Identification de motifs dans les réseaux métaboliques

Vincent Lacroix

N° d'ordre 202-2007 Année 2007

Thèse

Présentée

devant l'Université Claude Bernard - Lyon 1

pour l'obtention

du Diplôme de Doctorat

(arrêté du 7 août 2006)

et soutenue publiquement le
26 octobre 2007

par

Vincent Lacroix

Identification de motifs dans les réseaux métaboliques

Définitions, algorithmes, et application au métabolisme d'*Escherichia coli*

Directrice de thèse : Marie-France Sagot

Jury : Christian Gautier
Michel Habib
Stéphane Robin
Marie-France Sagot
Stefan Schuster
Leen Stougie
Dominique de Vienne

UNIVERSITÉ CLAUDE BERNARD-LYON 1

Président de l'Université	**M. le Professeur L. COLLET**
Vice-Président du Conseil Scientifique	M. le Professeur J. F. MORNEX
Vice-Président du Conseil d'Administration	M. le Professeur J. LIETO
Vice-Président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. SIMON
Secrétaire Général	**M. G. GAY**

SECTEUR SANTÉ

Composantes

UFR de Médecine Lyon R.T.H. Laënnec	Directeur : M. le Professeur D. VITAL-DURAND
UFR de Médecine Lyon Grange-Blanche	Directeur : M. le Professeur X. MARTIN
UFR de Médecine Lyon-Nord	Directeur : M. le Professeur F. MAUGUIERE
UFR de Médecine Lyon-Sud	Directeur : M. le Professeur F.N. GILLY
UFR d'Ontologie	Directeur : M. O. ROBIN
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. LOCHER
Institut Techniques de Réadaptation	Directeur : M. le Professeur MATILLON
Département de Formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. FARGE

SECTEUR SCIENCES

Composantes

UFR de Physique	Directeur : M. le Professeur A. HOAREAU
UFR de Biologie	Directeur : M. le Professeur H. PINON
UFR de Mécanique	Directeur : M. le Professeur H. BEN HADID
UFR de Génie Electrique et des Procédés	Directeur : M. le Professeur A. BRIGUET
UFR de Sciences de la Terre	Directeur : M. le Professeur P. HANTZPERGUE
UFR de Mathématique	Directeur : M. le Professeur M. CHAMARIE
UFR d'Informatique	Directeur : M. le Professeur M. EGEA
UFR de Chimie Biochimie	Directeur : Mme. le Professeur H. PARROT
UFR STAPS	Directeur : M. le Professeur R. MASSARELLI
Observatoire de Lyon	Directeur : M. le Professeur R. BACON
Institut des Sciences et des Techniques de l'Ingénieur de Lyon	Directeur : M. le Professeur J. LIETO
IUT A	Directeur : M. le Professeur M. C. COULET
IUT B	Directeur : M. le Professeur R. LAMARTINE
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J. C. AUGROS

Remerciements

Je voudrais remercier ici tous ceux qui ont compté pour moi pendant ces trois années et qui ont su me donner l'envie de faire ce travail.

Tout d'abord, merci à mes rapporteurs, Michel Habib, Stefan Schuster et Dominique de Vienne d'avoir accepté de lire et d'évaluer cette thèse.

Merci aux membres de mon comité de pilotage, Thomas Schiex, Gilles Curien, Alain Viari et Laurent Duret qui ont permis de valider et critiquer les étapes intermédiaires de ma thèse.

Merci à Cris, ma première collaboratrice, c'était un plaisir de travailler ensemble.

Merci à Leen et Alberto, tant de simplicité et d'intelligence réunies, c'est rare, et avec l'esprit d'équipe !

Merci à Christian pour sa sagesse et son humanisme.

Merci à Anne et Alain qui ont été à l'origine de ce travail et qui m'ont proposé un sujet si riche.

Merci à Sophie, Stéphane, Jean-Jacques et Franck pour leur clarté et leur pédagogie dans un domaine aussi difficile à appréhender que celui des graphes aléatoires.

Merci à Daniel d'avoir pris le temps de discuter de mes motifs.

Merci à Fabien de m'avoir fait découvrir le dessin de graphes, j'espère qu'on continuera à travailler ensemble.

Merci à Misou puis Agnès et Nathalie pour leur patience pour toutes les missions de dernière minute.

Merci à Stéphane, Bruno et Lionel pour leur disponibilité et leur sérénité face à des pannes généralement incompréhensibles.

Merci à Léo pour les hauts et les bas, les répètes entre midi et deux, et son humour à toute épreuve.

Merci à Émilie pour les cafés-allongés, les p'tites bières en terrasse, le temps des discussions.

Merci à Maud pour les bonnes vieilles fêtes de ces dernières années.

Merci à Vicente pour les preuves mathématiques et les innovations linguistiques.

Merci à Ludo d'avoir été un soutien sur le métabolisme, avec quelques mails douteux parlant de coude de la glycolyse.

Merci à Patricia et Manu d'avoir été présents et à l'écoute dans les moments difficiles.

Merci à Odile pour sa bonne humeur et ses bonnes idées.

Merci à Christelle pour les enseignements, c'était un plaisir de les préparer ensemble.

Merci à Paulo pour les discussions sur la science.

Merci à Pierre qui grâce à ses contrepetries m'a permis de faire une belle thèse.

Merci à Jeane et Saïd d'avoir été là au début et d'avoir apporté l'impulsion initiale à l'équipe.

Merci à tous les Baobabs, Marilia, Claire, Élise, Nuno, Vincent, Éric, Laurent, Sandrine... et les autres à venir.

Merci à Marie-France, fondatrice du baobab, tu ne comptes pas ton temps pour les autres. Merci pour cet environnement scientifique et humain.

Et merci à tout lecteur potentiel, cette thèse prend son sens si elle peut vous être utile.

Table des matières

Introduction

L'identification du support de l'hérédité (la molécule d'ADN) a marqué un tournant décisif dans l'histoire de la biologie ouvrant les portes au développement de la biologie moléculaire. Plus récemment, le séquençage de génomes complets a permis d'accélérer significativement le processus de détection et d'annotation des gènes encodés dans l'ADN. Mais cette démarche (dite réductionniste) qui consiste à identifier chaque gène dans un organisme présente des limites. Un organisme vivant ne peut être réduit à la liste de ses gènes.

En effet, il ne suffit pas qu'un gène soit contenu dans un génome pour conférer une capacité à un individu. La relation entre génotype (ensemble des gènes) et phénotype (ensemble des caractères observés) est beaucoup plus complexe. Une des voies qui se développe actuellement pour expliquer ce lien est l'étude systématique des interactions entre les molécules présentes dans une cellule (gènes, protéines, petites molécules). Ces interactions peuvent être directes (interactions physiques de protéines) ou indirectes (un facteur de transcription régule un autre gène en se fixant en amont de la séquence codante, ou deux enzymes catalysent des étapes consécutives dans une voie métabolique). Si on considère toutes ces interactions ensemble, on parle de réseau d'interaction. L'idée est que chaque gène n'est plus isolé dans un coin de la cellule mais au centre d'un complexe réseau d'interaction et c'est en comprenant mieux ces interactions qu'on parviendra à mieux comprendre la fonction de ce gène et son influence sur le phénotype.

Dans ce contexte, la recherche de motifs dans un réseau d'interaction moléculaire a pour but d'identifier des parties importantes de ce grand réseau. La recherche de motifs correspond à la recherche d'éléments répétés au sein du réseau. L'intérêt de trouver des éléments répétés est de deux ordres :

- D'un point de vue fonctionnel, on peut ainsi quantifier la redondance d'un réseau. Plusieurs répétitions d'un motif peuvent correspondre à plusieurs moyens de réaliser une même fonction. Si l'une des voies vient à défaillir, une voie de secours peut être utilisée.
- D'un point de vue évolutif, on peut faire l'hypothèse que les groupes de molécules détectés comme similaires ont une origine évolutive commune. Par le passé, il n'y avait alors qu'un seul groupe de molécules qui accomplissait cette fonction.

Détecter des motifs dans un réseau permet donc à la fois d'obtenir des informations sur le fonctionnement du réseau mais aussi sur la façon dont il

a évolué.

Ainsi, au cours de cette thèse, nous avons proposé une nouvelle définition de motif dans le cadre des réseaux métaboliques. D'un point de vue formel, un réseau métabolique est représenté par un graphe coloré et un motif est défini comme un multiensemble de couleurs. Une occurrence d'un motif correspond à un ensemble de noeuds connectés et colorés par les couleurs du motif. La recherche de motifs colorés constitue en fait un problème original en théorie des graphes. Nous avons donc caractérisé sa complexité et proposé un algorithme pour le résoudre. Nous avons aussi travaillé sur le problème d'inférence de tels motifs, en réfléchissant notamment à une mesure de sur-représentation statistique.

Il y a eu au cours de cette thèse un va-et-vient entre développement de méthodes et applications de ces méthodes à des exemples concrets. Ce va-et-vient a permis d'une part d'affiner les méthodes proposées pour les rendre plus opérationnelles et d'autre part d'augmenter la finesse des questions biologiques auxquelles ces méthodes pouvaient répondre.

Ainsi, nous verrons que les motifs extraits par notre méthode peuvent être interprétés. Nous argumenterons qu'ils peuvent correspondre à des unités fonctionnelles et évolutives. Le cadre formel que nous avons introduit permet de dégager des parties du réseau qui se ressemblent, et donc de générer des candidats pour une analyse plus approfondie. Une fois qu'un motif est détecté, on peut étudier la position de ses gènes sur le génome, ainsi que l'évolution des enzymes qui constituent ce motif.

Ce manuscrit présente les résultats obtenus au cours de ces trois années. Il regroupe à la fois des idées bien explorées mais aussi des pistes qui ont simplement été esquissées. Il est organisé autour de la notion de motif coloré qui a été le fil directeur de mon travail de thèse. Des liens avec des travaux connexes comme la visualisation de graphes, le calcul de modes élémentaires ou les modèles de graphes aléatoires sont également présentés. Une partie des résultats décrits dans cette thèse est publiée et est donc accessible via des articles. Ce manuscrit diffère dans sa forme d'un article scientifique en cela qu'il donne une perspective plus large et mentionne également des idées en cours.

Le manuscrit est organisé de la manière suivante : après un préambule présentant un questionnement sur la méthode scientifique, nous présentons un état de l'art de l'analyse structurelle du métabolisme, puis nous introduisons le concept de motif coloré ainsi que le cadre méthodologique permettant de rechercher un motif efficacement. Nous proposons ensuite une méthode permettant d'extraire tous les motifs d'un réseau ainsi qu'une mesure de sur-représentation permettant de décider si un motif est exceptionnel. Enfin, nous présentons des applications de cette recherche de motif au réseau métabolique de la bactérie *Escherichia coli* et nous tentons de montrer dans quelle mesure la notion de motif permet d'apporter de nouveaux éclairages sur l'évolution du métabolisme.

En annexe A et B est fourni un glossaire bref regroupant des notions de base de biologie moléculaire ainsi que des notions de base de complexité algorithmique.

Enfin, en annexe C, le lecteur pourra trouver trois articles correspondant à différents travaux réalisés au cours de ces trois années. Ces travaux sont tous liés d'une façon ou d'une autre à ceux présentés dans ce manuscrit. Ils seront introduit au cours du texte mais ne seront pas décrits en détail.

Chapitre 1

Préambule

J'ai souhaité commencer mon manuscrit par une réflexion sur la méthode scientifique suivie au cours de cette thèse.

En effet, mon sujet de thèse se trouve au carrefour entre plusieurs disciplines : la combinatoire, la biologie évolutive et les statistiques. Chacune de ces disciplines a son mode de fonctionnement, ses méthodes propres. Travailler à l'interface entre disciplines amène ainsi à se poser diverses questions. Suis-je en train de faire de la recherche en mathématique ou en biologie ? La méthodologie que j'utilise est-elle spécifique à une discipline ? Est-elle universelle ? Est-elle scientifique ?

Pour tenter de répondre en partie à ces questions, qui partent d'interrogations personnelles mais qui s'avèrent être aussi des questions relativement générales, je me suis intéressé à la philosophie des sciences.

Je manque clairement de compétences pour parler en profondeur de ce sujet mais je veux tout de même recenser quelques notions centrales qui ont servi à alimenter ma réflexion sur l'attitude qu'il me semblait bon d'adopter dans mes travaux. Ma principale source bibliographique pour cette partie a été [Vergez et Huisman, 1990].

1.1 Qu'est-ce qu'une méthode scientifique ?

Il semble tout d'abord utile de définir les concepts de déduction et d'induction.

La *déduction* logique consiste à déduire d'un énoncé général un énoncé particulier. Un exemple de déduction est le raisonnement syllogistique. À partir des premisses (les hommes sont mortels, Socrate est un homme), on déduit les conclusions (Socrate est mortel).

On note que la déduction est indépendante de l'expérience. Je prends certaines propositions pour prémisses et j'en tire, par la seule vertu du raisonnement, des conclusions. Par ailleurs, ces conclusions sont rigoureusement nécessaires. Une fois que j'ai adopté les prémisses, la conclusion logique ne peut pas être refusée.

À l'inverse, l'*induction* logique consiste à passer du particulier au général. C'est à partir de l'observation d'un grand nombre de faits qu'on affirme une loi générale. Dans le traité de logique de Port-Royal, on trouve l'exemple suivant : "lorsqu'on a éprouvé sur beaucoup de mers que l'eau est salée et sur beaucoup de rivières qe l'eau est douce, on conclut par induction que l'eau de mer est salée et l'eau de rivière douce". L'induction est dite amplifiante, car elle consiste à affirmer au-delà de ce qui est constaté.

Au 18^e siècle, David Hume [Hume, 1999] signale les dangers du raisonnement par induction. En effet, l'induction est basée sur l'observation, et la généralisation de cette observation à une loi. Du point de vue de la logique, ce n'est qu'un faux raisonnement car on conclut de quelques-uns à tous.

Un raisonnement par induction peut aboutir à des conclusions valides dans certains cas mais il peut également aboutir à des conclusions erronées. Par exemple, de multiples constations nous permettaient autrefois d'induire que tous les cygnes sont blancs. Nous savons maintenant que c'est faux car nous connaissons les cygnes noirs d'Australie.

Hume a constaté que l'induction était cependant communément employée et ceci pour deux raisons principales : l'habitude et l'association. L'habitude consiste à considérer que par le passé, le raisonnement par induction a donné de bons résultats et donc qu'il continuera à en donner. L'association consiste à dire que le futur doit ressembler au passé et qu'il existe une nécessité logique dans l'induction (les cygnes seront toujours blancs). Mais pour Hume, ces deux raisons sont insuffisantes à défendre le raisonnement par induction, et puisqu'aucune autre méthode rigoureuse ne semble pouvoir remplacer l'induction, elle continue d'être employée (tantôt à tort, tantôt à raison) et le problème reste ouvert.

Une première réponse est donnée par les positivistes au 19^e siècle. Selon les positivistes, pour qu'une théorie soit vraie scientifiquement, il faut qu'elle soit vérifiable. Le critère de scientificité est donc pour eux la *vérifiabilité*.

Mais cette idée de vérification semble incomplète à Popper qui lui préfère la notion de corroboration [Popper, 1992]. Si des observations sont en accord avec une théorie, alors ils la corroborent, mais ne la prouvent en aucun cas. En effet, une seule observation qui soit en contradiction avec la théorie suffit à l'infirmer. Une théorie ne peut donc pas être prouvée.

Le critère déterminant que Popper introduit alors est la falsifiabilité (ou en français *réfutabilité*). Ainsi, il énonce que les théories ne peuvent pas être prouvées mais bien plutôt réfutées, et c'est précisément cette réfutabilité qui les démarquent de propositions non scientifiques. Un exemple célèbre de proposition non réfutable et donc non scientifique est "Dieu existe".

Popper a eu une influence majeure dans l'élaboration d'une méthode scientifique extrêmement féconde : la méthode hypothético-déductive.

Ainsi, cette méthode consiste à poser une hypothèse (*i.e.* proposer une théorie) et puis voir si les observations la corroborent ou la réfutent. Si les observations la corroborent, l'hypothèse est acceptée. Si au contraire ils la réfutent, l'hypothèse est rejetée. Une théorie reste ainsi valide jusqu'à ce

qu'elle soit réfutée.

Nous avons donc pu voir que cette critique de l'induction a permis de fonder un critère de démarcation entre les énoncés scientifiques et non scientifiques.

Ainsi, on peut retenir que l'induction en elle-même ne peut servir de base solide à la démarche scientifique. On parle dans ce cas d'inductivisme.

Une question reste ouverte cependant : si l'induction est "mauvaise" et la déduction est "bonne", comment fait-on pour acquérir de nouvelles connaissances ? Comment fait-on pour générer les hypothèses, les fameuses prémisses du syllogisme ?

On peut sans doute dire que le critère de Popper permet de décider si un travail est scientifique quand il est fait mais qu'il ne décrit pas nécessairement l'intégralité du travail scientifique (la science en train de se faire), qui elle peut faire appel à l'induction.

On peut d'ailleurs avancer que la part inductive du processus d'acquisition de la connaissance est souvent cachée et que dans les articles scientifiques, on ne publie généralement que la deuxième partie : celle qui consiste à tester l'hypothèse qui a été formulée en amont.

Reprenons maintenant l'exemple des cygnes. Si on observe des cygnes et qu'on constate qu'ils sont tous de couleur blanche, Hume nous dit qu'on ne peut en aucun cas généraliser et tenir pour vérité que tous les cygnes sont blancs.

Par contre, le fait de constater que tous les cygnes qu'on a observés sont de couleur blanche peut nous inciter à formuler une hypothèse : tous les cygnes sont blancs. On se retrouve alors dans le cadre familier de la démarche hypothético-déductive. Chaque nouveau cygne blanc observé corroborera notre hypothèse mais l'observation d'un seul cygne de couleur différente la réfutera.

1.2 Différentes approches en bioinformatique

Au cours de ma thèse, j'ai pu voir qu'il existait un débat très actif dans le domaine de la bioinformatique opposant l'approche "hypothesis driven" (ou guidée par une hypothèse) à l'approche "data-driven" (guidée par les données). Cette question m'est apparue pour la première fois quand j'ai assisté à une conférence satellite d'ECCB (European Conference on Computational Biology) à Paris en 2003. Certains auditeurs pointaient du doigt les travaux présentés lors de la conférence principale comme essentiellement "data-driven" alors qu'ils considéraient que des travaux plus "hypothesis-driven" étaient présentés à cette conférence satellite.

Il me semble que ce débat pose typiquement un problème de démarche scientifique. S'affrontent ici deux tendances qu'on pourrait assimiler à la démarche hypothético-déductive et à la démarche inductive. J'ai par la suite retrouvé ce débat à plusieurs reprises, au cours de conférences ou dans la littérature scientifique.

Ainsi, dans un article assez provocateur, John Allen met en garde contre l'utilisation de l'approche inductive en bioinformatique [Allen, 2001]. Il se base sur une identification de l'induction à l'utilisation de programmes et de machines. Selon lui, l'induction seule ne permet pas de générer des hypothèses.

En réponse à John Allen, Douglas Kell argumente que l'approche inductive et l'approche hypothético-déductive ne s'excluent pas nécessairement [Kell et Oliver, 2004]. Ainsi, certaines disciplines, comme la biologie des systèmes, sont à l'heure actuelle riches en données mais pauvres en hypothèses. Une approche inductive permet de générer des hypothèses, ou de formuler des théories, dont les prédictions peuvent ensuite être testées par expériences.

En outre, Kell suggère que cette complémentarité entre les deux approches peut être itérative et propose le concept de cycle d'acquisition de connaissances. Ainsi, les expériences qui servent à tester les hypothèses précédemment formulées constituent également des jeux de données qui potentiellement dépassent le strict cadre de l'hypothèse testée et peuvent donc être utilisés pour induire de nouvelles hypothèses.

Enfin, on peut argumenter qu'il existe un risque à n'utiliser qu'une démarche hypothético-déductive. Ainsi, si notre recherche est exclusivement guidée par une théorie précédemment formulée, on peut être amené à rechercher les données qui corroborent notre hypothèse et à laisser de côté les données qui seraient susceptibles de l'invalider. Si on en croit [Gilbert et Mulkay, 1984], ce processus qui consiste à ignorer les données ou expériences, qui, si elles étaient considérées, invalideraient la théorie en vigueur, n'est pas rare dans l'histoire des sciences.

1.3 La démarche suivie dans cette thèse

Au cours de cette thèse, la démarche que nous avons suivie a été la suivante : à partir de travaux de la littérature montrant des similarités entre fragments de voies métaboliques, nous avons posé la question : existe-t-il d'autres exemples de similarités dans le métabolisme ?

Pour répondre à cette question, nous sommes d'abord passés par une étape de modélisation, qui a consisté à formaliser la notion de similarité que nous observions ; c'est alors que nous avons introduit et défini le concept de motif coloré. Nous avons ensuite développé des méthodes pour identifier ces motifs dans le métabolisme. Le développement de méthodes a été suivi de l'application de ces méthodes au métabolisme d'un organisme modèle : *Escherichia coli*. Cette application a fourni une première réponse à notre question puisque nous avons effectivement pu détecter de nouvelles similarités dans le métabolisme, mais elle nous a aussi incité à affiner notre modélisation en introduisant de nouvelles contraintes dans notre définition de motif. Nous avons alors développé de nouvelles méthodes découlant de notre nouvelle modélisation que nous avons à nouveau appliqué pour mettre en évidence de

nouvelles similarités.

Dans cette démarche, on peut noter que nous sommes partis d'une observation initiale (certaines voies métaboliques présentent des similarités) et nous avons voulu la généraliser (il existe d'autres similarités dans le métabolisme). Cette généralisation a par la suite été notre hypothèse de travail, que nous avons ensuite testée, en confrontant nos modèles au métabolisme d'*Escherichia coli*. En ce sens, notre démarche a d'abord été inductive, puis hypothético-déductive.

Un point qu'il nous semble important de souligner dans notre démarche est le va-et-vient entre modélisation, développement de méthodes d'une part et application, confrontation des modèles à la réalité biologique d'autre part. Ce va-et-vient nous a permis d'affiner petit à petit nos modèles et méthodes, et donc notre capacité à aborder des questions biologiques plus précises.

On peut ensuite signaler que dans cette thèse, nous avons choisi de donner une importance particulière aux méthodes. Mais les modèles et méthodes que nous développons trouvent toujours leur origine dans une question biologique. Et une fois développées, ces méthodes sont de nouveau confrontées à la biologie (ce qui permet éventuellement d'affiner les modèles ou de suggérer le développement de nouvelles méthodes). Ainsi, le cycle d'acquisition de connaissances est aussi dans notre cas un cycle entre biologie et mathématiques.

Enfin, on peut dire que la démarche suivie dans cette thèse est exploratoire. La notion de motif dans le cadre des réseaux métaboliques n'avait pas été étudiée auparavant. Nous ne savions pas à l'avance comment définir de telles structures. Nous avons donc dû proposer une définition initiale que nous avons par la suite raffinée.

Dans cette thèse, nous avons essayé de poursuivre le plus loin que nous pouvions l'idée qu'un réseau peut être appréhendé par une décomposition en briques fonctionnelles et/ou évolutives : les motifs.

Chapitre 2

État de l'art sur les réseaux biologiques

2.1 Introduction

2.1.1 Qu'est-ce qu'un réseau biologique ?

Un réseau biologique est une représentation abstraite d'un système biologique. Lorsque l'on parle de réseau, on se place dans une démarche de modélisation. La notion de réseau biologique est une notion très large, elle représente plutôt un niveau d'étude qu'une réalité biologique. L'idée principale qui motive l'utilisation de cette abstraction est la suivante : pour comprendre un processus biologique, il ne suffit pas de donner la liste des éléments qui y participent, cette liste ne constitue qu'une première étape à laquelle il faut ajouter l'étude des interactions entre ces éléments. Le terme de réseau en biologie est donc lié au terme d'interaction. De fait, la notion de réseau se retrouve dès lors qu'on veut modéliser des interactions en biologie, et ce, à différents niveaux de détails : depuis les interactions atomiques dans un repliement protéique jusqu'aux relations entre organismes dans une population ou un écosystème.

Dans cette thèse, nous allons nous concentrer plus précisément sur les réseaux d'interactions moléculaires, qu'on peut définir comme un ensemble de noeuds, représentant des gènes, des produits de gènes ou des métabolites, et un ensemble de connections représentant les interactions entre ces entités [Alm et Arkin, 2003]. Au sein même des réseaux d'interactions moléculaires (on parle aussi de réseaux cellulaires), il apparaît nécessaire d'opérer des subdivisions afin d'accéder à un niveau de description satisfaisant. En effet, si on analyse les réseaux d'interaction moléculaire dans leur ensemble, on se condamne à ne faire que des remarques très générales.

Le terme de réseau biologique a en particulier été utilisé pour désigner un des processus suivants :

- le métabolisme ;
- la régulation des gènes ;

– la transduction de signaux.

Chacun de ces processus met en jeu différents types de molécules : les gènes (fragments d'ADN) ; les transcrits (ARN) ; différents types de protéines : facteurs de transcription, enzymes ; et enfin des petites molécules : les métabolites.

Dans une approche qui n'est pas centrée sur les processus, on peut aussi définir les réseaux d'interaction de protéines. Ils rendent compte de toutes les interactions physiques entre protéines. Ils regroupent aussi bien la formation de complexes que des cascades de phosphorylation.

On peut noter que les interactions indirectes, comme des enzymes catalysant des étapes successives d'une voie métabolique ou un facteur de transcription agissant sur la régulation transcriptionnelle d'une protéine, ne seront généralement pas couvertes par ce type de données.

Cependant, certaines enzymes ou facteurs de transcription sont des complexes de plusieurs polypeptides. En outre, certains facteurs de transcription interagissent directement (de façon synergique) pour influer sur la transcription d'un gène. Les réseaux d'interaction de protéines recouvrent donc différents types de processus.

Quand on modélise ces processus par des réseaux, on opère généralement des simplifications, d'une part parce qu'on ne sait pas quelles sont toutes les molécules qui y sont réellement impliquées (les reconstructions sont incomplètes), et d'autre part, pour pouvoir faire des calculs, on construit des modèles qui sont le plus souvent des simplifications des réelles interactions moléculaires. Ainsi, un réseau de régulation de gènes va indiquer des interactions directes entre gènes alors que le mécanisme est indirect : l'un des gènes code pour un facteur de transcription, qui va être transcrit puis traduit pour pouvoir agir sur la transcription d'un second gène. Ces simplifications sont parfois acceptables mais parfois insuffisantes, suivant l'application qu'on considère.

Ainsi, dans la suite de ce manuscrit, on distinguera entre métabolisme (le processus biologique) et réseau métabolique (l'abstraction mathématique incomplète qui tente de modéliser ce processus).

Dans la suite, on se propose de définir un peu mieux chaque type de réseau, en insistant plus particulièrement sur les réseaux métaboliques.

2.1.2 Les réseaux métaboliques

Le métabolisme est généralement défini comme le processus à travers lequel les organismes vivants acquièrent et utilisent de l'énergie pour accomplir différentes tâches. Le métabolisme a quatre fonctions principales [Nelson et Cox, 2004] :

1. obtenir de l'énergie à partir de la dégradation de nutriments ;
2. convertir les nutriments en précurseurs (briques élémentaires) nécessaires à la synthèse de macromolécules ;

3. assembler ces briques élémentaires en composants cellulaires (protéines, acides nucléiques, lipides et polysaccharides) ;
4. synthétiser et dégrader les molécules requises pour des fonctions spécifiques de la cellule.

Chaque tâche peut être découpée en étapes élémentaires qu'on nomme réactions. Chaque réaction transforme des métabolites d'entrée (les substrats) en métabolites de sortie (les produits). Certaines réactions sont spontanées (*i.e.* les substrats se transforment en produits sans catalyseur) mais la plupart nécessite la présence d'une enzyme pour avoir lieu à une vitesse observable. Le rôle de l'enzyme est d'accélérer la réaction.

Une réaction est intégralement définie par ses substrats et ses produits, indépendamment de l'enzyme qui la catalyse. En pratique, une réaction peut être catalysée par plusieurs enzymes (isozymes). Les isozymes diffèrent généralement par leurs propriétés cinétiques et la façon dont elles sont régulées.

Chaque enzyme peut aussi catalyser plusieurs réactions. Certaines enzymes sont très spécifiques et ne catalysent qu'une seule réaction. D'autres sont peu spécifiques (enzymes dites "à large spectre") et peuvent catalyser toute une classe de réactions. Les propriétés cinétiques d'une enzyme varient selon la réaction qu'elle catalyse (on parle de différence d'affinité pour le substrat).

La plupart des réactions sont réversibles, c'est-à-dire qu'en présence d'un excès de substrat, la transformation se fait dans le sens classique (du substrat vers le produit) mais en présence d'un excès de produit, la transformation se fait en sens inverse (du produit vers le substrat). Le sens d'une réaction dépend donc de la concentration du produit et du substrat. Certaines réactions sont cependant considérées par certains auteurs comme irréversibles. C'est par exemple le cas de réactions faisant intervenir du CO_2 qui se dissout rapidement et ne reste donc pas disponible pour équilibrer la réaction.

La vitesse d'une réaction enzymatique n'est pas constante. Dans le cas d'une enzyme ne fixant qu'une molécule de substrat par molécule d'enzyme (enzyme michaelienne), cette vitesse dépend de la concentration de substrat ainsi que de constantes cinétiques caractéristiques de l'enzyme. Pour décrire l'évolution de la vitesse en fonction de la concentration du substrat, on utilise (dans le cas d'enzymes michaeliennes) l'équation de Michaelis-Menten :

$$V_i = \frac{V_{max}[S]}{K_m + [S]}$$

avec V_i vitesse de la réaction, V_{max} vitesse maximum de la réaction (pour une concentration saturante de substrat), $[S]$ concentration du substrat (en mol/L) et K_m constante de Michaelis spécifique de l'enzyme.

Le métabolisme est classiquement étudié par groupes de réactions participant à une même tâche. Ces groupes de réactions sont appelés voies métaboliques. La glycolyse, qui correspond à la dégradation du glucose en pyruvate par une série de huit étapes, est un exemple de voie métabolique.

Bien que certaines voies aient été bien étudiées, il n'existe pas de définition formelle de la notion de voie métabolique. Cette notion est plutôt historique et reflète la façon dont le métabolisme a été découvert. En suivant le devenir de certains métabolites d'intérêt tels que le glucose ou le pyruvate, on a regroupé des réactions sous le nom de glycolyse ou cycle de Krebbs. Les frontières des voies métaboliques sont généralement assez mal définies. En effet, selon les bases de données, une voie métabolique peut contenir plus ou moins de réactions. À part peut-être pour certaines voies très étudiées, il ne semble pas exister de consensus à ce sujet.

En outre, on peut noter qu'une même réaction peut appartenir à plusieurs voies métaboliques. L'ensemble des voies métaboliques ne constitue donc pas une partition du réseau au sens mathématique du terme (les voies ne sont pas des ensembles disjoints).

On retiendra que la notion de voie métabolique est mal définie et est donc délicate à utiliser pour modéliser un réseau. Cependant, cette notion correspond à un niveau d'analyse pertinent puisque c'est à ce niveau que se font l'essentiel des analyses métaboliques.

2.1.3 Les réseaux de régulation de gènes

Une question fondamentale en physiologie et en embryologie est de savoir pourquoi une cellule n'exprime pas en permanence toutes les potentialités présentes dans son génome [Jacob et Monod, 1961]. En effet, sur les 4000 gènes d'un génome bactérien ou les 35000 du génome humain, seulement une fraction est exprimée dans une cellule à un instant donné [Nelson et Cox, 2004]. Certaines protéines sont présentes en très larges quantités comme les facteurs d'élongation nécessaires à la synthèse de protéines, d'autres en plus faible quantité, comme les enzymes impliquées dans la réparation de l'ADN. Étant donné le fort coût énergétique associé à la synthèse de protéines, la régulation de l'expression est essentielle pour faire un usage optimal de l'énergie disponible.

La concentration d'une protéine peut être contrôlée à différents niveaux :
- synthèse de l'ARN transcrit primaire (transcription) ;
- modification post-transcriptionnelle de l'ARN ;
- dégradation de l'ARN messager ;
- synthèse de la protéine (traduction) ;
- modification post-traductionnelle de la protéine ;
- adressage et transport de la protéine ;
- dégradation de la protéine.

Quand on parle de réseaux de régulation des gènes dans la littérature, on parle généralement du premier point, à savoir la régulation de l'initiation de la transcription. On devrait donc dire réseaux de régulation de la transcription (et non réseaux de gènes).

De manière générale, la transcription d'un gène peut être régulée par un ou plusieurs facteurs de transcription. Ces facteurs de transcription sont des

protéines qui peuvent elle-mêmes être régulées au niveau transcriptionnel par d'autres facteurs de transcription. Lorsqu'on considère l'ensemble des régulations transcriptionnelles, on obtient un réseau complexe.

On n'a aujourd'hui qu'une vision très incomplète du réseau de régulation de la transcription d'un organisme. Cependant, certaines études globales ont déjà été menées [Milo *et al.*, 2002, Babu *et al.*, 2004].

Une nouvelle direction très intéressante qui se développe dans le domaine de la régulation des gènes et qui n'est pas beaucoup prise en compte actuellement est la régulation des gènes par modifications de la structure de la chromatine. En effet, pour qu'un gène puisse être transcrit, il est nécessaire qu'il soit accessible à la machinerie de transcription et donc qu'il soit dans une zone de chromatine ouverte.

2.1.4 Les réseaux de transduction du signal

La transduction de signaux désigne l'intégration d'un message d'origine extracellulaire par une cellule. La transduction de signaux est un point commun de la communication cellulaire des systèmes endocrinien, nerveux et immunitaire chez les mammifères. Chez ceux-ci, il existe plus de 200 types cellulaires différents et spécialisés.

Comme pour les réseaux de régulation de la transcription, on a actuellement une vision très partielle du réseau de transduction du signal complet d'un organisme. Dans la littérature, les études se concentrent généralement sur des voies particulières (comme la voie de $NF\kappa b$ ou la cascade des map-kinases).

Ces processus, quand ils sont bien caractérisés, sont modélisés sur le plan quantitatif par des modèles à base d'équations différentielles [Yue *et al.*, 2006]. Plus récemment, les réseaux de Petri ont été utilisés pour étudier la structure de ces réseaux [Klamt *et al.*, 2006, Sackmann *et al.*, 2006].

2.1.5 Intégration des réseaux

Au sein d'une cellule, le réseau métabolique, le réseau de régulation des gènes et le réseau de transduction du signal ne sont bien sûr pas disjoints. Les protéines impliquées dans les voies métaboliques et les voies de transduction du signal peuvent être régulées au niveau transcriptionnel et les métabolites impliqués dans les voies de signalisation sont impliquées également dans les voies métaboliques.

À l'heure actuelle, il existe peu d'études qui prennent en compte plusieurs réseaux à la fois. Une explication à cela est que les données sont encore trop partielles (et de qualité inégale suivant le type de réseaux). Une autre explication est qu'il n'existe actuellement pas de consensus sur un formalisme adapté pour modéliser conjointement ce réseau cellulaire intégré.

2.2 Données

On a pu voir que les données étaient amenées à jouer un rôle central dans une démarche d'acquisition de connaissances en bioinformatique. Aussi, il est nécessaire de bien comprendre de quelles données on dispose en pratique et de savoir comment elles ont été obtenues pour mieux décider du niveau de confiance qu'on peut leur accorder.

Dans cette partie, nous parlerons essentiellement de données disponibles publiquement dans des bases puisque c'est ce type de données que nous avons principalement utilisé. Certaines bases de données sont généralistes (KEGG), d'autres sont spécifiques d'un organisme (EcoCyc). Il existe également des jeux de données disponibles publiquement qui sont spécifiques à une question biologique et qui ont été constitués précisément pour y répondre.

Au sein des données disponibles dans les bases, on distingue ainsi 3 types :
- les données issues de la littérature (ou données bas débit) ;
- les données issues d'expériences à haut débit ;
- les données inférées.

On peut également faire la différence entre données qualitatives (quelles sont les molécules qui interagissent) et données quantitatives (quelles sont les constantes d'association). On parlera dans cette partie essentiellement de données qualitatives. Ces données permettent de travailler avec des modèles qualitatifs afin de poser des questions d'ordre structurel. On note qu'en pratique, les données quantitatives ne sont disponibles que pour des voies très étudiées.

Enfin, parmi les types de données dont nous ne parlerons pas, on peut mentionner les données issues de la métabolomique, nouveau domaine en expansion. Une expérience de métabolomique consiste à mesurer tous les métabolites présents à un instant donné dans une cellule. Ces données peuvent être qualitatives ou quantitatives. Pour une introduction à ce domaine, voir [Nobeli et Thornton, 2006].

2.2.1 Les reconstructions de réseaux

Pour certains organismes modèles, telles que la bactérie *Escherichia coli* ou la levure *Saccharomyces cerevisiae*, on dispose de reconstructions assez complètes de leur réseau métabolique. Ces exemples constituent encore des exceptions et les autres données disponibles publiquement sont à prendre avec beaucoup de précautions. En effet, si on dispose aujourd'hui de plus de 200 génomes complètement séquencés, on n'a pas pour autant accès à leurs réseaux métaboliques. Il ne suffit pas d'avoir la séquence d'un organisme pour connaître ses capacités métaboliques.

Cette section traite des méthodes de reconstruction de réseaux métaboliques. On appelle reconstruction le processus qui consiste à obtenir, à partir du génome (annoté) d'un organisme, la liste des réactions qui constituent son réseau métabolique.

Plusieurs types de données peuvent être utilisées pour reconstruire un réseau : en premier lieu, la séquence complète et ses annotations, mais aussi des données de métabolomique ([Tucker et Moulton, 2005]), ou encore des données de protéomique, et enfin des données issues de la littérature (bibliomique) [Duarte *et al.*, 2007].

2.2.1.1 La première étape de reconstruction

Toutes les méthodes de reconstruction du réseau métabolique d'un organisme prennent pour point de départ la séquence complète de son génome. Un point clé de la reconstruction est la détection puis l'annotation des gènes codant pour des enzymes le long du génome.

L'annotation des gènes est souvent faite par recherche d'homologie avec des gènes existants chez d'autres espèces (pour lesquelles le métabolisme est mieux caractérisé). Ainsi, pour déterminer si une réaction a lieu dans un organisme, on vérifie que son génome contient les gènes homologues codant pour l'enzyme correspondante. Dans le cas où tous les gènes qui participent à la formation d'une enzyme sont "présents" dans le génome d'un organisme, on considère que la réaction fait partie de son réseau métabolique. Dans ce processus, on fait une hypothèse forte. On considère que des enzymes homologues catalysent la même réaction dans des organismes différents.

On retiendra ainsi que la qualité de la reconstruction métabolique dépend fortement de la qualité de l'annotation du génome.

2.2.1.2 Le raffinement du modèle

Suite à cette première étape de reconstruction, une étape de raffinement est nécessaire. En effet, certaines réactions ont été ajoutées au réseau alors qu'elles ne devaient pas l'être (faux positifs) mais surtout certaines réactions manquent (faux négatifs). On peut soupçonner la présence de faux-négatifs notamment dans le cas de voies métaboliques incomplètes et tenter de les corriger. Ce raffinement est généralement fait manuellement ou de manière semi-automatique.

Les travaux présentés dans [Green et Karp, 2004] permettent d'automatiser la recherche de nouvelles réactions quand elles sont détectées comme manquante dans une voie métabolique.

Enfin, un réseau métabolique, une fois reconstruit, peut être testé pour sa consistence. Il existe ainsi des techniques de validation de modèle basées sur les invariants du système à l'état stationnaire. On peut alors détecter des inconsistences qui, pour être résolues, nécessitent, soit le retrait, soit l'ajout de réactions.

Un problème notable dans les méthodes de reconstruction est la question de l'indépendance entre les réseaux métaboliques de différents organismes. En effet, si on utilise principalement la similarité de séquences pour inférer les gènes (et par extension les réactions) présents dans un organisme, alors

on est réduit à manquer les réactions qui n'ont été observées chez aucun organisme auparavant. Cette dépendance entre les jeux de données pose des problèmes qui ne sont pas souvent abordés quand on cherche à faire des études comparatives entre différents réseaux.

Pour s'affranchir de ce problème, nous avons choisi de travailler essentiellement sur un organisme, *E. coli*, qui est sans doute celui pour lequel on dispose des informations les plus complètes.

2.2.1.3 Les outils disponibles

Il existe différents outils disponibles qui permettent de faire une reconstruction métabolique.

On peut mentionner le programme PathoLogic du package "Pathway Tools" [Karp *et al.*, 2002]. La reconstruction est automatique à partir d'un fichier de génome annoté. Des utilitaires de raffinage (PathwayHoleFiller) [Green et Karp, 2004] sont aussi disponibles dans le package.

D'autre part, PRIAM [Claudel-Renard *et al.*, 2003] est un programme qui permet de reconstruire un réseau métabolique en prenant en compte l'information des domaines de protéines.

Enfin, de nombreuses reconstructions sont disponibles via les bases de données traditionnelles (KEGG, BioCyc) et aussi via le site du groupe de Bernhard Palsson (`http://gcrg.ucsd.edu/`).

2.2.2 Bases de données

Dans une approche simplifiée, on distingue différents types de bases de données qui s'alimentent les unes les autres [Wittig et Beuckelaer, 2001] (voir Fig. 2.1). Au plus bas niveau, on a la base génomique qui alimente la base protéique, puis la base d'enzymes, et enfin la base de voies (voies métaboliques ou de transduction). De plus, une autre base alimente l'ensemble de ces bases de données : c'est la base issue de la littérature (qui correspond aux données provenant d'expériences humides).

De manière générale, les bases de données métaboliques (qu'on étudiera plus particulièrement ici) sont souvent intimement liées à des bases d'enzymes. Elles peuvent être vues comme des bases de réactions dont les réactants sont caractérisés (entre autres) grâce à des bases d'enzymes.

Les bases de données métaboliques les plus couramment utilisées sont les suivantes : KEGG (`http://www.genome.ad.jp/kegg/`), EcoCyc (`http://ecocyc.org/`), ExPASy-Biochemical Pathways (`http://us.expasy.org/`), ERGO (anciennement WIT) (`http://www.ergo-light.com/`), Reactome (`http://www.reactome.org/`) et UM-BBD (`http://umbbd.ahc.umn.edu/`).

Nous allons ici en détailler quelques-unes :

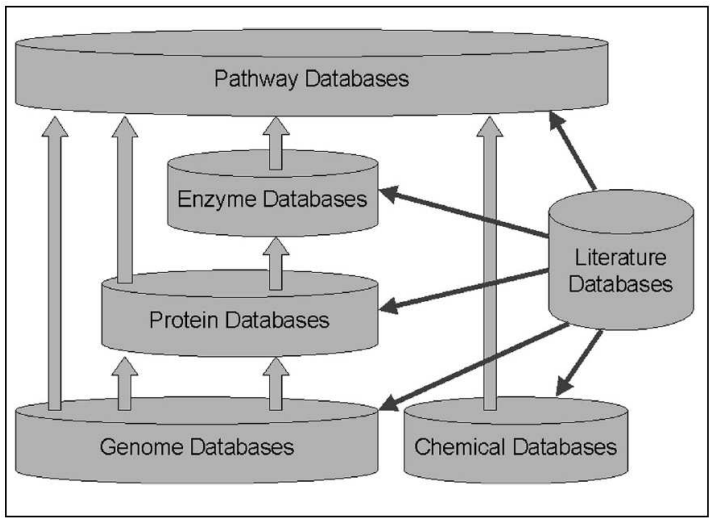


FIG. 2.1 – Connections entre les principaux types de bases de données. La figure est issue de [Wittig et Beuckelaer, 2001].

• **KEGG** ("Kyoto Encyclopedia of Genes and Genomes") contient toutes les voies métaboliques connues et un nombre limité de voies de régulation et de mécanismes de transport.
Le système KEGG est constitué principalement de trois bases de données intimement connectées : LIGAND qui regroupe des informations sur les composés, les enzymes et les réactions ; PATHWAY qui comprend les représentations graphiques des voies et la liste des enzymes et réactions associées ; et GENES, qui contient le génome et les informations sur les gènes pour un organisme ou une voie données. Les voies métaboliques sont classées en fonction de la structure chimique de leurs principaux composants, *e.g.* carbohydrates, lipides, acides aminés. Elles sont visualisables via des cartes statiques faites à la main. Ces cartes contiennent toutes les réactions connues catalysées par une enzyme mais les réactions non-enzymatiques sont généralement absentes de la représentation. Les composés auxiliaires (ATP ou NADH) ne sont pas représentés. On peut reprocher à KEGG son manque de références à la littérature, la majeure partie des données est issue de prédiction et de déduction à partir de comparaisons inter-organismes et inter-voies. En outre, il existe de nombreux problèmes de consistance des données liés à la structure même de KEGG [Wittig et Beuckelaer, 2001]. KEGG est la plus complète des bases de données mais est peut-être aussi la moins rigoureuse.
On peut signaler ici une initiative de Ma et Zeng [Ma et Zeng, 2003] qui ont constitué une version filtrée de KEGG. Ce travail a été réalisé pour 80 organismes, et outre la correction d'inconsistances, il introduit l'information de réversibilité des réactions.

• **ECOCYC/BIOCYC** EcoCyc est une base de données métaboliques qui décrit le génome et la biochimie d'*Escherichia coli* à partir de la base EcoGene, de SWISS-PROT et de la littérature scientifique. La base contient toutes les séquences et annotations fonctionnelles des gènes d'*E. coli*. Les voies métaboliques ainsi que les réactions et enzymes sont annotées à l'aide de références à la littérature. De plus, BioCyc décrit les voies, réactions et enzymes de plusieurs autres organismes microbiens en utilisant le même schéma de base de données (*i.e.* la même ontologie) et les mêmes utilitaires de visualisation.

Des diagrammes des voies sont générés à l'aide d'un algorithme de dessin de graphe et sont stockés en tant qu'images statiques. Pour chaque réaction, on peut distinguer les substrats principaux des substrats auxiliaires, information souvent manquante dans les bases mais qui s'avère très utile dans le cadre d'une analyse de réseau. EcoCyc comprend également des informations sur des gènes qui n'ont pas encore été clonés mais dont la fonction a été caractérisée par des approches génétiques ou biochimiques. Du fait que la base est dédiée à un unique organisme qui a été l'objet de nombreuses études, les données d'EcoCyc sont considérées comme fiables.

Au cours de ma thèse, j'ai surtout travaillé avec des données disponibles publiquement (KEGG puis EcoCyc). Dans un premier temps, j'ai pensé, à tort, que ces données étaient exhaustives.

Je suis aujourd'hui toujours intéressé par ce type de données mais je pense qu'elles doivent être manipulées avec précaution, en ayant conscience que les données sont parfois erronées et le plus souvent incomplètes. Le cas idéal serait en fait d'avoir des données qui sont spécialement collectées pour répondre à une question que l'on se pose. Le rôle des données publiques (inévitables pour avoir une "vision d'ensemble") serait ainsi de dégager des tendances générales qui pourraient ensuite être confirmées par des expériences complémentaires.

2.3 Modélisation du métabolisme

Le métabolisme peut être étudié d'un point de vue structurel ou d'un point de vue dynamique. Les modèles utilisés sont généralement différents. Nous allons ici en détailler certains.

On distingue deux grandes catégories de modèles [Deville *et al.*, 2003] :

1. **les modèles pour l'analyse structurelle** : ils regroupent principalement les modèles issus de la théorie des graphes. Ces modèles nécessitent généralement peu (ou pas) de données quantitatives et sont plutôt dédiés à une analyse qualitative des réseaux.
2. **les modèles pour l'analyse dynamique** : cette catégorie regroupe des modèles plutôt orientés vers la simulation et l'étude de propriétés dynamiques des réseaux. L'étude de la dynamique requiert généralement des informations quantitatives. Ces modèles n'excluent pas l'utilisation

de graphes, mais le graphe est ici un intermédiaire dans la démarche de modélisation.

Au cours de ma thèse, j'ai surtout travaillé sur le premier type de modèles. On peut argumenter que l'approche qualitative est à la fois une première étape vers une analyse quantitative mais est aussi utile en soi et a ses questions propres.

Pour obtenir des informations complémentaires sur les modèles, on peut consulter la synthèse de J. Stelling [Stelling, 2004] qui différencie 3 types de modèles, les graphes, les modèles de flux et les équations différentielles. Son exposé sur les graphes est particulièrement succinct, il les réduit aux graphes simples (sans considérer les hypergraphes par exemple). Un autre travail dont je me suis inspiré est celui de Deville et collaborateurs [Deville *et al.*, 2003] qui donne plus de détails sur les graphes, mais ne mentionne pas le problème de la distinction entre composés de droite et composés de gauche pour les réactions réversibles (ce problème sera expliqué dans la section 2.3.1.3).

Dans cette section, nous tenterons de faire une synthèse des différents articles et de mieux mettre en évidence les liens qui existent entre les différents modèles qui sont classiquement utilisés pour analyser le métabolisme.

2.3.1 Graphes

Un graphe est une structure mathématique utilisée pour modéliser des relations binaires entre les objets d'une collection donnée.

Formellement, un graphe G est défini comme un couple (V, E) avec :
- V un ensemble fini de sommets (nous utiliserons aussi indifféremments le terme de noeuds) ;
- E une partie de V^2, un ensemble d'arêtes.

Modéliser un réseau métabolique par un graphe revient à choisir quels sont les entités biologiques qu'on associe aux sommets et aux arêtes. Dans le cadre du métabolisme, les entités peuvent être les composés, les réactions ou les enzymes. Nous allons détailler les trois modèles de graphes correspondants.

2.3.1.1 Le graphe des composés et le graphe des réactions

Le graphe des composés est un modèle de réseau métabolique où les sommets correspondent aux composés et où il existe une arête entre deux composés A et B s'il existe une réaction dont ils sont respectivement substrat et produit.

Le graphe des réactions est le graphe dual. Les sommets correspondent aux réactions et il existe une arête entre deux réactions s'il existe un composé qui est produit par l'une et consommé par l'autre.

Dans le graphe d'enzymes, les sommets correspondent aux enzymes et il existe une arête entre deux enzymes si elles catalysent deux réactions qui ont des composés en commun. Ce graphe est moins utilisé que les précédents à cause de son manque de clarté. En effet, son utilisation est limitée par

le fait qu'une enzyme peut catalyser plusieurs réactions, et notamment des réactions qui sont distantes en termes de nombre de composés pour passer de l'une à l'autre. Considérer le graphe d'enzymes introduit donc des court-circuits dans le réseau. Une autre difficulté vient de la situation inverse où une réaction est catalysée par plusieurs enzymes ; dans ce cas, la réaction sera dupliquée.

Cependant, selon l'application qu'on considère, on peut ne pas être gêné par ces biais. En effet, les graphes d'enzymes ont été utilisés pour diverses analyses génomiques [Horne *et al.*, 2004]. On pourra retenir que l'utilisation du graphe d'enzymes entraîne une perte d'information structurelle. Cette perte d'information peut ne pas être dommageable si on s'intéresse strictement aux enzymes et aux relations entre elles.

Dans notre cas, nous favoriserons l'utilisation du graphe des réactions ou du graphe des composés qui reflètent plus une vision biochimique du métabolisme. Nous reviendrons sur l'utilisation des enzymes comme étiquettes des réactions dans la section 3.3.3. En effet, pour ne pas perdre l'information de structure du réseau mais pour avoir l'information des enzymes, on peut étiquetter les noeuds du graphe de réactions par les enzymes qui catalysent ces réactions.

