dans la mitochondrie. Abréviations : 1,3BPGA, 1,3-bisphosphoglycerate ; DHAP, dihydroxyacetone phosphate ; FBP, fructose 1,6-bisphosphate ; FUM, fumarate ; Gly3P, glycerol 3-phosphate ; G3P, glyceraldéhyde 3-phosphate ; MAL, malate ; OAA, oxaloacétate ; PEP, phosphoénolpyruvate ; PYR, pyruvate ; SUC, succinate. Les enzymes individuelles inclues dans le modèle sont : 1, hexokinase : 2, glucose-6-phosphate isomérase ; 3, phosphofructokinase ; 4, aldolase ; 5, triose-phosphate isomérase ; 6, glyceraldéhyde-3-phosphate déshydrogénase ; 7, phosphoglycérate kinase ; 8, phosphoglycérate mutase ; 9, énolase ; 10, pyruvate kinase ; 11, glycosomal glyceraldéhyde-3-phosphate déshydrogenase ; 12, FAD-dependent glycerol-3-phosphate déshydrogenase ; 13, ubiquinone ; 14, SHAM-sensitive alternative oxidase ; 15, glycérol kinase ; 16, pyruvate phosphate dikinase ; 17, pyruvate déshydrogenase complex ; 18, acetate :succinate CoA-transférase and acetyl-CoA thioesterase ; 19, phosphoenolpyruvate carboxykinase ; 20, glycosomal malate déshydrogenase ; 21, fumarase cytoloque et glycosomale ; 22, glycosomal NADH-dependent fumarate réductase ; 23, mitochondrial fumarase ; 24, mitochondrial NADH-dependent fumarate réductase ; 25, enzyme malique cytosolique ; 26, enzyme malique mitochondrial.

3.3.2 Metaboflux

La définition d'un formalisme de modélisation et d'une procédure d'optimisation compatible avec les contraintes de notre système ont constitué une question importante lors de cette étude. Nous décrivons dans cette section l'architecture de Metaboflux (Fig. 3.4), qui est le logiciel développé durant cette thèse, pour la simulation et l'estimation des paramètres d'un réseau métabolique.

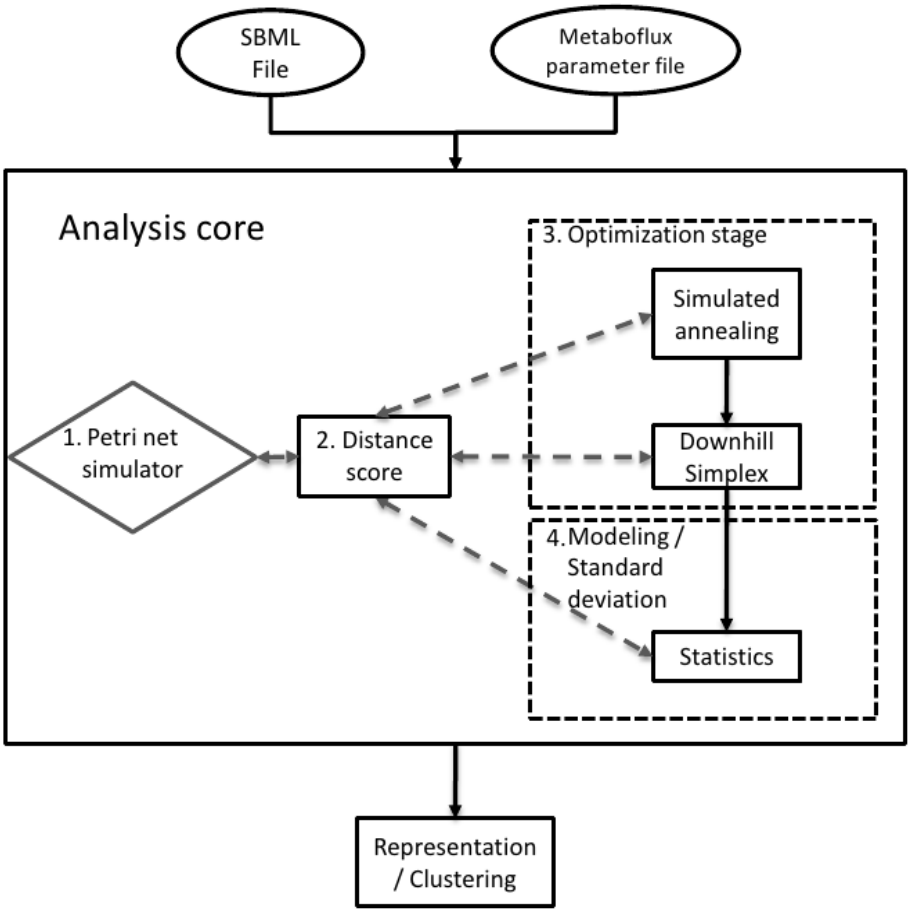


Figure 3.4 – Schéma de la procédure d'analyse de Metaboflux.
Metaboflux prend deux types de fichier en entrée, un fichier SBML et un fichier XML au format de Metaboflux pour indiquer les paramètres de la simulation. La première étape de l'analyse consiste au chargement de N-cœur d'analyse grâce à librairie MPI. Chacun de ses cœurs d'analyse résout le problème d'optimisation à l'aide du recuit simulé et du downhill simplex. Chacune de ses méthodes propose une distribution du flux optimale qui est soumise au simulateur. Le simulateur détermine le marquage final du réseau observé pour la distribution du flux donnée et retourne le score de distance correspondant. Ces étapes sont répétées jusqu'à ce que la procédure d'optimisation converge vers une solution.

Metaboflux est un programme développé pour Linux constitué d'un cœur en C pour la partie calcul, d'une interface en python et d'un script R pour l'analyse statistique des résultats. Quatre librairies sont utilisées par ce programme :

(i) la librairie libSBML [Bornstein *et al.* 2008], permettant de prendre en charge les réseaux métaboliques décrits au format SBML (*System Biology Markup Language*) [Hucka *et al.*

2003],

(ii) la librairie libXML2, pour prendre en charge les fichiers XML décrivant la simulation,

(iii) la librairie GSL (*GNU Scientific library*) [Galassi *et al.* 2009], permettant de réaliser l'optimisation des modèles,

(iv) et la libraire MPI (*Message Passing Interface*), qui est utilisée pour distribuer les calculs.

Metaboflux est automatiquement compilé et installé en utilisant la suite logiciel cmake. Deux versions sont disponibles selon l'utilisation avec son interface ou non (Fig. 3.5).

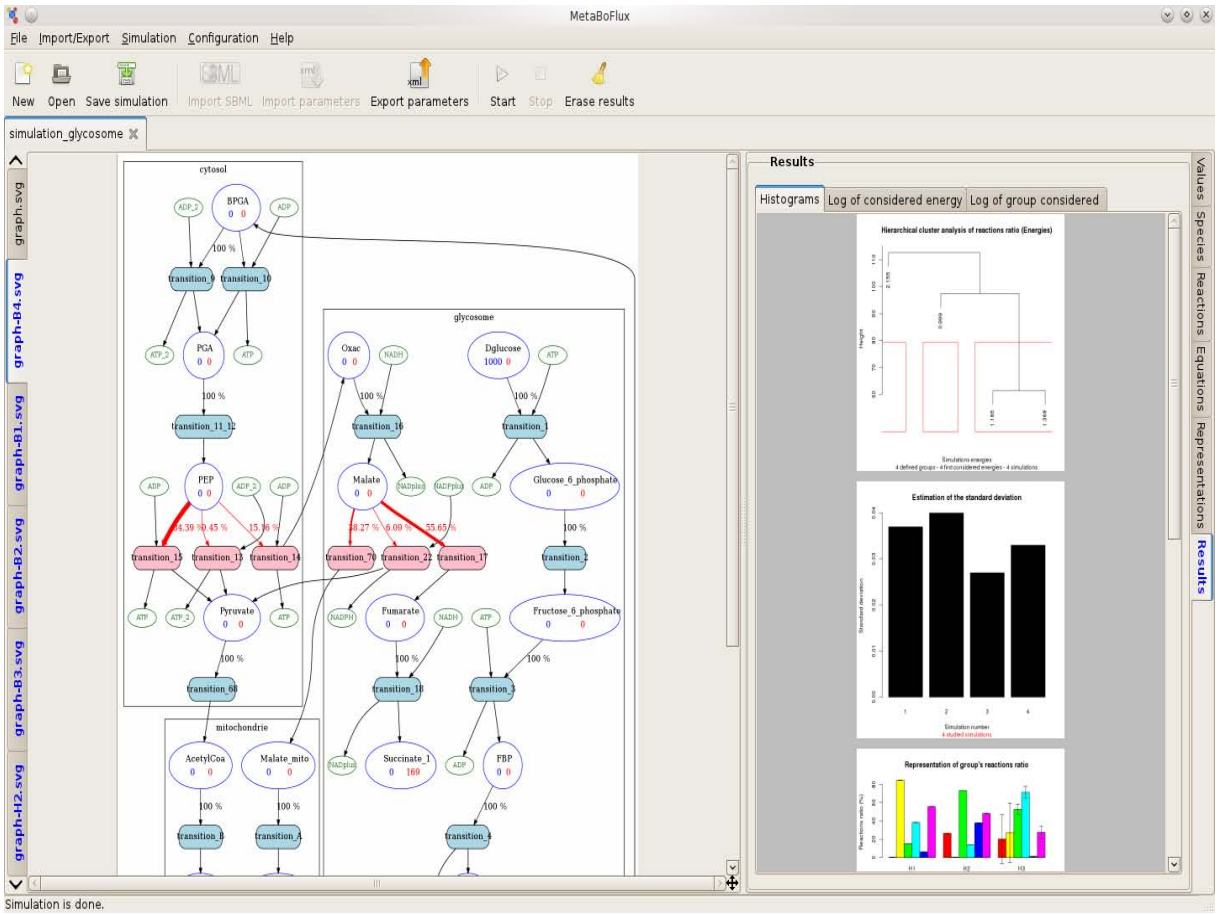


Figure 3.5 – Interface de Metaboflux

Afin de présenter le travail réalisé au sein de Metaboflux, nous définissons en premier lieu le variant des réseaux de Petri que nous avons développé pour modéliser les réseaux métaboliques. Nous présentons ensuite l'algorithme de simulation pour générer les séquences de marquage de ces modèles et la procédure d'optimisation pour estimer les paramètres de nos modèles. Enfin nous présentons une alternative rapide de l'algorithme de simulation que nous avons utilisé pour simuler le modèle durant les étapes d'optimisation.

Flux Petri Net (FPN)

Lors de cette étude, nous disposions de données génomiques pour définir la topologie du réseau (Fig. 3.3) et de données issues d'expériences de RMN pour définir les contraintes de notre modèle (voir Section 3.1). Les réseaux de Petri, en particulier sous la forme de leur

extension GSPN (voir Section 1.2.2), constituent un formalisme particulièrement intéressant pour notre étude. Ces modèles permettent, comme nous l'avons vu,

(i) de modéliser la distribution des ressources par des jetons, ce qui constitue un formalisme convenant parfaitement à la définition de contraintes liées aux proportions des métabolites dans le réseau,

(ii) et de modéliser l'évolution du système selon les paramètres des réactions.

Les profils ainsi obtenus présentent une distribution du flux dépendant donc des propriétés intrinsèques du réseau, à savoir la disponibilité des métabolites, la vitesse des réactions exprimée sous la forme d'un coefficient et la loi de transition employée. Cependant ces paramètres ne correspondent pas à une information disponible pour notre modèle d'étude. Nous ne disposons en effet d'aucune information cinétique pour le modèle PF à laquelle nous pourrions les rapporter. Nous avons donc tenu à simplifier comme FBA notre modèle. Ainsi en combinant les idées de GSPN et de FBA, nous avons étendu le formalisme des PN par l'introduction des poids de flux. Nous avons appelé ce nouveau formalisme *Flux Petri Net* (FPN).

Un *Flux Petri Net* est un graphe biparti orienté et étiqueté, composé comme un PN de places $p \in P$ et de transition $t \in T$. Chaque place du FPN peut contenir des jetons. Le nombre de jetons $i : P \to \mathbb{N}$ correspond au marquage de ce réseau. Le nombre de jetons dans une place p est indiqué par $m(p, i)$ ou $m(p)$. Chaque arc de ce réseau permet de connecter une place et une transition. Les transitions correspondent, quant à elles, à des entités qui pour les FPN peuvent être activées, sélectionnées et tirées, lorsque les pré-places contiennent suffisamment de jetons. Lors de ce processus, les jetons des pré-places sont consommés et ajoutés aux post-places selon les proportions correspondant à la multiplicité des arcs en jeu. En effet, chaque arc est annoté par une valeur de multiplicité $f(p, t)$ ou $f(t, p)$ (pour l'arc entrant ou sortant d'une transition) correspondant à la stœchiométrie de la réaction. Un FPN comporte comme les GSPNs un poids de flux labellisant les transitions. Un poids de flux est une fonction $w : T \to \mathbb{R}^+$ qui assigne un flux strictement positif pour chaque transition du FPN. Pendant la simulation, les poids de flux sont utilisés pour calculer la probabilité d'une distribution parmi les transitions possibles, si une ou plusieurs transitions sont activées selon un certain marquage. Plus la valeur du flux est élevée, plus la fréquence à laquelle cette réaction sera tirée est élevée. Ainsi une forte proportion des métabolites situés sur les pré-places emprunteront la transition du poids le plus élevé. En outre, les FPN ne comportent pas de priorités et de transitions temporelles, qui ne relèvent pas de données dont on dispose généralement sur les réseaux métaboliques. On peut ainsi définir un FPN de la manière suivante :

Définition 25 (***Flux Petri net***). *Un Flux Petri net ($\mathcal{FPN}$) est un 6-tuple (P, T, f, u, v, m_0) où :*

- *P et T sont des ensembles finis, non vides et disjoints. P est l'ensemble des places et T est l'ensemble des transitions.*
- *$f : ((P \times T) \cup (T \times P)) \to \mathbb{N}$ définit l'ensemble des arcs.*
- *$u : T \to E$ est une fonction qui assigne une fonction de poids w_t à chaque transition t, tel que :*
  $E = \cup_{t \in T}\{w_t | w_t : \mathbb{N}^{|\bullet t|} \to \mathbb{R}^+\}$ est l'ensemble des fonctions poids, et $u(t) = w_t$ pour toutes les transitions $t \in T$.
- *$m_0 : P \to \mathbb{N}$ est le marquage initial du réseau de Petri.*

Comme exemple de ce formalisme, nous prendrons le FPN $\mathcal{F}$ présenté Figure 3.6. Dans cette figure, les métabolites (A,B,C,D) sont associés aux places (représentées par des cercles) et les réactions (t_1, t_2, t_3) sont associées aux transitions (représentées par des rectangles). Les transitions sont étiquetées par le numéro de la transition ou de la réaction à laquelle ils

correspondent. La multiplicité des arcs est de 1 par défaut, en l'absence d'étiquetage spécifique.

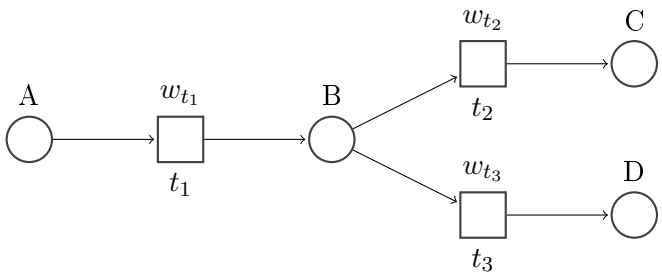


FIGURE 3.6 – Un *Flux Petri Net* $\mathcal{F}$ tels que $P = \{A, B, C, D\}$ et $T = \{t_1, t_2, t_3\}$.

Pour simuler ce réseau, nous échantillonnons une possible série de marquage avec un simulateur stochastique. La simulation démarre avec l'état initial du marquage du réseau. À chaque itération, la boucle de simulation met à jour l'état courant du marquage en sélectionnant les transitions actives et en les tirant. D'un point de vue plus formel, soit l'état courant $i : P \to \mathbb{N}$, on a l'ensemble des transitions actives dans le marquage i, tel que :

$$E(i) = \{t \in T | \forall p \in P, f(p,t) \leq m(p)\} \tag{3.1}$$

La probabilité $w_t(i)$ qu'une transition t soit tirée dans le marquage courant i est donnée par le poids normé :

$$w_t(i) = \frac{w(t)}{\Sigma_{t' \in E(i)} w(t')} \tag{3.2}$$

Ainsi pour une transition t sélectionnée par un choix aléatoire, le marquage suivant i' est défini par soustraction des jetons consommés au(x) pré-place(s) et addition des jetons produits au(x) post-place(s) (voir Section 1.2.2), correspondant à :

$$m(p, i') = m(p, i) - f(p, t) + f(t, p) \tag{3.3}$$

Cette boucle de simulation est répétée jusqu'à ce qu'un nombre maximum d'itérations fixé par l'utilisateur soit atteint ou qu'il n'y ait plus de transition activable dans le marquage courant. Dans le cas des modèles FPN de BSF et PF du trypanosome, toutes les séquences de marquages pouvant être atteintes tendent vers un marquage final où aucune transition n'est activable. En effet, ces FPN ne comportent aucun motif sous la forme de piège étant donné que l'absence de cycle de Krebs fonctionnel chez les trypanosomes. On peut aussi identifier la présence de places qui ne sont liées à aucune transition (pyruvate, glycérol et eau), où les jetons vont s'accumuler jusqu'à la fin de la simulation (appelée *sink place*). On notera enfin la présence d'un motif de type siphon (étapes 11, 12, 13, 14, T1 et T2) qui ne constituera pas non plus d'opposition pour atteindre le marquage final tant que le poids des étapes 5 et/ou 15 est différent de 0. Compte-tenu du marquage initial ou du poids des flux assigné aux transitions, tous les jetons de ces FPN atteindront donc les *sink place* après un nombre fini de transitions activées, ce qui se passe systématiquement dans toutes nos simulations.

Définition des paramètres du FPN

On considère à présent le problème du calcul des paramètres des FPN d'après les contraintes du réseau. Les données dont nous disposons, à savoir les proportions des produits

excrétés et l'équilibre des rapports ATP/ADP et NAD^+/NADH dans le glycosome, correspondent à des observations phénotypiques qui nous donnent non seulement des indications sur l'état d'équilibre du réseau mais aussi sur ses propriétés physiques générales. Nous souhaitons déterminer si une distribution du flux entre les branches du réseau permet d'expliquer ces observations. En conséquence, on s'est intéressé au marquage final du FPN. Compte-tenu du marquage initial de glucose dans le réseau, nous voulons identifier une distribution du flux qui satisfasse l'équilibre de ces rapports et les proportions des produits finaux. Pour un FPN, le marquage final du réseau dépend de la valeur des flux $\overrightarrow{f}$ (correspondant au poids du flux de chaque transitions). Le calcul de la distribution du flux est ainsi associé à l'état attendu du réseau, ce qui correspond donc à chercher une distribution $\overrightarrow{f}$ telle que le marquage final soit en accord avec ce qui a été mis en évidence expérimentalement.
Pour adresser ce problème, nous avons intégré les données expérimentales par une distance $D(\overrightarrow{f})$ entre le marquage final et les quantités attendues, dépendant deux paramètres $d_{x,y}(\overrightarrow{f})$ et $\mathcal{E}(\overrightarrow{f})$. $d_{x,y}(\overrightarrow{f})$ est la distance euclidienne normée permettant d'adresser les contraintes dites "molles" portant sur la différence entre la quantité attendue et simulée :

$$d_{x,y}(\overrightarrow{f}) = \sqrt{\sum_{i=0}^{m} a_i \times \left(\frac{x_i(\overrightarrow{f}) - y_i(\overrightarrow{f})}{y_i(\overrightarrow{f})} \right)^2} \tag{3.4}$$

où x représente la quantité (du produit) i simulée par le FPN compte-tenu de la distribution du flux $\overrightarrow{f}$, y sa quantité attendue et a_i le poids associé à cette contrainte. $\mathcal{E}(\overrightarrow{f})$ représente, quant à lui, les contraintes binaires (ou "dures") sur les quantités attendues :

$$\mathcal{E}(\overrightarrow{f}) = \sum_{j=0}^{n} c_j(\overrightarrow{f}) \tag{3.5}$$

où la variable binaire c_j tient le rôle de facteur de pénalité si la quantité attendue n'est pas obtenue lors de la simulation du FPN :

$$c_j(\overrightarrow{f}) = \begin{cases} 1000, \ si\ y_j\ ne\ satisfait\ pas\ la\ contrainte\ x_j \\ 0, \ sinon. \end{cases} \tag{3.6}$$

Cette valeur élevée de c_j, en cas de pénalisation, permet de contraindre fortement le modèle au respect des contraintes binaires, dont la satisfaction est ainsi considérée comme essentielle lors de l'optimisation. L'ensemble du problème d'optimisation peut ainsi être formulé de la manière suivante :

$$D(\overrightarrow{f}) = d_{x,y}(\overrightarrow{f}) + \mathcal{E}(\overrightarrow{f}) \tag{3.7}$$

où la fonction $D(\overrightarrow{f})$ mesure le degré d'erreur de la prédiction correspondant à la distance entre les quantités prédites et attendues.
Les contraintes ainsi modélisées pour la BSF de *T. brucei* (Fig. 3.3A,B) sont :

- les équilibres glycosomales de l'ATP/ADP et du NAD^+/NADH dans le marquage final,
- une quantité maximale d'ATP cytosolique attendue.

Ces contraintes ne sont exprimées que par des contraintes molles (quantité finale d'ATP glycosomale égale à sa quantité initiale, quantité d'ATP dans le cytosol égale à la quantité maximale d'ADP pouvant être phosphorylée). Pour la PF de *T. brucei* (Fig. 3.3C), selon l'analyse, les contraintes modélisées sont composées :

1. des équilibres glycosomales de l'ATP/ADP et du NAD^+/NADH dans le marquage final,

2. de la proportion attendue de l'acétate excrété, par exemple 50% de la quantité du produit attendue doit être de l'acétate,
3. un intervalle de valeurs acceptables pour le succinate dans le glycosome et la mitochondrie,
4. une valeur de flux minimale de la réaction PEP→OAA (étape 19),
5. une valeur de flux attendue pour la réaction MAL→PYR (étapes 25,26),
6. une valeur de flux équivalente pour les réactions MAL→PYR (étapes 25,26) et PEP→PYR (étapes 10 et 16).

Toutes les contraintes de ce modèle sont à nouveau exprimées par des contraintes molles exceptées les contraintes 2 et 3. Pour la contrainte 2, nous souhaitons en effet forcer le système à produire des quantités comprises dans cet intervalle et n'attribuer une pénalité que si ces dernières sont en dehors de l'intervalle. Pour la contrainte 3, il s'agit de s'assurer qu'un flux minimal passe par l'étape 19 pour que le flux attendu aux étapes 25 et 26 soit possible (voir Section 3.4.2). Nous minimisons ensuite la fonction $D(\overrightarrow{f})$ de manière à obtenir une conjonction entre ces contraintes sur le marquage et les flux.

Optimisation des paramètres du FPN

Pour trouver une distribution des flux minimisant l'erreur de prédiction, nous résolvons ce problème de minimisation en combinant deux optimisations non-linéaires basées sur des méthodes d'optimisation non dérivantes : le recuit simulé (*simulated annealing*) et le *downhill simplex* (Fig. 3.7). Comme notre modèle et nos simulations sont probabilistes par nature, une seule distribution du flux peut générer de multiples marquages finaux. Nous considérons donc l'erreur de prédiction associée à une estimation du marquage moyen obtenu à l'issue de 100 simulations.

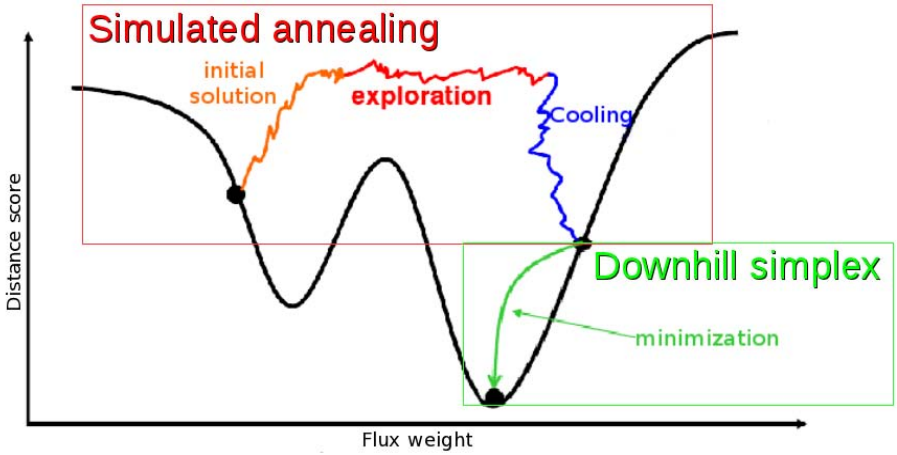


Figure 3.7 – Schéma de la procédure d'optimisation.
L'optimisation est initiée par l'utilisation de l'algorithme du recuit simulé qui à l'aide de la mesure du score de distance va rechercher la meilleur distribution des flux possibles. Les solutions obtenues par le recuit simulé sont ensuite raffinées à l'aide du downhill simplex pour tenter d'atteindre l'optimum global.

La technique du recuit simulé [Kirkpatrick *et al.* 1983] consiste en une optimisation globale dont l'efficacité et la robustesse ont été démontrées [Del Moral et Miclo 1999]. Cet algorithme

permet de rechercher une solution optimale de distribution des flux à partir d'une distribution aléatoire de départ selon la règle de Métropolis :

$$P(\overrightarrow{f'}) = e^{\frac{-(D(\overrightarrow{f'})-D(\overrightarrow{f}))}{kT}} \tag{3.8}$$

où P est la probabilité d'acceptation d'une nouvelle distribution du flux $\overrightarrow{f'}$, D est la fonction de distance, K est la constante de Boltzmann et T est la température du système. Chaque nouvelle distribution du flux $\overrightarrow{f'}$ est calculée à partir de $\overrightarrow{f}$ à laquelle est additionnée ou soustrait une valeur dépendant de la taille du pas défini par l'utilisateur.
Son déroulement s'effectue en 2 étapes selon la température associée au système. Dans la première itération, la température du système est élevée, ce qui correspond à une haute probabilité d'acceptation des nouvelles distributions du flux. Dans cette configuration, le système explore la surface d'énergie correspondant à l'erreur de prédiction obtenue pour chaque distribution du flux. A chaque itération, une descente graduelle de la température est observée selon un paramètre ΔT défini par l'utilisateur. Lorsque la température est basse, le système entre dans la phase de refroidissement. On tend donc vers un optimum dans le voisinage de la solution observée en fin d'exploration. Dans cette étude, nous avons utilisé l'implémentation du recuit simulé de la GSL, que nous avons lancé avec une température de 10000. Nous avons observé que même à cette température (relativement) très élevée le recuit simulé pouvait converger vers un optimum local. Pour palier ce problème, nous avons utilisé MPI pour lancer en parallèle 300 instances individuelles de recuit simulé avec des distributions de flux initiales aléatoires. Pour chacune de ces instances, nous avons retourné un optimum candidat comme point de départ d'une recherche amélioration basée sur l'algorithme du *downhill simplex* implémenté dans GSL. L'algorithme du *downhill simplex* [Nelder et Mead 1965] estime le minimum d'une fonction objective dans un espace à n dimensions en utilisant le concept de simplexe (polytope de $n+1$ sommets). Chaque sommet $(D(\overrightarrow{f_0}), D(\overrightarrow{f_1}), ..., D(\overrightarrow{f_n}))$ représente la distance d'erreur d'une distribution du flux obtenue par un recuit simulé auquel on a ajouté un paramètre aléatoire :

$$\begin{array}{rcl} D(\overrightarrow{f_0}) & = & (w(t_0), w(t_1), ..., w(t_h), w(t_n)) \\ D(\overrightarrow{f_1}) & = & (w(t_0)+s_0, w(t_1), ..., w(t_h), w(t_n)) \\ D(\overrightarrow{f_2}) & = & (w(t_0), w(t_1)+s_1, ..., w(t_h), w(t_n)) \\ ... & & \\ D(\overrightarrow{f_h}) & = & (w(t_0), w(t_1), ..., w(t_h)+s_h, w(t_n)) \\ D(\overrightarrow{f_n}) & = & (w(t_0), w(t_1), ..., w(t_h), w(t_n)+s_n) \end{array} \tag{3.9}$$

Le sommet ayant la distance d'erreur la plus élevée $D(\overrightarrow{f_h})$ est "pivoté". Cette opération est réalisée en trois étapes. La première consiste à calculer le centre de gravité c de tous les sommets du simplexe sauf de $D(\overrightarrow{f_h})$:

$$c = \frac{1}{n}\sum_{j=0}^{n} xj, \ avec \ j \neq h \tag{3.10}$$

Dans la seconde étape, on transforme le simplexe par correction du plus mauvais sommet énergétique. Pour cela, $D(\overrightarrow{f_h})$ est réfléchi par rapport au sommet de meilleure énergie (ici $D(\overrightarrow{f_1})$) et au centre de gravité :

$$\overrightarrow{f_r} = c + (c - \overrightarrow{f_h} + \overrightarrow{f_1}) \tag{3.11}$$

où $\overrightarrow{f_r}$ correspond alors aux paramètres du sommet réfléchi, $\overrightarrow{f_1}$ aux paramètres du meilleur sommet, et $\overrightarrow{f_h}$ aux paramètres du plus mauvais sommet. La troisième étape est une phase de réarrangement du simplexe, par des processus d'étirement, d'expansion ou de contraction, réalisés afin de maintenir la forme du simplexe. Ainsi, à chaque itération, la transformation du simplexe nous permet de nous rapprocher du minimum local. L'algorithme s'arrête lorsqu'il a convergé ou lorsque le nombre maximal d'itérations est atteint (100 itérations sont réalisées au maximum).

Approximation de l'algorithme de simulation de FPN

Pour simuler de manière satisfaisante la distribution des ressources au sein des modèles BSF et PF, nous avons disposé une quantité initiale de 1000 jetons de glucose, point d'entrée du réseau. Cette grande quantité de jetons implique des milliers d'itérations du simulateur de FPN sous sa forme exacte. En effet, une seule itération du simulateur ne va activer et tirer qu'une seule transition du système, ce qui ne bouge qu'un nombre limité de jetons entre les places. Dans les modèles BSF et PF, les stœchiométries sont de 1, excepté pour l'étape 12,13,14 où $2H^+$ et $0.5O_2$ sont consommés. Toutes les transitions ne vont déplacer qu'un seul jeton. Pour un modèle présentant n jetons de glucose en entrée, et en assumant que les places sont liées par $O(n)$ jetons, le nombre d'itérations requis pour atteindre le marquage final est de l'ordre $n \times |T|$ où $|T|$ est le nombre de transitions dans le modèle. De plus, pendant la procédure de minimisation, plusieurs milliers de flux différents sont évalués avant de trouver une solution optimale.
À cause de ce nombre important de simulations réalisé, il est impossible d'utiliser un algorithme de simulation exact. Pour palier à ces problèmes, nous avons implémenté dans Metaboflux un simulateur "glouton" qui approxime ces dynamiques. Cet algorithme est utilisé durant les étapes d'optimisation pour simuler le FPN. Il est basé sur l'idée que les marquages intermédiaires ne sont pas requis pour la procédure de minimisation. A chaque étape, le simulateur glouton tire le plus possible de jetons pour toutes les transitions actives du système, de manière à déplacer le plus de jetons en une étape. Nous identifions trois situations où les multiples transitions peuvent être regroupées et tirées simultanément. Le premier cas apparaît lorsqu'une seule transition est activée. Si on considère l'exemple où les jetons sont présents dans une place et que cette place est l'entrée d'un seul cofacteur. Dans ce cas, la totalité des jetons peut être déplacée aux places de sorties. La seconde situation est obtenue quand de multiples transitions sont activées et compatibles, c'est à dire qu'elles ne partagent pas leurs pré-places. Dans ce cas le simulateur glouton tire le maximum de jetons des transitions activées. Dans cette situation, on calcule alors pour chaque transition le nombre maximum de jetons pouvant être tiré, en tenant compte de la stœchiométrie et du nombre de jetons disponibles sur les pré-places. Soit n correspondant au nombre minimum des maximums de transitions pouvant être tiré, le simulateur glouton détermine le nombre de fois que chaque transition peut être tirée en échantillonnant une distribution multinomiale avec n essais. La probabilité des évènements dépend alors du poids normé des transitions activées. Ces deux heuristiques permettent de réduire 1000 fois le nombre d'itérations requises pour atteindre le marquage final. Nous avons vérifié que ces heuristiques rendent des probabilités de transition par un marquage final comparable à une simulation exacte (d'après un test de Student, les différences entre la marquage final moyen obtenu par la méthode exacte et la simulation approximée n'étaient pas d'un niveau significatif à un risque $\lambda = 0.05$).

3.4 Applications

L'objectif de cette analyse est de déterminer le fonctionnement et la distribution des flux au sein du modèle PF d'après la formulation actuelle des connaissances de ce métabolisme. Avant d'appliquer notre approche au modèle PF, nous avons évalué les performances de Metaboflux pour estimer la distribution des flux du modèle connu BSF. Nous avons ensuite étudié le profil de fonctionnement et la flexibilité de la distribution du flux du modèle PF. Enfin, nous avons comparé l'apport de Metaboflux par rapport à une approche par FBA du problème posé par le modèle PF.

3.4.1 *Trypanosoma brucei* forme sanguine (BSF)

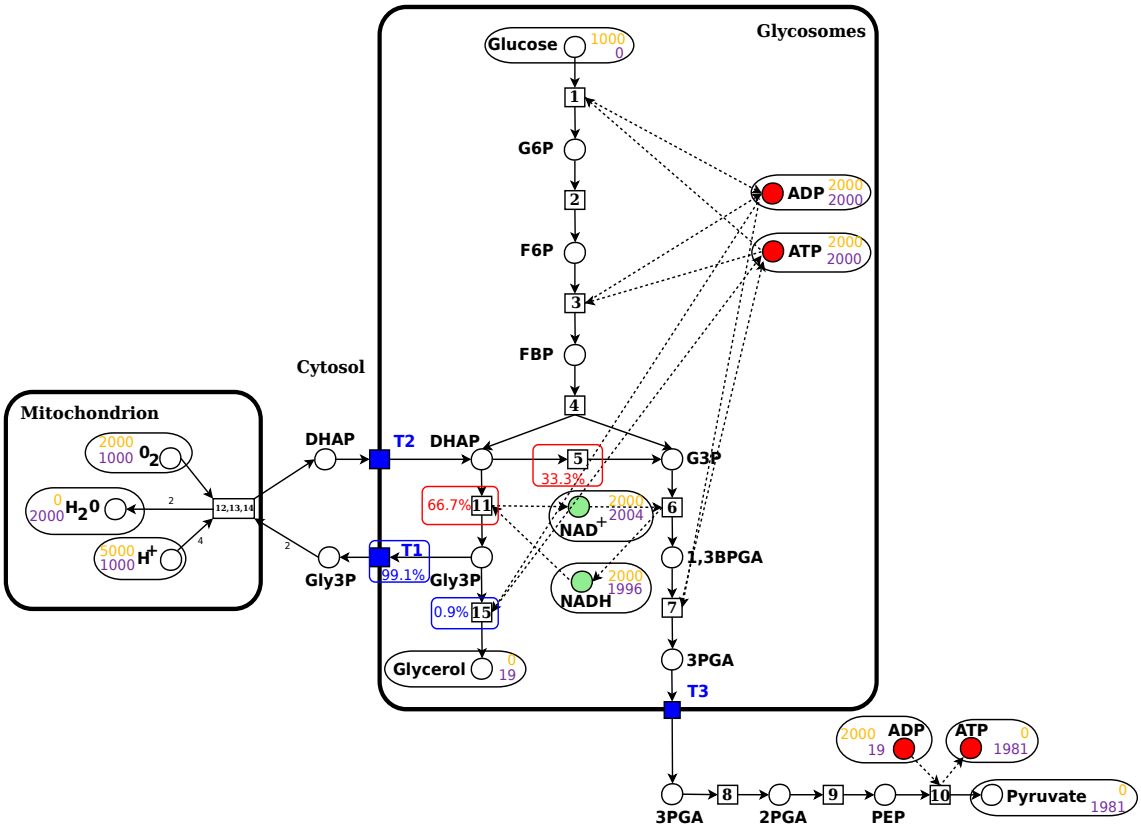
Le métabolisme du glucose de BSF a été caractérisé par deux études [Bakker *et al.* 1997, 1999] au travers desquelles les paramètres cinétiques de ce métabolisme ont été identifiés. Nous avons utilisé Metaboflux pour modéliser ce métabolisme dans les conditions de croissance en aérobiose et en anaérobiose. Pour initier le simulateur, nous avons pris une quantité de 1000 jetons pour le glucose en entrée, de 2000 jetons de O_2 et de H^+ en condition d'aérobiose (en excès) et de 2000 jetons pour les stock d'ADP, d'ATP, de NAD^+ et de NADH dans le glycosome. Deux types de contraintes ont été implémentées dans la fonction objective de Metaboflux :

(i) le maintien des rapports ADP/ATP et NAD^+/NADH à l'équilibre. Comme décrit plus tôt, ces conditions sont vitales pour le parasite puisqu'aucun transporteur glycosomal n'a été identifié pour ces métabolites, ce qui implique leur séquestration dans les glycosomes.

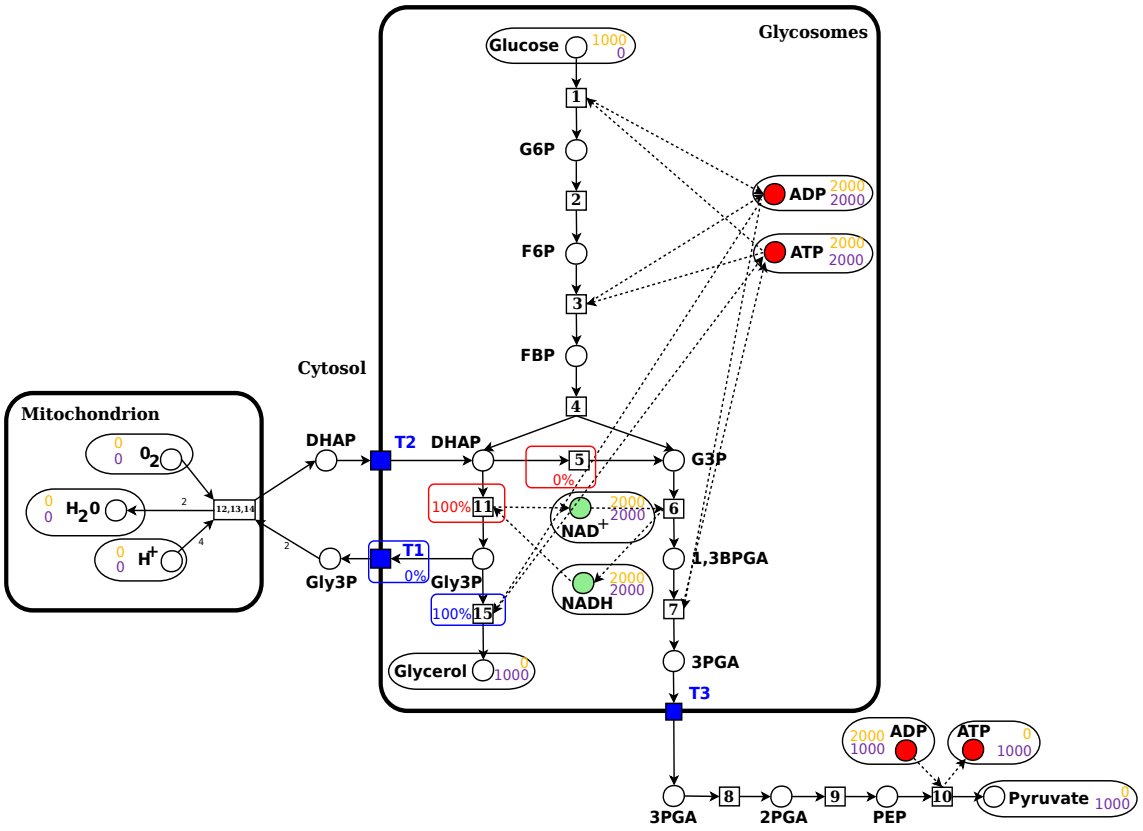
(ii) une contrainte additionnelle est spécifiée pour maximiser la quantité d'ATP synthétisée par la pyruvate kinase cytosolique (étape 10), qui est la seule source connue d'ATP chez les BSF [Bakker *et al.* 1997, 1999].

La simulation et l'optimisation réalisées par Metaboflux (Fig. 3.8) ont permis d'identifier, parmi les meilleurs solutions, une seule distribution du flux satisfaisant ces deux conditions. Trois observations peuvent être réalisées pour ses solutions. Premièrement, en aérobiose, seul le pyruvate est produit, avec un ratio de 2 molécules produites par molécule de glucose consommée (la quantité de jetons donnée par Metaboflux est le double du stock de glucose initial), tandis qu'aucune molécule de glycérol n'est produite dans les glycosomes (Fig. 3.8a). Deuxièmement, en condition d'anaérobiose, Metaboflux prédit une production équimolaire de pyruvate et de glycérol (Fig. 3.8b). Il indique également une production de presque deux molécules d'ATP (1995 jetons Fig. 3.8a) en condition d'aérobiose et d'une molécule d'ATP (1000 jetons Fig. 3.8b) en anaérobiose par molécule de glucose consommée, ce qui correspond parfaitement aux ratios observés expérimentalement pour le modèle BSF. Enfin, la distribution du flux aux points d'embranchement du DHAP et du glycérol 3-phosphate (Gly3P) est également cohérente avec le modèle. En condition d'anaérobiose, des flux identiques sont requis entre les branches du glycérol et du pyruvate (rouge et bleu Fig. 3.3B respectivement) pour maintenir les rapports ATP/ADP et NAD^+/NADH à l'équilibre. En conséquence, toutes les molécules de DHAP doivent être transformées en glycérol pour maintenir le rapport ATP/ADP à l'équilibre (Fig. 3.8b). En condition d'aérobiose, la distribution du flux est différente aux points d'embranchement. Le cycle du DHAP/Gly3P implique en effet qu'aucun glycérol ne soit produit, avec une forte contribution de l'export de Gly3P comme prédit par Metaboflux (99.8% de flux Fig. 3.8a). Au nœud du DHAP, une part importante de DHAP est convertie en G3P, puisque le DHAP produit à partir de Gly3P (étapes 11, 12 Fig. 3.8a) ré-entre dans les glycosomes. L'ensemble de ces résultats est en accord avec les données ex-

périmentales des BSF dans les deux conditions d'incubation et valident Metaboflux comme outil pour étudier la distribution des flux du métabolisme du glucose chez les trypanosomes.



(a) BSF en condition d'aérobiose



(b) BSF en condition d'anaérobiose

FIGURE 3.8 – FPN représentant la distribution du flux prédite pour les BSF du trypanosome en condition d'aérobiose (a) et d'anaérobiose (b).
Le métabolisme du glucose des BSF implique 3 compartiments cellulaires : la mitochondrie, les glycosomes et le cytosol. Les cercles représentent les métabolites, les carrés représentent les réactions, et les carrés bleus représentent les pores glycosomaux qui permettent (comme pour les peroxisomes) l'échange de petites molécules [Gualdron-López et al. *2012]. La stœchiométrie est de 1 par défaut, sauf pour le H^+, H_2O et O_2 qui ont été multipliées par 2, de manière à obtenir respectivement 4,2 et 1. La quantité des métabolites du FPN est spécifiée en orange pour l'état initial et violet pour l'état final. Le pourcentage de flux indiqué aux points d'embranchement du réseau est donné en rouge pour le nœud DHAP (étape 5 contre l'étape 11) et en bleu pour le nœud Gly3P (pore T1 contre l'étape 15). Les valeurs en pourcentages aux points d'embranchement correspondent à des sous-ensembles (-parties) d'un flux de 100% donné pour l'embranchement où deux sous-parties vont consommer du même métabolite. Par exemple, l'étape 11 et 5 utilisent toutes deux le DHAP, respectivement 66.7% et 32.2% du DHAP emprunte chacune de ses voies.*

3.4.2 *Trypanosoma brucei* forme procyclique (PF)

Analyse du fonctionnement du modèle PF

À partir des données du modèle PF, nous avons réalisé une étude du flux (Fig. 3.9). Nous avons soumis le système à 3 types de contraintes :

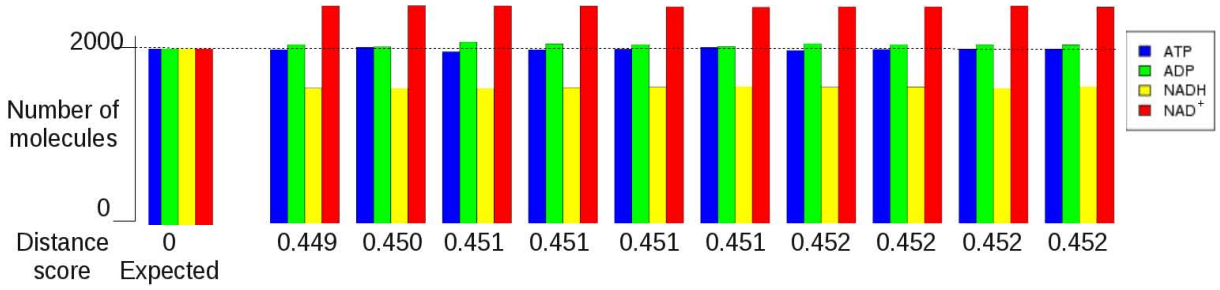
(i) un maintien des rapports ATP/ADP et NAD^+/NADH à l'équilibre dans les glycosomes,

(ii) une distribution des produits excrétés pour une quantité de glucose donnée correspondant à 50% de succinate dans le glycosome, 20% de succinate dans la mitochondrie, et 30% d'acétate transformé dans la mitochondrie. Nous simplifions ici volontairement les contraintes en ne prenant que les mesures observées par [Coustou *et al.* 2005], qui correspondent aux mesures actuellement obtenues dans l'équipe de Frédéric Bringaud.

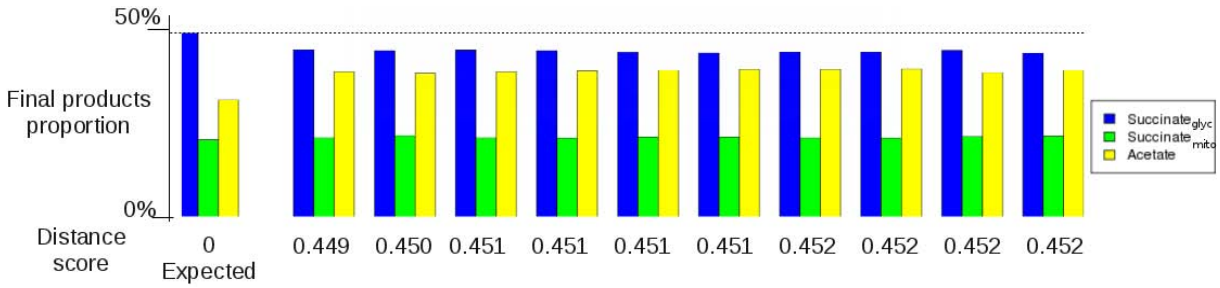
(iii) Une équivalence entre le flux passant par les enzymes maliques et la somme des flux résultats de la pyruvate kinase et de la pyruvate phosphate dikinase (étapes 10,16 et 25,26 Fig.3.3). Cette hypothèse a été déduite à partir des données de spectrométrie de masse non publiées par les équipes de Frédéric Bringaud et Jean-Charles Portais.

Nous avons considéré ici la distribution des ressources au niveau des rapports ATP/ADP et NAD^+/NADH et la proportion des métabolites finaux pour les 10 meilleures solutions identifiées par Metaboflux ($\Delta D < 0.1$).

Pour toutes ces simulations, la contrainte sur l'enzyme malique a été satisfaite (données non publiées). Les scénarios identifiés nous indiquent que le rapport NAD^+/NADH n'est pas équilibrée dans le glycosome pour cette formulation du schéma (Fig. 3.9a). Il y a une surproduction de 23.3% de NAD^+ lors de la dégradation du glucose. La proportion des produits finaux ne correspond pas non plus à la proportion attendue (Fig. 3.9b). En effet, la production de succinate dans le glycosome est diminuée de 15% au profit d'une augmentation de la production d'acétate. Seul le rapport ATP/ADP est équilibré dans de ce modèle. Deux raisons permettraient d'expliquer ce résultat : soit la formulation du modèle ne permet pas de répondre aux contraintes imposées à ce système, soit une ou plusieurs contraintes formulées n'existent pas.



(a) État des proportions en ATP/ADP et en NAD^+/NADH

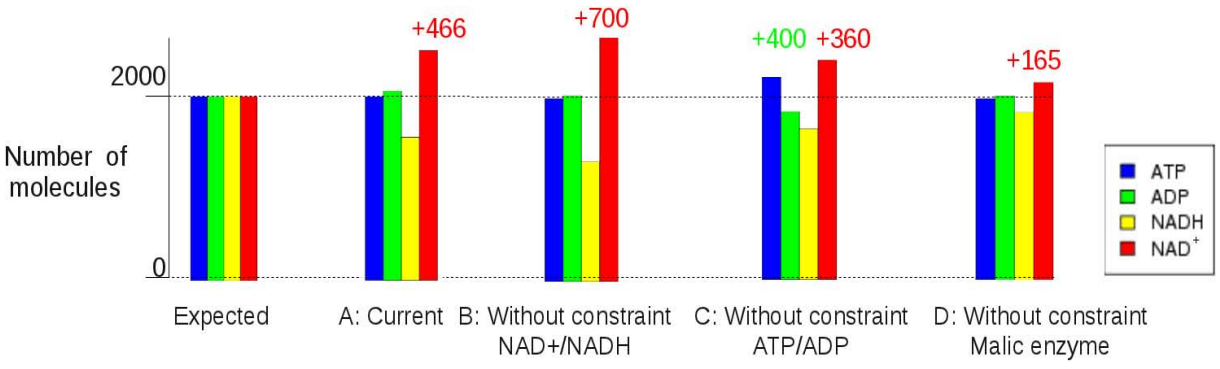


(b) Proportions des métabolites finaux

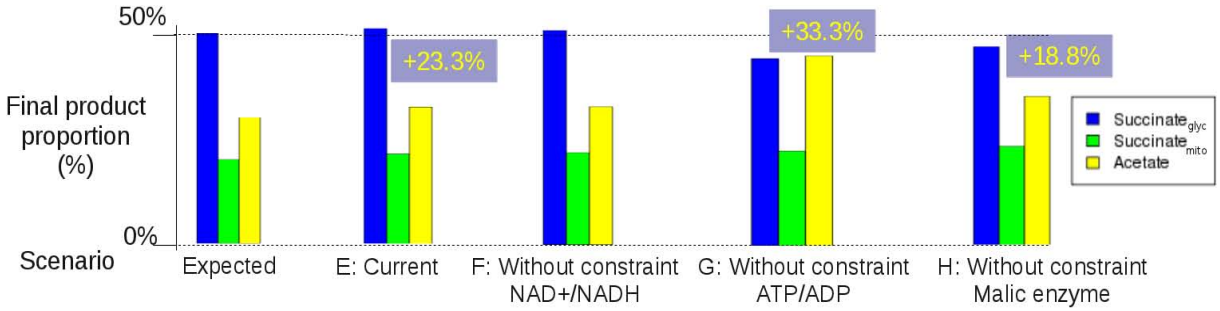
Figure 3.9 – Histogramme représentant le marquage final du FPN du modèle PF obtenu pour (a) l'ADP, l'ATP, le NAD^+ et le NADH et (b) le succinate glycosomal, mitochondrial et l'acétate. *Le modèle a été simulé avec 3 contraintes : 1) l'équilibre des rapports ATP/ADP et NAD^+/NADH, 2) des proportions fixes d'acétate et succinate excrétées correspondant à 50% de succinate dans le glycosome, 20% de succinate dans la mitochondrie et 30% d'acétate, et 3) l'équivalence du flux entre les enzymes maliques et la pyruvate kinase et la pyruvate phosphate dikinase. Les dix meilleures solutions identifiées par Metaboflux sont considérées par rapport au résultat attendu.*

Détermination de l'impact des contraintes

Pour déterminer si le déséquilibre détecté est le résultat de la formulation du modèle ou si une des contraintes formulées s'oppose à l'équilibre du rapport $NAD^+/NADH$ dans le glycosome, nous avons analysé la contribution de chaque contrainte dans le modèle PF (Fig. 3.10). Les résultats des simulations A-D et E-H de la Figure 3.10 ont été obtenus en retirant, une à une, les trois contraintes formalisées dans la fonction objective. Nous avons ensuite comparé les résultats par rapport au modèle PF soumis aux trois contraintes initiales.
En accord avec les profils obtenus précédemment, le retrait de la contrainte sur le rapport $NAD^+/NADH$ (Fig. 3.10B,F) a accentué le déséquilibre de ce rapport. La proportion des métabolites finaux a, cependant, été rétablie à son niveau requis. Les contraintes sur le rapport ATP/ADP et sur l'enzyme malique ont aussi été satisfaites (données non publiées). Le rapport $NAD^+/NADH$ a donc un impact sur la disproportion des métabolites finaux dans ce modèle. Le retrait de la contrainte sur le rapport ATP/ADP (Fig. 3.10C,G) a engendré une diminution du déséquilibre du rapport $NAD^+/NADH$. La surproduction d'acétate n'est pas compensée et est encore accentuée par un facteur de 10% par rapport au modèle initial. Un déséquilibre important du rapport ATP/ADP peut être constaté. Enfin, le dernier modèle lève la contrainte de l'enzyme malique (Fig. 3.10D,H) et réduit ainsi, de 15% la surproduction de NAD^+, et de 18% la surproduction d'acétate par rapport au modèle initial, sans perturber le rapport ATP/ADP.
Deux scénarios semblent donc particulièrement favorables pour la formulation d'un nouveau modèle respectant toutes les contraintes du système. Le premier serait en faveur d'un modèle sans la contrainte sur le rapport $NAD^+/NADH$, dont le retrait pourrait se justifier par l'intégration de voies métaboliques encore non identifiées. Dans un second modèle, il serait possible de retirer la contrainte sur l'enzyme malique sans modifier la structure du réseau. Suite à ces premières conclusions, de nouvelles expérimentations sur le flux de la PF de *T. brucei* ont été réalisées par l'équipe de Frédéric Bringaud et ont permis de mettre en évidence que l'importance de la contrainte sur le flux des enzymes maliques était effectivement surestimée. Cette contrainte est donc à présent retirée du modèle PF.