2.3.1.2 Le graphe biparti et l'hypergraphe des composés

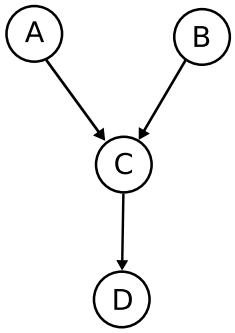
Le graphe des composés et le graphe des réactions peuvent parfois être ambigus. Ainsi, les deux réseaux suivants auront le même graphe des composés :

Réseau 1

Reaction 1 : A → C
Reaction 2 : B → C
Reaction 3 : C → D

Réseau 2

Reaction 1 : A+B → C
Reaction 2 : C → D



On peut résoudre cette ambiguïté en ajoutant des étiquettes sur les arêtes. En effet, si on indique sur chaque arête la réaction qui lui correspond, alors on lève cette ambiguïté.

Une autre manière de procéder est d'utiliser un modèle de graphe plus expressif : un graphe biparti ou un hypergraphe.

Formellement, un graphe biparti est un graphe dont l'ensemble des noeuds peut être divisé en deux ensembles disjoints U et V tel que chaque arête a un sommet dans U et un sommet dans V. L'utilisation d'un graphe biparti est assez intuitive pour modéliser un réseau métabolique : un des ensembles de noeuds correspond aux composés et l'autre correspond aux réactions. On n'a alors des arêtes qu'entre composés et réactions (voir Fig. 2.2).

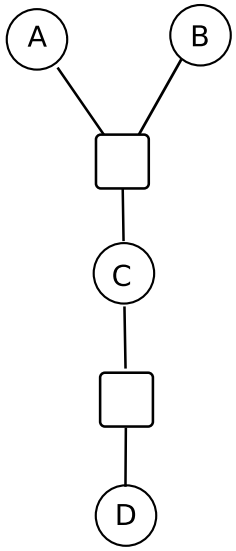


FIG. 2.2 – Graphe biparti correspondant au réseau : A+B ↔ C, C ↔ D.

L'utilisation d'un graphe biparti permet de résoudre l'ambiguïté observée pour les graphes simples. Mais cette ambiguïté peut aussi être levée en utilisant un hypergraphe. Ces deux structures de données (graphe biparti et hypergraphe) sont strictement équivalentes en termes de quantité d'information modélisée. On peut passer de l'une à l'autre très simplement.

Intuitivement, un hypergraphe est un graphe où les arêtes sont des hyperarêtes et peuvent lier plus de 2 noeuds.

Formellement, un hypergraphe H est une paire (V, E), où $V = \{v_1, v_2, ..., v_n\}$ est l'ensemble des sommets (ou noeuds) et $E = \{E_1, E_2, ..., E_m\}$, avec $E_i \subseteq V$, pour $i = 1, ..., m$, est l'ensemble des hyperarêtes. Clairement, si $|E_i| = 2$, pour tout $i = 1, ..., m$, alors l'hypergraphe est un simple graphe.

Pour modéliser un réseau métabolique par un hypergraphe, on associe généralement les composés aux noeuds et les réactions aux hyperarêtes.

2.3.1.3 L'orientation des arêtes

Nous avons pour l'instant parlé exclusivement de graphes non orientés.

Un graphe orienté G est un couple (V, E), où :

– V est un ensemble de sommets ;

– E est un ensemble de couples (ordonnés) de sommets, appelés arcs.

Tous les modèles mentionnés jusqu'ici peuvent être orientés ou non orientés, selon que les réactions sont réversibles ou pas. Un graphe orienté est un graphe où chaque arête a une origine et une destination. Pour un graphe simple, chaque arête est alors une paire orientée. Pour un hypergraphe, une hyperarête orientée correspond à plusieurs sources et plusieurs destinations (voir Fig. 2.3).

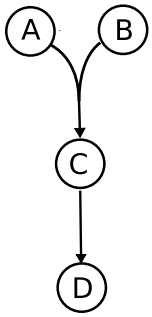


FIG. 2.3 – Hypergraphe orienté correspondant au réseau : A+B → C, C → D.

Dans le cas d'un réseau où toutes les réactions sont irréversibles (resp. réversibles), on choisit de modéliser le réseau par un graphe orienté (resp. non orienté). Dans le cas intermédiaire où certaines réactions sont réversibles et d'autres ne le sont pas, alors on utilise un graphe mixte (certaines arêtes sont orientées et d'autres ne le sont pas).

Enfin, on peut noter que l'utilisation d'arêtes non orientées dans le cas d'une réaction réversible laisse encore une ambiguïté (plusieurs réseaux peuvent avoir la même représentation), même quand on utilise un graphe biparti ou un hypergraphe. Un exemple est donné dans la figure 2.4.

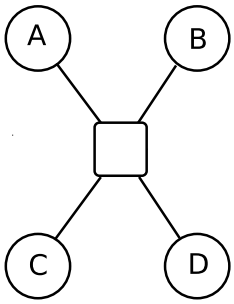


FIG. 2.4 – Graphe biparti correspondant à deux réseaux possibles : A+B ↔ C+D ou A+C ↔ B+D.

On constate que l'orientation ne rend pas compte de la différence entre substrats et produits, ou devrait-on dire entre composés de gauche et composés de droite (puisque les termes substrats et produits impliquent une

orientation). Pour lever cette ambiguïté, on peut proposer plusieurs solutions.

Une première solution consiste à dupliquer le noeud correspondant à chaque réaction réversible dans le graphe biparti, puis à orienter le graphe (de manière équivalente, dupliquer l'hyperarête dans le cas de l'hypergraphe puis orienter l'hypergraphe). Un inconvénient de cette méthode est qu'elle crée des cycles artificiels.

Une autre solution consiste à ajouter des étiquettes sur les arêtes du graphe biparti pour différencier les composés de gauche des composés de droite.

Enfin, une dernière solution consiste à combiner hypergraphe et graphe biparti. On obtient alors un hypergraphe biparti. Les hyperarêtes joignent réaction et substrats ou réaction et produits, mais ne mélangent jamais substrats et produits.

Pour chacune de ces solutions, on dispose de l'information nécessaire pour différencier substrats et produits.

Pour finir, il reste à préciser la notion de chemin réaliste dans un graphe métabolique. En effet, si on conserve une définition classique de chemin, on peut se retrouver dans une situation où un chemin "entre dans une réaction par un substrat" et "ressort par un autre substrat". Ce type de chemin n'est pas réaliste sur le plan biologique et ne peut pas être assimilé à un processus biologique car un substrat ne peut être obtenu à partir d'un autre substrat.

Cette situation ne se présente pas dans le cas de l'hypergraphe biparti. Elle est facile à régler dans le cas du graphe étiqueté, mais elle est plus délicate pour le graphe contenant des réactions dupliquées.

2.3.1.4 Le choix du modèle de graphe

Dans avons donc présenté plusieurs possibilités pour modéliser un réseau métabolique à l'aide de graphes. Nous allons à présent répondre à la question : dans quelle situation utiliser quel modèle ?

Dans cette thèse, nous utiliserons le graphe de réactions et le graphe biparti. Une idée simple qui a prévalu à nos choix est qu'il est toujours préférable d'utiliser le modèle le plus simple qui rende compte de la réalité à modéliser.

Ainsi, dans la plus grande partie de cette thèse, nous recherchons des ensembles de réactions connectées (sans nous intéresser à la façon dont elles sont connectées). Le graphe de réactions est suffisant pour modéliser cela. Quand, par contre, nous nous intéressons à la façon dont ces réactions sont connectées (par quels métabolites), alors il devient nécessaire d'utiliser le graphe biparti.

En ce qui concerne l'orientation des arêtes, nous avons choisi de travailler avec des graphes non orientés. Ce choix revient à faire l'hypothèse que toutes les réactions du réseau sont réversibles. Cette hypothèse paraît forte mais

on peut la justifier. En effet, l'information de réversibilité n'est pas toujours disponible selon les organismes étudiés, et quand elle est disponible, elle est parfois contradictoire selon les bases de données. Il nous a semblé préférable de considérer que des réactions étaient réversibles alors qu'en réalité elles ne le sont pas, plutôt que l'inverse (qui correspondrait à une perte d'information).

Enfin, un dernier point qui est crucial quand on parle de topologie dans les réseaux métaboliques, est la question des composés ubiquitaires.

2.3.1.5 Composés ubiquitaires

Dans tous les formalismes décrits précédemment, on considère que toutes les réactions et tous les composés sont équivalents.

Or, dans le métabolisme, certains composés peuvent être considérés comme centraux et d'autres comme périphériques, certains ont un statut de composés auxiliaires (cofacteurs) et d'autres de composés principaux. Ne pas prendre en compte cette diversité, c'est s'exposer à de mauvaises interprétations, notamment lorsqu'on s'intéresse à des distances entre composés dans un réseau. Un exemple est donné dans la figure 2.5 où, si on traite tous les composés de manière égale, on trouve une distance de taille 2 entre A' et B' alors que le lien se fait à travers l'ATP qui est une molécule participant à un grand nombre de réactions et dont le rôle principal est d'apporter de l'énergie en libérant un groupement phosphate.

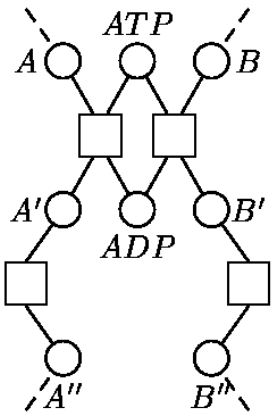


Fig. 2.5 – Exemple de l'effet de composés ubiquitaires sur la topologie du réseau.

Nous avons identifié 3 façons de traiter les composés génants :
- retirer les composés les plus fréquents ;
- retirer les composés secondaires (ou auxiliaires) ;

– retirer les composés qui ne sont pas pertinents (du point de vue biologique) pour l'analyse qu'on veut faire.

Ces trois méthodes sont implémentées dans notre programme, MOTUS (qui sera présenté dans le chapitre 7). La première nous semble assez critiquable. En effet, cette méthode permet de retirer des composés qu'il est effectivement utile de retirer comme CO_2 ou H_2O mais elle conduit aussi à retirer de l'analyse des composés comme le L-glutamate et le pyruvate qui sont des composés centraux autour desquels s'organisent le métabolisme. Cette méthode est cependant très utilisée (c'est en fait la plus utilisée) et nous l'avons donc implémentée dans notre outil dans l'optique de pouvoir faciliter la comparaison de nos résultats avec ceux d'autres équipes.

La seconde méthode est celle que nous avons le plus employée. Elle a l'avantage de ne pas retirer systématiquement un composé et de considérer le contexte dans lequel il est employé. Ainsi, l'ATP, qui est un composé qui intervient dans de nombreuses réactions, sera conservé dans les réactions où il intervient comme composé principal (les réactions impliquées dans la synthèse de l'ATP par exemple) et retiré dans les réactions où il intervient comme composé auxiliaire (les réactions où l'ATP agit comme fournisseur d'énergie).

Une point clé de cette seconde méthode est la façon dont on définit composés primaire et secondaire. En effet, cette partition reflète un point de vue sur une réaction. Ce choix peut être fait automatiquement en utilisant la notion de voie métabolique : un composé est primaire s'il est produit et consommé dans la voie ou s'il est substrat initial ou produit final [Karp et Paley, 1994] (c'est le cas pour les données d'EcoCyc) ou manuellement (c'est le cas pour les cartes métaboliques de KEGG).

La troisième méthode est sans doute la plus satisfaisante mais elle nécessite une expertise avancée qu'il est difficile d'avoir pour un grand système. Jean-Pierre Mazat (spécialiste du métabolisme mitochondrial, qui travaille au laboratoire "Physiologie mitochondriale" à Bordeaux) m'a aidé à trouver quels composés il était pertinent d'écarter pour le réseau de la mitochondrie. Une des limites de ce traitement est qu'il est dépendant de l'application biologique.

Plusieurs travaux traitant de questions liées aux composés ubiquitaires sont parus ces dernières années. Parmi les travaux qui s'intéressent au calcul de distance entre composés dans un réseau métabolique, on peut mentionner ceux de Didier Croes [Croes *et al.*, 2006]. Les auteurs s'intéressent à trois situations : 1. tous les composés sont conservés ; 2. les composés les plus fréquents sont retirés du graphe ; et 3. tous les composés sont conservés mais les noeuds sont pondérés par leur degré. Pour chacun des graphes, les auteurs recherchent le chemin le plus court (ou de poids le plus faible) entre deux composés. Un chemin est considéré comme valide s'il correspond à une voie métabolique connue (telle que définie dans une base de données). Selon ce critère de validation, les auteurs montrent que la solution 3 est meilleure que la 2 qui est meilleure que la 1. On peut remarquer que ce travail n'est

pas totalement satisfaisant, d'une part car le critère de validation proposé dépend fortement de la notion de voie métabolique (qui est mal définie), et d'autre part car la solution proposée (pondérer par le degré du noeud) apparaît comme une heuristique qui manque de fondement biologique.

Toujours concernant le calcul de chemins réalistes entre composés, on peut citer le travail de Arita [Arita, 2004]. Dans ce travail, chaque composé est lui-même vu comme un graphe où les noeuds sont les atomes de carbones et les arêtes sont les liaisons entre atomes. Une réaction est alors considérée comme une opération de transfert d'atomes. Dès lors qu'on travaille à ce niveau, on peut déterminer quelle part d'un substrat est transféré dans un produit. L'auteur définit alors une voie métabolique comme un chemin entre deux composés tel qu'au moins un atome de carbone est transféré de l'un à l'autre via la série de réactions empruntées. Les distances ainsi calculées s'avèrent plus longues que dans le graphe non traité, remettant ainsi en question des travaux plus récents de Fell et Wagner [Wagner et Fell, 2001] sur les propriétés petit-monde (en anglais "small-world") des réseaux métaboliques dont nous parlerons un peu plus longuement dans la section 2.4.3.

Cette approche par transfert d'atomes de carbone a par la suite été utilisée par [Boyer et Viari, 2003] pour calculer les voies alternatives possibles entre deux composés.

Enfin, on peut noter que, pour dépasser la notion de chemin, qui est fondamentalement ambiguë dans un graphe métabolique, on peut lui préférer la notion d'hyperchemin. La notion d'hyperchemin équilibré minimal (ou mode élémentaire) constitue une des notions centrales de la section suivante.

2.3.2 Matrice stochiométrique et Modèles basés sur des contraintes

Les modèles basés sur les contraintes (en anglais "constraint-based models") constituent un domaine de recherche très actif autour du métabolisme.

Du point de vue mathématique, l'objet étudié reste un graphe (un hypergraphe) auquel on rajoute des coefficients (les coefficients stochiométriques). L'ensemble (l'hypergraphe et ses coefficients) est généralement représenté par une matrice, la matrice stochiométrique.

Jusqu'à présent, nous avons toujours considéré (implicitement) qu'une réaction qui transforme A en B transformait une molécule de A en une molécule de B. Or dans certains cas, on peut avoir 1 molécule de A transformée en 2 molécules de B. Ce sont ces proportions que l'on appelle les coefficients stochiométriques (le coefficient stoichiométrique de B est donc égal à 2 pour cette réaction, celui de A est égal à 1).

La matrice stochiométrique S contient ces coefficients. Elle est constituée de n lignes et m colonnes, n étant le nombre de métabolites internes et m le nombre de réactions. La case $S(i, j)$ contient le coefficient stochiométrique

signé $s_{i,j}$ du métabolite i pour la réaction j avec la convention de signe suivante : négatif pour les substrats et positif pour les produits. Enfin, l'ensemble des réactions est divisé en deux sous-ensembles : Rev et $Irrev$, respectivement l'ensemble de réactions réversibles et irréversibles.

Ce qui est étudié dans les modèles basés sur les contraintes, ce sont les distributions de flux de matière à travers les réactions, sous certaines contraintes. Les principales contraintes considérées sont 1. l'hypothèse d'état stationnaire (chaque métabolite interne qui est produit doit être consommé) et 2. les contraintes dues à l'irréversibilité de certaines réactions. D'autres contraintes peuvent être ajoutées.

Par la suite, on utilise la notion de vecteur de flux qu'on note v. Un vecteur de flux (ou distribution de flux) est un m-vecteur de l'espace des réactions $\Re^m$, dont l'élément v_i décrit le flux à travers la réaction i. Dans ce contexte le terme de flux est équivalent à la vitesse nette de la réaction (en anglais "net rate"), c'est-à-dire la différence entre la vitesse de la réaction directe et celle de la réaction inverse.

Les deux contraintes mentionnées précédemment s'expriment de la manière suivante :

1. $Sv = 0$
2. $v_i > 0, \forall i \in Irrev$

et définissent un cône convexe dans l'espace de flux. Le cône est un élément central des modèles basés sur les contraintes.

Deux problèmes principaux sont généralement étudiés à partir de ce cône. Le premier est connu sous le nom de "Flux Balance Analysis" et consiste à trouver un vecteur de flux admissible qui optimise une certaine fonction objective (le flux à travers une réaction par exemple). Dans la littérature, la fonction objective considérée est généralement une pseudo-réaction symbolisant la biomasse. On note que ce problème d'optimisation peut être résolu par programmation linéaire. Parmi les nombreuses applications de ce type d'analyse, on peut mentionner [Edwards *et al.*, 2001].

Le second problème consiste à caractériser le cône en déterminant un ensemble de vecteurs capables de le générer. Cette question de caractérisation du cône a été posée à plusieurs reprises et dans des domaines différents par le passé [Colom et Silva, 1991]. Plusieurs concepts similaires coexistent aujourd'hui. On peut mentionner les notions de T-invariant minimal [Colom et Silva, 1991], courant extrême (en anglais "extreme currents") [Clarke, 1981], mode élémentaire [Schuster et Hilgetag, 1994], et voie extrême (en anglais "extreme pathway") [Schilling *et al.*, 1999].

Les deux premiers ne sont définis que dans le cas où toutes les réactions sont irréversibles (le cône est alors un cône pointé). On peut d'ailleurs noter que dans le cas d'un cône pointé, tous ces concepts sont équivalents et forment une base convexe du cône. Cette base est unique dans le cas d'un cône pointé.

Les concepts de mode élémentaire et de voie extrême ont été introduits

pour traiter le cas plus général d'un réseau contenant des réactions réversibles (cône non pointé). Les deux définitions sont différentes précisément dans leur façon de traiter les réactions réversibles.

L'ensemble des voies extrêmes constitue une base convexe du cône non pointé. Cette base n'est pas unique dans le cas général, mais l'unicité peut être assurée dans certains cas particuliers (qui nécessitent une reconfiguration du réseau décrite dans [Schilling *et al.*, 2000, Klamt et Stelling, 2003]).

L'ensemble des modes élémentaires ne constitue pas une base convexe mais représente cependant une famille génératrice et l'unicité est toujours garantie [Schuster et Hilgetag, 1994].

Un mode élémentaire est défini comme un vecteur de flux admissible (*i.e.* satisfaisant les contraintes 1 et 2) de support minimal. Formellement, si on note $R(v) = \{ j \mid v_j \neq 0\}$ l'ensemble des réactions participant (avec un flux non nul) à v, alors la condition de support minimal s'écrit :

3. il n'existe pas de vecteur de flux admissible non trivial r tel que $R(r) \subset R(v)$.

La figure 2.6, issue de [Papin *et al.*, 2003] illustre la différence entre les concepts de mode élémentaire et voie extrême sur un exemple :

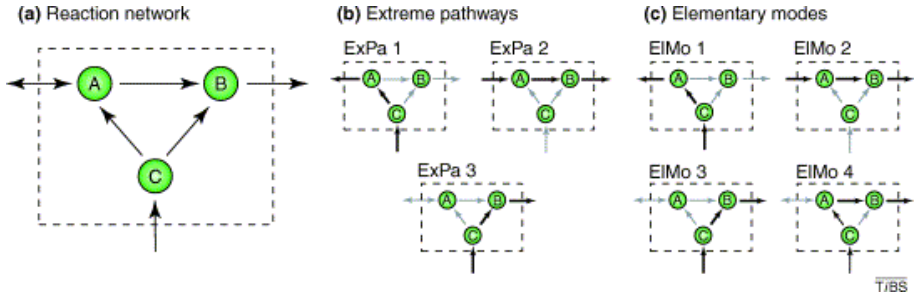


FIG. 2.6 – Modes élémentaires et voies extrêmes dans un réseau de réactions.

Le calcul des modes élémentaires est un problème difficile. De nombreux algorithmes ont été proposés pour le résoudre.

On peut noter que dans [Gagneur et Klamt, 2004], les auteurs proposent une transformation simple du cône non pointé en un cône pointé de dimension supérieure (la transformation consiste à séparer chaque réaction réversible en deux réactions irréversibles) et montrent que les modes élémentaires du réseau reconfiguré contiennent les modes élémentaires du réseau initial, ainsi que des modes élémentaires futiles constitués de cycles de taille 2 qui correspondent aux réactions réversibles. On se retrouve donc dans le cas favorable d'un cône pointé. N'importe quel algorithme d'énumération des rayons extrêmes d'un cône peut alors être appliqué.

Pour plus de détails sur l'analyse de la complexité de l'énumération des modes élémentaires (et de problèmes liés), le lecteur peut consulter l'article joint en annexe C. Le travail présenté dans cet article est le fruit d'une col-

laboration avec Leen Stougie, Alberto Marchetti-Spaccamela, Flavio Chierichetti et Marie-France Sagot.

En pratique, le nombre de modes élémentaires croît très rapidement avec la taille du réseau métabolique considéré. Pour donner un ordre d'idée, un réseau de 110 réactions peut contenir jusqu'à 550 000 modes élémentaires [Klamt et Stelling, 2002]. Certains auteurs [Schuster *et al.*, 2002] proposent de retirer les composés les plus fréquents de l'analyse, ce qui permet de réduire significativement le temps d'exécution et le nombre de solutions.

De manière générale, devant un si grand nombre de solutions, se pose la question de leur structure. Peut-on les regrouper en classes ? Selon quels critères ? Ces questions sont actuellement ouvertes. Nous sommes en train de tester différentes méthodes de regroupement.

2.3.3 Réseaux de Petri

Historiquement, Carl Adam Petri a introduit les premiers concepts des réseaux de Petri lors de sa thèse en 1962 dans le contexte des systèmes techniques [Petri, 1962].

Depuis, de nombreux théorèmes et algorithmes ont été proposés pour analyser les réseaux de Petri, voir par exemple [Murata, 1989]. Les premières applications des réseaux de Petri à la biologie ont été publiés dans [Reddy *et al.*, 1993, Hofestädt, 1994, Reddy *et al.*, 1996]. Pour une revue récente sur ces applications, voir [Hardy et Robillard, 2004] et pour un travail récent concernant l'application des réseaux de Petri au métabolisme, on peut consulter [Voss *et al.*, 2003].

Il existe de nombreux liens entre les réseaux de Petri et les modèles présentés précédemment. Ainsi, un réseau de Petri peut être vu comme une généralisation d'un graphe biparti. On a en effet deux types de noeuds, les places et les transitions. Dans le contexte du métabolisme, les places représentent les métabolites et les transitions les réactions. Les places et les transitions sont reliées par des arcs orientés pondérés par les coefficients stochiométriques. Enfin, le marquage définit pour chaque place le nombre de jetons qui y sont présents (un jeton peut être interprété comme une certaine quantité de métabolites). Un marquage définit ainsi un état du système. On peut donc étudier la dynamique (discrète) du système en étudiant la succession d'états. La notion de T-invariant minimal, qui a été introduite initialement pour caractériser les invariants d'un réseau de Petri, correspond exactement au concept de mode élémentaire (dans le cas d'un réseau de réactions irréversibles), et plus généralement au rayon extrême du cône convexe caractérisant le système [Schuster *et al.*, 2002].

De manière générale, le formalisme des réseaux de Petri est assez riche. Au même titre qu'on trouve une grande diversité de modèles de graphes, il existe aussi une grande diversité à l'intérieur des réseaux de Petri. Il y a ainsi des extensions permettant d'inclure le temps, de différencier les places et d'introduire de la hiérarchie.

Par ailleurs, de nombreux outils d'édition et de simulation faciles à utiliser sont disponibles pour travailler avec les réseaux de Petri, ce qui peut faciliter le dialogue entre modélisateurs et expérimentateurs.

Enfin, les réseaux de Petri dits continus permettent d'assurer un trait d'union entre les modèles statiques et les modèles dynamiques.

2.3.4 Équations différentielles

Les modèles présentés précédemment permettent essentiellement d'étudier la structure d'un réseau métabolique. Lorsqu'on souhaite étudier la dynamique du réseau, on utilise généralement des équations différentielles.

Nous avons peu travaillé sur ce type de modèles, nous les présenterons donc ici de manière plus succincte. Le lecteur intéressé peut notamment consulter [Szallasi *et al.*, 2006] pour une présentation générale de méthodes permettant de travailler sur la dynamique des réseaux biologiques. Dans ce type de modèle, on considère en effet généralement que la structure du réseau est connue et on s'intéresse uniquement à sa dynamique.

Il existe en fait différents types de modèles à base d'équations différentielles. L'idée commune est de décrire par des équations les variations de concentrations de chacun des composants du système. La variante la plus simple utilise des équations différentielles ordinaires. Cependant, pour ajouter du réalisme, on peut jouer sur au moins trois paramètres qui sont le temps, l'espace et la matière. On peut considérer chacun d'eux comme discret ou continu. On distingue ainsi entre modèles à temps discret et continu. Parmi les modèles qui prennent en compte l'espace, on distingue de même entre espace discret (compartimentalisation) et continu (diffusion). Enfin, la matière peut être prise sous forme discrète (modèles stochastiques) pour rendre compte de phénomènes dus à un faible nombre de molécules, ou continu (on traite alors avec des concentrations).

Un frein à l'utilisation des équations différentielles pour étudier le métabolisme est le nombre de paramètres à estimer. De manière générale, plus le modèle est réaliste, plus il contient de paramètres. Certains auteurs argumentent cependant qu'il n'est pas nécessaire d'estimer avec précision tous les paramètres du système [Szallasi *et al.*, 2006]. On peut mentionner ici les travaux de [Tucker et Moulton, 2005] qui, pour estimer les paramètres d'un système d'équations différentielles à partir de profils de métabolites, proposent d'utiliser l'analyse par intervalle pour déterminer les plages de valeurs que les paramètres ne peuvent pas prendre. Par ailleurs, différents travaux essaient d'éviter d'avoir à évaluer la valeur de tous les paramètres du modèle, notamment en travaillant sur des modèles possédant moins de paramètres [Fiévet *et al.*, 2006] ou en utilisant des modèles linéaires par morceaux qui nécessitent d'ordonner les paramètres sans avoir à les estimer [Jong *et al.*, 2004].

On peut noter que, historiquement, les études dynamiques sur le métabolisme portaient sur des enzymes isolées. La régulation du métabolisme était principalement abordée enzyme par enzyme. Or comprendre la régulation métabolique n'est pas seulement comprendre ses composants. La théorie du contrôle métabolique développée par [Kacser et Burns, 1973] en Ecosse et [Heinrich et Rapoport, 1974] en Allemagne peut être conçue comme une réaction contre une insistance excessive sur le rôle des enzymes individuelles. Cette théorie, décrite plus récemment dans [Cornish-Bowden, 1995], permet d'expliquer l'insuccès des efforts entrepris pour accroître les rendements des produits utiles par augmentation des activités des enzymes dites limitantes (une enzyme limitante est une enzyme clé dont l'activité détermine la vitesse de l'ensemble d'une voie métabolique). Ces auteurs estiment que le concept d'enzyme limitante d'une voie métabolique n'a pas de sens. En réalité, chaque enzyme contribue au contrôle du flux dans des proportions qui ne peuvent pas être prévues *a priori*. De plus, même si une seule enzyme possède une part importante du contrôle total du flux, cette part diminue toujours si l'activité de l'enzyme est augmentée.

Pour une introduction plus formelle de la théorie du contrôle métabolique, on peut se reporter à [Reder, 1988].

2.4 Analyse structurelle du métabolisme, les grandes questions

La section précédente a présenté les différents modèles disponibles pour étudier le métabolisme. Cette section va tenter de résumer quelles sont les principales questions qui ont pu être posées concernant la structure du métabolisme (la régulation ne sera pas abordée).

Certaines questions découlent directement des méthodes qui ont été employées et ne correspondent donc pas toujours à une réelle question biologique. À l'inverse, certaines questions biologiques n'ont peut-être pas encore trouvé de réponse car les modèles précédemment introduits ne sont pas assez réalistes. Un des objectifs de cet exposé est de faire un état des lieux du décalage qui peut exister aujourd'hui entre questions et méthodes.

Le type de questions qui vont être abordées dans cette section sont : qu'est-ce que la structure du réseau peut nous apprendre sur son fonctionnement et son évolution ? y a-t-il un lien entre topologie et fonction ? entre topologie et évolution ?

2.4.1 Analyses topologiques classiques

Dès lors qu'on peut modéliser un objet par un graphe, on peut utiliser des outils classiques de la théorie des graphes (précédemment développés pour d'autres applications) et voir s'ils sont pertinents dans ce nouveau contexte.

Une synthèse récente [Zhu *et al.*, 2007] répertorie différentes mesures classiquement utilisées : le degré, la distance, le diamètre, le coefficient de regroupement (en anglais "clustering"), la "betweenness centrality" (une nouvelle mesure de centralité). On peut noter que ces mesures ont été plus utilisées dans le cadre de l'analyse des réseaux d'interaction de protéines que dans celui des réseaux métaboliques.

Le **degré** d'un noeud i, noté k_i, est le nombre d'arêtes qui le lient à d'autres noeuds du graphe. Pour un graphe orienté, on peut différencier entre degré entrant et degré sortant.

La **distance** entre deux noeuds i et j, notée d_{ij}, est la longueur du plus court chemin entre ces deux noeuds (éventuellement le plus court chemin n'est pas unique).

Le **diamètre** d'un réseau, noté D, est la distance maximum entre toutes les paires de noeuds du réseau ; $D = max\{d_{ij}|i,j \in V\}$, avec V l'ensemble des noeuds du réseau.

Le **coefficient de clustering** d'un noeud i, noté c_i, est la proportion de liens existants entre les voisins d'un noeud ; $c_i = \frac{2e_i}{k_i(k_i-1}$ avec e_i le nombre d'arêtes entre voisins de i.

La **"betweenness centrality"** d'un noeud i, notée $C_B(i)$, est la proportion, parmi tous les plus courts chemins entre toutes les paires de noeuds du réseau, de ceux qui passent par ce noeud ; $C_B(i) = \sum_{jk} \sigma_{jk}(i)/\sigma_{jk}$, avec σ_{jk} le nombre de plus courts chemins entre j et k et σ_{jk} le nombre de plus courts chemins entre j et k qui passent par i.

Nous allons introduire à présent deux travaux qui se fondent sur de telles mesures structurelles globales et en donnent une interprétation biologique. Ces travaux ont rencontré un franc succès dans la communauté car ils proposent des outils simples pour aborder des phénomènes complexes. Nous essaierons de montrer cependant que les conclusions biologiques que l'on peut tirer de tels travaux sont parfois discutables.

2.4.2 Les réseaux "scale-free"

Dans un article de 2000 [Jeong *et al.*, 2000], Barabasi et collaborateurs ont proposé un modèle générique qui s'ajusterait à la distribution des noeuds de différents réseaux métaboliques. Les réseaux métaboliques feraient ainsi partie d'une classe de réseaux robustes, tolérants aux erreurs [Albert *et al.*, 2000], les réseaux "scale-free" (cette classe de réseaux comprend aussi notamment le réseau internet). Depuis, de nombreux articles sont parus à ce sujet, montrant que différents types de réseaux biologiques avaient également cette propriété [Wuchty, 2001].

Afin de décrire ce qu'est un réseau "scale-free", considérons tout d'abord la fonction de probabilité $p(k)$, qui donne la probabilité pour un noeud choisi aléatoirement d'avoir k arêtes qui lui sont connectées. Formellement, un réseau est considéré "scale-free" si pour tout k_1, k_2, le rapport $p(k_1)/p(k_2)$

est invariant par multiplication de k_1 et k_2 :

$$\frac{p(k_1)}{p(k_2)} = \frac{p(\alpha k_1)}{p(\alpha k_2)} = F(\frac{k_1}{k_2})$$

où α est une constante positive et F est nommée fonction de changement d'échelle (en anglais "rescaling"). On peut montrer que cette propriété est satisfaite si et seulement si la fonction de probabilité $p(k)$ suit une loi puissance, *i.e.* $p(k) \propto k^{-\gamma}$ [Przytycka et Yu, 2004].

Une des particularités de ce type de réseaux (et qui les différencient par exemple des graphes d'Erdös-Rényi pour lesquels la distribution du degré des noeuds suit une loi de Poisson) est que peu de noeuds sont très connectés et un grand nombre de noeuds sont très peu connectés. Les noeuds très connectés sont parfois appelés "hubs" et sont présumés avoir un rôle particulier dans le réseau. Par exemple, dans [Jeong *et al.*, 2001], les auteurs argumentent que, dans le cadre des réseaux d'interaction de protéines, les "hubs" correspondent à des protéines essentielles. Récemment, plusieurs travaux sont parus qui remettent en cause une partie de ces résultats.

Tout d'abord, on peut se poser la question : les réseaux considérés ont-ils réellement une distribution de degrés qui suit une loi puissance ?

Dans un travail récent [Khanin et Wit, 2006], les auteurs montrent que plusieurs réseaux considérés comme "scale-free" par des tests statistiques simplifiés (ajustement d'une droite pour des données transformées en log), peuvent ne plus être considérés comme tels lorsqu'on utilise des tests d'ajustement plus rigoureux.

Par ailleurs, les réseaux considérés correspondent généralement à des données partielles (reflétant notre connaissance incomplète du processus étudié). Dès lors, si on prouve que le réseau qu'on étudie possède une propriété, cela ne signifie pas nécessairement que le réseau complet (dont il est extrait) possède cette propriété. Ainsi, dans [Stumpf *et al.*, 2005], les auteurs montrent qu'un sous-réseau peut très bien être "scale-free" alors que le réseau dont il est extrait ne l'est pas.

Enfin, l'interprétation qui consiste à dire qu'un réseau scale-free est plus fragile à des attaques ciblées qu'à des attaques aléatoires est peut-être valide quand on parle du web mais est sans doute plus discutable quand on parle de réseaux biologiques. Ainsi, [Coulomb *et al.*, 2005] ont montré qu'il y avait une corrélation très faible entre l'essentialité d'une protéine et ses caractéristiques topologiques dans un réseau d'interaction de protéines.

Pour finir, certains auteurs [Keller, 2005] argumentent que le fait de savoir qu'un réseau appartient à la classe des réseaux "scale-free" ne nous apprend rien sur ce réseau, puisque cette classe est trop générale. Il semble donc nécessaire de développer des outils plus fins pour parvenir à un niveau de description qui devienne explicatif.

2.4.3 Les réseaux petit-monde

Le second travail dont nous aimerions parler à présent est issu des travaux de Watts et Strogatz [Watts et Strogatz, 1998] qui, dans le but de modéliser de manière réaliste la structure globale de réseaux réels (le réseau de neurones du nématode *Caenorabditis elegans*, le réseau électrique de l'ouest des États-Unis, et le graphe de collaboration des acteurs de films), proposent un modèle de réseau intermédiaire entre des réseaux réguliers [1] et des réseaux aléatoires. Une des caractéristiques de leur modèle est que les réseaux peuvent avoir un coefficient de clustering élevé (comme des réseaux réguliers) et pourtant avoir des longueurs de chemin très courtes (comme des réseaux aléatoires). Ce type de réseau a été baptisé réseau petit-monde (en anglais "small-world"), par analogie avec le phénomène petit-monde (connu sous le nom de 6 degrés de séparation [2]).

Dans leurs travaux, Watts et Strogatz proposent une méthode de construction de réseaux petit-monde. On prend pour point de départ un réseau régulier avec n noeuds et k arêtes par noeud puis on déconnecte et on rebranche chaque arête au hasard avec une probabilité p. Cette méthode de construction permet d'ajuster le graphe entre régularité ($p = 0$) et désordre ($p = 1$). Les auteurs montrent que c'est principalement l'ajout d'arêtes entre des noeuds distants (court-circuits) qui cause une diminution brutale du diamètre du graphe et lui attribue cette propriété petit-monde (diamètre faible, coefficient de clustering élevé) [3].

Peu de temps après, Fell et Wagner ont montré que le réseau métabolique d'*Escherichia coli* possédait les propriétés des réseaux petit-monde [Fell et Wagner, 2000, Wagner et Fell, 2001]. Les auteurs argumentent par la suite que ce type d'architecture permet de minimiser les temps de transition entre états métaboliques et contient des indices sur l'histoire évolutive du métabolisme.

En 2004, [Arita, 2004] montre cependant que le modèle de graphe utilisé par Fell et Wagner pour calculer le diamètre du réseau métabolique n'est pas suffisamment réaliste. Comme nous l'avons mentionné dans la section 2.3.1.5, Arita propose de calculer des chemins non pas entre composés mais entre les atomes de ces composés. Il montre alors que le diamètre du réseau est beaucoup plus grand qu'estimé initialement.

[1] Un réseau régulier est un réseau où chaque noeud a le même nombre de voisins.

[2] L'idée de 6 degrés de séparation évoque la possibilité que toute personne sur le globe peut être reliée à n'importe quelle autre au travers d'une chaîne de relations individuelles comprenant au plus cinq autres maillons.

[3] Une autre définition de la notion de réseau petit-monde a par la suite été donnée par Jon Kleinberg [Kleinberg, 2000]. Un graphe petit-monde est un graphe où le routage glouton peut être fait en temps *logn*. Cette définition s'appuie sur la notion de décision locale et permet d'enrichir la définition donnée par Watts et Strogatz. Cependant, ainsi définie, la notion de réseau petit-monde semble moins adaptée au métabolisme pour lesquels les noeuds ne sont pas des acteurs pouvant effectuer un choix (contrairement aux réseaux sociaux).

Ces travaux sont illustratifs de l'enthousiasme initial de la communauté bioinformatique autour de résultats généraux sur la structure des réseaux biologiques. On peut commenter que ce type d'approche qui consiste à rechercher des lois générales (comme il en existe en physique) n'est pas nécessairement très adaptée dans le cas des réseaux biologiques. La structure du métabolisme semble ne pas pouvoir se réduire à de grands principes, certes séduisants, mais souvent peu explicatifs.

Dans la section suivante, nous allons discuter un certain nombre de concepts qu'il semble intéressant de définir pour appréhender la structure du métabolisme.

2.4.4 Complexité, robustesse et modularité

Parmi les concepts biologiques qui apparaissent de manière récurrente dans la littérature, on retrouve les notions de **complexité**, **robustesse** et **modularité**. Dans cette section, nous tenterons de donner une définition à chacun de ces concepts et de les illustrer avec des exemples pris dans la littérature. On peut noter que ces concepts peuvent être définis à la fois dans un contexte structurel et dans un contexte dynamique ; nous discuterons des deux.

2.4.4.1 Complexité

Le terme de complexité apparaît dans de nombreux domaines scientifiques. Les acceptions du terme varient selon le domaine mais un trait commun est l'opposition entre complexité et simplicité. Un phénomène complexe est un phénomène qu'on ne peut pas comprendre simplement. Dans [Szallasi *et al.*, 2006], on trouve la définition suivante : un système complexe est un système dont les propriétés ne sont pas totalement expliquées par la compréhension des parties qui le constituent. On peut noter que le terme émergence est parfois employé avec un sens proche. On parle par exemple de propriété émergente pour qualifier une propriété d'un système qui n'apparaît pas quand on étudie les parties séparément.

Pour mieux appréhender cette notion de complexité, on peut comparer les différences entre disciplines.

La complexité dans les systèmes biologiques peut ainsi être comparée à la complexité dans les systèmes physiques (modèles de tas de sable par exemple). Dans les deux cas, on a des entités en interactions et ces interactions produisent des comportements complexes. Dans le cas de la biologie, les entités sont **hétérogènes** et **structurées spatialement**.

Par ailleurs, récemment, de nombreux auteurs ont comparé la complexité des systèmes biologiques avec la complexité des systèmes conçus par l'homme ("engineered systems") [Csete et Doyle, 2002, Alon, 2003]. Un des arguments pour rapprocher ces deux types de systèmes est que, contrairement aux

systèmes physiques, ils sont tous les deux soumis à des contraintes fonctionnelles, et ne sont donc pas organisés aléatoirement. Le parallèle entre systèmes biologiques et systèmes conçus par l'homme semble être relativement fécond lorsqu'on cherche à modéliser le comportement dynamique du système. On peut cependant d'ores et déjà signaler un obstacle majeur à cette analogie.

En effet, comme le suggère François Jacob [Jacob, 1977], l'évolution n'est pas un ingénieur mais un bricoleur (en anglais "tinkerer"). L'ingénieur est capable de savoir *a priori* l'utilisation future de son travail, l'évolution ne le peut pas. L'ingénieur peut choisir un plan de conception optimal parmi une gamme de possibilités, alors que l'évolution doit composer avec des structures préexistantes.

On peut d'ailleurs souligner le fait que (en partie à cause d'un parallèle incorrect entre systèmes biologiques et systèmes faits par l'homme) de nombreux auteurs cherchent à montrer des propriétés d'optimalité pour les systèmes biologiques. Or si les systèmes biologiques sont amenés à développer des solutions qui peuvent se rapprocher de solutions optimales, c'est avec la condition de réutiliser des structures préexistantes, ce qui est souvent mal pris en compte dans la modélisation.

2.4.4.2 Robustesse

La notion de robustesse est liée à la notion de perturbation. Quand on parle de robustesse, on doit spécifier deux choses : quelle fonctionalité est robuste et contre quelle perturbation elle est robuste. Par exemple, la composition en acides aminés d'une protéine est robuste et elle est robuste aux mutations. De manière générale, on dira qu'un système biologique est robuste s'il continue à fonctionner même sous l'effet de perturbations. Les perturbations peuvent être de deux types : environnementales et génétiques. Dans le cadre des réseaux biologiques, on peut ajouter un troisième type de perturbation : la stochasticité (due à un faible nombre de molécules) [Szallasi *et al.*, 2006].

Dans un ouvrage récent [Wagner, 2005], Andreas Wagner propose de parcourir tous les niveaux d'étude du vivant, de la molécule à la population, pour illustrer l'omniprésence de la robustesse en biologie.

Une idée générale qu'il développe dans son livre, et qu'il semble intéressant d'exposer ici, est la notion d'*espace neutre*. En effet, si on considère un problème biologique, par exemple, coder une protéine capable de catalyser une réaction métabolique, il existe bien sûr plusieurs solutions à ce problème (plusieurs séquences d'acides aminés sont possibles). On peut même dire qu'il existe plusieurs solutions pour obtenir le même résultat. Un ensemble de solutions équivalentes constitue, selon Wagner, un espace neutre. L'auteur ajoute qu'un espace neutre n'est pas nécessairement homogène et que certaines solutions peuvent, dans certaines conditions, être meilleures que d'autres. De manière générale, un espace neutre large constitue donc un réservoir de solutions qui peut être avantageux à l'échelle de temps de l'évolution. La ro-

bustesse s'apparente ici à la notion d'évolutivité. On peut noter que cette idée d'espace neutre peut être mise en parallèle avec la notion de paysage adaptatif, introduite par [Wright, 1932].

La notion de robustesse telle qu'elle est proposée dans [Szallasi *et al.*, 2006] est plus opérationnelle. Il s'agit toujours d'assurer le maintien de fonctionalités clés sous l'effet de perturbations, mais à une échelle de temps beaucoup plus courte. Une question essentielle qui y est abordée est : comment mesurer la robustesse d'un système ? Les auteurs proposent de prendre pour mesure la sensibilité des paramètres d'un système d'équations différentielles. Un exemple d'étude de robustesse sur le réseau de régulation du cycle circadien de la drosophile illustre le fait que tous les paramètres d'un système ne sont pas aussi sensibles. Les auteurs parlent de compromis (en anglais "trade-off") entre robustesse des paramètres globaux et robustesse des paramètres locaux. Une des implications de la robustesse pour la modélisation est donc que si le système est très robuste, alors on n'a pas besoin de connaître précisément les valeurs des paramètres pour un modèle quantitatif. Le plus important est de connaître la structure du réseau. Les valeurs exactes ne sont nécessaires que pour certains paramètres clé. Cette observation a des conséquences majeures pour la modélisation quantitative de grands réseaux.

Pour résumer, on peut voir la robustesse à deux échelles de temps différentes. La première, à l'échelle de l'évolution, peut être reliée à la notion d'évolutivité (un système est robuste dans le temps s'il a une propension à évoluer, si son espace neutre est suffisamment large). La seconde est plus opérationnelle et concerne la robustesse des paramètres d'un modèle décrivant la dynamique d'un système. Cette seconde acception a des conséquences importantes pour la modélisation quantitative de grands réseaux.

2.4.4.3 Modularité

Nous verrons que la notion de modularité est à la fois une notion fondamentale (le fait même qu'un système soit modulaire est informatif, et une appréhension de haut niveau en modules permet de donner un éclairage pertinent sur le système complet) et une notion opérationnelle (le découplage d'un système en modules indépendants permet d'opérer des simplifications de calculs importantes, et notamment d'étudier les modules indépendamment les uns des autres).

Cette section est organisée de la manière suivante : nous allons d'abord donner quelques propriétés générales de la modularité et montrer que la modularité est présente en biologie, puis nous discuterons plus précisément de la définition qu'on peut donner du concept de module et des techniques d'identification qui en découlent. Enfin, nous aborderons la question de la validation des modules identifiés.

Introduction

Il existe plusieurs définitions de modules en biologie mais toutes ont un point en commun : un module est caractérisé par son indépendance avec le reste du système. L'indépendance peut être spatiale ou temporelle, chimique ou génétique, structurelle ou dynamique, selon le point de vue que l'on souhaite adopter.

La modularité est présente en biologie de manière évidente à un haut niveau. La transplantation d'organe en fournit un très bon exemple. Une population est formée d'individus, un individu est formé d'organes, un organe est formé de cellules. À bas niveau, une molécule peut également être vue comme un module, les protéines peuvent ainsi être découpées en domaines qui constituent des modules indépendants.

Mais au niveau intermédiaire entre molécule et cellule, l'image est moins claire. De nombreux auteurs argumentent pourtant en faveur de l'omniprésence de la modularité en biologie, à tous les niveaux d'organisation du vivant, même au niveau cellulaire [Wolf et Arkin, 2003]. Ainsi, dans [Hartwell *et al.*, 1999], de nombreux exemples de modules sont donnés (la réplication de l'ADN, la glycolyse, la synthèse des protéines) dont certains ont pu être reconstruits *in vitro*, ce qui en soi constitue un excellent critère de validation de la modularité. Avec un point de vue de biologiste évolutif, Gunter Wagner prend pour preuve de modularité le fait même que l'évolution phénotypique puisse être étudiée caractère par caractère [Wagner, 1996]. En effet, si un caractère évolue indépendamment des autres, cela montre que les bases génétiques qui sont responsables de ce caractère constituent un module évolutif. L'auteur propose alors un modèle expliquant l'apparition et le maintien de modules au cours de l'évolution.

Identification

Mais au-delà de ces manifestations de la modularité et de la question de leur origine, nous aimerions poser la question : étant donné un système, quels en sont les modules ?

On peut d'ores et déjà signaler que le cadre théorique pour définir la modularité du point de vue dynamique est très peu développé pour l'instant. Ainsi, on se centrera surtout sur la notion de modularité structurelle.

Il semble exister essentiellement deux types d'approches pour identifier des modules dans le contexte des réseaux biologiques : une approche "bottom-up" et une approche "top-down" [Szallasi *et al.*, 2006].

L'approche "bottom-up" consiste à assembler des composants jusqu'à observer une indépendance vis-à-vis du reste. Un exemple réussi de ce type d'approche est la reconstruction du réseau des gènes de polarité segmentaire impliqués dans le développement embryonnaire de la drosophile [von Dassow *et al.*, 2000]. On peut noter que, dans ce cas, le module était en fait déjà presque connu car

le système avait été très étudié et chacun des composants bien caractérisé. L'essentiel du travail correspondait donc à l'assemblage des connaissances et à la validation de l'autonomie de ce réseau. Ce type d'approche ne peut donc pas être appliqué à n'importe quel système, il requiert un long travail préliminaire de caractérisation des composants du système.

Les approches "top-down" prennent un point de vue inverse. L'idée est de partir du réseau complet (issu d'une expérience à haut débit pour un réseau d'interaction de protéines ou un réseau de régulation de gènes, ou d'un travail de reconstruction pour un réseau métabolique) et d'en dégager des modules. Nous allons discuter cette approche plus en détail en précisant les définitions de modules employées et les méthodes d'identification qui en découlent. L'idée commune de ces méthodes est de dégager des sous-structures qui soient peu connectées avec le reste du réseau mais dont les éléments sont très connectés. Nous allons en particulier distinguer deux cas : 1. les modules identifiés couvrent tout le réseau (chaque élément du réseau est classé dans un module) et 2. seules certaines parties sont couvertes.

Concernant le second type d'approche, on peut mentionner les travaux de [Spirin et Mirny, 2003] qui recherchent des modules dans des réseaux d'interaction de protéines (modélisés par des graphes où les noeuds représentent les protéines et les arêtes représentent les interactions entre protéines). Un module est alors défini comme un sous-graphe dense. La densité d'un sous-graphe est donnée par la fonction $Q(m, n) = 2m/(n(n-1))$, où m est le nombre d'interactions entre les n noeuds du sous-graphe. Un critère statistique permet ensuite de décider si la valeur prise par Q est exceptionnelle (le modèle de graphe aléatoire utilisé, ou modèle nul, est un graphe aléatoire ayant la même séquence de degrés des noeuds que le graphe d'intérêt). On note que, pour pouvoir faire ce travail sur des sous-graphes de grande taille, les auteurs ont recours à des heuristiques (recherche locale, recuit simulé) pour la partie comptage et à des approximations et simulations pour la partie statistique. Il a par ailleurs été montré que le problème de la recherche de sous-graphe induit de poids maximum est un problème NP-difficile [Shamir *et al.*, 2002].

De nombreux travaux ont également porté sur la détection de structures dans les réseaux d'interactions de protéines. On peut mentionner notamment ceux du groupe de Roded Sharan qui portent, d'une part sur la détection efficace de voies de signalisation dans des graphes d'interactions [Scott *et al.*, 2006], et d'autre part sur la détection de complexes protéiques conservés entre organismes [Sharan *et al.*, 2005, Hirsh et Sharan, 2007].

Par ailleurs, le programme Cytoscape implémente un certain nombre de méthodes permettant d'analyser des réseaux d'interaction de protéines et notamment d'identifier des modules [Shannon *et al.*, 2003].

Nous venons donc de décrire brièvement une série de travaux ayant pour but de détecter des modules dans un réseau biologique dans le cas où les modules ne couvrent pas tout le réseau.

Avant de passer à la catégorie suivante, nous pouvons mentionner ici une notion proche de la notion de module que nous venons de décrire : la no-

tion de motif, introduite par le groupe de Uri Alon [Milo *et al.*, 2002]. Un motif est défini par les auteurs comme un pattern de connections répété de manière exceptionnelle (*i.e.* significativement plus que dans des réseaux aléatoires). Comme pour un module, une occurrence du motif est un sous-graphe connexe. La différence principale entre un module et un motif est que pour un motif, ce n'est pas la densité qui est importante mais la répétition. Certains auteurs argumentent que la différence principale entre motif et module est la taille (les motifs étant de petits sous-graphes et les modules de grands sous-graphes)[Alm et Arkin, 2003]. On peut cependant argumenter que la petite taille des motifs étudiés jusqu'à présent est due aux limitations des méthodes utilisées pour leur détection. On retiendra pour l'instant que la différence principale entre module et motif est qu'un motif est répété et qu'un module est autonome.

Revenons maintenant à la situation où tout le réseau doit être couvert de modules (chaque noeud appartient à au moins un module). Dans ce cas, il est à nouveau nécessaire de faire une distinction entre deux types de méthodes : 1. chaque noeud peut appartenir à plusieurs modules (les modules ne forment pas une partition) et 2. chaque noeud appartient à au plus un module (les modules forment une partition)

Nous avons déjà mentionné plusieurs cas où des modules couvraient le réseau mais ne formaient pas une partition dans le cadre des réseaux métaboliques. Par exemple, l'organisation d'un réseau en voies métaboliques en fait partie. En effet, si on considère le découpage d'un réseau en voies métaboliques, une réaction peut appartenir à plusieurs voies.

La notion de mode élémentaire a été proposée comme définition objective de voie métabolique [Schuster *et al.*, 2000]. Les modes élémentaires ne forment pas non plus une partition (une réaction peut appartenir à plusieurs modes élémentaires).

Une difficulté est que le nombre de modes élémentaires peut être très grand (par exemple un réseau de 100 réactions peut avoir 500 000 modes élémentaires). Il en découle que le recouvrement entre modes élémentaires peut être très important, remettant en question la notion d'indépendance entre modules.

On peut aussi utiliser un raffinement des modes élémentaires, les réactions dites couplées [Burgard *et al.*, 2004], pour définir les modules. Ainsi, 2 réactions sont parfaitement couplées si elles participent aux mêmes modes élémentaires. Cette définition de module fait sens (elle peut notamment correspondre à des enzymes qui participent toujours aux mêmes processus métaboliques), mais elle est très restreinte et ne permet pas de classer toutes les réactions dans des modules.

Nous allons maintenant traiter de la question de la recherche d'une partition du réseau (modules non recouvrants). Deux types de méthodes sont ici disponibles, le partitionnement de graphes et la détection de communautés.

Il existe une littérature assez étendue concernant le problème du partition-

nement de graphes qui trouve des applications notamment dans le domaine du calcul parallèle. Pour une synthèse, on peut consulter [Fjallstrom, 1998]. Dans sa formulation la plus simple, le problème de partitionnement de graphe consiste à séparer les noeuds d'un graphe en p sous-ensembles disjoints de même taille, en minimisant le nombre d'arêtes entre sous-ensembles. Ce problème est NP-complet même dans le cas où $p = 2$ [Garey *et al.*, 1976]. On peut noter que, dans les problèmes de partitionnement de graphes, tel que ce problème a été traité jusqu'à maintenant, le nombre de modules et la taille des modules est connue. Ainsi, une application typique de ce type de problème concerne le calcul parallèle où on cherche à répartir de manière égale une charge de calculs sur plusieurs processeurs.

Dans le cadre de la recherche de modules dans les réseaux métaboliques, on ne se trouve généralement pas dans cette situation où le nombre de modules est connu. Les méthodes de partitionnement de graphes ont donc peu été appliquées aux réseaux biologiques en général.

D'autres méthodes sont généralement employées dans le cas où le nombre de modules n'est pas connu. On parle alors de détection de structures en communauté (les termes de "clustering" hiérarchique et de "block modelling" ont également été employés). Ces techniques ont été initialement développées dans le domaine de la sociologie [Wasserman et Faust, 1994].

Le problème consiste toujours à séparer les noeuds du graphe en sous-ensembles disjoints mais le critère à minimiser n'est plus le nombre d'arêtes inter-groupes. En effet, un tel critère impliquerait que la solution optimale serait d'avoir un unique module comprenant le réseau complet. Le critère à optimiser, nommé modularité, correspond, pour chaque module, à la différence entre le nombre d'arêtes intra-module observé et le nombre d'arêtes intra-module attendu sous un modèle neutre [Newman et Girvan, 2004] (le modèle de graphe aléatoire généralement utilisé est un modèle de graphe qui maintient la séquence de degrés des noeuds). En pratique, le nombre d'arêtes observées est donné par la matrice d'adjacence du graphe, et le nombre d'arêtes attendues est donné par $\frac{k_i \times k_j}{2m}$ où k_i est le degré du noeud i et m le nombre d'arêtes du graphe. Le problème de trouver la partition qui maximise ce critère est un problème difficile. Plusieurs méthodes basées sur des heuristiques ont été proposées.

[Guimerà et Amaral, 2005] utilisent ainsi une approche par recuit simulé. Récemment, [Newman, 2006] a proposé une réécriture du problème sous forme matricielle qui lui permet d'utiliser une technique efficace utilisée pour résoudre des problèmes de bipartition de graphes (algorithme spectral).

Dans leur article, [Guimerà et Amaral, 2005] proposent une application de leur méthode au réseau métabolique d'*Escherichia coli*. Les modules identifiés (la méthode en trouve 19) sont ensuite comparés aux voies métaboliques telles qu'elles sont définies dans la base de données KEGG. Pour certains modules, une fonction majoritaire peut être dégagée (métabolisme des acides aminés, métabolisme des lipides...).

La méthode d'optimisation utilisée par Guimerà et Amaral (recuit simulé) est en plusieurs points insatisfaisante : elle ne garantit pas de trouver un optimum global et il est difficile de savoir quel type d'optimum local on a favorisé. On note que la comparaison des modules identifiés avec les voies métaboliques n'est pas non plus une méthode de validation satisfaisante dans la mesure où la notion de voie métabolique n'est pas clairement définie. On peut noter que c'est cependant une des seules méthodes disponible actuellement. La question du critère de validation des modules identifiés est donc encore ouverte.

À la lecture de ces travaux sur la recherche de communautés, il en ressort que, d'une part la définition du critère de modularité est centrale et reste peu discutée, et d'autre part la méthode d'optimisation est encore un problème ouvert.

Par ailleurs, on peut noter que ces mesures de modularité sont calculées pour des graphes. Pour plus de réalisme, il serait intéressant de pouvoir les étendre à des hypergraphes.

Enfin, on peut mentionner une dernière approche qui a été très étudiée en théorie des graphes et qui a été appliquée aux réseaux d'interaction de protéines [Gagneur *et al.*, 2004]. Il s'agit de la décomposition modulaire d'un graphe. Nous allons présenter ici le problème brièvement, pour une synthèse sur le sujet, on peut consulter [Möhring et Radermacher, 1984]. Un module, dans ce cas, est défini comme un ensemble de noeuds tel que, tout noeud extérieur au module et voisin d'un noeud du module, est également voisin de tous les autres noeuds du module. Formellement :

$$\forall x \notin M, \forall y, z \in M, (x, y) \in E \iff (x, z) \in E$$

En général, le nombre de modules d'un graphe peut être exponentiel. Si on se restreint aux modules non recouvrants, alors le nombre de modules reste linéaire. L'ordre d'inclusion sur cet ensemble définit un arbre, l'arbre de décomposition modulaire, qui suffit à reconstruire l'ensemble des modules [Möhring et Radermacher, 1984]. La racine de l'arbre correspond au module trivial V (l'ensemble des noeuds du graphe), et les n feuilles correspondent aux modules triviaux $\{x\}, x \in V$.

La Figure 2.7 illustre un graphe et son arbre de décomposition modulaire.

Un résultat notable est que la décomposition modulaire d'un graphe peut aujourd'hui être obtenue en temps linéaire avec des algorithmes simples à implémenter [Habib *et al.*, 2004].

Pour finir sur les approches dites "top-down", on peut sans doute signaler une critique générale faite par [Szallasi *et al.*, 2006]. Nous avons exposé de nombreuses méthodes qui permettent de dégager des structures d'intérêt dans un graphe. Cependant, ce n'est pas parce qu'un algorithme trouve des modules que le réseau est modulaire et par ailleurs, les modules identifiés doivent être validés biologiquement.

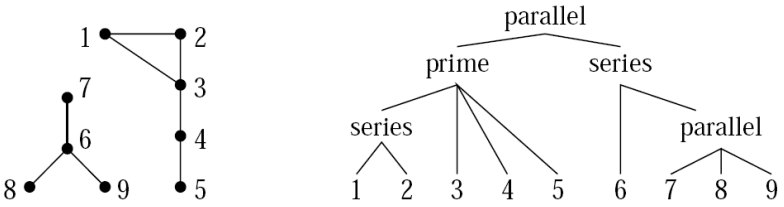


FIG. 2.7 – Un graphe et son arbre de décomposition modulaire. Les ensembles $\{1, 2\}$ et $\{7, 8, 9\}$ sont des modules non recouvrants. Les modules $\{7, 8\}$ et $\{8, 9\}$ sont recouvrants.

Comme nous l'avons vu précédemment pour l'approche "bottom-up", une validation consiste à reproduire *in vitro* le module pour tester son autonomie. Un problème majeur est que ce type de validation peut avoir un coût important.

Wang et Zhang [Wang et Zhang, 2007] ont récemment écrit un article sur la validation biologique des modules dans les réseaux d'interactions protéine-protéine. Les auteurs utilisent l'approche par recherche de communauté pour identifier des modules au sein d'un réseau d'interaction de protéines, puis comparent les modules obtenus pour différents organismes (levure, drosophile, nématode). Une des conclusions de l'étude est que les modules identifiés ne sont pas conservés au cours de l'évolution. Les auteurs présentent alors un point de vue sceptique sur la valeur biologique des modules trouvés dans les réseaux d'interaction de protéines. Ils argumentent qu'il n'est en tout cas pas nécessaire de faire appel à la sélection naturelle pour expliquer l'organisation du réseau en modules. En effet, ce type d'architecture semble très bien s'expliquer par un modèle neutre incluant la duplication de noeuds. On peut noter que cette question de modèle neutre incluant la duplication de gènes avait déjà été mentionnée par [Solé et Valverde, 2006].

Enfin, nous aimerions commenter sur un dernier point : la plupart des auteurs dans le domaine s'accordent à dire que la modularité est présente au niveau des réseaux cellulaires. On pourrait également prendre le parti inverse, en particulier lorsqu'on considère le métabolisme, pour lequel la notion de modularité n'est pas forcément évidente. En effet, on constate souvent qu'il existe un noyau de noeuds dense et une périphérie plus lâche. On pourrait donc préférer au concept de modularité (qui sous-entend qu'on peut découper le réseau en morceaux indépendants) celui de centre et périphérie (certains morceaux peuvent être extraits et d'autres sont inextricablement mêlés). Une question qui pour l'instant est ouverte serait : comment alors définir le centre et la périphérie d'un réseau ? On note que ces notions de centre et périphérie sont notamment utilisées dans [Pál *et al.*, 2005] mais ne sont pas formellement définies.

2.4.5 Évolution du métabolisme

Nous allons à présent aborder des questions plus spécifiques à l'évolution du métabolisme. On peut noter que la notion d'évolution a déjà été abordée dans les parties précédentes de manière indirecte (évolution et modularité, évolution et robustesse) mais nous avons choisi de faire une section séparée pour les travaux dont le thème principal est l'évolution. Les travaux présentés dans cette partie sont tous liés à la question : comment le métabolisme a-t-il évolué ?

Dans cette section, nous allons tout d'abord présenter différents travaux qui traitent de la comparaison de réseaux métaboliques, puis nous tenterons de résumer les différents modèles qui ont été proposés pour modéliser l'évolution du métabolisme, et enfin, nous finirons par les travaux qui traitent à la fois du génome et du métabolisme.

2.4.5.1 Analyse comparative

La comparaison de réseaux métaboliques peut avoir plusieurs applications. Une des principales est sans doute de comprendre l'évolution du métabolisme.

On peut ici faire un parallèle avec la comparaison de séquences. Une des applications de la comparaison de séquences (nucléiques ou protéiques) est la détection d'homologie (deux séquences sont homologues si elles sont issues d'une même séquence ancestrale). L'homologie est généralement déduite de l'information de similarité entre séquences. Des méthodes de reconstruction phylogénétique permettent ensuite de proposer une histoire évolutive pour un ensemble de séquences homologues. Cette histoire évolutive se présente généralement sous la forme d'un arbre.

Le même processus a récemment été appliqué au métabolisme. Nous allons maintenant présenter les travaux qui concernent l'alignement de voies métaboliques et de réseaux métaboliques. Nous présenterons ensuite les travaux qui ont pour but de reconstruire un scénario d'évolution à partir d'un ensemble de voies métaboliques.

D'un point de vue méthodologique, plusieurs travaux ont été menés en ce qui concerne l'alignement de voies métaboliques. Un alignement de voies métaboliques diffère d'un alignement de séquences par deux points principaux : 1. les unités à aligner ne sont plus des nucléotides ou des acides aminés mais des réactions et 2. la structure n'est plus linéaire mais est généralement représentée par un graphe.

Pour aborder le premier point, il est nécessaire de déterminer une distance entre réactions. On peut noter que, dans le cadre d'un alignement de séquences, les distances entre nucléotides ou entre acides aminés sont généralement choisies pour être proportionnelles à la probabilité de mutation d'un nucléotide vers un autre (coût de substitution). Dans un alignement de séquences, une substitution correspond à un mécanisme biologique :

la mutation (fixée par la sélection). Lorsqu'on compare des réactions, on ne peut pas faire le même parallèle entre substitution et mutation. Ainsi, les distances entre réactions sont généralement des distances fonctionnelles qui ne reflètent pas nécessairement une relation évolutive. Cette remarque n'a pas de conséquence si on s'intéresse à la comparaison structurale de voies métaboliques, mais elle peut en avoir si on s'intéresse à leur histoire évolutive.

Dans leurs travaux, [Tohsato *et al.*, 2000] proposent une méthode simple de comparaison de réactions. La méthode de comparaison est basée sur la notion de numéro EC (pour "Enzyme Commission"). Un numéro EC est un code à 4 numéros attribué par l'"International Nomenclature Committee" à chaque enzyme nouvellement découverte. Ce code identifie la chimie de la réaction que l'enzyme catalyse. L'ensemble de ces codes forme une classification hiérarchique qui peut être représentée par un arbre. Le score de similarité entre deux réactions est alors défini comme une fonction de la distance entre les numéros EC dans l'arbre.

Cette distance entre réactions présente l'avantage d'être simple à calculer. Les auteurs l'utilisent dans le cadre de l'alignement de voies métaboliques linéaires. Ils mettent ainsi en évidence des similarités structurelles entre les voies de synthèse de différents acides aminés.

[Pinter *et al.*, 2005] proposent ensuite une généralisation du travail de [Tohsato *et al.*, 2000] pour le cas de voies métaboliques branchées. Le problème sous-jacent qui est traité est lié au problème d'isomorphisme de sous-arbre qui peut être résolu en temps polynomial.

Une limite de cette approche est qu'elle ne s'applique pas à toutes les voies métaboliques. En effet, certaines voies contiennent des cycles et ne peuvent donc pas être représentées par des arbres. Dans ce cas, [Pinter *et al.*, 2005] suggèrent de supprimer arbitrairement une arête et d'appliquer leur méthode. Cette approche n'est bien sûr pas satisfaisante mais elle permet de donner une réponse dans le cadre de la comparaison de voies, où le nombre de cycles est limité. Lorsqu'on passe à la comparaison de réseaux, cette approche n'est plus du tout praticable, le nombre d'arêtes à enlever étant bien plus important.

Toujours concernant la comparaison de voies métaboliques, on peut encore mentionner le travail de [Clemente *et al.*, 2005]. Dans ce travail, les auteurs prennent le parti de ne pas considérer la topologie des voies métaboliques. Ainsi, une voie est vue comme un sac de réactions, et les réactions sont toutes comparées deux à deux, indépendamment de leurs positions dans la voie. En plus de la distance basée sur les numéros EC, est également considérée une distance basée sur la classification GO (pour "Gene Ontology"). On note qu'on reste dans le domaine de la comparaison fonctionnelle entre réactions. En effet, la classification GO permet de décrire la fonction moléculaire, le processus biologique ainsi que le compartiment cellulaire de chaque gène classé.

Nous allons maintenant présenter différents travaux qui traitent de la reconstruction de scénarios d'évolution de voies métaboliques.

En ce qui concerne les méthodes de reconstruction phylogénétique basées

sur les distances (neighbour-joining), il suffit pour obtenir un arbre de fournir en entrée une matrice de distances entre voies métaboliques. Cependant, pour que cet arbre reflète des relations de parenté et non des similarités structurelles, il est nécessaire que les distances entre voies intègrent un modèle réaliste d'évolution du métabolisme (ce qui n'est pas le cas pour les méthodes présentées précédemment).

[Cunchillos et Lecointre, 2003] ont proposé une méthode qui a pour objectif de prédire un ordre d'apparition des voies métaboliques au cours de l'évolution. Une voie métabolique est simplement modélisée par une séquence de présence-absence des enzymes. La reconstruction est faite par la méthode du maximum de parcimonie. On peut souligner ici l'effort pour intégrer un modèle d'évolution biologique qui soit cohérent. On peut remarquer que, dans ce travail, la topologie des voies n'est pas prise en compte.

[Heymans et Singh, 2003] introduisent quant à eux l'information de topologie dans leur approche. De nouveau, une voie métabolique est vue comme l'ensemble des enzymes impliquées. Cette fois, la phylogénie est inférée à l'aide d'une méthode basée sur les distances. Une distance entre voies est donc introduite. Celle-ci prend en compte la similarité de séquences entre les enzymes mais aussi la similarité de séquences entre les enzymes voisines dans la voie métabolique. La topologie n'est donc pas considérée de manière très complète (seul le voisinage immédiat est examiné) mais cela constitue une première avancée dans ce sens.

Enfin, on peut encore signaler le travail de [Liao *et al.*, 2002] qui prend pour objet d'étude non plus une voie mais le réseau métabolique en entier. Une feuille de l'arbre phylogénétique est ici une séquence de présence-absence de voies métaboliques (cette séquence symbolise le réseau). On juge qu'une voie est présente chez un organisme si toutes les enzymes (ou plutôt des orthologues[4]) sont présentes dans son génome.

Pour conclure, on peut constater que plusieurs méthodes de comparaison de réseaux métaboliques commencent à être disponibles. On note qu'il existe encore actuellement des limitations pour utiliser ces travaux afin d'inférer des histoires évolutives. En effet, une des principales limites est que la comparaison de réactions utilisée est généralement une comparaison fonctionnelle. Or une similarité fonctionnelle entre réactions n'est pas toujours le signe d'une relation de parenté. Outre ces questions de comparaison de réactions, il n'est par ailleurs pas clair à l'heure actuelle quelles sont les opérations d'édition qu'il faut autoriser pour modéliser de manière réaliste l'évolution d'une voie métabolique, ou plus généralement d'un réseau métabolique.

Enfin, on retiendra que les méthodes de reconstruction phylogénétique qui sont disponibles actuellement n'utilisent pas l'information de topologie.

[4]Des enzymes orthologues sont des enzymes qui sont issues d'un ancêtre commun et ont divergé suite à un évènement de spéciation

2.4.5.2 Modèles d'évolution du métabolisme

Avec le double objectif de comprendre les mécanismes d'évolution du métabolisme et de fournir une base méthodologique réaliste à des analyses comparatives, nous allons nous intéresser à présent aux différents modèles d'évolution du métabolisme qui ont été proposés dans la littérature.

Modèles biologiques

Dans [Schmidt *et al.*, 2003], cinq scénarios d'évolution du métabolisme sont présentés (voir Figure 2.8).

À l'echelle de la voie métabolique, plusieurs scénarios d'évolution ont été proposés. Premièrement, une voie peut avoir évolué spontanément sans adopter d'enzymes existantes (Fig. 2.8a). Par exemple, différentes tRNA synthétases semblent avoir initialement évolué indépendamment, et ont été plus tard impliquées ensemble dans différentes voies comme la traduction des protéines, la transamidation tRNA dépendante ou l'acylation non discriminante [Min *et al.*, 2002]. Deuxièmement, l'hypothèse de rétro-évolution [Horowitz, 1945] propose que la pression de sélection sur une voie métabolique agit principalement sur la capacité à former le composé final de la voie (Fig. 2.8b). Le fait de pouvoir former ce composé à partir d'un métabolite intermédiaire augmente la fitness d'un organisme. Puisque le produit final peut être dérivé à partir de métabolites de plus en plus distants, la fitness augmente et la voie évolue à l'envers. Ce scénario de rétro-évolution a été proposé pour la glycolyse [Fothergill-Gilmore et Michels, 1993] et la voie du mandelate [Petsko *et al.*, 1993]. Troisièmement, une voie peut avoir évolué à partir d'enzymes multifonctionnelles [Roy, 1999] (Fig. 2.8c). À partir d'une enzyme multifonctionnelle catalysant des étapes consécutives, la voie peut avoir ensuite évolué par duplication et diversification de cette enzyme précurseur vers les enzymes plus spécifiques et efficaces qu'on trouve aujourd'hui, qui catalysent chacune une étape de la voie. Des enzymes multifonctionnelles, sont utilisées aujourd'hui dans diverses voies métaboliques, comme les β-D-glucan hydrolases chez les plantes supérieures, et pourraient constituer des précurseurs pour de nouvelles voies [Hrmova *et al.*, 2002]. Ce scénario suppose qu'une seule enzyme se spécialise, mais il est également possible que des voies entières soit dupliquées et divergent (Fig. 2.8d). Ce mécanisme d'acquisition de nouvelle fonction a été étudié depuis longtemps [Fisher, 1958] et peut aujourd'hui être identifié par génomique comparative [Rison *et al.*, 2002, Huynen et Snel, 2000]. La biosynthèse du tryptophane et de l'histidine fournissent un exemple de ces scénarios [Gerlt et Babbitt, 2001, Jensen, 1976] ; ces deux voies sont constituées de plusieurs étapes qui ont des mécanismes réactionnels similaires et qui sont catalysées par des enzymes homologues. Finalement, les voies métaboliques peuvent aussi avoir évolué par recrutement d'enzymes utilisées dans d'autres voies. Une voie qui a évolué selon ce processus est alors formée d'un "patchwork" d'en-

zymes homologues à des enzymes catalysant des réactions dans des voies différentes (Fig. 2.8e). Des observations indiquent que certains types de repliements protéiques (*e.g.* TIM barrel [Copley et Bork, 2000]) ou de familles d'enzymes [Nahum et Riley, 2001] pourraient catalyser des réactions similaires dans différentes voies. Une telle versatilité a été mise en évidence pour de nombreuses enzymes du métabolisme des petites molécules d'*Escherichia coli* [Teichmann *et al.*, 2001].

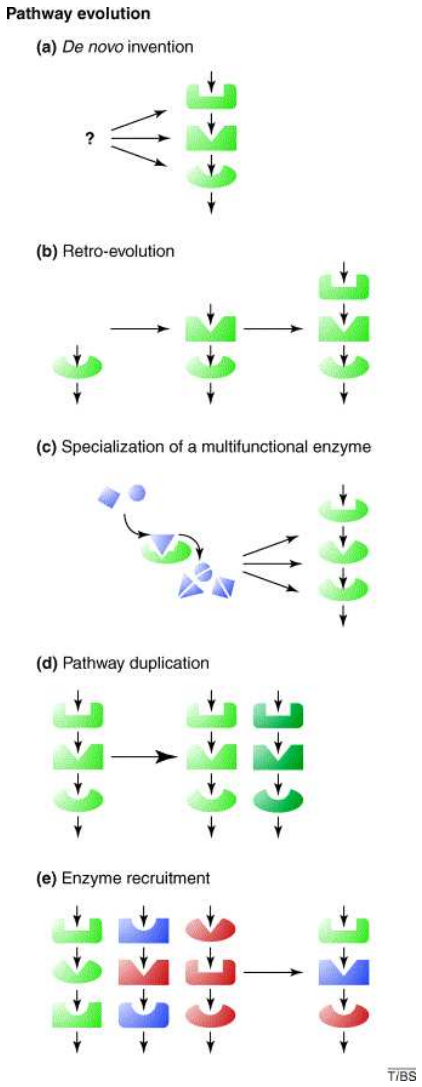


FIG. 2.8 – Modèles d'évolution des voies métaboliques. La figure est issue de [Schmidt *et al.*, 2003].

Modèles informatiques et statistiques

Outre les modèles biologiques qui ont été proposés dans la littérature, on trouve également des modèles statistiques. À la différence des modèles

biologiques, les modèles statistiques n'ont pas nécessairement de fondements mécanistiques. Ils sont jugés bons quand ils reproduisent les données observées.

Pour tenter de donner une base solide à leurs observations concernant la distribution des degrés des noeuds dans les réseaux réels qu'ils avaient observée, [Barabasi et Albert, 1999] ont proposé un modèle d'évolution de réseaux basé sur la notion d'attachement préférentiel.

Les deux règles d'évolution d'un réseau qu'ils considèrent sont :

1. Croissance : les réseaux croissent continuellement par addition de noeuds.
2. Attachement préférentiel : les nouveaux noeuds sont liés préférentiellement aux noeuds qui sont déjà très connectés.

Le résultat principal qu'ils obtiennent est que ce type de modèle reproduit effectivement une distribution du degré des noeuds qui suit une loi puissance. Ce résultat, qui reposait initialement sur des approximations, a par la suite été confirmé et généralisé par [Dorogovtsev *et al.*, 2000].

Deux réserves peuvent être émises. La première est que ce n'est pas le seul modèle d'évolution de réseaux qui permette de reproduire une telle distribution des degrés des noeuds [Keller, 2005]. La seconde est que la loi qu'ils proposent pour modéliser la distribution des degrés des noeuds observée n'est à nouveau pas la seule possible.

Ce modèle de graphe se place dans un cadre plus large de modélisation statistique de graphes sur lequel nous reviendrons dans la partie 5.3.2.

Dans un travail récent [Handorf *et al.*, 2005], Reinhart Heinrich, l'un des fondateurs de la théorie du contrôle métabolique, a proposé la notion de portée d'un composé (en anglais "scope of a compound"). L'idée est de savoir combien de composés on peut générer à partir d'un composé initial, dès lors qu'on dispose d'un certain nombre de réactions. C'est une situation tout à fait intéressante qui peut être vue comme une question évolutive puisque ce travail permet de caractériser quels composés sont primordiaux, et permet à terme d'inférer un ordre d'apparition des composés au cours de l'évolution. On peut noter que les auteurs utilisent un hypergraphe pondéré (donc équivalent à la matrice stochiométrique) pour faire leurs calculs.

Enfin, [Pál *et al.*, 2006] ont proposé un scénario d'adaptation du réseau à un nouvel environnement. En partant du réseau d'*Escherichia coli*, ils suppriment étape par étape les enzymes qui ne modifient pas substantiellement la production de biomasse. Afin de décider si une enzyme modifie la production de biomasse, les auteurs utilisent la technique de "Flux Balance Analysis" présentée brièvement à la Section 2.3.2.

Par ce processus, on obtient un réseau minimal. Ce réseau minimal est pour eux comparable au réseau métabolique de *Buchnera aphidicola*, bactérie de la même famille qu'*Escherichia coli* mais qui vit en symbiose avec son hôte et qui a un génome très réduit. Ils concluent alors qu'il est possible de prédire le contenu en gènes d'un organisme si on connaît son habitat.

Il existe plusieurs critiques qu'on peut faire à ce travail[5]. Tout d'abord, le réseau minimal obtenu n'est pas unique. D'un point de vue méthodologique, les auteurs utilisent (sans le dire explicitement) un algorithme glouton. En effet, chaque étape qui consiste à retirer une enzyme est un choix irréversible sur lequel ils ne reviennent pas. Ils obtiennent donc un réseau minimal (il y en a bien d'autres) mais pas le réseau minimum. Il serait sans doute plus pertinent d'explorer l'espace de tous les réseaux minimaux. D'autre part, la minimalité est ici une minimalité respectivement à une fonction objective, la production de biomasse. La production de biomasse est modélisée dans ce type d'approche par une réaction mettant en jeu un ensemble de substrats essentiels dans des proportions données. Ces proportions ont été établies expérimentalement pour *E. coli*. Dans l'approche, la fonction de biomasse est supposée constante au cours du temps. Cette limitation n'est cependant pas réaliste puisque si on cherche à modéliser la réduction d'un génome dû à un changement d'environnement (c'est le cas de l'évolution de *Buchnera*) alors il faut bien sûr prendre en compte ces changements d'environnements, et donc de concentrations de métabolites présents dans le milieu, qui ont un impact sur la fonction de biomasse.

2.4.5.3 Lien entre métabolisme et génome

Pour comprendre l'évolution du métabolisme, il est sans doute fondamental de revenir aux mécanismes qui génèrent cette nouveauté. Comme le rappellent [Pál *et al.*, 2005], l'un de ces mécanismes est le transfert horizontal de gènes qui est sans doute la principale source d'évolution dans les génomes procaryotes, et l'autre est la duplication de gènes, prépondérante chez les eucaryotes.

Plusieurs études se sont intéressées au lien entre génome et métabolisme. Dans [Rison *et al.*, 2002], les auteurs étudient les corrélations qui peuvent exister entre distance génomique et distance métabolique chez *Escherichia coli*. La distance génomique entre deux gènes est définie comme le nombre de gènes séparant deux gènes sur le génome. La distance métabolique est définie comme le nombre d'étapes séparant deux enzymes dans une voie métabolique. Les auteurs ajoutent que cette corrélation est principalement valable à faible distance et qu'elle s'explique quasiment intégralement par les structures connues d'opérons[6].

D'autre part, [Boyer *et al.*, 2005] introduisent la notion de métabolon. Un métabolon est un ensemble de gènes qui sont proches dans le réseau et proches sur le génome. Ils proposent une méthode générique permettant d'extraire automatiquement ces structures à partir d'un génome et d'un réseau. Cette notion de métabolon généralise l'approche de [Rison *et al.*, 2002] dans le sens

[5]La critique du travail de [Pál *et al.*, 2006] est le fruit d'une discussion avec Angela Douglas, spécialiste de *Buchnera* et professeur à l'université de York.

[6]Un opéron est un groupe de gènes colocalisés sur le génome qui sont transcrits ensemble et produisent un unique ARN messager.

que ce ne sont plus des paires d'enzymes qui sont étudiées mais des ensembles de taille quelconque.

Plusieurs auteurs ont par le passé utilisé le contexte génomique pour définir des modules dans les réseaux métaboliques. On peut mentionner à ce propos les travaux de [von Mering *et al.*, 2003] qui, dans le but d'annoter des génomes, utilisent la proximité des gènes sur le génome et dans le réseau métabolique.

D'autre part, [Yamada *et al.*, 2006] ont proposé la notion de module phylogénétique. Un module phylogénétique est un ensemble de gènes connectés dans le réseau métabolique et qui partagent le même profil phylogénétique (c'est-à-dire que ces gènes sont présents chez les mêmes espèces).

Enfin, [Spirin *et al.*, 2006] ont également travaillé sur la modélisation conjointe du métabolisme et du contexte génomique. Les auteurs explorent la notion de module dans ce cadre et suggèrent la notion de module évolutif mais cette fois en utilisant des relations évolutives entre les gènes d'un même organisme.

Pour conclure, on peut dire qu'il existe de nombreuses directions possibles de travail dès lors qu'on cherche à combiner les informations de localisation génomique et de position dans le réseau métabolique.

Le travail présenté dans cette thèse s'inscrit précisément dans le cadre de l'évolution du métabolisme par une analyse structurelle. On verra également les liens qui peuvent exister entre ces structures locales du métabolisme et des structures potentielles au niveau génomique.

Chapitre 3

Une nouvelle définition de motif dans le cadre des réseaux métaboliques

3.1 La notion de motif

Dans le dictionnaire, un motif est défini comme un thème, ou structure ornementale qui le plus souvent se répète.

En bioinformatique, le terme de motif désigne généralement un mot d'une séquence nucléique ou protéique qui est répété (parfois avec des erreurs) dans plusieurs séquences. La répétition de ce mot est généralement associée à une signification biologique ; on dit que le mot est conservé.

Une des applications principales de la recherche de motifs est la détection de sites de fixation de facteurs de transcription[1] dans les régions situées en amont des gènes (régions promotrices). Cette application s'intègre à la question plus large de la compréhension des mécanismes de régulation de la transcription des gènes.

D'un point de vue méthodologique, les premiers travaux concernant la recherche de motifs dans les séquences remontent à [Waterman *et al.*, 1984] qui n'utilisait pas le terme motif (qui s'est imposé dans les années 90) mais celui de "consensus pattern". Depuis, de nombreux travaux ont été menés sur ce sujet. La recherche et l'inférence de motifs dans les séquences nucléiques et protéiques est encore aujourd'hui un domaine très actif.

Plus récemment, la notion de motif a été proposée dans le cadre des réseaux biologiques [Milo *et al.*, 2002, Alon, 2006]. Dans ce travail, un motif est défini comme un "pattern" de connections qui apparaît de manière exceptionnelle dans un réseau (*i.e.* qui apparaît significativement plus qu'attendu dans un modèle de graphe aléatoire qui maintient la séquence de degrés des noeuds).

[1] Un facteur de transcription est une protéine nécessaire à l'initiation et à la régulation de la transcription de l'ADN en ARN.

Le travail a notamment été appliqué au réseau de régulation de la transcription d'*Escherichia coli* et les motifs extraits ont été interprétés comme des briques élémentaires du réseau ayant chacun une fonction spécifique dans le contrôle dynamique de l'expression des gènes (génération d'un programme temporel d'expression, gestion de la réponse à un signal externe fluctuant) [Shen-Orr *et al.*, 2002].

Dans cette thèse, nous proposons une définition de motif différente, qui puisse notamment être adaptée à des problématiques d'évolution du métabolisme.

Avant d'entrer dans le détail de ce travail, il est nécessaire de faire un point sur la terminologie.

Premièrement, dans la définition de [Milo *et al.*, 2002], la notion de sur-représentation est intégrée à la définition de motif. Il nous a semblé préférable dans notre approche de séparer la notion de pattern répété de la notion de sur-représentation. En effet, la mesure de sur-représentation dépend du modèle de graphe aléatoire que l'on se fixe. Il existe plusieurs modèles possibles (la question du choix du modèle est d'ailleurs une question largement ouverte que nous aborderons à la section 5.3.2), et le choix du modèle peut se faire indépendamment de la définition de motif. Dans le reste du manuscrit, nous ferons la distinction entre les termes de motif et motif sur-représenté.

Deuxièmement, nous allons par la suite traiter du problème de recherche de motifs et nous serons intéressés par les occurrences exactes de ce motif mais également par des occurrences approchées (*i.e.* qui contiennent des erreurs). Dans le cas d'occurrences approchées, le terme de modèle a parfois été préféré au terme de motif pour illustrer le fait qu'un motif est en fait extérieur aux objets d'étude [Sagot, 1998]. Dans ce manuscrit, nous nous en tiendrons au terme de motif pour désigner l'abstraction et on réservera le terme d'occurrence aux apparitions du motif dans l'objet d'étude.

3.2 Un exemple initial de motif dans le cadre des réseaux métaboliques - Cas de la synthèse de la lysine

Lorsqu'on considère une représentation du réseau métabolique complet d'un organisme (Fig. 3.1), la première impression qui se dégage est que tous les processus sont très entremêlés et qu'il semble difficile de dégager une structure claire. Malgré cette apparente complexité, il semble parfois possible de dégager des structures locales dans ce réseau.

Par exemple, certains auteurs, en tentant d'expliquer l'origine évolutive de la synthèse de la lysine chez les champignons, ont pu mettre en évidence des liens entre voies métaboliques [Velasco *et al.*, 2002, Irvin et Bhattacharjee, 1998, Miyazaki *et al.*, 2001]. Ainsi, si on regroupe les travaux de ces auteurs, on peut mettre en correspondance quatre étapes du cycle de Krebbs, de la synthèse de la leucine et de la synthèse de la lysine (Fig. 3.2).

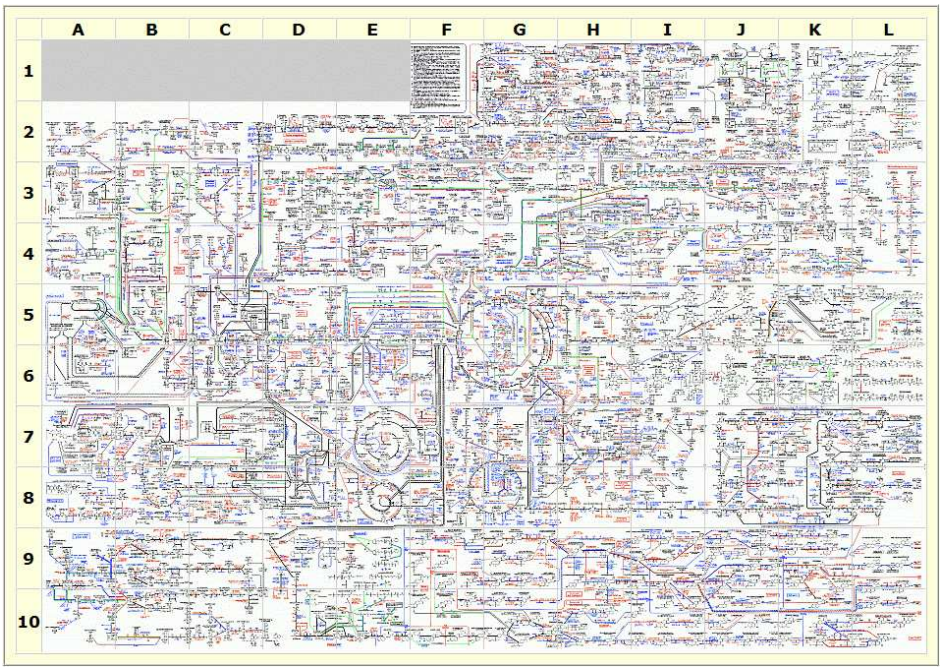


FIG. 3.1 – Carte de Boehringher du métabolisme. Ce poster représentant la quasi-totalité du métabolisme avait été sponsorisé par la compagnie Boeringher en 1992 et résulte de la compilation du travail de nombreux scientifiques coordonnée par Gerhard Michal.

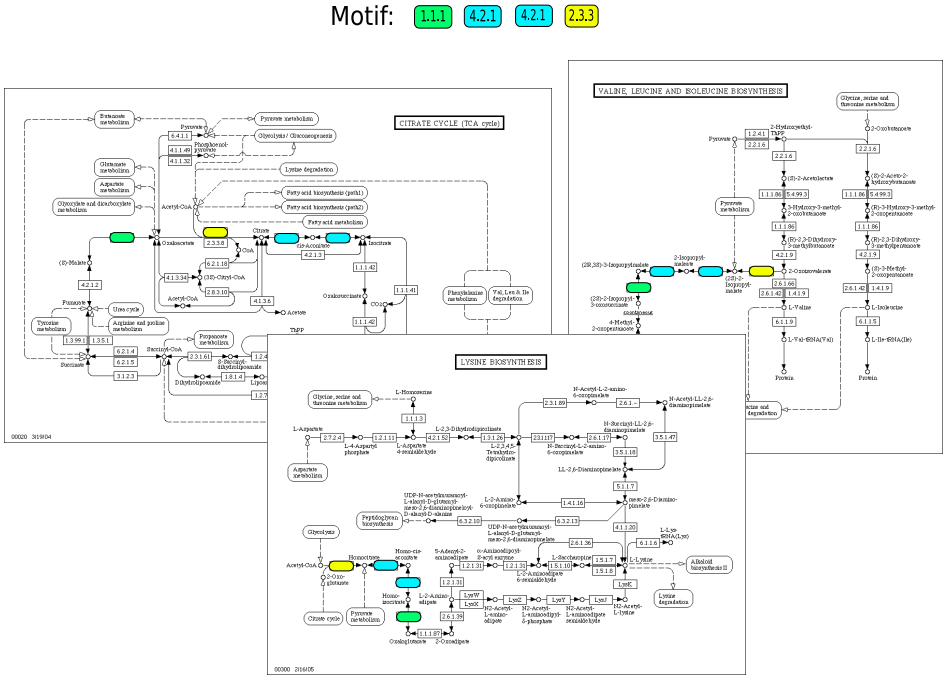


FIG. 3.2 – Correspondance de quatre étapes du cycle de Krebbs, de la synthèse de la leucine et de la synthèse de la lysine.

De cet exemple initial, nous avons dégagé plusieurs observations :
- seulement un fragment de la voie métabolique est impliqué dans la relation de similarité ;
- les numéros EC des réactions sont très similaires ;
- la structure (topologie) seule n'encode pas toute l'information (il existe beaucoup d'autres voies linéaires qui n'ont aucun lien avec ce motif) ;
- une similarité structurelle peut être liée à une similarité évolutive.

De ces observations, nous nous sommes demandés si nous pouvions rechercher de telles structures dans le métabolisme, de manière automatisée et sans *a priori* (en effet, le travail de [Velasco *et al.*, 2002] repose sur des alignements de séquences et de structures des gènes présélectionnés comme candidats pour expliquer l'histoire évolutive de la synthèse de la lysine).

Notre idée a donc été de développer un algorithme qui permette de proposer des candidats de manière automatique pour expliquer l'origine évolutive d'une voie métabolique d'intérêt. Dans ce contexte, un motif correspond à ce qu'il y a de commun entre ces candidats.

3.3 Définitions - Modélisation

3.3.1 Une nouvelle définition de motif

Quand nous avons commencé notre travail, la définition de motif proposée dans la communauté était celle de l'équipe de Uri Alon. Un motif était alors défini par sa topologie uniquement. Le constat de départ qui a motivé notre définition est que plusieurs auteurs s'accordent à dire que deux sous-parties d'un réseau qui ont une même topologie n'ont pas nécessairement la même fonction [Alm et Arkin, 2003, Guet *et al.*, 2002].

Dans le cadre des réseaux de régulation, une publication a récemment montré que le lien topologie-fonction n'était pas immédiat, notamment en ce qui concerne la dynamique du réseau ([Ingram *et al.*, 2006]). Les auteurs développent un modèle d'équations différentielles correspondant à un motif topologique et montrent que plusieurs comportements dynamiques peuvent être obtenus selon les paramètres utilisés.

En outre, notre but n'est pas de nous intéresser à la dynamique du réseau mais à sa structure et à son évolution. Une définition purement topologique semble moins justifiée dans le cas des réseaux métaboliques. En effet, deux voies métaboliques peuvent avoir la même topologie et correspondre à des processus complètement différents, aussi bien d'un point de vue biochimique que d'un point de vue évolutif.

Ainsi, par rapport à cette définition initiale de motif topologique, proposée par l'équipe de Uri Alon, nous allons introduire une autre définition qui diverge sur deux points principaux :
- tous les noeuds du graphe ne sont pas équivalents, on peut les différencier par classes fonctionnelles ;

– la topologie précise du motif n'est pas considérée, elle constitue une information secondaire, c'est la connection entre les noeuds qui compte.

Une nouveauté que nous apportons est précisément de ne pas compter sur la seule topologie pour résumer la fonction du sous-réseau. Dans notre cas, on considère que les noeuds ne sont pas tous équivalents mais qu'on leur attribue au préalable une fonction, qui sera symbolisée par une couleur.

Dans notre modélisation des réseaux métaboliques, les noeuds correspondent aux réactions et les couleurs correspondent à des classes de mécanismes réactionnels (oxydo-réduction, hydrolyse, ligation...). Nous reviendrons plus précisément sur la caractérisation de ces couleurs ultérieurement. On retiendra pour l'instant simplement que toutes les réactions ne sont pas équivalentes et qu'elles sont classées par groupe.

Pour aller plus loin et nous permettre d'explorer le poids que la topologie peut avoir dans la fonction d'un sous-réseau, nous avons en fait proposé d'ignorer la topologie exacte et de faire l'hypothèse que la fonction du sous-réseau était portée essentiellement par les fonctions des noeuds. Dans notre définition, la topologie est donc prise en compte de manière floue. La seule contrainte est que les noeuds soient connectés. La manière dont ils sont connectés ainsi que l'ordre dans lequel ils sont connectés n'a pas d'importance. Dans une optique de clarifier la terminologie, et pour différencier notre définition de la définition initialement proposée par l'équipe d'Uri Alon (motifs topologiques), nous parlerons de *motifs colorés*.

On note que dans un premier temps, nous avions introduit la notion de motif réactionnel [Lacroix *et al.*, 2005, Lacroix *et al.*, 2006] pour illustrer l'utilité de cette définition dans le cadre du métabolisme, mais il s'avère que ce terme reflète mal la généralité de la définition. On lui préfèrera donc le terme de motif coloré.

Nous allons à présent définir plus formellement la notion de motif coloré. Nous allons pour cela introduire la notion de couleur, de motif coloré et d'occurrence de motif coloré.

On note que les motifs colorés sont nécessairement définis dans le cadre de graphes colorés (*i.e.* les noeuds ont des étiquettes). Ainsi, pour pouvoir définir un motif, nous devons tout d'abord introduire l'ensemble des couleurs. On définit C un ensemble fini de couleurs. Un graphe coloré G est un graphe où les noeuds sont étiquetés par un ou plusieurs éléments de C.

Un **motif coloré** M est un multiensemble de couleurs prises dans C.

On peut noter à nouveau que cette définition ne contient aucune contrainte, ni sur l'ordre dans lequel ces couleurs apparaissent, ni sur la topologie du sous-graphe. Par ailleurs, on rappelle qu'un motif est extérieur au graphe. L'apparition d'un motif dans le graphe s'appelle une occurrence.

Nous allons à présent définir ce qu'est une occurrence d'un motif coloré.

3.3.2 Définition des occurrences

Intuitivement, une occurrence est un ensemble de noeuds connectés et colorés par les couleurs du motif.

Dans un premier temps, nous allons définir ce qu'est une occurrence exacte, puis nous nous intéresserons à la définition d'occurrences approchées (introduction de "mismatch" (ou mésappariement) et de "gap" (ou lacune)). La motivation pour introduire la notion d'occurrence approchée est que si on s'en tient aux occurrences exactes, on risque de ne trouver qu'une seule occurrence par motif. Or, nous sommes notamment intéressés ici pour trouver des fragments de réseaux qui ont une histoire évolutive commune mais qui ont pu diverger et donc n'être plus identiques aujourd'hui.

Pour rester dans le cadre général où les noeuds du graphe peuvent avoir plusieurs couleurs, il nous faut introduire le graphe biparti suivant :

Soit R un sous-ensemble des noeuds de G et soit M un motif de même taille que R. Soit $H(R, M)$ le graphe biparti dont l'ensemble des noeuds est $R \cup M$ et qui contient une arête entre un noeud v de R et un noeud c de M si et seulement si v possède c parmi ses couleurs.

Une occurrence exacte d'un motif M dans un graphe G est un ensemble R de noeuds de G tel que :

1. $H(R, M)$ admet un couplage parfait ;
2. R induit un sous-graphe connexe de G.

On peut d'ores et déjà noter que cette définition implique que motif et occurrence ont la même taille. Si un motif est constitué de k couleurs (pas nécessairement distinctes), alors l'occurrence contiendra k noeuds.

La Figure 3.3 illustre un cas où R ne constitue pas une occurrence de M car, dans ce cas, le graphe $H(R, M)$ n'admet pas de couplage parfait.

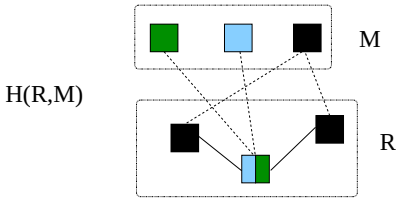


FIG. 3.3 – Les traits pleins indiquent les arêtes de G (le graphe d'où provient R). Les traits en pointillé correspondent aux arêtes de H.

On introduit maintenant le concept d'occurrence approchée. Pour cela, il nous faut définir une distance entre couleurs. Soit S une fonction qui assigne à chaque couple c_i, c_j de $C \times C$ un score qui exprime la similarité entre c_i et c_j. Deux couleurs vont être considérées comme *similaires* si leur score $S(c_i, c_j)$ est supérieur à un seuil s. On définit R_s la relation de similarité au seuil s.

On verra que, dans le cas particulier où R_s est une relation d'équivalence, on peut redéfinir une couleur par classe d'équivalence et appliquer la notion

d'occurrence exacte.

Dans le cas général, il faut adapter notre définition d'occurrence exacte en modifiant $H(R, M)$ de la façon suivante. À présent, il existe une arête entre un noeud v de R et une couleur c de M si et seulement si il existe une couleur c' de v telle que $S(c', c) \geq s$ (ou, dit autrement, $cR_s c'$).

De plus, on peut généraliser au cas où s est différent pour chaque élément c de M. Ceci est motivé par l'idée que certaines couleurs du motif peuvent être connues de manière plus précise que les autres.

Enfin, un autre type de flexibilité peut être ajouté pour autoriser des gaps dans les occurrences. En pratique, cela revient à autoriser l'occurrence à avoir plus de noeuds pour remplir la condition de connectivité. Ces noeuds supplémentaires ne sont pas appariés aux éléments du motif. Nous considérons par la suite deux types de bornes sur le nombre de gaps : une borne locale et une borne globale. Intuitivement, la borne locale permet de contrôler le nombre maximum de gaps consécutifs autorisés entre deux noeuds appariés de R. La borne globale contrôle le nombre total de gaps dans l'occurrence.

Cela nous mène à la définition suivante d'occurrence approchée d'un motif, où on dénote par G_R le sous-graphe de G induit par R, un ensemble de noeuds de G.

Soit l_b et g_b les bornes locale et globale et soit M un motif. Pour chaque élément c de M, soit s_c un nombre. Une occurrence de M est un ensemble R de noeuds qui est contenu dans un ensemble R' des noeuds de G satisfaisant les conditions suivantes :

1. le graphe biparti $H(M \cup R, E_H)$ avec $E_H = \{\{c, v\} \in M \times R \mid$ il existe une couleur c' de v telle que $S(c', c) \geq s_c\}$ admet un couplage parfait ;
2. pour tout sous-ensemble B de R tel que $B \neq \emptyset$ et $R \setminus B \neq \emptyset$, la longueur du plus court chemin dans $G_{R'}$ entre un élément de B et un élément de $R \setminus B$ est au plus lb ;
3. $|R'| - |R| \leq gb$.

On peut noter que, lorsqu'aucun gap n'est autorisé, alors $R = R'$ et la condition 2 signifie simplement que G_R est connecté. Un exemple est donné dans la Figure 3.4.

On peut remarquer que, contrairement à un alignement de séquences, où un score global est calculé prenant en compte mismatch et gaps, on sépare ici les deux pénalités qu'on soumet à un test de seuil. Une extension pourrait être de contrôler globalement les mismatches, en calculant un score global de mismatch (selon un modèle additif par exemple).

Enfin, on peut noter qu'on ne considère pas les délétions (c'est-à-dire la situation où le motif contient plus d'éléments que l'occurrence ne contient de noeuds). Prendre en compte les délétions et utiliser un score global de mismatch constituent deux extensions possibles.

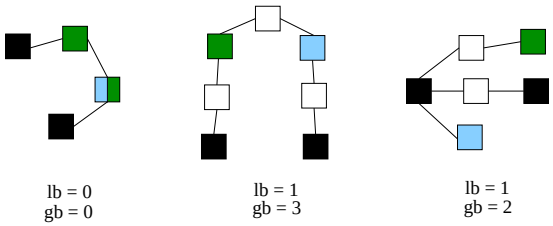


FIG. 3.4 – Sous-graphes induits par les occurrences du motif {noir, noir, gris foncé, gris clair}.