(a) État des proportions ATP/ADP et $NAD^+/NADH$



(b) Proportions des métabolites finaux

FIGURE 3.10 – Histogramme représentant le marquage final du FPN du modèle PF selon la condition retirée du modèle.
Les contraintes utilisées pour la Figure 3.9 sont les mêmes que celles qui sont appliquées ici. Le nombre de molécules supplémentaires par rapport au nombre attendu est représenté en rouge pour le NAD^+ et en vert pour l'ADP. La proportion d'acétate supplémentaire par rapport à celle attendue est représentée en jaune.

Analyse de la flexibilité du flux du modèle PF

Dans un second temps, nous nous sommes intéressés à la flexibilité du métabolisme du glucose du modèle PF. Différentes études de ce métabolisme [Van Weelden *et al.* 2003; Coustou *et al.* 2006] ont en effet mis en évidence une variation de la proportion des produits excrétés (le succinate et l'acétate). Le ratio succinate/acétate excrété varie dans ces études entre 2.6 (correspondant 28% d'acétate sur la proportion totale de produits excrétés) à 0.8 (55% d'acétate) et atteint même 0.43 (70% d'acétate) pour les mesures obtenues par l'équipe de Frédéric Bringaud (données non publiées). Ceci suggère une forte flexibilité dans la distribution du flux entre les différentes branches du réseau, malgré les fortes contraintes pour maintenir les rapports ADP/ATP et NAD^+/NADH à l'équilibre dans le glycosome.
L'objectif a donc été d'analyser avec l'aide de Metaboflux le degré de flexibilité de la production des produits excrétés, succinate et acétate, est flexible chez le modèle PF. Pour cela, nous avons testé le système avec 2 types de contraintes différentes :

(i) le maintien des rapports ADP/ATP et NAD^+/NADH à l'équilibre. Les molécules d'ATP aux étapes 1 et 3 dans les glycosomes doivent être régénérées par les étapes 16 et 19 et de manière similaire, les molécules de NAD^+ réduites à l'étape 6 glycosomale doivent être réoxydées par les étapes 20 et 22 comme dans le modèle BSF (FIG. 3.8c).

(ii) Entre 56 et 86% du total de succinate excrété doit être produit par les glycosomes et à l'inverse, entre 14 et 44% doit être produit dans la mitochondrie. Contrairement à l'étude du profil de fonctionnement (Section 3.4.2), nous prenons les bornes maximales et minimales de production des branches de ce métabolisme, déduites à partir des analyses génétiques de [Coustou *et al.* 2005].

Dans cette analyse, on considère qu'une différence de +/- 9% entre les distributions prédites et attendues est acceptable. D'un point de vue expérimental, cette variabilité de 9% correspond à des différences non-significatives. D'un point de vue numérique, nous avons déterminé que l'intervalle de valeurs retourné par la fonction de distance quand on permet cette variabilité entre le marquage prédit et le marquage attendu, correspond à une distance comprise entre 0 à 0.2 (FIG. 3.11). Dans cette analyse, on a donc fixé un seuil de 0.2 pour estimer les distributions de flux qui sont proches des résultats expérimentaux.

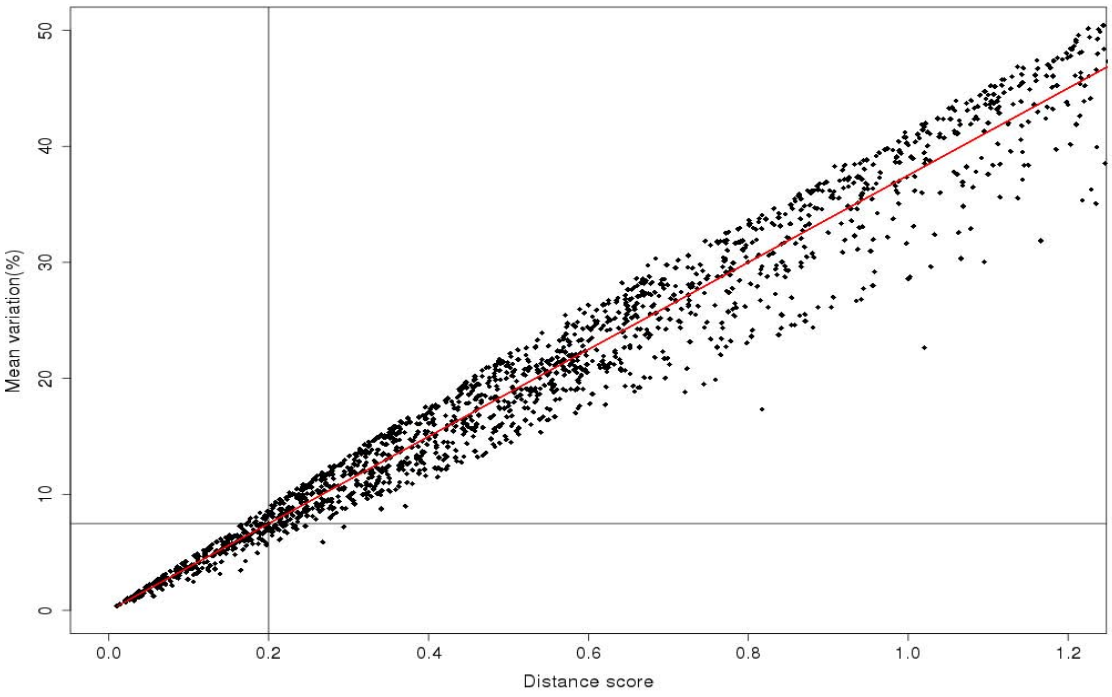


FIGURE 3.11 – Graphique représentant le pourcentage de variation moyen observé entre ce qui est attendu et le marquage final du FPN selon le score distance.
Les scores de distances considérés proviennent de l'ensemble des distributions de flux du modèle PF. Nous avons estimé le pourcentage de variance moyen entre le marquage attendu et le marquage final observé. Il apparaît qu'une variation de distribution de 9%, correspond à un score de distance compris entre 0 et 0.2.

Après la modélisation de ces contraintes dans Metaboflux, l'optimisation a été lancée avec une proportion fixe d'acétate excrété variant de 0 à 100% avec un pas de 5%. Le profil ainsi obtenu est représenté Figure 3.12. Nous pouvons observer ici que toutes les contraintes du système sont satisfaites, avec un score inférieur à 0.2, pour des proportions d'acétate excrété (d'après le nombre de jetons d'acétate par rapport au nombre total de jetons, donné par la somme des jetons de succinate et d'acétate) comprises entre 26 et 95%. Trois observations peuvent être alors effectuées en considérant ces résultats. Premièrement, cet intervalle d'acceptation correspond à ce qui est observé expérimentalement, puisque le pourcentage d'acétate excrété à partir du métabolisme du glucose varie entre 26 et 70% selon les analyses. Deuxièmement, une validation supplémentaire de ce profil nous est donnée en considérant les données expérimentales obtenues pour les formes "mutantes acétate et succinate" de *T. brucei* obtenues en laboratoire. Le mutant succinate (qui ne produit quasiment plus de succinate à partir du glucose) est obtenu par délétion du gène de la phospho*enol*pyruvate carboxykinase (PEPCK, étape 19 Fig.3.3) [Coustou *et al.* 2008; Ebikeme *et al.* 2010] ou par le blocage par ARN d'interférence de la fumarate réductase (étape 22 et 24 Fig.3.3) [Coustou *et al.* 2005]. Ce mutant est considéré comme favorable d'après le profil généré par Metaboflux, et il survit également *in vivo* après le blocage de sa production de succinate. Le mutant acétate (qui ne produit plus d'acétate) est, quant à lui, obtenu par l'inactivation de la succinate CoA-transferase et le blocage par ARN d'interférence de l'acetyl CoA thioesterase (étape 18 Fig.3.3) [Millerioux

et al. 2012]. On peut voir sur le profil qu'une faible production d'acétate est défavorable à cause du déséquilibre du rapport NAD^+/NADH, et que le mutant ne survit également pas *in vitro*. Enfin, de manière intéressante, il apparaît que la contrainte des rapports ATP/ADP et plus particulièrement NAD^+/NADH ont un fort impact dans l'établissement de cet intervalle de valeurs. En effet, l'intervalle d'acceptation correspond exactement à la région sur la courbe où ces équilibres sont parfaitement respectées (ratio de 1 pour les rapports). Metaboflux met ici en évidence que la variation de la proportion de succinate mitochondrial et glycosomal, dans l'intervalle maximum et minimum de ces flux, suffit pour obtenir l'équilibre des rapports. Le modèle confirme ainsi que le métabolisme du glucose est très flexible en terme de production d'acétate et de succinate, malgré les fortes contraintes imposées par l'organelle sur les rapports ATP/ADP et NAD^+/NADH.

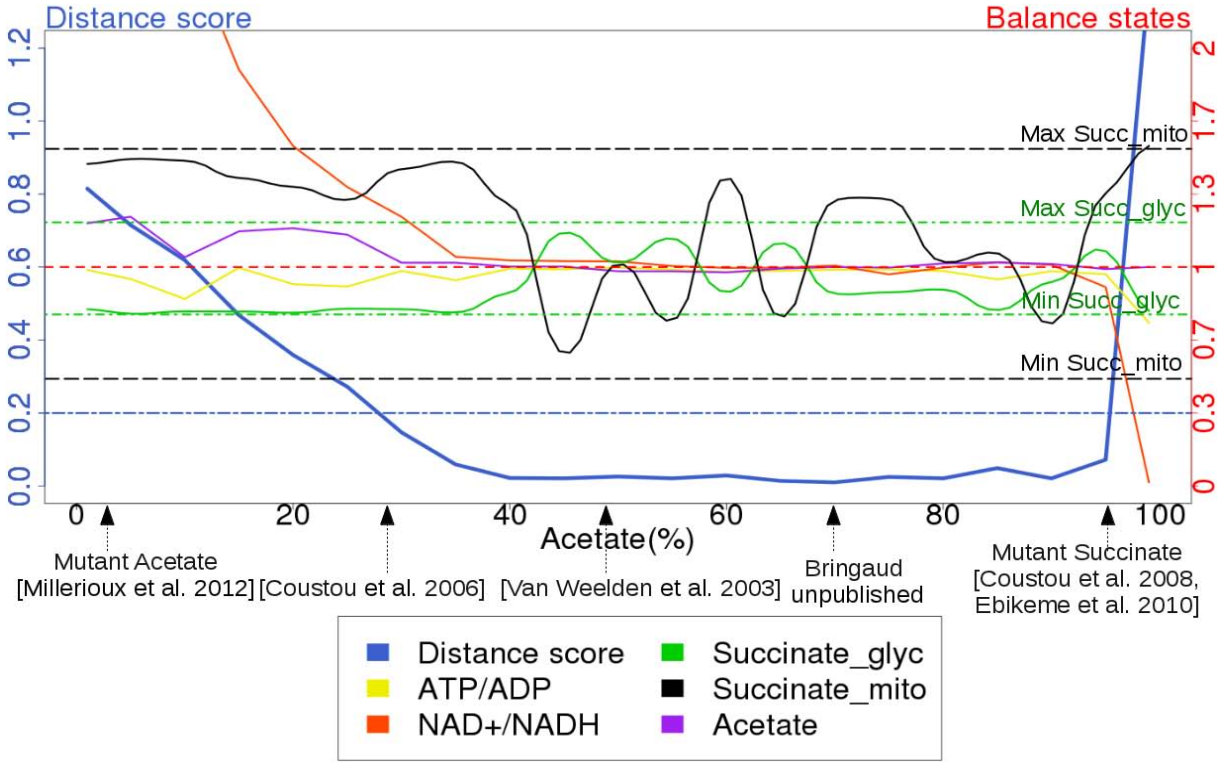
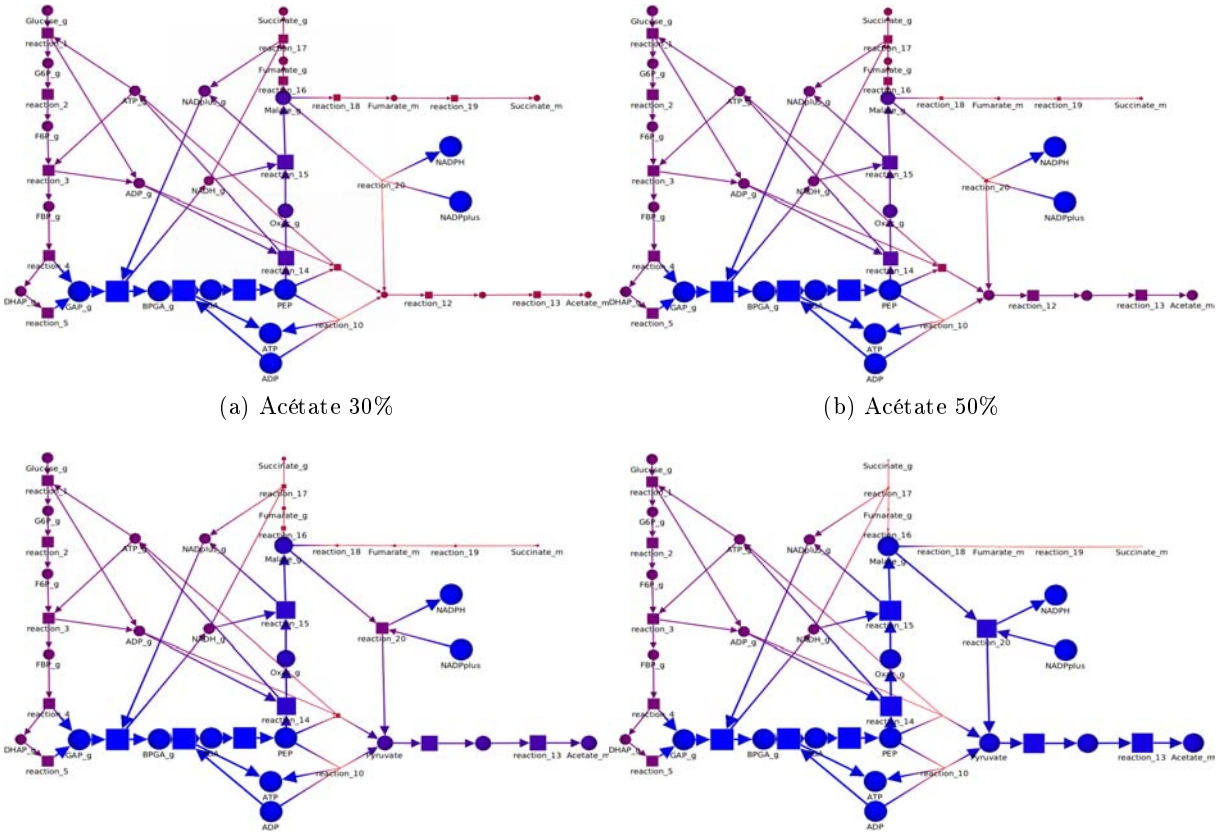


FIGURE 3.12 – Courbes représentant la valeur du score de distance, l'état des rapports ATP/ADP et NAD^+/NADH et les proportions des produits excrétés observées par le modèle selon la variation du ratio acétate/succinate attendue.

Ces courbes résument les proportions de plusieurs métabolites en fonction du ratio acétate/succinate. Pour cette expérience, le modèle a été simulé avec 3 contraintes : 1) la contrainte sur les proportions d'acétate/succinate dans un intervalle compris entre 1 et 99%, avec un pas de 5% ; 2) les proportions entre les deux marquages du succinate sont contraintes par des ratios compris entre 56-86% pour le succinate_glyc contre 14-44% pour le succinate_mito ; et 3) les rapports ATP/ADP et NAD^+/NADH sont requises à l'équilibre. Le score de distance (courbe bleu) est fonction de la satisfaction de ces contraintes. Quand le score de distance est inférieur à 0.2, nous considérons que toutes les contraintes sont satisfaites. On peut voir que les contraintes sont satisfaites pour des proportions d'acétate/succinate comprises entre 26-95%. Cela correspond à une flexibilité du flux de 69%. Dans cette aire de flexibilité, nous retrouvons la proportion acétate/succinate observée par 2 publications. Ces résultats expérimentaux sont indiqués par des flèches noires.

Identification des éléments de flexibilité au sein du modèle PF

Pour déterminer les raisons possibles de cette flexibilité, nous avons dans un premier temps analysé de manière visuelle la distribution du flux obtenue par le modèle précédent. Pour cela, nous avons visualisé à l'aide du logiciel Systrip [Dubois *et al.* 2012] qui est basé sur le logiciel Tulip [Auber 2003], la distribution du flux au sein du modèle PF pour chaque proportion fixe d'acétate excrété. Les graphes obtenus sont représentés Figure 3.13. L'élément principal pouvant être observé entre ces différentes distributions concerne la distribution du flux au travers des enzymes maliques (étape 20 de ce modèle correspondant aux étapes 25-26 Fig. 3.3). Il apparaît en effet que plus la proportion d'acétate excrétée est élevée, plus le flux de ces enzymes est important. Ceci suggère que la variation de distribution du flux entre les branches du succinate et de l'acétate serait liée à l'activité des enzymes maliques.



(a) Acétate 30% (b) Acétate 50%

(c) Acétate 70% (d) Acétate 90%

Figure 3.13 – Distribution du flux au sein du modèle PF pour des ratios d'acétate excrété de 30, 50, 70 et 90%

Les graphes représentés montrent la distribution du flux du modèle de la Figure 3.12 pour des ratios d'acétate excrétés de 30, 50, 70 et 90%. Ces profils ont été obtenus en utilisant le logiciel Systrip. La taille et la coloration des nœuds dépendent ici de la proportion des métabolites (en %) consommés par chaque réaction. Les réactions peu utilisées sont représentées en rouge, tandis que les réactions très utilisées sont représentées en bleu.

Pour vérifier cette première observation, nous avons inclus deux contraintes additionnelles au modèle pour ajuster le flux passant, d'une part par la PEPCK (étape 19), et d'autre

part, par les enzymes maliques (ME) (variation comprise entre 1 et 80% avec un pas de 5%), constituant le pont entre les branches succinate et acétate (étapes 25 et 26). La Figure 3.14 montre la courbe de distance selon la proportion d'acétate produite et pour différents flux imposés aux ME. Il apparaît que la contrainte sur le flux des ME réduit la flexibilité du système. Par exemple pour un flux de 25% au niveau des enzymes ME, la production d'acétate compatible avec les contraintes est seulement comprise entre 42 et 69% (Fig. 3.14). On constate également qu'il y a une relation directe entre le flux des ME (entre 1 à 80%) et la proportion d'acétate que peut excréter le modèle. Cependant, au delà de 80%, l'augmentation du flux des ME n'augmente pas la flexibilité du système. La corrélation entre la production d'acétate et la distribution du flux aux ME, implique que le flux des premières étapes de fermentation succinique (étape 19) augmente proportionnellement (Fig. 3.15). L'ensemble de ces résultats suggère ainsi que la flexibilité de la distribution du flux entre les branches du succinate et de l'acétate est considérablement accrue par l'activité des ME, qui sont des étapes essentielles pour la viabilité de la PF du trypanosome [Coustou *et al.* 2008].

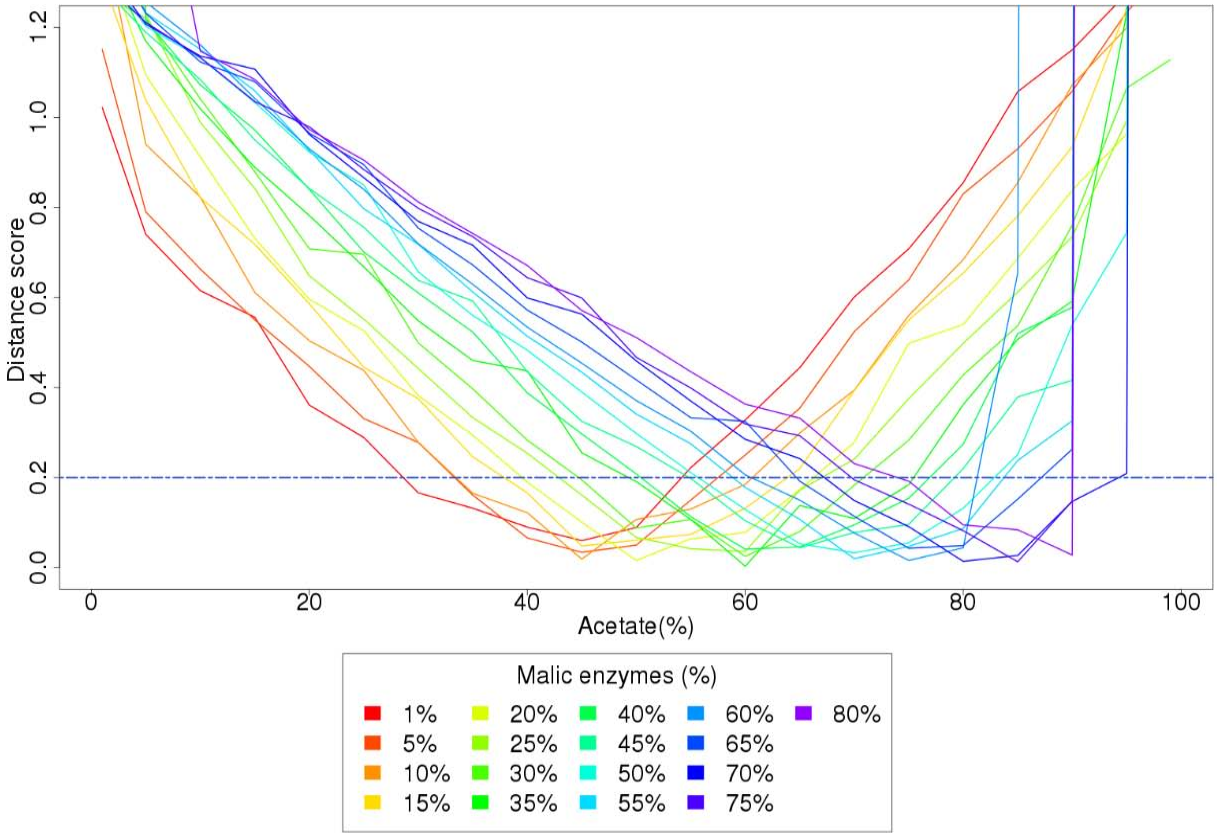


Figure 3.14 – Profil des scores de distance selon le flux accordé aux ME.
Le graphique représente la distance de score selon la proportion acétate/succinate attendue. Ce modèle comporte 2 contraintes supplémentaires, par rapport au modèle de la Figure 3.12, utilisées pour contraindre le flux de la PEPCK (étape 19) et des ME (étapes 25-26). Le flux de ces deux enzymes est en effet contraint à des valeurs comprises entre 1 et 80% par pas de 5. La contrainte sur la proportion d'acétate/succinate impose un intervalle de valeurs compris entre 1 à 99% par pas de 5%.

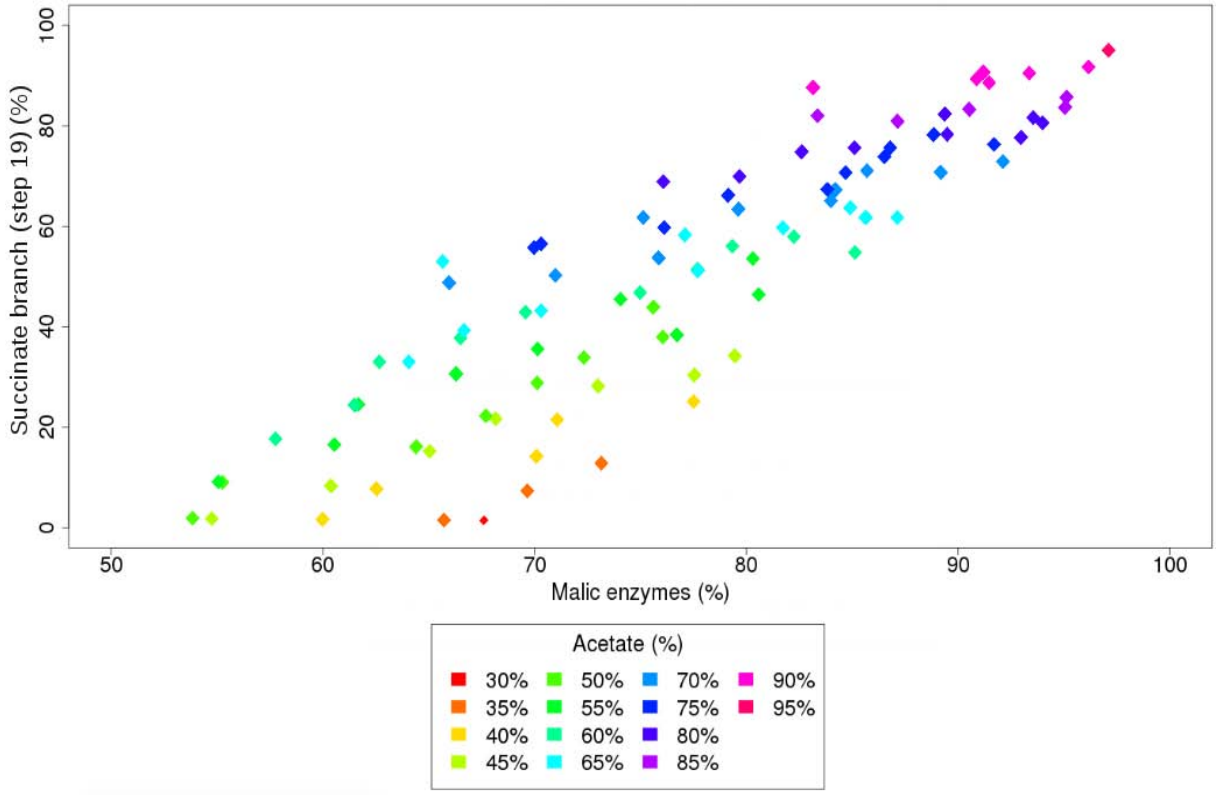


FIGURE 3.15 – Graphique représentant le flux observé pour les branches succinate par rapport au flux des ME.
Les contraintes utilisées pour la Figure 3.12 sont les mêmes que celles qui sont appliquées ici. Nous représentons le flux (%) entre la branche du succinate (PEPCK - étape 19) et des ME (étapes 25-26). Seuls les flux correspondant à des scores de distance inférieurs à 0.2 sont considérés. On peut observer ici que le flux à travers la branche succinate augmente de manière linéaire avec le flux des ME.

3.4.3 Analyse comparative FBA - Metaboflux

Dans le but de mesurer l'apport de l'approche proposée par Metaboflux, nous avons réalisé une analyse comparative avec FBA (implémenté dans le logiciel FBA-SimVis [Grafahrend-Belau *et al.* 2009]). Pour cette analyse, nous reprenons le problème posé par le modèle PF que nous soumettons à deux types de contraintes :

(i) l'équilibre des rapports ATP/ADP et NAD^+/NADH dans les glycosomes,

(ii) une distribution des produits excrétés correspondant aux données expérimentales de [Van Weelden *et al.* 2003], dont la proportion des métabolites finaux correspond pour une quantité de glucose donnée à 50% d'acétate, 36% de succinate dans le glycosome et 14% succinate dans la mitochondrie.

FBA-SimVis ne permet pas de spécifier plusieurs contraintes. Pour pouvoir palier à ce problème, nous avons spécifié la contrainte de proportion des métabolites finaux en ajoutant au modèle une réaction supplémentaire, qui consomme les métabolites excrétés à une stœchiométrie correspondant aux proportions attendues de ces métabolites (36, 14 et 50). Aucun moyen n'a en revanche été identifié pour indiquer dans le modèle la contrainte d'équilibre des rapports ATP/ADP et NAD^+/NADH. Pour Metaboflux, nous avons exploité les données expérimentales en deux temps. Nous avons considéré dans un premier modèle uniquement la

contrainte de proportion des métabolites excrétés, puis nous avons exploité la fonction multi-objectif de Metaboflux pour définir la contrainte supplémentaire sur les rapports ATP/ADP et NAD^+/NADH.
Deux tendances peuvent être identifiées dans les résultats de ces simulations (Fig. 3.16). Lorsqu'on ne considère qu'une seule des deux contraintes, FBA-SimVis et Metaboflux (Fig. 3.16a,b) permettent tous deux de satisfaire la contrainte de production de succinate et d'acétate, mais ceci se fait dans les deux cas au détriment des rapports ATP/ADP et NAD^+/NADH. La distribution du flux calculée par FBA-SimVis (Fig. 3.16a) entraine une surproduction de 25.1% d'ADP et de 14.3% NADH. De même, la distribution calculée par Metaboflux (Fig. 3.16b) engendre une surproduction d'ADP de 27.4%, qui impacte faiblement le rapport NAD^+/NADH (+1% de NADH). Dans un second temps, nous avons ajouté au modèle de Metaboflux (Fig. 3.16c) la contrainte d'équilibre des rapports ATP/ADP et NAD^+/NADH. Il apparaît que le modèle PF peut satisfaire les contraintes de proportions des métabolites finaux et d'équilibre des rapports. En effet, on peut observer que la solution calculée par Metaboflux correspond cette fois à un compromis, qui entraîne une légère diminution de la production d'acétate de 1.6% avec un faible impact sur les rapports, avec une surproduction d'ADP et de NADH ne dépassant pas les 3% par rapport aux proportions attendues. Ces résultats confirment ainsi l'apport favorable de la prédiction de la distribution des flux à l'aide de contraintes supplémentaires pour le modèle du trypanosome.

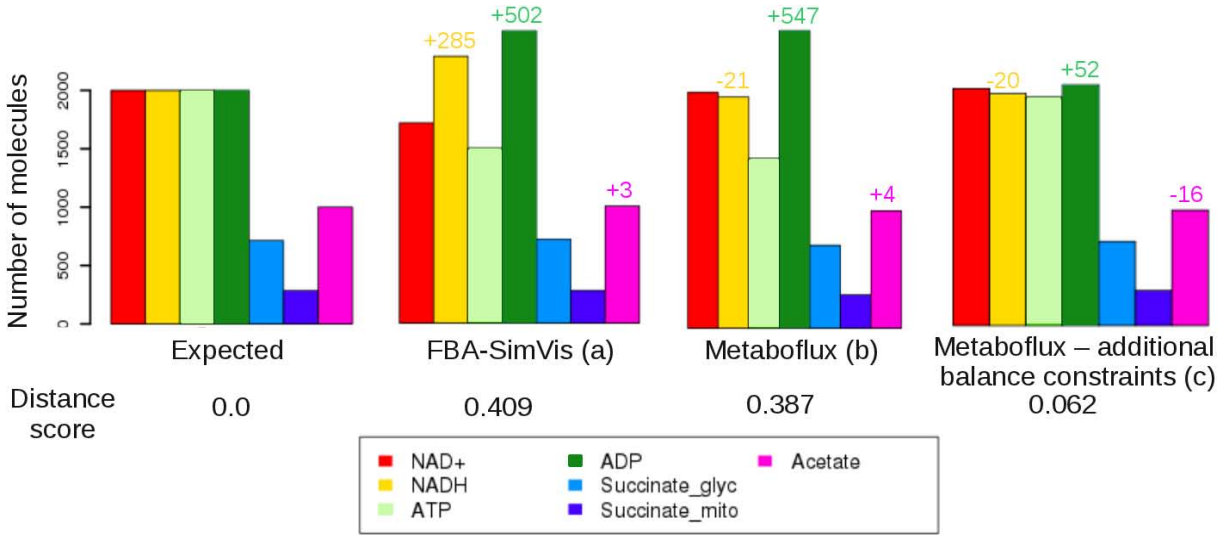
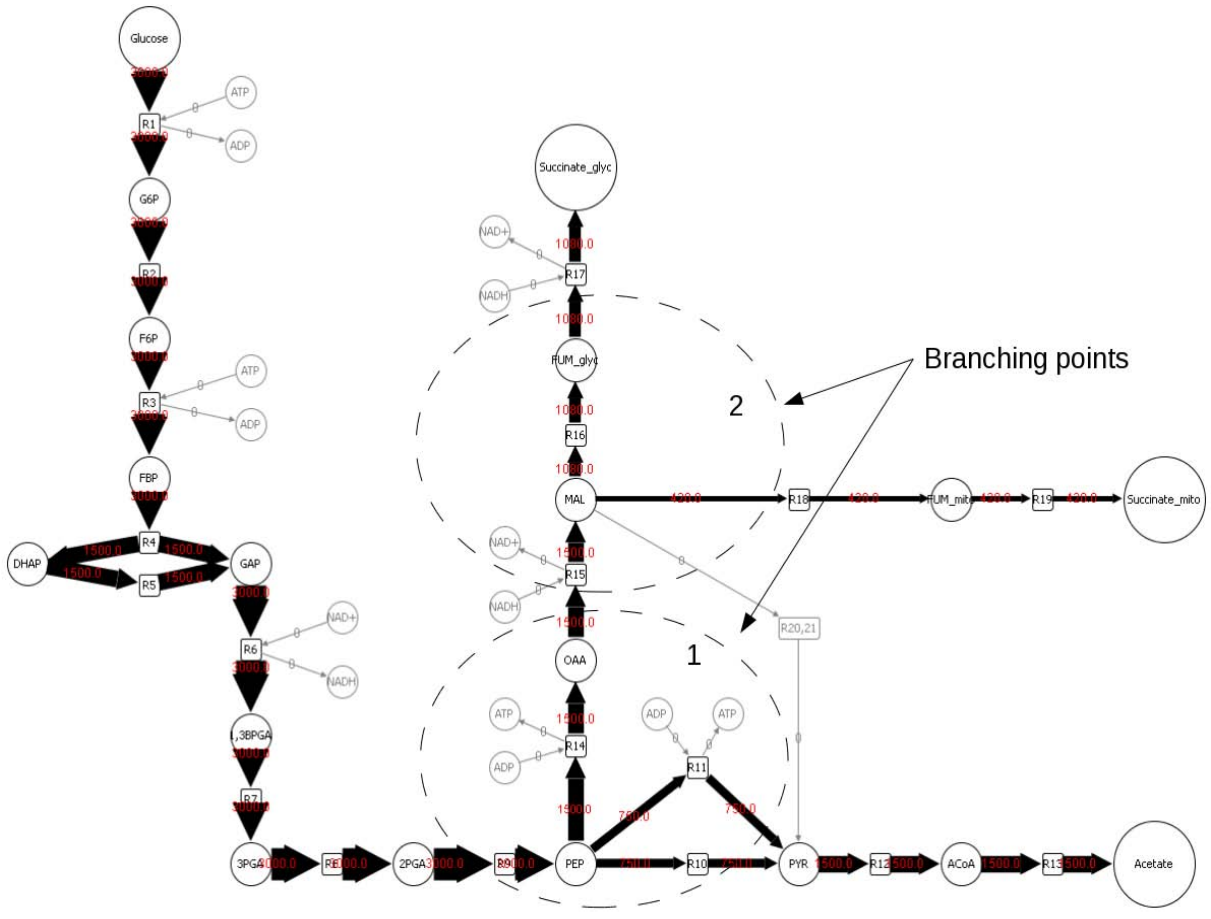
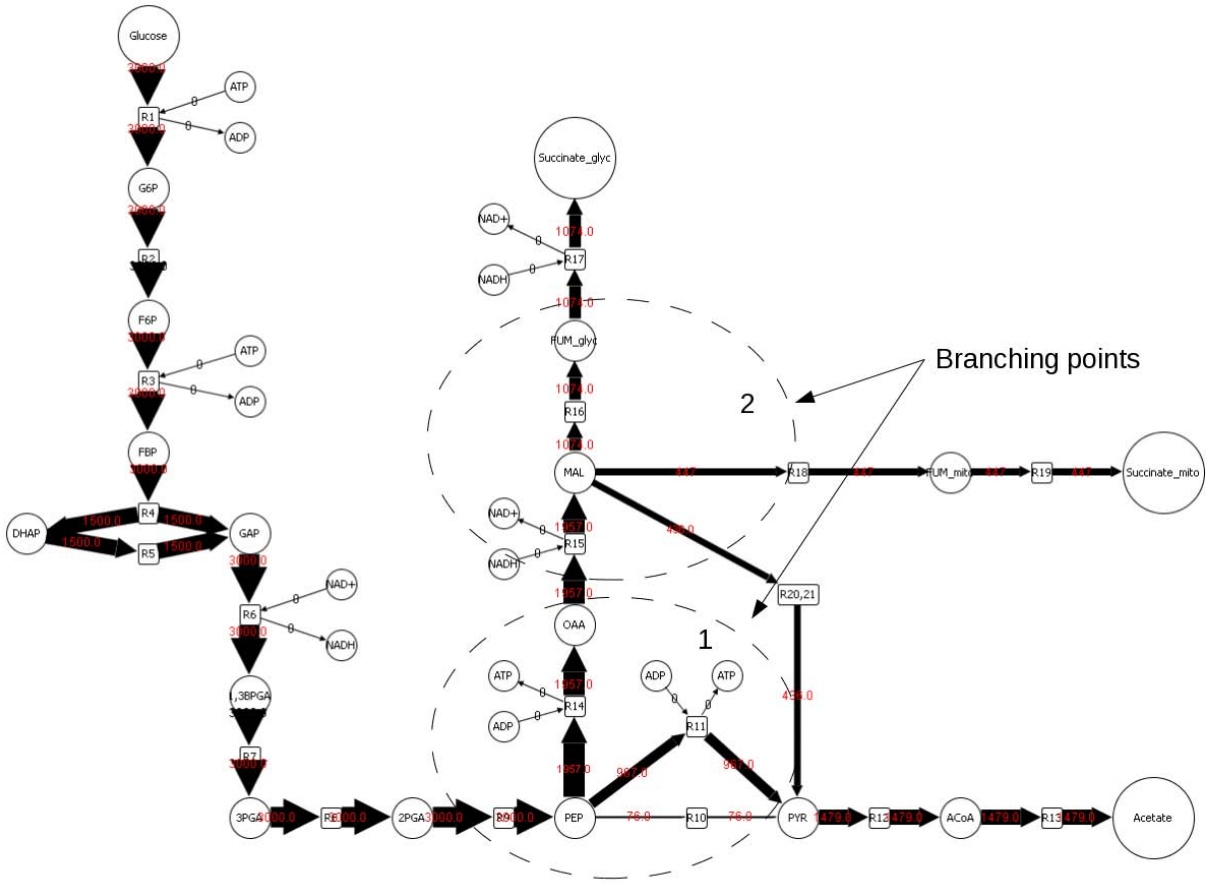


Figure 3.16 – Histogramme représentant le marquage final du FPN obtenu par FBA-SimVis et Metaboflux.
Nous considèrerons ici le marquage final observé en moyenne pour 100 simulations du modèle PF pour les distributions de flux calculées par FBA-SimVis et Metaboflux. 1000 molécules (correspondant à des jetons au sein du FPN) de glucose et 2000 molécules d'ATP, d'ADP, de NAD^+ et de NADH sont disposés en entrée du modèle. Le modèle de FBA-SimVis (a) et de Metaboflux (b) sont soumis à une contrainte sur la proportion des métabolites excrétés (succinate et acétate) correspondant respectivement à 36% pour succinate_glyc, 14% pour le succinate_mito et 50% l'acétate. Metaboflux (c) intègre une contrainte supplémentaire pour l'équilibre des rapports ATP/ADP et NAD^+/NADH. Le marquage final attendu correspond à 2000 molécules d'ATP, d'ADP, de NAD^+ et de NADH, 714 molécules de succinate_glyc, 286 molécules de succinate_mito et 1000 molécules d'acétate.

Pour aller plus loin dans cette étude, nous avons analysé visuellement la distribution du flux obtenue par FBA-SimVis et Metaboflux (Fig. 3.16a,c). Nous exploitons pour cela le logiciel de visualisation Vanted [Junker *et al.* 2006], utilisé par FBA-SimVis.
La Figure 3.17 représente la distribution des flux calculée par les deux logiciels. On peut observer que des différences entre ces deux distributions se présentent aux carrefours (*branching points* 1 et 2). Pour FBA-SimVis, la distribution du flux au carrefour 1 est proportionnellement égale entre la pyruvate kinase et la pyruvate phosphate dikinase (étape 10, 11 de ce modèle correspondant aux étapes 10, 16 Fig. 3.3) et le chemin métabolique impliquant les enzymes maliques (étapes 20, 21 de ce modèle correspondant aux étapes 25, 26 Fig. 3.3) n'est pas sollicité pour ajuster la production de succinate et d'acétate. Aussi, FBA-SimVis se focalise sur la distribution favorisant le chemin le plus court. Pour Metaboflux, la distribution du flux calculée donne la priorité à la pyruvate phosphate dikinase au niveau du carrefour 1, pour participer au rétablissement de la contrainte ATP/ADP dans le glycosome. De plus, 22% du flux total sont distribués aux enzymes maliques au carrefour 2 pour respecter la proportion attendue de succinate/acétate, ce qui n'était comme FBA-SimVis pas observé pour le modèle Metaboflux (Fig. 3.16b) (données non montrées). Cette analyse montre donc que l'intégration de plusieurs contraintes a un impact positif et concret sur la distribution des flux au sein du réseau métabolique de la PF, en permettant l'intégration de plus de données expérimentales. Ainsi, malgré la complexité accrue du modèle généré, le modèle proposé correspond mieux à la réalité biologique.



(a) Distribution des flux prédite par FBASimVis (a)



(b) Distribution des flux prédite par Metaboflux (c)

FIGURE 3.17 – Distributions des flux obtenues par FBASimVis et Metaboflux.
La proportion du flux calculée par Metaboflux et FBA-SimVis est reportée sous la forme de cartes de flux par le logiciel Vanted, où l'épaisseur des arcs varie positivement pour des flux plus importants. Les cartes de flux (a) et (b) correspondent aux distributions des flux calculées respectivement par FBA-SimVis et Metaboflux.

3.5 Discussion et conclusion

Le travail qui a été réalisé dans cette thèse pour l'étude du métabolisme, a consisté à définir une démarche permettant de mieux prendre en compte les connaissances biologiques disponibles pour l'analyse du métabolisme. Bien que la vérification du fonctionnement et l'intégration des données d'un réseau métabolique aient été abordées par les méthodes FBA, la définition d'une fonction multi-objectif n'avait jusque-là été abordée que de façon théorique, sans proposer d'outils pour le FBA. Nous avons donc développé et implémenté une nouvelle approche au sein du framework Metaboflux pour répondre à cette question. Cette méthode se base sur la prédiction de la distribution des flux au sein d'un réseau métabolique modélisé sous la forme d'un *Flux Petri net*, soumis à plusieurs contraintes formulées à partir des connaissances biologiques.
Pour déterminer si le formalisme adopté et cette capacité à intégrer plusieurs contraintes permettaient effectivement de reproduire un phénotype observé *in vitro*, nous avons testé notre

application sur le modèle connu du métabolisme du glucose des formes sanguines du trypanosome. Il est apparu lors de cette analyse que la distribution du flux mais aussi les proportions des métabolites prédites étaient parfaitement en accord avec les données expérimentales. Nous avons donc poursuivi notre étude sur le métabolisme énergétique des trypanosomes en prenant cette fois le modèle procyclique de ce métabolisme, dont la distribution des flux à l'intérieur du réseau connu n'avait pas été abordée. La simulation et l'analyse de ce métabolisme a mis en évidence dans un premier temps, un déséquilibre du rapport NAD^+/NADH par rapport à la formulation courante du modèle. Pour déterminer l'origine de ce déséquilibre, nous avons retiré une à une les contraintes du système et étudié la répartition des métabolites finaux. Nous avons observé que les contraintes sur le rapport NAD^+/NADH et sur les enzymes maliques étaient les contraintes les plus impliquées dans le déséquilibre du système. Les données expérimentales obtenues après cette analyse, ont confirmé que la contrainte sur les enzymes maliques était surestimée et devait être retirée du modèle.
Dans un second temps, nous nous sommes intéressés à l'analyse de la flexibilité de la distribution du flux du métabolisme énergétique du trypanosome. L'analyse de ce métabolisme a mis en évidence une flexibilité importante de ce métabolisme pour la redistribution du flux entre les branches succinate et acétate. L'analyse visuelle dans un premier temps, puis par la simulation, nous a fourni une explication rationnelle de cette flexibilité via la redistribution du flux par les étapes des ME. De manière intéressante, la littérature nous indique que la diminution de l'expression des gènes des ME par ARN interférence est létale pour les trypanosomes [Coustou *et al.* 2008]. Sachant que le principal rôle des ME est de fournir le NADPH requis pour les voies de biosynthèse et de répondre au stress oxydatif, on peut considérer que le flux au travers des ME dépend de la demande en NADPH. En condition de stress oxydatif, la redistribution du flux à travers les étapes des ME, pour augmenter la production de NADPH dans le cytosol et la mitochondrie, entraînerait une augmentation du ratio acétate/succinate.
Enfin, nous avons comparé l'apport de notre approche par rapport au FBA, en analysant le modèle PF avec Metaboflux et le logiciel FBA-SimVis qui incorporait cette méthode. Nous avons pu observer que la considération de plusieurs contraintes avait un réel impact sur la distribution des flux, en modifiant le chemin emprunté par le flux. L'emploi privilégié de la PK et surtout des ME essentielles à ce métabolisme étaient effectivement dépendante de considération combinées des différentes contraintes, que ne permettait pas FBA. Cette comparaison a aussi révélé l'apport des réseaux de Petri pour la définition de contraintes. En effet, leur définition au sein de FBA a été problématique en raison du formalisme proposé par ces modèles et il nous a été impossible de formaliser la contrainte d'équilibre des rapports ATP/ADP et NAD^+/NADH. Le choix de la modélisation par les réseaux de Petri s'est ainsi avéré intéressant durant toute l'analyse, pour exprimer simplement l'ensemble des contraintes du trypanosome.
L'application développée a ainsi permis d'étudier le métabolisme de glucose de *T. brucei*, en fournissant un modèle probant pour expliquer ce métabolisme. Le formalisme adopté s'est révélé en tout point satisfaisant pour l'étude du trypanosome. Metaboflux a donc été mis à disposition de la communauté [1] et publié dans le revue *Advances in Bioinformatics* [Ghozlane *et al.* 2012]. La visualisation des prédictions de Metaboflux dans le logiciel Systrip a constitué, quant à elle, l'objet d'un cas d'étude du logiciel Systrip, publié à la conférence *IV2012* (*16th International Conference Information Visualisation*) [Dubois *et al.* 2012]. Pour aller plus loin dans cette étude, une première perspective consisterait à vérifier expérimentalement l'implication des enzymes maliques dans la flexibilité de la distribution du flux chez

1. http ://www.cbib.u-bordeaux2.fr/metaboflux/

T. brucei. Cette analyse consisterait à étudier la relation entre le stress du trypanosome et le ratio acétate/succinate excrétés, qui confirmerait si cette relation était observée, l'implication des enzymes maliques dans la flexibilité du flux. Un second point porterait sur l'évaluation théorique des modèles. Le programme Metaboflux a été développé de manière à ce qu'il puisse fournir au-delà d'une simple identification de la répartition des flux, un nombre de scénarios limités et classés en fonction de leur qualité pour chaque modèle. Cette propriété pourrait être employée à l'avenir lorsque plus de données seront disponibles sur le métabolisme des trypanosomes.

Deuxième partie

Prédiction des cibles des sRNAs

Chapitre 4

Les ARN

Sommaire

Nous abordons dans cette seconde partie le travail qui a été réalisé sur les ARN non codants des bactéries. Dans ce premier chapitre, nous présentons les concepts biologiques sur lesquels s'appuient nos travaux sur les ARN. Nous commençons par définir les principales notions relatives aux ARN, leur synthèse et leurs fonctions. Nous nous intéressons ensuite plus particulièrement aux ARN non codants. Enfin nous présentons les méthodes de prédictions des interactions des ARN non codants. Les notions ici présentées sont issues de [Alberts *et al.* 2002] et [Mallick et Ghosh 2012].

4.1 Qu'est-ce que l'ARN ?

L'acide ribonucléique (ARN) est avec l'acide nucléique, les lipides, les glucides et les protéines, l'une des cinq macromolécules communes à tous les organismes. Cette molécule exerce un rôle essentiel dans le mécanisme d'expression de l'information génétique contenue dans l'ADN. Dans le dogme central de la biologie moléculaire [Crick 1970] (Fig. 4.1), nous pouvons identifier deux types de rôle pour l'ARN selon qu'il soit codant ou non codant.
L'ARN codant joue le rôle de structure transitoire de l'information par l'ARN messager (mRNA). L'information génétique de l'ADN est transmise à l'ARN par la transcription puis en protéines par la traduction. Les protéines forment dans ce modèle les constituants de base nécessaires au fonctionnement de la cellule. Ce modèle prévoit également un rôle non codant de l'ARN, qui est assuré par l'ARN ribosomique (rRNA) et l'ARN de transfert (tRNA). Ces ARN ne sont pas traduits en protéines, mais jouent un rôle fonctionnel dans le support de la traduction. Ce dogme a depuis été précisé [Shapiro 2009], en mettant en avant la possibilité de

l'ARN de se répliquer, de se rétro-transcrire en ADN, mais aussi d'exercer un rôle fonctionnel bien plus étendu, de nombreuses autres catégories d'ARN non codants ayant été découvertes. Avant de préciser le rôle de l'ARN non codant chez les bactéries (voir Section 4.2), nous présentons brièvement dans cette partie les caractéristiques et les processus biochimiques de cette molécule.

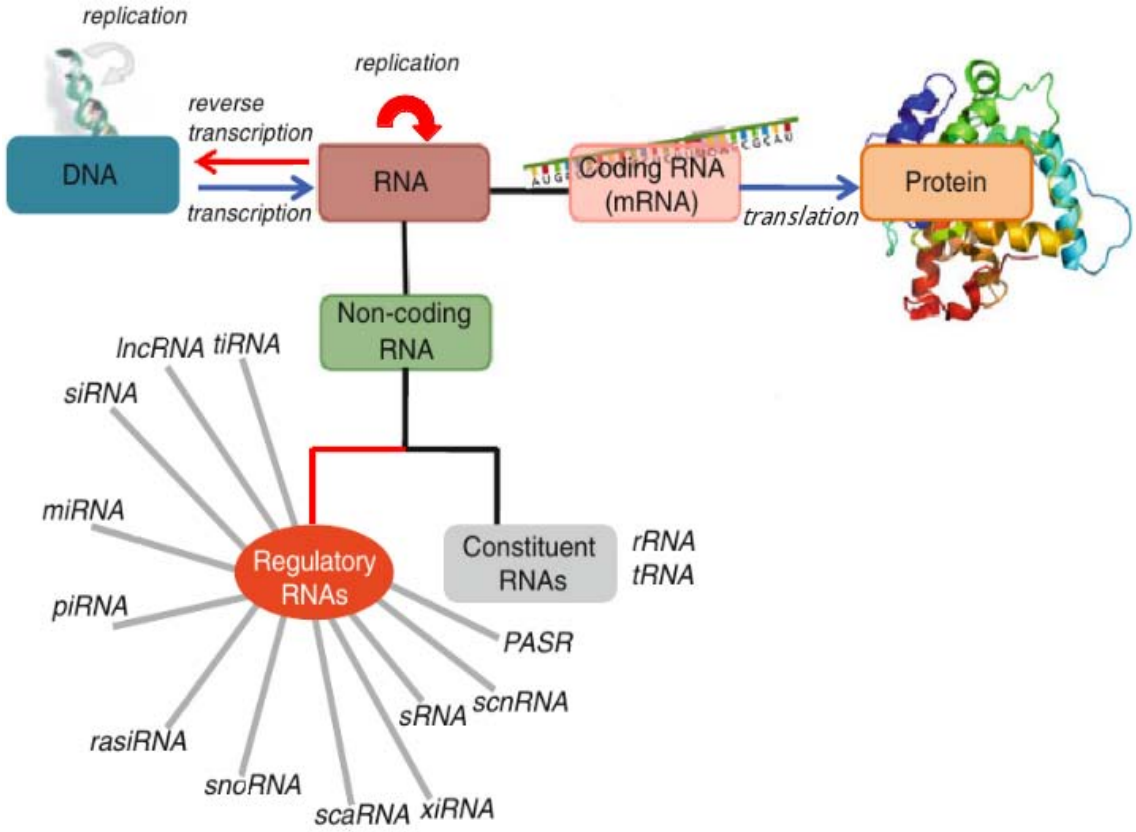


FIGURE 4.1 – Schéma du dogme central de [Crick 1970]. Les arcs en rouge représentent les éléments non-prévus par le modèle initial. Image issue de [Mallick et Ghosh 2012].

4.1.1 Structure de l'ARN

Principe

L'ARN est une molécule constituée d'un enchainement de quelques dizaines à quelques milliers de nucléotides (FIG. 4.2). Chaque nucléotide est constitué de deux parties : un sucre (le ribose) attaché à un groupement phosphate et une base azotée qui peut être de l'adénine (notée A), de la guanine (notée G), de la cytosine (notée C) ou de l'uracile (noté U) remplaçant la thymine (T) de l'ADN. Chaque sucre est lié au sucre suivant par l'intermédiaire d'une liaison phosphodiester, créant ainsi un polymère composé de manière répétitive de sucre lié à des bases.

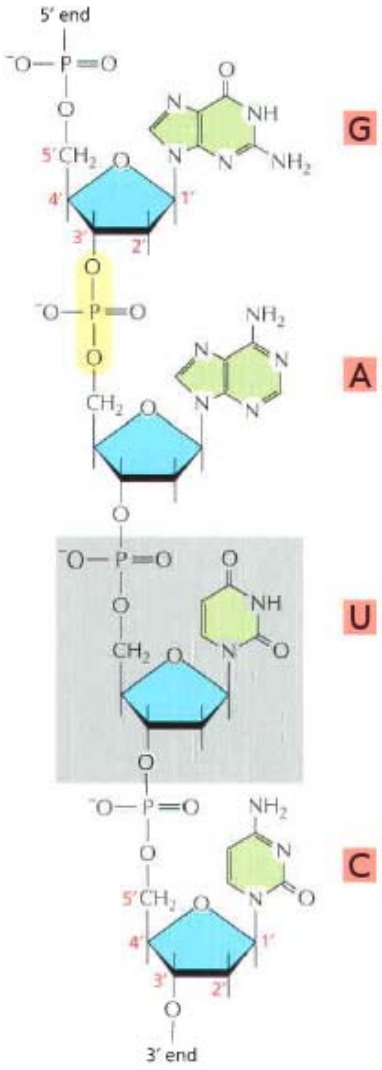
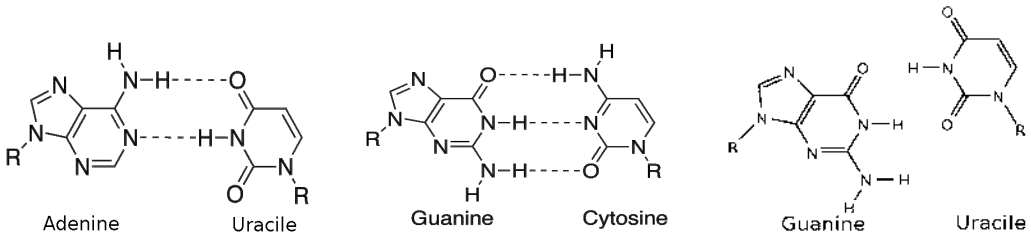


FIGURE 4.2 – Schéma de l'ARN [Alberts *et al.* 2002].
Quatre nucléotides peuvent être observé sur le schéma : G,A,U,C. Les portions en bleu, vert et jaune représentent respectivement le ribose, les bases azotées et une liaison phosphodiester.

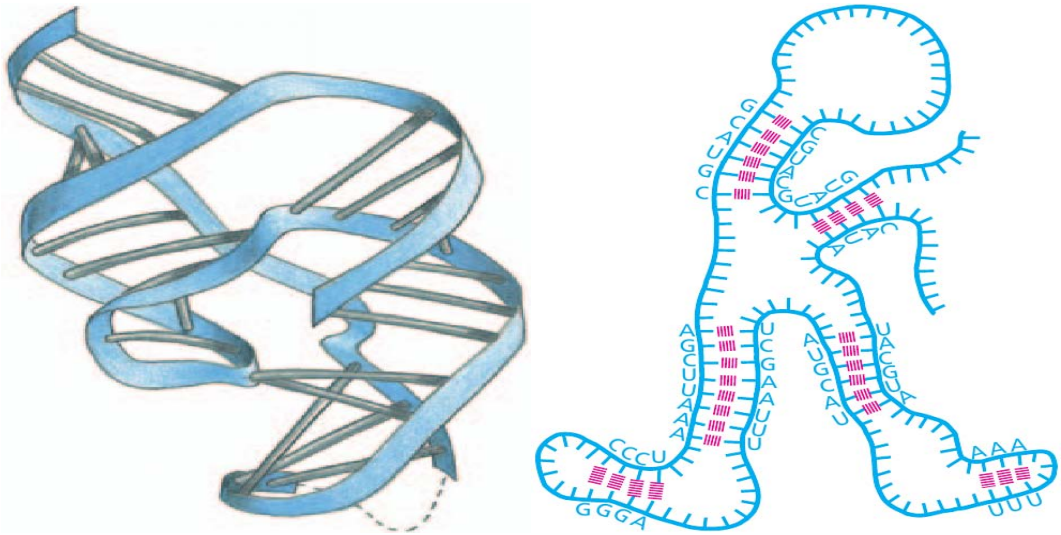
L'ARN est généralement formé d'un simple brin. Dans cet état, la molécule d'ARN est très malléable et tend à se replier sur elle-même en formant des liaisons hydrogènes entre ses bases. Deux types d'appariements peuvent être identifiés : les appariements canoniques et non canoniques. Les appariements canoniques correspondent aux interactions stables, elles mettent en jeu des interactions de type Watson-Crick entre les bases. Les bases A=U et G≡C forment entre elles respectivement deux et trois liaisons hydrogènes (FIG. 4.3a,b). Un troisième type d'arrangement survient aussi fréquemment chez les ARN pour la formation d'un appariement non-canonique de type G=U, appelé appariement de type Wobble (FIG. 4.3c). La stabilité de ses différents arrangements varie, ainsi le complexe G≡C est plus stable que le complexe A=U, lui-même plus stable que le complexe G=U.



(a) Liaison A=U (b) Liaison G≡C (c) Liaison G=U

FIGURE 4.3 – Liaisons hydrogènes observées entre les nucléotides. Seuls les appariements les plus fréquents sont ici représentés : les paires Watson-Crick (a) et (b), l'appariement non canonique G=U (c). L'ensemble des possibilités est référencé par [Westhof et Fritsch 2000].

L'ensemble de ces liaisons donne le repliement en trois dimensions correspondant à la structure tertiaire de la molécule (FIG. 4.4a), dont une simplification est représentée par la structure secondaire de l'ARN (FIG. 4.4b).



(a) Structure tertiaire (b) Structure secondaire

FIGURE 4.4 – Illustrations représentant (a) la structure tertiaire et (b) secondaire de l'ARN [Alberts *et al.* 2002].

Motifs

Différents types de motifs peuvent être identifiés dans la structure repliée d'un ARN (FIG. 4.5). Nous définissons ici quelques notions qui seront considérées lors de la prédiction des interactions des ARN (voir Section 4.2.4).

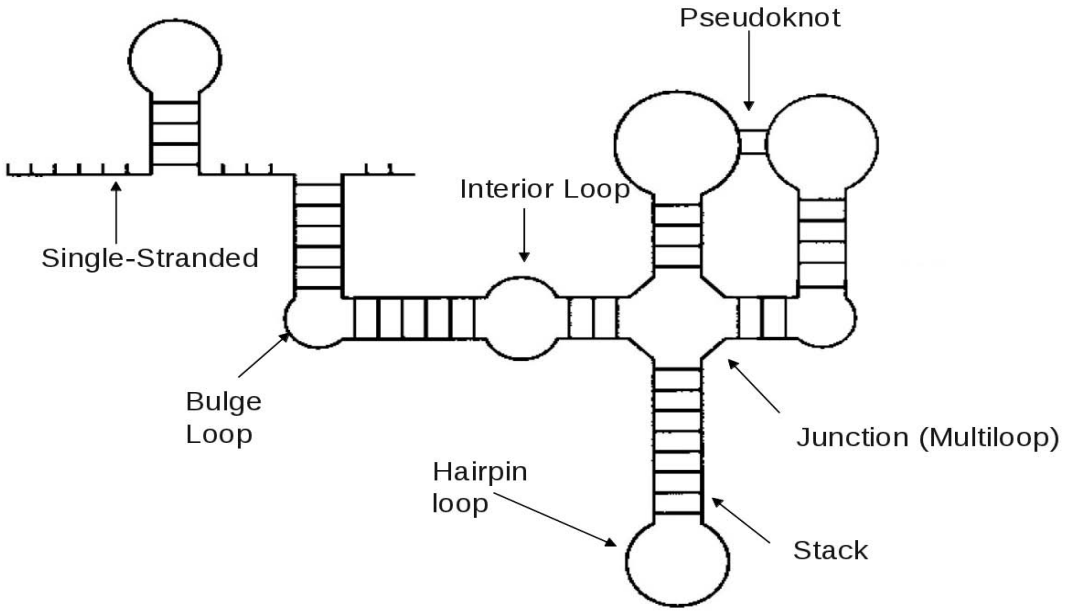


FIGURE 4.5 – Les différents motifs pouvant apparaître dans la structure secondaire d'un ARN [Wuchty *et al.* 1999].

Un empilement (*Stack*) Un empilement désigne un enchainement continu d'appariements. Cette structure constitue un élément favorable d'un point de vue de l'énergie et de la stabilité de la structure de l'ARN.

Un renflement (*bulge*) Un renflement survient lorsqu'une ou plusieurs bases ne peuvent former d'appariements d'un côté de la structure.

Une boucle interne (*interior loop*) Une boucle interne survient lorsqu'une ou plusieurs bases ne peuvent former d'appariement des deux côtés de la structure.

Une boucle terminale (aussi appelée tige-boucle ou *hairpin loop*) Une boucle terminale désigne une région d'un brin d'ARN complémentaire avec lui-même, qui se termine par une boucle d'au moins 4 nucléotides non-appariés.

Une boucle multiple (*junction*) Une boucle multiple (aussi appelée jonction) est une région où deux (ou plus) composantes en double brin se rejoignent pour former une structure fermée.

Un pseudonœud (*pseudoknot*) Un pseudonœud est une structure contenant au moins deux boucles terminales, dans laquelle un certain nombre de nucléotides de la première boucle effectue des appariements avec les nœuds de la seconde boucle [Staple et Butcher 2005].

4.1.2 Transcription

La transcription est le processus par lequel une molécule d'ARN est synthétisée à partir d'ADN (FIG. 4.6). Ce processus se base sur la complémentarité des nucléotides au brin tran-

scrit de l'ADN. Chez les procaryotes, la transcription a lieu dans le cytoplasme et se déroule en trois phases : l'initiation, l'élongation et la terminaison (Fig. 4.6).

Initiation L'initiation s'effectue au niveau du promoteur qui est en amont de la séquence codante (Figure 4.6 - étapes 1-3). La sous-unité σ de l'ARN polymérase bactérienne reconnait deux séquences consensus localisées en amont du site d'initiation de la transcription correspondant à "TTGACA" en -35 et à la boite *PRIBNOW* : "TATAAT" en -10. La liaison de ses deux composantes permet d'ouvrir la double hélice.

Élongation La phase d'élongation correspond à l'incorporation des nucléotides sur le brin transcrit (Figure 4.6 - étapes 4-5). Le brin d'ARN formé correspond à une copie du gène codant (de 5' $\rightarrow$ 3').

Terminaison La terminaison de la transcription survient lorsque l'ARN polymérase arrive au niveau du terminateur (Figure 4.6 - étapes 6-7). Chez les procaryotes, le terminateur (dit Rho indépendant) correspond à une séquence d'ADN riche en G et en C qui ralentit la progression de l'ARN polymérase, et par la formation d'une boucle terminale entre 2 régions complémentaires de l'ARN suivie d'une série de A sur le brin transcrit qui bloquent l'enzyme.

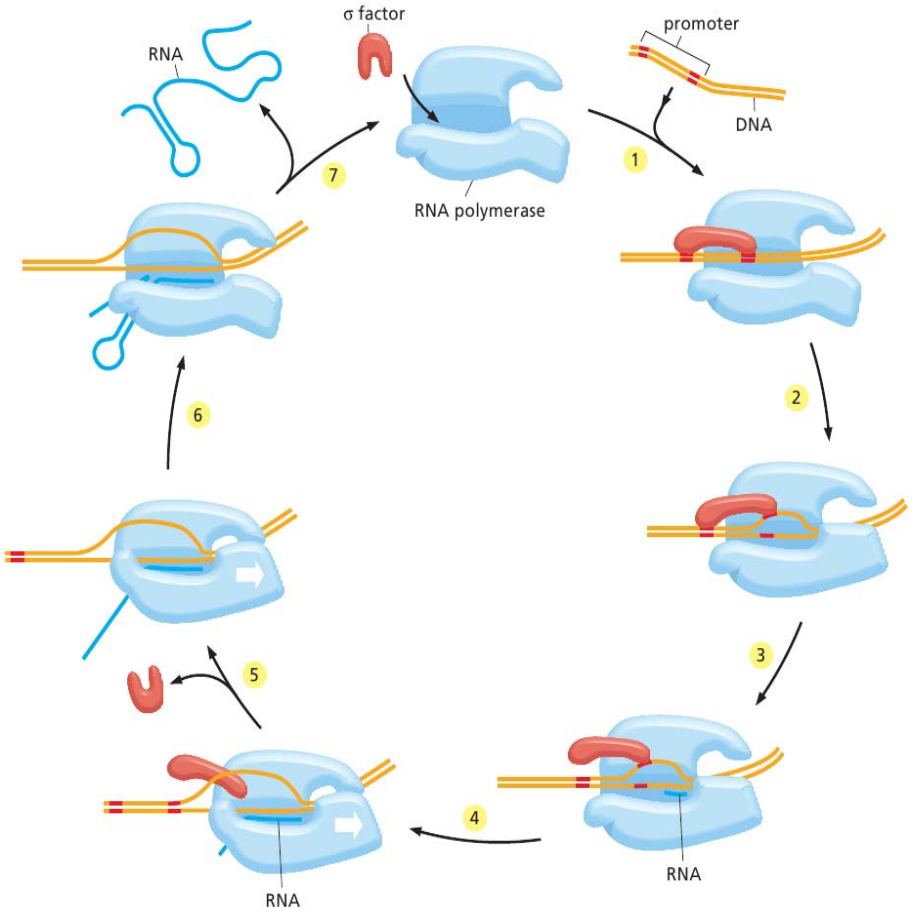


Figure 4.6 – Transcription de l'ARN chez les procaryotes [Alberts *et al.* 2002].
La transcription s'effectue en trois parties correspondant l'initiation (étapes 1 - 3), l'élongation (étapes 4-5) et enfin la terminaison (étapes 6-7).

4.1.3 Traduction

La traduction est le processus biologique au cours duquel un peptide (polymère d'acides aminés) est synthétisé à partir d'un ARN messager. Ce peptide forme après son repliement et parfois sa combinaison avec d'autres peptides, une protéine. La traduction repose sur la correspondance entre un acide aminé et un triplet de nucléotides, qui est donnée par le code génétique universel (Fig. 4.7).

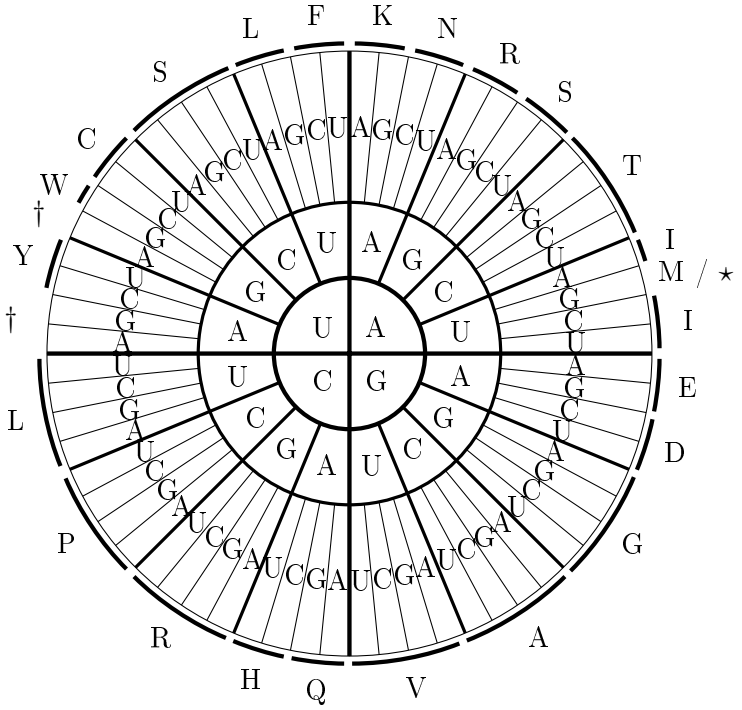


Figure 4.7 – Le code génétique universel.

Chez les procaryotes, la traduction se déroule dans le cytoplasme et débute dans la majorité des cas par la reconnaissance du codon de démarrage (AUG) (correspondant à une méthionine) et de la séquence de Shine-Dalgarno (aussi appelé *Ribosome Binding Site - RBS*) "UAAGGAGGU" [Shine et Dalgarno 1974] par la sous-unité 16S de l'ARN ribosomique. Cette séquence, longue de 4-5 nucléotides et située entre 5-8 nucléotides en amont du codon démarrage, joue un rôle essentiel pour l'ancrage du ribosome [Yusupova *et al.* 2006].
Le ribosome se déplace ensuite de codon en codon (dans le sens 5'→3') au niveau du mRNA (Fig. 4.8) et ajoute, à la lecture de chaque triplet, un acide aminé à la protéine en cours de construction à l'aide d'un ARN de transfert. La traduction s'arrête lorsque le ribosome parvient au niveau d'un codon stop. Il en existe trois dans le code génétique : UGA, UAG ou UAA, qui ne correspondent à aucun acide aminé. Les deux sous-unités du ribosome se détachent alors et le mRNA ainsi que le peptide synthétisé sont relâchés.

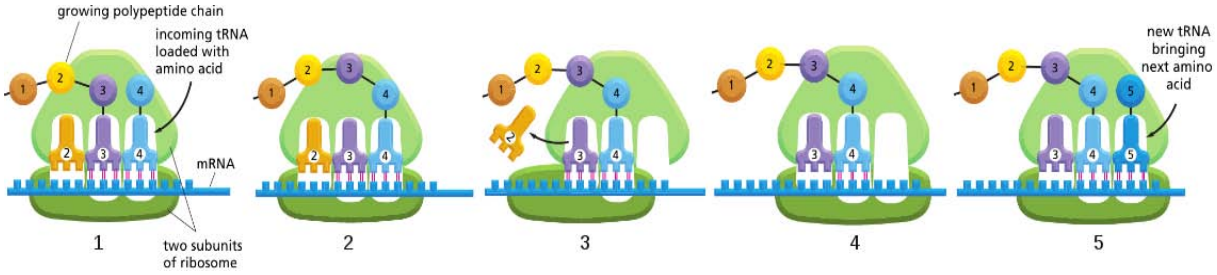


FIGURE 4.8 – Schéma du processus de traduction [Alberts *et al.* 2002].

Trois cadres de lectures potentiels sont possibles selon la position à laquelle débute la lecture de codons, auxquels s'ajoutent trois autres cadres lorsque le sens de lecture de la molécule est également indéterminé. Un seul cadre de lecture parmi les six permet d'obtenir la séquence codante de la protéine.

4.2 Les ARN non codants

Comme nous l'avons vu précédemment, l'ARN non codant désigne l'ensemble des ARN fonctionnels qui ne sont pas traduits en protéine. Depuis leur première découverte en 1965 grâce à leur haut niveau d'expression [Storz 2002], les ARN non codants ont constitué l'objet de découvertes majeures pour la compréhension de certains processus en jeux dans la cellule (FIG. 4.9). Il est estimé qu'ils pourraient jouer un rôle important dans la plupart des caractéristiques complexes des organismes et la variation génétique entre les espèces [Mattick et Makunin 2006]. Un enjeu majeur de ses dernières années a donc été de rechercher et de caractériser les processus dans lesquels ils sont impliqués. Dans cette thèse, nous nous sommes intéressés aux rôles des ARN non codants des bactéries. Nous présentons dans cette section les données initiales du contexte, en considérant les différentes familles d'ARN non codants. Nous verrons ensuite les caractéristiques et les mécanismes d'actions des ARN non codants des bactéries. Enfin, nous considèrerons les différentes méthodes développées pour identifier ces ARN et surtout déterminer leurs cibles.

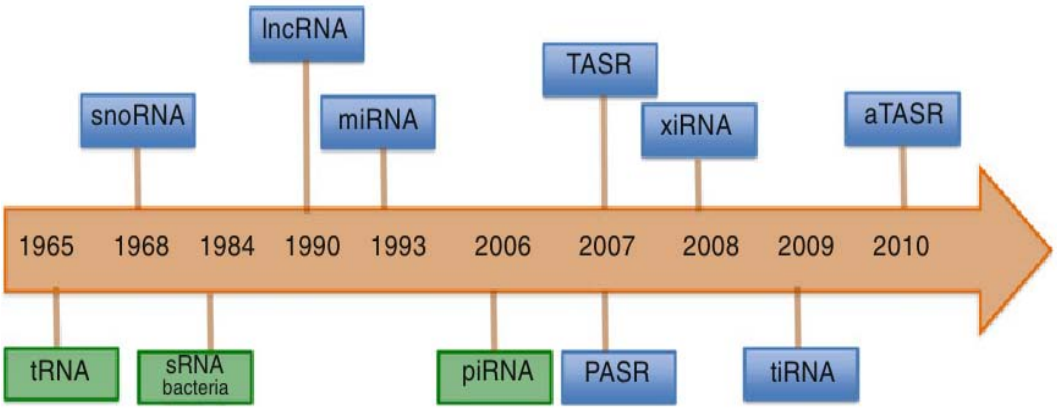


FIGURE 4.9 – Frise chronologique présentant les découvertes majeures liées aux ARN non codants [Mallick et Ghosh 2012].

4.2.1 Les différentes familles d'ARN non codants

Les ARN non codants sont classiquement divisés en deux catégories désignant les ARN dits constituants (ou transcriptionnels) regroupant les tRNA et rRNA, et les ARN régulateurs regroupant de nombreuses familles d'ARN (siRNA, miARN, snRNA...) (Fig. 4.10). Chacune de ses familles possède une structure et/ou une fonction particulière.

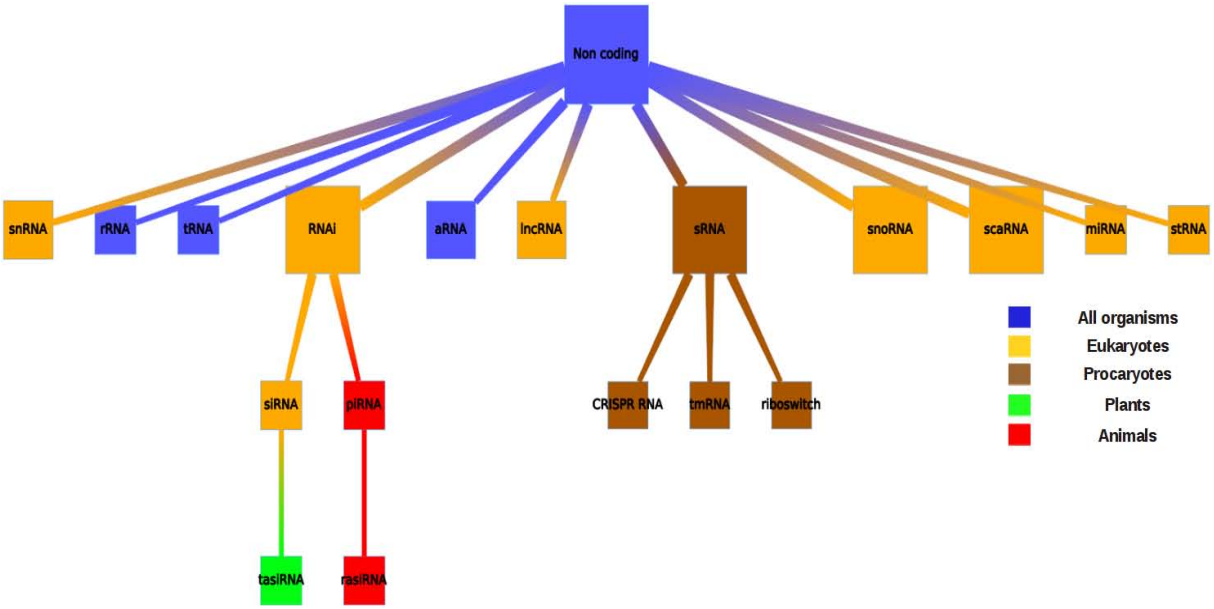


Figure 4.10 – Classification non-exhaustive des différentes catégories d'ARN non codants. *Les carrés bleus correspondent aux catégories d'ARN présentes d'un point de vue théorique chez tous les organismes vivants. Les carrés marrons et jaunes correspondent aux catégories uniquement présentes respectivement chez les procaryotes et les eucaryotes. Les carrés rouges et verts distinguent des catégories d'ARN uniquement identifiés chez les animaux ou les plantes.*

Chez les eucaryotes (et particulièrement les eucaryotes supérieurs), le système d'expression complexe des gènes implique une large variété de régulations utilisant les ARN non codants, tels que la régulation transcriptionnelle et l'épissage. Ces systèmes emploient souvent les ARN non codants pour cibler leur activité. Par exemple, le système d'inhibition transcriptionnel RISC (*RNA-induced silencing complex*) des eucaryotes met en jeu les ARN d'interférence (RNAi) pour reconnaitre et détruire les ARN messagers ciblés à l'aide de différentes familles d'ARN, tels que les micro-ARN (miRNA) [Bartel 2009] et les *small interfering RNA* (siRNA) [Elbashir *et al.* 2001]. L'épissage implique, quant à lui, les *small nuclear RNA* (snRNA) [Choudhuri 2010] regroupant notamment les *small nucleolar RNA* (snoRNA) et *Small Cajal body-specific RNAs* (scaRNA) [Jády et Kiss 2001] qui s'associent au complexe snRNP du spliceosome pour reconnaître et exciser de manière spécifique les introns des ARN pré-messagers.
Chez les bactéries, les ARN non codants (aussi appelés *small noncoding RNAs* - sRNAs) représentent la classe la plus importante de régulateurs post-transcriptionnels [Mallick et Ghosh 2012]. Depuis 2001, les sRNAs ont été recherchés et étudiés dans les plasmides, les phages et dans le chromosome bactérien. Ces études ont identifié de nombreux sRNAs (Fig. 4.11). Il est ainsi estimé qu'un génome bactérien encoderait pour environ 200 à 300 sRNAs avec diverses fonctions [Hershberg *et al.* 2003]. Ces ARN, longs de 50 à 500 nucléotides [Gottesman

et Storz 2011], agissent principalement pour réguler la réponse aux changements de conditions environnementales [Mallick et Ghosh 2012] en régulant les voies métaboliques ou les voies de réponses au stress. Différentes familles d'ARN sont regroupées dans cette catégorie, on identifie ainsi selon leur mécanisme d'action : les sRNAs agissant sur des protéines, les sRNAs agissant par appariement avec d'autres ARN, les riborégulateurs (aussi appelés *Riboswitches*) [Roth et Breaker 2009], et la nouvelle catégorie des CRISPRs [Sorek *et al.* 2008].

Organismes	sRNAs identifiés	N° mRNA	Publications
Streptococcus pneumoniae	88	2235	[Acebo *et al.* 2012; Kumar *et al.* 2010]
Escherichia coli	261	4201	[Raghavan *et al.* 2011; Shinhara *et al.* 2011]
Listeria monocytogenes	180	3000	[Izar *et al.* 2011]
Yersinia pseudotuberculosis	150	4447	[Koo *et al.* 2011]
Bacillus subtilis	261	4201	[Irnov *et al.* 2010]
Helicobacter pilori	60	1634	[Sharma *et al.* 2010]
Chlamydia trachomatis	9	889	[Albrecht *et al.* 2010]
Staphylococcus aureus	30	2625	[Bohn *et al.* 2010]
Listeria monocytogenes	50	3000	[Toledo-Arana *et al.* 2009]
Bacillus anthracis	2	5425	[Passalacqua *et al.* 2009]
Burkholderia cenocepacia	13	6512-6930	[Yoder-Himes *et al.* 2009]
Vibrio cholerae	147	3784	[Liu *et al.* 2009]
Salmonella enterica	40	4744	[Sittka *et al.* 2009]
Mycoplasma pneumoniae	117	689	[Guell *et al.* 2009]
Mycobacterium tuberculosis	9	4079	[Arnvig et Young 2009]
Mycobacterium leprae	68	141	[Akama *et al.* 2009]

FIGURE 4.11 – Liste référençant les publications qui identifient de nouveaux sRNAs chez des bactéries depuis 2009. Le nombre de sRNAs indiqués pour ces études correspond au nombre de sRNAs identifiés comme potentiels par *RNAseq* ou *tiling arrays* et/ou validés par *northern blot* (voir Section 4.2.3). Le nombre de mRNAs notifié [1]constitue un nombre de cibles potentielles pour ces sRNAs dans le cas d'interaction par appariement.

4.2.2 Mécanisme d'action des sRNAs régulateurs

La régulation par les sRNAs est avantageuse d'un point de vue métabolique, car les sRNAs sont courts et n'ont pas besoin d'étape supplémentaire de traduction. Leur effet est rapide et pourrait être aussi plus fiable et plus durable que celui des facteurs de transcription [Waters et Storz 2009]. Les sRNAs régulateurs présentent également une diversité importante d'activités et de mécanismes d'action que pointent de nombreuses revues [Storz *et al.* 2011; Romby et Charpentier 2010; Waters et Storz 2009]. Au cours de cette thèse, nous nous sommes particulièrement intéressés aux sRNAs agissant par appariement avec un mRNA, qui représentent la principale catégorie de sRNAs connus [Waters et Storz 2009]. Deux types d'actions peuvent être identifiées selon qu'ils soient en *cis* (localisation sur le brin opposé du gène qu'ils régulent) ou en *trans* (localisation sur un locus différent du gène qu'ils régulent) (FIG. 4.12).

1. https ://www.ebi.ac.uk/genomes/bacteria.html

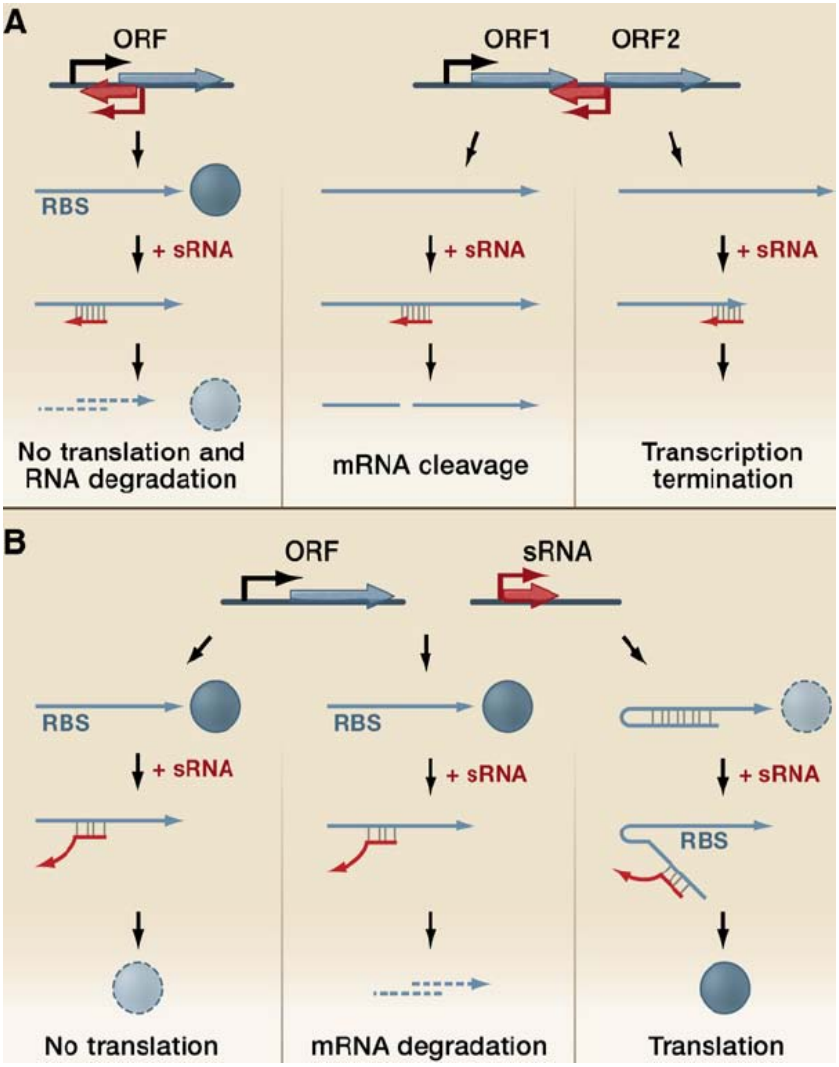


FIGURE 4.12 – Mécanisme d'action des sRNAs agissant par appariement [Waters et Storz 2009]. Deux types d'actions sont identifiées selon que le sRNA soit en *cis* (A) ou en *trans* (B) de sa cible.

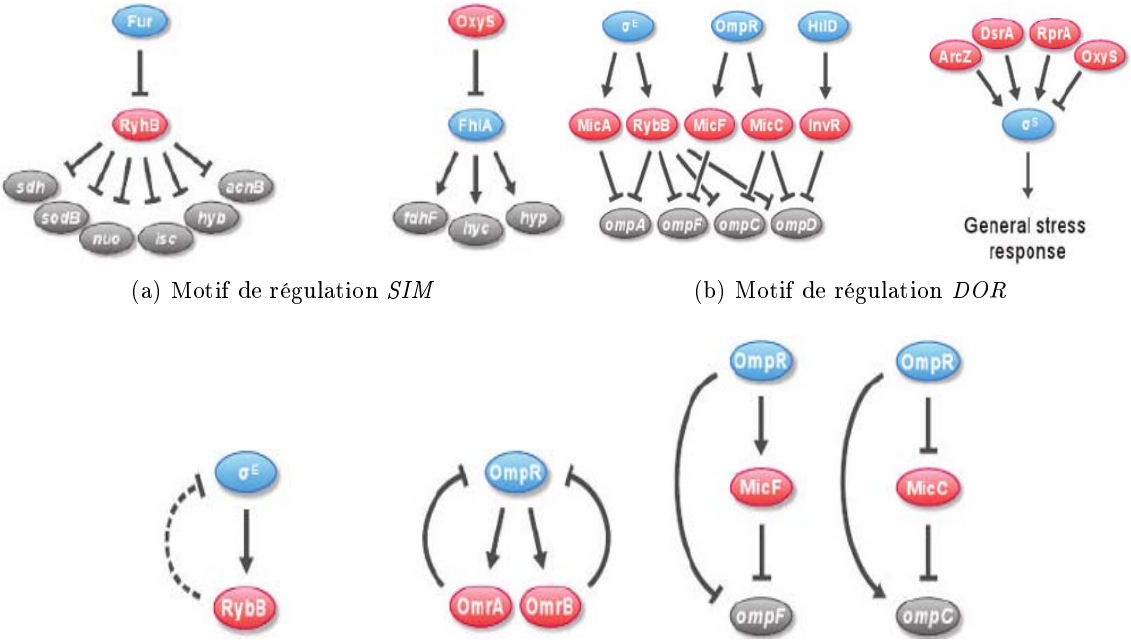
Les sRNAs *cis*-régulateurs

La majorité des sRNAs *cis*-régulateurs découverts sont exprimés de manière constitutive depuis des éléments mobiles : bactériophages, plasmides et transposons, où ils exercent principalement un rôle de contrôle de la réplication des gènes présents sur l'élément mobile [Brantl 2007]. Le mécanisme d'action de ces éléments repose sur la complémentarité étendue qu'ils partagent avec leur cible, souvent de 75 nucléotides ou plus [Brantl 2007; Wagner *et al.* 2002]. Cette complémentarité importante permet une régulation très spécifique de leur cible, leur permettant d'exercer différentes actions (FIG. 4.12A), telles que le blocage de la traduction de leur cible en masquant par leur appariement le RBS [Wagner *et al.* 2002]. Certains sRNAs vont ainsi permettre aux bactéries de contrôler la traduction d'ARN messagers toxiques à haut niveau [Gerdes et Wagner 2007; Fozo *et al.* 2008]. Ce cas est notamment observé chez *E.coli* pour différents systèmes antitoxine/toxine, tels que SymR/*SymE* et Hok/*Sok*, qui permettent à la cellule de ralentir sa croissance ou de se mettre en stase lorsque les conditions environ-

nementales sont stressantes d'un point de vue métabolique [Kawano *et al.* 2007; Unoson et Wagner 2008]. Les sRNAs *cis*-régulateurs peuvent aussi promouvoir ou inhiber la dégradation de leur cible en affectant la stabilité du mRNA [Brantl 2007]. Ce processus a par exemple été observé pour la régulation de l'expression des gènes dans un opéron, tel que l'appariement du sRNAs GadY au mRNA *gadXW* chez *E. coli* qui entraine le clivage du duplex entre *gadX* et *gadY* et une accumulation de *gadX* qui code pour une protéine intervenant dans la réponse au stress acide [Opdyke *et al.* 2011, 2004]. Des phénomènes de terminaison de transcription, spécifiques aux sRNAs *cis*-régulateurs, sont aussi observés. Ce cas est par exemple pour l'ARNβ et du mRNA *fatDCBangRT* [Stork *et al.* 2007].

Les sRNAs *trans*-régulateurs

Contrairement aux sRNAs *cis*-régulateurs, les sRNAs *trans*-régulateurs ont une complémentarité limitée et imparfaite avec leur(s) cible(s). Celle-ci est souvent comprise entre 10 à 25 paires de nucléotides [Storz *et al.* 2011; Waters et Storz 2009], dont seulement un nombre restreint est critique. Ce cas est par exemple observé pour l'appariement du sRNA Sgrs avec ptsG, qui ne présente que 6 paires indispensables sur 23 paires [Kawamoto *et al.* 2006]. Cette capacité limitée d'appariements permet aux sRNAs *trans*-régulateurs de réguler plusieurs gènes. Différents motifs de régulation peuvent alors apparaître [Beisel et Storz 2010] (Fig. 4.13). Les motifs *SIM* (*Single-Input Module*) et *DOR* (*Dense Overlapping Regulon*) correspondent ainsi à des motifs de régulations multiples qui surviennent respectivement lorsqu'un ou plusieurs sRNAs répondant à un ou à plusieurs signaux de stress croisés, régulent un ensemble de mRNAs (Fig. 4.13a,b). Ce cas est notamment observé pour le sRNA RybB, qui n'utilise que 5 à 7 nucléotides pour reconnaître plusieurs cibles impliqués dans la même voie métabolique [Papenfort *et al.* 2010; Balbontín *et al.* 2010]. Ceci est aussi observé pour OmrA/B [Guillier et Gottesman 2008] ou Spot42 [Møller *et al.* 2002]. D'autres motifs de régulation plus direct peuvent survenir, tels que les motifs *NF*(*Negative Feedback loop*) et *FF* (*FeedForward loop*) (Fig. 4.13c,d). Ces motifs correspondent à des régulations de l'activité et de l'expression des sRNAs. Ces cas sont par exemple observés pour les sRNAs OmrA et OmrB qui régulent l'expression du mRNA OmpR, qui peut également réguler leurs activités [Guillier et Gottesman 2006, 2008], ou encore pour la protéine de OmpR qui peut réguler les sRNAs MicF et MicC mais également les cibles de ces sRNAs : OmpF et OmpC [Shimoni *et al.* 2007]. Les sRNAs *trans*-régulateurs sont enfin très souvent synthétisés lors de conditions spécifiques de croissance. Ils exercent alors une variété importante de fonctions biologiques. Par exemple chez *E. coli*, le transport et le stockage du fer est lié au sRNA RyhB qui est exprimé lorsque la concentration en fer diminue [Massé *et al.* 2007] ; le stress lié au glucose-phosphate est régulé par l'expression du sRNA SgrS qui active SgrR [Rice et Vanderpool 2011] ; le changement de la concentration interne du glucose est contrôlé par les sRNAs Spot42 et CyaR qui respectivement, répriment et activent la traduction du mRNA CRP (pour une revue complète voir [Storz *et al.* 2011; Waters et Storz 2009]). La majorité de ces régulations sont négatives soit par l'inhibition de la traduction, soit par promotion de la dégradation du mRNA cible [Aiba 2007; Gottesman 2005] (Fig. 4.12B). Le blocage du RBS est le mécanisme le plus répandu, comme observé pour OxyS, Spot42, MicA, MicC, RyhB, RybB, OmrA/B et SgrS chez *E. coli*.



(a) Motif de régulation *SIM*

(b) Motif de régulation *DOR*

(c) Motif de régulation *NF*

(d) Motif de régulation *FF*

FIGURE 4.13 – Les différents motifs de régulation pouvant survenir lors des interactions des sRNAs [Beisel et Storz 2010].
Quatre motifs de régulation peuvent être identifiés pour les sRNAs (représentés par des cercles bleus). Les motifs SIM et DOR correspondent à des motifs de régulations multiples de l'expression d'un ou plusieurs mRNAs (représentés par des cercles rouges). Les motifs NF et FF correspondent quant à eux à des motifs de régulation de l'expression des sRNAs et de leur activité où une protéine est capable de réguler l'expression du sRNA et de sa cible ou qu'un mRNA régulé par un sRNA soit capable sous sa forme protéique de réguler également le sRNA.

Rôle de la protéine Hfq

Plusieurs régulations impliquant les sRNAs *trans*-régulateurs nécessitent la protéine Hfq [Brennan et Link 2007] (FIG. 4.14).

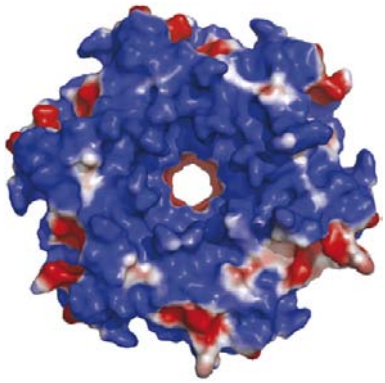


FIGURE 4.14 – Modélisation de la protéine Hfq de *Escherichia coli* [Brennan et Link 2007].

Cette protéine ARN-chaperon permet de faciliter l'interaction ARN-ARN. Son importance

a été mise en évidence chez un mutant d'*E. coli* ne produisant plus de protéine Hfq [Tsui *et al.* 1994]. Ce mutant présentait en effet un ratio de croissance ralenti, une longueur anormalement plus importante et une sensibilité accrue aux ultraviolets, mutagènes et oxydants. De nos jours, près de 22 sRNAs interagissant avec la protéine Hfq ont été identifiés chez *E. coli* [Majdalani *et al.* 2005]. La concentration de cette protéine semble plus importante chez les bactéries présentant une taille de génome et un taux de GC plus importants [Jousselin *et al.* 2009] (Fig. 4.15). Cette protéine n'a cependant pas été mise en évidence pour l'instant chez plusieurs espèces, telles que *Streptoccus pyogenes* et les *Prochlorococcus*. Des protéines aux fonctions similaires seraient possibles et restent à identifier chez ces espèces, telles que les protéines SMc01113 et YbeY identifiées chez *Sinorhizobium meliloti* affectant de manière similaire à la protéine Hfq la régulation des sRNAs [Pandey *et al.* 2011].

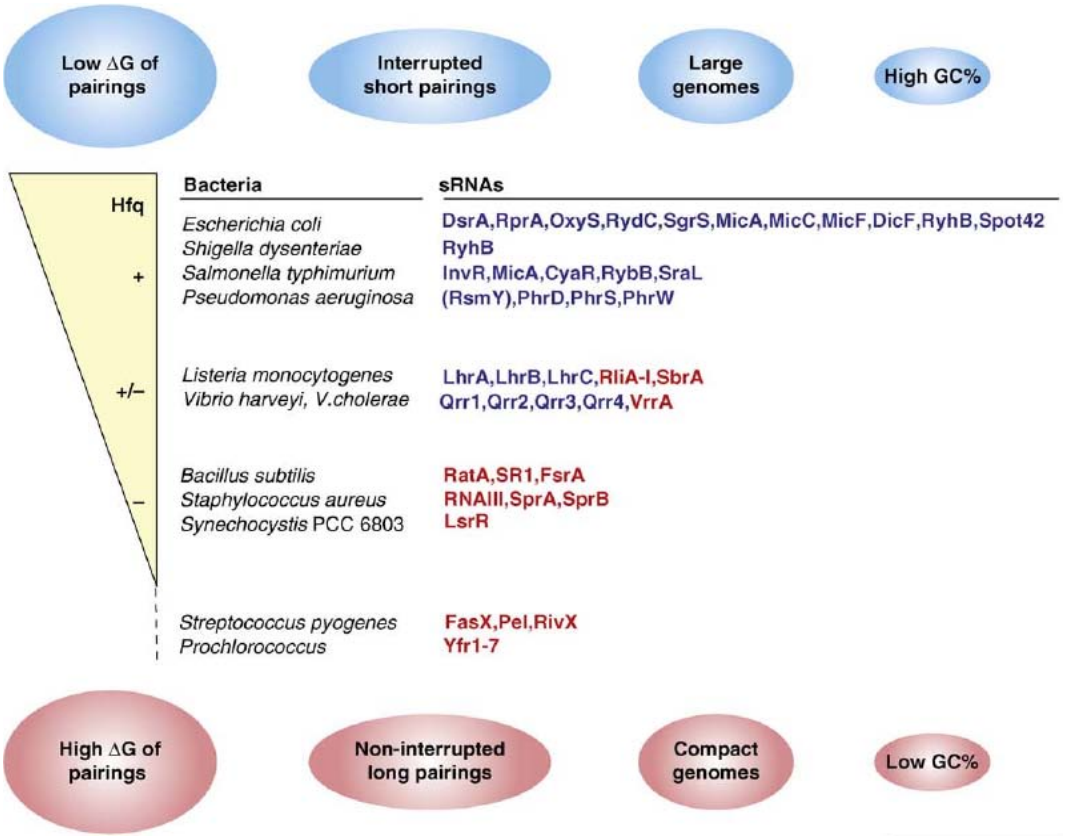
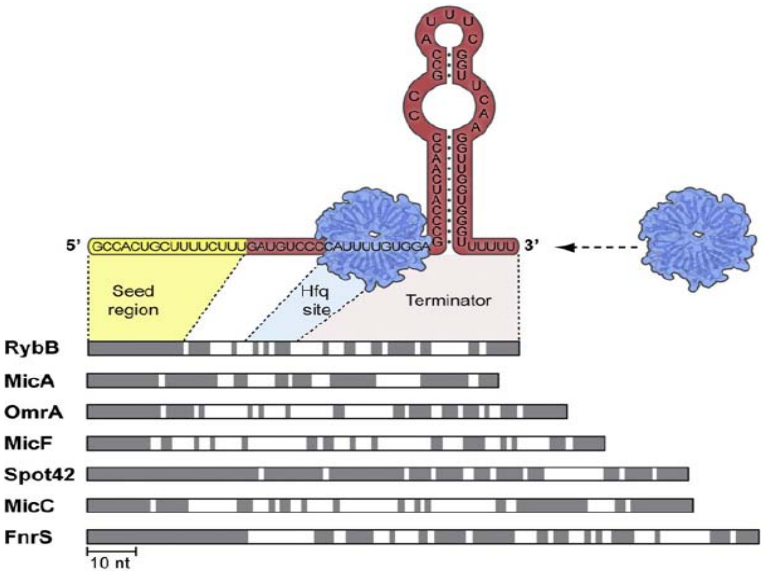


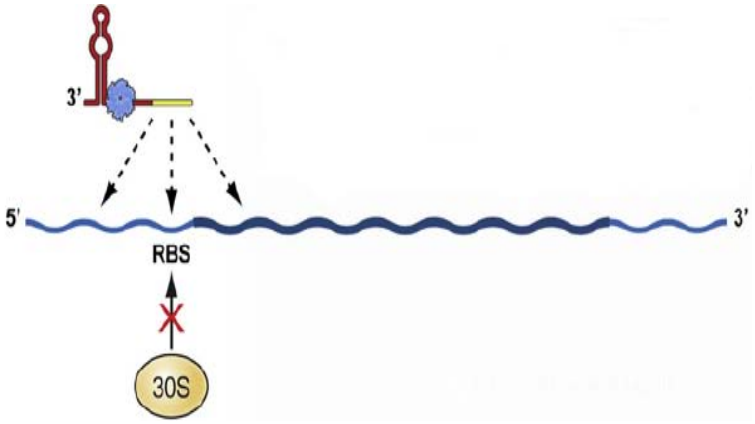
Figure 4.15 – Importance de la présence de Hfq pour les régulations des sRNAs selon l'organisme [Jousselin *et al.* 2009]. Les conditions favorables et défavorables à la régulation par Hfq sont représentées respectivement par des cercles bleus et rouges. Les cercles de la taille la plus importante indiquent les conditions les plus importantes.

Le processus d'action de la protéine Hfq repose principalement sur son interaction moléculaire avec les sRNAs, qui s'effectue au niveau des régions riches en A/U localisés près d'une structure en tige-boucle [Storz *et al.* 2011] (Fig. 4.16a). Le complexe ainsi formé sert à protéger le sRNA de la dégradation en absence d'appariement avec un mRNA [Rasmussen *et al.* 2005]. Il favorise la formation du complexe par la stabilisation du duplex partiel formé avec la cible [Aiba 2007; Brennan et Link 2007] (Fig. 4.16b). Enfin, il servirait également à recruter la machinerie de dégradation ribonucléase E dans certaines régulations pour cliver le mRNA

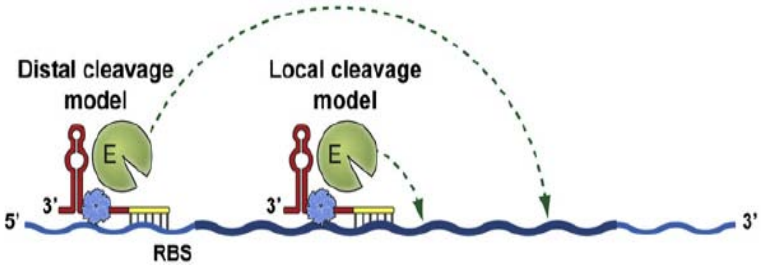
cible [Waters et Storz 2009; Brennan et Link 2007] (Fig. 4.16c).



(a) Interaction de la protéine Hfq et d'un sRNA



(b) Interaction du sRNA avec le RBS de son mRNA cible



(c) Recrutement de la ribonucléase E à l'aide de la protéine Hfq

Figure 4.16 – Mécanisme d'action de la protéine Hfq [Storz *et al.* 2011].
L'action de la protéine Hfq s'effectue en 3 étapes : (a) la protéine Hfq interagit avec une région riche en A/U près d'une structure en tige-boucle ; (b) l'interaction du sRNA avec la protéine Hfq stabilise sa conformation et favorise son interaction avec le mRNA cible ; (c) le complexe formé par la protéine Hfq permettrait également de recruter la ribonucléase E pour cliver les mRNA cibles.

4.2.3 Détection des sRNAs

La recherche des sRNAs dans un génome constitue une thématique relativement jeune dans le domaine de la génomique fonctionnelle, qui est liée aux récentes découvertes de leur intérêt fonctionnel, mais également au développement des nouvelles technologies de séquençage. La majorité des sRNAs identifiés sont ainsi le résultat de criblages génétiques réalisés par expérimentations mais aussi par analyses bioinformatiques. Nous décrivons ici ces deux approches avant d'aborder Section 4.2.4, la prédiction des cibles des sRNAs sur laquelle porte cette thèse.