3.3.3 Définition des couleurs et de la fonction de score

Nous allons ici discuter le choix d'un ensemble C de couleurs et d'une fonction S de comparaison entre couleurs qui soient pertinents dans le cadre du métabolisme. On peut remarquer que, par ailleurs, la notion de motif coloré n'est pas spécifique au métabolisme et peut très bien être appliquée dans un contexte tout à fait différent.

Une difficulté pour le choix de C va venir du fait qu'aucune classification systématique de réactions n'est disponible à l'heure actuelle. Par contre, il existe de nombreuses classifications des enzymes. Une solution consiste donc à choisir d'utiliser les enzymes pour étiqueter et comparer les réactions. Un problème auquel nous sommes alors confrontés est que plusieurs enzymes peuvent catalyser la même réaction et qu'une enzyme peut catalyser plusieurs réactions. Cette observation illustre la nécessité de se placer dans le cas où chaque noeud du graphe peut avoir plusieurs couleurs.

Nous allons ici discuter deux façons possibles de définir C et S. La première est basée sur l'alignement. En effet, pour comparer des réactions, on peut comparer les enzymes qui catalysent ces réactions en effectuant un alignement de leurs séquences. Un élément de C est alors une séquence protéique. La fonction S assigne un score d'alignement et s est un seuil défini par l'utilisateur qui doit être atteint pour que les séquences soient considérées comme similaires. La même méthode peut être employée en utilisant des structures de protéines plutôt que des séquences mais dans le cas d'étude d'un réseau métabolique complet, il est sans doute préférable d'utiliser les séquences puisque de nombreuses structures ne sont pas encore déterminées.

Le deuxième exemple est celui que nous avons principalement utilisé par la suite. Il est basé sur une classification hiérarchique des enzymes développée par l'IUBMB (International Union of Biochemistry and Molecular Biology) [Nomenclature, 1992]. Ainsi que nous l'avons vu à la section 2.4.5.1, cette classification consiste à assigner à chaque activité enzymatique nouvellement découverte un code à quatre numéros exprimant la chimie de la réaction catalysée. On rappelle que ce code est connu sous le terme de numéro EC ou "EC number" (pour Enzyme Commission Number) et est largement utilisé dans les bases de données. Le premier numéro d'un numéro EC peut avoir

une valeur comprise entre 1 et 6, chaque chiffre correspondant à une des 6 grandes classes d'activité enzymatique (1. Oxidoreductase, 2. Transferase, 3. Hydrolase, 4. Lyase, 5. Isomerase, 6. Ligase). Puis chacun des 3 chiffres restant du code indique un niveau de détail supplémentaire.

Par exemple, l'enzyme tripeptide aminopeptidase a le code EC 3.4.11.4 qui est construit comme suit : 3 signifie une hydrolase (enzymes qui utilisent l'eau pour détruire une autre molécule), 3.4 signifie hydrolases agissant sur des liens peptidiques, 3.4.11 implique celles qui détachent un acide aminé amino-terminal d'un polypeptide et 3.4.11.4 implique celles qui détachent cet acide aminé amino-terminal d'un tripeptide.

Un élément de C est dans ce cas un numéro EC. La fonction S assigne un score de similarité entre deux numéros EC qui correspond à la profondeur maximale de la classification pour laquelle les numéros coincident. Par exemple $S(1.1.1.2, 1.1.1.3) = 3$. Deux numéros EC sont considérés comme similaires si leur score de similarité est supérieur à un seuil s compris entre 0 et 4.

On peut noter que la classification EC est une classification hiérarchique, *i.e.* représentable par un arbre. Dès lors, la relation de similarité entre numéros EC qu'on a défini devient une relation d'équivalence. En effet, $\forall(c_1, c_2, c_3) \in C^3, c_1 R_s c_2 \wedge c_2 R_S c_3 \Rightarrow c_3 R_S c_1$. Une implication directe de cette observation est que, quand on connaît le seuil s de comparaison des couleurs, on peut recolorer le graphe en utilisant une couleur par classe d'équivalence. On peut noter que ces couleurs correspondent aux noeuds internes de la classification EC et sont en réalité des numéros EC partiels (par exemple, 1.1).

Ainsi, lorsqu'on utilise la classification EC, les noeuds internes peuvent aussi faire partie des couleurs autorisées. Ce choix a deux motivations : 1. en pratique certains numéros EC ne sont pas complètement caractérisés dans les données (le quatrième numéro n'est pas toujours spécifié), 2. le motif peut être constitué de mécanismes réactionnels généraux et c'est cela qui peut être intéressant.

On peut remarquer que, en comparaison de la mesure basée sur les alignements, un avantage de celle basée sur les numéros EC est qu'elle est plus proche de la notion de fonction. Les réactions comparées avec cette mesure seront donc plus susceptibles d'être proches fonctionnellement (et potentiellement liées par l'évolution aussi).

Cependant, il est important de noter que la notion de numéro EC est souvent mal utilisée dans la littérature car on la confond avec la notion d'enzyme ou la notion de réaction. Or un numéro EC n'identifie ni une réaction, ni une enzyme. Ainsi, un numéro EC ne détermine pas de façon unique une enzyme (en effet, plusieurs enzymes peuvent avoir un même numéro, le cas des isozymes est un exemple classique). Mais, de manière plus surprenante, un numéro EC ne détermine pas de manière unique une réaction (on rappelle qu'une réaction est déterminée de manière unique par l'ensemble de ses substrats et produits). Ainsi, même si les 4 numéros sont spécifiés, plusieurs réactions peuvent avoir le même numéro EC. En quelque sorte, il

faudrait un cinquième niveau pour discriminer ces ex-aequos. Dans ces cas, la chimie de la réaction est quasiment la même. Nous avons détecté une vingtaine d'exemples dans la base EcoCyc où un numéro EC correspond à deux réactions. Un exemple est le numéro EC 2.2.1.1 qui correspond à deux réactions de transketolation (transfert d'un groupe ketole (HOCH2CO-) d'un composé à un autre). Les deux réactions diffèrent par leur groupe accepteur qui est dans un cas l'erithrose, et dans l'autre le ribose.

La question du choix des couleurs est toujours ouverte dans la communauté bioinformatique. Les numéros EC ont été très utilisés [Pinter *et al.*, 2005, Tohsato *et al.*, 2000, Clemente *et al.*, 2005], notamment car ils sont faciles à manipuler. Un autre "avantage" est qu'ils maintiennent l'ambiguïté entre réactions et enzymes, et donc font le pont entre une approche génomique (basée sur les enzymes) et une approche chimique (basée sur les réactions).

Cependant, pour pouvoir aller plus loin, on pourrait vouloir séparer les approches chimique et génomique et selon la définition de couleur qu'on choisit, rechercher des motifs d'enzymes ou des motifs de réactions. Ce choix dépendrait de l'application considérée. On reviendra sur cette discussion dans la partie Perspectives.

Chapitre 4

La recherche de motifs colorés dans un graphe

Ce chapitre présente essentiellement des résultats que nous avons obtenu au cours de la thèse concernant le problème de la recherche de motifs colorés dans un graphe. D'autres auteurs se sont depuis intéressés à ce sujet et ont obtenu de nouveaux résultats [Hermelin *et al.*, 2007]. Nous les mentionnerons ici également.

4.1 Le problème de la recherche de motifs colorés

Le problème de la recherche de motifs colorés peut être formulé de la manière suivante :

RECHERCHE DE MOTIFS COLORÉS : Étant donné un graphe non-orienté coloré et un motif coloré M (multiensemble de couleurs), trouver toutes les occurrences de M dans G.

Il est important de rappeler que ce problème est différent de l'isomorphisme de sous-graphe puisque la topologie du motif n'est pas spécifiée. Dans la section suivante, nous considérerons, sans perte de généralité, que le graphe est connecté et que tous les sommets ont des couleurs qui apparaissent dans le motif. Si ce n'est pas le cas, alors on supprime les sommets qui n'ont pas de couleur qui apparaisse dans le motif et on résout le problème sur chacune des composantes connexes qui restent.

Cette simplification est valable quand aucun gap n'est autorisé. Des compléments sur la façon de gérer les gaps seront donnés dans la section 4.1.4 mais l'analyse de la complexité sera faite sur le cas exact.

Une variante naturelle du problème de recherche de motif consiste à, étant donné un motif et un graphe coloré, décider si le motif apparaît ou non dans le graphe. Il est facile de voir que la version décisionnelle du problème

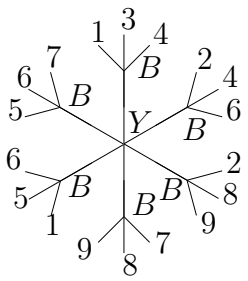


FIG. 4.1 – Arbre T et ses étiquettes pour $X = \{1, \ldots, 9\}$ et $\mathcal{C} = \{\{1,3,4\}, \{2,4,6\}, \{2,8,9\}, \{7,8,9\}, \{1,5,6\}, \{5,6,7\}\}$. Pour cet exemple, $M = \{Y, B, B, B, 1, \ldots, 9\}$.

de recherche est dans NP (si on donne une solution au problème, on peut la vérifier en temps polynomial). Nous allons montrer que le problème est NP-complet même dans le cas où G est un arbre, ce qui implique que le problème de RECHERCHE DE MOTIFS COLORÉS est NP-difficile pour les arbres.

4.1.1 L'étude de complexité

Nous avons la proposition suivante :

Proposition 1. *La version décisionnelle du problème de recherche de motifs colorés est NP-complet même si G est un arbre.*

Preuve. Nous proposons une réduction à partir de EXACT COVER BY 3-SETS (X3C).

INSTANCE : Ensemble X avec $|X| = 3q$ et une collection $\mathcal{C}$ de triplets de X.

QUESTION : Est-ce que $\mathcal{C}$ contient une couverture exacte de X, *i.e.*, une sous-collection $\mathcal{C}' \subseteq \mathcal{C}$ telle que chaque élément de X apparaît dans exactement un membre de $\mathcal{C}'$?

Soit $X = \{1, \ldots, 3q\}$ et $\mathcal{C} = \{C_1, \ldots, C_n\}$ une instance de X3C. L'instance de la version décisionnelle de problème de recherche consiste en un motif $M = \{Y, B, \ldots, B, 1, \ldots, 3q\}$, où B apparaît q fois dans M, et un arbre T défini comme suit (Voir Figure 4.1 pour un exemple). L'arbre T contient quatre sommets pour chaque i, $1 \leq i \leq n$, trois d'entre eux sont des feuilles de T, chacune étiquetée par un des éléments de C_i. Le quatrième sommet, nommé r_i, est adjacent aux trois feuilles et a pour couleur B. En dehors de ces $4n$ sommets, l'arbre contient seulement un autre sommet, qui est étiqueté par Y et est adjacent à chaque r_i. Ceci complète la description de l'instance. De façon évidente, elle est polynomiale dans la taille de X et $\mathcal{C}$.

Pour terminer la réduction, nous devons montrer que le motif M apparaît dans T si et seulement si il existe une sous-collection $\mathcal{C}'$ de $\mathcal{C}$ telle que chaque élément de X apparaît dans exactement un membre de $\mathcal{C}'$.

Supposons qu'il existe une telle sous-collection $\mathcal{C}'$. Clairement, $|\mathcal{C}'| = q$. Soit R un ensemble de sommets de T qui correspond au sommet étiqueté par Y et aux quatre sommets de chaque C de $\mathcal{C}'$. Le sous-graphe de T induit par R est connecté. De plus, dans R, il y a un sommet étiqueté par Y, q sommets étiquetés par B (un pour chaque C dans $\mathcal{C}'$) et un étiqueté par chaque élément de X (puisque c'est une propriété de $\mathcal{C}'$). Donc, R est une occurrence de M dans T.

À présent, supposons qu'il y a une occurrence de M dans T, c'est-à-dire qu'il existe un ensemble R de $1+4q$ sommets de T qui induit un sous-graphe connexe de T et a un sommet étiqueté par chacune des couleurs de M. Soit $\mathcal{C}'$ une sous-collection de $\mathcal{C}$, constituée d'ensembles C_i dont le sommet r_i de T est dans R. Nous allons prouver que chaque élément de X apparaît dans exactement un des ensembles de $\mathcal{C}'$. Tout d'abord, notons que le sommet étiqueté par Y est nécessairement dans R, puisque c'est le seul qui soit étiqueté par Y et qu'il y a un Y dans M. Alors, puisque R induit un sous-graphe connexe, une feuille d'un des ensembles C_i est dans R si et seulement si r_i est également dans R. Mais R doit contenir exactement q sommets étiquetés par B. En conséquence, $|\mathcal{C}'| = q$ et, puisque R doit contenir $1 + 4q$ sommets, les trois feuilles de chaque C de $\mathcal{C}'$ doivent être dans R, et cela constitue l'intégralité des sommets de R. Puisque R doit contenir un sommet étiqueté pour chaque élément de X, il doit y avoir exactement un ensemble de $\mathcal{C}'$ qui contient un élément de X. □

Cette preuve peut ensuite être généralisée au cas particulier de motifs où les couleurs ne sont pas répétées. La seule différence est dans la construction de l'instance. Dans ce cas, le motif est défini par $M = \{Y, B_1, B_2, \ldots, B_q, 1, \ldots, 3q\}$ et l'arbre est le même que précédemment à part que chacun des n sous-arbres de quatre sommets connectés au sommet central (couleur Y) apparaît maintenant en q copies. Les copies varient simplement dans la couleur de leur sommet racine qui a la couleur B_j, $1 \leq j \leq q$.

On peut enfin signaler qu'une autre preuve a par la suite été présentée dans [Hermelin *et al.*, 2007]. La réduction est cette fois faite à partir de 3-SAT. Le résultat obtenu est plus général puisque la preuve est faite pour un arbre de degré maximum égal à 3. On peut donc conclure que le problème est difficile même si G est un arbre de degré maximum 3.

Nous avons donc pu prouver que le problème de recherche de motif coloré était NP-difficile. Cela signifie qu'il n'existe *a priori* pas d'algorithme polynomial pour le résoudre.

Cependant, ceci ne nous renseigne pas sur l'origine du terme exponentiel. Est-ce que la difficulté provient essentiellement de la taille du motif ? Nous allons maintenant prouver que la réponse à cette question est oui, dans le cas particulier où le réseau est un arbre.

4.1.1.1 Complexité paramétrique

Trouver toutes les occurrences d'un motif M dans un arbre non orienté coloré T (RECHERCHE DE MOTIFS COLORÉS) peut être fait en temps polynomial quand on fixe la taille du motif (qu'on note k). En effet, un algorithme naïf consiste à générer toutes les topologies possibles pour le motif M, et à rechercher chaque topologie en utilisant un algorithme d'isomorphisme de sous-arbre. Puisqu'il est suffisant de générer toutes les topologies d'arbre pour M, le nombre de topologies à considérer dépend (exponentiellement) de k uniquement, et le problème de l'isomorphisme de sous-arbre est polynomial à la fois dans la taille du motif M et de l'arbre T dans lequel on recherche M. Ce raisonnement n'est plus valable lorsqu'on recherche le motif dans un graphe général G puisque le problème d'isomorphisme de sous-graphe est NP complet même si le motif est un arbre [Garey et Johnson, 1979].

Ce résultat a récemment été généralisé par [Hermelin *et al.*, 2007] au cas où le réseau est un graphe et pour des motifs de taille $\log n$, avec n le nombre de sommets de G. Ce résultat est prometteur dans le sens où il indique que le problème de recherche de motifs est traitable en temps polynomial pour des motifs de petite taille. L'algorithme proposé dans la section 4.1.3 ne tire pas parti de cette observation et n'est donc pas polynomial en n. Cependant, les instances de graphes considérées n'étant en pratique pas si grandes, il se trouve que le temps d'exécution n'est pas limitant.

Par ailleurs, les auteurs montrent que le problème est W[1]-dur en c si c est le nombre de couleurs du motif (en effet, pour une valeur de c fixée, k ne l'est pas forcément, du fait des répétitions de couleurs ; le problème peut donc bien être FPT en k et W[1]-dur en c).

4.1.2 Complexités relatives des motifs topologiques et des motifs colorés

Dans cette thèse, nous définissons un motif comme un multiensemble de couleurs (MOTIF COLORÉ). D'autres auteurs [Shen-Orr *et al.*, 2002] le définissent comme un sous-graphe (MOTIF TOPOLOGIQUE). Dans les deux cas, décider si un motif apparaît dans un graphe est NP-complet. Néanmoins, la combinaison des deux types de contraintes (topologie et couleurs) peut, dans certains cas, mener à des algorithmes polynomiaux. Dans cette section, nous allons comparer la complexité de différentes variantes de ces problèmes dans le but de localiser plus finement la frontière entre problèmes faciles et difficiles.

Dans la discussion qui suit, nous allons d'abord nous concentrer sur le cas général où le réseau est un graphe, puis nous commenterons le cas particulier où il s'agit d'un arbre.

Si nous considérons la topologie comme seule contrainte (MOTIFS TOPOLOGIQUES), toutes les variantes considérées du problème sont NP-complètes. En effet, l'isomorphisme de sous-graphe est NP-complet même si le motif est

un chemin. Ceci peut être prouvé par une réduction à partir du Chemin Hamiltonien. La réduction est basée sur une instance dans laquelle le motif est un chemin de taille n, avec n le nombre de sommets du graphe dans lequel le motif est recherché.

Si maintenant on considère à la fois la topologie et les couleurs comme contraintes (Motifs topologiques colorés), le problème est NP-complet dans tous les cas (réduction à partir du Chemin Hamiltonien en utilisant la même technique que précédemment), à part dans le cas où les couleurs sont fixées (chaque sommet du motif a une couleur qui lui est assignée) et qu'aucune répétition n'est autorisée. En effet, si les couleurs sont fixées, (le motif est ordonné), on peut transformer le graphe d'entrée en graphe orienté en ne conservant que les arêtes qui apparaissent dans le motif. Et puisqu'aucune répétition n'est autorisée, le graphe orienté qu'on obtient est en réalité un DAG. Il nous reste à montrer que rechercher un arbre orienté T coloré avec des couleurs distinctes dans un DAG coloré D peut se faire en temps polynomial. Ceci semble pouvoir être fait avec un algorithme de programmation dynamique. Ce travail est en cours.

Dans le cas où le réseau est un arbre, le problème est polynomial sous toutes ces variantes (l'isomorphisme de sous-graphe peut être résolu en temps polynomial) sauf dans le cas où la topologie n'est pas connue (Motifs colorés).

Ces résultats sont résumés dans le Tableau 4.1.2.

Motif \ Graphe en entrée		Arbre	Arbitraire
Motifs topologiques		polynomial	NP-complet
Motifs topologiques colorés	Cas Général	polynomial	NP-complet
	couleurs fixes, sans répétitions	polynomial	polynomial (conjecture)
Motifs colorés (cette thèse)		NP-complet	NP-complet

Tab. 4.1 – Résultats de complexité pour le problème de la recherche de motifs dans un graphe.

4.1.3 Un algorithme de comptage exact

Dans les sections précédentes, nous avons pu établir que le problème de la recherche de motifs colorés était un problème difficile. Face à un problème difficile, plusieurs solutions sont généralement envisagées :

- faire un algorithme exact ;
- recourir à des heuristiques ;
- reformuler le problème pour le rendre plus simple.

La principale limite de l'approche par algorithme exact est bien sûr le temps d'exécution dont on sait qu'il augmente exponentiellement avec la taille des entrées. Pour gagner en temps, on peut recourir à des heuristiques. Dans ce cas, on n'est plus garanti de trouver toutes les occurrences, on peut

en rater. Un problème majeur est d'ailleurs qu'on ne peut pas savoir combien on en rate.

Enfin, reformuler le problème consiste à se restreindre aux cas particuliers où le problème est facile. On se retrouve alors dans un cadre classique, mais on n'a pas réellement résolu le problème initial.

Il se trouve que, dans notre cas, les instances que nous considérons sont assez petites. En effet, le réseau métabolique d'*Escherichia coli* tel que nous l'avons reconstruit est constitué de 587 noeuds et 1667 arêtes. Il nous a donc semblé raisonnable d'adopter une approche exacte. Nous verrons par la suite que la question du temps d'exécution n'a en effet pas été une contrainte majeure, notamment parce que nous avons travaillé avec des motifs relativement petits (de taille 8 au maximum).

Nous allons présenter maintenant un algorithme exact qui permet de résoudre le problème de la recherche de motifs colorés. Nous expliquons tout d'abord comment cet algorithme fonctionne dans le cas simple où les paramètres de gaps lb et gb sont fixés à 0, et nous montrerons après comment il peut être étendu au cas général.

Soit M le motif recherché. Un algorithme très naïf consisterait à considérer chaque ensemble R de k sommets (où $k = |M|$) comme un candidat et de tester s'il remplit les conditions qui en font une occurrence. Nous rappelons que pour que R soit considéré comme une occurrence de M, le sous-graphe induit par R doit être connexe et le graphe biparti $H(R, M)$ (qui a une arête entre $r \in R$ et $c \in M$ si et seulement si c est similaire à une des couleurs du sommet r) doit admettre un couplage parfait. L'espace de toutes les combinaisons de k sommets parmi n étant très grand, nous proposons deux idées principales d'élagage qui proviennent des deux conditions que R doit remplir pour être validé comme occurrence de M.

L'idée générale de l'algorithme est de se centrer sur un sommet, de tester la condition de couleur sur tous les sous-graphes connexes qui contiennent ce sommet, puis d'éliminer ce sommet du graphe et de passer au sommet suivant.

La condition de connectivité peut être vérifiée avec une technique standard de parcours de graphe, comme un parcours en largeur (BFS). Dans notre cas, nous effectuons une recherche en largeur combinée avec du retour en arrière (en anglais "backtracking") en prenant successivement pour point de départ chaque sommet du graphe. Après chaque parcours, le noeud à partir duquel on est parti peut être supprimé du graphe sans risque. Lors d'un parcours, à chaque étape, un sous-ensemble des sommets qui se trouvent dans la file de la BFS sont marqués comme faisant partie de l'ensemble candidat R. La file, à chaque étape, contient seulement des sommets marqués et des voisins dans G des noeuds marqués. Un pointeur p indique le dernier sommet de la file à avoir été considéré.

À chaque étape, on peut se trouver dans deux situations : soit k noeuds sont marqués, soit il y en a moins. Si k noeuds sont marqués, on tient un ensemble candidat R. On soumet R au test des couleurs décrit plus loin, et

on backtracke pour trouver le prochain ensemble candidat. S'il y a moins de k noeuds marqués dans la file, alors deux cas sont à analyser : soit p pointe sur le dernier sommet de la file, soit il pointe sur un autre sommet. Si p ne pointe pas sur le dernier sommet de la file, on décale p d'une position, on marque le sommet nouvellement pointé et on ajoute ses voisins qui ne sont pas déjà dans la file (cette dernière opération peut être faite en temps constant en ajoutant un attribut à chaque sommet du graphe original). Puis on recommence une nouvelle étape. Si, par contre, p pointe sur le dernier sommet de la file, alors on redirige p vers le dernier sommet marqué (si un tel sommet n'existe pas alors la recherche est terminée) et on backtracke. Le backtracking consiste à 1. démarquer le sommet sur lequel p pointe, 2. retirer de la file les voisins qui ont été ajoutés lors de son marquage, et 3. recommencer une nouvelle étape. À présent, nous allons décrire le test correspondant à la condition de coloration.

Étant donné un ensemble candidat R, on peut vérifier s'il vérifie la condition de coloration en construisant un graphe H et en vérifiant s'il admet un couplage parfait. En fait, on peut déjà appliquer une variante de cette vérification à un ensemble candidat partiel. En effet, lors de la construction de l'ensemble candidat R, on peut vérifier si le graphe correspondant admet un couplage complet. Un couplage complet est un couplage qui couvre complètement l'ensemble candidat partiel. Si un tel couplage n'existe pas, alors on peut arrêter d'étendre ce candidat sans risque et continuer notre recherche à partir du prochain candidat partiel. On peut noter que cette vérification peut être faite en temps constant en utilisant une structure de données additionnelle de taille k.

Enfin, plusieurs optimisations peuvent être ajoutées à cette structure. Par exemple, au lieu d'utiliser chaque sommet comme graine pour la BFS, on peut se restreindre à un sous-ensemble : ceux qui sont colorés par une des couleurs du motif, de préférence la couleur la moins fréquente dans le graphe. Par ailleurs, en prétraitement, on peut supprimer les composantes connexes du graphe qui ne contiennent pas toutes les couleurs du motif.

4.1.4 La gestion des gaps

Contrôler le nombre de gaps locaux sans contrôler le nombre de gaps globaux ($lb > 0$ et $gb = \infty$) peut être fait facilement en calculant la fermeture transitive d'ordre lb du graphe initial G et en appliquant le même algorithme au graphe qu'on a rendu ainsi plus dense.

La fermeture transitive d'ordre q d'un graphe G est le graphe obtenu à partir de G en ajoutant une arête entre chaque paire de sommets u et v pour qui la longueur l du plus court chemin de u à v dans le graphe original satisfait $1 < l < q$. La fermeture transitive d'ordre p peut être calculée au début de l'algorithme.

Si on ne contrôle que le nombre de gaps globaux (c'est-à-dire $lb = \infty$ et $gb > 0$), le problème peut être reformulé comme un problème de recherche

d'arbre de Steiner minimum sur un ensemble de noeuds du graphe. L'ensemble de noeuds serait chaque ensemble R de sommets de G qui vérifient la condition de coloration. Si l'arbre de Steiner minimum pour R a une taille inférieure à $|R| + gb - 1$, alors R est une occurrence. Le problème de l'arbre de Steiner est NP-difficile dans le cas d'un graphe général, mais il existe plusieurs programmes disponibles qui le résolvent de manière exacte.

Ici encore, on peut adopter une stratégie similaire à la précédente pour éviter de tester tous les ensembles de k sommets parmi n. La situation est cependant plus délicate. En effet, lorsqu'un sommet v est incorporé à la file, c'est parce qu'il est situé à une distance d inférieure à gb d'un sommet marqué. Mais, au cours du déroulement de l'algorithme, sa contribution au gap global ne sera pas nécessairement de d puisque cette contribution dépend aussi des autres noeuds marqués de R qui peuvent être à une distance inférieure de v. On doit donc garder une trace de cette information lorsqu'on ajoute ou qu'on retire des sommets de la file (et de R). Cette information peut être stockée en utilisant un arbre balancé de taille proportionnelle à k associé à chaque sommet v de la file. Chaque sommet de l'arbre correspond à un sommet marqué u qui est responsable pour l'ajout de v dans la file et est étiqueté par la distance entre v et u (cette distance est au plus gb). Stocker, mettre à jour et utiliser ces informations supplémentaires ajoute un terme multiplicatif en $O(k \log k)$ à la complexité en temps de l'algorithme, ce qui semble raisonnable.

Avec ce système, on peut en fait facilement contrôler à la fois le gap local et le gap global (c'est-à-dire fixer $lb > 0$ et $gb > 0$). Dans ce cas en effet, on doit simplement vérifier que la distance entre v (le sommet non marqué qui est ajouté à la file) et u (le sommet marqué qui a mené à l'ajout de v) est au plus de $\min\{lb, gb\}$. Une autre façon de procéder consisterait à 1. trouver tous les motifs qui vérifient la condition de gap local et, 2. résoudre un problème d'arbre de Steiner minimum dans G avec les ensembles R donnés par les solutions de l'étape 1.

Enfin, on peut remarquer que dans le cas où on autorise des gaps, le filtrage initial, proposé dans le cas de l'algorithme exact, qui consiste à retirer *a priori* du graphe les noeuds qui ne sont pas colorés par les couleurs du motif ne s'applique pas directement. En fait, ce filtrage peut être appliqué après l'étape de fermeture transitive. En effet, dans le cas d'occurrence approchée, un noeud mal coloré n'appartiendra pas à une occurrence. Par contre, pour pouvoir satisfaire les contraintes de gap global, il faut garder l'information de distance entre noeuds dans le graphe d'origine. On a donc besoin d'une structure de données additionnelle de taille n^2.

De manière générale, concernant la gestion des gaps, il est important de noter que les solutions que nous proposons ici constituent des propositions initiales qui pourront être remplacées à terme par des solutions plus efficaces. En ce qui concerne les applications, nous aborderons succinctement le cas des gaps pour la recherche de motifs mais cette question sera laissée de côté lorsqu'on s'intéressera à l'inférence.

4.2 Application de la recherche de motifs à l'évolution de voies métaboliques

Nous allons maintenant donner un exemple d'application de recherche de motifs dans un réseau métabolique. La recherche de motifs sert ici à proposer des hypothèses sur l'évolution de certaines voies métaboliques.

La question posée est analogue à celle posée dans [Velasco *et al.*, 2002] où les auteurs cherchaient à comprendre l'origine évolutive de la synthèse de la lysine chez les champignons.

Nous nous sommes intéressés ici à la synthèse de la valine et nous avons procédé de la manière suivante. En se concentrant sur les cinq dernières étapes de la voie, nous avons dérivé un motif $M = \{1.1.1.86, 1.1.1.86, 4.2.1.9, 2.6.1.42, 6.1.1.9\}$ que nous avons recherché dans le réseau en utilisant initialement un seuil s de comparaison entre numéros EC égal à 4. Avec cette valeur de seuil, il se trouve que le motif n'apparaît qu'une seule fois.

À partir de ce motif défini de manière stricte, nous avons relâché des contraintes en fixant le seuil de comparaison s à 3, puis en fixant les paramètres de gap à 1 (motif M'). Avec ces valeurs, de nouvelles occurrences apparaissent (occurrences approchées). Trois d'entre elles ont particulièrement retenu notre attention [1] (voir Figure 4.2.)

La première correspond aux cinq dernières étapes de la synthèse de l'isoleucine. La seconde correspond aux cinq dernières étapes de la synthèse de la leucine. On obtient donc trois occurrences d'un même motif dans les voies de synthèse de ces trois acides aminés. On peut alors faire l'hypothèse que ces trois voies métaboliques ont une histoire évolutive commune.

Un point remarquable concernant la seconde occurrence est que l'ordre des réactions n'est pas le même que dans les autres voies. Cette occurrence n'aurait pas été trouvée si on avait utilisé une définition de motif plus stricte, où l'ordre des couleurs aurait été spécifiée.

Finalement, la troisième occurrence qui a attiré notre attention était formée de réactions qui faisaient partie de deux voies métaboliques distinctes : la synthèse de la valine et la synthèse du Panthotenate et du Coenzyme A. Cette dernière situation illustre une des limites de notre façon générale de penser le métabolisme : les voies métaboliques ne sont pas facilement séparables les unes aux autres mais bien entremêlées. Le fait de travailler sur le réseau métabolique complet et non sur les voies métaboliques séparées permet de détecter ce genre d'occurrences inter-voies. Dans ce cas, cette occurrence peut être interprétée comme une voie alternative de synthèse de la valine.

On peut ainsi mettre en évidence une proximité dans l'histoire évolutive des voies métaboliques.

[1] Les autres occurrences approchées du motif correspondaient à des groupes de réactions qui sont liées entre elles par des composés qu'on peut considérer comme secondaire (ATP, CO2). Nous avons choisi de ne pas les représenter sur la figure.

FIG. 4.2 – Occurrences du motif $M' = \{1.1.1, 1.1.1, 4.2.1, 2.6.1, 6.1.1\}$, avec un gap de taille 1.

Ce premier exemple est motivant dans le sens où, avec une définition simple de motif, on est capable d'extraire des résultats pertinents. C'est aussi l'occasion de faire un premier bilan sur les améliorations à apporter à la méthode pour pouvoir trouver des résultats plus novateurs. En effet, la découverte que la synthèse de la leucine, de la valine et de l'isoleucine sont proches n'est pas une découverte très surprenante. Du fait de la structure très proche de ces trois acides aminés, on s'attend à ce que leur synthèse se fasse selon le même mode.

À partir de cette première application, on peut faire les remarques suivantes :

1. le nombre d'occurrences croît rapidement quand on relaxe les paramètres. Certaines occurrences ne sont pas pertinentes (bruit).
2. certaines occurrences se recouvrent (*i.e.* ont des réactions en commun) tandis que d'autres sont clairement séparées. L'interprétation biologique sera différente selon ces configurations.
3. certaines occurrences sont incluses dans une voie métabolique (intra-voie) tandis que d'autres sont à cheval sur plusieurs voies (inter-voies). L'interprétation ne sera pas la même selon les cas.
4. faire de la recherche de motif implique qu'on connaît à l'avance le motif que l'on recherche.

Tout d'abord, concernant le premier point, il faut rappeller que cette première application a été menée avec les données de KEGG en utilisant l'ensemble des réactions de la base, tous organismes confondus. On peut donc potentiellement avoir des occurrences constituées de réactions qui ne sont jamais présentes ensemble dans le même organisme. Ce type de résultat est évidemment à évacuer (ce qui peut être fait facilement *a posteriori*) mais ceci ne constitue pas la plus grande source de bruit. La principale source de bruit vient vraisemblablement du traitement fait aux composés ubiquitaires dans KEGG. Ainsi, dans le cas des données de KEGG, nous avons considéré comme primaires les composés qui étaient desssinés sur les cartes métaboliques et comme secondaires les composés qui n'étaient pas dessinés. On rappelle que seuls les composés primaires sont utilisés pour construire le graphe (voir section 2.3.1.5). Ce choix s'est avéré discutable puisque ces cartes métaboliques ont été faites manuellement et il se trouve que certains composés auxiliaires comme le CO_2 ou l'eau y sont parfois indiqués (pour une meilleure compréhension). Ce type de composé crée des liens entre des réactions qui sont très différentes les unes des autres et a donc tendance à générer des occurrences artefactuelles (bruit).

Par ailleurs, de nombreuses erreurs ont été reportées dans KEGG dues à la propagation d'erreurs d'annotation. Pour plus de détails, on peut consulter [Green et Karp, 2005].

Enfin, nous voulions nous centrer sur un organisme particulier pour lequel les données étaient le plus fiable possible (pour éviter des problèmes dûs au

processus de reconstruction). Notre choix s'est donc porté sur *Escherichia coli* et sur la base EcoCyc.

Concernant le second point, se pose ici la question du schéma de comptage. Il existe plusieurs façons de compter les occurrences d'un motif dans un graphe suivant qu'on autorise les recouvrements entre occurrences :

- compter toutes les occurrences ;
- compter les paquets d'occurrences recouvrantes ;
- compter le nombre maximum d'occurrences non recouvrantes.

Par la suite, nous allons discuter des deux premiers types de comptage. Le troisième pose un problème d'optimisation, nous l'aborderons dans les Perspectives.

Concernant le point 3, en général, pour les motifs que nous avons testés, les occurrences inter-voies ont toujours été très nombreuses. Le fait qu'on trouve de telles occurrences illustre le fait qu'il existe des liens très forts entre voies métaboliques. Il est déjà connu que lorsqu'on centre une étude dynamique sur une unique voie métabolique, on peut être confronté à un phénomène de fuite de métabolites (*i.e.* un métabolite est en fait uilisé par une autre voie). En quelque sorte, on a une compétition entre voies métaboliques pour une même ressource et pour mieux appréhender ce phénomène, il est nécessaire de les modéliser conjointement.

Nous aimerions argumenter ici que cette idée de liens transversaux entre voies métaboliques est aussi intéressante à creuser pour les questions d'évolution. En effet, l'évolution de voies métaboliques ne se fait pas nécessairement en dupliquant une voie pour en donner une autre. On peut imaginer qu'une voie métabolique est composée de plusieurs blocs évolutifs et que ces blocs évolutifs peuvent éventuellement être à cheval sur plusieurs voies métaboliques.

Nous tenterons d'aller plus loin dans cette direction dans le chapitre 6.

Enfin, concernant le point 4, le chapitre suivant y apporte des réponses.

Chapitre 5

Inférence de motifs et sur-représentation

Dans le chapitre précédent, nous avons abordé la question suivante : à partir d'un fragment de réseau métabolique, quelles sont les fragments de réseaux qui lui sont proches ?

Cette question était abordée par le biais de la recherche de motif. Le motif était connu *a priori* et on recherchait ses occurrences dans le réseau. Dans l'application présentée (étude de la synthèse de la valine), nous nous trouvions en fait dans un cas particulier de recherche de motifs où on savait que le motif apparaissait au moins une fois. En effet, le motif était défini à l'aide d'une voie métabolique d'intérêt et on voulait savoir s'il apparaissait ailleurs.

Nous allons maintenant nous intéresser au cas où le motif que l'on cherche n'est pas connu. La question à laquelle on souhaite répondre est alors : quels sont tous les motifs qui apparaissent dans le graphe ? Nous nous trouvons à présent dans le cadre d'un problème d'inférence de motifs (on parle également d'extraction de motifs).

La différence entre recherche et inférence de motif est que, dans le premier cas, le motif est connu, et dans le second, il ne l'est pas. En pratique, l'inférence se fait rarement sans rien connaître du motif recherché et on spécifie généralement quelques-unes de ses caractéristiques. Dans notre cas, on devra spécifier la taille du motif ainsi que le seuil de comparaison utilisé pour les couleurs ("granularité" du motif).

Le nombre de contraintes qu'on peut considérer dans un problème d'inférence est variable. Plus on spécifie de contraintes, plus on s'éloigne d'un problème d'inférence pur. Un cas extrême est celui où on spécifie non seulement la taille et le seuil mais aussi toutes les couleurs du motif ; on se retrouve alors dans le cadre de la recherche de motifs. Ce petit raisonnement illustre le fait que, comme pour les motifs dans les séquences, il existe en fait un continuum entre un problème de recherche et un problème d'inférence de motifs.

Par la suite, outre les contraintes de taille et de seuil, on considérera également des contraintes sur les occurrences. Ainsi, on pourra se restreindre

aux motifs qui apparaissent au moins deux fois, ou encore aux motifs qui apparaissent au moins deux fois mais en ajoutant la contrainte que les occurrences soient disjointes.

5.1 Algorithme d'inférence

Le problème de l'inférence de motifs colorés peut être formulé de la manière suivante :

INFÉRENCE DE MOTIFS COLORÉS : Étant donné un graphe non-orienté coloré G et deux entiers k et s, trouver tous les motifs de taille k et de seuil s qui apparaissent dans G.

On rappelle que le seuil s correspond au score minimal pour considérer que deux couleurs sont similaires. Dans le cas d'un score basé sur la classification EC, ce seuil peut prendre les valeurs 1,2,3 ou 4 (la valeur 4 indique qu'on s'intéresse à des motifs exacts). Dans le cas plus général d'un score basé sur une classification hiérarchique, un seuil définit des classes d'équivalence entre couleurs. Le choix d'un seuil peut donc être vu comme une recoloration du graphe par de nouvelles couleurs (de "haut niveau") correspondant à ces classes d'équivalence. On retiendra qu'une diminution du seuil correspond à une diminution du nombre de couleurs.

Une solution naïve au problème d'inférence est de l'aborder comme une série de recherches de motifs (en utilisant de manière répétée l'algorithme de recherche de motifs présenté dans la section précédente). C'est par cette solution que nous avons commencé, et elle s'est avérée relativement performante sur les instances considérées et pour les tailles de motifs qui nous intéressaient. Ainsi, pour donner un ordre d'idée, avec cette méthode, on peut extraire du réseau métabolique d'*Escherichia coli* (600 noeuds et 1500 arêtes) tous les motifs de taille 6 et de seuil 2 en trois heures.

N'étant pas limité en temps, nous n'avons donc pas pour l'instant développé d'algorithme d'inférence de motif à proprement parler.

Dans les sections suivantes, nous allons tout d'abord présenter l'algorithme d'inférence tel qu'il est implémenté actuellement (série de recherches), puis nous donnerons quelques réflexions préliminaires qui pourront servir à l'avenir dans l'élaboration d'un algorithme d'inférence plus efficace.

5.1.1 La version actuelle de l'algorithme d'inférence

L'inférence de motifs peut être effectuée de façon naïve par une série de recherches de tous les motifs possibles. Le nombre de motifs possibles croît bien sûr exponentiellement avec la taille du motif. En effet, le nombre de motifs correspond au nombre de multiensembles de k éléments parmi l, avec $l = |C_s|$, le nombre de classes d'équivalences de couleurs au seuil s.

Le nombre de multiensembles n'est pas une quantité qu'on manipule très souvent. On peut noter que le nombre de multiensembles est différent du nombre de combinaisons de k éléments parmi l qui vaut C_l^k (la différence réside dans le fait que pour un multiensemble, les répétitions sont autorisées). Le nombre de multiensembles est aussi différent du nombre de k-listes l^k (car pour un multiensemble, l'ordre des éléments n'a pas d'importance). Le nombre de multiensembles de k parmi l est donné par le coefficient suivant : $M_l^k = C_{l+k-1}^k$.

À titre indicatif, pour $s = 2$, $|C_{s=2}| = 40$, et $C_{s=3} = 91$. Si on note $m(k, s)$ le nombre de motifs possibles de taille k avec un seuil s, on a $m(5, 2) = M_{40}^5 \simeq 10^6$, $m(6, 2) = M_{40}^6 \simeq 8.10^6$ et $m(5, 3) = M_{91}^5 \simeq 6.10^7$. La Figure 5.1 montre l'évolution du nombre de motifs possibles en fonction de k et s.

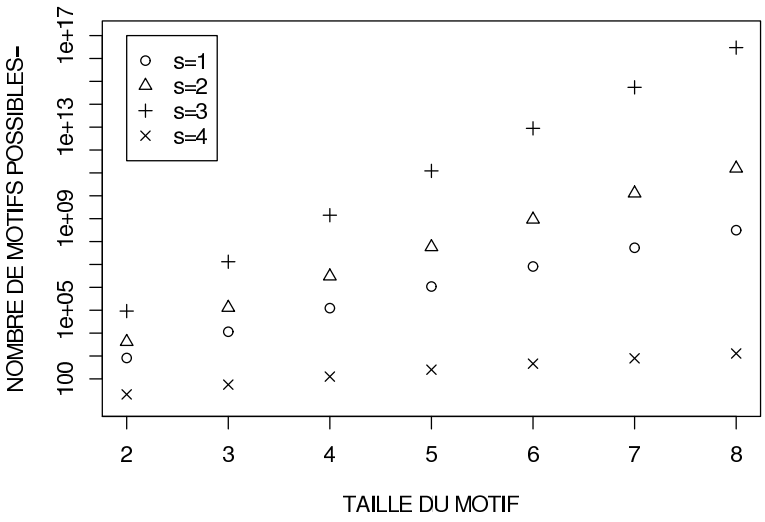


FIG. 5.1 – Nombre de motifs possibles selon la taille du motif et le seuil utilisé.

On constate ainsi que le nombre de motifs à tester croît très rapidement avec la taille du motif et le seuil considéré. Rechercher un à un tous les motifs peut être donc très long. Pour accélérer la méthode, on peut tirer parti de la structure hiérarchique de la classification EC et observer qu'une condition nécessaire pour qu'un motif de seuil 3 apparaisse est que le même motif défini au seuil 2 apparaisse aussi.

On peut donc se servir de la non-existence d'un motif au seuil s pour inférer la non-existence d'un motif au seuil $s + 1$. Par exemple, si le motif $\{1.1, 1.1\}$ n'apparaît pas, alors le motif $\{1.1.1, 1.1.2\}$ n'apparaîtra pas non plus. En pratique, on évite ainsi de parcourir de larges fractions de l'espace des motifs.

On peut d'ores et déjà remarquer que si on peut facilement utiliser des valeurs de s pour élaguer l'espace de recherche, il est plus difficile de faire un raisonnement semblable pour k. Ainsi, on ne peut pas facilement se servir de la non-existence d'un motif de taille k pour inférer la non-existence d'un

motif de taille $k+1$. Ceci est principalement dû au fait qu'on travaille sur des motifs non ordonnés. Plus de détails sont donnés dans la section 5.1.3.

Dans l'implémentation actuelle, si on souhaite inférer tous les motifs au seuil 4, l'algorithme commence par compter les motifs au seuil 2 puis au seuil 3. Nous avons délibérément choisi de ne pas inclure le seuil 1 dans l'élagage, car nous avons supposé que tous les motifs de seuil 1 apparaîtraient dans le réseau et donc que le gain de temps ne serait pas intéressant (en effet, l'élagage est inefficace si tous les motifs de seuil 1 apparaissent dans le graphe). Ce raisonnement s'avère en réalité discutable puisque dès la taille 4, on commence à avoir des motifs de seuil 1 qui n'apparaissent pas (par exemple, le motif $M = \{5, 5, 6, 6\}$ n'a pas d'occurrence).

Ce processus a l'avantage d'être facilement parallélisable. En effet, on peut lancer des recherches de motifs sur plusieurs machines.

Pour obtenir des découpages de tâches qui soient équitables, on peut se servir de la formule suivante :

Le nombre de motifs à tester M_n^k peut être décomposé en :
$M_n^k = M_n^{k-1} + M_{n-1}^{k-1} + M_{n-2}^{k-1} + M_{n-3}^{k-1} + ... + M_1^{k-1}$

Le premier terme de la somme correspond au nombre de motifs qui contiennent la première couleur (1.1 *), le deuxième au nombre de motifs qui contiennent la deuxième mais pas la première (1.2 *), etc. Les termes de la somme ne sont bien sûr pas tous égaux mais pour des valeurs de n et k connues, on peut choisir d'en regrouper certains (ou même d'en redécouper en sous-tâches, selon le même principe).

5.1.2 Temps d'exécution et limites actuelles

Dans cette section, nous présentons les résultats de temps d'exécution de l'algorithme d'inférence tel qu'il est implémenté. Toutes les estimations de temps ont été faites à l'aide de la commande GNU time, et correspondent au temps utilisateur (qui estime le temps d'exécution du programme sur une machine non stressée, *i.e.* qui n'a pas d'autre tâche à exécuter). Tous les essais ont été faits sur la même machine ayant les caractéristiques suivantes : AMD 64 bits, CPU 1.8GHz, 2Go de mémoire vive.

On constate que le temps d'exécution croît bien exponentiellement avec la taille du motif. On vérifie que le nombre de couleurs a une influence très importante sur le temps d'exécution.

L'inférence de motifs de taille 7 au seuil 1 n'a pas abouti, le point est donc absent. La limitation n'est pas due au temps mais à l'espace. En effet, nous verrons par la suite que le nombre d'occurrences par motif croît très rapidement et peut devenir un facteur limitant, en particulier dans le cas d'un petit nombre de couleurs.

Avec cette implémentation, et pour le jeu de données testé, nous sommes actuellement limités (en temps) à des motifs de taille 7. Pour une valeur de seuil égale à 2 (resp. 3), le temps d'exécution est de 25 heures (resp. 86

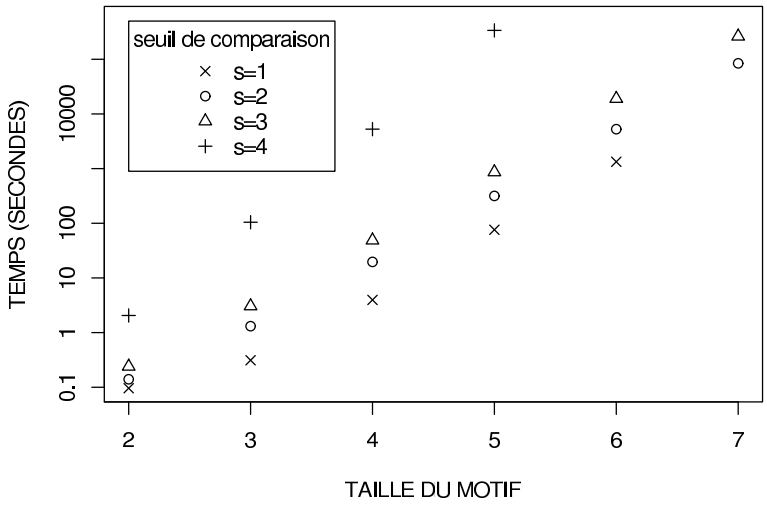


FIG. 5.2 – Temps d'exécution en fonction de la taille du motif.

heures). Pour une valeur de seuil égale à 4, le calcul n'a pas abouti pour des motifs de taille 6 et 7.

Lors de l'exécution du programme, nous avons pu remarquer que pour de longs motifs, l'algorithme passe plus de temps à chercher des motifs qui n'apparaissent pas plutôt qu'à compter des motifs qui sont effectivement présents. Éviter de rechercher des motifs qui n'apparaissent pas serait sans nul doute une piste intéressante à explorer.

5.1.3 Vers un algorithme d'inférence

Ce qui est coûteux dans l'algorithme précédent, c'est de tester tous les motifs possibles et de les rechercher un par un.

Nous allons présenter maintenant deux idées pour éviter de faire cela. La première consiste à énumérer tous les motifs à la fois (*i.e.* avec un seul algorithme qui a pour entrée le graphe et pour sortie tous les motifs). La seconde consiste à ne pas tester tous les motifs possibles mais seulement ceux dont on pense qu'ils apparaîtront et éviter de tester ceux pour lesquels on sait qu'il n'apparaîtront pas (condition nécessaire mais pas suffisante).

La première idée, qui consiste à énumérer tous les motifs en une seule fois, peut se faire à l'aide d'une adaptation de l'algorithme de recherche présenté dans la section précédente. On note que pour l'inférence, on ne se place plus dans le cas d'un graphe filtré ne contenant que les couleurs du motif recherché mais on travaille avec le graphe complet.

Pour rappel, l'idée générale de l'algorithme est d'énumérer tous les sous-graphes connexes de taille k du graphe dans lequel on recherche le motif, et de tester pour chacun si la condition de couleurs est satisfaite. L'énumération de tous les sous-graphes connexes se fait de la manière suivante : pour chaque noeud, on énumère tous les sous-graphes connexes qui le contiennent (par une

recherche en largeur mélangée à une stratégie de backtrack), puis on élimine le noeud et on passe au suivant.

Pour obtenir un algorithme d'inférence, on peut simplement modifier la partie concernant le test des couleurs. En effet, dans le cas de l'inférence de motif, on ne cherche pas un motif particulier mais tous les motifs à la fois. Ainsi, pour chaque sous-graphe connexe, il s'agit maintenant de décider de quel motif il est une occurrence. On note que dans le cas de graphes ayant une seule couleur par noeud, un sous-graphe connexe correspond à l'occurrence d'un unique motif.

La complexité d'un tel algorithme correspond à la complexité dans le pire cas de notre algorithme de recherche (le cas où tous les noeuds du graphe sont colorés de la même couleur et où le motif ne contient qu'une couleur répétée k fois).

Une difficulté nouvelle apparaît cependant, liée à l'espace que prend la solution. On peut utiliser une table de hachage ayant pour taille le nombre de motifs possibles. Pour chaque sous-graphe connexe parcouru, on met à jour la case de la table correspondant au motif dont le sous-graphe connexe est une occurrence.

On peut noter que ce schéma ne fonctionnerait que pour les motifs de petite taille car le nombre de motifs possibles croît exponentiellement alors que la mémoire vive d'un ordinateur est limitée. En pratique, sur une machine ayant 2 Go de mémoire vive, on peut estimer qu'on pourrait stocker des tables d'un milliard d'entiers au maximum. Le nombre de numéros EC au seuil 1 est 6, au seuil 2 est 40, au seuil 3 est 91, au seuil 4 est 428. On peut ainsi estimer traiter les tailles suivantes de motifs sans limitations de mémoire :

seuil de comparaison s	1	2	3	4
nombre de couleurs C_s	6	40	91	428
taille du motif k	161	8	6	3

Une limite évidente de cette approche est que beaucoup de cases du tableau resteraient vides.