Approches expérimentales

Historiquement, les premiers sRNAs découverts ont été obtenus grâce au fort signal de ces ARN sur des gels lors d'études des ARN marqués [Vogel et Sharma 2005]. L'approche expérimentale constitue à présent, le principal moyen employé pour détecter les sRNAs à l'échelle du génome. Deux méthodes sont communément employées pour cette recherche : les *tiling arrays* et les nouvelles technologies de séquençage (aussi appelées *RNA-seq*).
Les *tiling arrays* sont des puces à ADN de haute densité classiquement utilisées pour analyser le transcriptome des cellules. Elles rassemblent sur une surface de quelques centimètres carrés des fragments nucléotidiques synthétiques correspondant à l'ensemble du génome sur le brin sense et antisense en incluant notamment les régions intergéniques. Pour chaque région, plusieurs centaines de sondes sont ainsi réparties sur toute la séquence, avec près d'une sonde toutes les 6 bases. L'hybridation des ARN ou des ADN complémentaires (les cDNA correspondent à l'ARN rétro-transcrit) (totaux ou sélectionnés par leurs tailles) marqué à ces sondes permet d'établir le profil d'expression des gènes et de détecter dans les régions intergéniques les sRNAs. Ces puces permettent également d'estimer de manière approximative les bornes 5' et 3' des sRNAs et de rechercher leur expression sous différentes conditions [Tjaden *et al.* 2002]. Les *tiling arrays* ont ainsi été utilisées pour détecter les sRNAs de plusieurs bactéries, telles que *Listeria monocytogenes* [Toledo-Arana *et al.* 2009], *Mycobacterium leprae* [Akama *et al.* 2009] et *Streptococcus pneumoniae* [Kumar *et al.* 2010]. Cette approche est à présent de plus en plus en concurrencée par les approches de type *RNAseq*. En effet, les *tiling arrays* présentent plusieurs limitations liées à l'étape d'hybridation. Les signaux détectés par ces puces peuvent être bruités en raison d'interactions imparfaites des ARN à leurs sondes complémentaires ou encore d'interactions parasites des sondes avec d'autres ARN. Les données ainsi générées doivent ainsi être normalisées par rapport à ce bruit de fond, ce qui rend parfois difficile d'identifier et de quantifier l'expression des sRNAs.
Les approches de type *RNA-seq* ont donc été développées pour répondre à ces contraintes. Celles-ci reposent sur différentes technologies de séquençage ADN, tels que le *pyrosequencing* par 454, le séquençage par SOLiD ou illumina. Sans vouloir exposer le principe de ces différentes méthodes qui ne sont pas l'objet de cette thèse, nous souhaitons indiquer au lecteur que ces méthodes ont eu un impact important sur la recherche des sRNAs en permettant de simplifier et de diminuer significativement les coûts de ce type d'analyse, mais également d'améliorer la puissance de l'analyse. Les approches *RNA-seq* permettent en effet de mesurer la variation de la quantité de transcrits et de détecter avec un seul nucléotide de précision les bornes en 5' et en 3' des sRNAs [Sharma et Vogel 2009]. Ces méthodes ont ainsi permis d'augmenter significativement le nombre de sRNAs découverts [Sittka *et al.* 2009; Liu *et al.* 2009; Yoder-Himes *et al.* 2009; Albrecht *et al.* 2010; Sharma *et al.* 2010; Raghavan *et al.* 2011; Shinhara *et al.* 2011; Nicolas *et al.* 2012].

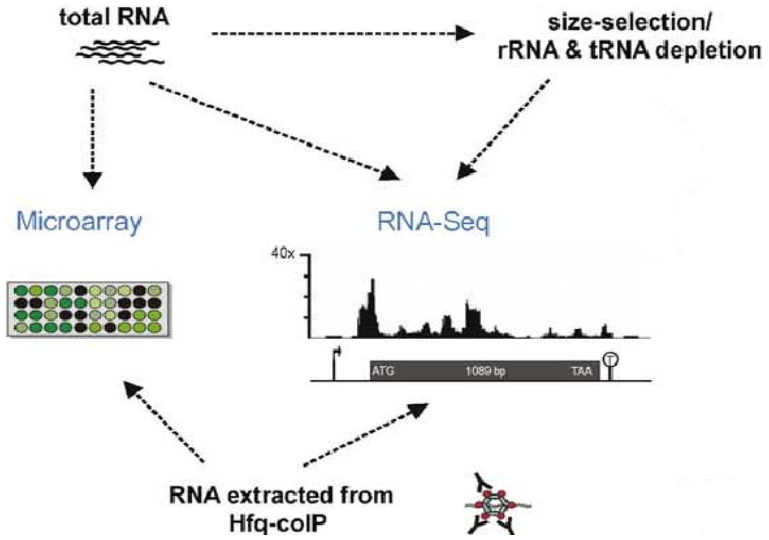


FIGURE 4.17 – La détection des sRNAs peut être réalisée en utilisant des *tiling arrays* (aussi appelées *microarrays*) ou par *RNA-seq* pour analyser les ARN totaux de la cellule ou seulement des ARN sélectionnés par leur taille [Sharma et Vogel 2009].

Approches bioinformatiques

L'approche bioinformatique est également une méthode intéressante pour identifier des sRNAs, car elle permet de rechercher ces éléments à partir des bases de données des génomes déjà disponibles (voir Section 2.2.2). Cependant, différentes stratégies ont dû être développées pour identifier spécifiquement les sRNAs. En effet, les séquences de ces éléments ne contiennent pas les motifs classiquement utilisés pour rechercher des gènes, tels qu'un RBS ou un biais de composition significatif relatif à l'utilisation de certains codons [Mallick et Ghosh 2012]. Les outils développés se sont donc basés pour les identifier sur un ou plusieurs des critères suivants : la séquence primaire, la structure secondaire ou la conservation de ces séquences entre les espèces proches, déjà observées chez *E. coli* [Zhang *et al.* 2004].
L'outil sRNAPredict3 [Livny *et al.* 2006] se base par exemple pour cette prédiction sur la présence de terminateurs Rho indépendants dans les séquences intergéniques et la conservation des séquences détectées chez des génomes proches (à l'aide de BLAST [Altschul *et al.* 1990]). Les outils, tels que eQRNA [Rivas et Eddy 2001] et RNAz [Washietl *et al.* 2005], considèrent quant à eux la conservation de la structure secondaire pour identifier les sRNAs. Ces séquences sont considérées comme conservées, tant qu'elles n'ont pas subi de modifications impactant leur structure secondaire. Le logiciel RNAz prend aussi en compte la stabilité thermodynamique des séquences qui est plus importante chez les sRNAs par rapport aux autres séquences qui n'encodent pas pour des ARN [Washietl *et al.* 2005]. Près d'une centaine de sRNAs ont ainsi pu être prédits et vérifiés expérimentalement [Backofen et Hess 2010]. Ces méthodes présentent cependant un nombre important de faux positifs [Lu *et al.* 2011], encourageant le développement de nouvelles pistes d'association avec les approches haut débit [Mallick et Ghosh 2012].

4.2.4 Détermination des cibles

La découverte de ces nouveaux sRNAs pose à présent la question de l'identification de leur rôle fonctionnel. Il est ainsi important d'identifier leurs cibles pour comprendre leur

action. Nous nous concentrerons ici sur les sRNAs agissant par appariement aux mRNAs qui représentent la principale catégorie de sRNA. À cet effet, les approches expérimentales et bioinformatiques sont à nouveaux employées en même temps pour répondre à cette question. Nous présentons ici ces deux approches, avant de voir Section 5, l'efficacité et l'intégration possible des différents outils bioinformatiques dans le pipeline d'analyse iRNA développé au cours de cette thèse.

Approches expérimentales

Les approches expérimentales pour la détection des cibles de sRNA reposent sur la manipulation génétique des bactéries, qui sont modifiées de manière à créer des mutants ne produisant plus le sRNA ciblé [Davis *et al.* 2005] ou au contraire en cultivant la bactérie dans des conditions permettant une surproduction du sRNA (par exemple en condition minimale de fer pour le sRNA RyhB). Le transcriptome de ces mutants peut être ensuite analysé comparativement à celui de la souche sauvage pour identifier les gènes cibles du sRNA qui sont sur- ou sous-exprimés selon sa présence. D'autres analyses modifient directement le sRNA en lui ajoutant une marque distinctive sous la forme d'un aptamère permettant de le filtrer parmi les ARN totaux [Said *et al.* 2009]. Ainsi après la culture de cette souche mutante, il est possible d'extraire et d'identifier les cibles attachées à ce sRNA à l'aide d'une puce à ADN. Ces méthodes encore coûteuses ne sont (comme pour les approches bioinformatiques) pas exemptes de faux positifs [Mallick et Ghosh 2012] et nécessitent de disposer de puces pour tout le génome de l'espèce étudiée. Elles offrent cependant une méthode de vérification de large échelle qui peut être utilisée avec les approches bioinformatiques.

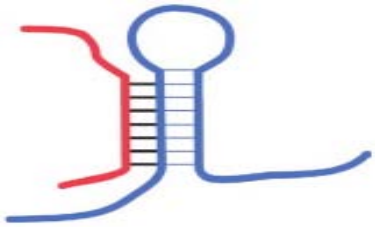
Approches bioinformatiques

Comme nous l'avons vu précédemment, les méthodes expérimentales d'identification des cibles des sRNAs sont plus difficiles à mettre à en place que celles employées pour détecter les sRNAs. Les approches bioinformatiques constituent donc ici un moyen intéressant pour orienter cette recherche. Quatre types d'approches peuvent être identifiées pour la prédiction d'une interaction entre un sRNAs et un mRNA selon qu'elles soient basées sur la séquence primaire ou la structure secondaire des ARN [Backofen et Hess 2010]. Deux types d'informations sont à considérer lors de cette analyse pour l'évaluation de la zone d'interaction et de la qualité de cet appariement.

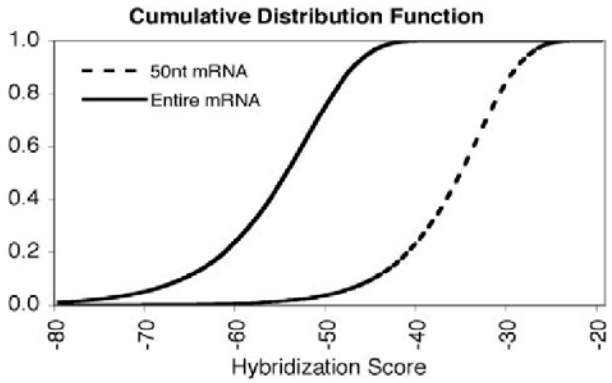
Approches basées sur la séquence - Le premier type d'approche se rapporte à la séquence primaire du sRNA et de sa cible. Elle consiste à rechercher une région complémentaire entre ces deux séquences qui correspond à la région d'appariement. Pour cela, différents outils basés sur des algorithmes d'alignements locaux équivalents de l'algorithme de Smith-Waterman peuvent être employés, tels que sRNAtarget [Cao *et al.* 2009], sTarPicker [Ying *et al.* 2011], [Zhang *et al.* 2006] ou Guugle [Gerlach et Giegerich 2006], ou en détournant des logiciels, tels que Blast [Altschul *et al.* 1990], Ssearch ou Yass [Noe et Kucherov 2005], pour l'identification de zones complémentaires entre le sRNA et le mRNA. Le résultat obtenu par ces méthodes correspond à un score de l'appariement. On obtient également pour Blast, Ssearch et Yass une e-valeur qui nous donne la confiance associée à ce score selon la probabilité de l'obtenir par chance. Ce type d'approche est confronté à la flexibilité importante des interactions sRNA-mRNA. Le logiciel spécialement dédiés à la recherche de cibles de sRNAs, sTarPicker [Ying *et al.* 2011], combine donc cette approche à une méthode d'apprentissage

Tclass [Wuju et Momiao 2002], entrainée sur un jeu de données d'interactions connues, pour mieux distinguer la qualité des appariements prédits.

Approches basées sur la thermodynamique - Les approches basées sur la thermodynamique utilisent la programmation dynamique pour résoudre la configuration de l'appariement. L'objectif de ces méthodes est d'identifier l'appariement ayant l'énergie minimum d'hybridation (*minimum free energy* - MFE) [Zuker et Stiegler 1981]. Cette énergie d'hybridation correspond à l'énergie associée des différents motifs de l'appariement, tels qu'un empilement, une boucle interne... La valeur énergétique de ces motifs est déterminée à partir d'un modèle physique défini expérimentalement par [Mathews *et al.* 1999]. Cette approche est employée par RNAhybrid [Rehmsmeier *et al.* 2004], RNAduplex, RNAplex [Tafer et Hofacker 2008] et TargetRNA [Tjaden *et al.* 2006]. Il est estimé que ces outils permettent de prédire correctement 73% des paires d'un appariement grâce au réalisme accru de leur modèle. Ce type d'approche permet également de prendre en compte la température, qui est un facteur important pour considérer la stabilité des duplexes [Backofen et Hess 2010].
Ces méthodes ont cependant pour désavantage d'ignorer les interactions intramoléculaires pouvant survenir au sein d'une même molécule, ce qui peut conduire à la prédiction de structures impossibles (Fig. 4.18a). Les interactions prédites par ces modèles peuvent être également trop longues [Backofen et Hess 2010]. Enfin le score de MFE calculé par ces méthodes présente un biais lié à la longueur des séquences [Rehmsmeier *et al.* 2004]. En effet, les séquences longues obtiennent des scores inférieurs aux séquences plus courtes (Fig. 4.18b), alors qu'il est plus probable que cela soit survenu par chance.



(a) Structure impossible qui peut être prédite.



(b) Effet de la longueur des séquences de mRNA sur le score d'hybridation [Tjaden *et al.* 2006].

Figure 4.18 – Biais observés lors de l'utilisation des approches basées sur la thermodynamique, (a) des structures impossibles où le sRNA est apparié à une région qui effectue normalement une interaction intra-moléculaire, (b) le score du MFE est dépendant de la longueur du sRNA et du mRNA comme observé ici.

Certaines de ces méthodes considèrent donc en plus du MFE, la p-valeur de l'appariement, qui correspond à la probabilité suivant une distribution des valeurs extrêmes, d'obtenir le score d'hybridation négatif normé h' selon deux paramètres u et s décrivant la distribution des résultats pour des séquences aléatoires, tel que :

$$h^{'} = -\frac{h}{\log(n * m)} \tag{4.1}$$

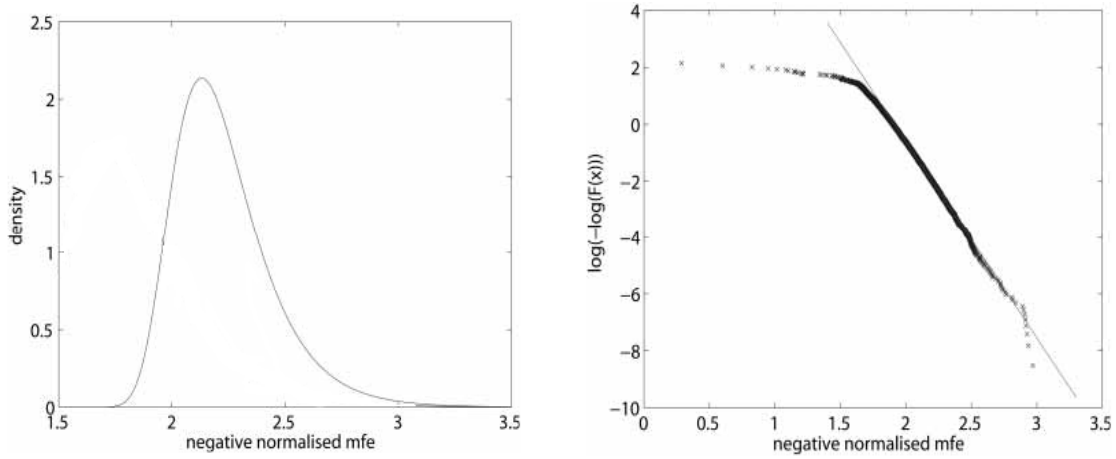
où h correspond au score d'hybridation obtenu par l'appariement d'un sRNA et d'un mRNA, n correspond à la longueur du sRNA et m correspond à la longueur du mRNA. Les paramètres u et s sont estimés quant à eux d'après la distribution cumulative des résultats $F_n(h')$ obtenus pour le sRNA donné hybridé à des séquences de mRNAs aléatoires (FIG. 4.19), tel que :

$$F_n(h') = \frac{1}{n}\sum_{i=1}^{n} card\{H'_i \leq h'\} \tag{4.2}$$

où h' correspond à un score d'hybridation négatif normé considéré, H'_i correspond au score négatif normé obtenu par l'interaction i entre le sRNA étudié et une séquence de mRNA aléatoire et n correspond au nombre total d'interactions considérées pour ces séquences aléatoires. Les paramètres u et s sont ainsi estimés d'après les paramètres de cette distribution, tel que :

$$\begin{aligned} \hat{s} &= -\frac{1}{a} \\ \hat{u} &= b * \hat{s} \end{aligned} \tag{4.3}$$

où a correspond au coefficient directeur de la droite et b à l'ordonnée à l'origine de la droite.



(a) Distribution des valeurs de MFE négatif normé d'un sRNA

(b) Distribution cumulative des valeurs de MFE négatif normé

FIGURE 4.19 – Estimation des paramètres de l'interaction d'un sRNA [Rehmsmeier *et al.* 2004].
(a) Les scores sont normés, seul les scores de MFE normés supérieur à la valeur 2.0 sans 1% des meilleurs points sont considérés. (b) La distribution cumulative des résultats obtenus est ensuite calculée et les paramètres de la droite sont estimés par régression linéaire.

La p-valeur pour le score optimal d'hybridation pour un couple sRNA-mRNA aléatoire, H, correspond ainsi :

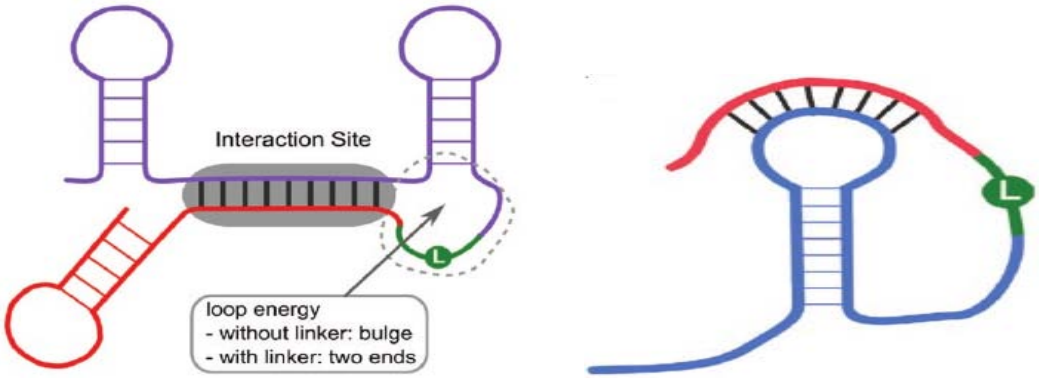
$$P(H' > h') = 1 - \exp(-\exp(\frac{h' - \hat{u}}{\hat{s}})) \tag{4.4}$$

Le logiciel TargetRNA considère ainsi (après mesure sur son jeu de test) que les appariements qui ont une p-valeur≤ 0.01 sont de vrais appariements.
D'autres logiciels, tels que sRNAtargetNB et sRNAtargetSVM [Cao *et al.* 2009; Zhao *et al.* 2008], utilisent en plus de cette prédiction, une classification des scores basée sur un classifieur bayésien ou SVM (séparateur à vaste marge) pour identifier les vrais appariements.

Approches basées sur la structure secondaire - Les approches basées sur la structure secondaire vont plus loin que les méthodes incluant la thermodynamique et incorporent directement des algorithmes de repliement de l'ARN. Ces méthodes tiennent ainsi compte de la structure interne du sRNA et de son mRNA cible. Pour cela, ces méthodes concatènent la séquence du sRNA à sa cible et replient la structure nouvellement formée par un algorithme de repliement équivalent de celui de MFOLD [Zuker 1994] ou de RNAfold [Hofacker *et al.* 1994]. Les structures prédites par ces méthodes sont ainsi imbriquées (Fig. 4.20a) et ne pourront présenter à cause de cette restriction d'autres formes (Fig. 4.20b). Il est estimé que ce type d'approche augmente à 79% le nombre de paires de bases correctement prédites. Elles permettent également de calculer la probabilité d'une structure selon la température à l'aide de la fonction de partition Z [McCaskill 1990] (Eq. 4.5).

$$Z_S = \sum_{Q \in S} e^{-\frac{E(Q)}{RT}} \tag{4.5}$$

où $E(Q)$ correspond à l'énergie d'une séquence S repliée dans la structure secondaire Q, R est la constante de Boltzmann et T est la température.
Cette approche est ainsi employée par les logiciels : RNAcofold [Bernhart *et al.* 2006] et Pairfold [Andronescu *et al.* 2005].



(a) Structure jointe des deux ARN.

(b) Structure jointe non imbriquée qui ne peut pas être prédite par cette approche.

Figure 4.20 – Deux structures jointes qui peuvent ou non être considérées par les approches basées sur la structure secondaire [Backofen et Hess 2010].
La prédiction des interactions entre deux ARN par ces méthodes considère ces ARN sous formes concaténées et séparées par un ligand ici en vert.

Approches basées sur l'accessibilité - Le dernier type d'approche se réfère à l'accessibilité des structures. Ces méthodes considèrent, en plus de l'évaluation de la thermodynamique, l'accessibilité des sites d'interaction du sRNA et de sa cible, qui doivent être disponibles pour que leurs séquences puissent s'apparier. Ces séquences ne doivent donc pas effectuer à cet endroit d'interactions intra-moléculaires. Pour cela, les méthodes développées considèrent, pour une séquence donnée (Fig. 4.21), la fonction d'énergie E définie par :

$$E = E^{hybrid} + ED^{mRNA}(i, i') + ED^{sRNA}(k, k') \tag{4.6}$$

où l'énergie d'hybridation E^{hybrid} correspond à l'énergie cumulée des différents motifs de l'appariement, et l'accessibilité dépend de deux paramètres, ED^{mRNA} et ED^{sRNA}, qui correspon-

dent respectivement à l'énergie nécessaire pour que le mRNA et le sRNA soient accessibles en simple brin.

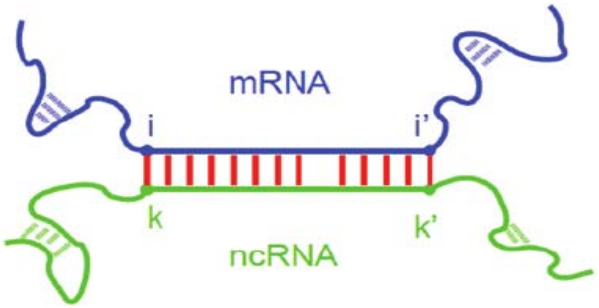


FIGURE 4.21 – Séquence exemple considérée pour un repliement par une approche par accessibilité [Backofen et Hess 2010].

Le calcul de la valeur de ED pour l'ensemble des structures S qui peuvent être formées par une séquence s compris entre les positions a et b (du site d'interaction) correspond à :

$$ED(a,b) = E^{ens}(S_{a,b}^{unpaired}) - E^{ens}(S) \tag{4.7}$$

où $E^{ens}(S)$ correspond à l'énergie de l'ensemble des structures de S et $E^{ens}(S_{a,b}^{unpaired})$ correspond à l'énergie des structures de S dont les nucléotides $\{S_a, S_{a+1}, ..., S_b\}$ ne sont pas appariés tel que :

$$\begin{aligned} E^{ens}(S) &= -RT\ln(Z_S) \\ E^{ens}(S_{a,b}) &= -RT\ln(Z_{S_{a,b}}) \end{aligned} \tag{4.8}$$

où R correspond à la constante de boltzmann, T correspond à la température, et Z_S et $Z_{S_{a,b}}$ correspondent à la fonction de partition de S et du site d'interaction $S_{a,b}$ considérée à partir des paramètres de la structure (FIG. 4.22).

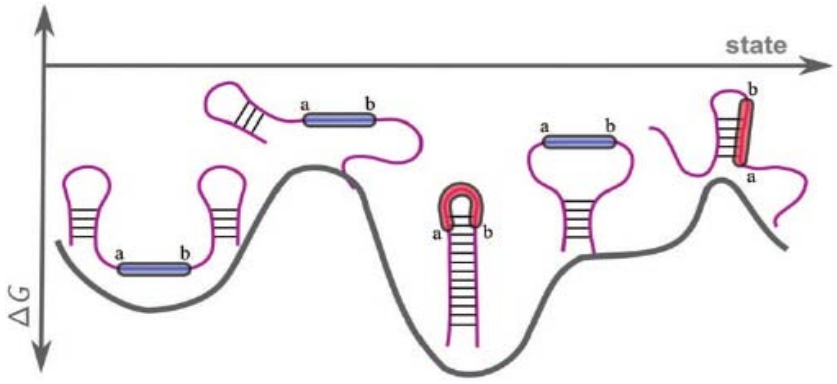
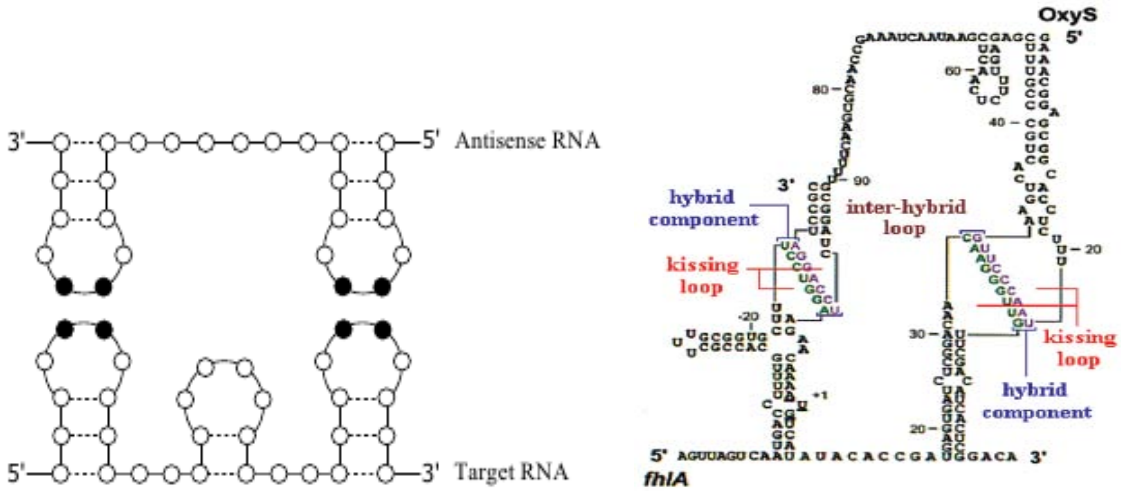


FIGURE 4.22 – Diagramme représentant l'énergie de différentes structures Q possibles de la séquence s [Backofen et Hess 2010].
Pour un site d'interaction donné entre les positions a et b de la séquence s, plusieurs structures peuvent être adoptées où le site est en simple brin (structures ovales bleues) ou en interactions intramoléculaires (structures ovales rouges). Pour le calcul de la fonction de partition $Z_{S_{a,b}}$, seules les structures qui ne sont pas appariées (en simple brin) seront considérées dans la somme des énergies $E(Q)$.

Ainsi, cette approche employée par les logiciels : IntaRNA [Busch *et al.* 2008], RNAup [Muckstein *et al.* 2006], BistaRNA [Poolsap *et al.* 2010] et Ractip [Kato *et al.* 2010], ne limite pas la forme des zones en interaction comme les approches basées sur la structure secondaire. Cependant, un seul site d'interaction entre les séquences pourra être considéré par ces méthodes. Cette restriction est en effet essentielle pour limiter la difficulté du problème à résoudre [Alkan *et al.* 2006]. Des interactions complexes entre des éléments structuraux, tels que les doubles *kissing-loop* qui correspondent à deux boucles terminales en interaction, ne pourront pas être prédits par ces méthodes (Fig. 4.23).



(a) Schéma d'interaction en *kissing-loop*

(b) Interaction en *Kissing loop* observée entre OxyS et fhlA

Figure 4.23 – Exemples d'interactions sous la forme de *kissing-loop* présentée (a) sous sa forme théorique [Poolsap *et al.* 2010] et (b) dans l'interaction du sRNA OxyS et du mRNA *fhlA* [Argaman et Altuvia 2000].

4.3 Conclusion

Les ARN non codants présentent un intérêt croissant de par leur importance dans le fonctionnement des cellules, mais également par le développement des technologies de séquençage, qui permettent de les détecter bien plus rapidement qu'auparavant avec un coût moindre. Comme nous l'avons vu, plusieurs méthodes bioinformatiques ont été développées pour détecter les sRNAs mais aussi prédire leurs cibles. Nous nous sommes intéressés dans cette thèse à la question de la sensibilité/spécificité de ces méthodes afin de déterminer si elles constituent des moyens réalistes pour l'identification de cibles. Nous avons également considéré l'efficacité de ces méthodes par rapport à la quantité d'information de plus en plus importante dont on disposait sur les sRNAs et sur les organismes étudiés. Dans ce contexte, nous avons développé le pipeline d'analyse iRNA qui se propose d'intégrer et d'évaluer ces différentes méthodes, mais également d'augmenter leur efficacité en intégrant le contexte de l'organisme. Nous proposons également au travers de ce système un logiciel de visualisation évolué permettant de mieux analyser ces prédictions. Ce système se base pour cela, sur la recherche de plusieurs caractéristiques connues des sRNAs, telles que la longueur restreinte des interactions des sRNAs, la présence de motifs de régulation multiples et conservés comme les motifs SIM ou DOR, la présence de zone d'interaction dans les sRNAs leur permettant d'interagir avec plusieurs mR-

NAs, leur implication dans la réponse au stress ou encore l'existence de boucle de régulation. Il s'agit enfin de s'appuyer aussi sur les caractéristiques communes d'interactions prédites avec des interactions déjà vérifiées expérimentalement pour certains sRNAs, pour cela nous proposons de considérer en parallèle la prédiction des cibles de plusieurs sRNAs.

Chapitre 5

iRNA : Pipeline dédié à la prédiction des cibles des sRNAs

Sommaire

Nous présentons dans ce chapitre le pipeline d'analyse iRNA qui a été développé durant cette thèse. iRNA permet de prédire des cibles potentielles pour les sRNAs d'un organisme donné. Il intègre pour cette prédiction les données issues des logiciels de prédiction des interactions (voir Section 4.2.4) et prend aussi en compte les données de contexte disponibles dans les bases de données de l'organisme (voir Section 2.2.2) grâce à une analyse d'enrichissement. iRNA fournit enfin un système de visualisation des prédictions basé sur les multigraphes, permettant de parcourir et d'affiner les prédictions. Nous verrons en premier lieu les raisons qui ont motivé ce développement, puis nous présenterons le principe de fonctionnement du pipeline et les résultats obtenus pour l'évaluation des performances des différentes étapes de iRNA. Nous discuterons enfin au travers d'exemples d'utilisations de iRNA de l'apport de ce pipeline pour la prédiction des cibles de sRNAs.

5.1 Motivations

Plusieurs approches bioinformatiques ont été développées pour prédire l'interaction d'un sRNA et d'un mRNA. Ces méthodes se basent pour cette prédiction sur la séquence (en recherchant une zone d'appariement), la thermodynamique, la structure secondaire ou encore l'accessibilité des sites d'interaction (voir Section 4.2.4). Ces méthodes permettent également de distinguer la qualité des appariements pour identifier les cibles potentielles. Elles se basent pour cela sur un score d'énergie, une p-valeur ou encore une classification directe des cibles. Cependant la caractérisation fonctionnelle d'un sRNA reste une tâche difficile car ces méthodes présentent plusieurs inconvénients. Une des principales constatations porte sur l'abondance

de cibles prédites pour un sRNA. En effet, plus d'une centaine de cibles potentielles peuvent être identifiées, et bien que la connaissance des sRNAs soit encore incomplète, il est estimé qu'une part importante de ces cibles constituent des faux positifs [Mallick et Ghosh 2012]. Deuxièmement, il apparaît que ces méthodes ne permettent pas une analyse à haut débit des sRNAs. Comme nous l'avons vu précédemment, l'évolution des techniques expérimentales et l'intérêt accru pour les sRNAs ont permis la recherche et l'identification de plus d'une centaines de sRNAs par organisme (voir Section 4.2.3). Il devient donc intéressant de pouvoir prédire parallèlement les cibles de ces sRNAs, et de rechercher dans ces prédictions la présence de motifs de régulations (voir Section 4.2.2). Enfin, la visualisation et la fouille des données constituent également des critères de plus en plus importants pour améliorer la prédiction des cibles des sRNAs. La combinaison de ces deux approches en même temps permettrait de donner plus de crédibilité à l'utilisateur pour restreindre le nombre de cibles prédites sur la base d'autres caractéristiques que celles prévues par l'analyse.
Différentes solutions ont ainsi été proposées pour améliorer la puissance de prédiction des méthodes actuelles. Une première solution porte sur l'intégration du contexte en considérant les informations disponibles sur les mRNAs dans les bases de données par une analyse d'enrichissement. L'idée de cette analyse consiste à filtrer parmi les cibles d'un sRNA, les mRNAs intervenant par exemple sur une même voie métabolique, en considérant les données issues du KEGG, ou encore en recherchant les mRNAs codants pour des protéines impliquées dans un même processus et/ou dans le même compartiment cellulaire, grâce aux données de la *Gene Ontology* [Ashburner *et al.* 2000]. Il a en effet été montré que par le biais d'un seul sRNA, la cellule pouvait en réponse à un seul signal réguler plusieurs mRNAs. Le logiciel myMIR [Corrada *et al.* 2011], dédié à la prédiction des cibles des micro-ARN, a ainsi combiné cette analyse aux méthodes d'identification de l'interaction, et observé une réduction importante du nombre de cibles prédites ainsi qu'une précision de prédiction accrue. Une deuxième solution porte sur la visualisation des prédictions sous la forme d'un graphe intégrant par le biais du dessin les différentes connaissances biologiques du réseau. Cette approche a été employée par [Modi *et al.* 2011] avec le logiciel CLR pour étudier sept sRNAs de *E. coli* à partir des données de la *Gene Ontology* et de la base de données EcoCyc (Fig. 5.1). Elle a montré que la mise en relation de l'ensemble de ces éléments permettait d'améliorer considérablement l'analyse des prédictions, et ainsi de proposer plus efficacement un rôle aux sRNAs. Cependant, ce dernier outil n'offre pas une méthode systématique et généralisable à l'analyse d'autres sRNAs que ceux étudiés. Par ailleurs, la visualisation générée ne permet pas d'interaction plus poussée avec le graphe pour filtrer sur la base d'autres critères les prédictions.
Le travail que nous avons réalisé se présente comme une méthode permettant une analyse plus approfondie des prédictions des cibles des sRNAs. Un première étape de ce travail a consisté à intégrer et à sélectionner l'outil de prédiction des cibles le plus performant. Pour cela, nous avons mis en place une méthode d'évaluation des performances des différents outils basée sur l'approche proposée par [Busch *et al.* 2008]. Une seconde étape a consisté à coupler cette prédiction avec une analyse par enrichissement des données. Nous avons pour cela interfacé notre outil avec la base de connaissances DAVID [Huang *et al.* 2007b]. Enfin, nous avons développé un système de visualisation avancé des prédictions basé sur le logiciel Tulip [Auber 2003]. Nous nous sommes particulièrement intéressés au cours de ce travail à développer des méthodes permettant une analyse de large échelle des cibles des sRNAs, en distribuant les calculs de prédictions et en créant des méthodes de visualisation pertinentes pour filtrer les données de prédiction des sRNAs. Ce travail a donc pour cela été réalisé en collaboration avec Jonathan Dubois et Romain Bourqui de l'équipe MaBioVis, du LaBRI à Bordeaux, pour le développement du système de visualisation.

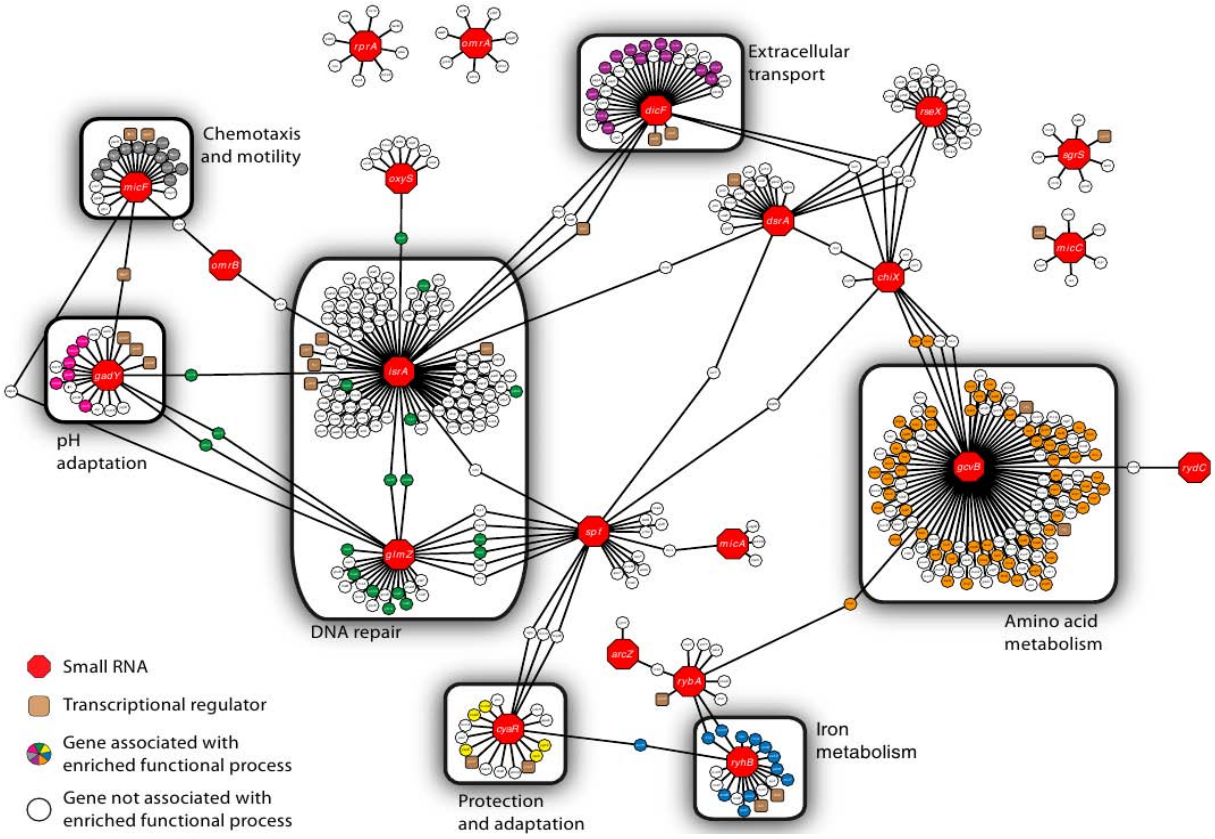


FIGURE 5.1 – Visualisation des interactions des sRNAs sous la forme de graphe [Modi *et al.* 2011].

L'analyse proposée par [Modi et al. *2011] se base sur des données d'une puce à ADN et d'analyse d'enrichissement pour inférer les cibles de sRNAs de E. coli. Les données expérimentales ont ainsi été étendues par les sRNAs partageant un même contexte. Les cercles rouges correspondent aux sRNAs, les cercles blancs et marrons correspondent aux mRNAs associés respectivement à partir des données de puces à ADN ou connus comme étant des régulateurs de la transcription. Les mRNAs d'autres couleurs correspondent à des mRNAs associés par le processus d'enrichissement, en vert pour les mRNAs impliqués dans les processus de réparation, en jaune pour la protection et adaptation cellulaire, et en rose pour adaptation au pH.*

5.2 Principe

5.2.1 Données biologiques de l'étude

Afin de réaliser cette étude, nous avons constitué deux jeux de données d'interactions. Nous présentons dans un premier temps le jeu de données de test utilisé pour l'évaluation et l'estimation des meilleurs paramètres des logiciels de l'étude. Ce jeu de données est exclusivement constitué de vraies et non- interactions démontrées expérimentalement. Dans un second temps, nous présentons le jeu de données de sRNAs de *E. coli* K12. Ce jeu de données est constitué de sRNAs connus ou nouvellement identifiés par expérimentation, dont nous avons cherché à prédire les cibles. Notre approche sera ainsi validée au travers de l'analyse des prédictions pour le jeu de données de test, avant de considérer la prédiction des cibles pour les sRNAs de *E. coli*.

Données de test

Les données de test que nous avons compilées sont issues de la sRNATarBase [Cao *et al.* 2010] et de [Peer et Margalit 2011]. Elles comprennent 43 sRNAs et 52 mRNAs qui réalisent entre eux, 62 vraies interactions et 41 non-interactions identifiées chez plusieurs bactéries : *Azotobacter vinelandii*, *Escherichia coli*, *Listeria monocytogenes*, *Pseudomonas aeruginosa*, *Perkinsus marinus* et *Salmonella enterica*. Nous disposons pour chacun des couples en interaction, de la position de début et de fin de l'interaction sur le sRNA et sur le mRNA. Ces régions seront considérées comme contigües au cours de l'analyse, en raison de l'incertitude relative aux nucléotides en interactions. En effet, les méthodes expérimentales employées pour caractériser ces interactions nous donnent les bornes de début et de fin de ces interactions à l'intérieur desquels les interactions de seuls quelques nucléotides ont été validées expérimentalement, les autres étant généralement identifiées par prédiction. Nous avons aussi raccourci les mRNAs pour se focaliser sur la région -150 à +50 nucléotides autour du codon d'initiation comme le propose [Busch *et al.* 2008], de manière à encadrer le RBS où se produit l'essentiel des interactions. Ceci permet également de diminuer la complexité du problème à résoudre et d'augmenter la précision des logiciels de prédiction. L'ensemble des informations pour ces données sont disponibles en annexe (voir Annexes A.1 et A.2).

Données de *Escherichia coli* K12

Nous avons pris pour *E. coli* les données issues de deux études : [Raghavan *et al.* 2011] et [Shinhara *et al.* 2011]. Ces analyses ont recherché par une approche expérimentale basée sur le *RNA-seq* et le *northern blot* de nouveaux sRNAs. 85 sRNAs déjà connus ont été retrouvés par ces études (Fig. 5.2). Dans le cas de l'étude de [Raghavan *et al.* 2011], 63 nouveaux sRNAs ont été également identifiés, dont 53 avaient été prédits auparavant par approche bioinformatique [Argaman *et al.* 2001; Rivas et Eddy 2001; Chen *et al.* 2002; Yachie *et al.* 2006; Tran *et al.* 2009], et 10 n'avaient jamais été observés. Pour l'étude de [Shinhara *et al.* 2011], 229 sRNAs candidats ont été identifiés parmi lesquels 44 sRNAs ont un promoteur σ identifié par prédiction et expérimentation, 69 sRNAs ont un promoteur σ identifié uniquement par expérimentation et 119 sRNAs ont un promoteur σ uniquement identifié par prédiction. Nous avons sélectionné pour cette étude uniquement les sRNAs qui avaient un promoteur identifié expérimentalement.
Nous avons donc obtenu une liste de 261 sRNAs, pour lesquels nous avons cherché des cibles parmi les 4142 séquences codantes de *E. coli* K12 MG1655 identifiées dans les données EMBL [Blattner *et al.* 1997]. Nous avons comme pour les données de test raccourci les séquences des mRNAs pour se focaliser sur la région -150 à +50. Ainsi, 1.081.062 interactions potentielles devront être considérées pour la prédiction. Les informations supplémentaires relatives à ces séquences sont disponibles en annexe (voir Annexe A.2).

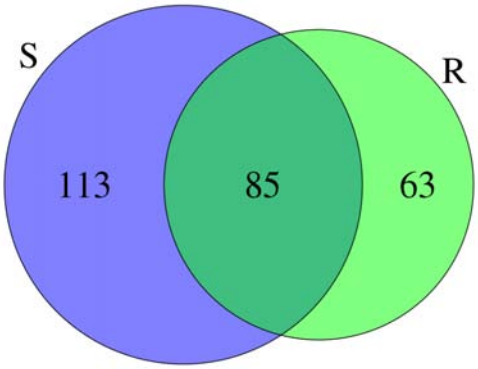


FIGURE 5.2 – Diagramme de Venn représentant la distribution des sRNAs identifiés par [Shinhara *et al.* 2011] en bleu (S) et [Raghavan *et al.* 2011] en vert (R).

5.2.2 iRNA

Nous décrivons dans cette partie les différentes étapes du pipeline iRNA (FIG. 5.3) qui a été développé durant cette thèse.

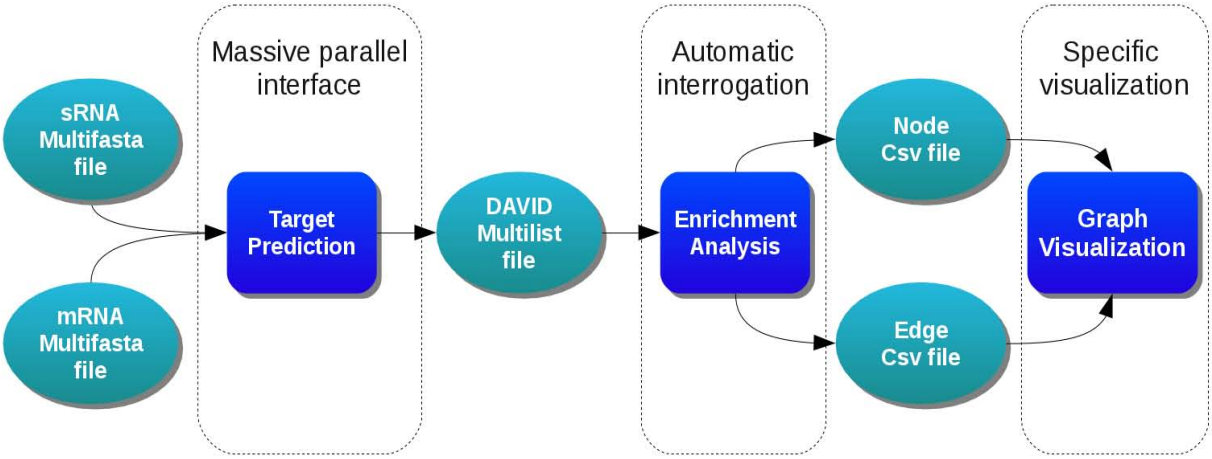


FIGURE 5.3 – Schéma de fonctionnement de l'analyse de iRNA.
iRNA prend deux types de fichier en entrée, un fichier multifasta contenant les séquences des sRNAs et un fichier multifasta ou le fichier EMBL/Genbank contenant tous les mRNAs de l'organisme étudié. La première étape de l'analyse consiste à prédire pour tous les sRNAs et les mRNAs, ceux qui sont en interaction grâce aux outils de prédiction. Cette étape est parallélisée à l'aide de la librairie MPI. Chacun des cœurs de calculs prédit l'interaction d'un couple de sRNA et mRNA. Les interactions sont ensuite sélectionnées d'après le seuil de significativité du résultat et les cibles de chaque sRNA sont soumises à une analyse par enrichissement. Les résultats obtenus sont ensuite visualisés sous la forme d'un graphe.

Ce pipeline se divise en trois parties. La première et seconde partie sont constituées de 4 programmes en Python permettant de réaliser les prédictions et l'enrichissement. Le premier logiciel iRNA_seq permet d'extraire les données nécessaires issues des fichiers Genbank ou

EMBL. Le second logiciel iRNA_pred prend en charge la prédiction des cibles des différents logiciels. Ce programme gère automatiquement l'exécution et l'extraction des résultats retournés par ces logiciels, qui sont enregistrés sous la forme d'une base de données Sqlite. Le troisième logiciel iRNA_stat réalise l'analyse statistique des résultats et enfin le logiciel David2tulip permet d'effectuer l'interrogation automatique du web-service de la base de connaissances DAVID [Huang *et al.* 2007b] et de produire les deux fichiers de données csv nécessaires à la seconde partie du pipeline. La troisième partie du pipeline désigne le logiciel iRNA_visu qui permet d'analyser et de filtrer de manière visuelle les résultats obtenus par les précédentes étapes (Fig. 5.4). Ce logiciel réalisé en C++ est basé sur le logiciel de visualisation de graphe Tulip [Auber 2003]. Il intègre des filtres décrits en C++ ou en Python.

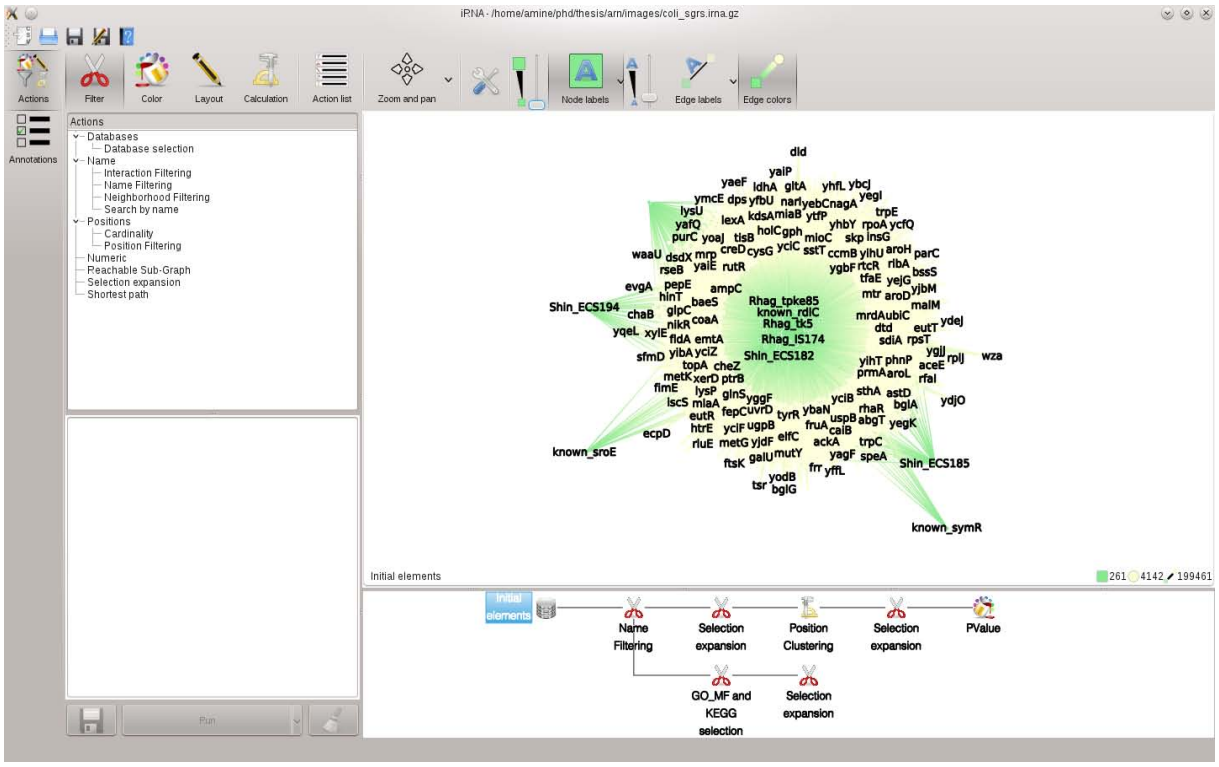


Figure 5.4 – Interface de iRNA

Les logiciels de la première et de la seconde partie dépendent de plusieurs librairies Python et R :

(i) les librairies scipy et numpy sont utilisées pour les étapes de calcul,

(ii) la librairie lxml, pour stocker transitoirement durant la prédiction les résultats des différents logiciels de prédiction sous forme de fichier XML,

(iii) la librairie sqlitebck (optionnel), pour optimiser l'utilisation du système de gestion de base de données Sqlite dans ce contexte de grande quantité de données

(iv) la librairie Rpy2 permettant d'interfacer R depuis Python,

(v) la librairie MPI4py est utilisé pour distribuer les calculs réalisés par iRNA_pred et iRNA_stat,

(vi) et enfin, la librairie pROC [Robin *et al.* 2011] pour R est pour analyser les données statistiques des courbes de ROC.

L'installation des 4 premiers logiciels du pipeline est automatiquement réalisée grâce aux

setuptools de Python sous Linux. Le logiciel iRNA_visu est, quant à lui, disponible en version pré-compilée pour Windows et Linux.
Afin d'expliciter le travail réalisé au sein de iRNA, nous présentons en premier lieu la procédure d'intégration et d'évaluation des différents logiciels de prédiction. Nous détaillons ensuite l'analyse d'enrichissement réalisée. Enfin nous présentons le système de visualisation qui a été développé pour analyser les résultats de prédiction.

Sélection et intégration des logiciels de prédiction

Nous avons sélectionné pour cette étude plusieurs logiciels de prédiction des interactions, en incluant des approches spécifiques et non-spécifiques aux sRNAs. Le principal critère de notre sélection se basait sur la disponibilité de ces logiciels et sur leur vitesse d'exécution.
Sur un panel de 25 logiciels potentiels, quatre logiciels ont été retirés du comparatif à cause de leur temps de calcul prohibitif : lara [Bauer *et al.* 2007], inteRNA [Alkan *et al.* 2006], piRNA [Chitsaz *et al.* 2009] et rip [Huang *et al.* 2009b]. Pour ce comparatif, nous avons également exclu les méthodes uniquement accessibles sous la forme d'application web : TargetRNA [Tjaden *et al.* 2006], RNApredator [Eggenhofer *et al.* 2011], sRNAtarget [Cao *et al.* 2009] et sTarPicker [Ying *et al.* 2011]. Enfin les logiciels : inRNA [Salari *et al.* 2010], biRNA [Chitsaz *et al.* 2009], RIG [Kato *et al.* 2009] et picTar [Krek *et al.* 2005] ont dû être exclus car ils n'étaient pas disponibles. Nous avons ainsi sélectionné 13 logiciels (Fig. 5.5) appartenant aux quatre types d'approches identifiées (voir Section 4.2.4).

Logiciels	e-valeur	Score	MFE	Énergie	Références
Blastall	X				[Altschul *et al.* 1990]
Ssearch	X				[Pearson 1991]
Yass	X				[Noe et Kucherov 2005]
BistaRNA		X			[Poolsap *et al.* 2010]
Guugle		X			[Gerlach et Giegerich 2006]
Ractip		X			[Kato *et al.* 2010]
Pairfold			X		[Andronescu *et al.* 2005]
RNAduplex			X		[Hofacker 2003]
RNAhybrid			X		[Rehmsmeier *et al.* 2004]
RNAplex			X		[Tafer et Hofacker 2008]
IntaRNA				X	[Busch *et al.* 2008]
RNAcofold				X	[Bernhart *et al.* 2006]
RNAup				X	[Muckstein *et al.* 2006]

Figure 5.5 – Liste des logiciels sélectionnés pour cette étude. Le type de résultat considéré pour les interactions prédites par ces logiciels est indiqué par une croix.

Les logiciels sélectionnés ont été utilisés avec les paramètres conseillés dans leurs publications et/ou avec les paramètres proposés par défaut. Nous avons ainsi obtenu un total de 19 comparaisons possibles, dont le détail est disponible en annexe (voir Annexe A.3).
Deux traitements sont réalisés sur ces résultats lors de leurs évaluations. Pour les résultats de type score, MFE et d'énergie, nous avons normé la valeur obtenue par rapport à la longueur du sRNA et du mRNA donné comme le propose [Rehmsmeier *et al.* 2004] (Eq. 4.1). Cette normalisation a montré avoir un impact positif sur la qualité des résultats en améliorant

légèrement la sensibilité des logiciels, pour une même spécificité. Concernant les positions d'interactions, nous les avons considérées comme contigües entre les bornes minimales et maximales de l'interaction, comme pour les données expérimentales (voir Section 5.2.1).

Procédure d'évaluation des logiciels

Deux critères sont considérés pour évaluer les logiciels [Busch *et al.* 2008] : la prédiction des vraies- et non-interactions, et la précision de la prédiction.

Prédiction des vraies- et non-interactions La prédiction des vraies- et non-interactions consiste à déterminer dans quelle mesure ces logiciels sont capables d'identifier les sRNAs et les mRNAs qui interagissent réellement ensemble. Cette mesure est essentielle puisqu'elle nous permet également d'estimer le pourcentage de vraies cibles présentes dans les prédictions de ces logiciels.
Pour réaliser cette étude, nous avons prédit avec chacun des logiciels, les appariements des sRNAs et des mRNAs de notre jeu de test. Le résultat de chaque logiciel associé au meilleur appariement prédit pour chaque couple sRNA et mRNA est considéré. Nous avons ensuite calculé à partir de ces données la courbe de ROC (*Receiver Operating Characteristic*) grâce au package R : pROC [Robin *et al.* 2011]. Cette courbe nous permet de mesurer la performance du logiciel en fonction de la sensibilité et de la spécificité à différents seuils (Fig. 5.6) (pour revue voir [Powers 2007]), tel que :

$$\begin{aligned} Sensitivity &= \frac{TP}{P} = \frac{TP}{TP+FN} \\ Specificity &= \frac{TN}{N} = \frac{TN}{TN+FP} \end{aligned} \tag{5.1}$$

avec TP (*True Positive*) qui correspond au nombre de vrais positifs, TN (*True Negative*) qui correspond au nombre de vrais négatifs, FP (*False Positive*) qui correspond au nombre de faux positifs et FN (*False Negative*) qui correspond au nombre de faux négatifs.

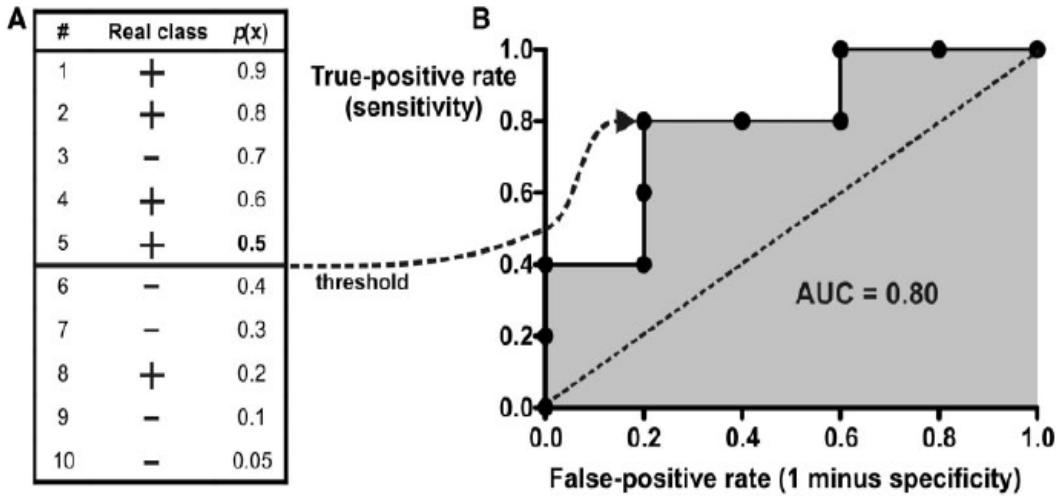


FIGURE 5.6 – Exemple de courbe de ROC [Berrar et Flach 2011].
Dix cas de tests sont classés de manière décroissante sur la base de leur score (A). Chaque seuil de score est associé à un taux spécifique de vrais et faux positifs. Par exemple pour un seuil de score de 0.45 *(ou n'importe quelle valeur entre 0.4 et 0.5), nous avons un cas négatif (#3) et un cas positif (#8) mal classé. Le taux de faux positifs résultant est de 0.2 ($False-positive\ rate = 1-Specificity = 1-(4/5)$) et de vrais positifs est de 0.8 ($Sensitivity = 4/5 = 0.8$). Lorsque l'on procède de manière analogue pour tous les seuils possibles, nous obtenons une courbe de ROC (B).*

L'interprétation de la courbe de ROC s'effectue selon son AUC (aire sous la courbe - *Area Under the Curve*) [Hand et Till 2001] dont on peut estimer l'intervalle de confiance par la méthode de [DeLong *et al.* 1988]. On estime ainsi qu'une classification est :

- Non satisfaisante, si $AUC < 0.5$,
- Peu satisfaisante, si $0.5 < AUC < 0.7$,
- Très satisfaisante, si $0.7 < AUC < 1$,
- Parfaite, si $AUC = 1$.

À cette étape, nous avons également estimé le seuil où l'AUC est maximale par l'index de Youden. Cet index correspond à la valeur où la sensibilité et la spécificité sont maximales. Il est ici estimé à partir du jeu de données de test. Ce seuil sera ensuite employé pour les données de *E. coli* et les autres organismes, afin d'identifier pour le logiciel sélectionné les vraies-interactions (dont la valeur d'énergie sera supérieure à ce seuil) et les non-interactions (dont la valeur d'énergie sera strictement inférieure à ce seuil) (FIG. 5.7).

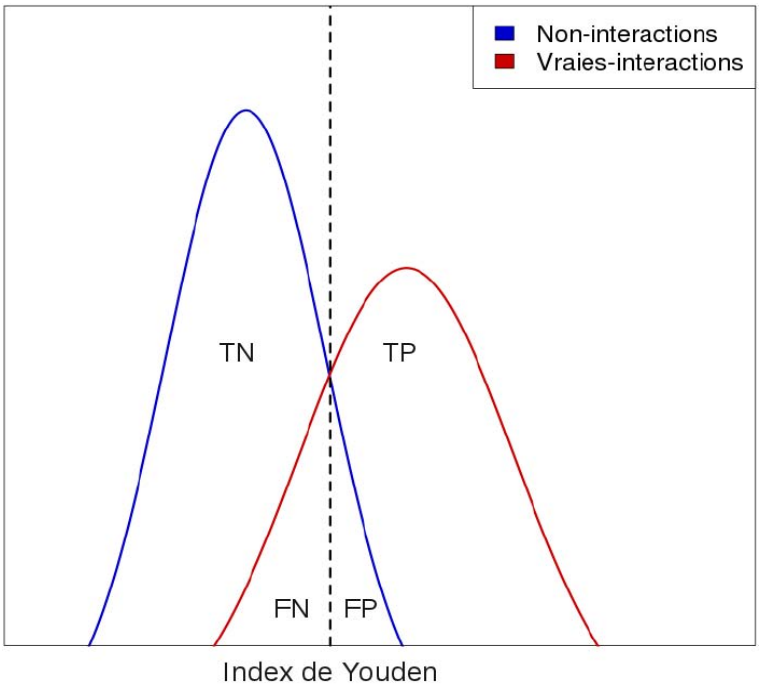


FIGURE 5.7 – L'index de Youden identifie le seuil de score discriminant au mieux les vraies- et non-interactions (TN - TP), tout en minimisant au mieux le taux de faux négatifs (FN) et de faux positifs (FP).

Estimation de la précision de la prédiction La précision de la prédiction se rapporte quant à elle à la capacité des logiciels à identifier la vraie zone d'interaction entre les sRNAs et les mRNAs. Cette mesure proposée par [Do *et al.* 2006] pour évaluer la précision des logiciels dépend de la sensibilité et la *Positive Predictive Value* (PPV), tel que :

$$\begin{aligned} Sensibility &= \frac{TP}{P} = \frac{Number\ of\ correctly\ predicted\ base\ pairings}{Number\ of\ true\ base\ pairings} \\ PPV &= \frac{TP}{TP+FP} = \frac{Number\ of\ correctly\ predicted\ base\ pairings}{Number\ of\ predicted\ base\ pairings} \end{aligned} \tag{5.2}$$

où la sensibilité dépend du nombre de bases appariées correctement prédit en fonction du nombre de bases de l'interaction et le PPV dépend du nombre de bases appariées correctement prédit par rapport au nombre de bases prédites par le logiciel.

Analyse par enrichissement

Pour les cibles identifiées comme étant des interactions potentielles lors de l'étape de prédiction des interactions, nous proposons au travers de iRNA de considérer également leur enrichissement. L'idée de cette étude est de rechercher parmi les cibles de chaque sRNA celles qui partagent une même annotation dans les bases de données. Cette analyse présente un intérêt particulier pour restreindre efficacement le nombre de cibles, en considérant par exemple les cibles qui partagent une annotation avec les interactions déjà connues ou encore en identifiant une annotation commune parmi les cibles d'un sRNA pouvant par exemple correspondre à un motif de régulation de type SIM ou DOR.
Parmi les 68 outils disponibles pour enrichir des groupes de gènes [Huang *et al.* 2009a], nous

avons sélectionné la base de connaissances DAVID [Huang *et al.* 2007b]. Cet outil constitue l'une des principales bases de connaissances en permettant de considérer pour l'enrichissement les données issues de 10 bases de données pour de nombreux organismes (FIG. 5.8). Contrairement à de nombreux outils d'enrichissement, DAVID n'est pas spécifique à quelques organismes. Il offre également un accès par web-service permettant d'interfacer automatiquement l'analyse d'enrichissement au sein de iRNA.

	Bases de données d'annotations	Liens
1 2 3	GO Biological Process (GO_BP) GO Molecular Function (GO_MF) GO Cellular Component (GO_CC)	http ://www.geneontology.org
4	COG/KOG Ontology	http ://www.ncbi.nlm.nih.gov/COG/new/
5	SMART Domains	http ://smart.embl-heidelberg.de/
6	InterPro Domains	http ://www.ebi.ac.uk/interpro/
7	KEGG Pathways	http ://www.genome.jp/kegg/
8	UniProt Sequence Features	http ://www.pir.uniprot.org/
9	Swiss-Prot Keywords	http ://www.pir.uniprot.org/
10	PIR SuperFamily Names	http ://pir.georgetown.edu/iproclass/

FIGURE 5.8 – Bases de données disponibles pour un enrichissement dans DAVID [Huang *et al.* 2007a].

Nous présentons donc ici le principe de l'analyse d'enrichissement réalisée par DAVID. L'enrichissement réalisé par DAVID repose sur une analyse de type SEA (*Singular enrichment analysis*) [Huang *et al.* 2009a], qui constitue l'approche classiquement employée pour l'enrichissement. Trois étapes sont réalisées :

Première étape - Calcul de la distance entre les annotations des cibles La première étape de l'étude consiste à déterminer la distance entre les cibles qui partagent une ou plusieurs annotations (FIG. 5.10A). Les gènes des mRNAs sont ici considérés. L'hypothèse initiale étant que si deux gènes partagent une annotation, ils doivent être liés d'un point de vue fonctionnel [Huang *et al.* 2007a]. Pour cela, DAVID identifie pour chaque cible les termes d'annotation qui lui sont associés dans les différentes bases de données (FIG. 5.9A), puis calcule la distance entre ces cibles en se basant sur le test du Kappa [Cohen 1960; Byrt *et al.* 1993]. Ce test permet de mesurer le degré de co-occurence des annotations entre les gènes. Ainsi, pour deux profils d'annotation des gènes m et n, la valeur du Kappa K_{mn} est égale à :

$$K_{mn} = \frac{O_{mn} - A_{mn}}{1 - A_{mn}} \tag{5.3}$$

où O_{mn} représente la co-occurrence observée et A_{mn} représente les chances de co-occurrence (FIG. 5.9B). La valeur du Kappa est maximale lorsqu'il est égal à 1.

A	Cell death	Apoptosis	Ph domain	Sh2 domain	Apoptosis pathway	Membrane
Gene a	1	1	0	0	1	0
Gene b	1	1	0	1	1	0
Gene c	1	0	0	1	1	1
Gene d	1	1	0	0	1	1
Gene e	0	1	1	1	1	1
Gene f	0	0	1	1	0	1
Gene g	0	0	1	1	0	1

B		Gene a: 1	Gene a: 0	Row total
Gene b	1	3 ($C_{1,1}$)	1 ($C_{0,1}$)	4 ($C_{1,\cdot}$)
	0	0 ($C_{0,1}$)	2 ($C_{0,0}$)	2 ($C_{0,\cdot}$)
Column total		3 ($C_{\cdot,1}$)	3 ($C_{\cdot,0}$)	6 (T_{ab})

$$O_{ab} = \frac{C_{1,1} + C_{0,0}}{T_{ab}} = \frac{3+2}{6} = 0.83$$

$$A_{ab} = \frac{C_{\cdot,1} \bullet C_{1,\cdot} + C_{\cdot,0} \bullet C_{0,\cdot}}{T_{ab} \bullet T_{ab}} = \frac{3 \bullet 4 + 3 \bullet 2}{6 \bullet 6} = 0.5$$

$$K_{ab} = \frac{O_{ab} - A_{ab}}{1 - A_{ab}} = \frac{0.83 - 0.5}{1 - 0.5} = 0.66$$

FIGURE 5.9 – Un exemple de détection des relations entre les gènes des cibles par le test du Kappa [Huang *et al.* 2007a].
Chaque gène est ici associé aux annotations qui lui sont attribuées dans la base de données considérée (voir Figure 5.8) (A). Ceci permet de construire une matrice binaire donnant le profil d'annotation unique de chaque gène, où une valeur 1 correspond à une annotation du gène pour ce terme et 0 sinon. On construit ensuite une matrice de contingence pour estimer le degré d'accord entre les gènes (B). Un score de Kappa élevé indique que les gènes présentent une annotation fortement en accord par rapport à la chance.

Deuxième étape - Identification de l'ensemble des voisins Après le calcul de la distance entre les cibles, DAVID recherche l'ensemble des voisins de chaque cible. Pour cela, il se base sur une méthode d'agrégation appelée *heuristic fuzzy multiple-linkage partitioning* (FIG. 5.10B-C). Cette méthode consiste tout d'abord à rechercher pour chaque élément d'autres voisins avec lesquels il partage un score de kappa > 0.35. Les groupes ainsi formés sont ensuite fusionnés s'ils partagent au moins 50% d'éléments communs [Huang *et al.* 2007a].

Troisième étape - Estimation de la significativité des enrichissements identifiés La dernière étape consiste à déterminer la significativité des groupes par rapport aux données de l'organisme. Pour cela, deux hypothèses sont considérées :

- H_0 : L'ensemble identifié constitue une annotation non significative par rapport à l'organisme.
- H_1 : L'ensemble identifié constitue une annotation significative.

Pour répondre à cette question, DAVID effectue un test de Fisher exact [Agresti 1992] modifié. Ce test consiste tout d'abord à calculer la table de contingence. Celle-ci estime le nombre de gènes dans la liste de cibles et dans le génome de l'organisme qui partagent une annotation donnée par rapport au nombre total de gènes dans la liste de cibles et dans le génome, tel

que :

	Liste de cibles	Génome de l'organisme	Total
Annotation	a-1	b	a+b
Hors de l'annotation	c	d	c+d
Total	a+c	b+d	n

(5.4)

Le décompte des éléments positifs est ici pénalisé d'un facteur 1 dans l'implémentation de DAVID. La probabilité d'obtenir cette annotation dans la liste de cibles par rapport à la chance est ensuite déterminée par la distribution hypergéométrique, tel que :

$$EASE\_score = P = \frac{(a+b)!(c+d)!(a+c)!(b+d)!}{a!b!c!d!n!} \tag{5.5}$$

Ainsi, les termes dont la p-valeur dépassent le seuil de p-valeur défini par l'utilisateur sont reportés. Pour cette étude, nous avons utilisé le seuil proposé par défaut par DAVID ($EASE\_score = 0.1$) qui s'est révélé être le meilleur compromis pour les données de test (données non montrées).

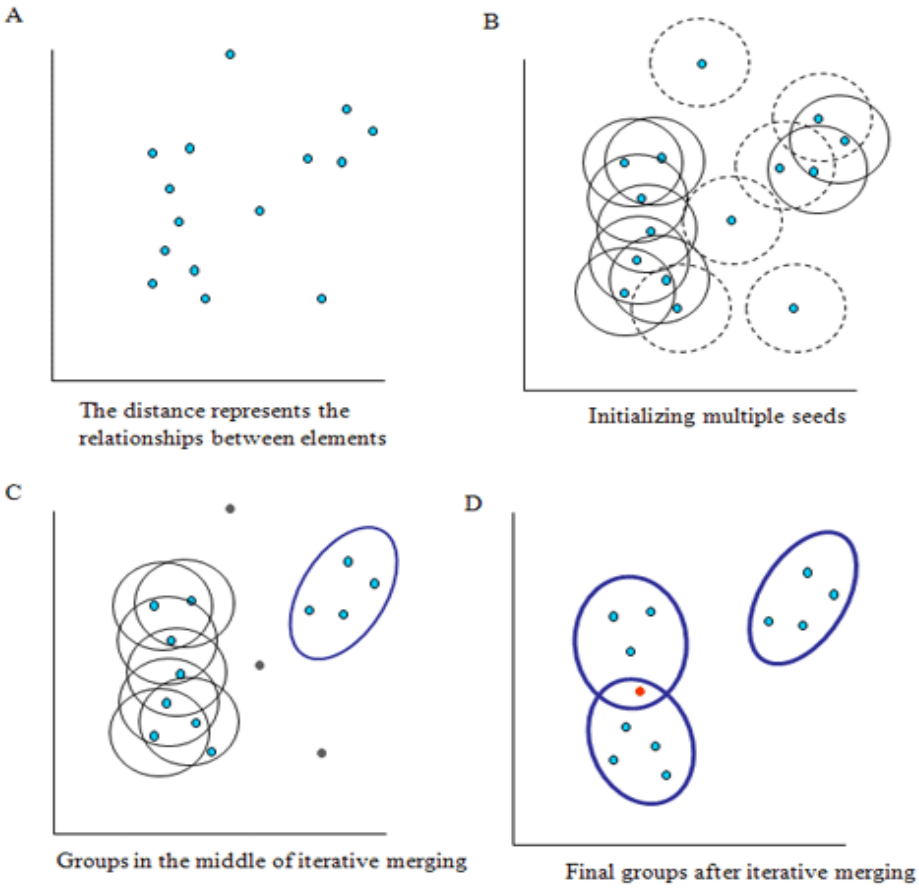


FIGURE 5.10 – Algorithme d'enrichissement de DAVID [Huang *et al.* 2007a].
L'algorithme d'enrichissement de DAVID se divise en trois étapes. La première étape consiste à estimer la distance entre les annotations des gènes des cibles grâce au test du Kappa (A). Les gènes sont ensuite regroupés avec leurs plus proches voisins par la méthode d'agrégation appelée heuristic fuzzy multiple-linkage partitioning sur la base d'un terme d'annotation (B-C). Enfin la significativité des groupes formés est enfin testée par un test de Fisher exact modifié (D).

Analyse visuelle des résultats

Afin d'analyser les résultats de prédiction, nous avons développé en collaboration avec Jonathan Dubois et Romain Bourqui de l'équipe MaBioVis, du LaBRI à Bordeaux, le logiciel iRNA-visu qui constitue une extension au logiciel Tulip [Auber 2003]. Tulip permet de visualiser, de dessiner et d'éditer des graphes, et le logiciel iRNA-visu propose d'adapter cette visualisation aux données des sRNAs et de leurs cibles. Il inclut pour cela différentes méthodes permettant de visualiser l'ensemble des données dont on dispose sur les cibles et leurs relations avec les sRNAs, mais également de pouvoir filtrer ces données en interagissant avec le graphe pour en sélectionner des cibles. Le travail que nous avons réalisé correspond à une réflexion commune des moyens pouvant être mis en place pour la visualisation, tels que le *Neighbors* et le filtre des annotations (décrits ci-dessous), ou encore la disposition globale de l'interface, le développement de ces outils ayant été réalisé par Jonathan Dubois. Nous avons donc fortement participé à l'élaboration du cahier des charges, à la conception et au test de chacun des éléments implémentés. Enfin, nous avons directement développé en Python, plusieurs filtres de sélection disponibles dans iRNA_visu (décrits ci-dessous).
Pour iRNA-visu, nous avons choisi de représenter les données d'interactions sous la forme d'un multigraphe biparti (Fig. 5.11). Les sRNAs sont représentées graphiquement par des carrés et les mRNAs sont représentées sous la forme de cercles. Ces éléments sont annotés par le nom du gène, par le Gene ID ou par les termes définis par l'utilisateur.

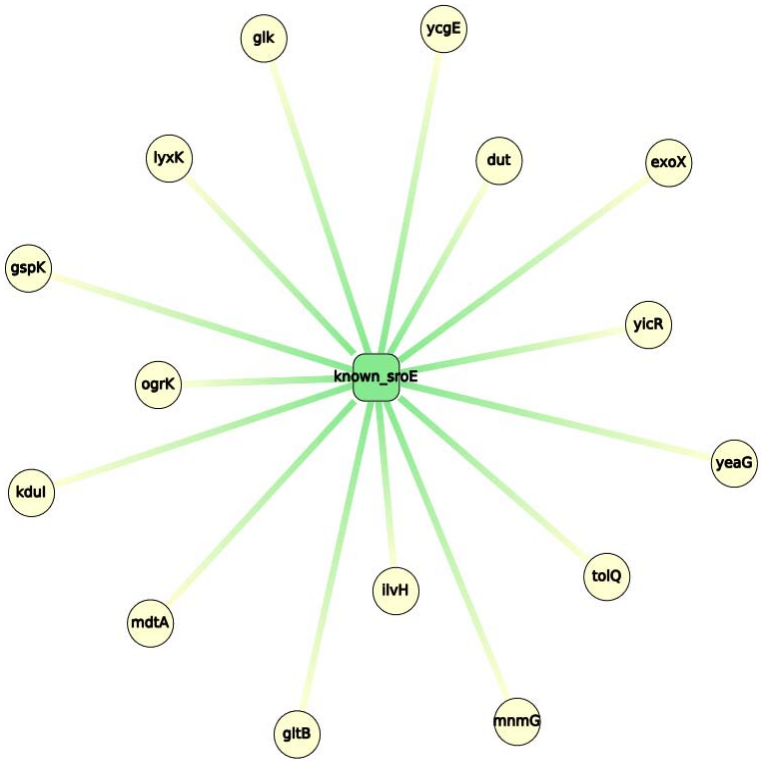


Figure 5.11 – Visualisation sous la forme d'un graphe biparti des cibles prédites du sRNA SroE chez *E. coli*, représentés respectivement par des cercles jaunes et un carré vert.

Les arêtes servent à représenter une interaction entre ces deux éléments (Fig. 5.12). Elles sont annotées sélectivement par les données d'enrichissement ou par la position d'interaction entre le sRNA et le mRNA ou encore par le score de l'enrichissement.

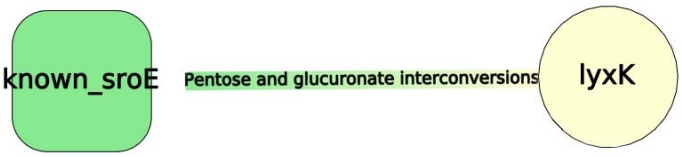


FIGURE 5.12 – Annotation selon l'attribut des données d'enrichissement issues du KEGG pour l'interaction du sRNA SroE et du mRNA lyxK.

Au sein de iRNA-visu, 11 graphes différents peuvent être considérés correspondant au graphe initial obtenu lors des prédictions des interactions et aux 10 sous-graphes correspondant aux éléments enrichis pour chacune des 10 bases de données considérés par DAVID (voir Figure 5.8). Nous obtenons donc un multigraphe contenant l'ensemble de ces données. Afin de visualiser cet ensemble important de données, différentes méthodes de visualisation ont été développées :

Visualisation spécifique des cibles Le premier point porte sur la création d'un visualiseur circulaire des cibles appelé *Neighbors* qui permet de centrer la vue sur un sRNA ou un mRNA et de visualiser sur un cercle respectivement les mRNAs ou les sRNAs interagissant avec le nœud sélectionné (FIG. 5.13). Cette visualisation comprend également une barre centrale (localisée sous le nœud sélectionné) qui représente la séquence de l'élément sélectionné. Elle permet de visualiser le nombre d'interactions dans lesquelles est impliqué chaque nucléotide (courbe en noire - Figure 5.13), mais également de visualiser et de sélectionner les sRNAs ou les mRNAs d'après la zone où ils interagissent avec leur cible. Pour cela, une classification des positions d'interactions des cibles sur l'élément considéré est réalisée. Celle-ci est calculée en deux étapes. Nous estimons tout d'abord le degré de recouvrement des interactions par le calcul de l'indice de Jaccard modifié. Pour deux interactions A et B de l'élément sélectionné qui surviennent respectivement aux positions (u, v) et (x, y) sur cet élément (avec $u < v$ et $x < y$), l'indice de Jaccard modifié J_m est égal à :

$$\begin{aligned} J_m(A,B) &= \frac{|A \cap B|}{min(A,B)} \\ J_m(A,B) &= \frac{min(v,y) - max(u,x)}{min(y-x, v-u)} \end{aligned} \tag{5.6}$$

Cette mesure range ainsi les interactions de 0 pour des ensembles disjoints, à 1 lors d'un recouvrement total des séquences entre elles. Appliquée à toutes les positions d'interaction pour chacune des cibles, nous obtenons une matrice de distance, où les éléments fortement liés entre eux sont recherchés par classification non supervisée via l'algorithme du MCL (*Markov Cluster Algorithm*) [Van Dongen 2000]. Une valeur par minimum de recouvrement de $J_m = 0.5$ est considérée par défaut pour que deux ensembles soient considérés comme joints.
Ce système de visualisation peut enfin être utilisé pour considérer les interactions d'un même élément selon plusieurs conditions (FIG. 5.14).

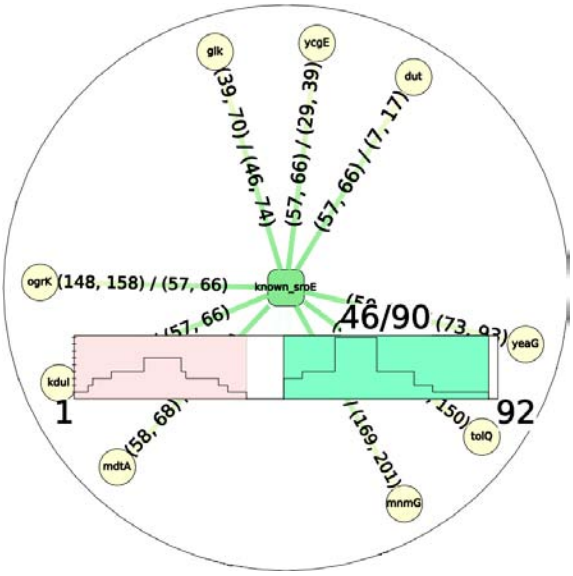


FIGURE 5.13 – Visualisation par le *Neighbors* des interactions du sRNA SroE.
La barre de séquence du sRNA SroE indique par une ligne noire le nombre d'interactions dans lesquelles est impliqué chaque nucléotide. On peut également observer que deux groupements pour les positions d'interactions ont identifiés au niveau des nucléotides en (1-38) et (46-90) (représentés par les boites rouges et vertes). On peut ainsi observer qu'il y a corrélation visuelle entre ces deux dessins.

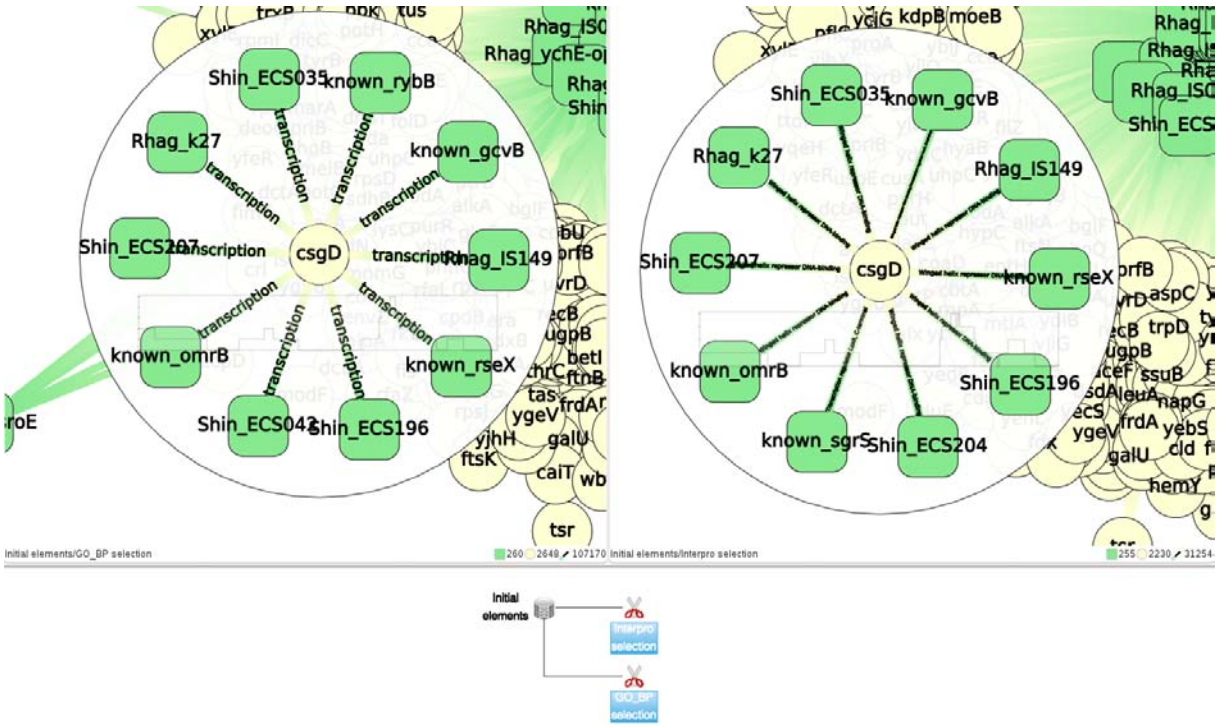


FIGURE 5.14 – Visualisation des sRNAs interagissant avec le mRNA modF.
Deux graphes sont ici considérés pour visualiser les interactions de csgD enrichies selon les bases de données GO_BP (à gauche) et Interpro (à droite). Ces visualisations sont automatiquement synchronisées.

Filtre des annotations Le second point porte sur l'intégration d'un système de filtre des annotations (Fig. 5.15). Ce filtre permet de sélectionner et de dessiner automatiquement le sous-graphe des interactions dont les attributs d'arêtes présentent une ou plusieurs des annotations sélectionnées. Plusieurs bases de données peuvent ainsi être considérées pour le dessin de ce sous-graphe.

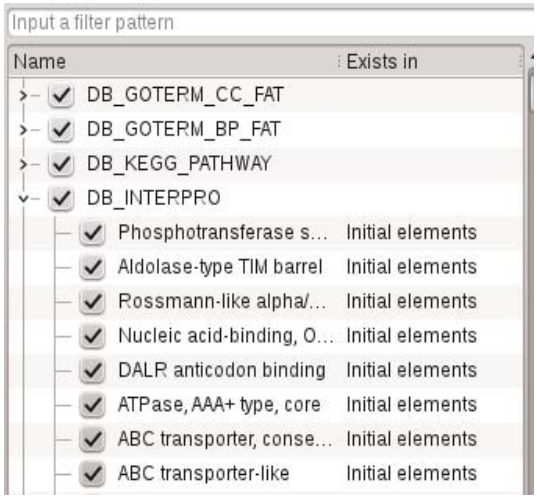


Figure 5.15 – Filtre des termes d'annotation du graphe courant de iRNA-visu. Seuls les éléments sélectionnés dans cette liste sont dessinés dans le graphe.

Filtres de sélection et de coloration Enfin, nous avons développé plusieurs autres filtres pour sélectionner et colorier les graphes selon d'autres paramètres (les filtres que nous avons directement développés sont marqués d'une étoile). Ces filtres permettent de considérer différents éléments du graphe. Les filtres *Position** et *Cardinality** sélectionnent ainsi le sous-graphe des sRNAs et des mRNAs respectivement selon la position de l'interaction ou le nombre d'interactions qu'ils effectuent à cette position. Nous avons aussi réalisé des filtres basés sur les paramètres du graphe. Les filtres *Name** et *Neighborhood** sélectionnent respectivement les éléments qui interagissent avec un élément d'intérêt (éléments à une distance 1), ainsi que les éléments distants interagissant avec ces derniers (éléments à une distance > 1) et les éléments impliqués dans un certain nombre d'interactions communes entre deux éléments d'intérêt. Il est également possible de colorier le graphe en fonction du degré, de l'excentricité (distance par rapport aux autres noeuds) [Hage et Harary 1995] ou de la centralité [Freeman 1977] des nœuds dans le graphe.
D'autres modules permettent de filtrer le sous-graphe obtenu par une ou plusieurs bases de données (filtre *Database*), de rechercher spécifiquement un sRNA ou un mRNA (filtre *Search**) ou encore de sélectionner tous le sous-graphe des interactions entre des éléments (filtre *Interaction**). Enfin, plusieurs filtres ont été réalisés pour sélectionner un sous-graphe selon la p-valeur des interactions ou encore l'indice de Jaccard modifié, tels que les filtres *Numeric* et *Color pValue*.
Le système développé reposant sur l'architecture de Tulip, il est possible pour les utilisateurs experts d'ajouter de nouveaux filtres en Python.

5.3 Applications

Nous présentons dans cette section le comparatif des logiciels de prédiction réalisé à partir des données du jeu de test. Nous appliquerons dans un second temps l'approche de iRNA avec les paramètres déterminés lors du comparatif pour prédire les cibles des sRNAs de notre jeu de données de *E. coli*. À cette étape, nous discuterons au travers de deux exemples d'applications de l'apport de iRNA pour cette analyse. Enfin nous conclurons sur les perspectives à venir pour iRNA.

5.3.1 Comparaison des différentes méthodes de prédiction

Avec iRNA, nous avons intégré 13 outils de prédictions des interactions et nous avons souhaité déterminer quel était le meilleur logiciel pour prédire les interactions de sRNAs et de mRNAs. Pour cela, nous avons réalisé un comparatif dont l'objectif est d'identifier parmi les logiciels de prédiction la méthode de prédiction la plus efficace ainsi que ses meilleurs paramètres d'utilisation. Trois éléments sont considérés pour cette étude : la prédiction des vraies- et non-interactions, la prédiction de la zone d'interaction et le temps d'exécution de ces logiciels.

Prédiction des vraies- et non-interactions

À partir des données de notre jeu de test, nous avons estimé la capacité des logiciels à identifier les sRNAs et les mRNAs interagissant ou non ensemble. Nous avons considéré l'e-valeur ou le résultat de score, de MFE ou d'énergie normé associés aux interactions effectivement prédites. Nous présentons ici la meilleure courbe de ROC obtenue ainsi que l'index de Youden des courbes de ROC des autres logiciels selon le type d'approche qu'ils implémentent (voir Section 4.2.4).
Les résultats obtenus (Fig. 5.16) nous indiquent que les logiciels IntaRNA et RNAup, effectuant une prédiction basée sur l'accessibilité, sont ceux qui obtiennent les meilleurs résultats avec respectivement une AUC de 88% et 83% (voir Annexe B.1 pour obtenir l'ensemble des résultats). Ils sont suivis par les logiciels guugle et ssearch qui se basent sur la séquence avec une AUC de 77% et 76% et de RNAplex qui est basé sur la thermodynamique avec une AUC de 74%. Les logiciels pairfold et RNAcofold basés sur la structure secondaire obtiennent le résultat le plus faible avec une AUC de 58% et 57%. Enfin, le logiciel Yass qui obtient l'AUC la plus élevée de cette étude avec un résultat de 92%, constitue un cas particulier de ce classement puisqu'il n'identifie que 30 interactions (parmi les 103 possibles) dont 20 constituent des vrais positifs et 10 des faux positifs. Il n'identifie ainsi que 32% des vraies interactions contrairement aux autres logiciels qui prédisent une interaction potentielle dans tous les cas. Hormis les logiciels basés sur la séquence, ce classement correspond donc au résultat également obtenu par [Busch *et al.* 2008]. Le logiciel IntaRNA, basé sur l'accessibilité, constitue ici la méthodes la plus efficace pour identifier dans notre jeu de test les vraies- et non-interactions et le logiciel IntaRNA_1 (avec le premier ensemble de paramètres) est celui qui ressort premier de cette analyse.

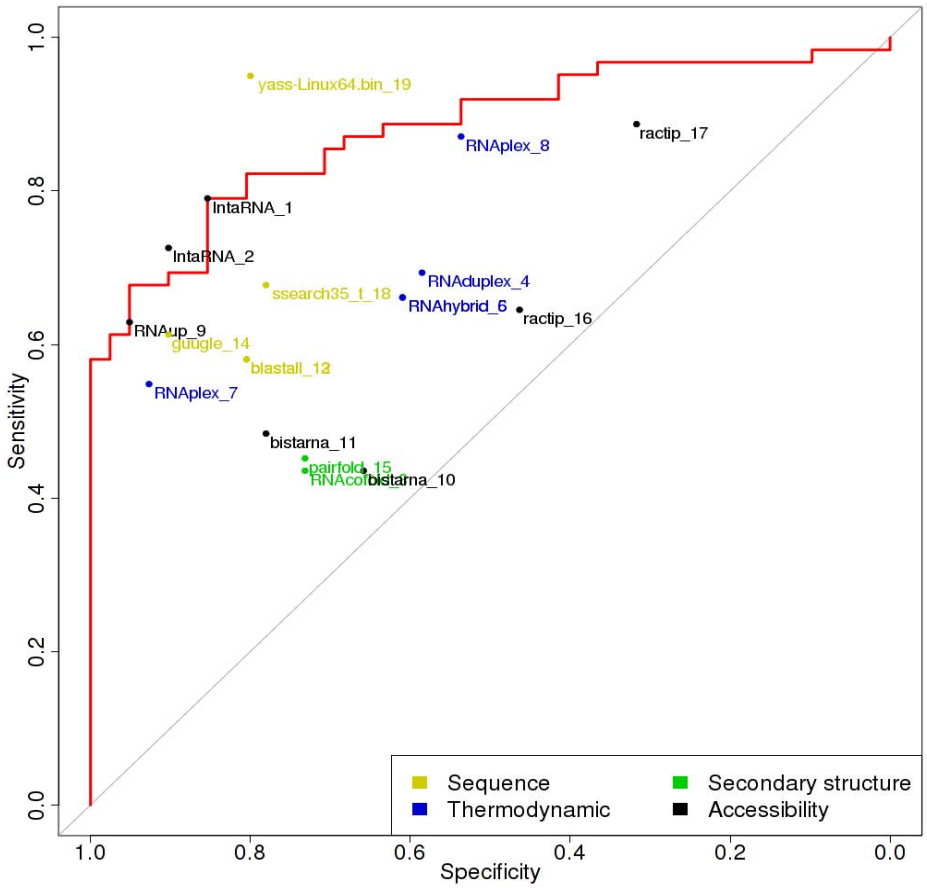
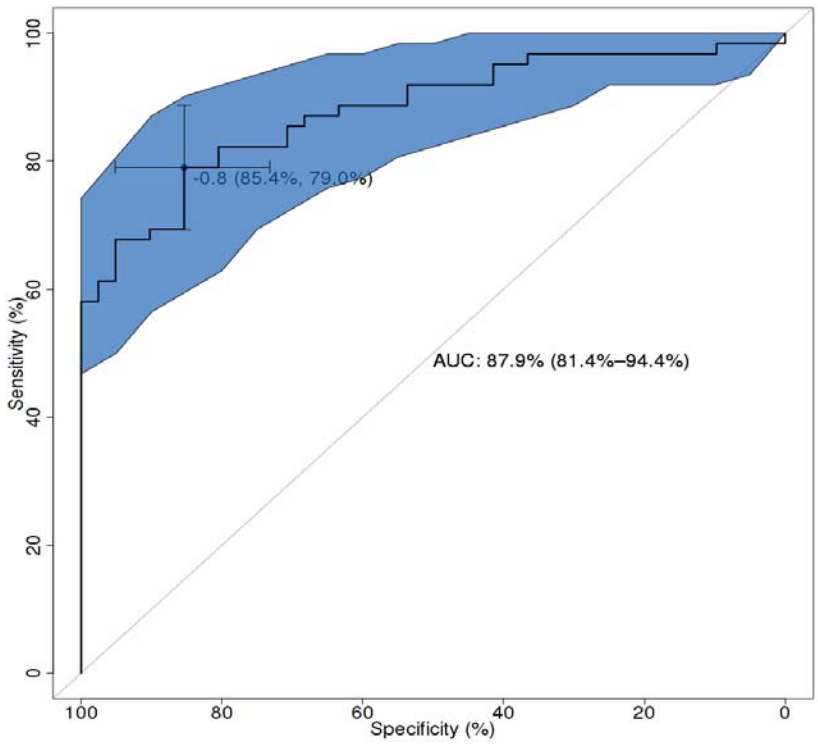


FIGURE 5.16 – Comparaison de la capacité des logiciels à prédire des vraies- et non-interactions par courbe de ROC.
Nous avons calculé la courbe de ROC de 13 logiciels selon différents paramètres (pour un total de 19 lancements) pour notre jeu de données de test à l'aide du package R : pROC. Nous avons représenté en rouge la courbe de ROC du logiciel ayant obtenu l'AUC la plus élevée : IntaRNA_1 et par un point correspondant à l'index de Youden les autres logiciels, dont la couleur dépend du type d'approche implémentée. Le profil de l'ensemble des courbes de ROC est disponible en annexe (Annexe B.2)

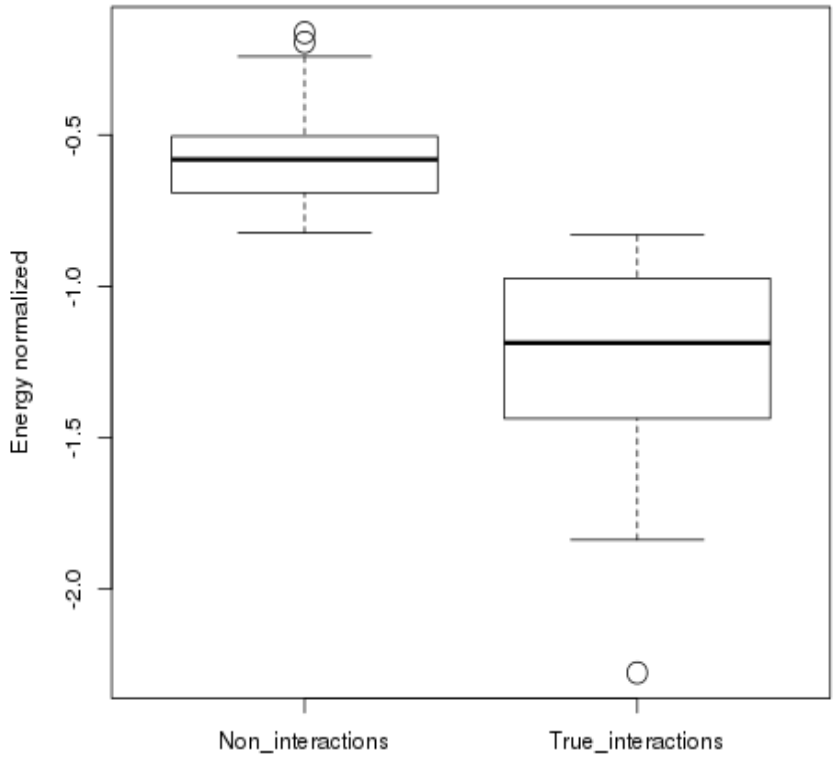
Nous avons ensuite estimé les paramètres de la prédiction de IntaRNA_1 en considérant l'intervalle de confiance associé à son AUC, ainsi que la confiance associée à son index de Youden (FIG. 5.17). À un risque α de 5%, l'AUC de IntaRNA_1 est comprise entre 81% et 94%, ce qui correspond à une classification très satisfaisante des vraies- et non-interactions (voir Section 5.2.2). La valeur de l'index de Youden calculée pour la courbe de ROC de IntaRNA_1 correspond à un seuil d'énergie normée de -0.8 avec une spécificité de 85.4% et une sensibilité de 79.0%. À ce seuil, nous avons estimé par test ANOVA la séparation entre les non-interactions et les vraies-interactions (FIG. 5.17b). Deux hypothèses sont considérées :

- H_0 : Les moyennes des scores obtenus pour les vraies- et non-interactions définies par ce seuil sont égales.
- H_1 : Les deux moyennes sont différentes.

À un risque $\alpha = 5\%$, il est apparu que les moyennes des scores obtenus pour les vraies- et non-interactions sont très significativement différentes avec une $p-valeur = 2.2e^{-16} < 0.001$.



(a) Courbe de ROC de IntaRNA_1



(b) Distribution des résultats des vraies et non interactions pour IntaRNA_1

FIGURE 5.17 – Estimation des paramètres de la prédiction des interactions par IntaRNA_1. *Nous avons calculé l'AUC de IntaRNA_1 à partir de la courbe de ROC obtenue pour le jeu de données de test. On peut observer que celle-ci est comprise entre 81.4% et 94.4% à un risque $\alpha = 5\%$ (a). Pour cette courbe de ROC, l'index de Youden de IntaRNA_1 correspond à un seuil d'énergie normé de -0.8 avec une spécificité 85.4% et la sensibilité de 79.0%. À ce seuil, il apparaît que les distributions des scores pour les vraies- et non-interactions sont significativement différentes avec une p-valeur au test d'ANOVA de $2.2e^{-16} < 0.001$ (b).*

Prédiction de la zone d'interaction

Dans un second temps, nous nous sommes intéressés à la capacité des logiciels à identifier la zone d'interaction sur le sRNA et le mRNA. Pour cela, nous avons estimé la sensibilité et le PPV moyens (voir Eq. 5.2) des interactions prédites pour les vraies interactions de notre jeu de test (Fig. 5.18) (l'ensemble des résultats est disponible Annexe B.3).
On peut observer que le logiciel IntaRNA_2 présente le meilleur compromis de l'étude avec un PPV moyen de 0.75 qui constitue le score le plus élevé du comparatif, mais également une sensibilité relativement élevée de 0.58. Le logiciel IntaRNA_1 obtient ici un résultat légèrement inférieur avec un PPV moyen de 0.68 et une sensibilité de 0.59. On peut également observer que le logiciel RNAduplex qui est basé sur la thermodynamique présente la plus forte sensibilité avec un score de 0.95. Ce résultat est dû à la longueur importante des interactions prédites, avec une longueur d'interaction moyenne de 141 nucléotides (Min 65 - Max 219) contre 18 nucléotides (Min 7 - Max 66) pour IntaRNA_2 et 20 nucléotides (Min 7 - Max 85) pour IntaRNA_1.

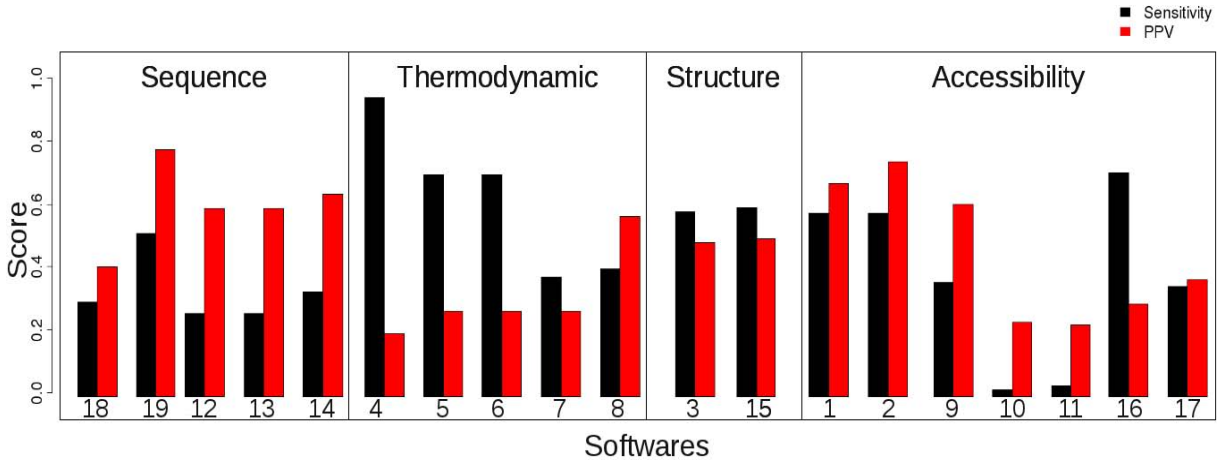


Figure 5.18 – Comparaison de la capacité des logiciels à prédire la zone effective de l'interaction.
Nous avons calculé la sensibilité moyenne (en noir) et le ppv moyen (en rouge) des logiciels pour la prédiction de la zone d'interaction sur le sRNA et le mRNA des vraies interactions du jeu de données de test.
Abréviations : 1.IntaRNA_1 ; 2.IntaRNA_2 ; 3.RNAcofold_3 ; 4.RNAduplex_4 ; 5.RNAhybrid_5 ; 6.RNAhybrid_6, 7.RNAplex_7 ; 8.RNAplex_8 ; 9.RNAup_9 ; 10.bistarna_10 ; 11.bistarna_11 ; 12.blastall_12 ; 13.blastall_13 ; 14.guugle_14 ; 15.pairfold_15 ; 16.ractip_16 ; 17.ractip_17 ; 18.ssearch_t_18 ; 19.Yass-Linux64.bin_19.

Temps d'exécution

Enfin, nous avons considéré le temps d'exécution des logiciels pour prédire les 103 interactions de l'étude sur une machine possédant 8 cœurs de calculs. Cette étape a été réalisé pour les 19 lancements en mode distribué où tous les cœurs de calculs sont exploités, et en mode séquentiel où un seul cœur est exploité.
Il apparut lors de ces tests que le mode distribué de iRNA était près de 20% plus rapide que le mode séquentiel en prenant 4 min 38 sec contre 23 min 31 sec dans le second cas. Nous avons donc considéré le temps de calcul de chacun des logiciels dans le mode distribué. Les résultats

obtenus pour le logiciel IntaRNA (Fig. 5.19) nous indiquent que le premier jeu de paramètres est légèrement plus avantageux que le second (variation de $\Delta 1$ *sec*). Cette courte avance est dû à la restriction de la fenêtre considérée pour prédire les interactions à 140 nucléotides, comme proposée par ses concepteurs [Busch *et al.* 2008].