On peut donc préférer utiliser une structure ayant pour taille le nombre de motifs trouvés et non le nombre de motifs possibles. En effet, le nombre de motifs observés est bien moindre que le nombre de motifs possibles. On peut donc espérer ne pas être limité aussi rapidement par la mémoire vive.

La Figure 5.3 montre la différence entre le nombre de motifs possibles et le nombre de motifs observés.

La deuxième idée que nous aimerions introduire consiste à ne pas tester tous les motifs possibles mais seulement ceux dont on pense qu'ils apparaîtront et éviter ceux pour lesquels on sait qu'il n'apparaîtront pas.

On peut observer que pour qu'un motif de taille k apparaisse, une condition nécessaire est que 2 des sous-motifs de taille $k-1$ qui le composent appa-

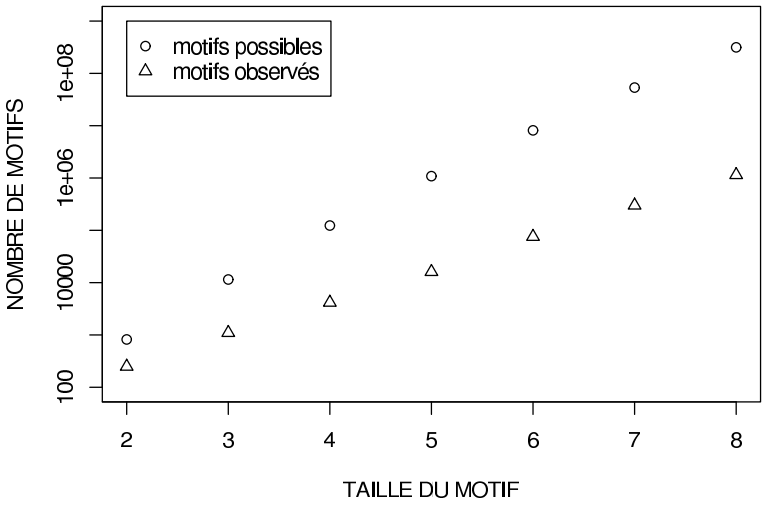


FIG. 5.3 – Nombre de motif possibles et nombre de motifs observés en fonction de la taille du motif.

raissent. Il est important de remarquer qu'il s'agit d'une condition nécessaire mais pas d'une condition suffisante. En effet, on peut très bien avoir deux motifs de taille $k-1$ qui apparaissent alors que le motif qui les rassemble n'apparaît pas. Un exemple est donné dans la figure 5.4 :

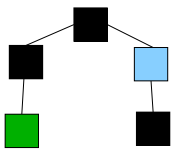


FIG. 5.4 – Dans ce graphe, le motif {noir, gris clair} et le motif {noir, gris foncé} apparaissent mais le motif {noir, gris clair, gris foncé} n'apparaît pas.

Pour un motif contenant plusieurs couleurs, on peut donc être certain qu'un motif de taille k n'apparaîtra pas si moins de 2 sous-motifs de taille $k-1$ qui le composent apparaissent.

On note que pour un motif unicolore (*i.e.* ne contenant qu'une seule couleur), la condition devient : pour qu'un motif unicolore apparaisse, il faut que son sous-motif de taille $k-1$ apparaisse 2 fois. En effet, un motif unicolore n'a qu'un seul sous-motif.

On peut ainsi tirer parti de cette observation pour décider quels sont les motifs qu'on va rechercher à partir des motifs de la taille inférieure.

Ainsi, si parmi les motifs de taille 2, on note que {rouge, vert} et {rouge, bleu} apparaissent, alors on peut décider de chercher {rouge, vert, bleu}. De même, si {rouge, vert, bleu} et {vert, bleu, jaune} apparaissent, alors on peut décider de chercher {rouge, vert, bleu, jaune}. De manière générale, la liste des motifs de taille k à rechercher peut être obtenue en faisant la jonction

des motifs de taille $k-1$ dont l'intersection a une taille $k-2$. La génération des motifs candidats de taille k à partir des motifs trouvés de taille $k-1$ est donc une étape clé de cet algorithme.

Si on note $Cand_k$ l'ensemble de motifs candidats de taille k et L_k l'ensemble des motifs trouvés de taille k, la relation entre L_k et $Cand_k$ s'écrit : $Cand_k = \{X \uplus X' | \, X; X' \in L_{k-1}, |X \cap X'| = k-2\}$, où $\uplus$ est la fonction de jonction pour des multiensembles (par exemple, $\{1,1\} \uplus \{1,2\} = \{1,1,1,2\}$ alors que $\{1,1\} \cup \{1,2\} = \{1,1,2\}$).

On peut noter des similarités entre la méthode proposée ici et l'algorithme "Apriori" proposé par [Agrawal *et al.*, 1993] et qui permet de détecter des ensembles fréquents dans un ensemble de transactions. La différence majeure réside dans la condition d'élagage qui est moins stricte dans notre cas. En effet, dans l'algorithme "Apriori", un ensemble n'est fréquent que si **tous** ses sous-ensembles sont fréquents.

Enfin, on note qu'on peut généraliser ce raisonnement à plusieurs occurrences. Pour qu'un motif de taille k apparaisse n fois, une condition nécessaire est que 2 des motifs de taille $k-1$ qui le composent apparaissent n fois.

5.2 Analyse préliminaire et nombre de solutions

On a pu voir que le temps d'exécution était "raisonnable" pour des motifs de taille inférieure à 7 (moins d'une journée de calcul). Le problème principal auquel on est confronté n'est pas un problème de temps mais plutôt un problème d'espace. En effet, le nombre de solutions croît très rapidement. Ce grand nombre de solutions pose un problème de stockage dans certains cas, mais il pose surtout un problème d'interprétation. On ne peut pas analyser manuellement chacune des occurrences.

Nous nous sommes donc posé les questions suivantes. Est-ce que toutes ces occurrences sont intéressantes biologiquement ? Peut-on en écarter certaines ? Selon quels critères ? Est-ce que toutes ces occurrences auront la même interprétation biologique ? Peut-on les classer par groupes ? Selon quels critères ?

Dans un premier temps, nous allons donner quelques chiffres concernant le nombre d'occurrences, puis nous discuterons quels critères de classement/filtrage peuvent être appliqués.

5.2.1 Le nombre de solutions

Tout d'abord, on peut noter que, pour un graphe où chaque noeud n'a qu'une couleur (c'est le cas du graphe construit à partir de la base EcoCyc), le nombre d'occurrences cumulé si on considère tous les motifs d'une taille k est égal au nombre de sous-graphes connexes de taille k. Cette valeur est indépendante du seuil utilisé (et donc du nombre de couleurs dans le graphe).

Le nombre de sous-graphes connexes de taille k dépend uniquement de la topologie du graphe. Plus ce graphe est dense, plus ce nombre sera élevé.

La Figure 5.5 montre que le nombre de sous-graphes connexes grandit exponentiellement avec la taille. L'enjeu de rechercher des motifs est en quelque

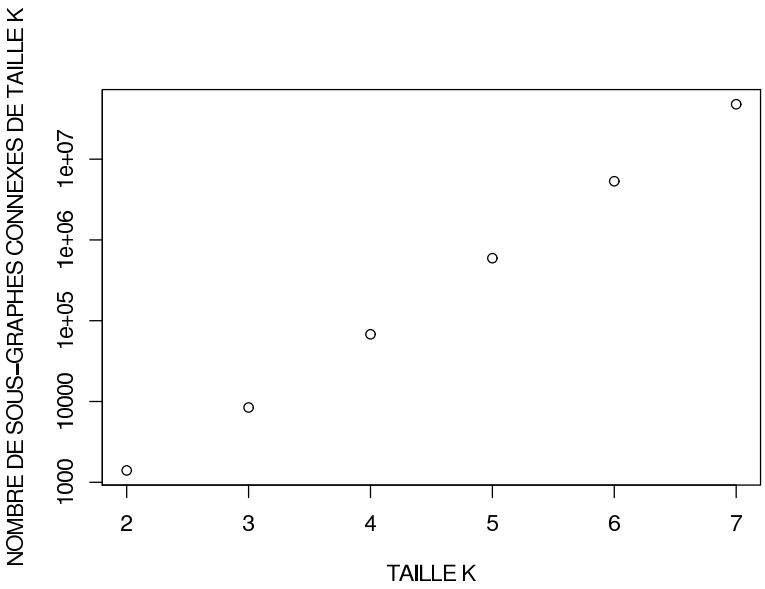


FIG. 5.5 – Nombre de sous-graphes connexes de taille k.

sorte de regrouper ces sous-graphes connexes par fonction commune.

Si on fait varier s, et donc C_s le nombre de couleurs, on va faire varier le nombre d'occurrences par motif. On joue ainsi sur la finesse du regroupement. Ainsi, le nombre moyen d'occurrences par motif, pour les motifs de taille 4, prend des valeurs très différentes selon les seuils : 551.1 ($s = 1$), 12.12 ($s = 2$), 4.14 ($s = 3$) et 1.13 ($s = 4$).

La Figure 5.6 montre le nombre d'occurrences par motif (pour un seuil 2, la tendance est la même pour d'autre seuils).

La première constatation qu'on peut faire est que le nombre d'occurrences par motif a tendance à augmenter avec la taille du motif. Cette observation est plutôt contre-intuitive puisqu'on s'attend à ce qu'un grand motif apparaisse moins de fois qu'un petit. En effet, quand on travaille avec les motifs dans les séquences, le fait d'agrandir un motif diminue son nombre d'occurrences. Il y a deux différences majeures dans notre cas : 1. on travaille sur des graphes (on peut donc étendre un motif de beaucoup plus que 2 manières car chaque noeud peut avoir plus de deux voisins), et 2. les couleurs ne sont pas ordonnées.

On peut cependant prédire que cet effet d'augmentation ne durera pas éternellement et que le nombre d'occurrences par motif va commencer à baisser pour des tailles de motif plus grandes. En effet, dans le cas limite où le motif a la taille du graphe, on n'a plus qu'un motif qui n'a plus qu'une occurrence.

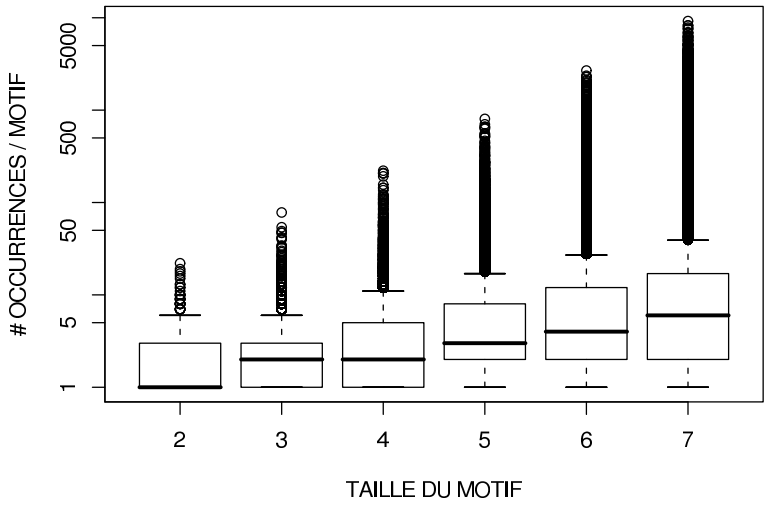


FIG. 5.6 – Nombre d'occurrences par motif en fonction de la taille du motif.

La seconde constatation qu'on peut faire est que le nombre d'occurrences est variable pour des motifs d'une même taille. Certains motifs sont très représentés, d'autres sont très peu présents. Une raison évidente est que les couleurs qui composent les motifs ne sont pas toutes aussi fréquentes. On s'attend donc à ce qu'un motif constitué de couleurs fréquentes soit très représenté. Mais la fréquence des couleurs n'explique sans doute pas toutes les différences de représentation des motifs. Dans la section 5.3.2, nous verrons comment définir la notion de motif sur-représenté (*i.e.* plus représenté qu'attendu par hasard).

5.2.2 Le regroupement, le dégroupement ou le filtrage des solutions, vers une autre définition de motif

Nous avons pu constater que le nombre d'occurrences augmentait très rapidement avec la taille du motif considéré. En effet, pour une taille donnée, le nombre total d'occurrences est égal au nombre de sous-graphes connexes de taille k. Suivant la valeur de s, on regroupe plus ou moins finement les occurrences. Une question qu'on peut alors se poser est : toutes les occurrences sont-elles intéressantes à examiner ?

Dans cette section, nous allons proposer deux méthodes de regroupement d'occurrences que nous allons exploiter par la suite. La première méthode, le regroupement par recouvrement, consiste à regrouper les occurrences qui partagent des noeuds. La deuxième méthode, le regroupement par topologie, consiste à regrouper les occurrences qui partagent la même topologie.

La motivation pour utiliser un regroupement par recouvrement vient du fait qu'on ne donne pas nécessairement la même interprétation à des occurrences qui sont disjointes et à des occurrences qui se recouvrent. Dans l'exemple de la synthèse de la valine (section 4.2), on avait plusieurs paquets

d'occurrences. L'un correspondait à la synthèse de la valine, un autre à la synthèse de la leucine et un troisième à la synthèse de l'isoleucine. Pour la synthèse de la valine, on avait en fait deux occurrences qui se recouvraient. L'interprétation qu'on en donnait était qu'elles constituaient des voies alternatives pour la synthèse de la valine.

Nous proposons donc de regrouper ensemble les occurrences qui se recouvrent. Il existe plusieurs manières de faire cela. Un groupe d'occurrences qui se recouvrent (ce que nous appellerons par la suite un *paquet d'occurrences*) est défini dans notre cas par simple lien, c'est-à-dire que, pour chaque occurrence du paquet, il existe une autre occurrence du paquet avec qui elle partage un noeud. On note que cela ne signifie pas que chaque paire d'occurrences d'un paquet partage des noeuds.

Du point de vue de la terminologie, les termes de paquet, ou de *train d'occurrences* (en anglais "clumps") ont déjà été employés dans le cadre des motifs dans les séquences pour désigner des concepts analogues [Schbath, 1995]. On préfèrera dans notre cas utiliser le terme paquet.

Enfin, si on considère le graphe formé par les noeuds des occurrences d'un motif, chaque paquet correspond à une composante connexe de ce graphe. On pourra également utiliser le terme de composante connexe d'occurrences pour désigner un paquet.

Le second regroupement que nous proposons consiste à regrouper les occurrences qui, en plus d'avoir les même couleurs, ont la même topologie. On peut noter que cela revient à modifier notre définition de motif coloré pour la raffiner en motif topologique coloré.

On peut remarquer que, lorsqu'on travaille dans le graphe des réactions, la topologie est relativement dégénérée (une même topologie dans le graphe des réactions peut correspondre à des topologies différentes dans le graphe biparti), ce qui ne pose pas de problème quand on s'intéresse simplement à la connectivité, mais qui en pose quand on veut discriminer finement différentes topologies. Ainsi, pour le regroupement topologique, on s'intéressera à la topologie de l'occurrence dans le graphe biparti. On rappelle qu'une occurrence est simplement définie par un ensemble de réactions (pas par sa topologie) et est donc indépendante du graphe dans lequel on travaille. Cette observation est illustrée dans la Figure 5.7.

En pratique, le regroupement par topologie est fait à l'aide de Nauty [McKay, 1990], un programme qui traite le problème d'isomorphisme de graphes, avec lequel nous avons interfacé notre algorithme.

On peut noter que, quand on compare la topologie de deux occurrences, on ne prend en compte que les composés internes (*i.e.* les composés qui lient des réactions de l'occurrence) et pas les composés externes.

Enfin, on observe que ces deux types de regroupement peuvent être appliqués l'un à la suite de l'autre.

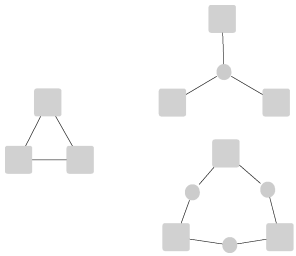


FIG. 5.7 – Illustration de la nécessité de travailler avec le graphe biparti pour classer plus finement les topologies. Différents graphes bipartis (graphes de droite) peuvent être représentés par un même graphe de réactions (graphe de gauche).

5.2.3 L'interprétation des solutions - Contexte voies métaboliques - Dessin de graphes

Le problème abordé dans cette section résulte du fait qu'une solution consiste en un ensemble de réactions connectées. Quand on est confronté aux sorties du programme, on peut être dans un premier temps désorienté, surtout si on ne connaît pas les noms des réactions.

Pour interpréter une occurrence (*i.e.* comprendre à quel(s) processus biologique(s) elle correspond), une méthode intéressante est de la replacer dans son contexte, et notamment essayer de voir à quelles voies métaboliques les réactions qui forment l'occurrence sont classiquement assignées. Par exemple, dans le cas de l'analyse de la synthèse de la valine (Section 4.2), nous avons pu interpréter facilement les occurrences qui étaient "intra-voie" (*i.e.* toutes les réactions appartenaient à la même voie). Pour cela, nous avons notamment utilisé les dessins de voies métaboliques qui sont disponibles sur le site de KEGG (de tels dessins sont également disponibles pour EcoCyc). Mais dans le cas d'occurrences "inter-voie" (*i.e.* toutes les réactions ne sont pas assignées à la même voie métabolique), aucun dessin n'est disponible et il faut naviguer entre les dessins des différentes voies métaboliques concernées.

Il existe des dessins qui représentent le réseau métabolique complet. La carte de Boeringher en est la représentation la plus connue (voir Figure **??**) et reste une référence dans le domaine. Mais ce type de dessin n'est pas facilement modifiable puisqu'il a été fait manuellement.

Pour aborder ce problème d'interprétation de nos occurrences, qui s'avère être lié à des problèmes de dessins de réseaux métaboliques, nous avons initié une collaboration avec Fabien Jourdan, (CR INRA travaillant au laboratoire de Xénobiotiques de Toulouse, UMR 1089 INRA-ENVT).

Il a résulté de cette collaboration la production d'une méthode de visualisation de réseaux métaboliques qui permette à la fois de voir l'intégralité du réseau tout en représentant conjointement la notion de voie métabolique.

L'originalité de l'approche, par rapport aux méthodes existantes, pro-

vient de la contrainte de non-duplication des noeuds du graphe. Pour plus de détails, le lecteur intéressé peut consulter l'article correspondant placé en annexe C [Bourqui *et al.*, 2007].

5.3 Statistiques dans les graphes

Cette section est commune autant à la recherche qu'à l'inférence de motifs. On l'introduit ici car la notion de sur-représentation devient d'autant plus cruciale qu'on s'intéresse à un grand nombre de motifs à la fois.

Nous avons pu constater dans la section précédente que certains motifs étaient très représentés (*i.e.* leur nombre d'occurrences était élevé). Or nous avons fait la remarque que le nombre d'occurrences d'un motif pouvait en partie être expliqué par la fréquence des couleurs qui composent le motif. Ainsi, il est attendu qu'un motif constitué de couleurs fréquentes apparaisse un grand nombre de fois dans le réseau. Le nombre d'occurrences d'un motif n'est donc pas nécessairement une mesure très informative. Plutôt que de simplement savoir si un motif est très présent, il peut être intéressant de savoir s'il est plus présent qu'attendu.

L'hypothèse qu'on fait alors est que les motifs qui sont plus présents qu'attendu, les motifs sur-représentés, ont une signification biologique particulière.

Définir ce qui est attendu est donc crucial dans cette approche. C'est le sujet de cette section.

5.3.1 L'importance du modèle de graphe aléatoire

Décider si un motif est sur-représenté revient à tester une hypothèse. De manière classique, on se fixe une hypothèse nulle H_0 et un risque α (risque de rejeter l'hypothèse alors qu'elle est vraie). L'hypothèse nulle dans ce cas est : le motif n'est pas plus présent qu'attendu. Si on rejette cette hypothèse, on décidera que le motif est sur-représenté (avec un risque α de se tromper). Tester la sous-représentation d'un motif se fait de la même manière. La seule différence est que l'hypothèse nulle est alors que le motif n'est pas moins présent qu'attendu.

Le terme "attendu" est bien sûr à définir. Par exemple dans notre cas, on s'attend à ce que les motifs constitués de couleurs fréquentes soient plus présents que les autres. Le comptage observé du motif est à comparer au comptage attendu du motif sous un modèle nul.

Le modèle de graphe aléatoire qu'on va utiliser pour déterminer le comptage attendu de notre motif va donc devoir incorporer ces informations de fréquences de couleurs (au même titre que pour décider si un motif est sur-représenté dans une séquence d'ADN, on utilise un modèle de séquence aléatoire qui préserve les fréquences des bases A,C,G,T, voire des mots d'une certaine longueur).

En plus des fréquences des couleurs, le modèle de graphe aléatoire devra également préserver certaines caractéristiques topologiques du graphe métabolique. On note que la quantité d'information à injecter dans le modèle nul est difficile à évaluer. Si on injecte trop peu d'informations, on concluera que tous les motifs sont sur-représentés. Si on injecte trop d'informations, on concluera qu'aucun motif n'est sur-représenté.

Mais on peut argumenter que la sur-représentation n'est pas la finalité à laquelle on s'intéresse. En effet, la raison pour laquelle on s'intéresse à la sur-représentation est parce qu'on fait l'hypothèse qu'un motif sur-représenté est un motif fonctionnel. Certains auteurs font même l'hypothèse qu'un motif est sur-représenté car il a été sélectionné au cours de l'évolution. Pour pouvoir tirer ce type de conclusions, le choix du modèle de graphe aléatoire est crucial.

En effet, quel que soit le modèle de graphe aléatoire qu'on utilise, si on rejette l'hypothèse H_0, on conclut que le motif est plus présent qu'attendu sous ce modèle nul. Si on veut étendre notre conclusion et dire que notre motif est fonctionnel, ou qu'il a été sélectionné au cours de l'évolution, alors il faut que l'abondance du motif ne puisse pas être expliquée par un processus évolutif neutre. Il faut donc que notre modèle de graphe aléatoire encapsule toute information relative à un processus neutre connu.

Ainsi, on pourra distinguer les modèles de graphes aléatoires qui servent pour poser des hypothèses nulles des modèles de graphes aléatoires qui servent à tester une hypothèse évolutive.

Pour illustrer l'importance du choix du modèle de graphe aléatoire, on peut mentionner un exemple qui concerne la recherche de motifs topologiques dans un graphe représentant le réseau de neurones du nématode *Caenorabditis elegans* [Milo *et al.*, 2002]. Dans ce travail, les auteurs utilisent un modèle de graphe aléatoire dans lequel la seule contrainte est de maintenir la séquence de degrés des noeuds du graphe. Avec ce modèle, ils détectent plusieurs motifs sur-représentés et tirent notamment la conclusion que le réseau de neurones du nématode présente des similarités structurales avec d'autres classes de réseau comme internet ou des réseaux électriques. Il s'avère en fait qu'en choisissant un modèle de graphe aléatoire adapté à la situation, dans ce cas un modèle prenant en compte l'organisation spatiale des neurones dans le réseau, ces motifs ne sont plus détectés comme sur-représentés [Artzy-Randrup *et al.*, 2004].

5.3.2 Différents modèles de graphe aléatoire

Aucun des modèles dont nous allons discuter ne prend en compte l'évolution du métabolisme. Par évolution, on entend prise en compte du temps.

Une question qu'on peut alors se poser est : est-ce qu'on cherche un modèle de graphe qui soit réaliste du point de vue évolutif ou un modèle de graphe qui modélise bien certaines propriétés observées ?

Les motifs sur-représentés seront des déviations au modèle. Plus le modèle est riche, moins on aura de déviations. L'idéal est d'incorporer dans le modèle

les choses qu'on ne veut pas retrouver dans les déviations.

Les trois modèles principaux que nous avons considéré sont :

- Modèle de Erdos-Renyi ;
- Modèle ERMG ;
- Modèle à topologie fixe.

5.3.2.1 Erdös-Rényi

Le modèle d'Erdös-Rényi (ER) est sans doute le modèle de graphe aléatoire qui a été le plus étudié dans la littérature. Il a l'avantage d'être très simple puisqu'il ne dépend que de deux paramètres : n le nombre de noeuds et p la probabilité que deux noeuds soient connectés par une arête.

Une limite majeure de ce modèle de graphe aléatoire est qu'il modélise mal la topologie des réseaux métaboliques (et notamment la distribution du degré des noeuds).

En effet, le modèle suppose que les arêtes sont indépendantes et sont présentes avec une probabilité p. Si on note X_{ij} la variable indiquant la présence d'une arête entre les noeuds i et j, on a :

$$X_{ij} \sim \mathcal{B}(p).$$

Dans ce modèle, le degré de chaque noeud suit une loi binomiale, qui peut être approchée par une loi de Poisson pour n grand et p petit. Si on note $\lambda = (n-1)p$ et K_i le degré du noeud i, on obtient :

$$K_i \sim \mathcal{B}(n-1, p) \approx \mathcal{P}(\lambda).$$

Or dans un réseau métabolique (comme dans beaucoup d'autres types de réseaux issus de données réelles), le degré d'un noeud ne suit pas du tout une loi de Poisson. La Figure 5.8 illustre cette observation pour le réseau métabolique d'*Escherichia coli*.

Comme alternative à la distribution de Poisson, la distribution "scale-free" (ou distribution de Zipf) a été largement utilisée pour modéliser la distribution des degrés des noeuds d'un réseau [Jeong *et al.*, 2000]. Cette distribution est définie comme suit :

$$Pr\{K_i = k\} = c(\rho)k^{-(\rho+1)}$$

où k est un entier positif, ρ un réel positif, $c(\rho) = \sum_{k\geq 1} k^{-(\rho+1)} = 1/\zeta(\rho+1)$ et $\zeta(\rho+1)$ est la fonction zeta de Riemann.

Il est important de noter que cette distribution est généralement utilisée pour modéliser une queue de distribution. En pratique, on peut montrer que l'ajustement est en effet meilleur pour la queue de distribution des degrés que pour la distribution complète (voir article joint en annexe C).

Comme alternative à la loi de Zipf, on peut également supposer que le mauvais ajustement de la loi de Poisson est simplement dû à une hétérogénéité entre les noeuds, certains étant plus connectés que d'autres.

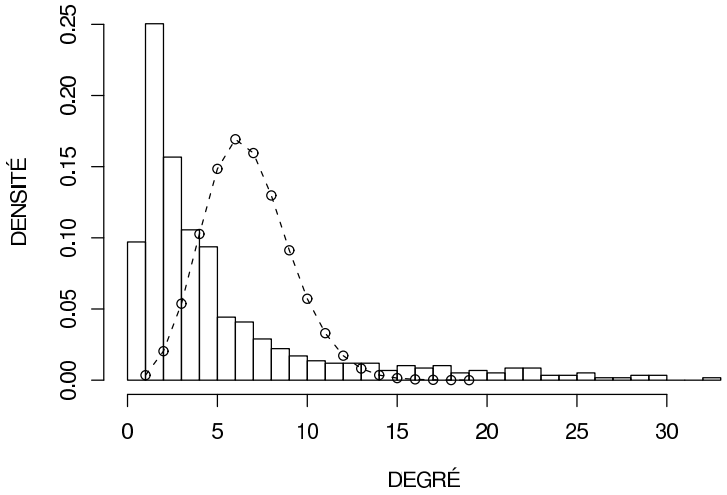


FIG. 5.8 – Histogramme des degrés des noeuds avec la distribution de Poisson ajustée.

Le modèle ERMG décrit dans la section suivante est une généralisation du modèle de Erdös qui permet de prendre en compte une telle hétérogénéité.

Ce modèle a été développé par des collaborateurs statisticiens : Stéphane Robin, Jean-Jacques Daudin et Franck Picard qui travaillent au laboratoire OMIP de l'INA-PG à Paris. Nous avons par ailleurs appliqué ce modèle au réseau métabolique d'*Escherichia coli* pour vérifier qu'il parvenait à capturer les caractéristiques structurales principales du réseau. Ce travail a été présenté aux journées thématiques "Réseaux d'interactions : analyse, modélisation et simulation" (RIAMS), l'article correspondant est disponible en annexe.

5.3.2.2 ERMG

Dans le modèle ERMG ("Erdös-Rényi Mixture for Graphs") proposé par [Daudin *et al.*, 2006], on suppose que les noeuds sont structurés en Q groupes, et que la probabilité de connexion de deux noeuds dépend des groupes auxquels ils appartiennent. Les noeuds appartiennent aux groupes avec une probabilité α_q.

Ainsi, si on note π_{ql} la probabilité pour un noeud du groupe q d'être connecté avec un noeud du groupe l, et qu'on suppose que les arêtes sont conditionnellement indépendantes étant donné les groupes, on obtient :

$$X_{ij} \mid \{i \in q, j \in \ell\} \sim \mathcal{B}(\pi_{q\ell}).$$

Ce modèle permet donc de décrire de manière fine la distribution des degrés, mais de manière plus générale, il décrit la topologie complète du réseau via la matrice de connectivité $\mathbf{\Pi} = (\pi_{q\ell})$.

La méthode d'estimation des paramètres (α_q et π_{ql}) est donnée dans [Daudin *et al.*, 2006] et une discussion sur la pertinence des groupes sélectionnés par la méthode est donnée dans l'article joint en annexe.

Les principales conclusions auxquelles nous sommes parvenus concernant l'application du modèle ERMG au réseau métabolique d'*Escherichia coli* sont les suivantes :

- le modèle ERMG est plus réaliste que d'autres modèles (par exemple ER, "scale-free") pour modéliser la distribution des degrés des noeuds et le coefficient de clustering du réseau étudié ;
- les groupes identifiés ont un sens biologique : ils correspondent à des composés ayant des rôles clés dans le métabolisme (ATP, pyruvate, L-glutamate) et autour desquels le reste des voies métaboliques s'organise.

En ce sens, ERMG est un modèle satisfaisant pour remplir le rôle de modèle servant à tester si un motif est sur-représenté car il capture des caractéristiques essentielles de la structure du métabolisme (distribution des noeuds, clustering, composés centraux).

Des formules exactes sont d'ores et déjà disponibles pour le comptage attendu de motifs topologiques dans le modèle ERMG, ce qui permet donc de décider si un motif topologique est sur-représenté pour un tel modèle de graphes [Picard *et al.*, 2007]. Le travail est actuellement en cours pour dériver une formule analytique qui donnerait le comptage attendu de motifs colorés dans ce type de modèles. On peut noter qu'une des difficultés dans ce cas, qui apparaît déjà dans le cas plus simple du modèle de Erdös-Rényi, est le calcul de la variance du comptage, et plus précisément le calcul de la probabilité que deux sous-graphes de taille k partageant entre 2 et $k-1$ sommets soient chacun connexe.

Jusqu'à présent, nous avons essentiellement discuté de la capacité des modèles de graphes aléatoires à rendre compte de la topologie du réseau étudié. Or dans notre cas, on a affaire à des graphes colorés. Il faut donc également se préoccuper de la gestion des couleurs. Dans chacun des cas mentionnés, les couleurs peuvent être ajoutées au modèle indépendamment de la topologie.

5.3.2.3 Modèle à topologie fixe

Une méthode simple pour s'affranchir des problèmes de réalisme du modèle de graphe aléatoire vis-à-vis de la topologie du réseau observé est de fixer la topologie.

En parallèle du développement de ERMG, nous avons décidé d'adopter une telle méthode, qui présente l'avantage d'être facile à mettre en pratique.

Dans ce modèle, la topologie est fixée et les couleurs sont mélangées. Le fait de travailler à topologie fixe entraîne qu'il est *a priori* difficile de dériver des formules analytiques pour le comptage des motifs dans un tel modèle. Ces formules ne seraient par ailleurs valables que dans le cadre très contraint du graphe qu'on étudie.

D'autre part, même dans le cas où on parvient à obtenir des formules pour l'espérance et la variance du comptage, il semble que dans ce cas, on ne puisse pas utiliser la loi de Poisson composée (proposée dans [Picard *et al.*, 2007]) pour modéliser la loi du comptage. En effet, le nombre de paquets d'occurrences (nombre de composantes connexes formées par les occurrences recouvrantes) ne suit pas une loi de Poisson.

En pratique, nous avons donc opté pour une approche par simulations. On génère un grand nombre de graphes par permutation des couleurs, puis on compte le motif recherché dans chacun des graphes simulés. On obtient ainsi une distribution approchée de celle du comptage attendu. On peut alors, soit estimer la moyenne et la variance du comptage et calculer un Z-score (score normalisé qui peut servir à classer les motifs), soit directement estimer une p-valeur, égale à la proportion de graphes simulés qui ont un comptage supérieur à celui observé dans le graphe métabolique. La seconde méthode nécessite un plus grand nombre de simulations mais permet de travailler directement avec des p-valeurs sans avoir à faire d'hypothèse sur la distribution du comptage, qui est *a priori* inconnue (voir section suivante).

5.3.3 La p-valeur sur le nombre de composantes connexes

Le nombre d'occurrences, ou plus exactement la déviation de ce comptage par rapport au comptage attendu est un critère classique pour évaluer l'importance d'un motif.

On peut cependant être également intéressé par la façon dont les occurrences s'arrangent les unes par rapport aux autres (recouvrement). Dès lors, on ne s'intéresse plus uniquement au nombre d'occurrences, mais au nombre de paquets d'occurrences. De la même manière qu'on teste le caractère exceptionel d'un comptage d'occurrences, on peut alors tester l'exceptionalité d'un comptage de paquets d'occurrences.

Par la suite, nous utiliserons à la fois le nombre d'occurrences et le nombre de paquets d'occurrences comme mesures pour analyser nos motifs.

Comme nous l'avons mentionné précédemment, le terme de paquet est également utilisé pour désigner les mots qui se recouvrent dans les séquences. Pour les mots dans les séquences, le nombre de paquets suit classiquement une loi de Poisson et la taille des trains (nombre de motifs qui se recouvrent) suit une loi géométrique. Au final, le nombre d'occurrences d'un mot dans une séquence est généralement bien modélisé par une loi de Poisson composée.

Dans [Picard *et al.*, 2007], le même type d'hypothèse est fait dans le cadre des motifs topologiques sur le comptage des occurrences, des paquets d'occurrences et de la taille des paquets.

Dans notre cas, on peut tester si ces hypothèses sont valides. Nous allons notamment tester si le nombre de paquets d'occurrences suit une loi de Poisson.

La Figure 5.9 montre la distribution du comptage du nombre de paquets d'occurrences pour un motif de taille 3. Un test d'ajustement rejette l'hy-

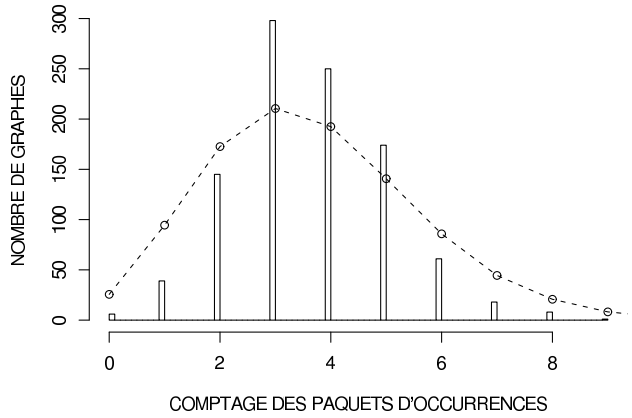


FIG. 5.9 – Histogramme des paquets d'occurrences avec la distribution de Poisson ajustée.

pothèse que ce comptage suit une loi de Poisson ($chi2 = 151, p < 10^{-16}$). Le nombre de composantes connexes (ou nombre de trains) ne suit pas une loi de Poisson dans un modèle de graphe aléatoire à topologie fixe.

Ce résultat indique que, au moins pour le modèle de graphe aléatoire à topologie fixe, l'hypothèse classiquement faite que le comptage d'un motif suit une loi de poisson composée risque de ne pas être vérifiée. En effet, cette hypothèse est généralement fondée sur l'idée que le comptage des trains suit une loi de Poisson.

5.3.4 Limitations actuelles de notre méthode

5.3.4.1 Précisions des estimations, nombre de simulations à effectuer

Dans le cadre des motifs dans les séquences, quand on veut décider si un motif est sur-représenté avec un risque α de 1%, et qu'on a recours à une estimation de la p-valeur, alors il est recommandé de faire 100 000 simulations (Stéphane Robin, communication personnelle)

En pratique, dans notre situation, 1000 simulations suffisent pour avoir une précision acceptable. On peut quantifier la qualité de l'estimation en donnant un intervalle de confiance autour de la p-valeur. La largeur de l'intervalle de confiance dépend directement du nombre de simulations faites. Pour ce faire, on peut utiliser la formule suivante de l'intervalle de confiance pour une proportion (formule de Wald) :

$$p = \hat{p} + -z_{\frac{\alpha}{2}} \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Cette formule utilise l'approximation de la loi binomiale par la loi normale qui est généralement considérée comme valable pour $n\hat{p} > 5$, $n(1-\hat{p}) > 5$ et n suffisamment grand.

Ainsi, à titre d'exemple, pour le motif {2.3, 2.7, 4.1}, on obtient les estimations suivantes (intervalles de confiance à 95%) :
$n = 100, \hat{p} = 0.34, p \in [0.2472; 0.4328]$
$n = 500, \hat{p} = 0.215, p \in [0.1968; 0.271]$
$n = 1000, \hat{p} = 0.268, p \in [0.2406; 0.2954]$
$n = 5000, \hat{p} = 0.2709, p \in [0.2586; 0.2832]$
$n = 10000, \hat{p} = 0.2697, p \in [0.2611; 0.2783]$

Pour un second exemple, le motif {1.1, 1.4, 2.6}, on obtient les estimations suivantes (intervalle à 95%) :
$n = 100, \hat{p} = 0$
$n = 500, \hat{p} = 0.006$
$n = 1000, \hat{p} = 0.0044$
$n = 5000, \hat{p} = 0.0043, p \in [0.0025; 0.0061]$
$n = 10000, \hat{p} = 0.0054, p \in [0.004; 0.0068]$

Aucun intervalle n'est calculé quand les conditions d'approximation de la loi normale ne sont pas remplies.

On peut ainsi conclure que l'erreur commise lors de l'estimation de la p-valeur est de l'ordre de 0.1 pour 100 simulations, de l'ordre de 0.03 pour 1000 simulations et de l'ordre de 0.015 pour 5000 simulations et de l'ordre de 0.01 pour 10000 simulations.

On note que l'erreur commise n'est pas quantifiable pour des p-valeurs faibles. En pratique, pour un motif ayant une p-valeur proche de 0.01, il faudra au moins 500 simulations pour obtenir un intervalle de confiance.

On retiendra que pour avoir une estimation précise de la p-valeur, il faut payer le prix en nombre de simulations et donc en temps de calcul.

Une solution pour éviter de faire autant de simulations est de ne pas estimer directement la p-valeur mais la moyenne et la variance du comptage (pour lesquels le nombre de simulations à effectuer est moins important). À partir de l'espérance et de la variance, on obtient un Z-score. Le problème qui reste à résoudre ensuite, si on veut obtenir une p-valeur, est de trouver une bonne approximation pour la loi suivie par le comptage.

5.3.4.2 Motifs sur-représentés à cause de motifs de taille inférieure

Aux cours de nos expériences, nous nous sommes aperçus que la p-valeur d'un motif était souvent significative lorsqu'il contenait un motif qui lui-même était sur-représenté.

Ce problème est bien connu dans le cadre des séquences et la méthode utilisée pour y faire face est d'utiliser un modèle de séquence aléatoire qui contrôle les mots de taille $k-1$ lorsqu'on cherche à travailler avec des mots de taille k (on utilise pour cela des modèles de Markov d'ordre $k-1$). Ainsi, lorsqu'un mot de taille k est considéré comme exceptionnel, cela n'est pas dû au fait qu'il contient des mots exceptionnels.

Dans le cadre des motifs dans les graphes, une telle approche n'est pas toujours possible. En effet, un modèle de graphe aléatoire à topologie fixe qui maintient le nombre de motifs de taille k est un modèle très contraint. En pratique, le nombre de graphes qui remplissent ces contraintes est très réduit (car on travaille avec de petits graphes (600 noeuds, 1500 arêtes) qui ont beaucoup de couleurs (40 lorsqu'on travaille au seuil 2)). Une conséquence d'utiliser un modèle trop contraint est qu'aucun motif ne peut être détecté comme sur-représenté.

On peut noter qu'il n'y a pas d'obstacle fondamental à utiliser une telle approche pour les motifs dans les graphes. Le problème vient principalement des caractéristiques des graphes qu'on considère. L'approche serait sans doute applicable pour de grands graphes possédant peu de couleurs.

Enfin, on peut noter que le même type de difficulté apparaît même pour des modèles d'ordre 0 (*i.e.* qui ne contrôlent que les motifs de longueur 1, c'est-à-dire les fréquences des couleurs) lorsque le nombre de couleurs est grand en comparaison du nombre de noeuds. Par exemple, dans le cas où on travaille avec une valeur de $s = 4$, le nombre de couleurs est $C_s = 428$. On se retrouve alors dans une situation où on a quasiment une couleur par noeud dans un graphe peu dense (1500 arêtes). Il se produit alors le phénomène suivant : un motif de taille 2 qui apparaît simplement une fois dans ce graphe a des chances très faibles d'apparaître dans un autre graphe de même topologie où les couleurs ont été mélangées. Cette probabilité est égale à $\frac{2m}{n(n-1)}$ dans le cas d'un graphe ayant m arêtes, n noeuds et une couleur par noeud. Dès lors, ce motif sera détecté comme sur-représenté par notre méthode (*i.e.* l'association de ces deux couleurs sera considérée comme exceptionnelle). Le problème que nous soulevons ici est que dans un graphe peu dense ayant beaucoup de couleurs, la majorité des associations de deux couleurs seront considérées comme exceptionnelles. La mesure de sur-représentation n'est donc plus informative dans ce cas.

La Figure 5.10 indique que la proportion de motifs sur-représentés grandit avec la taille du motif et que ceci est d'autant plus vrai que le seuil est élevé (beaucoup de couleurs).

À titre d'exemple, 80% des motifs de taille 4 au seuil 3 qu'on trouve sont sur-représentés (p-valeur inférieure à 0.05). Clairement, la notion de sur-représentation est peu informative dans ce cas. On observe donc bien un effet taille. Plus les motifs sont grands, plus la proportion de motifs qualifiés de sur-représentés est grande. On peut cependant noter qu'on trouve des motifs qui contiennent des motifs sur-représentés et qui eux-mêmes ne sont pas sur-représentés. On a également un effet seuil. En effet, pour un seuil élevé (beaucoup de couleurs), la plupart des motifs sont sur-représentés.

On note qu'une solution à ce problème consiste à utiliser des seuils de décision (risque α) différents selon les seuils de comparaison des numéros EC. En effet, quand on travaille avec une valeur de s élevée, certaines couleurs sont peu fréquentes. Une solution consiste à baisser le risque α pour repousser plus loin la décision qu'un motif est sur-représenté. Mais pour que

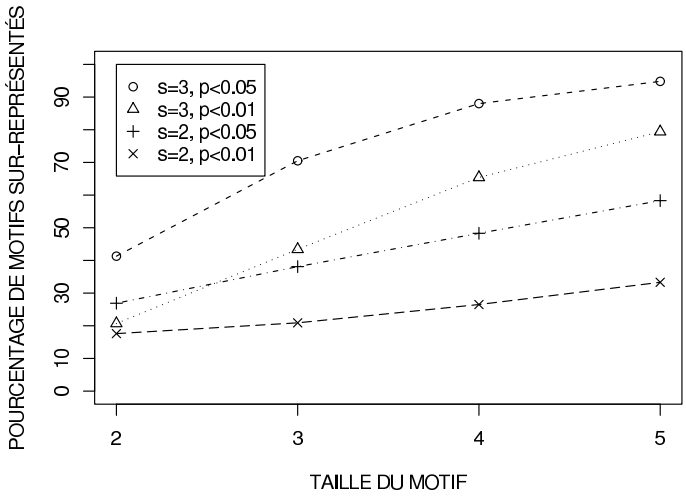


FIG. 5.10 – Pourcentage de motifs sur-représentés en fonction de la taille du motif.

les estimations des p-valeurs soient fiables, il faut augmenter le nombre de simulations (nombre de graphes aléatoires générés).

À l'heure actuelle, nous n'avons pas de moyen satisfaisant pour traiter ces problèmes. Deux pistes semblent intéressantes à suivre. La première consiste à prendre un modèle de graphe aléatoire moins contraint (ERMG par exemple). Dans un tel modèle, il devient sans doute plus envisageable de contrôler le nombre de motifs de taille 2, car le nombre de graphes satisfaisant ces contraintes devient plus important.

La seconde piste consiste, pour éviter les problèmes dus aux faibles comptages, à considérer des pseudo-compte (au lieu du comptage). Une approche bayésienne nous permettrait de nous placer dans un cadre formel rigoureux pour l'étude des pseudo-comptes.

Pour conclure, on dispose d'une mesure qui permet de classer les motifs et de décider s'ils sont ou non sur-représentés. On note que ce critère de classement est moins bon quand le motif est grand. On note aussi que ce critère est peu discriminant quand beaucoup de couleurs sont peu fréquentes (grandes valeurs de s). Pour pallier à ce problème, une solution consiste à adapter le seuil de décision (risque α) à la valeur de s.

On peut noter que, dès lors qu'on s'intéresse aux motifs répétés au moins deux fois (ce qu'on fait par la suite), le problème dû au nombre de couleurs se pose de manière beaucoup moins drastique. En effet, ce problème des motifs constitués de couleurs peu fréquentes concerne essentiellement des motifs qui n'apparaissent qu'une fois.

On retiendra de cette section que notre mesure de sur-représentation est à prendre avec des précautions pour les motifs de grande taille ou/et ceux contenant des couleurs de fréquence faible.

Chapitre 6

Applications de l'inférence de motifs

Dans les chapitres précédents, nous avons introduit des outils mathématiques qui nous paraissaient nécessaires pour répondre à une question centrale : existe-t-il une brique évolutive du métabolisme ? Si oui, comment la définit-on ? Comment l'identifie-t-on ?

Une partie importante du travail à présenter maintenant est d'évaluer la pertinence des définitions proposées. Cette étape d'évaluation est présentée après les développements méthodologiques. Ce plan est un peu trompeur car il ne reflète pas le va-et-vient qu'il y a eu entre modélisation et évaluation.

C'est au cours de ce va-et-vient que nous avons été amené à modifier le schéma de comptage des motifs pour priviliégier les solutions qui apparaissaient en plusieurs paquets. C'est ainsi également qu'on a pu se rendre compte des limites du modèle de graphe aléatoire qu'on avait. Enfin, c'est ainsi qu'on a pu voir qu'il pouvait être parfois intéressant de réintroduire la topologie comme contrainte.

Ce chapitre présente les analyses faites sur le réseau métabolique d'*Escherichia coli* et les discussions qui en découlent concernant la pertinence de notre définition de motif.

6.1 Présentation des données et du prétraitement

Nous avons fait l'essentiel de nos analyses sur le réseau métabolique de la bactérie *Escherichia coli*. La motivation principale pour le choix de cet organisme est que c'est celui pour lequel on dispose de données de meilleure qualité.

Les données que nous avons utilisées pour construire notre graphe proviennent de la base de données EcoCyc [Keseler *et al.*, 2005]. La version de la base utilisée est la 10.0 qui date du 13 Mars 2006.

Sans prétraitement (en considérant toutes les réactions et tous les composés de la base), le réseau comprend 812 réactions, 854 composés et 596

numéros EC. Le graphe des réactions reconstruit à partir de ces données contient 812 noeuds et 57213 arêtes.

Cependant, comme nous l'avons mentionné dans la section 2.3.1.5, travailler avec des données brutes pose problème. En effet, on risque de considérer comme connectées des réactions qui partagent simplement un de leurs substrats secondaires (par exemple H_2O ou ATP).

Nous procédons donc au prétraitement suivant : pour chaque réaction, les composés classés comme secondaires sont retirés de l'analyse. On note qu'un composé peut être classé comme secondaire dans une réaction et primaire dans une autre. Ce traitement est donc différent de retirer complètement un composé.

Ce classement des composés en primaire/secondaire n'est disponible dans EcoCyc que pour les réactions qui sont classées dans au moins une voie métabolique. Nous excluerons donc de notre analyse les réactions qui ne sont assignées à aucune voie métabolique.

Après le prétraitement, le réseau est constitué de 587 réactions, 553 composés et 463 numéros EC.

Le graphe des réactions reconstruit à partir de ces données contient alors 587 noeuds et 1667 arêtes.

D'autres prétraitements peuvent être considérés. Un type de prétraitement largement utilisé dans la littérature consiste à retirer du réseau les composés les plus fréquents (composés dits "ubiquitaires"). À titre indicatif, on peut noter que si on retire les 10 composés les plus fréquents, le graphe des réactions contient alors 9114 arêtes. Le choix du nombre de composés à retirer n'est pas évident. Pour choisir, on peut s'appuyer sur la distribution des degrés des noeuds dans le graphe biparti. De manière générale, nous avons choisi de ne pas privilégier cette méthode car elle nous a semblé supprimer des liens importants (des composés centraux comme le pyruvate ou le L-glutamate sont écartés dans ce type d'analyse).

6.2 Analyse systématique des motifs

Dans cette partie, nous allons présenter les résultats que nous avons obtenus en appliquant notre algorithme d'inférence de motifs au réseau métabolique d'*Escherichia coli*. Nous présenterons tout d'abord les résultats obtenus pour la recherche de motifs exacts (seuil de comparaison des numéros EC fixé à 4, c'est-à-dire que les numéros EC doivent être identiques). Puis nous nous intéresserons à la recherche de motifs approchés (seuil 3). Nous aurons alors recours aux techniques de filtrage présentées au chapitre précédent. Nous discuterons du regroupement par recouvrement et du regroupement par topologie. Au cours de cette analyse, certains motifs seront mis en avant et feront l'objet d'une étude séparée et approfondie.

L'analyse des motifs approchés est essentiellement centrée sur les motifs au seuil 3, on abordera succinctement les motifs au seuil 2. La principale

raison de ce choix est qu'un motif au seuil 2 a beaucoup plus d'occurrences qu'un motif au seuil 3 (les motifs au seuil 3 peuvent être vus comme des raffinements des motifs au seuil 2). Un motif au seuil 2 est donc d'une part plus long à analyser, et d'autre part il est susceptible de rassembler des occurrences assez différentes entre elles biologiquement.

Que ce soit pour les motifs exacts ou les motifs approchés, nous avons pu voir que le nombre de motifs extraits augmente rapidement avec la taille du motif (voir figure 5.3). Il est donc impossible de tous les analyser en détail. Nous avons ainsi choisi de nous centrer sur les motifs répétés (au moins deux occurrences) et plus particulièrement les motifs répétés en plusieurs paquets (au moins deux paquets d'occurrences).

6.2.1 Analyse des motifs exacts

Nous nous sommes intéressés aux motifs exacts de taille 1 à 5. Le Tableau 6.2.1 indique le nombre de motifs répétés et le nombre de motifs répétés en plusieurs paquets pour ces différentes tailles.

taille du motif	1	2	3	4	5
nombre de motifs répétés	21	56	460	4048	35682
nombre de motifs à plusieurs paquets	21	3	2	1	0

Tab. 6.1 – Nombre de motifs répétés et répétés en plusieurs paquets en fonction de la taille du motif.

On constate que le nombre de motifs répétés augmente très rapidement alors que le nombre de motifs répétés en plusieurs paquets a tendance à diminuer. Ce résultat est logique : plus on augmente la taille du motif, plus il a d'occurrences, et plus ces occurrences auront tendance à se recouvrir.

L'analyse des motifs répétés ne peut être faite à la main (il y en a trop). Cependant, on peut observer que pour qu'un motif soit répété, il faut qu'il contienne une couleur répétée. Ainsi, l'analyse des motifs exacts de taille 1 peut déjà nous donner des informations.

Un motif répété de taille 1 correspond au cas où plusieurs noeuds ont la même couleur, c'est-à-dire dans le cas du seuil 4, au cas où un numéro EC a été attribué à plusieurs réactions. Comme nous l'avons mentionné dans la section 3.3.3, ce cas est possible.

En voici la liste exhaustive : 1.1.1.23, 1.1.1.86, 2.2.1.1, 2.3.1.16, 2.4.1.1 (3), 2.5.1.48, 2.6.1.11, 2.6.1.42 (3), 2.6.1.57, 2.7.1.21, 2.7.1.35, 2.7.1.73, 2.7.4.6 (8), 2.7.6.5, 3.5.4.5, 4.2.1.3, 4.2.1.9, 4.3.2.2, 5.3.1.12, 5.4.2.7, 6.3.2.17 (3).

On peut noter que dans certains cas, les réactions sont catalysées par la même enzyme. C'est le cas des 8 réactions auxquelles on a attribué le numéro EC 2.7.4.6. L'enzyme en question est la nucleoside diphosphate kinase. Il s'agit d'une enzyme multifonctionnelle.

Dans d'autres cas, ce sont des enzymes distinctes qui catalysent les réactions. C'est le cas des 3 réactions auxquelles on a attribué le numéro EC 2.4.1.1. Les enzymes concernées sont la glycogène phosphorylase et la maltodextrine phosphorylase. Ce sont des isozymes qui ont des spécificités différentes. La maltodextrine phosphorylase a une forte affinité pour le maltotetraose alors que la glycogène phosphorylase a une forte affinité pour le glycogène.

Une différence majeure entre ces deux situations est que des isozymes peuvent être régulées différemment alors que dans le cas d'une enzyme multifonctionnelle, dès que l'enzyme est présente, toutes les réactions peuvent avoir lieu (pas de régulation différentielle des processus).

Le concept de motif exact peut donc correspondre à des réalités différentes au niveau des enzymes. Il est nécessaire de faire la différence entre les différents cas.

En ce qui concerne les motifs qui apparaissent en plusieurs paquets, étant donné qu'ils sont peu nombreux, nous avons pu les analyser plus en détail. Il apparaît que les 3 motifs de taille 2 se recouvrent (*i.e.* partagent des couleurs) deux à deux. Les motifs sont : {2.2.1.6, 1.1.1.86}, {1.1.1.86, 4.2.1.9} et {4.2.1.9, 2.6.1.42 }. Il en est de même pour les deux motifs de taille 3 ({2.2.1.6, 1.1.1.86, 4.2.1.9} et {1.1.1.86, 4.2.1.9, 2.6.1.42}). Ces motifs sont en réalité des sous-motifs de l'unique motif de taille 4 à plusieurs paquets : $M = \{2.2.1.6, 1.1.1.86, 4.2.1.9, 2.6.1.42\}$. On peut noter qu'à la taille 4, le motif est maximal au sens du nombre de paquets (*i.e.* il ne peut pas être étendu sans perdre la propriété d'avoir plusieurs paquets d'occurrences). Il n'existe aucun motif de taille 5 ayant plusieurs paquets.

Pour le motif de taille 4, les deux paquets d'occurrences (chaque paquet contient en fait une occurrence) peuvent être interprétés biologiquement. Le premier correspond à quatre étapes de la synthèse de la valine et le second à quatre étapes de la synthèse de l'isoleucine. On retrouve ici le résultat présenté dans la section 4.2. Par rapport à ce premier résultat, on note que le motif est maintenant allongé d'une étape en amont[1].

On se trouve dans un cas favorable pour l'interprétation car chaque occurrence est "intra-voie" (*i.e.* toutes les réactions de l'occurrence appartiennent à la même voie métabolique). Pour visualiser ces occurrences, on peut donc utiliser directement les dessins des voies métaboliques accessibles dans EcoCyc (voir les Figures 6.1 et 6.2).

On peut ainsi constater que non seulement les numéros EC coïncident mais que ce sont exactement les mêmes enzymes qui catalysent ces quatres étapes. On est donc dans le cas d'enzymes multifonctionnelles.

En outre, si on considère les positions sur le génome des 4 gènes impliqués

[1]Les autres différences sont dues à des différences entre les bases de données utilisées (KEGG dans le premier cas et EcoCyc maintenant). Ainsi dans KEGG, un substrat intermédiaire est identifié pour la réaction d'isomérorédution (1.1.1.86) ce qui entraîne que cette réaction y est représentée par deux réactions.

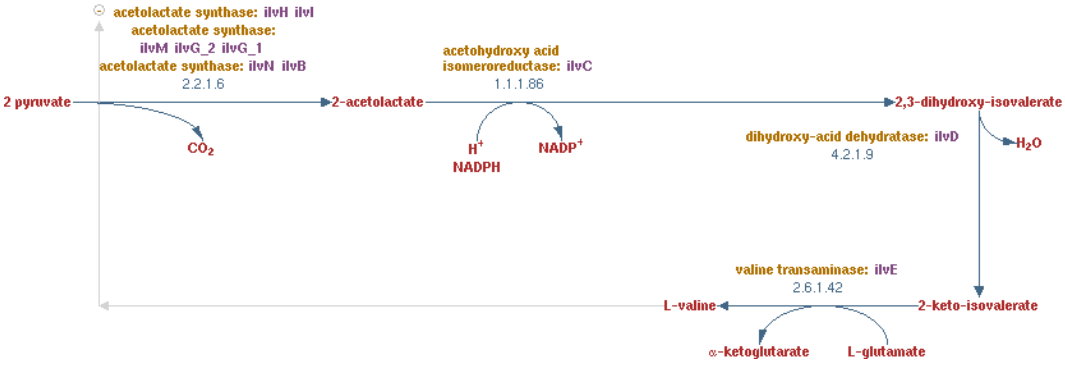


FIG. 6.1 – Dessin de la synthèse de la valine.

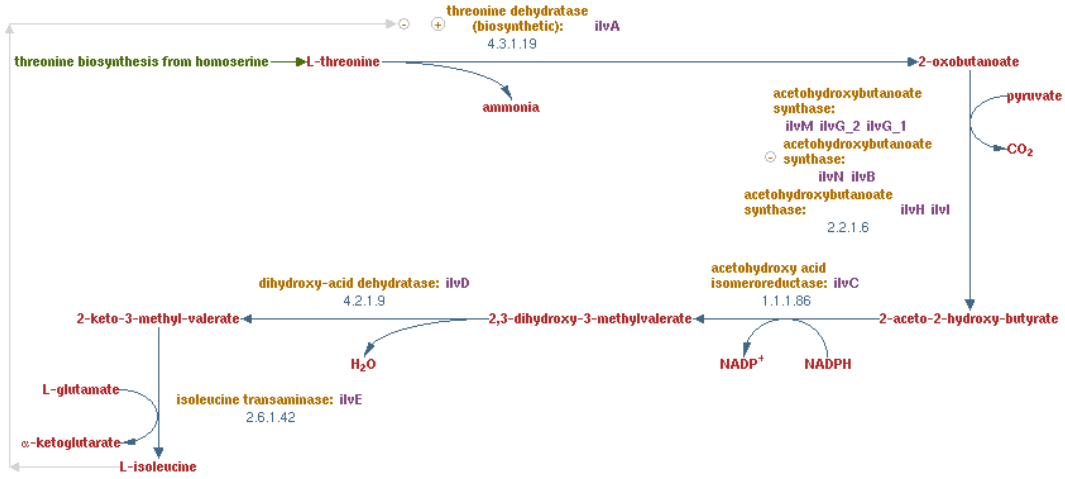


FIG. 6.2 – Dessin de la synthèse de l'isoleucine.

dans ces deux voies, on s'aperçoit qu'ils appartiennent à un même opéron[2]. On peut supposer que la pression de sélection pour maintenir cette opéron est très forte puisque plusieurs voies sont en jeu.

Par ailleurs, on peut remarquer que le fait d'obtenir ce résultat à partir de notre algorithme d'inférence de motif plutôt qu'avec notre algorithme de recherche nous permet d'affirmer qu'il s'agit d'un cas unique. Il n'existe pas d'autre motif exact de taille 4 qui soit répété en plusieurs paquets. Il n'existe d'ailleurs pas d'autre motif exact de taille 2 qui soit répété en plusieurs paquets. En effet, les seuls motifs de taille 2 et 3 sont des sous-motifs du motif de taille 4.

Finalement, on peut remarquer que le motif de taille 4 que nous avons considéré était un motif linéaire coloré et que l'ordre des couleurs était conservé entre les deux occurrences. Il s'agit donc d'un motif topologique coloré ordonné (*i.e.* en plus de partager les mêmes couleurs, les occurrences partagent la même topologie et l'ordre des couleurs).

6.2.2 Analyse de motifs approchés, au seuil 3

Concernant les motifs approchés, nous nous sommes intéressés aux motifs de taille 2 à 7.

Devant le grand nombre d'occurrences par motif, nous avons été amenés à utiliser les techniques de regroupement introduites au chapitre précédent : le regroupement par composantes connexes et le regroupement par topologie. De plus, nous ferons la différence entre occurrences qui partagent la même topologie ET le même ordre de couleurs et les occurrences qui partagent la même topologie en ayant les couleurs placées différemment. Nous parlerons alors de motifs colorés, de motifs topologiques colorés ordonnés et de motifs topologiques non ordonnés.

Les Tableaux 6.2, 6.3 et 6.4 indiquent le nombre de motifs trouvés classés par taille de motif et nombre de paquets d'occurrences, pour respectivement, les motifs colorés, les motifs topologiques colorés non ordonnés et les motifs topologiques colorés ordonnés.

De ces trois tableaux, on peut dégager une tendance générale : plus on considère de grands motifs, moins on a de paquets d'occurrences. Ce résultat n'est pas surprenant puisque, si on considère de grands motifs, les occurrences ont plus de chances de se recouvrir. Par ailleurs, on retrouve le fait que les motifs colorés sont moins contraints que les motifs topologiques colorés non ordonnés qui sont moins contraints que les motifs topologiques colorés ordonnés. En effet, le nombre de motifs à plusieurs paquets diminue au fur et à mesure qu'on ajoute des contraintes.

On peut noter que les comptages diffèrent légèrement pour les motifs de taille 2 selon la définition de motif (ce qui peut paraître surprenant au

[2]On rappelle qu'un opéron est un groupe de gènes colocalisés sur le génome qui sont transcrits ensemble et produisent un unique ARN messager.

Taille du motif \ Nombre de paquets	1	2	3	4	5	6	10	12
2	418	98	22	10	9	2	1	1
3	2475	163	15	5	1			
4	13416	244	12	1				
5	73656	395	4					
6	403628	635	1					
7	2178094	1017						

TAB. 6.2 – Chaque cellule du tableau indique le nombre de motifs colorés de taille k ayant p paquets d'occurrences. Par exemple, il y a 5 motifs de taille 4 qui ont 3 paquets d'occurrences. Les cellules vides indiquent qu'il n'y a aucun motif de cette taille ayant autant de paquets d'occurrences.

Taille du motif \ Nombre de paquets	2	3	4	5	6	10	12
2	100	21	10	9	2	1	1
3	145	16	5				
4	215	7					
5	283	1					
6	415						
7	747						

TAB. 6.3 – Chaque cellule du tableau indique le nombre de motifs topologiques colorés non ordonnés de taille k ayant p paquets d'occurrences.

Taille du motif \ Nombre de paquets	2	3	4	5	6	10	12
2	100	21	10	9	2	1	1
3	136	12					
4	108	1					
5	82						
6	60						
7	41						

TAB. 6.4 – Chaque cellule du tableau indique le nombre de motifs topologiques colorés ordonnés de taille k ayant p paquets d'occurrences.

premier abord). Ceci est dû au fait que pour définir un motif topologique, on considère la topologie de l'occurrence dans le graphe biparti. Or dans le graphe biparti, des réactions peuvent être liées par un ou plusieurs composés. Ces cas sont différenciés.

Chacun des motifs dégagés par cette première phase pourrait faire l'objet d'une analyse plus approfondie. Nous l'avons faite pour certains d'entre eux. Nous avons à nouveau choisi de nous centrer sur les motifs qui avaient plusieurs paquets d'occurrences. De plus, le choix des motifs à étudier s'est fait selon un critère de maximalité. On a choisi des motifs qui, si on les étendaient, perdent leur propriété d'avoir autant de composantes connexes.

Plus précisément les critères qui ont présidé au choix des motifs à analyser ont été :

1. avoir un nombre de paquets d'occurrences maximal ;
2. en cas de choix, on privilégie les voies métaboliques connues ;
3. quand le choix est trop vaste, on choisit au hasard.

Pour chaque motif, l'analyse a consisté à poser les questions suivantes :

- est-ce que les réactions font partie d'une même voie métabolique ?
- est-ce que les gènes appartiennent à un même opéron ?
- est-ce que les enzymes des différentes occurrences sont homologues[3] ?

On note que pour les motifs approchés, on ne peut *a priori* pas se retrouver dans le cas où ce sont les mêmes enzymes qui catalysent les réactions des différentes occurrences (enzymes multifonctionnelles). Par contre, il peut être intéressant de poser la question : est-ce que par le passé, ces réactions étaient catalysées par une même enzyme ? Répondre à cette question n'est pas immédiat.

Le fait de détecter de l'homologie entre enzymes permet dans un premier temps de distinguer entre évolution convergente[4] et évolution divergente[5]. Dans le cas d'une évolution divergente, d'autres tests peuvent être mis en place pour tester s'il s'agit de néofonctionalisation (l'enzyme ancestrale n'avait qu'une fonction, une nouvelle fonction a été acquise au cours de l'évolution) ou de subfonctionalisation (chaque enzyme a évolué pour se spécialiser dans une sous-fonction de l'enzyme ancestrale).

De manière générale, on retiendra que détecter de l'homologie entre enzymes permet de formuler des hypothèses sur l'histoire évolutive des occurrences.

6.2.2.1 Motif approché de taille 3

Pour la taille 3, on peut observer que le nombre de paquets maximal atteint est de 5 pour les motifs colorés (voir Table 6.2), 4 pour les motifs

[3]Des enzymes homologues sont des enzymes qui ont un ancêtre commun.

[4]Les enzymes de ces voies n'ont pas d'ancêtre commun mais ont évolué vers des fonctions proches.

[5]Les enzymes ont un ancêtre commun et ont divergé depuis.

topologiques colorés non ordonnés (voir Table 6.3) et 3 pour les motifs topologiques colorés ordonnés (voir Table 6.4).

Nous allons ici détailler un motif topologique coloré ordonné à 3 paquets puis un motif topologique coloré non ordonné à 4 paquets. Ces deux motifs ont particulièrement attiré notre attention car, en plus de la condition de maximalité qu'ils remplissaient, certaines de leurs occurrences concernent des voies métaboliques très étudiées.

Parmi les motifs topologiques colorés qui ont 3 composantes connexes (il y en a 12), on trouve {5.3.1, 2.7.1, 4.1.2} qui est un motif commun à différentes voies de dégradation des sucres. Il nous a semblé particulièrement intéressant de mettre ce motif en avant car une des occurrences concerne la glycolyse[6], qui est sans doute la voie métabolique la plus étudiée dans la littérature.

Ainsi, le motif coloré {5.3.1, 2.7.1, 4.1.2} apparaît 9 fois dans le réseau métabolique d'*Escherichia coli*. Les 9 occurrences sont réparties en 3 paquets (un paquet de 7 occurrences et deux paquets d'une occurrence). Quand on introduit la topologie comme critère de classement, les 9 occurrences se répartissent en deux groupes de topologie (un groupe de 2 occurrences avec trois composés internes et un groupe de 7 occurrences avec deux composés internes ; toutes les occurrences sont linéaires). Si on rajoute l'ordre des couleurs comme contrainte, on obtient 3 groupes (le groupe de 7 est divisé en deux selon l'ordre d'apparition des couleurs). Finalement, le motif coloré peut donc être divisé en 3 motifs topologiques colorés ordonnés. Seul l'un d'entre eux garde la propriété d'avoir 3 paquets d'occurrences.

La Figure 6.3 montre l'ensemble des occurrences du motif coloré.

Le motif topologique coloré ordonné {5.3.1, 2.7.1, 4.1.2} (linéaire et dans cet ordre) a 4 occurrences réparties en trois paquets. C'est un motif commun à la dégradation du glucose (glycolyse), du mannose, du rhamnose et du fucose.

Il peut être étendu "en aval" d'au moins une réaction car toutes les occurrences produisent un composé en commun, le dihydroxy-acetone phosphate, mais le motif étendu perdrait alors la propriété d'avoir plusieurs paquets disjoints d'occurrences.

On note que le sous-motif {2.7.1 4.1.2} est commun à 7 voies métaboliques : la dégradation du glucose, du galactitol, du fucose, du rhamnose, du fucose, du D-galactonate et du D-galacturonate.

Ces motifs communs illustrent les fortes similarités qui existent entre les voies de dégradation des sucres. On note que, contrairement aux motifs exacts, ce ne sont ici pas les mêmes enzymes qui sont impliquées dans chacune des étapes (à l'exception de pfkB qui catalyse à la fois la dégradation du fructose-6-phosphate en fructose-1,6-biphosphate et la dégradation du tagatose-6-phosphate (2.7.1.11) en tagatose-1,6-biphosphate (2.7.1.144)).

Si les enzymes ne sont pas identiques, on peut se demander si certaines

[6]La glycolyse est la voie de dégradation du glucose en pyruvate. Au cours de ce processus, 2 moles d'ATP sont produites pour une mole de glucose.

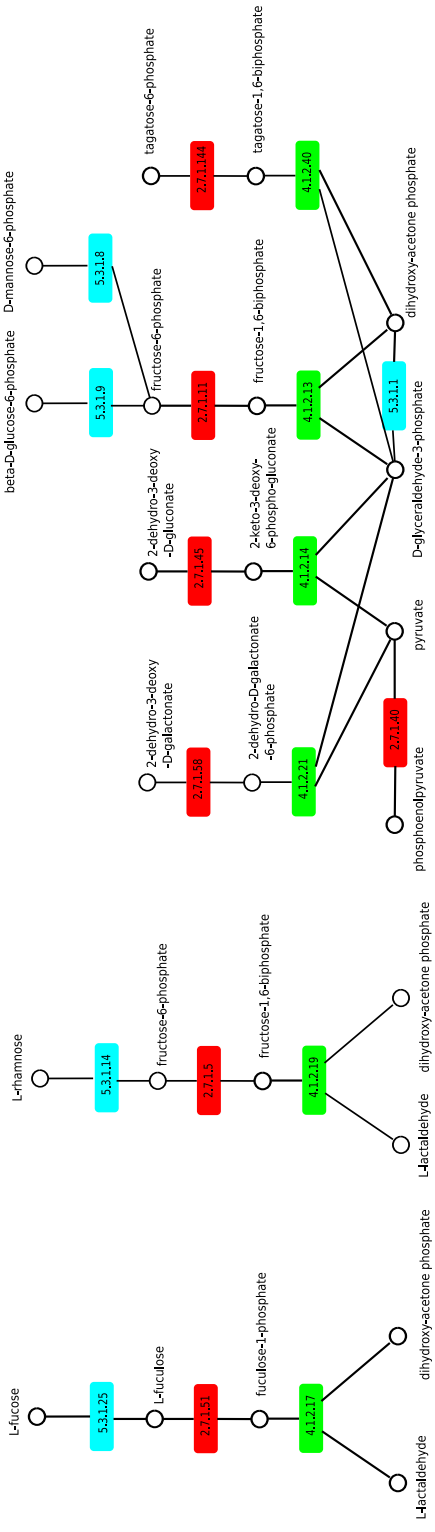


FIG. 6.3 – Occurrences du motif coloré {5.3.1, 2.7.1, 4.1.2}.

sont homologues (issues d'une même enzyme ancestrale). Nous avons donc recherché des cas d'homologie en comparant les séquences des protéines à l'aide de méthodes standards (BLAST contre la base de données SwissProt, en utilisant les paramètres par défaut, la E-valeur maximum considérée a été 10^{-5}).

Globalement, il y a peu de paires d'enzymes pour lesquelles on détecte de l'homologie. Ce résultat suggère une évolution convergente pour l'ensemble de ces voies (*i.e.* les enzymes de ces voies n'ont pas d'ancêtre commun mais ont évolué vers des fonctions proches). Cependant, si on considère les paires d'enzymes au cas par cas, on trouve les relations d'homologie suivante : kdgK (2.7.1.45) et pfkB (2.7.1.11, 2.7.1.144), dgoA (4.1.2.21) et eda (4.1.2.14) et enfin fbaA (4.1.2.13), gatY et kbaY (4.1.2.40). Localement, on trouve de l'homologie, ce qui suggère que les scénarios de recrutement des enzymes de ces voies métaboliques sont complexes et ne peuvent pas être expliquées par un seul mécanisme.

Enfin, on peut noter que dans les cas où on n'a pas détecté d'homologie, cela ne signifie pas que les enzymes ne sont pas homologues. En effet, on ne détecte plus de similarité de séquences entre des enzymes qui ont divergé il y a longtemps.

Le second motif de taille 3 qu'on se propose d'étudier (plus brièvement) est un motif topologique coloré non ordonné.

Il s'agit d'un des 5 motifs topologiques colorés non ordonnés qui ont 4 paquets d'occurrences (voir Table 6.3). Nous avons choisi d'analyser ce motif plus particulièrement pour illustrer la notion de motif topologique coloré non ordonné (les couleurs ne sont pas dans le même ordre selon les occurrences).

Il se trouve à nouveau que le motif est linéaire. Il est défini comme suit : {2.6.1, 1.1.1, 4.2.1}. Ce motif a 7 occurrences.

On peut noter que lorsqu'on recherche le motif coloré correspondant, on ne trouve pas d'occurrence supplémentaire. Ces couleurs ne sont donc jamais observées avec une autre topologie.

Les 7 occurrences sont regroupées en 4 paquets. Les trois premiers paquets correspondent à une partie des résultats présentés dans la Section 4.2 (le motif est un sous-motif du motif étudié précédemment). Ce sont trois étapes de la synthèse de la valine, l'isoleucine et la leucine. On note qu'il y a une inversion de couleur pour la leucine. On note qu'on a aussi une variante (voie alternative) pour la synthèse de la valine.

Le quatrième paquet d'occurrences concerne la synthèse de l'aspartate et la synthèse de l'alanine. On peut noter que les réactions de ces occurrences sont annotées dans EcoCyc comme faisant partie de plusieurs voies métaboliques : le cycle de Krebbs (1.1.1.37, 4.2.1.2), la gluconéogénèse (1.1.1.38, 1.1.1.40), la biosynthèse de l'alanine (2.6.1.2) et la biosynthèse de l'aspartate (2.6.1.1). Ce cas illustre le fait que la délimitation qui est classiquement faite entre voies métaboliques est limitée et que les frontières sont finalement arbitraires. Une partie du cycle de Krebbs peut donc ici être vue comme le début

de la synthèse d'un acide aminé.

Au final, on a donc 5 voies de synthèse d'acide aminé qui partagent ce motif. On note que pour la valine et l'isoleucine, l'ordre est {1.1.1, 4.2.1, 2.6.1} alors que pour la leucine, l'apartate et l'alanine, l'ordre est {4.2.1, 1.1.1, 2.6.1}. L'hydrolyase (4.2.1) et l'oxydoreductase (1.1.1) sont inversées. La fonction globale est pourtant la même.

La Figure 6.4 illustre ces occurrences.

6.2.2.2 Motif approché de taille 4

Comme motif de taille 4, nous avons choisi d'étudier le seul motif coloré de taille 4 qui a 4 paquets d'occurrences (voir Table 6.2). Il s'agit du motif {2.7.1, 2.7.4, 2.7.4, 2.7.7}. Ce motif a 7 occurrences, réparties en 4 paquets. Ce motif coloré peut être séparé en 3 motifs topologiques, deux linéaires (un avec 4 composés internes et l'autre avec 3) et un branché. Le motif linéaire a 3 occurrences qui sont réparties en 3 paquets (les occcurrences sont toutes disjointes).

On peut remarquer que pour cette exemple, l'ordre n'apporte aucune information supplémentaire (*i.e.* un seul ordre de couleur est observé par classe de topologie ; la succession est dans l'ordre suivant : {2.7.1, 2.7.4, 2.7.4, 2.7.7})

La Figure 6.5 illustre ces trois occurrences.

La thymine (T), l'uracil (U) et la cytosine (C) sont tous les trois des bases pyrimidiques. Avec les bases puriques, l'adénine (A) et la guanine (G), ils constituent les 5 bases azotées entrant dans la constitution des nucléotides, éléments de base de l'ADN et de l'ARN.

Les occurrences du motif considéré ici contiennent des réactions impliquées dans les voies de synthèse de ces nucléotides.

On constate ainsi les fortes similarités qui existent entre leurs voies de synthèse.

6.2.2.3 Motifs approchés de taille 5 et 6

Nous n'avons pas fait d'analyses approfondies pour les motifs de taille 5 et 6. Nous donnons ici quelques pistes initiales de motifs qu'il serait intéressant d'analyser à l'avenir.

À la taille 5, on a 4 motifs colorés qui ont 3 composantes connexes (voir Table 6.2). Il n'y en a qu'un parmi les motifs topologiques colorés non ordonnés. C'est le motif {2.4.2, 2.4.2, 2.7.1, 2.7.4, 3.5.4} qui est commun à plusieurs voies de synthèse des nucléotides.

À la taille 6, il n'y a plus qu'un motif coloré à 3 composantes connexes (voir Table 6.2). Ce motif est {1.1.1, 1.1.1, 2.6.1, 2.7.1, 2.7.1, 4.1.2}. Ce motif est commun à des processus branchés (*i.e.* non linéaires) autour du pyruvate, du fructose-phosphate et du glycerate.

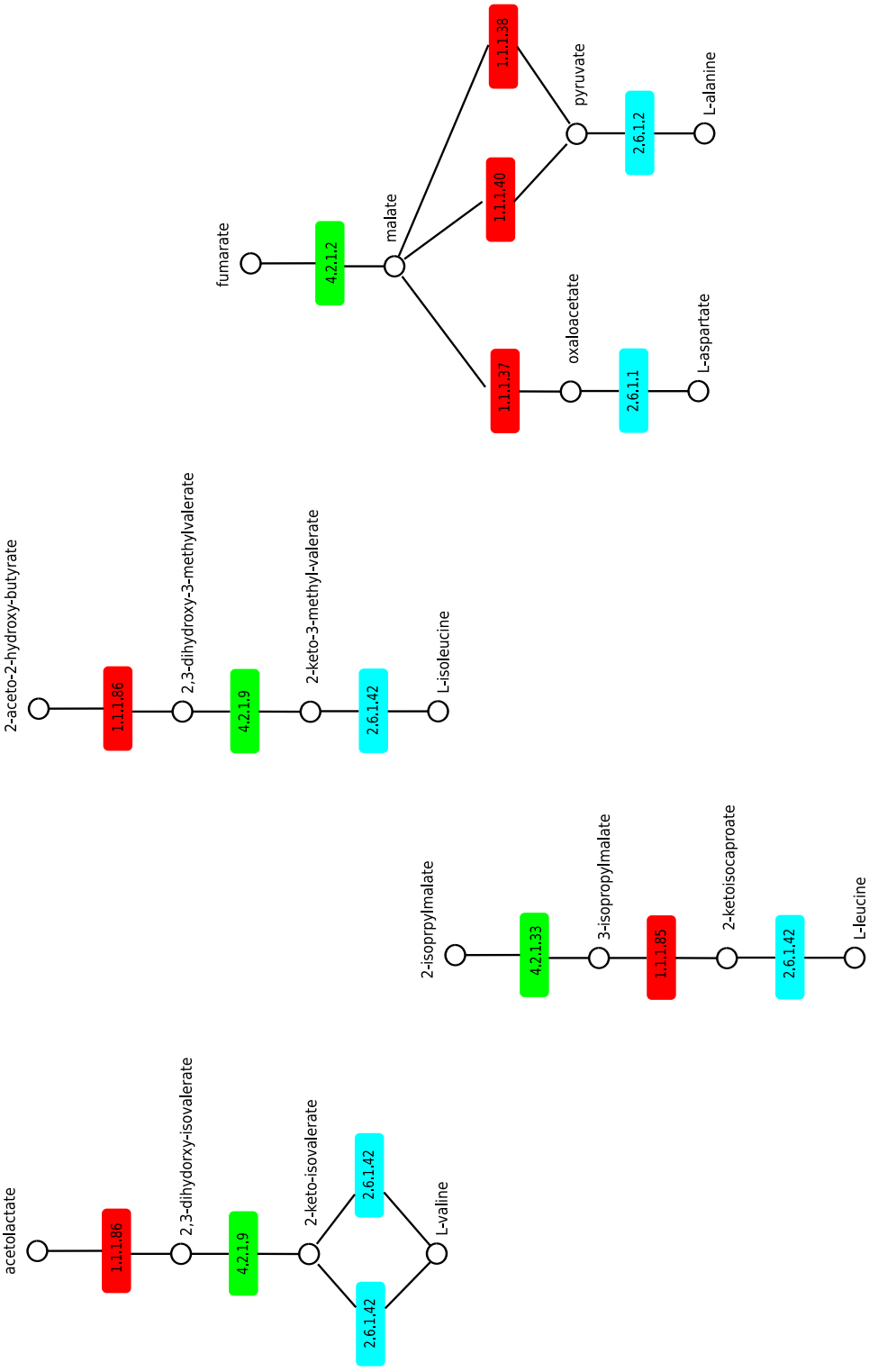


FIG. 6.4 – Occurrences du motif linéaire non ordonné {1.1.1, 4.2.1, 2.6.1}

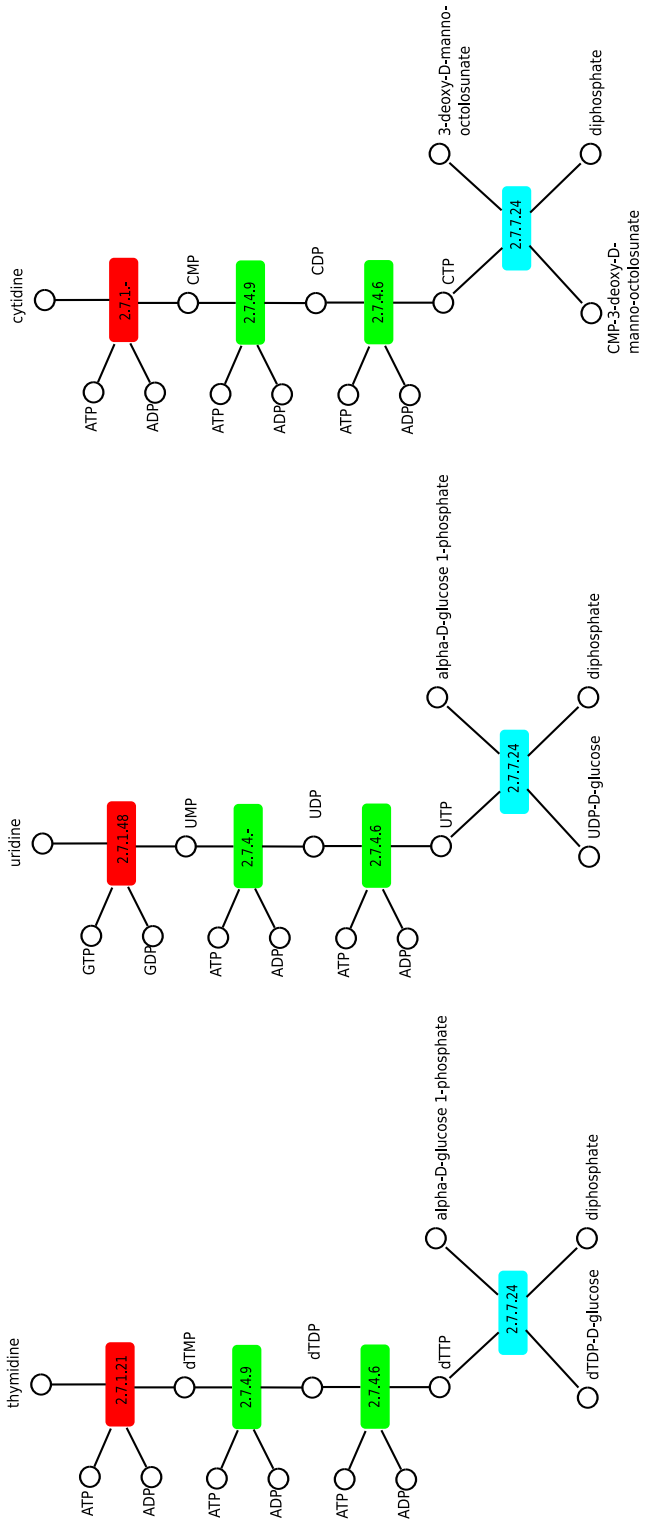


FIG. 6.5 – Occurrences du motif linéaire {2.7.1, 2.7.4, 2.7.4, 2.7.7}.

6.2.2.4 Motifs approchés de taille 7

À la taille 7, on n'a plus de motifs colorés qui aient 3 paquets d'occurrences (voir Table 6.2). Par contre, on trouve un nombre important de motifs à 2 paquets (1017 pour les motifs colorés, 747 pour les motifs topologiques colorés non ordonnés et 41 pour les motifs topologiques colorés ordonnés).

Nous avons choisi pour notre analyse un motif au hasard parmi les 41 motifs topologiques colorés ordonnés à deux paquets.

Le motif choisi a les couleurs suivantes : {2.3.3, 2.6.1, 2.6.1, 4.2.1, 5.1.1, 6.3.2, 6.3.2} et une topologie de forme branchée. La Figure 6.6 illustre les occurrences de ce motif.

Ce motif peut être vu comme composé de 3 blocs[7] : deux blocs linéaires connectés par une transaminase (2.6.1). La transaminase ne constitue pas la partie la plus intéressante du motif car c'est une classe d'enzyme très fréquente. En effet, il existe quasiment une transaminase pour chaque couple acide organique, acide aminé. Un rôle des transaminases est d'ailleurs d'équilibrer les concentrations entre acides aminés et acides organiques. Du point de vue topologique, on peut dire que les transaminases sont des sortes de hubs qui connectent un grand nombre de voies entre elles.

Ainsi, pour chaque occurrence, une transaminase connecte deux motifs linéaires : un qui part d'un acide aminé (L-alanine dans une des occurrences et L-glutamate dans l'autre) et un qui part d'un acide organique (2 keto-isovalerate dans une occurrence et oxaloacetate dans l'autre).

Le premier sous-motif, {2.3.3, 4.2.1}, est commun à la synthèse de la leucine et au cycle de Krebbs. On peut noter que la correspondance entre ces deux voies avait déjà été reportée par [Velasco *et al.*, 2002] alors qu'ils cherchaient à comprendre l'origine évolutive de la voie de synthèse de la lysine chez les champignons. En s'appuyant sur leurs résultats, on peut d'ailleurs suggérer une extension à ce motif. En effet, la synthèse de la leucine et le cycle de Krebbs partagent une étape supplémentaire : dans les deux cas, le produit final est transformé par une déhydrogénase (1.1.1) (voir Chapitre 3, Figure 3.2).

Le second sous-motif, {5.1.1, 6.3.2, 6.3.2}, est impliqué à deux reprises dans la voie de synthèse du peptidoglycane, une fois au début de la voie, une autre fois vers la fin[8]. Le peptidoglycane est un composant essentiel de la membrane cellulaire des bactéries Gram négatif. La synthèse du peptidoglycane est un processus complexe en deux parties (synthèse de l'unité de base puis polymérisation), seule la première est traitée dans EcoCyc.

D'un point de vue évolutif, on peut tester si ces voies ont une origine

[7]Cette analyse est le fruit d'une discussion avec Daniel Kahn, DR INRA spécialiste du métabolisme qui a rejoint le laboratoire BBE en 2006.

[8]Plus précisément, la réaction de transformation de la L-alanine en D-alanine (racémase 5.1.1.6) n'est pas classée dans la voie de synthèse du peptidoglycane, mais dans la voie de dégradation de l'alanine. Une fois de plus, on constate que les limites entre voies métaboliques sont discutables.

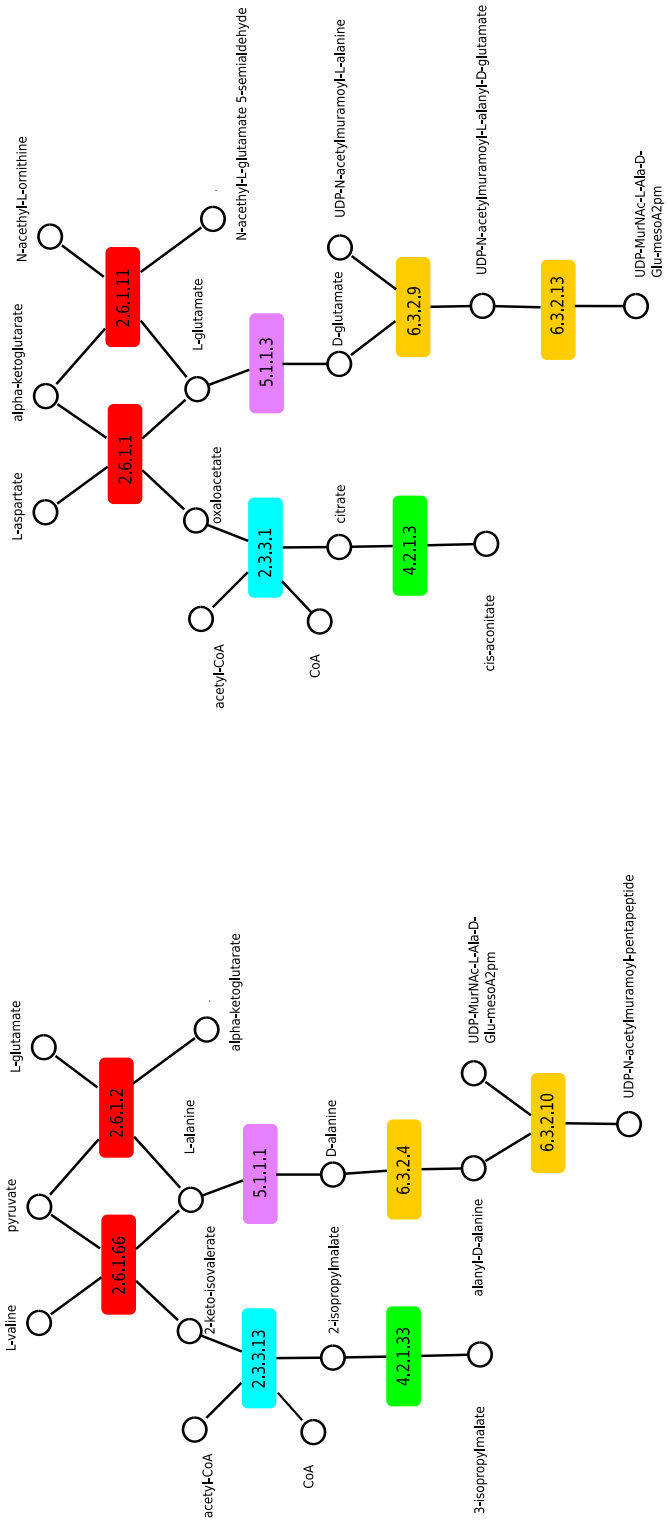


FIG. 6.6 – Occurrences du motif topologique coloré ordonné {2.3.3, 2.6.1, 2.6.1, 4.2.1, 5.1.1, 6.3.2, 6.3.2}

évolutive commune.

Concernant le premier sous-motif (synthèse de la leucine et cycle de Krebbs), on trouve de l'homologie entre acnA, acnB (4.2.1.3) et leuC (4.2.1.33) et aussi entre leuB (1.1.1.85) et icd (1.1.1.42) ; par contre, il n'y a pas d'homologie détectable entre gltA (2.3.3.1) et leuA (2.3.3.13). On peut ajouter que leuA, leuB et leuC appartiennent au même opéron.

Concernant le second sous-motif, entre les deux fragments de la biosynthèse du peptidoglycane, on détecte de l'homologie entre murD (6.3.2.9), murE (6.3.2.13) et murF (6.3.2.10) qui par ailleurs font partie du même opéron. Enfin, ddlA et ddlB sont homologues mais l'homologie n'est pas détectable entre ddlA/ddlB et murE/murD/murF. Cependant, murF et ddlA font partie du même opéron dans d'autres espèces (chez *Helicobacter pylori* par exemple). L'homologie entre murD, murE et murF est consistante avec une évolution rétrograde de la voie de biosynthèse du peptidoglycane. En effet, ces gènes catalysent trois étapes successives de la voie. Si on poursuit notre idée d'évolution rétrograde et qu'on suppose que l'ordre de ces gènes a été conservé au cours de l'évolution, on peut alors proposer un scénario de formation de l'opéron murD, murE, murF.

La non-détection d'homologie avec ddlA/ddlB et murD suggère un recrutement indépendant de cette enzyme. On peut faire la même remarque pour les racémases murI et alr/dadX. Au final, le motif peut être partiellement expliqué par une évolution divergente de murD, murE et murF, et une évolution convergente de ddlA/ddlB et murD d'une part, et alr/dadX et murI d'autre part.

On peut noter que les stuctures des protéines sont disponibles pour alr et murI. Nous avons donc cherché à aligner les structures de ces protéines[9] pour détecter plus finement une éventuelle homologie. Même avec cette méthode, nous n'avons pas pu détecter d'homologie. De même pour ddlA/ddlB et murD.

Enfin, si on adopte un point de vue plus fonctionnel sur ce motif de taille 7, on peut discuter la vraisemblance de l'enchaînement de certaines réactions. On rappelle que ce n'est pas parce que des réactions sont connectées dans le réseau qu'elles seront utilisées à la suite l'une de l'autre en condition normales. En effet, il est par exemple peu probable qu'en conditions normales, de la valine soit consommée pour que de la leucine soit produite (ces deux acides aminés sont généralement produits à partir du pyruvate). Cependant, même si cet enchaînement de réactions n'est pas observé (ou correspond à des flux très faibles) en conditions classiques, on peut imaginer que si on se place en condition d'excès de valine, alors cette voie (de la valine à la leucine) pourrait être empruntée.

[9]Pour aligner les structures de protéines, nous avons utilisé le programme Yakusa [Carpentier *et al.*, 2005] avec les paramètres par défaut (50 structures les plus proches), en utilisant successivement l'une puis l'autre structure comme requête.

6.2.3 Conclusion sur les motifs approchés et l'analyse systématique

Dans cette partie, nous avons établi une carte des motifs répétés. Nous nous sommes plus particulièrement intéressés aux motifs ayant plusieurs paquets d'occurrences.

Nous avons pu mettre en évidence que le nombre de motifs ayant plusieurs paquets d'occurrences augmente puis décroît quand la taille du motif augmente. Pour les motifs topologiques colorés ordonnés, la limite entre croissance et décroissance est obtenue à la taille 3. Pour les autres types de motifs, cette limite est après la taille 7.

Nous nous sommes ensuite interessés aux motifs qui avaient des propriétés de maximalité vis-à-vis de leur nombre de paquets d'occurrences. Nous avons alors choisi certains motifs que nous avons analysé un plus en détail.

De manière générale, leur analyse suggère que le concept de motif est pertinent dans le cadre des réseaux métaboliques. En effet, on peut trouver une interprétation biologique pour chaque motif trouvé.

On peut noter que les motifs étudiés étaient de petite taille. Même pour le motif de taille 7, on se rend compte lors de l'interprétation qu'il est préférable de le découper en 2 motifs de taille inférieure. Ces exemples suggèrent que l'organisation du métabolisme en motifs se fait à petite échelle.

Par ailleurs, la majorité des motifs que nous avons étudié sont en fait des motifs topologiques colorés (ordonnés ou non). La notion de motif coloré, plus large, contient le concept de motif topologique coloré, qui en est un raffinement. Ainsi, nous n'avons pas mis en évidence de motifs colorés qui ne soient pas des motifs topologiques colorés.

Ceci ne signifie pas que la notion de motif coloré n'est pas valide puisque nous n'avons pas étudié tous les exemples en détail.

Beaucoup de motifs restent inexplorés. Cette approche qui consiste à trouver des exemples à analyser plus en détails, a ses limites puisqu'on se focalise sur quelques motifs et qu'on ne montre pas de propriété générale sur tous les motifs. Les conclusions qu'on obtient ne sont que partielles. On ne peut donc pas tirer de conclusion générale sur les motifs. On peut simplement exhiber des cas particuliers qui marchent bien.

Cette étape d'exploration est cependant nécessaire et peut être vue comme une étape préliminaire visant à établir un protocole de validation des motifs. Ainsi, au cours de l'étude approfondie de quelques motifs, certains paramètres ont retenu notre attention.

Au cours de l'étude de ces exemples, nous avons en effet pu constater que certains facteurs étaient intéressants à regarder pour aller plus loin dans l'analyse du motif et de son évolution :

- Est-ce que les enzymes sont homologues ?
- Est-ce que les gènes qui codent pour ces enzymes sont proches sur le génome ?

Dans la section suivante, nous allons adopter une approche différente

puisque nous allons nous centrer sur une propriété et étudier tous les motifs à la lumière de cette propriété. En ce sens, nous allons pouvoir établir des propriétés générales à tous les motifs, qui ne sont pas spécifiques à certains exemples.

6.3 Comparaison entre motifs et opérons

Nous avons pu voir dans la section précédente que l'organisation des gènes sur le génome pouvait être un indice que le motif étudié correspondait bien à une unité fonctionnelle. C'est le cas des gènes impliqués dans la synthèse de la valine et de l'isoleucine qui sont situés proches les uns des autres sur le génome. C'est aussi le cas de murD, murE et murF (gènes de la synthèse du peptidoglycane) qui forment un opéron.

Comme précisé dans la section 2.4.5.3, plusieurs travaux se sont déjà interessés au lien entre organisation génomique et structure du réseau métabolique. On peut notamment mentionner les travaux de [Rison *et al.*, 2002] et ceux de [Boyer *et al.*, 2005]. On peut noter qu'aucun de ces travaux ne prend en compte la notion de répétition.

Des travaux de [Rison *et al.*, 2002], on peut dégager un résultat marquant : il existe une corrélation positive entre la distance entre les gènes dans le réseau et la distance entre les gènes sur le génome. Les auteurs ajoutent que cette corrélation est principalement valable à faible distance et qu'elle s'explique quasiment intégralement par les structures connues d'opérons.

Pour la suite, on peut donc retenir que les gènes qui sont proches dans le réseau métabolique ont plus tendance à faire partie du même opéron.

Ce résultat conforte l'idée que le fait d'être connecté dans le réseau augmente les chances de faire partie d'une unité fonctionnelle (dans ce cas un opéron). Mais nous avons voulu savoir si, en plus d'être connecté dans le réseau, le fait d'appartenir à un motif sur-représenté, augmentait les chances de faire partie d'une unité fonctionnelle.

Nous nous sommes donc posé la question suivante :

Parmi les gènes connectés dans le réseau, est-ce que ceux qui participent à des motifs répétés font plus (ou moins) partie d'opérons que les autres ?

Pour répondre à cette question, le protocole que nous avons suivi est le suivant : 1. nous avons récupéré l'ensemble des opérons connus chez *Escherchia coli* à partir de la base RegulonDb [Salgado *et al.*, 2006], puis 2. nous avons inféré tous les motifs de taille k et de seuil s du réseau métabolique d'*E. coli*, et 3. pour chaque occurrence de chaque motif, nous avons déterminé si l'occurrence était couverte par un opéron.

Pour simplifier, on peut dire dans un premier temps qu'une occurrence est couverte par un opéron si tous les gènes impliqués appartiennent à un même opéron. Avant de donner une définition plus précise, il est sans doute utile

de rappeler que le lien entre gène et réaction peut parfois être complexe. La Figure 6.7, issue de [Reed *et al.*, 2003], donne une idée de cette complexité.

On constate donc qu'il n'est pas toujours nécessaire que tous les gènes soient présents pour que la réaction puisse avoir lieu (par exemple dans la troisième situation).

Nous dirons donc qu'une occurrence est couverte par un opéron si toutes ses réactions sont couvertes par un opéron. Une réaction est couverte par un opéron si au moins une des enzymes catalysant cette réaction est couverte. Enfin, une enzyme est couverte si tous les gènes codant pour ses sous-unités sont couverts.

À l'issue de ces opérations, nous avons séparé les occurrences, d'une part entre occurrences couvertes par un opéron et occurrences non couvertes, et d'autre part entre occurrences de motifs répétés et occurrences de motifs non répétés. On rappelle qu'une occurrence ne peut être assignée à plusieurs motifs (à part dans le cas où le graphe contient des noeuds multicolores, ce qui n'est pas le cas ici).

Le résultat peut être représenté par une table de contingence 2×2 comme suit (résultats obtenus pour des motifs de taille 2, seuil 3) :

	opéron +	opéron -	
motif répété	85	1006	1091
motif non répété	37	251	288
	122	1257	1379

À partir de ces résultats nous allons répondre à notre question en 3 étapes :

1. Est-ce qu'il y a un lien entre opéron et motif répété ?
2. Quelle est le sens de ce lien ?
3. Quelle est l'intensité de ce lien ?

Pour tester l'indépendance des deux caractères (opéron et motif répété), un test de χ^2 d'indépendance paraît adapté mais n'est en réalité pas applicable dans ce cas car une condition d'application de ce test n'est pas remplie : les individus statistiques ne sont pas indépendants. En effet, les individus statistiques sont ici des occurrences (ensemble de noeuds) et les occurrences peuvent partager des noeuds (et donc les probabilités pour deux occurrences d'être chacune couverte par un opéron ne sont pas indépendantes).

Nous avons donc utilisé un test à base de permutations. L'hypothèse nulle à tester, qu'on note H_0 est, dans le cas d'un test bilatéral : il n'y a pas d'association entre appartenir à un opéron et appartenir à un motif à plusieurs paquets. Pour tester cette hypothèse, la statistique que nous étudions est le nombre d'occurrences couvertes parmi les occurrences de motifs répétés (la valeur observée de cette statistique est égale à 85 dans l'exemple précédent). Pour approcher la distribution de cette statistique sous l'hypothèse H_0, nous

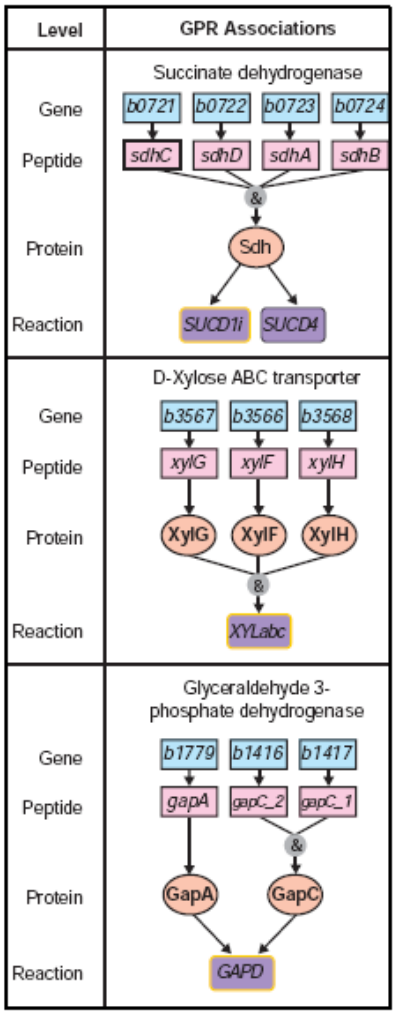


FIG. 6.7 – Représentation des associations gène-protéine-réaction. Dans la première situation, 4 gènes sont nécessaires à la formation d'une protéine qui catalyse seule deux réactions. Dans la seconde situation, trois gènes sont nécessaires à la formation de trois protéines qui ensemble forment un complexe enzymatique qui catalyse une réaction. Enfin, dans la troisième situation, trois gènes forment deux protéines qui sont chacune capables de catalyser la même réaction.

procédons par simulations en permutant aléatoirement les opérons un grand nombre de fois (en pratique 10000 permutations). La valeur observée (85 dans l'exemple proposé) est ensuite comparée à la distribution simulée (voir Figure 6.8).