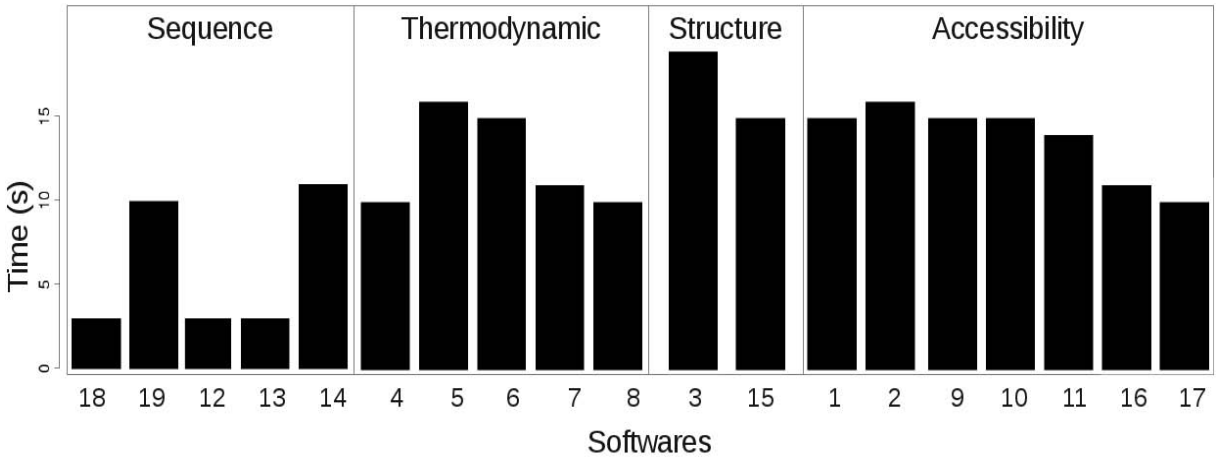


Figure 5.19 – Temps d'exécution des logiciels pour la prédiction des interactions des 103 interactions du jeu de données de test. Les prédictions ont été réalisées sur une machine comportant 8 cœurs de calcul.
Abréviations : 1.IntaRNA_1 ; 2.IntaRNA_2 ; 3.RNAcofold_3 ; 4.RNAduplex_4 ; 5.RNAhybrid_5 ; 6.RNAhybrid_6, 7.RNAplex_7 ; 8.RNAplex_8 ; 9.RNAup_9 ; 10.bistarna_10 ; 11.bistarna_11 ; 12.blastall_12 ; 13.blastall_13 ; 14.guugle_14 ; 15.pairfold_15 ; 16.ractip_16 ; 17.ractip_17 ; 18.ssearch_t_18 ; 19.Yass-Linux64.bin_19.

Ainsi, à l'issue de cette étude, deux logiciels obtiennent les meilleurs résultats pour nos données, que sont le logiciel IntaRNA avec la restriction de la fenêtre (IntaRNA_1) ou sans (IntaRNA_2). Ces logiciels présentent une sensibilité et une précision de prédiction des vraies et non interactions très satisfaisante pour notre jeu de test. Ils permettent également de prédire la zone d'interaction de manière satisfaisante. Les temps de calculs de ces logiciels dont la complexité est importante (avec par exemple une complexité $O(n^2m^2)$ pour IntaRNA_2, avec n et m correspondant respectivement aux longueurs des séquences du sRNA et mRNA), se sont également révélés relativement courts en mode distribué. La distribution des calculs par iRNA a un impact important sur le temps de calcul de ces interactions, ce qui nous permet d'envisager le calcul de jeux de données de tailles plus importantes. Compte-tenu de notre objectif de prédiction d'identification des cibles des sRNAs et de la variation observée dans chaque cas, nous privilégions IntaRNA_1 qui sera utilisé pour les prédictions pour le jeu de données de *E. coli*.

5.3.2 Application du pipeline d'analyse iRNA au jeu de données de *E. coli*

Afin de prédire les cibles des sRNAs du jeu de données de *E. coli*, nous avons utilisé iRNA sur la ferme de calcul Avakas [1]. La prédiction se divise en trois étapes (Fig. 5.20). Premièrement, les interactions des 261 sRNAs avec les 4142 mRNAs de *E. coli* sont calculées par IntaRNA_1. Cette prédiction est distribuée par iRNA sur 288 processeurs, pour un temps de prédiction correspondant à 0.8 heures. Les prédictions ainsi obtenues sont traitées pour ne

1. http ://www.mcia.univ-bordeaux.fr/

sélectionner que les interactions dont l'énergie normée est inférieure au seuil de -0.8 déterminé précédemment. Nous considérons également à cette étape les différentes caractéristiques des interactions prédites. Cette analyse prend 0.17 heures sur 12 processeurs. Enfin, nous considérons l'enrichissement des cibles prédites de chaque sRNA grâce au web-service de DAVID qui est interfacé avec iRNA. Cette étape prend 1.3 heures. Les données ainsi traitées sont ensuite représentées sous la forme d'un graphe dans iRNA-visu. L'ensemble de ces opérations est ainsi réalisé en 2.27 heures.

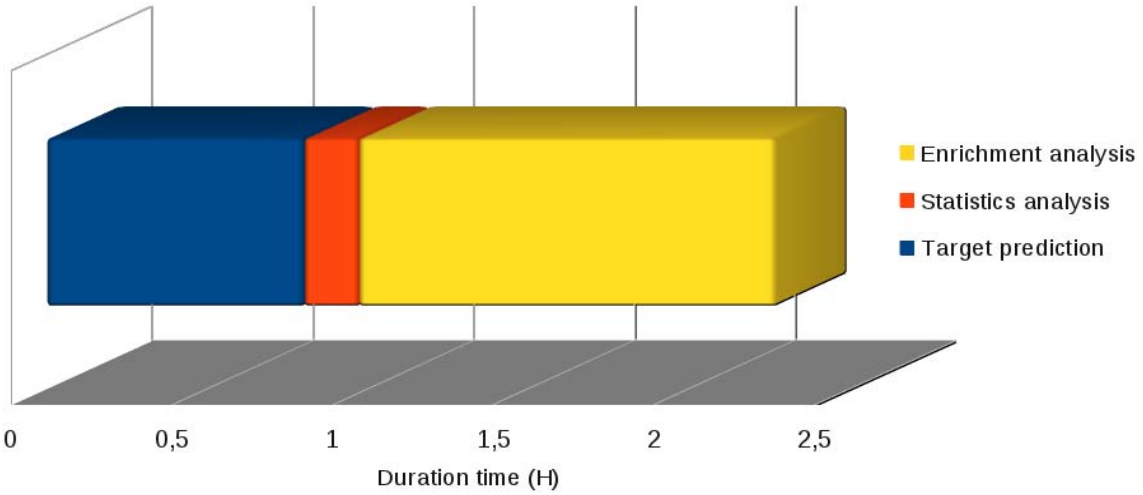


FIGURE 5.20 – Répartition du temps d'exécution pour chaque étape de la prédiction des cibles de *E. coli* par iRNA.
Les temps de calculs présentés correspondent au temps en heure nécessaire à la prédiction des cibles (en bleu), à l'analyse statistique des données (en orange) et l'analyse par enrichissement des cibles de chaque sRNA (en jaune).

Le graphe des interactions prédites de *E. coli* obtenu comporte 199461 interactions (contre 1081062 possibles) (FIG. 5.21). La longueur moyenne des interactions prédites est de 15 nucléotides (Min 6 - Max 140). Les 261 sRNAs ont entre 13 et 2913 cibles potentielles à l'issue de la première étape et partagent en moyenne 9.4% de leurs cibles (Min 0% - Max 57.8%). L'ensemble des caractéristiques de ce graphe est disponible en annexe (Annexe B.4).
Afin de présenter l'apport de iRNA, nous considérons à présent deux exemples de prédiction au sein de ce graphe pour le sRNA SgrS et la recherche des sRNAs impliqués dans la sensibilité au quorum chez *E. coli*.

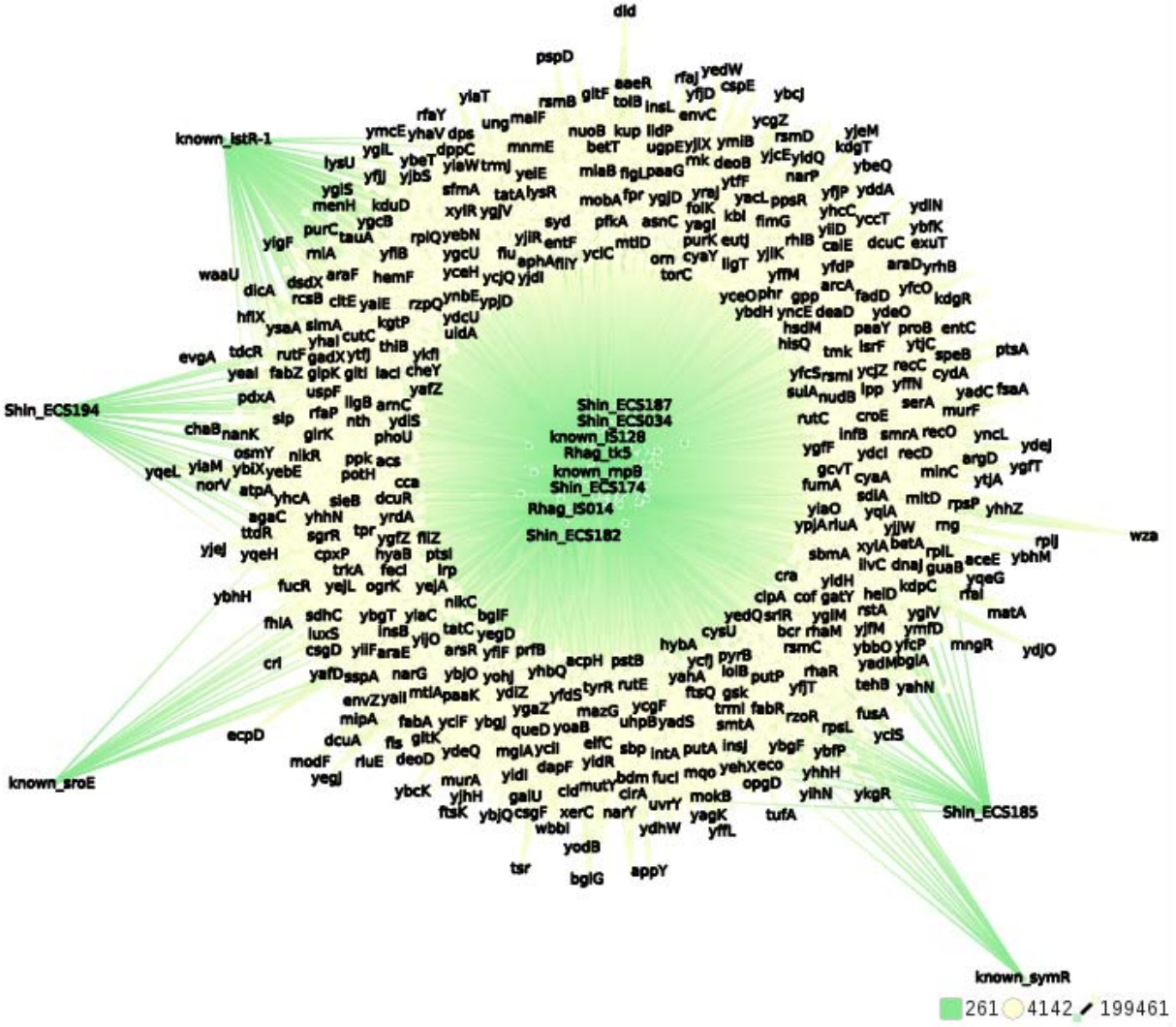
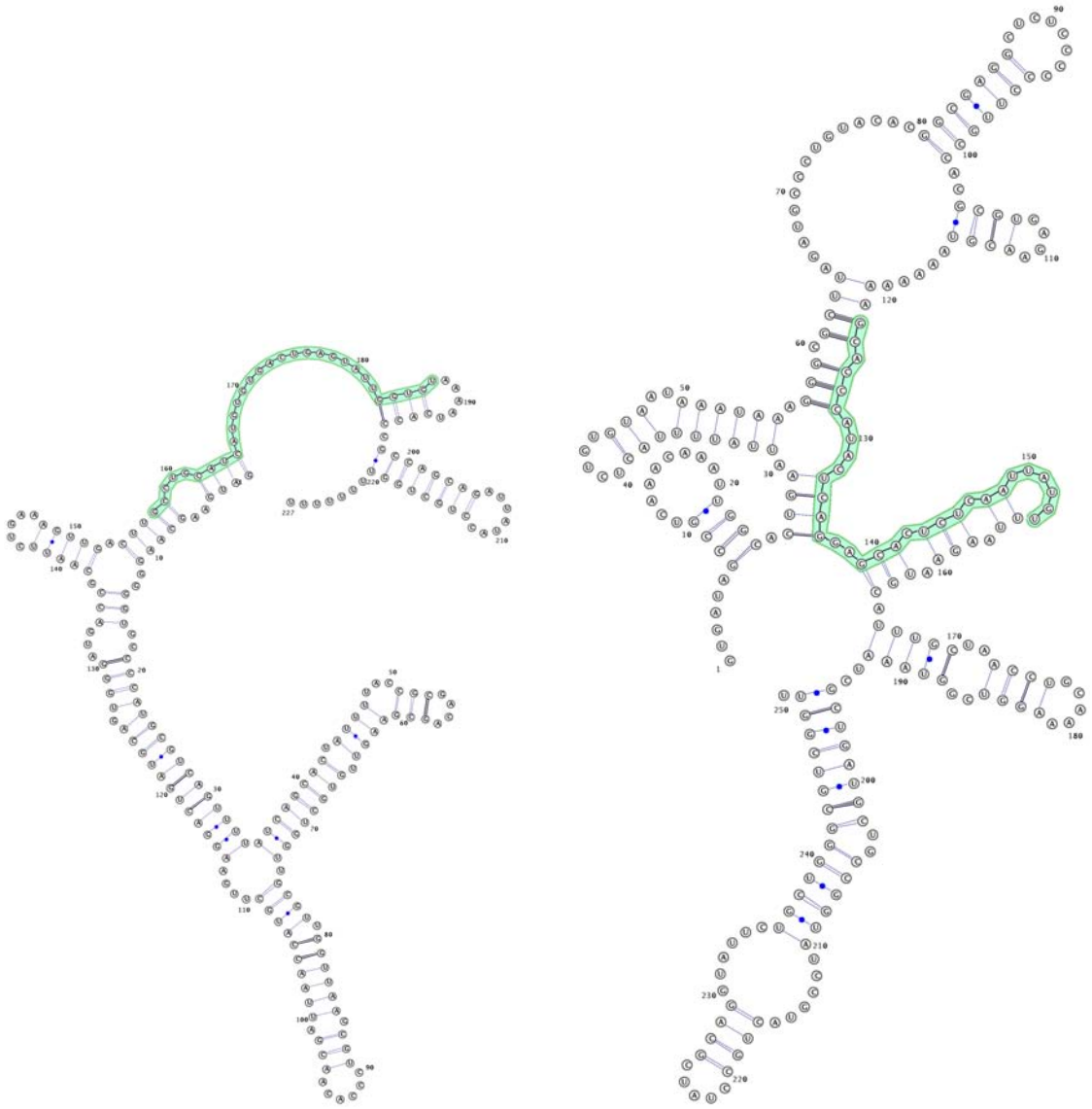


FIGURE 5.21 – Graphe initial des interactions prédites par iRNA pour le jeu de données de *E. coli*.
Les sRNAs sont identifiés par différents sigles selon qu'ils soient connus ("Known_"), identifiés par l'étude de [Shinhara et al. *2011] ("Shin_") ou identifiés par l'étude de [Raghavan* et al. *2011] ("Rhag_") (voir Section 5.2.1).*

Prédiction des cibles de SgrS

Nous considérons en premier lieu le cas des interactions du sRNA SgrS. Cet ARN de 227 nucléotides de long a été identifié par [Zhang *et al.* 2003] grâce à sa capacité à se lier à la protéine Hfq. Il est exprimé lorsqu'il y a un stress métabolique suite à une accumulation de glucose-6-phosphate ou de α-methyl-glucoside (αMG) [Vanderpool et Gottesman 2004; Kawamoto *et al.* 2005]. Sa transcription est activée par le facteur de transcription SgrR [Vanderpool et Gottesman 2004] et il présente également la particularité de coder pour un petit peptide : sgrT [Vanderpool *et al.* 2011], qui inhibe l'absorption de glucose par la cellule [Wadler et Vanderpool 2007]. Le sRNA SgrS réprime le mRNA de ptsG qui code pour la protéine $EIICB^{Glc}$, qui est la principale protéine responsable du transport de ces oses dans la bactérie. Il affecte pour cela la stabilité de ptsG et inhibe sa traduction en interagissant au

niveau du RBS. Pour cette interaction, 30 nucléotides sont requis dans la région 3' de SgrS, ainsi que le recrutement de la RNAse E (Fig. 5.22).
Afin de déterminer si d'autres cibles potentielles pouvaient être prédites pour le sRNA SgrS, nous avons recherché avec iRNA_visu si cette interaction était retrouvée dans les prédictions, et si elle partageait des caractéristiques communes avec d'autres mRNAs pour lesquels IntaRNA_1 aurait également prédit une interaction avec SgrS.



(a) sRNA SgrS (b) mRNA ptsG

Figure 5.22 – Structure secondaire prédite du sRNA SgrS et du mRNA ptsG obtenu sur sRNATarBase [Cao *et al.* 2010].
La structure secondaire des deux ARN présentée a été prédite par le programme RNAfold [Hofacker 2003] et visualisée à l'aide de VARNA [Darty et al. *2009]. La structure du mRNA ici prédite est donnée pour la séquence correspondant de -150 à +100 nucléotides autour du site d'initiation. Les sites d'interaction des deux ARN sont indiqués en vert aux positions 157 à 187 pour le sRNA et de -28 à 4 pour le mRNA d'après les données de [Kawamoto* et al. *2006].*

Nous avons ici considéré les prédictions sur la base des données des interactions et de l'enrichissement des séquences des deux premières étapes de iRNA. L'analyse a consisté en plusieurs étapes de filtres du graphe initial de *E. coli* qui sont toutes réalisées au sein de iRNA_visu (Fig. 5.23).

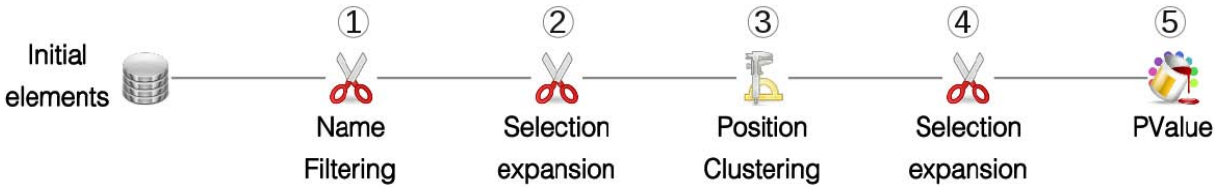


Figure 5.23 – Les différentes étapes de l'analyse des interactions de SgrS avec iRNA_visu.
Ce graphe, automatiquement dessiné lors de l'utilisation de iRNA_visu, présente l'ensemble des étapes d'une analyse réalisée. Chaque étape correspond à un graphe ou au sous-graphe obtenu après l'application d'un filtre, d'un calcul ou d'une coloration sur le graphe précédent. Six étapes ont ici été réalisées pour le filtrage du premier graphe selon le nom du sRNA, la sélection de la base de données, le regroupement selon les positions d'interaction et la coloration selon la p-valeur des interactions.

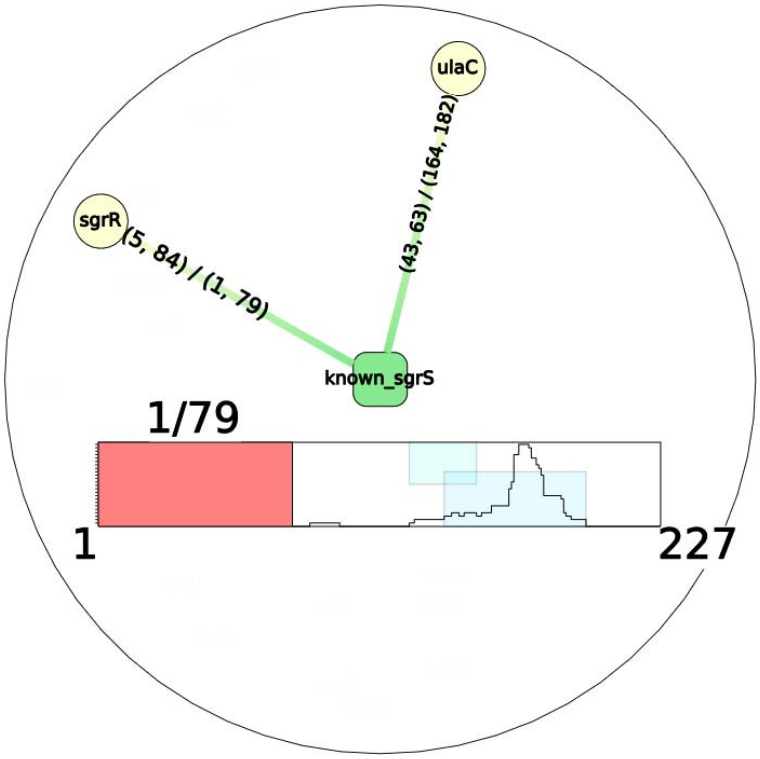
Parmi les 812 cibles prédites par IntaRNA_1 pour SgrS (étape 1 Fig. 5.23), il apparaît que l'interaction avec le mRNA ptsG est bien retrouvée aux positions de l'interaction réelle (Fig. 5.24).

	Predicted position	Real binding position*
sRNA SgrS	168..188	157..187
mRNA ptsG	122..143	123..147 (-27..-3)

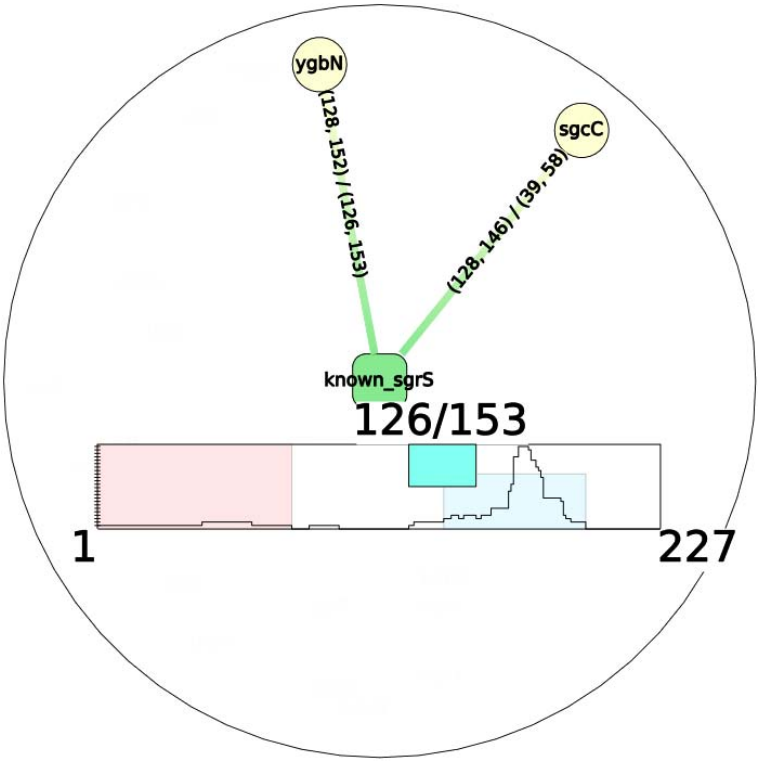
Figure 5.24 – Comparaison des zones d'interaction prédites par iRNA par rapport aux données expérimentales de [Kawamoto *et al.* 2006].
Les prédictions de position ont été identifiées par mutation spécifique des nucléotides de SrgS. Les positions d'interaction sur le mRNA sont ici reportées par rapport à la longueur des séquences : $\{-150..+50\}$.

Cette interaction est également enrichie dans la base de données des fonctions métaboliques de la *Gene Ontology* (GO_MF) pour l'annotation *"sugar transmembrane activity"* et dans la base de données KEGG pour l'annotation *"Phosphotransferase system"* (PTS). Nous avons donc filtré les cibles de SgrS enrichies par la GO_MF et le KEGG et sélectionné celles qui étaient marquées par ces annotations (étape 2). Nous avons ainsi obtenu pour la GO_MF et le KEGG un total de 29 cibles potentielles (Fig. 5.25), pour lesquelles nous avons étudié la localisation des interactions sur le sRNA grâce à l'algorithme de groupement des positions (étape 3). L'idée de cette analyse est d'identifier parmi les autres interactions prédites pour SgrS, celles qui se produisent à la même position que ptsG sur le sRNA. Cette caractéristique est en effet observée chez plusieurs sRNAs régulant plusieurs cibles (voir Section 4.2.2).
Nous retrouvons ainsi trois groupements comprenant pour la première position (1..79) deux cibles, ulaC et sgrR, pour la seconde position (126..153) deux cibles, ygbN et sgcC, et pour la troisième position (140..197) 24 cibles potentielles dont le mRNA ptsG. Ce dernier groupe constitue un ensemble très intéressant, car l'interaction des mRNAs ptsL (synonyme de manX) et manY avec SgrS a également été identifiée expérimentalement [Rice et Vanderpool 2011]. Parmi ces cibles, on constate également que seules 5 cibles dont ptsG et manY effectuent une interaction au niveau du RBS (dans région $-50..1$) : npr, yagG, et setC. Ce dernier présente

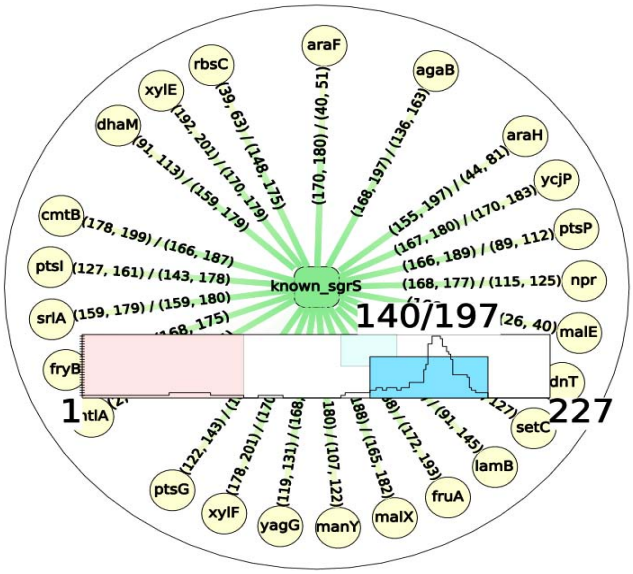
par ailleurs une forte homologie avec le gène setA, qui joue un rôle sur l'efflux de sucre dans la cellule et est co-exprimé avec SgrS [Sun et Vanderpool 2011].
Pour aller plus loin dans cette étude, nous avons enfin considéré l'énergie d'hybridation normée des cibles prédites pour le second groupe de positions (étape 4-5) (Fig. 5.26). Il apparaît que le meilleur résultat d'appariement avec SgrS est obtenu par le mRNA cmtB avec une énergie de -1.76, suivie du mRNA ptsG avec -1.64. Ces interactions impliquent 21 nucléotides dans chaque cas et semblent donc particulièrement favorables pour le couple SgrS-cmtB.



(a) Cibles interagissant au niveau du 1er groupe de positions.



(b) Cibles interagissant au niveau du 2e groupe de positions.



(c) Cibles interagissant au niveau du 3e groupe de positions.

FIGURE 5.25 – Sous-graphes des interactions de SgrS prédit à l'étape 3.
Ce sous-graphe des interactions de SgrS comporte 29 cibles potentielles. Trois groupements d'interaction sont identifiés par l'algorithme de classification des positions. Le premier et le second groupement en (1..79) et (126..153) comporte deux interactions dans chaque cas de SgrS, avec les mRNAs sgrR et ulaC, et ygbN et sgcC. Le troisième groupement en (140..197) comporte 24 interactions de SgrS, parmi lesquelles figure l'interaction avec ptsG qui est vérifiée expérimentalement. L'interaction de SgrS avec le mRNA bglF n'est ici retrouvée dans aucun groupement.

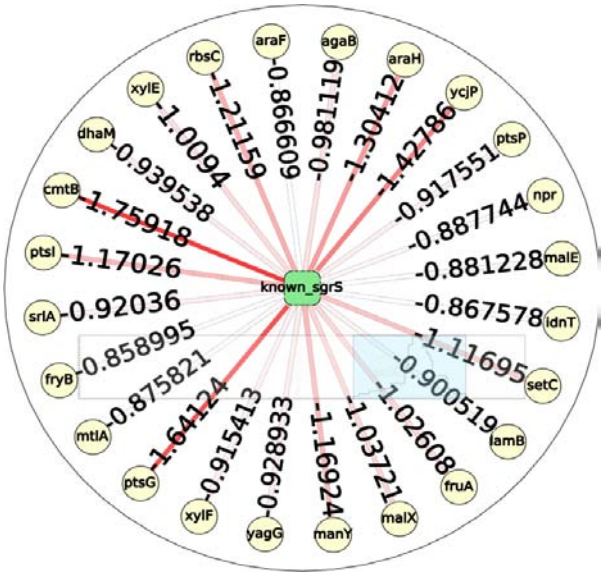


FIGURE 5.26 – Sous-graphe des interactions de SgrS prédites à l'étape 5.
Pour le groupement de positions 3 identifié à l'étape 3 (FIG. 5.25), nous avons colorié les arêtes selon l'énergie normée de chaque interaction. Les valeurs d'énergie les plus basses sont ici les plus favorables et représentées par la couleur rouge foncé, par rapport aux interactions d'énergie plus élevées en blanc.

Ainsi, à l'issue de cette première prédiction, iRNA nous a permis de retrouver en quelques étapes plusieurs cibles démontrées expérimentalement pour le sRNA SgrS, mais également d'identifier plusieurs autres cibles potentielles. Les résultats que nous avons obtenus, indiquent un possible effet de SgrS sur le transport de plusieurs oses, tels que l'arabinose avec araF et araH, le fructose avec fruA et fryB, le maltose avec cmtB, lamB, malX et malE, le xylose avec xylE et xylF et d'autres systèmes de régulation du transport du sucre avec ptsL et ptsG, ce qui est également observé pour SgrS chez *Salmonella typhimurium* [Papenfort *et al.* 2012]. L'analyse réalisée prédit aussi une région en 140..197 permettant d'effectuer ces multiples interactions.
Le pipeline iRNA présente donc un potentiel important pour prédire les interactions de sRNAs interagissant sur plusieurs cibles. Il est à présent intéressant de déterminer si d'autres motifs de régulation peuvent être prédits par iRNA. Nous nous sommes pour cela intéressés au mécanisme de la sensibilité au quorum chez *E. coli*.

Cas d'étude : Détection du quorum

La sensibilité au quorum est un mécanisme qui a été découvert pour la première fois chez *Vibrio fisheri* [Ng et Bassler 2009]. Il constitue un système de communication entre les bactéries basé sur la production d'acyl-homosérine lactone. Cette molécule auto-inductrice (AI) est utilisée par les bactéries Gram négatives pour synchroniser l'expression de leurs gènes selon la densité de la population [Vendeville *et al.* 2005; Williams *et al.* 2007]. Elle est ainsi responsable de différentes caractéristiques chez ces bactéries, telles que la virulence [Winzer *et al.* 2002], la mobilité [Atkinson *et al.* 2006; Daniels *et al.* 2004], la bioluminescence [Anetzberger *et al.* 2009] ou encore la formation de biofilm permettant aux cellules d'adhérer à la surface sur laquelle elles reposent et entre elles [González Barrios *et al.* 2006].
Chez la bactérie *Vibrio harveyi*, il a été montré que ce processus implique plusieurs sRNAs [Shao et Bassler 2012] (Fig. 5.27). En effet, à faible densité, la concentration des AIs est insuffisante pour activer le système. Les récepteurs membranaires luxN, cqsS et luxQ agissent comme des kinases et transfèrent un phosphate à luxO via luxU. La protéine phospho-luxO ainsi activée induit l'expression des sRNAs Qrr1,2,3,4, qui ont pour effet d'inhiber la traduction du régulateur de transcription luxR en interagissant au niveau du RBS. À forte densité, les récepteurs membranaires agissent en revanche comme des phosphatases qui engendrent la déphosphorylation de luxO [Freeman et Bassler 1999], qui n'active plus l'expression des sRNAs Qrr1,2,3,4. Le gène luxR est ainsi traduit et active l'opéron *lux*. L'expression des différents gènes contrôlés par cet opéron est ainsi activée, permettant notamment la bioluminescence de ces bactéries [Swartzman *et al.* 1992].

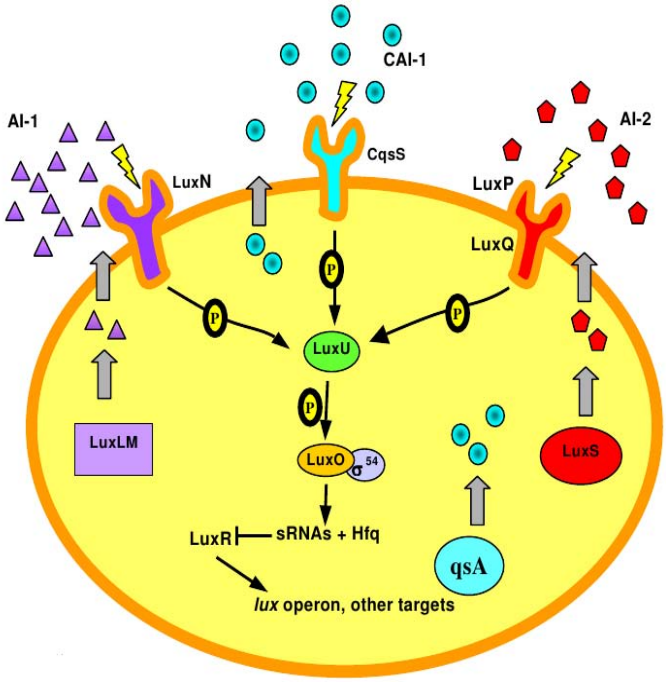


FIGURE 5.27 – Mécanisme de fonctionnement de la sensibilité au quorum chez *Vibrio harveyi* [Wang 2004].
La sensibilité au quorum de Vibrio harveyi repose sur trois molécules d'AIs : AI-1, CAI-1 et AI-2 qui sont synthétisées respectivement par luxL-M, qsA et luxS et détectées par les protéines membranaires luxN,cqsS et luxQ [Henke et Bassler 2004].

Chez *E. coli* K12 MG1655, la sensibilité au quorum permet notamment le contrôle de la formation du biofilm [Sung *et al.* 2006; Wood *et al.* 2006]. Ce système repose, comme pour *Vibrio Harveyi*, sur la synthèse de la molécule AI-2 par luxS, mais il ne comprend pas les autres éléments identifiés chez *Vibrio* [Surette *et al.* 1999]. L'implication d'autres gènes a cependant pu être identifiée, tels que le gène sdiA qui constitue un homologue à luxR impliqué dans l'activation de l'expression de plusieurs gènes notamment impliqués dans la formation de biofilm [Dyszel *et al.* 2010; Yao *et al.* 2006], les gènes rpoS et bolA qui interviennent dans la réponse général au stress ont aussi impliqué dans la formation de biofilm [Adnan *et al.* 2010], et les gènes tqsA et mqsR intervenant aussi tous deux dans le contrôle de la formation de biofilm [Rettner et Saier 2010; Herzberg *et al.* 2006; González Barrios *et al.* 2006]. Le sRNA CyaR a également été identifié comme agissant sur luxS chez *E. coli* [De Lay et Gottesman 2009]. Toutefois, l'ensemble des mécanismes impliqués dans la sensibilité au quorum chez *E. coli* n'ont pas encore été complètement identifiés.
Afin de déterminer si d'autres sRNAs et cibles potentielles pouvaient être prédites, nous avons recherché parmi les éléments connectés au mRNA luxS si des interactions connues étaient retrouvées dans les prédictions en considérant les paramètres des interactions mais aussi en analysant les paramètres du graphe par la recherche des éléments les plus fortement connectés.

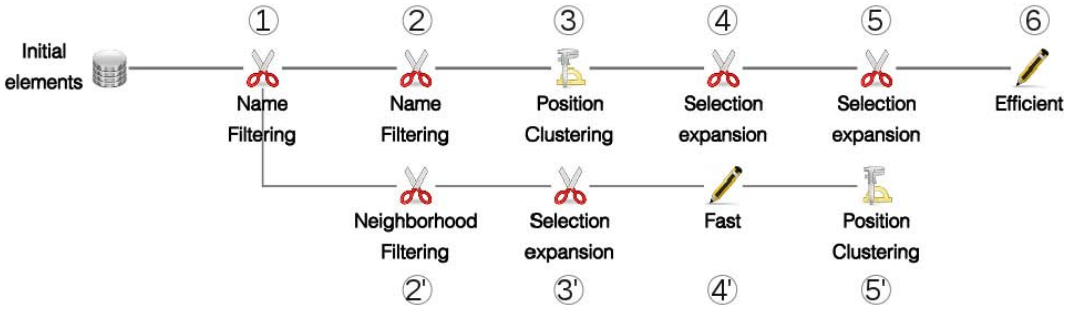


FIGURE 5.28 – Les différentes étapes de l'analyse de la sensibilité au quorum chez *E. coli*.

Pour cela, nous avons sélectionné tout d'abord les sRNAs ciblant luxS, ainsi que tous les éléments qui leur sont incidents (étape 1). On obtient ainsi un sous-graphe comportant 33 sRNAs dont 13 sRNAs connus, 16 sRNAs issus des données de [Shinhara *et al.* 2011] et 4 sRNAs issus des données de [Raghavan *et al.* 2011] et 4137 autres mRNAs potentiellement ciblés par ces sRNAs. Parmi les interactions de luxS, le sRNA CyaR a bien été prédit comme étant un de ces interacteurs potentiels, au niveau du même site d'interaction identifié par les expérimentations (FIG. 5.29).

	Predicted position	Real binding position*
sRNA CyaR	41..50	35..49
mRNA luxS	138..144	138..153 (-12..3)
sRNA CyaR	34..44	31..50
mRNA yqaE	154..163	146..166 (-4..16)
sRNA CyaR	1..49	38..48
mRNA ompX	142..177	141..162 (-9..12)

FIGURE 5.29 – Comparaison des zones d'interaction prédites par iRNA par rapport aux données de [De Lay et Gottesman 2009].
Les sites des interactions ont été identifiés par prédictions et vérifiés par la mutation ponctuelle de certains nucléotides impliqués dans l'interaction [De Lay et Gottesman 2009].

Nous avons donc considéré pour la première partie de l'étude uniquement les cibles du sRNA CyaR (étape 2). Nous filtrons également ces interactions, en ne sélectionnant que celles survenant au niveau du même site d'interaction que pour luxS (étapes 3-4). On retrouve ainsi 966 cibles potentielles de CyaR parmi lesquelles figurent, avec luxS, 3 cibles des 4 autres cibles connues de CyaR, avec yqaE et ompX (FIG. 5.29) et bolA interagissant à la position $(-62.. -42)$ avec le sRNA CyaR sur la position (20..37). Nous avons ensuite cherché s'il existait à cette position des interactions avec une annotation en lien avec la sensibilité au quorum. Nous avons ainsi sélectionné les mRNAs selon leur annotation dans Uniprot :"DNA binding region :H-T-H", qui correspond à la présence dans leur protéine d'un domaine Hélice-Tour-Hélice leur permettant de se lier à l'ADN. Cette même annotation caractérise en effet la protéine de type luxR chez les bactéries. Nous retrouvons 54 cibles potentielles, parmi lesquelles figurent les mRNAs sdiA et rpoS qui interagissent respectivement en (21..30) et $(-134.. -117)$ (correspondant aux positions réelles de l'interaction) avec CyaR sur la position (22..33) et (18..34) (FIG. 5.30). On peut également remarquer la présence du mRNA de la protéine crp qui interagit avec CyaR, évoquant un possible rétro-contrôle de CyaR sur le mRNA de la protéine crp qui active sa transcription.

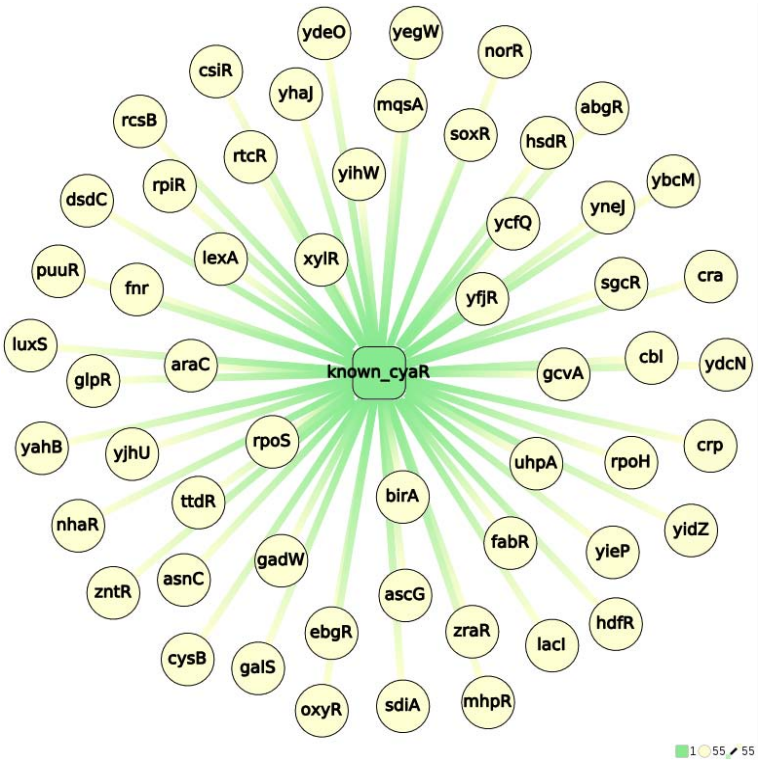


FIGURE 5.30 – Listes des cibles prédites pour le sRNA CyaR à l'étape 6.
Les cibles identifiées ont été sélectionnées selon la zone de leur interaction avec CyaR (identique à ptsG) et selon leur enrichissement pour un domaine H-T-H leur permettant de se lier à l'ADN. Parmi les 55 cibles prédites, trois cibles ont précédemment été identifiées expérimentalement : luxS, yqaE et ompX. On peut également constater la présence parmi ces cibles du mRNA de cbr, dont la protéine active CyaR.

Pour la deuxième partie de cette analyse, nous avons recherché la présence d'éléments fortement connectés à luxS parmi les interactions prédites à l'étape 1. Nous avons considéré pour cela l'adjacence entres les mRNAs prédits, en filtrant uniquement les mRNAs qui interagissent avec au moins 8 des sRNAs interagissant lors de l'étape 1 (étape 2'). Ce seuil correspond à la valeur la plus élevée où sont identifiés des éléments relatifs à la sensibilité au quorum. Le sous-graphe ainsi obtenu comporte 33 sRNAs et 2142 mRNAs. Parmi ces cibles, nous avons recherché celles qui présentaient l'annotation "*quorum sensing*" dans Swissprot et dans la base de données *biological pathway* de la *Gene Ontology* (GO_BP). Il est apparu que deux autres mRNAs, tqsA et mqsR, partagent cette annotation et sont les cibles de 8 sRNAs communs avec luxS (FIG. 5.31). Plusieurs de ces sRNAs interagissent plus spécifiquement au niveau du RBS de ces mRNAs, tels que les sRNAs ECS190 et ECS200 de l'étude [Shinhara *et al.* 2011] avec luxS, le sRNA rttR avec tqsA et les sRNAs sokA, rttR, rybA et NC021 de l'étude de [Raghavan *et al.* 2011] avec mqsR. Ces éléments présentent potentiellement les caractéristiques d'un motif DOR, ce qui rejoint la caractéristique de la sensibilité du quorum d'être sensible à plusieurs stimuli.

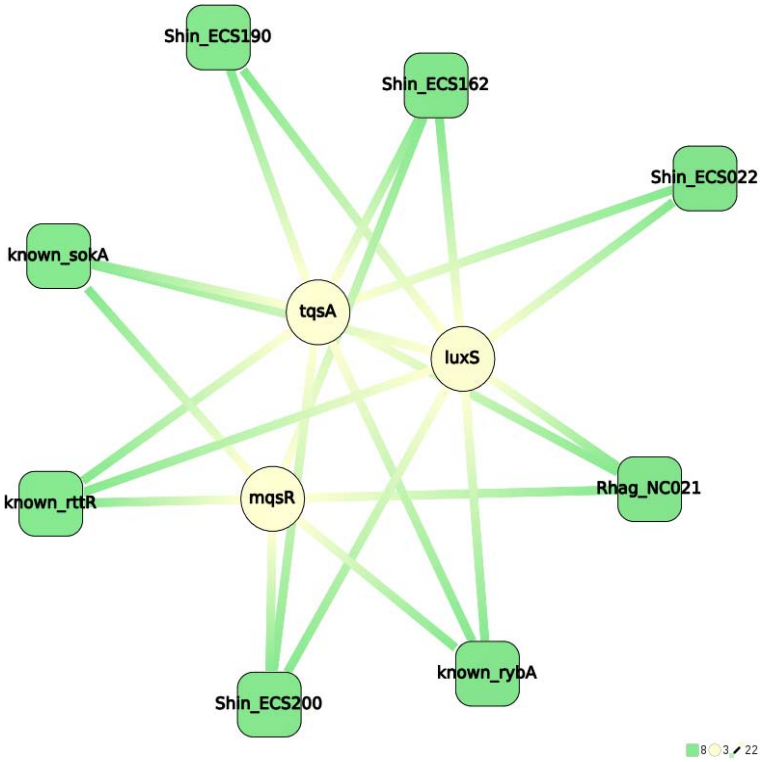


FIGURE 5.31 – Liste des éléments fortement connectés à luxS identifiés à l'étape 4'.
Nous avons recherché parmi les interactions de luxS les mRNAs qui partageaient au moins 8 interactions, et enrichi ces cibles selon leur annotation au "quorum sensing" dans swissprot et GO_MF. Deux autres mRNAs impliqués dans le contrôle de la formation de biofilm ont ainsi été identifiés : tqsA et mqsR. Huit sRNAs interagissant avec ces mRNAs ont également été identifiés dans cette analyse, appartenant au groupe des sRNAs connus (identifiés par le sigle "Known_"), de l'étude de [Shinhara et al. *2011] (identifiés par le sigle "Shin_") ou encore de l'étude de [Raghavan* et al. *2011] (identifiés par le sigle "Rhag_") (voir Section 5.2.1).*

Ainsi, à l'issue de cette seconde prédiction, iRNA nous a permis de retrouver plusieurs cibles démontrées expérimentalement pour le sRNA CyaR, qui partagent également plusieurs caractéristiques avec d'autres éléments nouvellement prédits. La recherche des éléments fortement connectés semble être également un moyen intéressant d'identifier des motifs complexes de régulation comme le motif DOR, bien que nous ne disposions pas ici d'assez de connaissances sur ces éléments pour conclure l'analyse.

5.4 Discussion et conclusion

Le travail qui a été réalisé dans cette thèse pour la prédiction des cibles des sRNAs, a consisté à définir une démarche permettant d'analyser de manière plus approfondie les données issues des méthodes de prédictions existantes. Bien que différentes études aient montré l'apport de l'analyse d'enrichissement ou de la visualisation pour la prédiction de cibles, l'intégration de l'ensemble de ces analyses en un seul système pour les sRNAs n'avait pas été réalisée jusque-là. De plus, ces méthodes ne permettent pas en l'état actuel de rechercher simultanément les cibles de plusieurs sRNAs, ce qui est pourtant nécessaire pour identifier des motifs de régulation. Nous avons donc développé une approche permettant de prendre en

compte la prédiction des interactions et le contexte des connaissances biologiques pour ces éléments, avec l'utilisation d'un système de visualisation adapté à ces données.
Pour déterminer les meilleurs paramètres des différentes étapes de l'analyse de iRNA, nous avons tout d'abord évalué l'efficacité des différentes approches de prédiction disponibles. Il est apparu lors de cette analyse que le logiciel IntaRNA permettait de prédire de manière très satisfaisante les vrais et non interactions, mais également d'identifier avec une bonne précision la zone d'interaction pour les mRNAs et les sRNAs. Nous avons ensuite considéré l'enrichissement des cibles. Parmi les multiples solutions disponibles, nous avons identifié que la base de connaissances DAVID était de par son ouverture à des programmes extérieurs et par le nombre important de bases de données disponibles la solution la plus adaptée pour l'analyse d'enrichissement des données des sRNAs. Enfin pour la visualisation des interactions, nous avons sélectionné le logiciel de visualisation de graphe Tulip, auxquels nous avons pu ajouter plusieurs extensions.
Dans un second temps, nous avons appliqué notre approche pour la recherche des cibles des sRNAs de *E. coli*. La procédure d'analyse de iRNA s'est révélée être un moyen de prédiction crédible à l'échelle d'un organisme, en permettant de considérer l'ensemble les cibles des sRNAs de *E. coli* en quelques heures. Nous avons ensuite analysé les prédictions fournies pour deux cas portant sur la recherche de nouvelles cibles du sRNA SgrS et la recherche de sRNAs et des mRNAs impliqués dans la sensibilité au quorum. Dans le premier cas, nous avons pu identifier plusieurs cibles potentielles de SgrS, partageant plusieurs caractéristiques communes avec des interactions connues de SgrS. La combinaison de la prédiction de IntaRNA à l'analyse d'enrichissement a joué ici un rôle important comme source d'informations permettant de réduire drastiquement le nombre de cibles potentielles. Dans le second cas, nous avons considéré une approche différente en partant du mRNA et en considérant des paramètres du graphe. Nous avons alors observé que plusieurs éléments prédits par iRNA pour le mRNA luxS et le sRNA SgrS étaient en accord avec les données expérimentales, mettant ainsi en avant les autres cibles partageant également ces paramètres. La visualisation a ici montré son importance pour considérer plus facilement les nombreuses données issues des précédentes étapes du pipeline iRNA, mais aussi pour la considération plus poussée des paramètres du graphe.
L'application développée a ainsi permis d'étudier la prédiction des cibles pour l'ensemble des sRNAs connus pour *E. coli*. L'approche a montré être un bon compromis en terme d'efficacité d'un point de vue du temps de calcul, mais également pour la qualité des prédictions, qui étaient en accord avec les données expérimentales disponibles. L'enrichissement et la visualisation des données ont mis en évidence l'intérêt de ce type d'approche comme sources d'information pour mieux comprendre les phénomènes complexes sous-jacents à la régulation des sRNAs. Le pipeline iRNA constitue donc un moyen crédible et efficace pour la prédiction des cibles des sRNAs à l'échelle du génome. Il sera mis prochainement à disposition de la communauté [2] et fera l'objet d'un article dans le domaine de la visualisation de part les différentes méthodes de visualisation qui ont été ici développées.
Étant donné le caractère encore préliminaire de ce travail, de nombreuses perspectives de développements peuvent être envisagées pour aller plus loin dans cette étude. Une première perspective consisterait à améliorer le système de prédiction des interactions et d'enrichissement. Le logiciel iRNA prend en charge plusieurs logiciels de prédiction et il serait intéressant de déterminer si une combinaison de certains de ces logiciels permettrait d'augmenter la sensibilité de la prédiction. Concernant l'enrichissement, DAVID propose en plus de l'analyse de

2. http ://www.cbib.u-bordeaux2.fr/irna/

l'annotation fonctionnelle des gènes, une classification des annotations enrichies dans les bases de données. L'intégration de ces données dans les graphes de iRNA_visu permettrait ainsi de rapprocher les annotations ayant la même signification et de simplifier la fouille naïve des données, en permettant de considérer des groupements sémantiques d'annotation. Une seconde perspective consisterait à vérifier expérimentalement, en collaboration avec des biologistes, les cibles prédites pour les cas d'étude. Ceci nous permettrait d'identifier le pourcentage de faux positifs subsistant dans les prédictions de iRNA, et ainsi d'améliorer les filtres pour réduire ce chiffre. Une troisième perspective consisterait à améliorer les critères de sélection. De nouveaux filtres peuvent être définis pour mieux rechercher la présence de motif de régulation en considérant la connectivité des éléments, la zone d'interaction ou encore la p-valeur des interactions. La littérature portant sur les graphes propose aussi de nombreux algorithmes pour rechercher des motifs, il serait intéressant de considérer ceux qui pourraient être appliqués à la prédiction des cibles des sRNAs. Une quatrième perspective porterait sur l'adaptation du pipeline iRNA au travers de l'éditeur de workflow Galaxy [Goecks *et al.* 2010]. Cet éditeur permet de simplifier la diffusion et l'utilisation d'un pipeline en fournissant une interface permettant à l'utilisateur d'entrer directement ces paramètres mais aussi de relier son analyse aux résultats issus d'autres pipelines. Enfin, iRNA ayant été testé pour prédire les cibles des sRNAs d'autres organismes (*Mycoplasma pneumoniae*, *Helicobacter pylori*, *Bacillus subtilis*), il serait intéressant d'analyser également ces données en collaboration avec les biologistes travaillant sur ces organismes.

Conclusion et perspectives

Dans cette thèse, nous nous sommes intéressés à la prédiction des mécanismes moléculaires survenant au niveau du métabolisme et des ARN non codants. Nous avons privilégié pour ces études une approche intégrative basée sur la modélisation des éléments biologiques et de leurs interactions sous la forme de graphes et d'automates.

Prédiction de la distribution des flux au sein d'un réseau métabolique

Nous avons développé dans un premier temps, une méthode permettant de prédire la distribution du flux au sein d'un réseau métabolique. L'objectif de cette approche était de tenir compte pour cette prédiction d'une ou de plusieurs contraintes formulées à partir des données expérimentales, sans recherche à priori de l'objectif cellulaire. Cette méthode ne devait pas requérir de données cinétiques qui ne sont généralement pas disponibles, mais permettre d'intégrer différentes propriétés d'un réseau métabolique qui peuvent être mesurées expérimentalement.
Nous avons pu atteindre ce but en deux étapes. Tout d'abord, nous avons développé un nouveau formalisme de réseau de Petri appelé *Flux Petri Net* (FPN). Ces réseaux permettent de considérer, grâce à l'introduction de la notion de poids de flux, la distribution différentielle des métabolites entre les réactions et de définir des contraintes basées sur la distribution des jetons dans le réseau. Nous avons ensuite couplé ce système à une méthode d'optimisation capable d'identifier les paramètres du FPN selon une à plusieurs contraintes formulées par l'utilisateur. L'ensemble de cette approche a ainsi été implémenté dans le logiciel Metaboflux [3].
Pour valider notre approche, nous avons prédit la distribution des flux pour le modèle du métabolisme énergétique de la forme sanguine de *T. brucei* dont le fonctionnement était connu. Nous avons ensuite considéré la distribution du flux de la forme procyclique de *T. brucei*. La formulation de ce réseau métabolique a été étudiée pour déterminer si elle était compatible avec les contraintes de ce métabolisme.
Les résultats que nous avons obtenus pour la forme sanguine de *T. brucei* ont révélé que la distribution des flux et la proportion des métabolites excrétés prédites par notre méthode étaient parfaitement en accord avec les données expérimentales de cet organisme. Nous avons donc appliqué notre approche à la forme procyclique de *T. brucei*. Les prédictions réalisées ont indiqué qu'une des contraintes formulées à partir de données préliminaires de ce métabolisme pour l'équivalence du flux entre deux branches de ce réseau n'était pas compatible avec la formulation de ce réseau. Son retrait permettait de satisfaire sans modification de la formulation du réseau métabolique les autres contraintes du système. Cette hypothèse a ensuite été validée expérimentalement permettant de retirer du modèle cette contrainte. La fonctionnalité de ce réseau métabolique a ainsi été validée d'un point de vue théorique. Nous nous sommes ensuite

3. http ://www.cbib.u-bordeaux2.fr/metaboflux/

intéressés à l'analyse de la flexibilité du réseau. Différentes études avaient mis en avant une variation de la proportion des produits excrétés de ce métabolisme. Les prédictions réalisées étaient également en accord avec cette flexibilité et ont montré une implication des enzymes maliques, également liées au stress de cet organisme, dans la flexibilité de cette distribution. Il sera ainsi intéressant de déterminer si cette relation entre le stress du trypanosome et le ratio des produits excrétés est également observée expérimentalement, ce qui confirmerait l'implication des enzymes maliques dans la flexibilité du flux.
La prise en compte des données expérimentales par notre approche a ainsi eu un impact positif sur la qualité de la prédiction de la distribution du flux. Nous avons pu observer que le chemin emprunté par ce dernier était modifié selon les contraintes considérées. Nous avons également observé cet apport par rapport à l'approche de FBA qui contrairement à Metaboflux se concentre sur la réalisation d'un objectif cellulaire. L'évaluation théorique plus précise de la distribution des flux du modèle du trypanosome pourrait être ensuite réalisée lorsque plus de données seront disponibles sur le métabolisme des trypanosomes. Ceci permettrait d'exploiter les capacités de Metaboflux pour identifier différents scénarios de distribution de flux d'un réseau métabolique et d'améliorer ainsi la compréhension du fonctionnement du métabolisme des trypanosomes.

Prédiction des cibles des sRNAs

Nous avons développé dans un second temps une méthode pour la prédiction des cibles des sRNAs. L'objectif de cette approche était de combiner à la prédiction des interactions entre les sRNAs et les mRNAs de l'organisme étudié une analyse d'enrichissement. Cette méthode doit ainsi permettre d'exploiter les connaissances dont on dispose sur les cibles des sRNAs, pour focaliser par exemple la recherche sur des processus cellulaires bien précis (certains sRNAs régulant plusieurs cibles d'une même voie de régulation). Nous nous sommes également intéressés à l'exploitation de ces prédictions par un système de visualisation des graphes. Cette solution présente l'intérêt de pouvoir considérer plus facilement les caractéristiques de ces ARN, pour mieux filtrer les cibles potentielles des sRNAs. Il s'agit enfin de disposer d'une solution logicielle suffisamment performante d'un point de vue calculatoire pour considérer parallèlement les interactions de plusieurs sRNAs d'un organisme, et être en mesure d'étudier spécifiquement la question des motifs de régulation opérés par ces ARN.
Pour déterminer quelles étaient les meilleurs méthodes pour faire cette analyse, nous avons évalué l'efficacité des différents logiciels disponibles. Nous avons débuté cette analyse par la recherche de la méthode de prédiction des interactions la plus efficace. Nous avons utilisé pour cela un jeu de données de test de vraies et de non-interactions validées expérimentalement entre des sRNAs et des mRNAs de différents organismes. Nous avons identifié par cette analyse que le logiciel IntaRNA était le logiciel le plus performant pour nos données. Celui-ci permettait de détecter de manière satisfaisante les vraies et non-interactions et prédisait de manière satisfaisante la zone d'interaction des sRNAs avec leurs cibles. Nous avons ensuite considéré les différentes solutions logicielles permettant de faire une analyse d'enrichissement. Nous avons identifié que la base de connaissances DAVID était la solution la plus adaptée à l'enrichissement des données des sRNAs en permettant de considérer plusieurs organismes et différentes bases de données pour l'enrichissement. Enfin, nous avons développé un système de visualisation des interactions en nous basant sur le logiciel de visualisation de graphe Tulip. Ce logiciel présentait plusieurs caractéristiques intéressantes pour notre analyse en intégrant un système de visualisation de très grands graphes et la possibilité de développer des plugins permettant d'adapter son utilisation au contexte de la prédiction des interactions des sRNAs.

L'ensemble de cette approche a ainsi été intégré dans le pipeline d'analyse iRNA [4].
Nous avons ensuite appliqué notre méthode pour la recherche des cibles des sRNAs de *E. coli*. Le système s'est révélé efficace d'un point de vue calculatoire en permettant de considérer la prédiction des interactions de l'ensemble des sRNAs avec tous les mRNAs de cet organisme en quelques heures. Nous avons donc pu considérer deux exemples de prédictions, pour la recherche des cibles du sRNA SgrS et la recherche des sRNAs impliqués dans la sensibilité au quorum chez *E. coli*. Les résultats obtenus pour le premier modèle d'étude nous ont permis d'étudier le motif de régulation SIM (consistant à une régulation par un sRNA de plusieurs mRNAs). Nous avons retrouvé plusieurs cibles connues de ce sRNA et proposé d'autres cibles potentielles sur la base des caractéristiques de leurs interactions qu'elles partageaient avec d'autres cibles connues. Pour le second cas d'étude, nous nous sommes intéressés à la recherche d'un motif de régulation DOR (impliquant une réponse de la cellule à plusieurs stimuli). Nous avons étudié les interactions du mRNA luxS, qui est impliqué dans le contrôle de ce processus. Nous avons retrouvé une interaction déjà identifiée pour ce mRNA avec le sRNA CyaR et identifié grâce aux critères topologiques du graphe de nouvelles cibles candidates, ainsi que d'autres sRNAs potentiellement impliqués dans ce processus.
Le système que nous avons mis en place s'est ainsi révélé pertinent pour considérer la prédiction des cibles des sRNAs. Ce travail étant encore préliminaire, nous avons identifié plusieurs perspectives à court terme d'évolution pour améliorer la prédiction les cibles des sRNAs, tels que :

(i) l'utilisation de plusieurs logiciels de prédiction pour améliorer la sensibilité de notre méthode,
(ii) le développement de filtres plus spécifiquement dédiés à la recherche de motifs via la topologie du graphe,
(iii) la mise à disposition de la communauté du travail réalisé sur les autres organismes testés pour développer notre approche auprès d'experts de ces organismes,
(iv) et l'intégration du pipeline dans un gestionnaire de *workflow* pour faciliter sa diffusion et son utilisation.

Le lecteur a ainsi pu voir dans ce manuscrit l'apport théorique et pratique que peut représenter les travaux d'intégration pour la biologie et pour l'informatique. Une combinaison de ces travaux pourrait être à présent envisagée comme le montrent les efforts actuels de la communauté pour intégrer à plus grande échelle les différentes méthodes de modélisation développées. On peut notamment citer la réalisation de la première simulation d'un modèle cellulaire complet [Karr *et al.* 2012] paru durant l'écriture de cette thèse. Cette modélisation intègre les résultats et les méthodes de simulations développées pour l'ADN, l'ARN et le métabolisme. Elle a montré des résultats prometteurs en prédisant de nouvelles fonctionnalités de ces cellules à l'échelle macroscopique, et ouvre ainsi une nouvelle voie vers une approche plus globale de la modélisation.

4. http ://www.cbib.u-bordeaux2.fr/irna/

Références bibliographiques

Acebo, P., Martin-Galiano, A. J., Navarro, S., Zaballos, A. et Amblar, M. (2012). Identification of 88 regulatory small rnas in the tigr4 strain of the human pathogen streptococcus pneumoniae. *RNA*, 18(3):530–546.

Adnan, M., Morton, G., Singh, J. et Hadi, S. (2010). Contribution of rpos and bola genes in biofilm formation in escherichia coli k-12 mg1655. *Mol Cell Biochem*, 342(1-2):207–213.

Agresti, A. (1992). A survey of exact inference for contingency tables. *Statistical Science*, 7(1):131–153.

Aiba, H. (2007). Mechanism of rna silencing by hfq-binding small rnas. *Curr Opin Microbiol*, 10(2):134–139.

Akama, T., Suzuki, K., Tanigawa, K., Kawashima, A., Wu, H., Nakata, N., Osana, Y., Sakakibara, Y. et Ishii, N. (2009). Whole-genome tiling array analysis of mycobacterium leprae rna reveals high expression of pseudogenes and noncoding regions. *J Bacteriol*, 191(10):3321–3327.

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K. et Walter, P. (2002). *Molecular biology of the cell, Garland Science.*

Albrecht, M., Sharma, C. M., Reinhardt, R., Vogel, J. et Rudel, T. (2010). Deep sequencing-based discovery of the chlamydia trachomatis transcriptome. *Nucleic Acids Res*, 38(3):868–877.

Alkan, C., Karakoc, E., Nadeau, J. H., Sahinalp, S. C. et Zhang, K. (2006). Rna-rna interaction prediction and antisense rna target search. *J Comput Biol*, 13(2):267–282.

Altschul, S. F., Gish, W., Miller, W., Myers, E. W. et Lipman, D. J. (1990). Basic local alignment search tool. *J Mol Biol*, 215(3):403–410.

Andronescu, M., Zhang, Z. C. et Condon, A. (2005). Secondary structure prediction of interacting rna molecules. *J Mol Biol*, 345(5):987–1001.

Anetzberger, C., Pirch, T. et Jung, K. (2009). Heterogeneity in quorum sensing-regulated bioluminescence of vibrio harveyi. *Mol Microbiol*, 73(2):267–277.

Argaman, L. et Altuvia, S. (2000). fhla repression by oxys rna : kissing complex formation at two sites results in a stable antisense-target rna complex. *J Mol Biol*, 300(5):1101–1112.

Argaman, L., Hershberg, R., Vogel, J., Bejerano, G., Wagner, E. G., Margalit, H. et Altuvia, S. (2001). Novel small rna-encoding genes in the intergenic regions of escherichia coli. *Curr Biol*, 11(12):941–950.

Arnvig, K. B. et Young, D. B. (2009). Identification of small rnas in mycobacterium tuberculosis. *Mol Microbiol*, 73(3):397–408.

Ashburner, M., Ball, C. A., Blake, J. A., Botstein, D., Butler, H., Cherry, J. M., Davis, A. P., Dolinski, K., Dwight, S. S., Eppig, J. T., Harris, M. A., Hill, D. P., Issel-Tarver, L., Kasarskis, A., Lewis, S., Matese, J. C., Richardson, J. E., Ringwald, M., Rubin, G. M. et Sherlock, G. (2000). Gene ontology : tool for the unification of biology. the gene ontology consortium. *Nat Genet*, 25(1):25–29.

Atkinson, S., Chang, C.-Y., Sockett, R. E., Cámara, M. et Williams, P. (2006). Quorum sensing in yersinia enterocolitica controls swimming and swarming motility. *J Bacteriol*, 188(4):1451–1461.

Auber, D. (2003). Tulip-a huge graph visualization framework. *Graph Drawing Software*, 2265:105–126.

Aziz, R. K., Bartels, D., Best, A. A., DeJongh, M., Disz, T., Edwards, R. A., Formsma, K., Gerdes, S., Glass, E. M., Kubal, M., Meyer, F., Olsen, G. J., Olson, R., Osterman, A. L., Overbeek, R. A., McNeil, L. K., Paarmann, D., Paczian, T., Parrello, B., Pusch, G. D., Reich, C., Stevens, R., Vassieva, O., Vonstein, V., Wilke, A. et Zagnitko, O. (2008). The rast server : rapid annotations using subsystems technology. *BMC Genomics*, 9:75.

Backofen, R. et Hess, W. R. (2010). Computational prediction of srnas and their targets in bacteria. *RNA Biol*, 7(1):33–42.

Bakker, B., Mensonides, F., Teusink, B., Van Hoek, P., Michels, P. et Westerhoff, H. (2000). Compartmentation protects trypanosomes from the dangerous design of glycolysis. *Proceedings of the National Academy of Sciences of the United States of America*, 97(5):2087.

Bakker, B. M., Michels, P. A., Opperdoes, F. R. et Westerhoff, H. V. (1997). Glycolysis in bloodstream form trypanosoma brucei can be understood in terms of the kinetics of the glycolytic enzymes. *J Biol Chem*, 272(6):3207–3215.

Bakker, B. M., Michels, P. A., Opperdoes, F. R. et Westerhoff, H. V. (1999). What controls glycolysis in bloodstream form trypanosoma brucei ? *J Biol Chem*, 274(21):14551–14559.

Balbontín, R., Fiorini, F., Figueroa-Bossi, N., Casadesús, J. et Bossi, L. (2010). Recognition of heptameric seed sequence underlies multi-target regulation by rybb small rna in salmonella enterica. *Mol Microbiol*, 78(2):380–394.

Banga, J. R. (2008). Optimization in computational systems biology. *BMC Syst Biol*, 2:47.

Bartel, D. P. (2009). Micrornas : target recognition and regulatory functions. *Cell*, 136(2): 215–233.

Bauer, M., Klau, G. W. et Reinert, K. (2007). Accurate multiple sequence-structure alignment of rna sequences using combinatorial optimization. *BMC Bioinformatics*, 8:271.

Beard, D. A., dan Liang, S. et Qian, H. (2002). Energy balance for analysis of complex metabolic networks. *Biophys J*, 83(1):79–86.

Becker, S. A. et Palsson, B. O. (2005). Genome-scale reconstruction of the metabolic network in staphylococcus aureus n315 : an initial draft to the two-dimensional annotation. *BMC Microbiol*, 5:8.

Beisel, C. L. et Storz, G. (2010). Base pairing small rnas and their roles in global regulatory networks. *FEMS Microbiol Rev*, 34(5):866–882.

Bergmann, S., Ihmels, J. et Barkai, N. (2004). Similarities and differences in genome-wide expression data of six organisms. *PLoS Biol*, 2(1):E9.

Bernhart, S. H., Tafer, H., Mückstein, U., Flamm, C., Stadler, P. F. et Hofacker, I. L. (2006). Partition function and base pairing probabilities of rna heterodimers. *Algorithms Mol Biol*, 1(1):3.

Berrar, D. et Flach, P. (2011). Caveats and pitfalls of roc analysis in clinical microarray research (and how to avoid them). *Brief Bioinform*, 13(1):83–97.

Besteiro, S., Barrett, M., Rivičre, L. et Bringaud, F. (2005). Energy generation in insect stages of trypanosoma brucei : metabolism in flux. *Trends in parasitology*, 21(4):185–191.

Blattner, F. R., Plunkett, 3rd, G., Bloch, C. A., Perna, N. T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J. D., Rode, C. K., Mayhew, G. F., Gregor, J., Davis, N. W., Kirkpatrick, H. A., Goeden, M. A., Rose, D. J., Mau, B. et Shao, Y. (1997). The complete genome sequence of escherichia coli k-12. *Science*, 277(5331):1453–1462.

Bochud-Allemann, N. et Schneider, A. (2002). Mitochondrial substrate level phosphorylation is essential for growth of procyclic trypanosoma brucei. *J Biol Chem*, 277(36):32849–32854.

Bohn, C., Rigoulay, C., Chabelskaya, S., Sharma, C. M., Marchais, A., Skorski, P., Borezée-Durant, E., Barbet, R., Jacquet, E., Jacq, A., Gautheret, D., Felden, B., Vogel, J. et Bouloc, P. (2010). Experimental discovery of small rnas in staphylococcus aureus reveals a riboregulator of central metabolism. *Nucleic Acids Res*, 38(19):6620–6636.

Bornstein, B., Keating, S., Jouraku, A. et Hucka, M. (2008). Libsbml : an api library for sbml. *Bioinformatics*, 24(6):880.

Bossi, L. et Figueroa-Bossi, N. (2007). A small rna downregulates lamb maltoporin in salmonella. *Mol Microbiol*, 65(3):799–810.

Bouvier, M., Sharma, C. M., Mika, F., Nierhaus, K. H. et Vogel, J. (2008). Small rna binding to 5' mrna coding region inhibits translational initiation. *Mol Cell*, 32(6):827–837.

Boyer, F. (2004). *Reconstruction ab initio de voies métaboliques-Formalisation et approches combinatoires*. Thèse de doctorat, Ph. D. Thesis, Université Joseph Fourier, Grenoble.

Boysen, A., Møller-Jensen, J., Kallipolitis, B., Valentin-Hansen, P. et Overgaard, M. (2010). Translational regulation of gene expression by an anaerobically induced small non-coding rna in escherichia coli. *J Biol Chem*, 285(14):10690–10702.

Brantl, S. (2007). Regulatory mechanisms employed by cis-encoded antisense rnas. *Curr Opin Microbiol*, 10(2):102–109.

Brennan, R. G. et Link, T. M. (2007). Hfq structure, function and ligand binding. *Curr Opin Microbiol*, 10(2):125–133.

Briggs, G. et Haldane, J. (1925). A note on the kinematics of enzyme actions. *Biochemical Journal*, 19:338–339.

Bringaud, F., Ebikeme, C. et Boshart, M. (2010). Acetate and succinate production in amoebae, helminths, diplomonads, trichomonads and trypanosomatids : common and diverse metabolic strategies used by parasitic lower eukaryotes. *Parasitology*, 137(9):1315–1331.

Bringaud, F., Rivière, L. et Coustou, V. (2006). Energy metabolism of trypanosomatids : adaptation to available carbon sources. *Mol Biochem Parasitol*, 149(1):1–9.

Burgard, A. P. et Maranas, C. D. (2003). Optimization-based framework for inferring and testing hypothesized metabolic objective functions. *Biotechnol Bioeng*, 82(6):670–677.

Burgard, A. P., Pharkya, P. et Maranas, C. D. (2003). Optknock : a bilevel programming framework for identifying gene knockout strategies for microbial strain optimization. *Biotechnol Bioeng*, 84(6):647–657.

Busch, A., Richter, A. S. et Backofen, R. (2008). Intarna : efficient prediction of bacterial srna targets incorporating target site accessibility and seed regions. *Bioinformatics*, 24(24):2849–2856.

Byrt, T., Bishop, J. et Carlin, J. (1993). Bias, prevalence and kappa. *Journal of clinical epidemiology*, 46(5):423–429.

Cameron, D. E., Urbach, J. M. et Mekalanos, J. J. (2008). A defined transposon mutant library and its use in identifying motility genes in vibrio cholerae. *Proc Natl Acad Sci U S A*, 105(25):8736–8741.

Cao, Y., Wu, J., Liu, Q., Zhao, Y., Ying, X., Cha, L., Wang, L. et Li, W. (2010). srnatarbase : a comprehensive database of bacterial srna targets verified by experiments. *RNA*, 16(11):2051–2057.

Cao, Y., Zhao, Y., Cha, L., Ying, X., Wang, L., Shao, N. et Li, W. (2009). srnatarget : a web server for prediction of bacterial srna targets. *Bioinformation*, 3(8):364–366.

Casagrande, A., Mysore, V., Piazza, C. et Mishra, B. (2005). Independent dynamics hybrid automata in systems biology. *In Proceedings of the First International Conference on Algebraic Biology (AB'05)*, pages 61–73.

Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S. Y., Tissier, C., Zhang, P. et Karp, P. D. (2006). Metacyc : a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34(Database issue):D511–D516.

Chang, R. (2000). *Physical chemistry for the chemical and biological sciences*. Univ Science Books.

Chen, L. et Vitkup, D. (2006). Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol*, 7(2):R17.

Chen, S., Lesnik, E. A., Hall, T. A., Sampath, R., Griffey, R. H., Ecker, D. J. et Blyn, L. B. (2002). A bioinformatics based approach to discover small rna genes in the escherichia coli genome. *Biosystems*, 65(2-3):157–177.

Chen, S., Zhang, A., Blyn, L. B. et Storz, G. (2004). Micc, a second small-rna regulator of omp protein expression in escherichia coli. *J Bacteriol*, 186(20):6689–6697.

Cherry, J., Adler, C., Ball, C., Chervitz, S., Dwight, S., Hester, E., Jia, Y., Juvik, G., Roe, T., Schroeder, M. *et al.* (1998). Sgd : Saccharomyces genome database. *Nucleic acids research*, 26(1):73–79.

Chiola, G., Dutheillet, C., Franceschinis, G. et Haddad, S. (1993). Stochastic well-formed colored nets and symmetric modeling applications. *Computers, IEEE Transactions on*, 42(11):1343–1360.

Chitsaz, H., Salari, R., Sahinalp, S. C. et Backofen, R. (2009). A partition function algorithm for interacting nucleic acid strands. *Bioinformatics*, 25(12):i365–i373.

Choudhuri, S. (2010). Small noncoding rnas : biogenesis, function, and emerging significance in toxicology. *J Biochem Mol Toxicol*, 24(3):195–216.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Coornaert, A., Lu, A., Mandin, P., Springer, M., Gottesman, S. et Guillier, M. (2010). Mica srna links the phop regulon to cell envelope stress. *Mol Microbiol*, 76(2):467–479.

Corrada, D., Viti, F., Merelli, I., Battaglia, C. et Milanesi, L. (2011). mymir : a genome-wide microrna targets identification and annotation tool. *Brief Bioinform*, 12(6): 588–600.

Coustou, V., Besteiro, S., Biran, M., Diolez, P., Bouchaud, V., Voisin, P., Michels, P. A. M., Canioni, P., Baltz, T. et Bringaud, F. (2003). Atp generation in the trypanosoma brucei procyclic form : cytosolic substrate level is essential, but not oxidative phosphorylation. *J Biol Chem*, 278(49):49625–49635.

Coustou, V., Besteiro, S., Rivière, L., Biran, M., Biteau, N., Franconi, J.-M., Boshart, M., Baltz, T. et Bringaud, F. (2005). A mitochondrial nadh-dependent fumarate reductase involved in the production of succinate excreted by procyclic trypanosoma brucei. *J Biol Chem*, 280(17):16559–16570.

Coustou, V., Biran, M., Besteiro, S., Rivière, L., Baltz, T., Franconi, J.-M. et Bringaud, F. (2006). Fumarate is an essential intermediary metabolite produced by the procyclic trypanosoma brucei. *J Biol Chem*, 281(37):26832–26846.

Coustou, V., Biran, M., Breton, M., Guegan, F., Rivière, L., Plazolles, N., Nolan, D., Barrett, M. P., Franconi, J.-M. et Bringaud, F. (2008). Glucose-induced remodeling of intermediary and energy metabolism in procyclic trypanosoma brucei. *J Biol Chem*, 283(24):16342–16354.

Covert, M. W., Knight, E. M., Reed, J. L., Herrgard, M. J. et Palsson, B. O. (2004). Integrating high-throughput and computational data elucidates bacterial networks. *Nature*, 429(6987):92–96.

Covert, M. W. et Palsson, B. O. (2002). Transcriptional regulation in constraints-based metabolic models of escherichia coli. *J Biol Chem*, 277(31):28058–28064.

Crick, F. (1970). Central dogma of molecular biology. *Nature*, 227(5258):561–563.

Daniels, R., Vanderleyden, J. et Michiels, J. (2004). Quorum sensing and swarming migration in bacteria. *FEMS Microbiol Rev*, 28(3):261–289.

Darty, K., Denise, A. et Ponty, Y. (2009). Varna : Interactive drawing and editing of the rna secondary structure. *Bioinformatics*, 25(15):1974–1975.

David, R. et Alla, H. (1987). Continuous petri nets. *In 8th European Workshop on Application and Theory of Petri nets*, volume 340, pages 275–294.

Davidsen, T., Beck, E., Ganapathy, A., Montgomery, R., Zafar, N., Yang, Q., Madupu, R., Goetz, P., Galinsky, K., White, O. *et al.* (2010). The comprehensive microbial resource. *Nucleic acids research*, 38(suppl 1):D340–D345.

Davis, B. M., Quinones, M., Pratt, J., Ding, Y. et Waldor, M. K. (2005). Characterization of the small untranslated rna ryhb and its regulon in vibrio cholerae. *J Bacteriol*, 187(12):4005–4014.

De Lay, N. et Gottesman, S. (2009). The crp-activated small noncoding regulatory rna cyar (ryee) links nutritional status to group behavior. *J Bacteriol*, 191(2):461–476.

DeJongh, M., Formsma, K., Boillot, P., Gould, J., Rycenga, M. et Best, A. (2007). Toward the automated generation of genome-scale metabolic networks in the seed. *BMC Bioinformatics*, 8:139.

Del Moral, P. et Miclo, L. (1999). On the convergence and the applications of the generalized simulated annealing. *SIAM Journal on Control and Optimization*, 37:1222–1250.

DeLong, E. R., DeLong, D. M. et Clarke-Pearson, D. L. (1988). Comparing the areas under two or more correlated receiver operating characteristic curves : a nonparametric approach. *Biometrics*, 44(3):837–845.

Denielou, Y. (2010). Alignement multiple de données génomiques et post-génomiques : Approches algorithmiques.

Desnoyers, G., Morissette, A., Prévost, K. et Massé, E. (2009). Small rna-induced differential degradation of the polycistronic mrna iscrsua. *EMBO J*, 28(11):1551–1561.

Do, C. B., Woods, D. A. et Batzoglou, S. (2006). Contrafold : Rna secondary structure prediction without physics-based models. *Bioinformatics*, 22(14):e90–e98.

Douchin, V., Bohn, C. et Bouloc, P. (2006). Down-regulation of porins by a small rna bypasses the essentiality of the regulated intramembrane proteolysis protease rsep in escherichia coli. *J Biol Chem*, 281(18):12253–12259.

Duarte, N., Becker, S., Jamshidi, N., Thiele, I., Mo, M., Vo, T., Srivas, R. et Palsson, B. (2007). Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proceedings of the National Academy of Sciences*, 104(6):1777.

Duarte, N. C., Herrgard, M. J. et Palsson, B. O. (2004). Reconstruction and validation of saccharomyces cerevisiae ind750, a fully compartmentalized genome-scale metabolic model. *Genome Res*, 14(7):1298–1309.

Dubois, J., Cottret, L., Ghozlane, A., Auber, D., Bringaud, F., Thébault, P. et Bourqui, R. (2012). Systrip : a visual environment for the investigation of time-series data in the context of metabolic networks. *In IV2012 : 16th International Conference Information Visualisation*.

Dyszel, J. L., Soares, J. A., Swearingen, M. C., Lindsay, A., Smith, J. N. et Ahmer, B. M. M. (2010). E. coli k-12 and ehec genes regulated by sdia. *PLoS One*, 5(1):e8946.

Ebikeme, C., Hubert, J., Biran, M., Gouspillou, G., Morand, P., Plazolles, N., Guegan, F., Diolez, P., Franconi, J.-M., Portais, J.-C. et Bringaud, F. (2010). Ablation of succinate production from glucose metabolism in the procyclic trypanosomes induces metabolic switches to the glycerol 3-phosphate/dihydroxyacetone phosphate shuttle and to proline metabolism. *J Biol Chem*, 285(42):32312–32324.

Edwards, J. S. et Palsson, B. O. (2000). The escherichia coli mg1655 in silico metabolic genotype : its definition, characteristics, and capabilities. *Proc Natl Acad Sci U S A*, 97(10): 5528–5533.

Eggenhofer, F., Tafer, H., Stadler, P. F. et Hofacker, I. L. (2011). Rnapredator : fast accessibility-based prediction of srna targets. *Nucleic Acids Res*, 39(Web Server issue): W149–W154.

Elbashir, S. M., Harborth, J., Lendeckel, W., Yalcin, A., Weber, K. et Tuschl, T. (2001). Duplexes of 21-nucleotide rnas mediate rna interference in cultured mammalian cells. *Nature*, 411(6836):494–498.

Feist, A. M., Henry, C. S., Reed, J. L., Krummenacker, M., Joyce, A. R., Karp, P. D., Broadbelt, L. J., Hatzimanikatis, V. et Palsson, B. O. (2007). A genome-scale metabolic reconstruction for escherichia coli k-12 mg1655 that accounts for 1260 orfs and thermodynamic information. *Mol Syst Biol*, 3:121.

Feist, A. M., Herrgård, M. J., Thiele, I., Reed, J. L. et Palsson, B. Ø. (2009). Reconstruction of biochemical networks in microorganisms. *Nat Rev Microbiol*, 7(2):129–143.

Feist, A. M. et Palsson, B. O. (2010). The biomass objective function. *Curr Opin Microbiol*, 13(3):344–349.

Feist, A. M., Scholten, J. C. M., Palsson, B. O., Brockman, F. J. et Ideker, T. (2006). Modeling methanogenesis with a genome-scale metabolic reconstruction of methanosarcina barkeri. *Mol Syst Biol*, 2:2006.0004.

Fell, D. A. et Small, J. R. (1986). Fat synthesis in adipose tissue. an examination of stoichiometric constraints. *Biochem J*, 238(3):781–786.

Figueroa-Bossi, N., Valentini, M., Malleret, L., Fiorini, F. et Bossi, L. (2009). Caught at its own game : regulatory small rna inactivated by an inducible transcript mimicking its target. *Genes Dev*, 23(17):2004–2015.

Fischer, E. et Sauer, U. (2003). A novel metabolic cycle catalyzes glucose oxidation and anaplerosis in hungry escherichia coli. *J Biol Chem*, 278(47):46446–46451.

Fozo, E. M., Hemm, M. R. et Storz, G. (2008). Small toxic proteins and the antisense rnas that repress them. *Microbiol Mol Biol Rev*, 72(4):579–89, Table of Contents.

Freeman, J. A. et Bassler, B. L. (1999). Sequence and function of luxu : a two-component phosphorelay protein that regulates quorum sensing in vibrio harveyi. *J Bacteriol*, 181(3): 899–906.

Freeman, L. (1977). A set of measures of centrality based on betweenness. *Sociometry*, pages 35–41.

Fromm, J. (2004). *The emergence of complexity*. Kassel university press.

Galassi, M., Davies, J., Theiler, J., Gough, B., Jungman, G., Alken, P., Booth, M. et F., R. (2009). *GNU Scientific Library Reference Manual - Third Edition*. Network Theory.

Gallagher, L. A., Ramage, E., Jacobs, M. A., Kaul, R., Brittnacher, M. et Manoil, C. (2007). A comprehensive transposon mutant library of francisella novicida, a bioweapon surrogate. *Proc Natl Acad Sci U S A*, 104(3):1009–1014.

Geissmann, T. A. et Touati, D. (2004). Hfq, a new chaperoning role : binding to messenger rna determines access for small rna regulator. *EMBO J*, 23(2):396–405.

Gerdes, K. et Wagner, E. G. H. (2007). Rna antitoxins. *Curr Opin Microbiol*, 10(2):117–124.

Gerlach, W. et Giegerich, R. (2006). Guugle : a utility for fast exact matching under rna complementary rules including g-u base pairing. *Bioinformatics*, 22(6):762–764.

Ghozlane, A., Bringaud, F., Soueidan, H., Dutour, I., Jourdan, F. et Thebault, P. (2012). Flux analysis of the trypanosoma brucei glycolysis based on a multi-objective criteria bioinformatic approach. *Advances in bioinformatics*.

Gillespie, D. (1976). A general method for numerically simulating the stochastic time evolution of coupled chemical reactions. *Journal of computational physics*, 22(4):403–434.

Gillespie, D. (1977). Exact stochastic simulation of coupled chemical reactions. *The journal of physical chemistry*, 81(25):2340–2361.

Goecks, J., Nekrutenko, A., Taylor, J. et , G. T. (2010). Galaxy : a comprehensive approach for supporting accessible, reproducible, and transparent computational research in the life sciences. *Genome Biol*, 11(8):R86.

Gonzalez, O., Oberwinkler, T., Mansueto, L., Pfeiffer, F., Mendoza, E., Zimmer, R. et Oesterhelt, D. (2010). Characterization of growth and metabolism of the haloalkaliphile natronomonas pharaonis. *PLoS Comput Biol*, 6(6):e1000799.

González Barrios, A. F., Zuo, R., Hashimoto, Y., Yang, L., Bentley, W. E. et Wood, T. K. (2006). Autoinducer 2 controls biofilm formation in escherichia coli through a novel motility quorum-sensing regulator (mqsr, b3022). *J Bacteriol*, 188(1):305–316.

Gottesman, S. (2005). Micros for microbes : non-coding regulatory rnas in bacteria. *Trends Genet*, 21(7):399–404.

Gottesman, S. et Storz, G. (2011). Bacterial small rna regulators : versatile roles and rapidly evolving variations. *Cold Spring Harb Perspect Biol*, 3(12).

Grafahrend-Belau, E., Klukas, C., Junker, B. H. et Schreiber, F. (2009). Fbasimvis : interactive visualization of constraint-based metabolic models. *Bioinformatics*, 25(20):2755–2757.

Green, M. L. et Karp, P. D. (2007). Using genome-context data to identify specific types of functional associations in pathway/genome databases. *Bioinformatics*, 23(13):i205–i211.

Gross, C., Neidhardt, F., Curtiss, R. et Lin, E. (1996). Escherichia coli and salmonella : cellular and molecular biology. *Escherichia coli and Salmonella : Cellular and Molecular Biology*.

Gross, J. et Yellen, J. (2004). *Handbook of graph theory*. CRC.

Gualdron-López, M., Vapola, M. H., Miinalainen, I. J., Hiltunen, J. K., Michels, P. A. M. et Antonenkov, V. D. (2012). Channel-forming activities in the glycosomal fraction from the bloodstream form of trypanosoma brucei. *PLoS One*, 7(4):e34530.

Gudmundsson, S. et Thiele, I. (2010). Computationally efficient flux variability analysis. *BMC Bioinformatics*, 11:489.

Guell, M., van Noort, V., Yus, E., Chen, W.-H., Leigh-Bell, J., Michalodimitrakis, K., Yamada, T., Arumugam, M., Doerks, T., Kühner, S., Rode, M., Suyama, M., Schmidt, S., Gavin, A.-C., Bork, P. et Serrano, L. (2009). Transcriptome complexity in a genome-reduced bacterium. *Science*, 326(5957):1268–1271.

Guillier, M. et Gottesman, S. (2006). Remodelling of the escherichia coli outer membrane by two small regulatory rnas. *Mol Microbiol*, 59(1):231–247.

Guillier, M. et Gottesman, S. (2008). The 5' end of two redundant srnas is involved in the regulation of multiple targets, including their own regulator. *Nucleic Acids Res*, 36(21):6781–6794.

Hage, P. et Harary, F. (1995). Eccentricity and centrality in networks. *Social networks*, 17(1):57–63.

Hand, D. J. et Till, R. J. (2001). *A Simple Generalisation of the Area Under the ROC Curve for Multiple Class Classification Problems*.

Hatzimanikatis, V., Li, C., Ionita, J. A., Henry, C. S., Jankowski, M. D. et Broadbelt, L. J. (2005). Exploring the diversity of complex metabolic networks. *Bioinformatics*, 21(8):1603–1609.

Heiner, M., Gilbert, D. et Donaldson, R. (2008). Petri nets for systems and synthetic biology. *In Proceedings of the Formal methods for the design of computer, communication, and software systems 8th international conference on Formal methods for computational systems biology*, pages 215–264. Springer-Verlag.

Henke, J. M. et Bassler, B. L. (2004). Quorum sensing regulates type iii secretion in vibrio harveyi and vibrio parahaemolyticus. *J Bacteriol*, 186(12):3794–3805.

Henry, C. S., Zinner, J. F., Cohoon, M. P. et Stevens, R. L. (2009). ibsu1103 : a new genome-scale metabolic model of bacillus subtilis based on seed annotations. *Genome Biol*, 10(6):R69.

Herrgård, M. J., Fong, S. S. et Palsson, B. O. (2006). Identification of genome-scale metabolic network models using experimentally measured flux profiles. *PLoS Comput Biol*, 2(7):e72.

Hershberg, R., Altuvia, S. et Margalit, H. (2003). A survey of small rna-encoding genes in escherichia coli. *Nucleic Acids Res*, 31(7):1813–1820.

Herzberg, M., Kaye, I. K., Peti, W. et Wood, T. K. (2006). Ydgg (tqsa) controls biofilm formation in escherichia coli k-12 through autoinducer 2 transport. *J Bacteriol*, 188(2):587–598.

Hofacker, I., Fontana, W., Stadler, P., Bonhoeffer, L., Tacker, M. et Schuster, P. (1994). Fast folding and comparison of rna secondary structures. *Monatshefte für Chemie/Chemical Monthly*, 125(2):167–188.

Hofacker, I. L. (2003). Vienna rna secondary structure server. *Nucleic Acids Res*, 31(13): 3429–3431.

Holmqvist, E., Reimegård, J., Sterk, M., Grantcharova, N., Römling, U. et Wagner, E. G. H. (2010). Two antisense rnas target the transcriptional regulator csgd to inhibit curli synthesis. *EMBO J*, 29(11):1840–1850.

Hoppe, A., Hoffmann, S., Gerasch, A., Gille, C. et Holzhütter, H.-G. (2011). Fasimu : flexible software for flux-balance computation series in large metabolic networks. *BMC Bioinformatics*, 12:28.

Huang, D. W., Sherman, B. T. et Lempicki, R. A. (2009a). Systematic and integrative analysis of large gene lists using david bioinformatics resources. *Nat Protoc*, 4(1):44–57.

Huang, D. W., Sherman, B. T., Tan, Q., Collins, J. R., Alvord, W. G., Roayaei, J., Stephens, R., Baseler, M. W., Lane, H. C. et Lempicki, R. A. (2007a). The david gene functional classification tool : a novel biological module-centric algorithm to functionally analyze large gene lists. *Genome Biol*, 8(9):R183.

Huang, D. W., Sherman, B. T., Tan, Q., Kir, J., Liu, D., Bryant, D., Guo, Y., Stephens, R., Baseler, M. W., Lane, H. C. et Lempicki, R. A. (2007b). David bioinformatics resources : expanded annotation database and novel algorithms to better extract biology from large gene lists. *Nucleic Acids Res*, 35(Web Server issue):W169–W175.

Huang, F. W. D., Qin, J., Reidys, C. M. et Stadler, P. F. (2009b). Partition function and base pairing probabilities for rna-rna interaction prediction. *Bioinformatics*, 25(20):2646–2654.

Hucka, M., Finney, A., Sauro, H. M., Bolouri, H., Doyle, J. C., Kitano, H., Arkin, A. P., Bornstein, B. J., Bray, D., Cornish-Bowden, A., Cuellar, A. A., Dronov, S., Gilles, E. D., Ginkel, M., Gor, V., Goryanin, I. I., Hedley, W. J., Hodgman, T. C., Hofmeyr, J.-H., Hunter, P. J., Juty, N. S., Kasberger, J. L., Kremling, A., Kummer, U., Le Novère, N., Loew, L. M., Lucio, D., Mendes, P., Minch, E., Mjolsness, E. D., Nakayama, Y., Nelson, M. R., Nielsen, P. F., Sakurada, T., Schaff, J. C., Shapiro, B. E., Shimizu, T. S., Spence, H. D., Stelling, J., Takahashi, K., Tomita, M., Wagner, J., Wang, J. et , S. B. M. L. F. (2003). The systems biology markup language (sbml) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4):524–531.

Iliopoulos, I., Tsoka, S., Andrade, M. A., Janssen, P., Audit, B., Tramontano, A., Valencia, A., Leroy, C., Sander, C. et Ouzounis, C. A. (2001). Genome sequences and great expectations. *Genome Biol*, 2(1):INTERACTIONS0001.

Irnov, I., Sharma, C. M., Vogel, J. et Winkler, W. C. (2010). Identification of regulatory rnas in bacillus subtilis. *Nucleic Acids Res*, 38(19):6637–6651.

Izar, B., Mraheil, M. et Hain, T. (2011). Identification and role of regulatory non-coding rnas in listeria monocytogenes. *International Journal of Molecular Sciences*, 12(8):5070–5079.

Johansen, J., Rasmussen, A. A., Overgaard, M. et Valentin-Hansen, P. (2006). Conserved small non-coding rnas that belong to the sigmae regulon : role in down-regulation of outer membrane proteins. *J Mol Biol*, 364(1):1–8.

Jou, W., Haegeman, G., Ysebaert, M. et Fiers, W. (1972). Nucleotide sequence of the gene coding for the bacteriophage ms2 coat protein. *Nature*, 237:82–88.

Jousselin, A., Metzinger, L. et Felden, B. (2009). On the facultative requirement of the bacterial rna chaperone, hfq. *Trends Microbiol*, 17(9):399–405.

Jung, Y.-S. et Kwon, Y.-M. (2008). Small rna arrf regulates the expression of sodb and fesii genes in azotobacter vinelandii. *Curr Microbiol*, 57(6):593–597.

Junker, B. H., Klukas, C. et Schreiber, F. (2006). Vanted : a system for advanced data analysis and visualization in the context of biological networks. *BMC Bioinformatics*, 7:109.

Jády, B. E. et Kiss, T. (2001). A small nucleolar guide rna functions both in 2'-o-ribose methylation and pseudouridylation of the u5 spliceosomal rna. *EMBO J*, 20(3):541–551.

Kanehisa, M. et Goto, S. (2000). Kegg : kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 28(1):27–30.

Karp, P., Paley, S., Altman, T., Paulsen, I., Keseler, I., Caspi, R., Krummenacker, M., Latendresse, M., Dale, J., Lee, T. *et al.* (2010). Pathway tools version 13.0 : integrated software for pathway/genome informatics and systems biology| macquarie university researchonline.

Karr, J. R., Sanghvi, J. C., Macklin, D. N., Gutschow, M. V., Jacobs, J. M., Bolival, Jr, B., Assad-Garcia, N., Glass, J. I. et Covert, M. W. (2012). A whole-cell computational model predicts phenotype from genotype. *Cell*, 150(2):389–401.

Kato, Y., Akutsu, T. et Seki, H. (2009). A grammatical approach to rna-rna interaction prediction. *Pattern Recognition*, 42(4):531–538.

Kato, Y., Sato, K., Hamada, M., Watanabe, Y., Asai, K. et Akutsu, T. (2010). Ractip : fast and accurate prediction of rna-rna interaction using integer programming. *Bioinformatics*, 26(18):i460–i466.

Kauffman, K. J., Prakash, P. et Edwards, J. S. (2003). Advances in flux balance analysis. *Curr Opin Biotechnol*, 14(5):491–496.

Kawamoto, H., Koide, Y., Morita, T. et Aiba, H. (2006). Base-pairing requirement for rna silencing by a bacterial small rna and acceleration of duplex formation by hfq. *Mol Microbiol*, 61(4):1013–1022.

Kawamoto, H., Morita, T., Shimizu, A., Inada, T. et Aiba, H. (2005). Implication of membrane localization of target mrna in the action of a small rna : mechanism of post-transcriptional regulation of glucose transporter in escherichia coli. *Genes Dev*, 19(3):328–338.

Kawano, M., Aravind, L. et Storz, G. (2007). An antisense rna controls synthesis of an sos-induced toxin evolved from an antitoxin. *Mol Microbiol*, 64(3):738–754.

Keseler, I. M., Collado-Vides, J., Santos-Zavaleta, A., Peralta-Gil, M., Gama-Castro, S., Muñiz-Rascado, L., Bonavides-Martinez, C., Paley, S., Krummenacker, M., Altman, T., Kaipa, P., Spaulding, A., Pacheco, J., Latendresse, M., Fulcher, C., Sarker, M., Shearer, A. G., Mackie, A., Paulsen, I., Gunsalus, R. P. et Karp, P. D. (2011). Ecocyc : a comprehensive database of escherichia coli biology. *Nucleic Acids Res*, 39(Database issue):D583–D590.

Kim, J. et Reed, J. L. (2010). Optorf : Optimal metabolic and regulatory perturbations for metabolic engineering of microbial strains. *BMC Syst Biol*, 4:53.

Kim, T. Y., Sohn, S. B., Kim, Y. B., Kim, W. J. et Lee, S. Y. (2011). Recent advances in reconstruction and applications of genome-scale metabolic models. *Curr Opin Biotechnol.*

Kircher, M. et Kelso, J. (2010). High-throughput dna sequencing–concepts and limitations. *Bioessays*, 32(6):524–536.

Kirkpatrick, S., Gelatt, C. D. et Vecchi, M. P. (1983). Optimization by simulated annealing. *Science*, 220(4598):671–680.

Kitano, H. (2002). Computational systems biology. *Nature*, 420(6912):206–210.

Klamt, S., Haus, U.-U. et Theis, F. (2009). Hypergraphs and cellular networks. *PLoS Comput Biol*, 5(5):e1000385.

Koch, I., Junker, B. H. et Heiner, M. (2005). Application of petri net theory for modelling and validation of the sucrose breakdown pathway in the potato tuber. *Bioinformatics*, 21(7):1219–1226.

Koch, I., Reisig, W. et Schreiber, F. (2011). *Modeling in Systems Biology The Petri Net Approach*. Springer London.

Koo, J. T., Alleyne, T. M., Schiano, C. A., Jafari, N. et Lathem, W. W. (2011). Global discovery of small rnas in yersinia pseudotuberculosis identifies yersinia-specific small, non-coding rnas required for virulence. *Proc Natl Acad Sci U S A*.

Koshland, D. E. (1958). Application of a theory of enzyme specificity to protein synthesis. *Proc Natl Acad Sci U S A*, 44(2):98–104.

Krek, A., Grün, D., Poy, M. N., Wolf, R., Rosenberg, L., Epstein, E. J., MacMenamin, P., da Piedade, I., Gunsalus, K. C., Stoffel, M. et Rajewsky, N. (2005). Combinatorial microrna target predictions. *Nat Genet*, 37(5):495–500.

Kulikova, T., Akhtar, R., Aldebert, P., Althorpe, N., Andersson, M., Baldwin, A., Bates, K., Bhattacharyya, S., Bower, L., Browne, P. *et al.* (2007). Embl nucleotide sequence database in 2006. *Nucleic acids research*, 35(suppl 1):D16–D20.

Kumar, R., Shah, P., Swiatlo, E., Burgess, S. C., Lawrence, M. L. et Nanduri, B. (2010). Identification of novel non-coding small rnas from streptococcus pneumoniae tigr4 using high-resolution genome tiling arrays. *BMC Genomics*, 11:350.

Kumar, S., Vinay, Dasika, M. S. et Maranas, C. D. (2007). Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8:212.

Lamour, N., Rivière, L., Coustou, V., Coombs, G. H., Barrett, M. P. et Bringaud, F. (2005). Proline metabolism in procyclic trypanosoma brucei is down-regulated in the presence of glucose. *J Biol Chem*, 280(12):11902–11910.

Lamprecht, R., Smith, G. et Kemper, P. (2011). Stochastic petri net models of ca 2+ signaling complexes and their analysis. *Natural Computing*, 10(3):1045–1075.

Lease, R. A., Cusick, M. E. et Belfort, M. (1998). Riboregulation in escherichia coli : Dsra rna acts by rna :rna interactions at multiple loci. *Proc Natl Acad Sci U S A*, 95(21):12456–12461.

Lee, J. M., Gianchandani, E. P. et Papin, J. A. (2006). Flux balance analysis in the era of metabolomics. *Brief Bioinform*, 7(2):140–150.

Lee, K., Berthiaume, F., Stephanopoulos, G. N. et Yarmush, M. L. (1999). Metabolic flux analysis : a powerful tool for monitoring tissue function. *Tissue Eng*, 5(4):347–368.

Lei, J. (2010). Stochastic modeling in systems biology. *Journal of Advanced Mathematics and Applications*.

Lespinet, O. et Labedan, B. (2005). Orphan enzymes ? *Science*, 307(5706):42.

Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P.-O., Han, J.-D. J., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J.-F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang, L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano,

F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E. et Vidal, M. (2004). A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657):540–543.

Liu, J. M., Livny, J., Lawrence, M. S., Kimball, M. D., Waldor, M. K. et Camilli, A. (2009). Experimental discovery of srnas in vibrio cholerae by direct cloning, 5s/trna depletion and parallel sequencing. *Nucleic Acids Res*, 37(6):e46.

Livny, J., Brencic, A., Lory, S. et Waldor, M. K. (2006). Identification of 17 pseudomonas aeruginosa srnas and prediction of srna-encoding genes in 10 diverse pathogens using the bioinformatic tool srnapredict2. *Nucleic Acids Res*, 34(12):3484–3493.

Lladser, M. E., Betterton, M. D. et Knight, R. (2008). Multiple pattern matching : a markov chain approach. *J Math Biol*, 56(1-2):51–92.

Llaneras, F. et Picó, J. (2008). Stoichiometric modelling of cell metabolism. *J Biosci Bioeng*, 105(1):1–11.

Loh, K. D., Gyaneshwar, P., Markenscoff Papadimitriou, E., Fong, R., Kim, K.-S., Parales, R., Zhou, Z., Inwood, W. et Kustu, S. (2006). A previously undescribed pathway for pyrimidine catabolism. *Proc Natl Acad Sci U S A*, 103(13):5114–5119.

Lu, X., Goodrich-Blair, H. et Tjaden, B. (2011). Assessing computational tools for the discovery of small rna genes in bacteria. *RNA*, 17(9):1635–1647.

Ma, H. et Zeng, A.-P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2):270–277.

Maglott, D., Ostell, J., Pruitt, K. et Tatusova, T. (2011). Entrez gene : gene-centered information at ncbi. *Nucleic acids research*, 39(suppl 1):D52–D57.

Majdalani, N., Vanderpool, C. K. et Gottesman, S. (2005). Bacterial small rna regulators. *Crit Rev Biochem Mol Biol*, 40(2):93–113.

Mallick, B. et Ghosh, Z. (2012). *Regulatory RNAs : basics, methods and applications*. Springer Verlag.

Mandin, P. et Gottesman, S. (2010). Integrating anaerobic/aerobic sensing and the general stress response through the arcz small rna. *EMBO J*, 29(18):3094–3107.

Markowitz, V., Korzeniewski, F., Palaniappan, K., Szeto, E., Werner, G., Padki, A., Zhao, X., Dubchak, I., Hugenholtz, P., Anderson, I. *et al.* (2006). The integrated microbial genomes (img) system. *Nucleic Acids Research*, 34(suppl 1):D344–D348.

Marsan, A., Conte, G. et Balbo, G. (1984). A class of generalized stochastic petri nets for the performance evaluation of multiprocessor systems. *ACM Transactions on Computer Systems (TOCS)*, 2(2):93–122.

Marschall, T. (2011). Construction of minimal deterministic finite automata from biological motifs. *Theoretical Computer Science*, 412(8):922–930.

Massé, E. et Gottesman, S. (2002). A small rna regulates the expression of genes involved in iron metabolism in escherichia coli. *Proc Natl Acad Sci U S A*, 99(7):4620–4625.

Massé, E., Salvail, H., Desnoyers, G. et Arguin, M. (2007). Small rnas controlling iron metabolism. *Curr Opin Microbiol*, 10(2):140–145.

Mathews, D. H., Sabina, J., Zuker, M. et Turner, D. H. (1999). Expanded sequence dependence of thermodynamic parameters improves prediction of rna secondary structure. *J Mol Biol*, 288(5):911–940.

Mattick, J. S. et Makunin, I. V. (2006). Non-coding rna. *Hum Mol Genet*, 15 Spec No 1:R17–R29.

McCaskill, J. S. (1990). The equilibrium partition function and base pair binding probabilities for rna secondary structure. *Biopolymers*, 29(6-7):1105–1119.