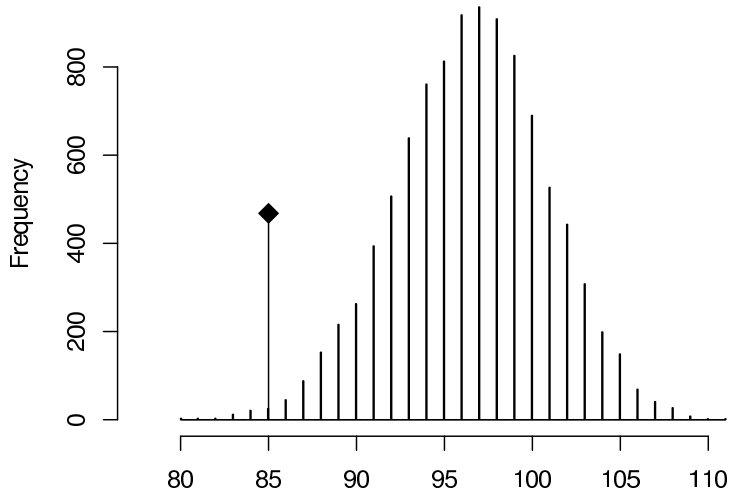


FIG. 6.8 – Distribution simulée et valeur observée.

Dans l'exemple précédent, la p-valeur estimée est de 0.01. Si on se fixe un risque α de 5% (risque de rejeter l'hypothèse H_0 alors qu'elle est vraie), alors on rejette H_0 et on conclut qu'il y a une association entre appartenir à un opéron et appartenir à un motif répété.

On note que le test qu'on a réalisé est un test bilatéral qui ne permet pas de conclure sur le sens de cette association. En observant la Figure 6.8, on peut cependant constater que la valeur observée 85 est significativement inférieure à la valeur attendue.

On peut donc conclure que le fait d'appartenir à un motif répété diminue les chances d'appartenir à un opéron.

Ce résultat semble tout d'abord surprenant mais il peut s'expliquer de la manière suivante : un motif répété peut avoir beaucoup d'occurrences et parmi ces occurrences, certaines sont couvertes par des opérons et certaines ne le sont pas. Le résultat qu'on obtient semble indiquer que les occurrences couvertes (le signal) sont "noyées" parmi celles qui ne le sont pas (bruit). La notion de répétition du motif ne semble donc pas permettre de dégager des occurrences fonctionnelles (au sens où les gènes sont regroupés dans le même opéron).

Pour aller plus loin, on peut quantifier le lien qu'on a mis en évidence. Dans ce cas, on peut utiliser des mesures inspirées des notions de sensibilité et de spécificité. Si on considère que "ne pas appartenir à un motif répété" est un prédicteur pour "appartenir à un opéron", on peut par exemple quantifier la proportion d'occurrences couvertes par un opéron qui sont prédites. Cette proportion est de 37/122=30.3%. On peut également quantifier la pro-

portion d'occurrences qui sont effectivement couvertes par un opéron, parmi celles qui sont prédites. Cette proportion est de 37/288=12.8%. On constate donc qu'avec 30.3% de sensibilité, il reste un nombre important d'opérons qui ne sont pas prédits.

Dans un deuxième temps, nous avons voulu savoir si on obtenait le même résultat en considérant, non pas le nombre d'occurrences, mais le nombre de paquets d'occurrences.

Nous avons donc posé la question suivante :

Parmi les gènes connectés dans le réseau, est-ce que ceux qui participent à des motifs répétés en plusieurs paquets font plus (ou moins) partie d'opérons que les autres ?

Les résultats sont consignés dans la table de contingence suivante :

	opéron +	opéron -	
plusieurs paquets	77	612	689
un seul paquet	45	645	690
	122	1257	1379

Le test d'indépendance conclut qu'il existe une association entre le fait d'appartenir à un motif à plusieurs paquets et le fait d'être couvert par un opéron. La p-valeur estimée est de 0.003. La Figure 6.9 montre la position de la valeur observée par rapport à la distribution simulée.

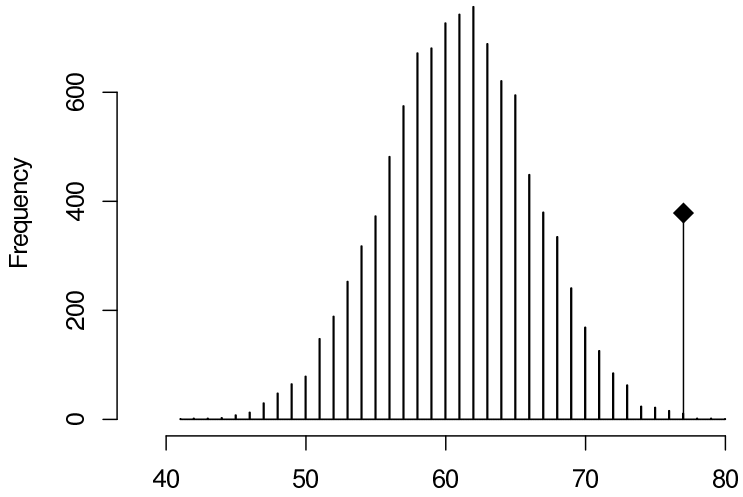


FIG. 6.9 – Distribution simulée et valeur observée.

Cette fois, le lien est dans le sens différent. La valeur observée (77) est supérieure à la valeur attendue. On peut donc conclure que le fait d'appartenir à un motif qui a plusieurs paquets d'occurrences augmente les chances pour une occurrence d'être couverte par un opéron.

De même que précédemment, on peut quantifier ce lien. Si on considère que “appartenir à un motif à plusieurs paquets” est un prédicteur pour “appartenir à un opéron”, alors la sensibilité est de 77/122=63.1% et la spécificité est de 77/689=11.2%.

La notion de répétition en plusieurs paquets du motif semble donc permettre de dégager des occurrences fonctionnelles (au sens où les gènes sont regroupés dans le même opéron).

Dans un troisième temps, nous avons voulu tester si la notion de sur-représentation nous permettrait de tirer des conclusions différentes. Nous avons donc reconduit deux séries de tests pour les deux mesures de sur-représentation dont nous disposons (l'une sur le nombre d'occurrences, l'autre sur le nombre de paquets d'occurrences). Les conclusions auxquelles nous parvenons suivent les mêmes tendances. Ne pas appartenir à un motif sur-représenté (au sens du nombre d'occurrences) augmente les chances d'appartenir à un opéron (p=0.003, sensibilité : 59%, spécificité : 11.5%). Appartenir à un motif sur-représenté (au sens du nombre de paquets) augmente les chances d'appartenir à un opéron (p=0.03, sensibilité : 29.5%, spécificité : 11.8%).

Nous avons ensuite voulu combiner ces tests et nous avons posé des questions du type : parmi les occurrences de motifs à plusieurs paquets, est-ce que celles qui sont des occurrences de motifs qui ne sont pas sur-représentés ont plus de chances de faire partie d'opérons ? La réponse à cette question a été non (p=0.28). De manière générale, les réponses ont été négatives pour ce type de tests.

Enfin, nous avons refait ces tests pour des valeurs différentes de k et s. Nous avons en particulier testé $k = 2, 3, 4$ et $s = 2, 3$. De manière générale, les résultats que nous avons présenté précédemment ont été confirmés. On note que lorsqu'on considère des motifs de taille supérieure à 4, très peu d'occurrences sont complètement couvertes par des opérons. Pour pouvoir travailler avec des tailles de motifs plus grandes, il nous faudrait revoir notre définition d'occurrence couverte en considérant non pas si une occurrence est couverte ou pas, mais quelle est la proportion de ses réactions qui sont couvertes.

Pour finir, nous avons également refait une partie des tests pour les motifs topologiques colorés (ordonnés ou non ordonnés). Nous avons en particulier testé l'influence du nombre d'occurrences et du nombre de paquets d'occurrences pour les motifs de taille 3 et 4. On note que pour les motifs topologiques colorés non ordonnés, on obtient des résultats similaires aux motifs colorés alors que pour les motifs topologiques colorés ordonnés, l'une des conclusions principales change. Pour ce type de motifs, appartenir à un motif à plusieurs paquets n'influence pas le fait d'appartenir à un opéron (p=0.11).

Pour récapituler, ces tests nous ont permis d'évaluer le lien entre le fait

d'appartenir à un motif (répété ou non) et le fait d'appartenir à un opéron. Il apparaît que parmi les gènes qui sont connectés dans le réseau :

1. Ceux qui appartiennent à des motifs répétés (ou sur-représentés) font moins partie d'opérons que les autres.
2. Ceux qui appartiennent à des motifs répétés en plusieurs paquets (ou sur-représentés au sens du nombre de paquets) font plus partie d'opérons que les autres.

On peut donc avancer que les gènes qui appartiennent à des opérons font soit partie d'occurrences uniques, soit partie d'occurrences de motifs répétés en plusieurs paquets.

Concernant l'évaluation de notre définition de motif, on pourra retenir que, en ce qui concerne les motifs colorés, le nombre d'occurrences du motif ne semble pas être une mesure qui permette de dégager des informations fonctionnelles. Les motifs qui ont beaucoup d'occurrences (réunies en un seul paquet) semblent contenir beaucoup de bruit (*i.e.* beaucoup d'occurrences qui ne sont pas couvertes).

Par contre, lorsqu'on considère non plus le nombre d'occurrences, mais le nombre de paquets d'occurrences, le concept de motif coloré permet de dégager des occurrences fonctionnelles (dans le sens où elles correspondent à des opérons).

Enfin, si on compare les concepts de motif coloré, motif topologique coloré non ordonné et motif topologique coloré ordonné, on peut constater que le dernier, plus contraint, ne permet plus de dégager clairement d'occurrences fonctionnelles.

Finalement, si on revient aux conclusions de [Rison *et al.*, 2002], on constate que nos résultats permettent de les affiner dans plusieurs directions. Ainsi, parmi les gènes qui sont connectés dans le réseau, ce sont ceux qui appartiennent à des motifs non répétés ou à des motifs répétés en plusieurs paquets qui ont plus de chances d'appartenir à des opérons.

6.4 Lien entre topologie des réseaux métaboliques et expression des enzymes

Enfin, on peut signaler que notre algorithme a été utilisé par Patricia Thébault lors de son post-doc dans l'équipe dans le cadre de l'analyse de données d'expression (Patricia Thébault a effectué un post-doc de deux ans dans l'équipe Baobab et a été recrutée en 2007 comme Maître de Conférences à l'Université de Bordeaux 2.)

Ce travail est en cours et fera l'objet d'une publication dans les mois qui viennent. Nous présentons ici les résultats principaux.

Les données analysées correspondent à des données d'expression collectées dans diverses conditions chez la levure *Saccharomyces cerevisiae*. Les données

ont été collectées à partir du site SGD (Saccharomyces Genome Databank, (`http://www.yeastgenome.org/`)

De manière générale, cette analyse s'inscrit dans la question de l'exploration du lien entre structure du réseau métabolique et expression des enzymes.

Plus concrètement, la question que nous nous sommes posée est :

Est-ce que les enzymes qui sont connectées dans le réseau métabolique ont plus de chances d'être exprimées de manière coordonnée ?

Pour répondre à cette question, nous avons développé une approche similiaire à celle présentée dans la section précédente. À l'aide de MOTUS, nous avons identifié tous les ensembles d'enzymes connectés dans le réseau. Nous avons séparé cet ensemble en deux sous-ensembles, selon que les enzymes avaient une expression coordonnée ou non. Nous avons considéré qu'un ensemble de k enzymes avaient une expression coordonnée s'ils étaient tous sur-exprimés ou tous sous-exprimés dans les conditions analysées. À l'aide d'un test à base de permutations, nous avons alors testé notre hypothèse, pour différentes valeurs de k.

Il apparaît que, pour toutes les valeurs de k testées, la réponse à notre question est oui. Les enzymes qui sont connectées dans le réseau métabolique ont plus de chances d'avoir une expression coordonnée.

Par la suite, nous avons également testé si ces enzymes qui avaient une expression coordonnée avaient plus tendance à partager un facteur de transcription. Les résultats sont en cours d'analyse.

On peut noter que dans ce travail, MOTUS n'a pas été utilisé pour rechercher des motifs mais pour une tâche plus simple : identifier dans un graphe tous les sous-graphes connexes d'une certaine taille.

Une question que l'on pourrait se poser à l'avenir est : parmi les enzymes qui sont connectées, est-ce que les enzymes qui font partie de motifs répétés en plusieurs paquets ont plus de chances d'être exprimées de manière coordonnée ?

Chapitre 7

MOTUS

Ce chapitre présente l'outil MOTUS qui regroupe toutes les définitions présentées dans les chapitres précédents. En effet, afin de tester les définitions de motifs, il nous a semblé utile d'avoir une plateforme qui, en outre, peut être mise à disposition de la communauté.

Le travail présenté dans cette thèse est exploratoire par nature. En effet, nous avons proposé une nouvelle définition de motif pour laquelle nous avons développé des méthodes algorithmiques et statistiques de comptage. Nous avons appliqué ces méthodes au réseau métabolique d'*Escherichia coli*. Cela nous a permis à la fois de questionner/affiner notre méthode et de dégager des résultats intéressants du point de vue biologique. L'aller-retour entre développement de méthodes et applications à des cas concrets nous a semblé particulièrement fécond. C'est un processus itératif qui peut être poursuivi, à la fois pour affiner les méthodes, mais aussi parallèlement pour affiner les questions biologiques auxquelles les méthodes permettent de répondre.

Un moyen efficace pour que les méthodes proposées dans cette thèse soient utilisées et critiquées est de les diffuser. C'est pourquoi nous avons attaché une grande importance au développement d'un logiciel, MOTUS, disponible via une interface web ou en ligne de commande.

Cet outil a été développé avec l'aide de Ludovic Cottret (doctorant dans l'équipe Baobab) pour la première version de l'interface web et le traitement des données, de Fabien Jourdan (chercheur INRA au laboratoire de Xénobiotiques de Toulouse, UMR 1089 INRA-ENVT) pour le travail de visualisation des occurrences, et enfin d'Odile Rogier (ingénieur au PRABI -Pôle Rhône-Alpin de BIoinformatique) pour la version actuelle de l'interface web, la gestion des applets et de manière générale la coordination des différentes fonctionnalités.

L'interface web est accessible à l'adresse suivante : `http://pbil.univ-lyon1.fr/software/motus/`

L'outil n'étant pas encore publié, l'accès est restreint aux collaborateurs proches. Cette paire compte/mot de passe :
Compte : baobab
Mot de passe : baobab

permet d'y accéder.

Ce dernier chapitre récapitule les fonctionnalités disponibles dans MOTUS. Plusieurs options sont disponibles à l'utilisation. On peut les découper en 4 sous-rubriques : données, recherche, inférence et dessin.

7.1 Traitement des données

Plusieurs jeux de données sont disponibles à l'analyse, ainsi que plusieurs méthodes de pré-traitement des données.

Les données proviennent essentiellement de la base BioCyc (http ://biocyc.org) où sont disponibles au téléchargement les reconstructions de réseaux métaboliques de différents organismes. En pratique, l'information téléchargeable est sous forme de liste de réactions (avec pour chaque réaction les composés et les enzymes qui lui sont associées) à partir de laquelle on peut facilement reconstuire un graphe. Les jeux de données disponibles sont classés par niveau de qualité de données (correspondant à différents efforts de réannotation manuelle après reconstruction automatique). Au premier niveau (qualité maximale, au moins un an de réannotation manuelle), on trouve le réseau d'*Escherichia coli*. Au second niveau (qualité moyenne : données inférées avec entre 1 et 4 mois d'effort de réannotation manuelle), on trouve les réseaux de 13 organismes. Il s'agit essentiellement de procaryotes (*Agrobacterium tumefaciens*, *Helicobacter pylori*) à l'exception de *Homo sapiens*. Au troisième niveau (données inférées, aucune réannotation manuelle), on trouve 242 organismes. Seules les 2 premières catégories sont disponibles à l'analyse.

Pour chaque jeu de données en provenance de BioCyc, plusieurs prétraitements sont disponibles :

- retrait de composés ubiquitaires ;
- retrait de composés primaires ;
- retrait de réactions.

Pour constater l'impact du prétraitement sur les données brutes, on peut consulter les statistiques descriptives qui donnent le nombre de réactions, d'enzymes et de composés restant après prétraitement.

Accessible uniquement en ligne de commande : On peut également charger un graphe coloré soi-même et donc potentiellement avec des applications complètement différentes.

7.2 Mode recherche

MOTUS est disponible en deux modes : recherche et inférence. Le mode recherche correspond à la situation où on connaît le motif qu'on recherche. Si on n'a pas d'idée *a priori* sur un motif à rechercher, le mieux est de commencer avec le mode inférence.

En mode recherche, on doit spécifier les couleurs (qu'on peut choisir dans une liste de couleurs possibles, relativement au jeu de données sélectionné).

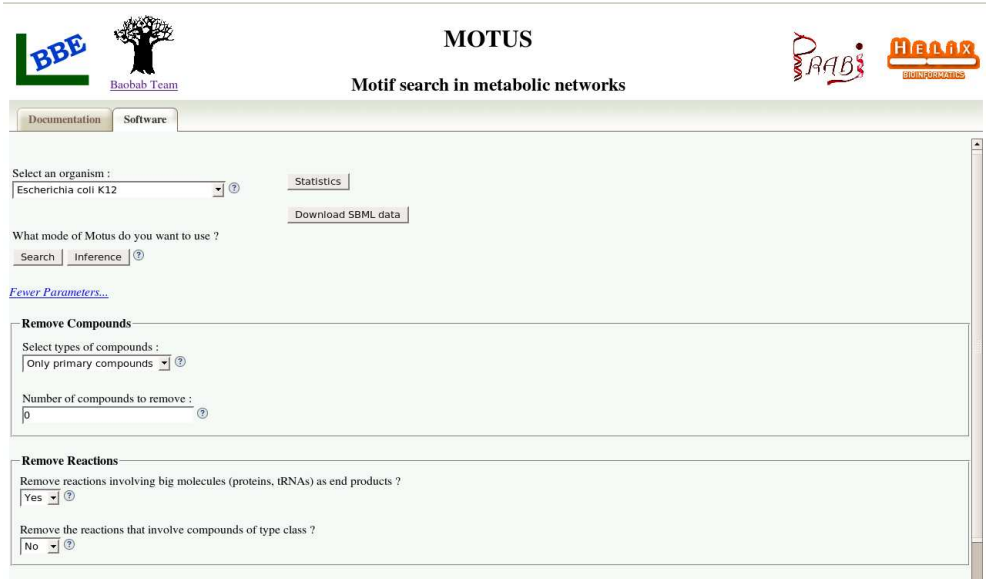


FIG. 7.1 – MOTUS : Sélection du jeu de données.

Le seuil de comparaison des couleurs est défini de manière implicite en donnant un numéro EC incomplet. On note qu'il est possible de sélectionner des seuils de comparaison de couleurs différents selon les couleurs.

On doit également spécifier le nombre de simulations à faire pour estimer la p-valeur.

Dans les résultats, le regroupement par paquet est disponible par défaut. Les occurrences sont regroupées par paquets et chaque paquet est séparé par un trait épais.

Pour chaque occurrence intra-voie (*i.e.* toutes les réactions sont dans la même voie métabolique), on affiche le nom de la voie métabolique concernée.

En outre, les noms des réactions et des voies métaboliques sont cliquables et mènent à la page correspondante de BioCyc. On peut ainsi avoir des informations complémentaires (enzyme, gène, régulation, position sur le génome). ACCESSIBLE UNIQUEMENT EN LIGNE DE COMMANDE : Regroupement par topologie. Intervalle de confiance sur la p-valeur.

7.3 Mode inférence

Pour le mode inférence, l'utilisateur doit spécifier la taille des motifs à extraire ainsi que la granularité des motifs (seuils de comparaison des couleurs).

Pour chaque motif, on dispose de :

– son nombre d'occurrences et la p-valeur correspondante ;
– son nombre de paquets d'occurrences et la p-valeur correspondante ;

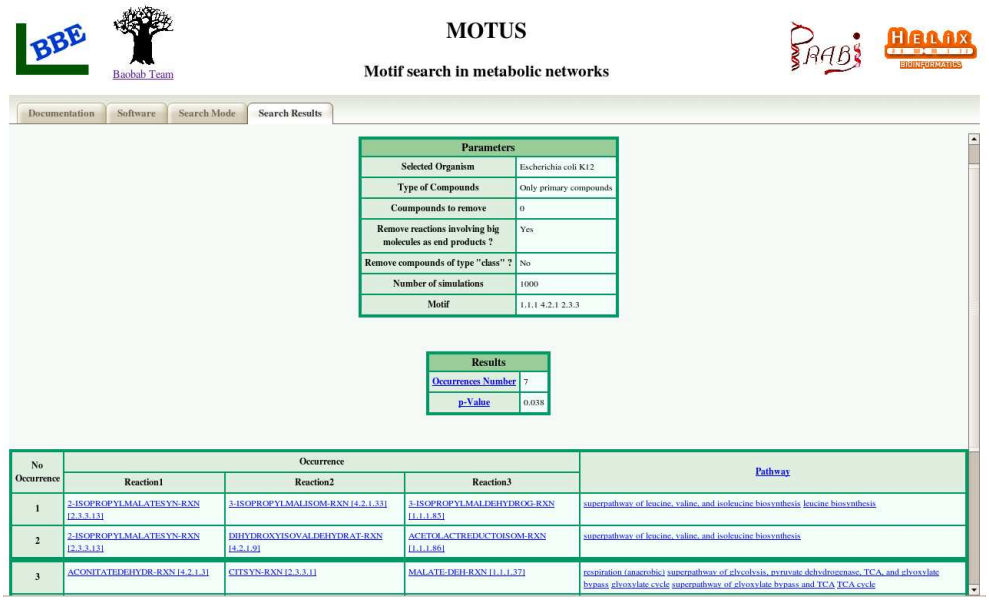


FIG. 7.2 – MOTUS : Mode recherche.

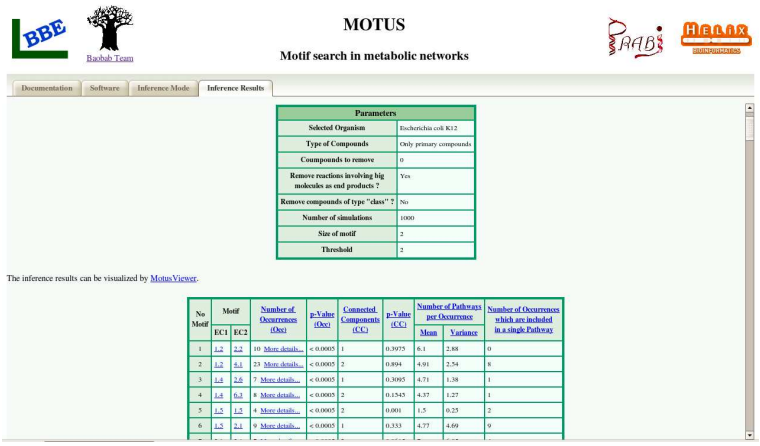


FIG. 7.3 – MOTUS : Mode inférence.

– le nombre de voies métaboliques couvertes par occurrence (moyenne, variance) et le nombre d'occurrences intra-voie.

Les résultats sont classés par p-valeur croissante.

Il est important de noter que les temps d'exécution du programme ne sont pas les mêmes pour l'interface web et pour la version en ligne de commande. En effet, pour l'interface web, le temps d'exécution est fortement dépendant de l'utilisation éventuelle des ressources du serveur par d'autres programmes.

On note que pour gagner du temps, certains fichiers de résultats sont conservés sur le serveur. Ainsi, si la requête qu'on veut faire a déjà été faite par un autre utilisateur, on obtient directement le résultat.

Lorsqu'on veut extraire des motifs de taille supérieure à 4, la version en ligne de commande est vivement recommandée.

7.4 Mode visualisation

Pour faciliter la navigation dans la liste des motifs inférés, Fabien Jourdan (laboratoire de Xénobiotiques de Toulouse) a développé un outil de visualisation, MOTUSViewer, qui permet de dessiner, pour chaque motif, un graphe représentant le lien entre occurrences et voies métaboliques (voir Fig. 7.4).

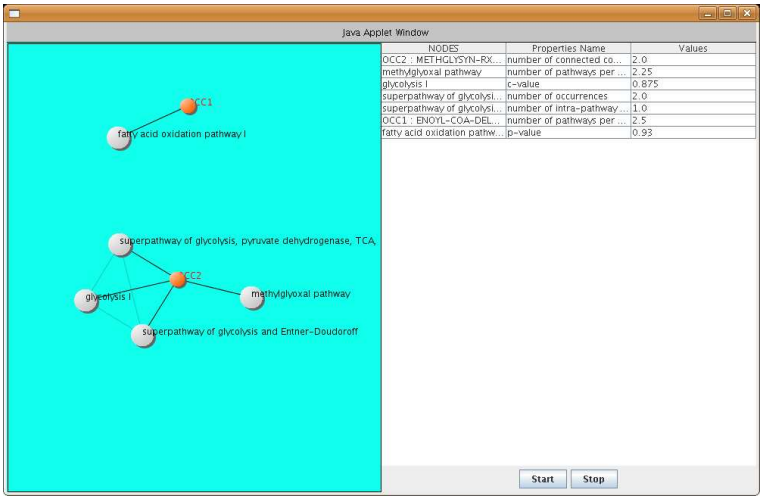


FIG. 7.4 – Graphe des voies et des occurrences.

Ce graphe a deux types de noeuds représentant respectivement les occurrences (noeuds rouges) et les voies métaboliques (noeuds noirs). Un recouvrement entre occurrences est visualisé par une arête rouge entre noeuds rouges. Un recouvrement entre voies métaboliques est symbolisé par une arête grise entre noeuds noirs. Enfin, l'appartenance d'une occurrence à une voie métabolique est symbolisée par une arête noire entre un noeud rouge et un noeud noir. On note que la longueur de cette arête est proportionnelle à

l'inclusion de l'occurrence dans la voie (toutes les réactions de l'occurrence ne sont pas nécessairement inclues dans la voie).

En mode inférence, le nombre de motifs extraits est souvent très important. MOTUSViewer permet de visualiser un grand nombre de motifs à la fois (voir Fig. 7.5). En plus de dessiner les graphes indiquant le lien entre occurrences et voies métaboliques, le programme colore chaque motif selon sa p-valeur. Une p-valeur faible (sur-représentation) est symbolisée par la couleur jaune et une p-valeur forte (sous-représentation) par la couleur bleue.

FIG. 7.5 – MOTUSViewer : un outil de visualisation pour explorer les résultats de l'inférence de motifs.

On peut noter que cet outil de visualisation est pour l'instant limité à de petits motifs.

Conclusions et perspectives

Conclusions

Au cours de cette thèse, nous avons donc proposé une nouvelle définition de motif dans le cadre des réseaux métaboliques.

Dans notre approche, un réseau métabolique est modélisé par un graphe coloré et un motif est défini comme un ensemble de couleurs. Une occurrence d'un motif correspond à un ensemble de noeuds connectés et colorés par les couleurs du motif. La recherche de motifs colorés constitue en fait un problème original en théorie des graphes. Nous avons donc caractérisé sa complexité et proposé un algorithme pour le résoudre. Une première application de cet algorithme à la recherche de voies similaires à la synthèse de la valine nous a montré que notre approche permettait de dégager des structures pertinentes. Une des limites de cette méthode de recherche de motif est qu'il faut savoir à l'avance quel est le motif qu'on recherche.

Nous avons alors décidé de passer au problème d'inférence de motifs qui est une généralisation du problème de recherche. Dans un problème d'inférence, le motif recherché n'est pas connu, seules certaines caractéristiques sont précisées (dans notre cas la taille et le seuil de similarité des couleurs). Nous avons ainsi proposé un algorithme pour résoudre ce problème d'inférence. L'application de cet algorithme au métabolisme d'*Escherichia coli* permet de dégager un très grand nombre de motifs. Tous les motifs extraits ne sont sans doute pas pertinents biologiquement (dans le sens où ils ne correspondent pas tous à des unités fonctionnelles et/ou évolutives explicatives de la structure du réseau). Pour nous aider dans l'interprétation de ces motifs, nous avons développé des techniques de regroupement consistant à mettre ensemble d'une part les occurrences qui se recouvrent, et d'autre part les occurrences qui ont la même topologie. En outre, nous avons proposé un critère statistique permettant de déterminer si un motif est sur-représenté.

À l'aide de ces critères supplémentaires nous avons pu extraire des motifs qui nous semblaient particulièrement intéressants (plusieurs occurrences, peu de recouvrement) et que nous avons analysés plus en détail. Nous avons pu mettre en évidence que ces motifs correspondaient pour certains à des unités fonctionnelles (les gènes appartenaient à un même opéron) et/ou à des unités évolutives (détection d'homologie entre les occurrences du motif).

Finalement, nous avons testé, pour l'ensemble des motifs extraits, la cor-

respondance qu'on pouvait trouver entre motifs et opérons. Il apparaît que les motifs ayant plusieurs paquets sont enrichis en occurrences couvertes par des opérons (*i.e.* les gènes concernés sont dans un même opéron). On note que cette observation n'est plus vraie si on considère le nombre d'occurrences et non le nombre de paquets d'occurrences.

Si on revient maintenant à notre question initiale qui était, existe-t-il un niveau d'organisation du métabolisme qui soit intermédiaire entre la réaction et la voie métabolique, il semble que nous puissions répondre par l'affirmative.

Il faut cependant préciser la définition de motif qu'on utilise. On peut notamment faire la différence, d'une part entre motifs colorés et motifs topologiques colorés, ordonnés ou non ordonnés, et d'autre part entre motifs répétés et motifs répétés en plusieurs paquets. Il semble que la notion de motif répété en plusieurs paquets soit plus explicative que la simple notion de motif répété. Par ailleurs, les exemples que nous avons analysé plus en détail dans la Section 6.2 étaient tous des motifs topologiques colorés (ordonnés ou non ordonnés).

Ma conviction personnelle est que le concept de motif coloré est moins explicatif que le concept de motif topologique coloré. La raison principale de cette conviction résulte de l'appréhension du métabolisme que j'ai pu obtenir au cours de ces trois ans. Ainsi, si l'ordre précis dans lequel s'enchaînent des réactions n'est pas une information déterminante, la topologie de cet enchaînement (linéaire ou branché) semble être une information primordiale qu'il est nécessaire de prendre en compte.

La notion de motif coloré valait cependant la peine d'être développée, d'une part parce qu'elle est plus générale que la notion de motif topologique coloré et permet donc de donner un cadre mathématique général à la recherche de motif dans les graphes, et d'autre part car elle peut trouver des applications dans d'autres domaines.

Si on devait donner une définition de motif métabolique aujourd'hui, on priviligierait la notion de motif topologique coloré non ordonné répété en plusieurs paquets.

Perspectives

Cette thèse constitue un premier travail autour de la notion de brique fonctionnelle et/ou évolutive du métabolisme. À l'issue de ces trois années, de nombreuses questions restent ouvertes, autant sur le plan méthodologique que biologique. Nous allons à présent suggérer des pistes qui pourraient être suivies pour améliorer et prolonger ce travail.

À court terme, une direction intéressante serait de tester des définitions de couleurs alternatives. En effet, la classification EC est insatisfaisante sous plusieurs aspects. Nous l'avions choisie d'une part parce qu'elle est simple à utiliser et d'autre part parce qu'elle représente un compromis entre une approche génomique et une approche biochimique. Une approche génomique

consisterait à comparer les séquences des enzymes. Une approche biochimique (ou plutôt chimique) consisterait à comparer les mécanismes réactionnels. Les deux approches sont pertinentes et complémentaires mais il peut être important de les séparer. Les numéros EC servent à comparer des réactions dans notre cas. On pourrait vouloir utiliser directement les numéros RC (un numéro RC définirait de manière non ambiguë une réaction). Parmi les travaux récents sur la comparaison de réactions, ceux de [Kotera *et al.*, 2004] méritent une attention particulière car ils pourraient permettre de définir une distance chimique entre réactions.

Toujours en ce qui concerne la modélisation, une perspective intéressante serait d'introduire un nouveau schéma de comptage des occurrences. Nous avons pu constater que compter le nombre de paquets d'occurrences pouvait être plus pertinent que de compter simplement le nombre d'occurrences. Or le nombre de paquets d'occurrences est une mesure assez grossière puisqu'on peut, par exemple, avoir dans un même paquet deux occurrences qui ne se recouvrent pas. Ainsi, il pourrait être pertinent de compter le nombre maximum d'occurrences non recouvrantes d'un motif. Ce nouveau type de comptage pose par ailleurs un problème combinatoire intéressant.

Concernant l'inférence de motif, nous avons déjà donné des pistes de réflexion pour l'élaboration d'un algorithme d'inférence (voir Section 5.1.3). Un algorithme d'inférence efficace nous permettrait de traiter des motifs de plus grande taille et ainsi de constater à partir de quelle taille le nombre d'occurrences commence à décroître, et en particulier de déterminer quel est le plus grand motif répété. Cette question du plus grand motif répété est revenue de manière récurrente lorsque j'ai présenté mon travail dans des séminaires.

Concernant cette question de maximalité, on peut d'ailleurs être intéressé non seulement par le motif de taille maximum mais aussi par tous les motifs maximaux (un motif est maximal si lorsqu'on lui ajoute une couleur, son nombre d'occurrences diminue). On note que nous avons déjà répondu à la question du motif maximum quand on travaille au seuil 4. Il s'agit du motif de taille 4 commun à la synthèse de la valine et de l'isoleucine.

Par ailleurs, nous avons signalé les limites d'utilisation du critère statistique de sur-représentation que nous avons introduit. Pour rappel, deux biais ont été identifiés : premièrement, les motifs qui contiennent des motifs sur-représentés ont tendance à être détectés comme sur-représentés, et deuxièmement, les motifs constitués de couleurs peu fréquentes ont tendance à être détectés comme sur-représentés (même s'ils n'apparaissent qu'une fois). Pour dépasser ces problèmes, des approches bayésiennes utilisant la notion de pseudo-compte semblent prometteuses.

En ce qui concerne les analyses des résultats obtenus sur le réseau métabolique d'*E. coli*, nous avons exploré le recouvrement entre occurrences mais nous n'avons pas encore examiné le recouvrement entre motifs (*i.e.* les motifs partagent des couleurs) ou entre occurrences de motifs différents (*i.e.* les occurrences partagent des noeuds). Ce type d'analyse permettrait ainsi de

déterminer si certaines couleurs sont impliquées dans beaucoup de motifs (comme les transaminases par exemple).

Par ailleurs, il serait intéressant d'appliquer notre méthode à plusieurs organismes différents et comparer les résultats obtenus. Nous pourrions par exemple nous intéresser à la comparaison du réseau métabolique d'*Escherichia coli* et de *Buchnera aphidicola*. *Buchnera* présente la particularité d'avoir un génome très réduit et il se trouve que chacun de ses gènes (sauf 3) a un orthologue chez *E. coli*. On s'attend *a priori* à ce que *Buchnera* ait beaucoup moins de motifs que *coli*. Nous pourrions alors quantifier si cette différence est réellement en termes de nombre de motifs, ou plutôt en nombre d'occurrences par motif.

La notion de motif coloré pourrait par ailleurs être appliquée à d'autres types de réseaux biologiques. Par exemple, dans le cas des réseaux d'interaction de protéines, les données d'interaction sont souvent bruitées (nombreux faux-négatifs ou faux-positifs), ce qui entraîne qu'il manque généralement des arêtes ou que des arêtes sont en trop dans ce type de graphe. La notion de motif coloré (qui ignore la topologie exacte) pourrait être tout à fait pertinente dans ce cadre.

Enfin, concernant l'étude de l'évolution du métabolisme, nous avons pu mettre en évidence pour certains motifs des relations d'homologie entre enzymes. Nous pourrions tester si ces cas sont fréquents, notamment parmi les motifs répétés en plusieurs paquets.

D'autre part, pour les cas détectés, une piste intéressante serait de confronter l'histoire évolutive de ces motifs aux scénarios d'évolution des voies métaboliques présentés à la Section 2.4.5.2. En effet, une des conclusions des travaux de [Rison *et al.*, 2002] est qu'un seul type de scénario ne peut expliquer l'évolution de tout le métabolisme. Localement, on peut avoir de la rétroévolution, puis de la duplication de voies, puis de l'évolution en patchwork. La notion de motif peut sans doute aider à explorer certains scénarios.

En ce qui concerne l'étude du lien entre métabolisme et génome, nous n'avons pour l'instant exploré que le lien entre motifs et opérons. Si on veut faire le même type d'analyses sur des réseaux eucaryotes, la notion d'opéron ne peut plus être utilisée. On peut donc vouloir généraliser à la notion de gènes "proches sur le génome", ou proches spatialement dans le noyau.

Enfin, nous nous sommes intéressés pour l'instant exclusivement à la structure du métabolisme sans tenir compte de la régulation des enzymes. Or toutes les enzymes d'un organisme ne sont pas présentes dans une cellule au même instant. Comprendre les mécanismes complexes qui sont en jeu dans la régulation de la transcription de ces enzymes, notamment au moment de l'épissage, constitue aujourd'hui un enjeu majeur auquel nous souhaiterions à présent nous intéresser.

Bibliographie

[Agrawal *et al.*, 1993] Rakesh Agrawal, Tomasz Imieliński, et Arun Swami. Mining association rules between sets of items in large databases. *SIGMOD '93 : Proceedings of the 1993 ACM SIGMOD international conference on Management of data*, 207–216, New York, NY, USA, 1993. ACM Press.

[Albert *et al.*, 2000] Albert, Jeong, et Barabasi. Error and attack tolerance of complex networks. *Nature*, 406(6794) :378–382, Jul 2000.

[Allen, 2001] J. F. Allen. Bioinformatics and discovery : induction beckons again. *Bioessays*, 23(1) :104–107, Jan 2001.

[Alm et Arkin, 2003] Eric Alm et Adam P Arkin. Biological networks. *Curr Opin Struct Biol*, 13(2) :193–202, Apr 2003.

[Alon, 2003] U. Alon. Biological networks : the tinkerer as an engineer. *Science*, 301(5641) :1866–1867, Sep 2003.

[Alon, 2006] Uri Alon. *An introduction to systems biology : design principles of biological networks*. Chapman and Hall, 2006.

[Arita, 2004] Masanori Arita. The metabolic world of Escherichia coli is not small. *Proc Natl Acad Sci U S A*, 101(6) :1543–1547, Feb 2004.

[Artzy-Randrup *et al.*, 2004] Yael Artzy-Randrup, Sarel J Fleishman, Nir Ben-Tal, et Lewi Stone. Comment on "Network motifs : simple building blocks of complex networks" and "Superfamilies of evolved and designed networks". *Science*, 305(5687) :1107 ; author reply 1107, Aug 2004.

[Babu *et al.*, 2004] M. Madan Babu, Nicholas M Luscombe, L. Aravind, Mark Gerstein, et Sarah A Teichmann. Structure and evolution of transcriptional regulatory networks. *Curr Opin Struct Biol*, 14(3) :283–291, Jun 2004.

[Barabasi et Albert, 1999] Barabasi et Albert. Emergence of scaling in random networks. *Science*, 286(5439) :509–512, Oct 1999.

[Bourqui *et al.*, 2007] Romain Bourqui, Ludovic Cottret, Vincent Lacroix, David Auber, Patrick Mary, Marie-France Sagot, et Fabien Jourdan. Metabolic network visualization eliminating node redundance and preserving metabolic pathways. *BMC Syst Biol*, 1(1) :29, Jul 2007.

[Boyer *et al.*, 2005] Frédéric Boyer, Anne Morgat, Laurent Labarre, Joël Pothier, et Alain Viari. Syntons, metabolons and interactons : an exact

graph-theoretical approach for exploring neighbourhood between genomic and functional data. *Bioinformatics*, 21(23) :4209–4215, Dec 2005.

[Boyer et Viari, 2003] Frédéric Boyer et Alain Viari. Ab initio reconstruction of metabolic pathways. *Bioinformatics*, 19 Suppl 2 :II26–II34, Oct 2003.

[Burgard *et al.*, 2004] Anthony P Burgard, Evgeni V Nikolaev, Christophe H Schilling, et Costas D Maranas. Flux coupling analysis of genome-scale metabolic network reconstructions. *Genome Res*, 14(2) :301–312, Feb 2004.

[Carpentier *et al.*, 2005] Mathilde Carpentier, Sophie Brouillet, et Joël Pothier. YAKUSA : a fast structural database scanning method. *Proteins*, 61(1) :137–151, Oct 2005.

[Clarke, 1981] B. L. Clarke. Complete set of steady states for the general stoichiometric dynamical system. *J. Chem. Phys.*, 75 :4970–4979, 1981.

[Claudel-Renard *et al.*, 2003] Clotilde Claudel-Renard, Claude Chevalet, Thomas Faraut, et Daniel Kahn. Enzyme-specific profiles for genome annotation : PRIAM. *Nucleic Acids Res*, 31(22) :6633–6639, Nov 2003.

[Clemente *et al.*, 2005] José C Clemente, Kenji Satou, et Gabriel Valiente. Reconstruction of phylogenetic relationships from metabolic pathways based on the enzyme hierarchy and the gene ontology. *Genome Inform*, 16(2) :45–55, 2005.

[Colom et Silva, 1991] J. M. Colom et M. Silva. Convex geometry and semiflows in p/t nets. a comparative study of algorithms for computation of minimal p-semiflows. *Proceedings of the 10th International Conference on Applications and Theory of Petri Nets*, 79–112, London, UK, 1991. Springer-Verlag.

[Copley et Bork, 2000] R. R. Copley et P. Bork. Homology among (betaalpha)(8) barrels : implications for the evolution of metabolic pathways. *J Mol Biol*, 303(4) :627–641, Nov 2000.

[Cornish-Bowden, 1995] A. Cornish-Bowden. Metabolic control analysis in theory and practice. *Adv Mol Cell Biol*, 11(21-64), 1995.

[Coulomb *et al.*, 2005] Stéphane Coulomb, Michel Bauer, Denis Bernard, et Marie-Claude Marsolier-Kergoat. Gene essentiality and the topology of protein interaction networks. *Proc Biol Sci*, 272(1573) :1721–1725, Aug 2005.

[Croes *et al.*, 2006] Didier Croes, Fabian Couche, Shoshana J Wodak, et Jacques van Helden. Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol*, 356(1) :222–236, Feb 2006.

[Csete et Doyle, 2002] Marie E Csete et John C Doyle. Reverse engineering of biological complexity. *Science*, 295(5560) :1664–1669, Mar 2002.

[Cunchillos et Lecointre, 2003] Chomin Cunchillos et Guillaume Lecointre. Evolution of amino acid metabolism inferred through cladistic analysis. *J Biol Chem*, 278(48) :47960–47970, Nov 2003.

[Daudin *et al.*, 2006] Jean-Jacques Daudin, Franck Picard, et Stéphane Robin. A mixture model for random graphs. , INRIA, 2006.

[Deville *et al.*, 2003] Yves Deville, David Gilbert, Jacques van Helden, et Shoshana J Wodak. An overview of data models for the analysis of biochemical pathways. *Brief Bioinform*, 4(3) :246–259, Sep 2003.

[Dorogovtsev *et al.*, 2000] S. N. Dorogovtsev, J. F. Mendes, et A. N. Samukhin. Structure of growing networks with preferential linking. *Phys Rev Lett*, 85(21) :4633–4636, Nov 2000.

[Duarte *et al.*, 2007] Natalie C Duarte, Scott A Becker, Neema Jamshidi, Ines Thiele, Monica L Mo, Thuy D Vo, Rohith Srivas, et Bernhard Ø Palsson. Global reconstruction of the human metabolic network based on genomic and bibliomic data. *Proc Natl Acad Sci U S A*, 104(6) :1777–1782, Feb 2007.

[Edwards *et al.*, 2001] J. S. Edwards, R. U. Ibarra, et B. O. Palsson. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol*, 19(2) :125–130, Feb 2001.

[Fell et Wagner, 2000] D. A. Fell et A. Wagner. The small world of metabolism. *Nat Biotechnol*, 18(11) :1121–1122, Nov 2000.

[Fisher, 1958] R. A. Fisher. *The Genetical Theory of Natural Selection*. Oxford University Press, 2nd , April 1958.

[Fiévet *et al.*, 2006] Julie B Fiévet, Christine Dillmann, Gilles Curien, et Dominique de Vienne. Simplified modelling of metabolic pathways for flux prediction and optimization : lessons from an in vitro reconstruction of the upper part of glycolysis. *Biochem J*, 396(2) :317–326, Jun 2006.

[Fjallstrom, 1998] P. Fjallstrom. Algorithms for graph partitioning : A survey. 1998.

[Fothergill-Gilmore et Michels, 1993] L. A. Fothergill-Gilmore et P. A. Michels. Evolution of glycolysis. *Prog Biophys Mol Biol*, 59(2) :105–235, 1993.

[Gagneur et Klamt, 2004] Julien Gagneur et Steffen Klamt. Computation of elementary modes : a unifying framework and the new binary approach. *BMC Bioinformatics*, 5 :175, 2004.

[Gagneur *et al.*, 2004] Julien Gagneur, Roland Krause, Tewis Bouwmeester, et Georg Casari. Modular decomposition of protein-protein interaction networks. *Genome Biol*, 5(8) :R57, 2004.

[Garey et Johnson, 1979] M. R. Garey et D. S. Johnson. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. Freeman, 1979.

[Garey *et al.*, 1976] M. R. Garey, David S. Johnson, et Larry J. Stockmeyer. Some simplified np-complete graph problems. *Theor. Comput. Sci.*, 1(3) :237–267, 1976.

[Gerlt et Babbitt, 2001] J. A. Gerlt et P. C. Babbitt. Divergent evolution of enzymatic function : mechanistically diverse superfamilies and functionally distinct suprafamilies. *Annu Rev Biochem*, 70 :209–246, 2001.

[Gilbert et Mulkay, 1984] G. N. Gilbert et M. Mulkay. *Opening Pandora's box : a sociological analysis of scientists' discourse.* Cambridge University Press, 1984.

[Green et Karp, 2004] Michelle L Green et Peter D Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5 :76, Jun 2004.

[Green et Karp, 2005] M. L. Green et P. D. Karp. Genome annotation errors in pathway databases due to semantic ambiguity in partial EC numbers. *Nucleic Acids Res*, 33(13) :4035–4039, 2005.

[Guet *et al.*, 2002] Calin C Guet, Michael B Elowitz, Weihong Hsing, et Stanislas Leibler. Combinatorial synthesis of genetic networks. *Science*, 296(5572) :1466–1470, May 2002.

[Guimerà et Amaral, 2005] Roger Guimerà et Luís A Nunes Amaral. Functional cartography of complex metabolic networks. *Nature*, 433(7028) :895–900, Feb 2005.

[Habib *et al.*, 2004] M. Habib, F. de Montgolfier, et C. Paul. A simple linear-time modular decomposition algorithm for graphs, using order extension. *SWAT'04, 9th Scandinavian Workshop on Algorithm Theory*, 2004.

[Handorf *et al.*, 2005] Thomas Handorf, Oliver Ebenhöh, et Reinhart Heinrich. Expanding metabolic networks : scopes of compounds, robustness, and evolution. *J Mol Evol*, 61(4) :498–512, Oct 2005.

[Hardy et Robillard, 2004] Simon Hardy et Pierre N Robillard. Modeling and simulation of molecular biology systems using petri nets : modeling goals of various approaches. *J Bioinform Comput Biol*, 2(4) :595–613, Dec 2004.

[Hartwell *et al.*, 1999] L. H. Hartwell, J. J. Hopfield, S. Leibler, et A. W. Murray. From molecular to modular cell biology. *Nature*, 402(6761 Suppl) :C47–C52, Dec 1999.

[Heinrich et Rapoport, 1974] R. Heinrich et T. A. Rapoport. A linear steady-state treatment of enzymatic chains. General properties, control and effector strength. *Eur J Biochem*, 42(1) :89–95, Feb 1974.

[Hermelin *et al.*, 2007] D. Hermelin, M. Fellows, G. Fertin, et S. Vialette. Sharp tractability borderlines for finding connected motifs in vertex-colored graphs. *Proc. 34th International Colloquium on Automata, Languages and Programming (ICALP), Wroclaw, Poland*, Lecture Notes In Computer Science, 2007.

[Heymans et Singh, 2003] Maureen Heymans et Ambuj K Singh. Deriving phylogenetic trees from the similarity analysis of metabolic pathways. *Bioinformatics*, 19 Suppl 1 :i138–i146, 2003.

[Hirsh et Sharan, 2007] Eitan Hirsh et Roded Sharan. Identification of conserved protein complexes based on a model of protein network evolution. *Bioinformatics*, 23(2) :e170–e176, Jan 2007.

[Hofestädt, 1994] Ralf Hofestädt. A petri net application to model metabolic processes. *Syst. Anal. Model. Simul.*, 16(2) :113–122, 1994.

[Horne *et al.*, 2004] A. B. Horne, T. C. Hodgman, H. D. Spence, et A. R. Dalby. Constructing an enzyme-centric view of metabolism. *Bioinformatics*, 20(13) :2050–2055, Sep 2004.

[Horowitz, 1945] N. H. Horowitz. On the Evolution of Biochemical Syntheses. *Proc Natl Acad Sci U S A*, 31(6) :153–157, Jun 1945.

[Hrmova *et al.*, 2002] Maria Hrmova, Ross De Gori, Brian J Smith, Jon K Fairweather, Hugues Driguez, Joseph N Varghese, et Geoffrey B Fincher. Structural basis for broad substrate specificity in higher plant beta-D-glucan glucohydrolases. *Plant Cell*, 14(5) :1033–1052, May 2002.

[Hume, 1999] David Hume. *An enquiry concerning human understanding. 1748*. Oxford University Press, 1999.

[Huynen et Snel, 2000] M. A. Huynen et B. Snel. Gene and context : integrative approaches to genome analysis. *Adv Protein Chem*, 54 :345–379, 2000.

[Ingram *et al.*, 2006] Piers J Ingram, Michael P H Stumpf, et Jaroslav Stark. Network motifs : structure does not determine function. *BMC Genomics*, 7 :108, 2006.

[Irvin et Bhattacharjee, 1998] S. D. Irvin et J. K. Bhattacharjee. A unique fungal lysine biosynthesis enzyme shares a common ancestor with tricarboxylic acid cycle and leucine biosynthetic enzymes found in diverse organisms. *J Mol Evol*, 46(4) :401–408, Apr 1998.

[Jacob, 1977] F. Jacob. Evolution and tinkering. *Science*, 196 :116–166, 1977.