Menten, L. et Michaelis, M. (1913). Die kinetik der invertinwirkung. *Biochem Z*, 49:333–369.

Meyer, J., Movaghar, A. et Sanders, W. (1985). Stochastic activity networks : Structure, behavior, and application. *In Proceedings of the International Workshop on Timed Petri Nets*, pages 106–115. IEEE Computer Society Press.

Michels, P. A. M., Bringaud, F., Herman, M. et Hannaert, V. (2006). Metabolic functions of glycosomes in trypanosomatids. *Biochim Biophys Acta*, 1763(12):1463–1477.

Millerioux, Y., Morand, P., Biran, M., Mazet, M., Moreau, P., Wargnies, M., Ebikeme, C., Deramchia, K., Gales, L., Portais, J.-C., Boshart, M., Franconi, J.-M. et Bringaud, F. (2012). Atp synthesis-coupled and -uncoupled acetate production from acetyl-coa by mitochondrial acetate :succinate coa-transferase and acetyl-coa thioesterase in trypanosoma. *J Biol Chem*, 287(21):17186–17197.

Modi, S. R., Camacho, D. M., Kohanski, M. A., Walker, G. C. et Collins, J. J. (2011). Functional characterization of bacterial srnas using a network biology approach. *Proc Natl Acad Sci U S A*, 108(37):15522–15527.

Molloy, M. K. (1982). Performance analysis using stochastic petri nets. *IEEE Trans. Comput.*, 31(9):913–917.

Muckstein, U., Tafer, H., Hackermuller, J., Bernhart, S. H., Stadler, P. F. et Hofacker, I. L. (2006). Thermodynamics of rna-rna binding. *Bioinformatics*, 22(10):1177–1182.

Muppala, J., Ciardo, G. et Trivedi, K. (1994). Stochastic reward nets for reliability prediction. *Communications in reliability, maintainability and serviceability*, 1(2):9–20.

Murata, T. (1989). Petri nets : Properties, analysis and applications. *Proceedings of the IEEE*, 77(4):541–580.

Møller, T., Franch, T., Udesen, C., Gerdes, K. et Valentin-Hansen, P. (2002). Spot 42 rna mediates discoordinate expression of the e. coli galactose operon. *Genes Dev*, 16(13): 1696–1706.

Nabli, F. (2011). Finding minimal siphons and traps as a constraint satisfaction problem. *In The Seventeenth International Conference on Principles and Practice of Constraint Programming (CP 2011)*, page 67.

Nagrath, D., Avila-Elchiver, M., Berthiaume, F., Tilles, A. W., Messac, A. et Yarmush, M. L. (2010). Soft constraints-based multiobjective framework for flux balance analysis. *Metab Eng*, 12(5):429–445.

Navarro, G. et Raffinot, M. (2002). *Flexible pattern matching in strings : practical on-line search algorithms for texts and biological sequences*. Cambridge Univ Pr.

Nelder, J. A. et Mead, R. (1965). A simplex method for function minimization. *The Computer Journal*, 7(4):308–313.

Ng, W.-L. et Bassler, B. L. (2009). Bacterial quorum-sensing network architectures. *Annu Rev Genet*, 43:197–222.

Nicolas, P., Mäder, U., Dervyn, E., Rochat, T., Leduc, A., Pigeonneau, N., Bidnenko, E., Marchadier, E., Hoebeke, M., Aymerich, S., Becher, D., Bisicchia, P., Botella, E., Delumeau, O., Doherty, G., Denham, E. L., Fogg, M. J., Fromion, V., Goelzer, A., Hansen, A., Härtig, E., Harwood, C. R., Homuth, G., Jarmer, H., Jules, M., Klipp, E., Le Chat, L., Lecointe, F., Lewis, P., Liebermeister, W., March, A., Mars, R. A. T., Nannapaneni, P., Noone, D., Pohl, S., Rinn, B., Rügheimer, F., Sappa, P. K., Samson, F., Schaffer, M., Schwikowski, B., Steil, L., Stülke, J., Wiegert, T., Devine, K. M., Wilkinson, A. J., van Dijl, J. M., Hecker, M., Völker, U., Bessières, P. et Noirot, P. (2012). Condition-dependent transcriptome reveals high-level regulatory architecture in bacillus subtilis. *Science*, 335(6072):1103–1106.

Nielsen, J. S., Lei, L. K., Ebersbach, T., Olsen, A. S., Klitgaard, J. K., Valentin-Hansen, P. et Kallipolitis, B. H. (2010). Defining a role for hfq in gram-positive bacteria : evidence for hfq-dependent antisense regulation in listeria monocytogenes. *Nucleic Acids Res*, 38(3):907–919.

Niittylae, T., Chaudhuri, B., Sauer, U. et Frommer, W. B. (2009). Comparison of quantitative metabolite imaging tools and carbon-13 techniques for fluxomics. *Methods Mol Biol*, 553:355–372.

Noe, L. et Kucherov, G. (2005). Yass : enhancing the sensitivity of dna similarity search. *Nucleic Acids Res*, 33(Web Server issue):W540–W543.

Oh, Y., Lee, D., Lee, S. et Park, S. (2009). Multiobjective flux balancing using the nise method for metabolic network analysis. *Biotechnology progress*, 25(4):999–1008.

Opdyke, J. A., Fozo, E. M., Hemm, M. R. et Storz, G. (2011). Rnase iii participates in gady-dependent cleavage of the gadx-gadw mrna. *J Mol Biol*, 406(1):29–43.

Opdyke, J. A., Kang, J.-G. et Storz, G. (2004). Gady, a small-rna regulator of acid response genes in escherichia coli. *J Bacteriol*, 186(20):6698–6705.

Opperdoes, F. et Borst, P. (1977). Localization of nine glycolytic enzymes in a microbody-like organelle in trypanosoma brucei : the glycosome. *FEBS letters*, 80(2):360.

Orth, J. D. et Palsson, B. O. (2010). Systematizing the generation of missing metabolic knowledge. *Biotechnol Bioeng*, 107(3):403–412.

Orth, J. D., Thiele, I. et Palsson, B. (2010). What is flux balance analysis ? *Nat Biotechnol*, 28(3):245–248.

Osterman, A. (2006). A hidden metabolic pathway exposed. *Proc Natl Acad Sci U S A*, 103(15):5637–5638.

Overbeek, R., Larsen, N., Walunas, T., D'Souza, M., Pusch, G., Selkov, Jr, E., Liolios, K., Joukov, V., Kaznadzey, D., Anderson, I., Bhattacharyya, A., Burd, H., Gardner, W., Hanke, P., Kapatral, V., Mikhailova, N., Vasieva, O., Osterman, A., Vonstein, V., Fonstein, M., Ivanova, N. et Kyrpides, N. (2003). The ergo genome analysis and discovery system. *Nucleic Acids Res*, 31(1):164–171.

Palsson, B. (2006). *Systems biology : properties of reconstructed networks*. Cambridge Univ Pr.

Pandey, S. P., Minesinger, B. K., Kumar, J. et Walker, G. C. (2011). A highly conserved protein of unknown function in sinorhizobium meliloti affects srna regulation similar to hfq. *Nucleic Acids Res*, 39(11):4691–4708.

Papenfort, K., Bouvier, M., Mika, F., Sharma, C. M. et Vogel, J. (2010). Evidence for an autonomous 5' target recognition domain in an hfq-associated small rna. *Proc Natl Acad Sci U S A*, 107(47):20435–20440.

Papenfort, K., Podkaminski, D., Hinton, J. C. D. et Vogel, J. (2012). The ancestral sgrs rna discriminates horizontally acquired salmonella mrnas through a single g-u wobble pair. *Proc Natl Acad Sci U S A*, 109(13):E757–E764.

Papenfort, K., Said, N., Welsink, T., Lucchini, S., Hinton, J. C. D. et Vogel, J. (2009). Specific and pleiotropic patterns of mrna regulation by arcz, a conserved, hfq-dependent small rna. *Mol Microbiol*, 74(1):139–158.

Papin, J., Price, N., Wiback, S., Fell, D. et Palsson, B. (2003). Metabolic pathways in the post-genome era. *Trends in Biochemical Sciences*, 28(5):250–258.

Passalacqua, K. D., Varadarajan, A., Ondov, B. D., Okou, D. T., Zwick, M. E. et Bergman, N. H. (2009). Structure and complexity of a bacterial transcriptome. *J Bacteriol*, 191(10):3203–3211.

Patil, K. R., Rocha, I., Förster, J. et Nielsen, J. (2005). Evolutionary programming as a platform for in silico metabolic engineering. *BMC Bioinformatics*, 6:308.

Pearson, W. R. (1991). Searching protein sequence libraries : comparison of the sensitivity and selectivity of the smith-waterman and fasta algorithms. *Genomics*, 11(3):635–650.

Peer, A. et Margalit, H. (2011). Accessibility and evolutionary conservation mark bacterial small-rna target-binding regions. *J Bacteriol*, 193(7):1690–1701.

Petri, C. (1962). *Communicating with Automata*. Thèse de doctorat, PhD thesis, Technical University Darmstadt.

Pfeiffer, V., Sittka, A., Tomer, R., Tedin, K., Brinkmann, V. et Vogel, J. (2007). A small non-coding rna of the invasion gene island (spi-1) represses outer membrane protein synthesis from the salmonella core genome. *Mol Microbiol*, 66(5):1174–1191.

Poolsap, U., Kato, Y. et Akutsu, T. (2010). Dynamic programming algorithms for rna structure prediction with binding sites. *Pac Symp Biocomput*, pages 98–107.

Powers, D. (2007). Evaluation : From precision, recall and f-factor to roc, informedness, markedness & correlation. *School of Informatics and Engineering, Flinders University of South Australia Adelaide*.

Price, N. D., Papin, J. A., Schilling, C. H. et Palsson, B. O. (2003). Genome-scale microbial in silico models : the constraints-based approach. *Trends Biotechnol*, 21(4):162–169.

Prévost, K., Salvail, H., Desnoyers, G., Jacques, J.-F., Phaneuf, E. et Massé, E. (2007). The small rna ryhb activates the translation of shia mrna encoding a permease of shikimate, a compound involved in siderophore synthesis. *Mol Microbiol*, 64(5):1260–1273.

Pulvermacher, S. C., Stauffer, L. T. et Stauffer, G. V. (2009). Role of the srna gcvb in regulation of cyca in escherichia coli. *Microbiology*, 155(Pt 1):106–114.

Pyla, R., Kim, T.-J., Silva, J. L. et Jung, Y.-S. (2009). Overproduction of poly-beta-hydroxybutyrate in the azotobacter vinelandii mutant that does not express small rna arrf. *Appl Microbiol Biotechnol*, 84(4):717–724.

Raghavan, R., Groisman, E. A. et Ochman, H. (2011). Genome-wide detection of novel regulatory rnas in e. coli. *Genome Res*, 21(9):1487–1497.

Rahman, S. A. et Schomburg, D. (2006). Observing local and global properties of metabolic pathways : 'load points' and 'choke points' in the metabolic networks. *Bioinformatics*, 22(14):1767–1774.

Rasmussen, A. A., Eriksen, M., Gilany, K., Udesen, C., Franch, T., Petersen, C. et Valentin-Hansen, P. (2005). Regulation of ompa mrna stability : the role of a small regulatory rna in growth phase-dependent control. *Mol Microbiol*, 58(5):1421–1429.

Reddy, V. N., Mavrovouniotis, M. L. et Liebman, M. N. (1993). Petri net representations in metabolic pathways. *Proc Int Conf Intell Syst Mol Biol*, 1:328–336.

Reed, J. L., Famili, I., Thiele, I. et Palsson, B. O. (2006). Towards multidimensional genome annotation. *Nat Rev Genet*, 7(2):130–141.

Reed, J. L., Vo, T. D., Schilling, C. H. et Palsson, B. O. (2003). An expanded genome-scale model of escherichia coli k-12 (ijr904 gsm/gpr). *Genome Biol*, 4(9):R54.

Rehmsmeier, M., Steffen, P., Hochsmann, M. et Giegerich, R. (2004). Fast and effective prediction of microrna/target duplexes. *RNA*, 10(10):1507–1517.

Reichenbach, B., Maes, A., Kalamorz, F., Hajnsdorf, E. et Görke, B. (2008). The small rna glmy acts upstream of the srna glmz in the activation of glms expression and is subject to regulation by polyadenylation in escherichia coli. *Nucleic Acids Res*, 36(8):2570–2580.

Repoila, F., Majdalani, N. et Gottesman, S. (2003). Small non-coding rnas, co-ordinators of adaptation processes in escherichia coli : the rpos paradigm. *Mol Microbiol*, 48(4):855–861.

Rettner, R. E. et Saier, Jr, M. H. (2010). The autoinducer-2 exporter superfamily. *J Mol Microbiol Biotechnol*, 18(4):195–205.

Rice, J. B. et Vanderpool, C. K. (2011). The small rna sgrs controls sugar-phosphate accumulation by regulating multiple pts genes. *Nucleic Acids Res*, 39(9):3806–3819.

Richter, A. S., Schleberger, C., Backofen, R. et Steglich, C. (2010). Seed-based intarna prediction combined with gfp-reporter system identifies mrna targets of the small rna yfr1. *Bioinformatics*, 26(1):1–5.

Rivas, E. et Eddy, S. R. (2001). Non-coding rna gene detection using comparative sequence analysis. *BMC Bioinformatics*, 2:8.

Rivière, L., van Weelden, S. W. H., Glass, P., Vegh, P., Coustou, V., Biran, M., van Hellemond, J. J., Bringaud, F., Tielens, A. G. M. et Boshart, M. (2004). Acetyl :succinate coa-transferase in procyclic trypanosoma brucei. gene identification and role in carbohydrate metabolism. *J Biol Chem*, 279(44):45337–45346.

Robin, X., Turck, N., Hainard, A., Tiberti, N., Lisacek, F., Sanchez, J.-C. et Müller, M. (2011). proc : an open-source package for r and s+ to analyze and compare roc curves. *BMC Bioinformatics*, 12:77.

Rocha, I., Maia, P., Evangelista, P., Vilaça, P., Soares, S., Pinto, J. P., Nielsen, J., Patil, K. R., Ferreira, E. C. et Rocha, M. (2010). Optflux : an open-source software platform for in silico metabolic engineering. *BMC Syst Biol*, 4:45.

Rohr, C., Marwan, W. et Heiner, M. (2010). Snoopy-a unifying petri net framework to investigate biomolecular networks. *Bioinformatics*.

Romby, P. et Charpentier, E. (2010). An overview of rnas with regulatory functions in gram-positive bacteria. *Cell Mol Life Sci*, 67(2):217–237.

Roth, A. et Breaker, R. R. (2009). The structural and functional diversity of metabolite-binding riboswitches. *Annu Rev Biochem*, 78:305–334.

Said, N., Rieder, R., Hurwitz, R., Deckert, J., Urlaub, H. et Vogel, J. (2009). In vivo expression and purification of aptamer-tagged small rna regulators. *Nucleic Acids Res*, 37(20):e133.

Sakarovitch, J. (2003). *Eléments de théorie des automates*. Vuibert.

Salari, R., Backofen, R. et Sahinalp, S. C. (2010). Fast prediction of rna-rna interaction. *Algorithms Mol Biol*, 5:5.

Sauer, U. (2006). Metabolic networks in motion : 13c-based flux analysis. *Mol Syst Biol*, 2:62.

Scheer, M., Grote, A., Chang, A., Schomburg, I., Munaretto, C., Rother, M., Söhngen, C., Stelzer, M., Thiele, J. et Schomburg, D. (2011). Brenda, the enzyme information system in 2011. *Nucleic acids research*, 39(suppl 1):D670–D676.

Schellenberger, J., Park, J. O., Conrad, T. M. et Palsson, B. O. (2010). Bigg : a biochemical genetic and genomic knowledgebase of large scale metabolic reconstructions. *BMC Bioinformatics*, 11:213.

Schellenberger, J., Que, R., Fleming, R. M. T., Thiele, I., Orth, J. D., Feist, A. M., Zielinski, D. C., Bordbar, A., Lewis, N. E., Rahmanian, S., Kang, J., Hyduke, D. R. et Palsson, B. O. (2011). Quantitative prediction of cellular metabolism with constraint-based models : the cobra toolbox v2.0. *Nat Protoc*, 6(9):1290–1307.

Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S. et Palsson, B. O. (2002). Genome-scale metabolic model of helicobacter pylori 26695. *J Bacteriol*, 184(16):4582–4593.

Schmidt, M., Zheng, P. et Delihas, N. (1995). Secondary structures of escherichia coli antisense micf rna, the 5'-end of the target ompf mrna, and the rna/rna duplex. *Biochemistry*, 34(11):3621–3631.

Schuster, S., Dandekar, T. et Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks : a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2):53–60.

Segrè, D., Vitkup, D. et Church, G. M. (2002). Analysis of optimality in natural and perturbed metabolic networks. *Proc Natl Acad Sci U S A*, 99(23):15112–15117.

Shao, Y. et Bassler, B. L. (2012). Quorum-sensing non-coding small rnas use unique pairing regions to differentially control mrna targets. *Mol Microbiol*, 83(3):599–611.

Shapiro, J. A. (2009). Revisiting the central dogma in the 21st century. *Ann N Y Acad Sci*, 1178:6–28.

Sharan, R. et Ideker, T. (2006). Modeling cellular machinery through biological network comparison. *Nat Biotechnol*, 24(4):427–433.

Sharma, C. M., Darfeuille, F., Plantinga, T. H. et Vogel, J. (2007). A small rna regulates multiple abc transporter mrnas by targeting c/a-rich elements inside and upstream of ribosome-binding sites. *Genes Dev*, 21(21):2804–2817.

Sharma, C. M., Hoffmann, S., Darfeuille, F., Reignier, J., Findeiss, S., Sittka, A., Chabas, S., Reiche, K., Hackermüller, J., Reinhardt, R., Stadler, P. F. et Vogel, J. (2010). The primary transcriptome of the major human pathogen helicobacter pylori. *Nature*, 464(7286):250–255.

Sharma, C. M. et Vogel, J. (2009). Experimental approaches for the discovery and characterization of regulatory small rna. *Curr Opin Microbiol*, 12(5):536–546.

Shimoni, Y., Friedlander, G., Hetzroni, G., Niv, G., Altuvia, S., Biham, O. et Margalit, H. (2007). Regulation of gene expression by small non-coding rnas : a quantitative view. *Mol Syst Biol*, 3:138.

Shine, J. et Dalgarno, L. (1974). The 3'-terminal sequence of escherichia coli 16s ribosomal rna : complementarity to nonsense triplets and ribosome binding sites. *Proc Natl Acad Sci U S A*, 71(4):1342–1346.

Shinhara, A., Matsui, M., Hiraoka, K., Nomura, W., Hirano, R., Nakahigashi, K., Tomita, M., Mori, H. et Kanai, A. (2011). Deep sequencing reveals as-yet-undiscovered small rnas in escherichia coli. *BMC Genomics*, 12:428.

Shlomi, T., Berkman, O. et Ruppin, E. (2005). Regulatory on/off minimization of metabolic flux changes after genetic perturbations. *Proc Natl Acad Sci U S A*, 102(21):7695–7700.

Sittka, A., Sharma, C. M., Rolle, K. et Vogel, J. (2009). Deep sequencing of salmonella rna associated with heterologous hfq proteins in vivo reveals small rnas as a major target class and identifies rna processing phenotypes. *RNA Biol*, 6(3):266–275.

Smith, A. C. et Robinson, A. J. (2011). A metabolic model of the mitochondrion and its use in modelling diseases of the tricarboxylic acid cycle. *BMC Syst Biol*, 5:102.

Smoot, M. E., Ono, K., Ruscheinski, J., Wang, P.-L. et Ideker, T. (2011). Cytoscape 2.8 : new features for data integration and network visualization. *Bioinformatics*, 27(3):431–432.

Sorek, R., Kunin, V. et Hugenholtz, P. (2008). Crispr–a widespread system that provides acquired resistance against phages in bacteria and archaea. *Nat Rev Microbiol*, 6(3):181–186.

Staple, D. W. et Butcher, S. E. (2005). Pseudoknots : Rna structures with diverse functions. *PLoS Biol*, 3(6):e213.

Stapley, B. J. et Benoit, G. (2000). Biobibliometrics : information retrieval and visualization from co-occurrences of gene names in medline abstracts. *Pac Symp Biocomput*, pages 529–540.

Stelling, J. (2004). Mathematical models in microbial systems biology. *Curr Opin Microbiol*, 7(5):513–518.

Stork, M., Di Lorenzo, M., Welch, T. J. et Crosa, J. H. (2007). Transcription termination within the iron transport-biosynthesis operon of vibrio anguillarum requires an antisense rna. *J Bacteriol*, 189(9):3479–3488.

Storz, G. (2002). An expanding universe of noncoding rnas. *Science*, 296(5571):1260–1263.

Storz, G., Vogel, J. et Wassarman, K. M. (2011). Regulation by small rnas in bacteria : expanding frontiers. *Mol Cell*, 43(6):880–891.

Sun, Y. et Vanderpool, C. K. (2011). Regulation and function of escherichia coli sugar efflux transporter a (seta) during glucose-phosphate stress. *J Bacteriol*, 193(1):143–153.

Sung, B. H., Lee, C. H., Yu, B. J., Lee, J. H., Lee, J. Y., Kim, M. S., Blattner, F. R. et Kim, S. C. (2006). Development of a biofilm production-deficient escherichia coli strain as a host for biotechnological applications. *Appl Environ Microbiol*, 72(5):3336–3342.

Surette, M. G., Miller, M. B. et Bassler, B. L. (1999). Quorum sensing in escherichia coli, salmonella typhimurium, and vibrio harveyi : a new family of genes responsible for autoinducer production. *Proc Natl Acad Sci U S A*, 96(4):1639–1644.

Suthers, P. F., Dasika, M. S., Kumar, V. S., Denisov, G., Glass, J. I. et Maranas, C. D. (2009). A genome-scale metabolic reconstruction of mycoplasma genitalium, ips189. *PLoS Comput Biol*, 5(2):e1000285.

Swartzman, E., Silverman, M. et Meighen, E. A. (1992). The luxr gene product of vibrio harveyi is a transcriptional activator of the lux promoter. *J Bacteriol*, 174(22):7490–7493.

Tafer, H. et Hofacker, I. L. (2008). Rnaplex : a fast tool for rna-rna interaction search. *Bioinformatics*, 24(22):2657–2663.

Teusink, B., Walsh, M. C., van Dam, K. et Westerhoff, H. V. (1998). The danger of metabolic pathways with turbo design. *Trends Biochem Sci*, 23(5):162–169.

Thiele, I. et Palsson, B. O. (2010). A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc*, 5(1):93–121.

Thiele, I., Vo, T. D., Price, N. D. et Palsson, B. O. (2005). Expanded metabolic reconstruction of helicobacter pylori (iit341 gsm/gpr) : an in silico genome-scale characterization of single- and double-deletion mutants. *J Bacteriol*, 187(16):5818–5830.

Tjaden, B., Goodwin, S. S., Opdyke, J. A., Guillier, M., Fu, D. X., Gottesman, S. et Storz, G. (2006). Target prediction for small, noncoding rnas in bacteria. *Nucleic Acids Res*, 34(9):2791–2802.

Tjaden, B., Saxena, R. M., Stolyar, S., Haynor, D. R., Kolker, E. et Rosenow, C. (2002). Transcriptome analysis of escherichia coli using high-density oligonucleotide probe arrays. *Nucleic Acids Res*, 30(17):3732–3738.

Toledo-Arana, A., Dussurget, O., Nikitas, G., Sesto, N., Guet-Revillet, H., Balestrino, D., Loh, E., Gripenland, J., Tiensuu, T., Vaitkevicius, K., Barthelemy, M., Vergassola, M., Nahori, M.-A., Soubigou, G., Régnault, B., Coppée, J.-Y., Lecuit, M., Johansson, J. et Cossart, P. (2009). The listeria transcriptional landscape from saprophytism to virulence. *Nature*, 459(7249):950–956.

Tompa, M., Li, N., Bailey, T. L., Church, G. M., De Moor, B., Eskin, E., Favorov, A. V., Frith, M. C., Fu, Y., Kent, W. J., Makeev, V. J., Mironov, A. A., Noble, W. S., Pavesi, G., Pesole, G., Régnier, M., Simonis, N., Sinha, S., Thijs, G., van Helden, J., Vandenbogaert, M., Weng, Z., Workman, C., Ye, C. et Zhu, Z. (2005). Assessing computational tools for the discovery of transcription factor binding sites. *Nat Biotechnol*, 23(1):137–144.

Tran, T. T., Zhou, F., Marshburn, S., Stead, M., Kushner, S. R. et Xu, Y. (2009). De novo computational prediction of non-coding rna genes in prokaryotic genomes. *Bioinformatics*, 25(22):2897–2905.

Tsui, H. C., Leung, H. C. et Winkler, M. E. (1994). Characterization of broadly pleiotropic phenotypes caused by an hfq insertion mutation in escherichia coli k-12. *Mol Microbiol*, 13(1):35–49.

Udekwu, K. I., Darfeuille, F., Vogel, J., Reimegård, J., Holmqvist, E. et Wagner, E. G. H. (2005). Hfq-dependent regulation of ompa synthesis is mediated by an antisense rna. *Genes Dev*, 19(19):2355–2366.

Unoson, C. et Wagner, E. G. H. (2008). A small sos-induced toxin is targeted against the inner membrane in escherichia coli. *Mol Microbiol*, 70(1):258–270.

Urban, J. H. et Vogel, J. (2007). Translational control and target recognition by escherichia coli small rnas in vivo. *Nucleic Acids Res*, 35(3):1018–1037.

Valette, R. (2007). Introduction aux réseaux de petri.

Van Dongen, S. (2000). A cluster algorithm for graphs. *Report-Information systems*, (10):1–40.

Van Noorden, C. J. F. (2010). Imaging enzymes at work : metabolic mapping by enzyme histochemistry. *J Histochem Cytochem*, 58(6):481–497.

van Weelden, S. W. H., Fast, B., Vogt, A., van der Meer, P., Saas, J., van Hellemond, J. J., Tielens, A. G. M. et Boshart, M. (2003). Procyclic trypanosoma brucei do not use krebs cycle activity for energy generation. *J Biol Chem*, 278(15):12854–12863.

Vanderpool, C. K., Balasubramanian, D. et Lloyd, C. R. (2011). Dual-function rna regulators in bacteria. *Biochimie*, 93(11):1943–1949.

Vanderpool, C. K. et Gottesman, S. (2004). Involvement of a novel transcriptional activator and small rna in post-transcriptional regulation of the glucose phosphoenolpyruvate phosphotransferase system. *Mol Microbiol*, 54(4):1076–1089.

Varma, A. et Palsson, B. O. (1994). Stoichiometric flux balance models quantitatively predict growth and metabolic by-product secretion in wild-type escherichia coli w3110. *Appl Environ Microbiol*, 60(10):3724–3731.

Vecerek, B., Moll, I. et Bläsi, U. (2007). Control of fur synthesis by the non-coding rna ryhb and iron-responsive decoding. *EMBO J*, 26(4):965–975.

Vendeville, A., Winzer, K., Heurlier, K., Tang, C. M. et Hardie, K. R. (2005). Making 'sense' of metabolism : autoinducer-2, luxs and pathogenic bacteria. *Nat Rev Microbiol*, 3(5):383–396.

Vilar, J. M. G., Guet, C. C. et Leibler, S. (2003). Modeling network dynamics : the lac operon, a case study. *J Cell Biol*, 161(3):471–476.

Vogel, J., Argaman, L., Wagner, E. G. H. et Altuvia, S. (2004). The small rna istr inhibits synthesis of an sos-induced toxic peptide. *Curr Biol*, 14(24):2271–2276.

Vogel, J. et Sharma, C. M. (2005). How to find small non-coding rnas in bacteria. *Biol Chem*, 386(12):1219–1238.

Wadler, C. S. et Vanderpool, C. K. (2007). A dual function for a bacterial small rna : Sgrs performs base pairing-dependent regulation and encodes a functional polypeptide. *Proc Natl Acad Sci U S A*, 104(51):20454–20459.

Wagner, E. G. H., Altuvia, S. et Romby, P. (2002). Antisense rnas in bacteria and their genetic elements. *Adv Genet*, 46:361–398.

Wang, L. (2004). *Autoinducer-2 (AI-2) mediated quorum sensing in Escherichia coli*. Thèse de doctorat, University of Maryland, College Park.

Washietl, S., Hofacker, I. L. et Stadler, P. F. (2005). Fast and reliable prediction of noncoding rnas. *Proc Natl Acad Sci U S A*, 102(7):2454–2459.

Waters, L. S. et Storz, G. (2009). Regulatory rnas in bacteria. *Cell*, 136(4):615–628.

Westhof, E. et Fritsch, V. (2000). Rna folding : beyond watson-crick pairs. *Structure*, 8(3):R55–R65.

Wiechert, W. (2001). 13c metabolic flux analysis. *Metabolic Engineering*, 3(3):195–206.

Wilderman, P. J., Sowa, N. A., FitzGerald, D. J., FitzGerald, P. C., Gottesman, S., Ochsner, U. A. et Vasil, M. L. (2004). Identification of tandem duplicate regulatory small rnas in pseudomonas aeruginosa involved in iron homeostasis. *Proc Natl Acad Sci U S A*, 101(26):9792–9797.

Wilkinson, D. (2006). *Stochastic modelling for systems biology*, volume 44. CRC press.

Williams, P., Winzer, K., Chan, W. C. et Cámara, M. (2007). Look who's talking : communication and quorum sensing in the bacterial world. *Philos Trans R Soc Lond B Biol Sci*, 362(1483):1119–1134.

Winzer, K., Sun, Y.-h., Green, A., Delory, M., Blackley, D., Hardie, K. R., Baldwin, T. J. et Tang, C. M. (2002). Role of neisseria meningitidis luxs in cell-to-cell signaling and bacteremic infection. *Infect Immun*, 70(4):2245–2248.

Wood, T. K., González Barrios, A. F., Herzberg, M. et Lee, J. (2006). Motility influences biofilm architecture in escherichia coli. *Appl Microbiol Biotechnol*, 72(2):361–367.

Wuchty, S., Fontana, W., Hofacker, I. L. et Schuster, P. (1999). Complete suboptimal folding of rna and the stability of secondary structures. *Biopolymers*, 49(2):145–165.

Wuju, L. et Momiao, X. (2002). Tclass : tumor classification system based on gene expression profile. *Bioinformatics*, 18(2):325–326.

Yachie, N., Numata, K., Saito, R., Kanai, A. et Tomita, M. (2006). Prediction of non-coding and antisense rna genes in escherichia coli with gapped markov model. *Gene*, 372: 171–181.

Yao, Y., Martinez-Yamout, M. A., Dickerson, T. J., Brogan, A. P., Wright, P. E. et Dyson, H. J. (2006). Structure of the escherichia coli quorum sensing protein sdia : activation of the folding switch by acyl homoserine lactones. *J Mol Biol*, 355(2):262–273.

Yeh, I., Hanekamp, T., Tsoka, S., Karp, P. D. et Altman, R. B. (2004). Computational analysis of plasmodium falciparum metabolism : organizing genomic information to facilitate drug discovery. *Genome Res*, 14(5):917–924.

Ying, X., Cao, Y., Wu, J., Liu, Q., Cha, L. et Li, W. (2011). starpicker : A method for efficient prediction of bacterial srna targets based on a two-step model for hybridization. *PLoS One*, 6(7):e22705.

Yizhak, K., Benyamini, T., Liebermeister, W., Ruppin, E. et Shlomi, T. (2010). Integrating quantitative proteomics and metabolomics with a genome-scale metabolic network model. *Bioinformatics*, 26(12):i255–i260.

Yoder-Himes, D. R., Chain, P. S. G., Zhu, Y., Wurtzel, O., Rubin, E. M., Tiedje, J. M. et Sorek, R. (2009). Mapping the burkholderia cenocepacia niche response via high-throughput sequencing. *Proc Natl Acad Sci U S A*, 106(10):3976–3981.

Yusupova, G., Jenner, L., Rees, B., Moras, D. et Yusupov, M. (2006). Structural basis for messenger rna movement on the ribosome. *Nature*, 444(7117):391–394.

Zhang, A., Wassarman, K. M., Rosenow, C., Tjaden, B. C., Storz, G. et Gottesman, S. (2003). Global analysis of small rna and mrna targets of hfq. *Mol Microbiol*, 50(4):1111–1124.

Zhang, Y., Sun, S., Wu, T., Wang, J., Liu, C., Chen, L., Zhu, X., Zhao, Y., Zhang, Z., Shi, B., Lu, H. et Chen, R. (2006). Identifying hfq-binding small rna targets in escherichia coli. *Biochem Biophys Res Commun*, 343(3):950–955.

Zhang, Y., Zhang, Z., Ling, L., Shi, B. et Chen, R. (2004). Conservation analysis of small rna genes in escherichia coli. *Bioinformatics*, 20(5):599–603.

Zhao, Y., Li, H., Hou, Y., Cha, L., Cao, Y., Wang, L., Ying, X. et Li, W. (2008). Construction of two mathematical models for prediction of bacterial srna targets. *Biochem Biophys Res Commun*, 372(2):346–350.

Zuker, M. (1994). Prediction of rna secondary structure by energy minimization. *Methods Mol Biol*, 25:267–294.

Zuker, M. et Stiegler, P. (1981). Optimal computer folding of large rna sequences using thermodynamics and auxiliary information. *Nucleic Acids Res*, 9(1):133–148.

Annexes

Annexe A

Données de l'étude de comparaison des logiciels de prédiction

TABLE A.1 – Liste des interactions identifiées expérimentalement.

Pour chaque interaction, nous indiquons le numéro d'accession Genbank du génome, le type de régulation, ainsi que la position de l'interaction sur le sRNA (BP sRNA) et sur le mRNA (BP mRNA) déterminée expérimentalement.

Accession	Génome	sRNA	mRNA	Régulation	BP sRNA	BP mRNA	Références
NC_012560	A. vinelandii	ArrF	phbR	Activation	(75,105)	(-34,-4)	[Pyla *et al.* 2009]
NC_012560	A. vinelandii	ArrF	sodB	Activation	(74,107)	(-42,-8)	[Jung et Kwon 2008]
NC_000913	E. coli	CyaR	luxS	Repression	(35,49)	(-12,3)	[De Lay et Gottesman 2009]
NC_000913	E. coli	CyaR	nadE	Repression	(35,48)	(-11,3)	[De Lay et Gottesman 2009]
NC_000913	E. coli	CyaR	ompX	Repression	(38,48)	(-9,2)	[De Lay et Gottesman 2009]
NC_000913	E. coli	CyaR	yqaE	Repression	(31,50)	(-4,16)	[De Lay et Gottesman 2009]
NC_000913	E. coli	DsrA	hns	Repression	(31,43)	(7,19)	[Lease *et al.* 1998]
NC_000913	E. coli	DsrA	rpoS	Activation	(10,32)	(-119,-97)	[Repoila *et al.* 2003]
NC_000913	E. coli	FnrS	metE	Repression	(37,67)	(-18,13)	[Boysen *et al.* 2010]
NC_000913	E. coli	FnrS	sodA	Repression	(11,47)	(-19,16)	[Boysen *et al.* 2010]
NC_000913	E. coli	FnrS	sodB	Repression	(40,74)	(-30,4)	[Boysen *et al.* 2010]
NC_000913	E. coli	GcvB	cycA	Repression	(125,161)	(-26,7)	[Pulvermacher *et al.* 2009]
NC_000913	E. coli	GcvB	sstT	Repression	(64,99)	(-34,2)	[Pulvermacher *et al.* 2009]
NC_000913	E. coli	GlmZ	glmS	Activation	(150,169)	(-40,-22)	[Reichenbach *et al.* 2008]
NC_000913	E. coli	IstR	tisB	Repression	(57,95)	(-135,-92)	[Vogel *et al.* 2004]
NC_000913	E. coli	MicA	phoP	Repression	(6,31)	(-15,8)	[Coornaert *et al.* 2010]
NC_000913	E. coli	MicC	ompC	Repression	(1,30)	(-41,-15)	[Chen *et al.* 2004]
NC_000913	E. coli	MicF	ompF	Repression	(1,33)	(-16,10)	[Schmidt *et al.* 1995]
NC_011601	E. coli	OmrA	cirA	Repression	(2,24)	(-35,-10)	[Guillier et Gottesman 2008]
NC_000913	E. coli	OmrA	csgD	Repression	(2,20)	(-79,-61)	[Holmqvist *et al.* 2010]
NC_011601	E. coli	OmrA	ompR	Repression	(1,19)	(-29,-11)	[Guillier et Gottesman 2008]
NC_011601	E. coli	OmrA	ompT	Repression	(1,33)	(-12,20)	[Guillier et Gottesman 2008]
NC_011601	E. coli	OmrB	cirA	Repression	(2,24)	(-35,-10)	[Guillier et Gottesman 2008]

Suite page suivante ...

Accession	Génome	sRNA	mRNA	Régulation	BP sRNA	BP mRNA	Références
NC_000913	E. coli	OmrB	csgD	Repression	(2,20)	(-79,-61)	[Holmqvist *et al.* 2010]
NC_011601	E. coli	OmrB	ompR	Repression	(1,19)	(-29,-11)	[Guillier et Gottesman 2008]
NC_011601	E. coli	OmrB	ompT	Repression	(1,33)	(-12,20)	[Guillier et Gottesman 2008]
NC_000913	E. coli	OxyS	fhlA	Repression	(98,104)	(-15,-9)	[Argaman et Altuvia 2000]
NC_000913	E. coli	RprA	rpoS	Activation	(33,62)	(-117,-94)	[Mandin et Gottesman 2010]
NC_000913	E. coli	RseX	ompA	Repression	(37,50)	(-22,-8)	[Douchin *et al.* 2006]
NC_000913	E. coli	RseX	ompC	Repression	(30,55)	(-31,-1)	[Douchin *et al.* 2006]
NC_000913	E. coli	RybB	ompC	Repression	(1,30)	(-53,-4)	[Johansen *et al.* 2006]
NC_000913	E. coli	RybB	ompW	Repression	(1,31)	(-13,20)	[Muckstein *et al.* 2006]
NC_000913	E. coli	RyhB	fur	Repression	(38,76)	(-96,-47)	[Vecerek *et al.* 2007]
NC_000913	E. coli	RyhB	iscS	Repression	(43,68)	(-26,-1)	[Desnoyers *et al.* 2009]
NC_000913	E. coli	RyhB	sdhD	Repression	(9,50)	(-42,-3)	[Massé et Gottesman 2002]
NC_000913	E. coli	RyhB	shiA	Activation	(19,75)	(-77,-27)	[Prévost *et al.* 2007]
NC_000913	E. coli	RyhB	sodB	Repression	(34,64)	(-17,9)	[Geissmann et Touati 2004]
NC_000913	E. coli	SgrS	ptsG	Repression	(157,187)	(-28,4)	[Kawamoto *et al.* 2006]
NC_011601	E. coli	SPOT42	galK	Repression	(1,62)	(-20,56)	[Møller *et al.* 2002]
NC_003210	L. monocytogenes	LhrA	lmo0850	Repression	(152,165)	(-35,-22)	[Nielsen *et al.* 2010]
NC_002516	P. aeruginosa	PrrF1	PA4880	Repression	(85,118)	(-26,7)	[Wilderman *et al.* 2004]
NC_002516	P. aeruginosa	PrrF1	sodB	Repression	(84,118)	(-40,-6)	[Wilderman *et al.* 2004]
NC_002516	P. aeruginosa	PrrF2	PA4880	Repression	(82,116)	(-26,7)	[Wilderman *et al.* 2004]
NC_002516	P. aeruginosa	PrrF2	sodB	Repression	(82,116)	(-40,-6)	[Wilderman *et al.* 2004]
NC_005072	P. marinus	Yfr1	PMM1119	Repression	(19,28)	(-3,7)	[Richter *et al.* 2010]
NC_005072	P. marinus	Yfr1	PMM1121	Repression	(20,28)	(-3,6)	[Richter *et al.* 2010]
NC_003197	S. enterica	ArcZ	sdaC	Repression	(62,71)	(-13,-3)	[Papenfort *et al.* 2009]
NC_003197	S. enterica	ArcZ	STM3216	Repression	(63,87)	(-25,-5)	[Papenfort *et al.* 2009]
NC_003197	S. enterica	ArcZ	tpx	Repression	(66,83)	(10,26)	[Mandin et Gottesman 2010]
NC_003197	S. enterica	ChiX	ybfM	Repression	(42,53)	(-19,-8)	[Figueroa-Bossi *et al.* 2009]
NC_003197	S. enterica	GcvB	argT	Repression	(70,91)	(-57,-37)	[Sharma *et al.* 2007]

Suite page suivante ...

Accession	Génome	sRNA	mRNA	Régulation	BP sRNA	BP mRNA	Références
NC_003197	S. enterica	GcvB	dppA	Repression	(65,82)	(-31,-14)	[Sharma *et al.* 2007]
NC_003197	S. enterica	GcvB	gltI	Repression	(66,76)	(-38,-27)	[Sharma *et al.* 2007]
NC_003197	S. enterica	GcvB	livJ	Repression	(63,87)	(-51,-28)	[Sharma *et al.* 2007]
NC_003197	S. enterica	GcvB	livK	Probing	(62,88)	(-39,-15)	[Sharma *et al.* 2007]
NC_003197	S. enterica	GcvB	oppA	Repression	(65,89)	(-8,16)	[Sharma *et al.* 2007]
NC_003197	S. enterica	GcvB	STM4351	Repression	(62,87)	(-41,-16)	[Sharma *et al.* 2007]
NC_003197	S. enterica	InvR	nmpC	Repression	(34,82)	(15,65)	[Pfeiffer *et al.* 2007]
NC_003197	S. enterica	MicA	lamB	Repression	(14,42)	(-9,18)	[Bossi et Figueroa-Bossi 2007]
NC_003198	S. enterica	MicA	ompA	Repression	(9,25)	(-21,-6)	[Udekwu *et al.* 2005]
NC_003197	S. enterica	MicC	nmpC	Repression	(1,12)	(67,78)	[Pfeiffer *et al.* 2007]
NC_003197	S. enterica	RybB	ompN	Repression	(1,16)	(5,20)	[Bouvier *et al.* 2008]
NC_000913	E. coli	DsrA	ompA	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	DsrA	ompC	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	DsrA	ompF	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	DsrA	ptsG	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	GcvB	hns	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	GcvB	ompA	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	GcvB	ompC	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	GcvB	ompF	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	GcvB	ptsG	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	GcvB	sodB	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicA	hns	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicA	ompC	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicA	ompF	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicA	ptsG	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicA	sodB	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicC	hns	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicC	ompA	Non interaction	-	-	[Urban et Vogel 2007]

Suite page suivante ...

Accession	Génome	sRNA	mRNA	Régulation	BP sRNA	BP mRNA	Références
NC_000913	E. coli	MicC	ompF	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicC	ptsG	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicC	sodB	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicF	hns	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicF	ompA	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicF	ompC	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	MicF	ptsG	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	OmrA	ompA	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	OmrB	ompA	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RprA	hns	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RprA	ompA	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RprA	ompC	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RprA	ompF	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RprA	sodB	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RyhB	hns	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RyhB	ompA	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RyhB	ompC	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RyhB	ompF	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	RyhB	ptsG	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	SgrS	hns	Non interaction	-	-	[Urban et Vogel 2007]
NC_000913	E. coli	SgrS	ompA	Non interaction	-	-	[Urban et Vogel 2007]

TABLE A.2 – Informations générales sur les séquences des jeux de données de test et d'*E. coli*.

Données	Type	Nombre de séquences	GC(%)	Longueur moyenne (nt)
Test	sRNAs	43	41.7	141
	mRNA	52	42.6	200
E. coli	sRNAs	261	44.8	119
	mRNA	4142	45.7	200

TABLE A.3 – Lignes de commande employées au sein de iRNA pour tester les différents logiciels.
Les chemins vers les fichiers fasta des sRNAs et des mRNAs sont indiqués par les marques : %sRNA et %mRNA. Les marques %couple %couple-2 réfèrent à des fichiers comportant les deux séquences respectivement au format fasta ou sous la forme d'une seule séquence séparant les deux éléments par un liguant "&".

	Ligne de commande
1	IntaRNA -t %mRNA -m %sRNA -o -s 4 -w 140
2	IntaRNA -t %mRNA -m %sRNA -o -s 4
3	RNAcofold -a -p -d2 -noPS -noLP < %couples-2
4	RNAduplex -noPS < %couples
5	RNAhybrid -t %mRNA -q %sRNA -m 10000 -n 10000 -s 3utr_human
6	RNAhybrid -t %mRNA -q %sRNA -m 10000 -n 10000 -d 1
7	RNAplex -e -10 -c 30 -l 20 -t %mRNA -q %sRNA
8	RNAplex -t %mRNA -q %sRNA
9	RNAup -b -d2 -noLP < %couples
10	bistarna %couples
11	bistarna -w 5 %couples
12	blastall -p blastn -W 4 -e 10000 -S 2 -m 8 -d %mf_mRNA -i %sRNA
13	blastall -p blastn -W 4 -e 10000 -S 2 -m 8 -d %mf_mRNA -i %sRNA -r 4 -q 5 -G 11 -E 11
14	guugle -d 7 %mRNA %sRNA
15	pairfold %sRNA %mRNA
16	ractip %mRNA %sRNA
17	ractip -m %mRNA %sRNA
18	ssearch35_t -Q -B -H -m 9 -E 10000 -d 0 -f 0 -g -11 -i -n -a -s dna_matrix.mat %sRNA %mf_mRNA
19	yass-Linux64.bin -d 2 -r 1 -G -11,-11 -E 10000 -C 5,-11 -p "##-#,###" %sRNA %mRNA

Annexe B

Résultat de prédiction des logiciels

TABLE B.1 – Résultats de prédiction des vraies- et non-interactions.
Les mesures ici présentées ont été calculées à partir des courbes de ROC obtenues pour les résultats de prédiction du jeu de données de test en utilisant le package R : pROC [Robin et al. *2011] (voir Section 5.2).*

Classement	Logiciels	Sensibilité	Spécificité	Index de Youden	AUC
1	IntaRNA_1	0,79	0,85	-0,83	0,88
2	RNAup_9	0,63	0,95	-0,96	0,83
3	IntaRNA_2	0,73	0,90	-0,83	0,83
4	guugle_14	0,61	0,90	4,48	0,77
5	ssearch35_t_18	0,68	0,78	1,48	0,76
6	RNAplex_7	0,55	0,93	-2,07	0,74
7	RNAplex_8	0,87	0,54	-1,57	0,72
8	blastall_12	0,58	0,80	9,75	0,67
9	blastall_13	0,58	0,80	9,75	0,67
10	RNAduplex_4	0,69	0,59	-4,68	0,65
11	bistarna_11	0,48	0,78	1,50	0,62
12	RNAhybrid_5	0,66	0,61	-5,38	0,61
13	RNAhybrid_6	0,66	0,61	-5,38	0,61
14	ractip_17	0,89	0,32	6,38	0,61
15	pairfold_15	0,45	0,73	-9,68	0,58
16	RNAcofold_3	0,44	0,73	-10,67	0,57
17	ractip_16	0,65	0,46	25,63	0,52
18	bistarna_10	0,44	0,66	1,32	0,47
19	Yass-Linux64.bin_19*	0,95	0,80	0,16	0,92

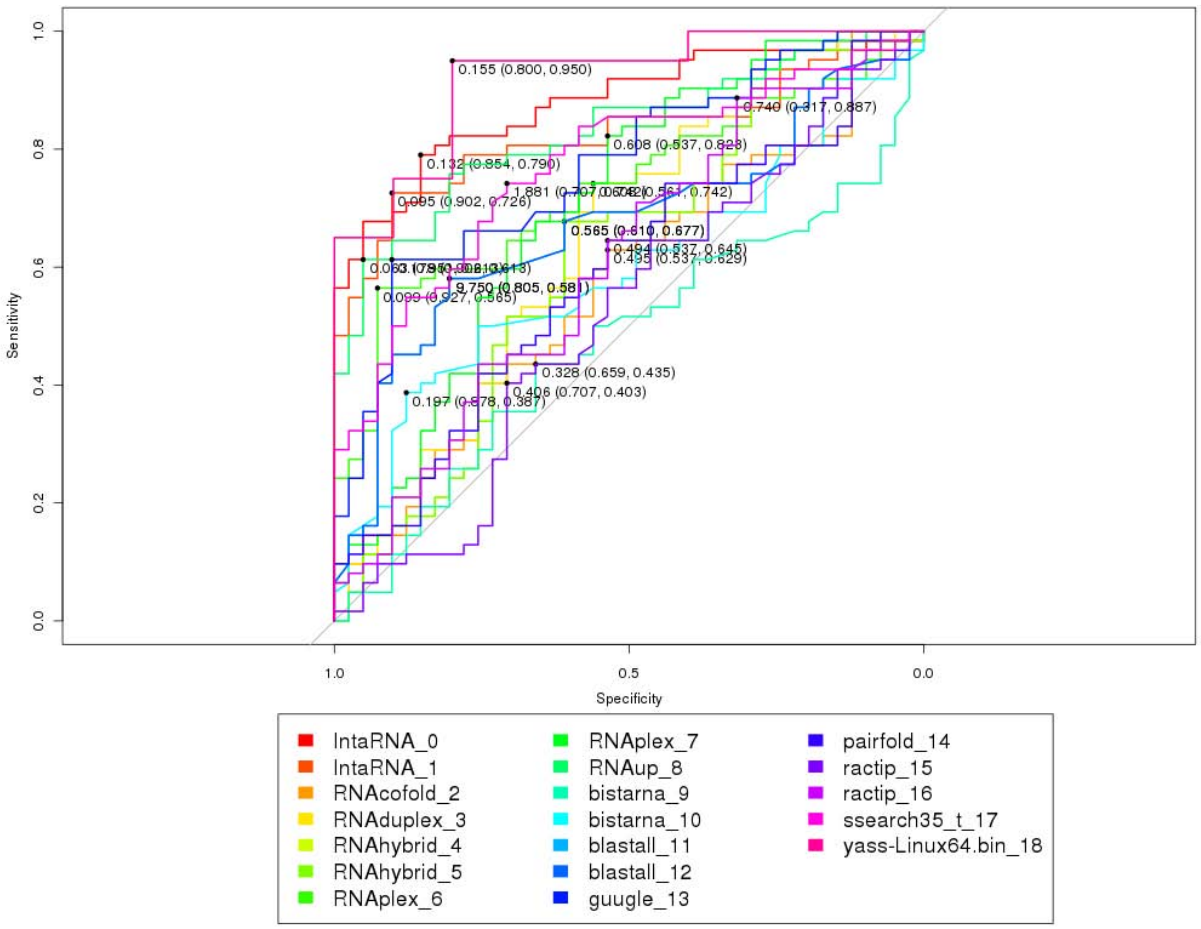


FIGURE B.2 – Comparaison de la capacité des logiciels à prédire des vraies- et non-interactions par courbe de roc.
Les courbes de ROC ici présentées ont été calculées à partir des résultats de prédiction du jeu de données de test à l'aide du package R : pROC (voir Section 5.2).

TABLE B.3 – Résultats de prédiction de la zone d'interaction des vraies interactions.
Les mesures ici présentées ont été calculées à partir des résultats de prédiction du jeu de données de test (voir Section 5.2).

Logiciels	Sensibilité moyenne	PPV moyen
IntaRNA_1	0.585	0.678
IntaRNA_2	0.583	0.746
RNAcofold_3	0.589	0.488
RNAduplex_4	0.950	0.199
RNAhybrid_5	0.706	0.269
RNAhybrid_6	0.706	0.269
RNAplex_7	0.379	0.269
RNAplex_8	0.404	0.573
RNAup_9	0.361	0.611
bistarna_10	0.022	0.236
bistarna_11	0.034	0.227
blastall_12	0.264	0.600
blastall_13	0.264	0.600
guugle_14	0.332	0.644
pairfold_15	0.602	0.502
ractip_16	0.713	0.294
ractip_17	0.349	0.369
ssearch35_t_18	0.300	0.410
yass-Linux64.bin_19	0.520	0.784

TABLE B.4 – Caractéristiques des graphes de prédiction obtenus.
Le graphe des interactions correspond au graphe initial dans iRNA pour les données de E. coli. Il correspond aux interactions prédites par IntaRNA_1 comme de vraies interactions. Différents sous-graphes sont ensuite obtenus à partir de ces données, après l'enrichissement des cibles par DAVID, puis le filtrage par iRNA_visu des interactions obtenant effectivement des enrichissements de l'une des 10 bases de données considérées par DAVID.

Source	sRNAs	mRNAs	Nombre Interactions
Graphe des interactions	261	4142	199461
Graphe enrichi	261	3713	174258
Gene Ontology Biological Pathways	260	2648	107170
Gene Ontology Metabolic Functions	259	2259	75188
Gene Ontology Cellular Component	260	1593	71374
Uniprot	258	2403	72210
KEGG	258	1373	60239
Swissprot	261	3562	170640
Interpro	255	2230	31254
SMART	197	412	5053
COG	180	326	4052
PIR	240	792	3780