[Jacob et Monod, 1961] F. Jacob et J. Monod. Genetic regulatory mechanisms in the synthesis of proteins. *J Mol Biol*, 3 :318–356, Jun 1961.

[Jensen, 1976] R. A. Jensen. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*, 30 :409–425, 1976.

[Jeong *et al.*, 2001] H. Jeong, S. P. Mason, A. L. Barabási, et Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(6833) :41–42, May 2001.

[Jeong *et al.*, 2000] H. Jeong, B. Tombor, R. Albert, Z. N. Oltvai, et A. L. Barabási. The large-scale organization of metabolic networks. *Nature*, 407(6804) :651–654, Oct 2000.

[Jong *et al.*, 2004] Hidde De Jong, Jean-Luc Gouzé, Céline Hernandez, Michel Page, Tewfik Sari, et Johannes Geiselmann. Qualitative simulation of genetic regulatory networks using piecewise-linear models. *Bull Math Biol*, 66(2) :301–340, Mar 2004.

[Kacser et Burns, 1973] H. Kacser et JA. Burns. Control of enzyme flux. *Symp Soc Exp Biol*, 27 :65–104, 1973.

[Karp *et al.*, 2002] Peter D Karp, Suzanne Paley, et Pedro Romero. The Pathway Tools software. *Bioinformatics*, 18 Suppl 1 :S225–S232, 2002.

[Karp et Paley, 1994] P. D. Karp et S. M. Paley. Representations of metabolic knowledge : pathways. *Proc Int Conf Intell Syst Mol Biol*, 2 :203–211, 1994.

[Kell et Oliver, 2004] Douglas B Kell et Stephen G Oliver. Here is the evidence, now what is the hypothesis ? The complementary roles of inductive and hypothesis-driven science in the post-genomic era. *Bioessays*, 26(1) :99–105, Jan 2004.

[Keller, 2005] Evelyn Fox Keller. Revisiting "scale-free" networks. *Bioessays*, 27(10) :1060–1068, Oct 2005.

[Keseler *et al.*, 2005] Ingrid M Keseler, Julio Collado-Vides, Socorro Gama-Castro, John Ingraham, Suzanne Paley, Ian T Paulsen, Martín Peralta-Gil, et Peter D Karp. EcoCyc : a comprehensive database resource for Escherichia coli. *Nucleic Acids Res*, 33(Database issue) :D334–D337, Jan 2005.

[Khanin et Wit, 2006] Raya Khanin et Ernst Wit. How scale-free are biological networks. *J Comput Biol*, 13(3) :810–818, Apr 2006.

[Klamt *et al.*, 2006] Steffen Klamt, Julio Saez-Rodriguez, Jonathan A Lindquist, Luca Simeoni, et Ernst D Gilles. A methodology for the structural and functional analysis of signaling and regulatory networks. *BMC Bioinformatics*, 7 :56, 2006.

[Klamt et Stelling, 2002] Steffen Klamt et Jörg Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2) :233–236, 2002.

[Klamt et Stelling, 2003] Steffen Klamt et Jörg Stelling. Two approaches for metabolic pathway analysis ? *Trends Biotechnol*, 21(2) :64–69, Feb 2003.

[Kleinberg, 2000] Jon Kleinberg. The small-world phenomenon : an algorithm perspective. *STOC '00 : Proceedings of the thirty-second annual ACM symposium on Theory of computing*, 163–170. ACM Press, 2000.

[Kotera *et al.*, 2004] Masaaki Kotera, Yasushi Okuno, Masahiro Hattori, Susumu Goto, et Minoru Kanehisa. Computational assignment of the EC numbers for genomic-scale analysis of enzymatic reactions. *J Am Chem Soc*, 126(50) :16487–16498, Dec 2004.

[Lacroix *et al.*, 2005] V. Lacroix, C. G. Fernandes, et M.-F. Sagot. Reaction motifs in metabolic networks. *Proceedings of 5th Workshop on Algorithms for BioInformatics (WABI'05), Lecture Notes in BioInformatics, subseries Lecture Notes in Computer Science*, 3692, 178–191, 2005.

[Lacroix *et al.*, 2006] Vincent Lacroix, Cristina G Fernandes, et Marie-France Sagot. Motif search in graphs : application to metabolic networks. *IEEE/ACM Trans Comput Biol Bioinform*, 3(4) :360–368, 2006.

[Liao *et al.*, 2002] L. Liao, S. Kim, et J.F. Tomb. Genome comparisons based on profiles of metabolic pathways. *Sixth international conference on knowledge-based intelligent informationand engineering systems*, Crema, Italy, 2002.

[Ma et Zeng, 2003] Hong-Wu Ma et An-Ping Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11) :1423–1430, Jul 2003.

[McKay, 1990] Brendan D. McKay. nauty *user's guide (version 2.2), Technical report TR-CS-90-02*. Computer Science Department, Australian National University,, 1990.

[Milo *et al.*, 2002] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, et U. Alon. Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–827, Oct 2002.

[Min *et al.*, 2002] Bokkee Min, Joanne T Pelaschier, David E Graham, Debra Tumbula-Hansen, et Dieter Söll. Transfer RNA-dependent amino acid biosynthesis : an essential route to asparagine formation. *Proc Natl Acad Sci U S A*, 99(5) :2678–2683, Mar 2002.

[Miyazaki *et al.*, 2001] J. Miyazaki, N. Kobashi, M. Nishiyama, et H. Yamane. Functional and evolutionary relationship between arginine biosynthesis and prokaryotic lysine biosynthesis through alpha-aminoadipate. *J Bacteriol*, 183(17) :5067–5073, Sep 2001.

[Möhring et Radermacher, 1984] Rolf H. Möhring et Franz J. Radermacher. Substitution decomposition for discrete structures and connections with combinatorial optimization. *Ann. Discrete Math.*, 19 :257–356, 1984.

[Murata, 1989] Tadao Murata. Petri nets : Properties, analysis and applications. *Proceedings of the IEEE*, 541–580, 1989. NewsletterInfo : 33Published as Proceedings of the IEEE, volume 77, number 4.

[Nahum et Riley, 2001] L. A. Nahum et M. Riley. Divergence of function in sequence-related groups of Escherichia coli proteins. *Genome Res*, 11(8) :1375–1381, Aug 2001.

[Nelson et Cox, 2004] David L. Nelson et Michael M. Cox. *Lehninger Principles of Biochemistry, Fourth Edition*. W. H. Freeman, April 2004.

[Newman, 2006] M. E J Newman. Modularity and community structure in networks. *Proc Natl Acad Sci U S A*, 103(23) :8577–8582, Jun 2006.

[Newman et Girvan, 2004] M. E J Newman et M. Girvan. Finding and evaluating community structure in networks. *Phys Rev E Stat Nonlin Soft Matter Phys*, 69(2 Pt 2) :026113, Feb 2004.

[Nobeli et Thornton, 2006] Irene Nobeli et Janet M Thornton. A bioinformatician's view of the metabolome. *Bioessays*, 28(5) :534–545, May 2006.

[Nomenclature, 1992] Nomenclature. *Enzyme nomenclature. Recommendations 1992*. Academic Press, August 1992.

[Papin *et al.*, 2003] Jason A Papin, Nathan D Price, Sharon J Wiback, David A Fell, et Bernhard O Palsson. Metabolic pathways in the post-genome era. *Trends Biochem Sci*, 28(5) :250–258, May 2003.

[Petri, 1962] C. A. Petri. *Communication with automata (in German)*. PhD thesis, Institut für instrumentelle Mathematik, Bonn : Schriften des IIM Nr. 3, 1962.

[Petsko *et al.*, 1993] G. A. Petsko, G. L. Kenyon, J. A. Gerlt, D. Ringe, et J. W. Kozarich. On the origin of enzymatic species. *Trends Biochem Sci*, 18(10) :372–376, Oct 1993.

[Picard *et al.*, 2007] Franck Picard, Jean-Jacques Daudin, Sophie Schbath, et Stéphane Robin. Assessing the exceptionality of network motifs. , SSB, 2007.

[Pinter *et al.*, 2005] Ron Y Pinter, Oleg Rokhlenko, Esti Yeger-Lotem, et Michal Ziv-Ukelson. Alignment of metabolic pathways. *Bioinformatics*, 21(16) :3401–3408, Aug 2005.

[Popper, 1992] Karl Popper. *Conjectures and refutations : the growth of scientific knowledge*. Routledge and Kegan Paul, London, 5^{th}*edition*, 1992.

[Przytycka et Yu, 2004] Teresa M Przytycka et Yi-Kuo Yu. Scale-free networks versus evolutionary drift. *Comput Biol Chem*, 28(4) :257–264, Oct 2004.

[Pál *et al.*, 2005] Csaba Pál, Balázs Papp, et Martin J Lercher. Adaptive evolution of bacterial metabolic networks by horizontal gene transfer. *Nat Genet*, 37(12) :1372–1375, Dec 2005.

[Pál *et al.*, 2006] Csaba Pál, Balázs Papp, Martin J Lercher, Péter Csermely, Stephen G Oliver, et Laurence D Hurst. Chance and necessity in the evolution of minimal metabolic networks. *Nature*, 440(7084) :667–670, Mar 2006.

[Reddy *et al.*, 1996] V. N. Reddy, M. N. Liebman, et M. L. Mavrovouniotis. Qualitative analysis of biochemical reaction systems. *Comput Biol Med*, 26(1) :9–24, Jan 1996.

[Reddy *et al.*, 1993] V. N. Reddy, M. L. Mavrovouniotis, et M. N. Liebman. Petri net representations in metabolic pathways. *Proc Int Conf Intell Syst Mol Biol*, 1 :328–336, 1993.

[Reder, 1988] C. Reder. Metabolic control theory : a structural approach. *J Theor Biol*, 135(2) :175–201, Nov 1988.

[Reed *et al.*, 2003] Jennifer L Reed, Thuy D Vo, Christophe H Schilling, et Bernhard O Palsson. An expanded genome-scale model of Escherichia coli K-12 (iJR904 GSM/GPR). *Genome Biol*, 4(9) :R54, 2003.

[Rison *et al.*, 2002] Stuart C G Rison, Sarah A Teichmann, et Janet M Thornton. Homology, pathway distance and chromosomal localization of the small molecule metabolism enzymes in Escherichia coli. *J Mol Biol*, 318(3) :911–932, May 2002.

[Roy, 1999] S. Roy. Multifunctional enzymes and evolution of biosynthetic pathways : retro-evolution by jumps. *Proteins*, 37(2) :303–309, Nov 1999.

[Sackmann *et al.*, 2006] Andrea Sackmann, Monika Heiner, et Ina Koch. Application of Petri net based analysis techniques to signal transduction pathways. *BMC Bioinformatics*, 7 :482, 2006.

[Sagot, 1998] Marie-France Sagot. Spelling approximate repeated or common motifs using a suffix tree. *LATIN '98 : Proceedings of the Third Latin American Symposium on Theoretical Informatics*, 374–390, London, UK, 1998. Springer-Verlag.

[Salgado *et al.*, 2006] Heladia Salgado, Socorro Gama-Castro, Martín Peralta-Gil, Edgar Díaz-Peredo, Fabiola Sánchez-Solano, Alberto Santos-Zavaleta, Irma Martínez-Flores, Verónica Jiménez-Jacinto, César Bonavides-Martínez, Juan Segura-Salazar, Agustino Martínez-Antonio, et Julio Collado-Vides. RegulonDB (version 5.0) : Escherichia coli K-12 transcriptional regulatory network, operon organization, and growth conditions. *Nucleic Acids Res*, 34(Database issue) :D394–D397, Jan 2006.

[Schbath, 1995] S. Schbath. Compound poisson approximation of word counts in DNA sequences. *ESAIM : Probability and Statistics*, 1 :1–16. (`http ://www.emath.fr/ps/`)., 1995.

[Schilling *et al.*, 2000] C. H. Schilling, D. Letscher, et B. O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, 203(3) :229–248, Apr 2000.

[Schilling *et al.*, 1999] C. H. Schilling, S. Schuster, B. O. Palsson, et R. Heinrich. Metabolic pathway analysis : basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog*, 15(3) :296–303, 1999.

[Schmidt *et al.*, 2003] Steffen Schmidt, Shamil Sunyaev, Peer Bork, et Thomas Dandekar. Metabolites : a helping hand for pathway evolution ? *Trends Biochem Sci*, 28(6) :336–341, Jun 2003.

[Schuster *et al.*, 2000] S. Schuster, D. A. Fell, et T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol*, 18(3) :326–332, Mar 2000.

[Schuster et Hilgetag, 1994] S. Schuster et C. Hilgetag. On elementary flux modes in biochemical reactions systems at steady state. *J. Biol. Syst.*, 2 :165–182, 1994.

[Schuster *et al.*, 2002] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, et T. Dandekar. Exploring the pathway structure of metabolism : decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, 18(2) :351–361, Feb 2002.

[Scott *et al.*, 2006] Jacob Scott, Trey Ideker, Richard M Karp, et Roded Sharan. Efficient algorithms for detecting signaling pathways in protein interaction networks. *J Comput Biol*, 13(2) :133–144, Mar 2006.

[Shamir *et al.*, 2002] R. Shamir, R. Sharan, et D. Tsur. Cluster graph modification problems. 2002.

[Shannon *et al.*, 2003] Paul Shannon, Andrew Markiel, Owen Ozier, Nitin S Baliga, Jonathan T Wang, Daniel Ramage, Nada Amin, Benno Schwikowski, et Trey Ideker. Cytoscape : a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11) :2498–2504, Nov 2003.

[Sharan *et al.*, 2005] Roded Sharan, Trey Ideker, Brian Kelley, Ron Shamir, et Richard M Karp. Identification of protein complexes by comparative analysis of yeast and bacterial protein interaction data. *J Comput Biol*, 12(6) :835–846, 2005.

[Shen-Orr *et al.*, 2002] Shai S Shen-Orr, Ron Milo, Shmoolik Mangan, et Uri Alon. Network motifs in the transcriptional regulation network of Escherichia coli. *Nat Genet*, 31(1) :64–68, May 2002.

[Solé et Valverde, 2006] Ricard V Solé et Sergi Valverde. Are network motifs the spandrels of cellular complexity ? *Trends Ecol Evol*, 21(8) :419–422, Aug 2006.

[Spirin *et al.*, 2006] Victor Spirin, Mikhail S Gelfand, Andrey A Mironov, et Leonid A Mirny. A metabolic network in the evolutionary context : multiscale structure and modularity. *Proc Natl Acad Sci U S A*, 103(23) :8774–8779, Jun 2006.

[Spirin et Mirny, 2003] Victor Spirin et Leonid A Mirny. Protein complexes and functional modules in molecular networks. *Proc Natl Acad Sci U S A*, 100(21) :12123–12128, Oct 2003.

[Stelling, 2004] Jörg Stelling. Mathematical models in microbial systems biology. *Curr Opin Microbiol*, 7(5) :513–518, Oct 2004.

[Stumpf *et al.*, 2005] Michael P H Stumpf, Carsten Wiuf, et Robert M May. Subnets of scale-free networks are not scale-free : sampling properties of networks. *Proc Natl Acad Sci U S A*, 102(12) :4221–4224, Mar 2005.

[Szallasi *et al.*, 2006] Zoltan Szallasi, Jörg Stelling, et Vipul Periwal. *System Modeling in Cellular Biology : From Concepts to Nuts and Bolts*. The MIT Press, 2006.

[Teichmann *et al.*, 2001] S. A. Teichmann, S. C. Rison, J. M. Thornton, M. Riley, J. Gough, et C. Chothia. The evolution and structural anatomy of the small molecule metabolic pathways in Escherichia coli. *J Mol Biol*, 311(4) :693–708, Aug 2001.

[Tohsato *et al.*, 2000] Y. Tohsato, H. Matsuda, et A. Hashimoto. A multiple alignment algorithm for metabolic pathway analysis using enzyme hierarchy. *Proc Int Conf Intell Syst Mol Biol*, 8 :376–383, 2000.

[Tucker et Moulton, 2005] Warwick Tucker et Vincent Moulton. Reconstructing metabolic networks using interval analysis. *Proceedings of 5th Workshop on Algorithms for BioInformatics (WABI'05), Lecture Notes in BioInformatics, subseries Lecture Notes in Computer Science*, 3692, 192–203, 2005.

[Velasco *et al.*, 2002] A. M. Velasco, J. I. Leguina, et A. Lazcano. Molecular evolution of the lysine biosynthetic pathways. *J Mol Evol*, 55(4) :445–459, Oct 2002.

[Vergez et Huisman, 1990] André Vergez et Denis Huisman. *Cours de Philosophie*. Nathan, 1990.

[von Dassow *et al.*, 2000] G. von Dassow, E. Meir, E. M. Munro, et G. M. Odell. The segment polarity network is a robust developmental module. *Nature*, 406(6792) :188–192, Jul 2000.

[von Mering *et al.*, 2003] Christian von Mering, Evgeny M Zdobnov, Sophia Tsoka, Francesca D Ciccarelli, Jose B Pereira-Leal, Christos A Ouzounis, et Peer Bork. Genome evolution reveals biochemical networks and functional modules. *Proc Natl Acad Sci U S A*, 100(26) :15428–15433, Dec 2003.

[Voss *et al.*, 2003] Klaus Voss, Monika Heiner, et Ina Koch. Steady state analysis of metabolic pathways using Petri nets. *In Silico Biol*, 3(3) :367–387, 2003.

[Wagner, 2005] Andreas Wagner. *Robustness and Evolvability in Living Systems (Princeton Studies in Complexity)*. Princeton University Press, August 2005.

[Wagner et Fell, 2001] A. Wagner et D. A. Fell. The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478) :1803–1810, Sep 2001.

[Wagner, 1996] G. Wagner. Homologues, natural kinds and the evolution of modularity. 1996.

[Wang et Zhang, 2007] Zhi Wang et Jianzhi Zhang. In Search of the Biological Significance of Modular Structures in Protein Networks. *PLoS Comput Biol*, 3(6) :e107, Jun 2007.

[Wasserman et Faust, 1994] S. Wasserman et K. Faust. *Social network analysis*. Cambridge University Press, Cambridge, 1994.

[Waterman *et al.*, 1984] M. S. Waterman, R. Arratia, et D. J. Galas. Pattern recognition in several sequences : consensus and alignment. *Bull Math Biol*, 46(4) :515–527, 1984.

[Watts et Strogatz, 1998] D. J. Watts et S. H. Strogatz. Collective dynamics of 'small-world' networks. *Nature*, 393(6684) :440–442, Jun 1998.

[Wittig et Beuckelaer, 2001] U. Wittig et A. De Beuckelaer. Analysis and comparison of metabolic pathway databases. *Brief Bioinform*, 2(2) :126–142, May 2001.

[Wolf et Arkin, 2003] Denise M Wolf et Adam P Arkin. Motifs, modules and games in bacteria. *Curr Opin Microbiol*, 6(2) :125–134, Apr 2003.

[Wright, 1932] Sewall Wright. The roles of mutation, inbreeding, crossbreeding and selection in evolution. *Proceedings of the Sixth International Congress of Genetics*, 1 :356–366, 1932.

[Wuchty, 2001] S. Wuchty. Scale-free behavior in protein domain networks. *Mol Biol Evol*, 18(9) :1694–1702, Sep 2001.

[Yamada *et al.*, 2006] Takuji Yamada, Minoru Kanehisa, et Susumu Goto. Extraction of phylogenetic network modules from the metabolic network. *BMC Bioinformatics*, 7 :130, 2006.

[Yue *et al.*, 2006] Hong Yue, Martin Brown, Joshua Knowles, Hong Wang, David S Broomhead, et Douglas B Kell. Insights into the behaviour of systems biology models from dynamic sensitivity and identifiability analysis : a case study of an NF-kappaB signalling pathway. *Mol Biosyst*, 2(12) :640–649, Dec 2006.

[Zhu *et al.*, 2007] Xiaowei Zhu, Mark Gerstein, et Michael Snyder. Getting connected : analysis and principles of biological networks. *Genes Dev*, 21(9) :1010–1024, May 2007.

Annexe A

Éléments de base sur le métabolisme

La **cellule** (en latin *cellula* signifie petite chambre) est l'unité structurale, fonctionnelle et reproductrice constituant tout ou partie d'un être vivant. Chaque cellule est une entité vivante qui, dans le cas d'organismes multicellulaires, fonctionne de manière autonome mais coordonnée avec les autres. Les cellules de même type sont réunies en tissus, eux même réunis en organes. Selon que l'ADN est séparé du cytoplasme ou non, la cellule sera eucaryote ou procaryote.

L'acide désoxyribonucléique (souvent abrégé en **ADN**) est une molécule que l'on retrouve dans tous les organismes vivants. On dit que l'ADN est le support de l'hérédité car il constitue le génome des êtres vivants et se transmet en totalité ou en partie lors des processus de reproduction. Il est à la base de la synthèse des protéines.

Un **gène** est défini comme un enchaînement de nucléotides (dite aussi séquence), c'est-à-dire comme une portion d'acide désoxyribonucléique (ADN), destiné à être transcrit en acide ribonucléique (ARN) ; si c'est le cas la séquence est dite « codante ». La plupart du temps, un gène commence par une séquence de nucléotides appelée promoteur, dont le rôle est de permettre l'initiation mais surtout la régulation (tous les gènes ne sont pas exprimés dans toutes les cellules) de la transcription de l'ADN en ARN, et se termine par une séquence terminatrice, qui marque la fin de la transcription. La molécule d'ARN ainsi produite peut soit être traduite en protéine (elle est dans ce cas appelée ARN messager), soit être directement fonctionnelle (c'est le cas pour les ARN ribosomaux ou les ARN de transfert). Il y a environ 13000 gènes dans l'ADN des cellules d'une drosophile et entre 23000 et 27000 gènes chez l'Homme.

Une **protéine**, ou aussi appelé protide, est une macromolécule composée par une chaîne (ou séquence) d'acides aminés liés entre eux par des liaisons peptidiques. L'enchaînement des acides aminés est codé par le génome et constitue la structure primaire d'une protéine.

Une **enzyme** (nom féminin, souvent utilisé au masculin) est une molécule

(protéine ou ARN dans le cas des ribozymes) permettant d'accélérer jusqu'à des millions de fois les réactions chimiques du métabolisme se déroulant dans le milieu cellulaire ou extracellulaire. Les enzymes agissent à faible concentration et elles se retrouvent intactes en fin de réaction : ce sont des catalyseurs biologiques (ou biocatalyseurs).

Un **métabolite** est une petite molécule. On fait généralement la distinction en biochimie entre petites molécules et macromolécules. Les macromolécules sont la plupart du temps un assemblage de petites molécules par des liaisons covalentes (ADN, ARN, protéines, polysaccharides). Dans cette thèse, on utilise indifféremment les termes de composé, métabolite ou petite molécule.

Une **réaction** chimique est une transformation de la matière. Au cours d'une réaction chimique, les espèces chimiques (atomiques, ioniques ou moléculaires) qui constituent la matière sont modifiées : les espèces qui sont consommées sont appelées réactifs. Les espèces formées au cours de la réaction sont appelées produits (de réaction).

Annexe B

Éléments de base sur la complexité algorithmique

Ce glossaire regroupe des notions de base sur la théorie de la complexité. Nous les introduisons ici de façon informelle.

Un **algorithme** est un énoncé dans un langage bien défini d'une suite d'opérations permettant de résoudre par calcul un problème.

Un **problème** est un ensemble de questions, où chaque question est une chaîne de caractères de longueur finie. Par exemple, le problème FACTORISER est : "étant donné un entier, donner tous ses facteurs premiers". Une question particulière est une **instance**. Par exemple, "donner tous les facteurs du nombre 15" est une instance du problème FACTORISER.

Un **problème de décision** est un problème dont la réponse est oui ou non. Les problèmes de décision sont souvent étudiés car de nombreux problèmes peuvent être réduits à des problèmes de décision.

Une **réduction** est la transformation d'un problème en un autre problème. Selon le type de transformation considérée, le concept de réduction peut être utilisé pour définir une classe de complexité pour un ensemble de problèmes.

La théorie de la **complexité algorithmique** s'intéresse à l'estimation de l'efficacité des algorithmes. Elle s'attache à la question : entre différents algorithmes réalisant une même tâche, quel est le plus rapide et dans quelles conditions ?

Pour qu'une analyse ne dépende pas de la vitesse d'exécution de la machine ni de la qualité du code produit par le compilateur, il faut utiliser comme unité de comparaison des « opérations élémentaires » en fonction de la taille des données en entrée. On évalue le nombre d'opérations élémentaires en fonction de cette taille : si n est la taille, on calcule une fonction $t(n)$.

La **complexité en temps d'un algorithme** est le nombre d'étapes qu'il est nécessaire de faire pour résoudre une instance du problème, en fonction de la taille de l'entrée.

La **complexité en temps d'un problème** est la complexité en temps de l'algorithme le plus efficace pour le résoudre.

Une **classe de complexité** est un ensemble qui regroupe des problèmes

de même complexité. Une classe de complexité a typiquement la définition suivante : l'ensemble des problèmes qui peuvent être résolus par une machine abstraite M en utilisant $O(f(n))$ unités d'une ressource R, n étant la taille de l'entrée.

P est la classe de complexité contenant les problèmes de décision qui peuvent être résolus par une machine de Turing déterministe en utilisant une quantité de temps polynomiale, ou en temps polynomial.

NP ("Non-deterministic Polynomial time") est la classe de complexité contenant les problèmes de décision qui peuvent être résolus en temps polynomial par une machine de Turing non-déterministe. De manière équivalente, on peut dire que cette classe contient l'ensemble des problèmes dont la solution peut être vérifiée en temps polynomial sur une machine de Turing déterministe.

La relation entre les classes de complexité P et NP est une question ouverte en informatique théorique. Il est généralement admis que c'est le problème ouvert le plus important de cette discipline.

En essence, la question "est-ce que P = NP ?" demande : si les solutions positives à un problème de décision peuvent être vérifiées rapidement (où rapidement signifie "en temps polynomial"), est-ce que les solutions peuvent aussi être calculées rapidement ?

Les problèmes **NP-durs** sont les problèmes auxquels on peut réduire en temps polynomial n'importe quel problème de NP.

Les problèmes **NP-complets** sont les problèmes qui sont NP-durs et qui sont dans NP. Pour prouver qu'un problème A de NP est un problème NP-complet, il suffit de montrer qu'il existe une réduction polynomiale d'un problème NP-complet déjà connu à A.

Les problèmes NP-complets peuvent être vus comme les problèmes les plus durs de NP, dans le sens que ce sont ceux qui ont le moins de chance d'être dans P.

Annexe C

Articles

Cette annexe contient 4 publications : la première correspond aux travaux présentés dans les chapitres 3 et 4 et les trois autres correspondent à des sujets connexes au sujet de la thèse mais qui ne sont pas détaillés dans le manuscrit.

Le premier article traite de la modélisation de la notion de motif dans le contexte des réseaux métaboliques. Une définition est proposée et le problème du recherche de motif est étudié (sa complexité est caractérisée et un algorithme exact est donné). Une application à l'étude de l'évolution des voies métaboliques est proposée.

Le second traite du problème d'énumération des modes élémentaires dans un réseau métabolique. Plusieurs résultats de complexités sont donnés, concernant le problème d'identification et de comptage de mode élémentaire. Ce travail est en cours de soumission.

Le troisième traite de la visualisation de réseaux métaboliques. Un algorithme de dessin est proposé, qui prend en compte la structuration du réseau en voies métaboliques mais qui conserve la topologie vraie du graphe (*i.e.* les noeuds ne sont pas dupliqués). Ce travail a été publié dans la revue *BMC Systems Biology*.

Le quatrième traite d'un modèle de graphe aléatoire qui généralise le modèle de Erdös-Rényi. On montre que ce modèle permet de modéliser fidèlement certaines caractéristiques essentielles de la structure d'un réseau métabolique. Ce travail a été présenté à la conférence RIAMS.

Motif Search in Graphs: Application to Metabolic Networks

Vincent Lacroix, Cristina G. Fernandes, and Marie-France Sagot

Abstract—The classic view of metabolism as a collection of metabolic pathways is being questioned with the currently available possibility of studying whole networks. Novel ways of decomposing the network into modules and motifs that could be considered as the building blocks of a network are being suggested. In this work, we introduce a new definition of motif in the context of metabolic networks. Unlike in previous works on (other) biochemical networks, this definition is not based only on topological features. We propose instead to use an alternative definition based on the functional nature of the components that form the motif, which we call a *reaction motif*. After introducing a formal framework motivated by biological considerations, we present complexity results on the problem of searching for all occurrences of a reaction motif in a network and introduce an algorithm that is fast in practice in most situations. We then show an initial application to the study of pathway evolution. Finally, we give some general features of the observed number of occurrences in order to highlight some structural features of metabolic networks.

Index Terms—Reaction motif, network motif, metabolic network, combinatorics, graph algorithms, subgraph isomorphism, evolution, leucine biosynthesis.

1 Introduction

Network biology is a general term for an emerging field that concerns the study of interactions between biological elements [1]. The term *molecular interaction networks* may designate several types of networks, depending on the kind of molecules involved. Classically, one distinguishes between gene regulatory networks, signal transduction networks, and metabolic networks. Protein-protein interaction networks represent yet another type of network, but this term is rather linked to the techniques (such as Yeast-2-hybrid) used to produce the data and possibly covers several biological processes (including, for example, the formation of complexes and phosphorylation cascades) [16].

One of the declared objectives of network biology (or systems biology in general) is whole cell simulation [8]. However, studying the dynamics of a network requires knowledge of reaction mechanisms such as the kinetic parameters describing a Michaelis-Menten equation. Besides the fact that such knowledge is often unavailable or unreliable, the study of the static set of reactions that constitute a biochemical network is equally important, both as a first step toward introducing dynamics and in itself. Indeed, such a static set represents not what is happening at a given time in a given cell but, instead, the capabilities of the cell, including capabilities the cell does not use. A careful analysis of this set of reactions for a given organism, alone or in comparison with the set of other organisms may also help in arriving at a better understanding of how metabolism evolves. It is this set we propose to study in this paper. More precisely, in the following sections, the term "metabolism" should be understood as the static set of reactions involved in the synthesis and degradation of small molecules. Regulation information is not taken into consideration for now. It may be added in a later step, as the "software" running on the "hardware" of a metabolic network [15].

A major issue concerning the study of biochemical networks is the problem of their organization. Several attempts have been made to decompose complex networks into parts. These "parts" have been called modules or motifs, but no definition of such terms seems to be completely satisfying.

Modules were first mentioned by Hartwell et al. [5] who outlined the general features a module should have but provided no clear definition for it. In the context of metabolic networks, a natural definition of modules could be based on the decomposition of a metabolic network into the metabolic pathways one can find in databases: Modules would thus be the pathways that have been established. The advantage of this definition of module is that it reflects the way metabolism has been discovered experimentally (starting from key metabolites and studying the ability of an organism to synthesize or degrade them). The drawback is that it is not based on objective criteria and, therefore, is not universal

- *V. Lacroix is with the Équipe BAOBAB, Laboratoire de Biométrie et Biologie Evolutive (UMR 5558);CNRS; Univ. Lyon 1, 43 bd 11 nov, 69622, Villeurbanne Cedex, France. He is also with the Projet Helix, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France. E-mail: lacroix@biomserv.univ-lyon1.fr.*
- *C.G. Fernandes is with the Instituto de Matemática e Estatística, Universidade de São Paulo, Rua do Matão, 1010, Cidade Universitária, CEP 05508-090, São Paulo, Brazil. E-mail: cris@ime.usp.br.*
- *M.-F. Sagot is with the Équipe BAOBAB, Laboratoire de Biométrie et Biologie Evolutive (UMR 5558);CNRS; Univ. Lyon 1, 43 bd 11 nov, 69622, Villeurbanne Cedex, France. She is also with the Projet Helix, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France and Department of Computer Science, King's College London, Strand, London WC2R 2LS, England. E-mail: sagot@biomserv.univ-lyon1.fr.*

Manuscript received 13 Feb. 2006; revised 30 May 2006; accepted 6 June 2006; published online 31 Oct. 2006.
For information on obtaining reprints of this article, please send e-mail to: tcbb@computer.org, and reference IEEECS Log Number TCBBSI-0013-0206.

(indeed, the number of metabolic pathways and the frontiers between them varies from one database to the other).

Several attempts to give systematic and practical definitions have been made using graph formalisms [4], [9], [14] and constraint-based approaches [10]. Graph-based methods range from a simple study of the local connectivity of metabolites in the network [14] to the maximization of a criterion expressing modularity (number of links within modules) [4]. The only information used in these methods is the topology of the network. In the case of constraint-based approaches, the idea is quite different. First, a decomposition of the network into functional sets of reactions is performed (by analysis of the stoichiometric matrix [11]) and then modules are defined from the analysis of these functional states. The result is not a partition in the sense that a single reaction might belong to several modules.

Unlike the definition of module, the notion of motif has not been studied in the context of metabolic networks. In general, depending on what definition is adopted for modules and motifs, there is no clear limit between the two notions besides the difference in size. In the context of regulatory networks, motifs have been defined as small, repeated, and perhaps evolutionary conserved subnetworks. In contrast with modules, motifs do not function in isolation. Furthermore, they may be nested and overlapping [24]. This definition refers to general features that regulatory motifs are believed to share, but it provides no practical way to find them. A more practical definition has been proposed, still in the context of gene regulatory networks (and other types of nonbiological networks such as the Web or social networks). These are "network motifs" and represent patterns of interconnections that recur in many different parts of a network at frequencies much higher than those found in randomized networks [17]. This definition is purely topological and disregards the nature of the components in a motif. It assumes that the local topology of the network is sufficient to model function (which is understood here as the dynamic behavior of the motif). This assumption seems acceptable when studying the topology of the Internet and may also hold when analyzing gene regulatory networks, but it appears to not be adapted to metabolic networks. In a static context, a topological definition of motif indeed seems inappropriate as similar topologies can give rise to very different functions.

In the definition of motif that we introduce, the components of the network play the central part and the topology can be added as a further constraint only. This is the main biological contribution of this paper.

Its main algorithmical contribution comes from the fact that the definition of motif we adopt leads to new questions. Indeed, if searching for "purely" topological motifs may be formally modeled as a subgraph isomorphism problem, this no longer applies when searching for motifs where the features describing the components are the important elements and topology is initially indifferent (connectivity only is taken into account). Besides, the problem we address is also different from pathway alignment because we wish to go beyond the notion of pathway in order to study the network as a whole, therefore leading to noticeable differences in the modeling decisions. Indeed, in [20] and [12], the pathways are modeled as, respectively, chains and trees to simplify the problem. This simplification may seem reasonable when modeling pathways, but it is no longer so in the case of general networks.

The paper addresses complexity issues related to this new definition of a graph motif, providing hardness results on the problem, and then presents an exact algorithm that is fast in practice for searching for such motifs in networks representing the whole metabolism of an organism. The paper ends with two applications of the algorithm: one concerning the formulation of hypotheses on the evolution of pathways, the other related to the study of some structural features of metabolic networks.

This paper is an extended version of a previous paper presented at WABI in 2005.

2 Preliminaries

2.1 Data

The metabolic network analyzed in this work was obtained from the PATHWAY database from KEGG [7]. Data describing reactions, compounds, and enzymes were downloaded and stored locally using a relational database management system (postgreSQL). The KEGG database contains metabolic data concerning 209 sequenced organisms. The network we built from such data is therefore a consensus of our current knowledge on the metabolisms of all those organisms. As a consequence, sequences of reactions present in the network may have been observed in no organism. To avoid this configuration, one can "filter" the consensus network by an organism of interest, keeping in only the data set reactions catalyzed by enzymes the organism is considered to be able to synthesize. We adopt a different strategy by choosing to perform our motif search on the consensus network and possibly filtering the results in a second step, allowing for easier comparative analysis among organisms.

Moreover, we use additional information present in KEGG: the notion of primary/secondary metabolites. Indeed, in the KEGG reference pathway diagrams (maps), only primary metabolites are represented and connect reactions together, whereas secondary metabolites are not drawn (even though they participate in the reaction). A typical example of a secondary metabolite is the ATP molecule in an ATP-consuming reaction. (Observe that, unlike the notion of ubiquitous compound [14], the notion of the primary/secondary metabolite is relative to a reaction.) Keeping all metabolites in the network leads to the creation of artifact links between reactions and the bias introduced can lead to inaccurate results such as considering metabolic networks as small-world networks as shown in [2]. Withdrawing secondary metabolites may not be the best strategy to adopt, but it represents a simple way of avoiding this bias.

2.2 Graph Models

Several formal models have been used to study metabolic networks. The choice of a formal model seems to depend mainly on the nature of the hypotheses one wishes to test (qualitative or quantitative, static or dynamic) and on the size of the network under study. Differential equations seem well adapted to studying the dynamic aspects of very

small networks, whereas graphs enable the static study of very large networks.

Between these two ends of the spectrum, semiquantitative models have been proposed. For example, Petri nets allow for the simulation and dynamical analysis of small networks [22], while constraint-based models provide a mathematical framework enabling us to decompose the network into functional states starting only from information on stoichiometry and making the assumption that the network is at steady-state [11].

As our goal is to deal with large networks and work with the least possible a priori, graph models seem appropriate. In previous genome-scale studies [6], graphs have been used mainly for topological analyses regardless of the nature of their components (reactions, compounds, and enzymes). We propose enriching the graph models and taking into consideration some of the features of such components.

Formally, a graph G is defined as a pair (V, E), with V a set of *vertices* and $E \subseteq V \times V$ a set of *edges*. The edges represent the relations between the vertices and may be directed or undirected. The vertices and edges of the graph can be labeled.

The most intuitive graph representation of a metabolic network is provided by a bipartite graph. A bipartite graph has two types of vertices which, in the context of metabolic networks, represent, respectively, reactions and chemical compounds, and edges only between these compounds. The compound graph is a compact version of the bipartite graph where only compound vertices are kept and information on the reactions is stored as edge labels. The reaction graph is the symmetric representation of a compound graph (i.e., reaction vertices are kept and information on the compounds is stored as edge labels). Directed versions of these graphs can be drawn expressing the irreversibility of some reactions. The information concerning the reversibility of reactions is generally not well-known. Indeed, contradictions may be found within the same database. We therefore consider this information as uncertain and, in an initial step, assume that all reactions are reversible. This apparently strong hypothesis seems preferable to considering a reaction as irreversible when it actually is reversible (leading to a loss of information).

For practical reasons, the algorithm developed uses the reaction graph as model for the metabolic network. In the following sections, the graph $G(V, E)$ we consider will be undirected and labeled at the vertices. Vertices of G may be labeled by more than one element from a finite set C whose elements we refer to as colors. In the context of metabolic networks, these colors correspond to reaction types. Details will be given in Section 2.5.

2.3 Motif Definition

We define a motif using the nature of the components it contains.

Definition of a motif. *A* motif *is a multiset of elements from the set C of colors.*

As mentioned earlier, we choose in this definition to not introduce any constraint on the order of the reactions or on the topology. This choice is motivated by the wish to explore the network with the least possible a priori information on what we are searching for. Topology and order of the reactions can be used later as further constraints. The advantage of this strategy is that the impact of each additional constraint can then be measured.

2.4 Occurrence Definition

Intuitively, an occurrence is a connected set of vertices labeled by the colors of the motif. For a precise definition, let R be a set of vertices of G and let M be a motif of the same size as R. Let $H(R, M)$ denote the bipartite graph whose set of vertices is $R \cup M$ and where there is an edge between a vertex v of R and a vertex c of M if and only if v has c as one of its colors.

Definition of an exact occurrence of a motif. *An* exact occurrence *of a motif M is a set R of vertices of G such that $H(R, M)$ has a perfect matching and R induces a connected subgraph of G.*

If one searches for exact occurrences of a motif, the risk is of finding a single occurrence or none [2]. Moreover, since studying the evolution of what the graph G represents is one of our main objectives, it seems relevant to allow for flexibility in the search.

With this in mind, we introduce a function S (detailed later) that assigns, to each pair c_i, c_j in $C \times C$, a score which measures the similarity between c_i and c_j. Two colors are considered similar if this score is superior to a threshold s. We then adapt our definition of exact occurrence by modifying $H(R, M)$ in the following way: There will be an edge between a vertex v in R and a color c in M if and only if there exists a color c' of v such that the value of $S(c', c) \geq s$. Further, we generalize this to the case where the threshold s is different for every element c in M. The latter is motivated by the idea that some elements in the motif we are searching for may be more crucial than others.

Another type of flexibility can then be added that allows for gaps in the occurrences. By this we mean, roughly, allowing the occurrence to have more vertices just to achieve the connectivity requirement. These extra vertices are not matched to the elements of the motif. Two types of control on the number of gaps are considered: local and global. Intuitively, a local gap control policy bounds the maximum number of consecutive gaps allowed between a pair of matched vertices of R. A global control policy bounds the total number of gaps in an occurrence.

This leads to the following definition of an approximate occurrence of a motif, where we denote by G_R the subgraph of G induced by a set R of vertices of G:

Definition of an approximate occurrence of a motif. *Let lb and gb be the local and global gap control bounds and let M be a motif. For each c in M, let s_c be a number. An* approximate occurrence *of M (with respect to lb, gb, and the thresholds s_c) is a set R of vertices which is contained in a set R' of vertices of G that satisfies the following conditions:*

1. *for the bipartite graph $H(M \cup R, E_H)$ with $E_H = \{\{c, v\} \in M \times R |$ there exists a color c' of v such that $S(c', c) \geq s_c\}$ contains a perfect matching;*

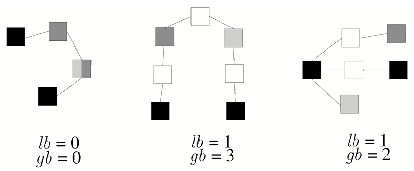


Fig. 1. Subgraphs induced by occurrences for the motif (black, black, dark gray, and light gray).

2. *for each subset B of R such that $B \neq \emptyset$ and $R \setminus B \neq \emptyset$, the length of a shortest path in $G_{R'}$ between B and $R \setminus B$ is at most lb;*
3. $|R'| - |R| \leq gb$.

Observe that, when no gaps are allowed, then $R = R'$ and condition 2 means simply that G_R is connected. An example is given in Fig. 1.

2.5 Colors and Similarity

We now discuss the choice of a set C and a function S for the problem of metabolic networks and reaction motifs in such networks. A major obstacle in this task is that no systematic classification of reactions is available. Instead, enzyme classifications can be used, but a problem one has to face is that several enzymes can catalyze the same reaction (and a single enzyme can catalyze several reactions). This observation explains the necessity of allowing for multiple labeling of nodes in the reaction graph.

We will describe here two possible ways of defining C and S. The first one is based on alignment. Indeed, in order to compare reactions, which is what function S is used for, one can compare the enzymes that catalyze these reactions by performing an alignment of their sequences. An element of C would then be a protein sequence. The function S assigns a sequence alignment score and s is a user-defined threshold that has to be met to consider the sequences similar. The same method can be applied using protein structures instead of sequences, but, in the case of whole networks, sequences are preferable since many structures are not known.

The second example is the one we adopt in this paper. It is based on a hierarchical classification of enzymes developed by the International Union of Biochemistry and Molecular Biology (IUBMB) [23]. It consists of assigning to each enzyme a code with four numbers expressing the chemistry of the reaction it catalyzes. This code is known as the enzyme's EC number (for Enzyme Commission number). The first number of the EC number can take values in [1..6], each number symbolizing the six broad classes of enzymatic activity. (1. Oxidoreductase, 2. Transferase, 3. Hydrolase, 4. Lyase, 5. Isomerase, 6. Ligase.) Then, each of the three remaining numbers of the EC number provides additional levels of detail. For example, the EC number 1.1.1.1 refers to an oxidoreductase (1) with CH-OH as the donor group and NAD+ as the acceptor group.

An element of C is, in this case, an EC number. The function S then assigns a similarity score between two EC numbers that corresponds to the index of the deepest level down to which they remain identical. For example, $S(1.1.1.2, 1.1.1.3) = 3$. Two EC numbers are considered similar if their similarity score is above a user-defined cut-off value s in $[0 \ldots 4]$. The advantage of this definition of similarity between colors, i.e., reaction labels, is that it is more directly linked to the notion of function. Reactions compared with this measure are likely to be functionally related (and possibly evolutionarily related also).

3 Algorithmics

3.1 Hardness Results

The formal problem we address is the following:

Search Problem. Given a motif M and a labeled undirected graph G, find all occurrences of M in G.

As mentioned earlier, this problem is different from subgraph isomorphism because the topology is not specified for the motif.

In the following section, we may assume the graph is connected and all vertices have colors that appear in the motif. Otherwise, we preprocess the graph withdrawing all the vertices having no color appearing in the motif and solve the problem in each component of the resulting graph.

This assumption only holds when no gaps are allowed. Some indications on how to deal with gaps will be given in Section 3.2.2, but the complexity analysis will be done on the exact case.

A natural variant of the Search Problem consists of, given a motif and a labeled graph, deciding whether the motif occurs in the graph or not. As before, we may assume the graph is connected, all vertices are labeled with colors, and all colors appear in the motif. It is easy to see that the decision version of the Search Problem is in NP. We show next that it is NP-complete even if G is a tree, which implies that the Search Problem is NP-hard for trees.

3.1.1 NP-Complete for Trees

We have the following proposition:

Proposition 1. *The decision version of the Search Problem is NP-complete even if G is a tree.*

Proof. We present a reduction from EXACT COVER BY 3-SETS (X3C):

- **INSTANCE:** Set X with $|X| = 3q$ and a collection $\mathcal{C}$ of 3-element subsets of X.
- **QUESTION:** Does $\mathcal{C}$ contain an exact cover for X, i.e., a subcollection $\mathcal{C}' \subseteq \mathcal{C}$, such that every element of X occurs in exactly one member of $\mathcal{C}'$?

Let $X = \{1, , 3q\}$ and $\mathcal{C} = \{C_1, \ldots, C_n\}$ be an instance of X3C. The instance for the decision version of the Search Problem consists of a motif $M = \{Y, B, \ldots, B, 1, \ldots, 3q\}$, where B appears q times in M and a tree T is as follows. (See Fig. 2 for an example.) There are four vertices in T for each i, $1 \leq i \leq n$, three of them are leaves in T, each one labeled by one of the elements of C_i. The fourth vertex, named r_i, is adjacent to the three leaves and has color B. Besides these $4n$ vertices, there is only one more vertex in T, which is labeled Y and is adjacent to each r_i. This completes the description of the instance. Clearly, it has size polynomial in the size of X and $\mathcal{C}$.

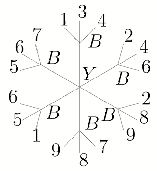


Fig. 2. Tree T and its labels for $X = \{1,\ldots,9\}$ and $\mathcal{C} = \{\{1,3,4\},\{2,4,6\},\{2,8,9\},\{7,8,9\},\{1,5,6\},\{5,6,7\}\}$. For this example, $M = \{Y, B, B, B, 1, \ldots, 9\}$.

To complete the reduction, we need to argue that the motif M occurs in T if and only if there is a subcollection $\mathcal{C}'$ of $\mathcal{C}$ such that each element of X occurs in exactly one member of $\mathcal{C}'$.

Suppose there is such a $\mathcal{C}'$. Clearly, $|\mathcal{C}'| = q$. Let R be the set of vertices of T consisting of the vertex labeled Y and the four vertices of each C in $\mathcal{C}'$. The subgraph of T induced by R is connected. Also, in R, there is a vertex labeled Y, q vertices labeled B (one for each C in $\mathcal{C}'$), and one labeled by each element in X (because of the property of $\mathcal{C}'$). That is, R is an occurrence of M in T.

Now, suppose there is an occurrence of M in T, that is, there is a set R of $1+4q$ vertices of T that induces a connected subgraph of T and has a vertex labeled by each of the colors in M. Let $\mathcal{C}'$ consist of the sets C_i in $\mathcal{C}$ whose vertex r_i in T is in R. Let us prove that each element of X appears in exactly one of the sets in $\mathcal{C}'$. First, note that the vertex labeled Y is necessarily in R because it is the only one labeled Y and there is a Y in M. Then, as R induces a connected graph, a leaf from a set C_i is in R if and only if r_i is also in R. But, R must contain exactly q vertices labeled B. Consequently, $|\mathcal{C}'| = q$ and, as R must contain $1+4q$ vertices, all three leaves of each C in $\mathcal{C}'$ must be in R, and these are all vertices in R. As R must contain a vertex labeled after each element in X, there must be exactly one set in $\mathcal{C}'$ containing each element in X. □

3.1.2 Fixed Parameter Tractability

Finding all occurrences of a motif M in an undirected labeled tree T (Search Problem in trees) is fixed-parameter tractable with a parameter the size of the motif (denoted by k). Indeed, a naive fixed-parameter algorithm consists of generating all possible topologies for the input motif M and then searching for each topology by using a subtree isomorphism algorithm. Since it is enough to generate all possible tree topologies for M, the number of topologies to consider depends (exponentially) on k only and subtree isomorphism is polynomial in the size of both the motif M and the tree T where M is sought. This reasoning is not valid anymore when the motif must be searched in a general graph G as subgraph isomorphism is NP-complete even when the motif is a tree [3].

3.1.3 General Complexity Results

In this paper, we define a motif as a multiset of colors. Other works [17] have defined a motif as a subgraph. In both cases, deciding if a motif occurs in a graph is NP-complete. Nevertheless, using a combination of constraints on the topology and the colors of the motif may, in some cases, lead to polynomial algorithms. In this section, we compare the complexity of several variants of those problems in order to determine which constraint is responsible for NP-completeness.

In the following discussion, we shall first concentrate on the general case where the network is a graph and then comment on the cases where it is a tree.

If we only consider the topology as a constraint (TOPOLOGICAL MOTIFS), all considered variants of the problem are NP-complete. Indeed, subgraph isomorphism is NP-complete even if the motif is a path. This can be shown by a reduction from the Hamiltonian Path. The reduction is based on an instance in which the motif is a path of size n, with n being the number of vertices of the graph in which the motif is being sought.

If we consider both the topology and the colors as constraints (TOPOLOGICAL COLORED MOTIFS), the problem is NP-complete in all cases (reduction from the Hamiltonian Path using the same technique as before), except when the colors are fixed (each node has a color assigned to it) and no repetition of colors is allowed. This last statement is still a conjecture for which we have strong evidence in the case where the topology is a path or a tree and partial results for the case of a general graph. The proof is not trivial and will be the subject of a forthcoming paper.

In the case where the network is a tree, every variant of the problem is polynomial (subtree isomorphism can be solved in polynomial time) except the variants where the topology is not known (COLORED MOTIFS). This result holds even when we do not allow for repetitions of the colors. The latter can be shown using the same reduction as in Section 3.1.1. The only difference is in the building of the instance of the Search Problem. In this case, the motif is defined as $M = \{Y, B_1, B_2, \ldots, B_q, 1, \ldots, 3q\}$ and the tree is the same as in Section 3.1.1 except that each of the n subtrees of four nodes connected to the central node (color Y) now appears in q copies. The copies vary only in the color of their root node, which has color B_j, $1 \leq j \leq q$.

Table 1 summarizes the complexity of the decision version of the Search Problem for the variants mentioned above. Reaction motifs, which are the subject of this paper, fall into the larger category of COLORED MOTIFS, which also includes the case where no color is repeated.

In practice, metabolic networks are graphs and not trees. But, fortunately, they are relatively small (3,184 vertices and 17,642 edges for the network built from the KEGG Pathway database). Even though our initial problem is NP-complete, it is nevertheless conceivable to choose to solve it exactly, provided some efficient pruning is applied. This is described in the next section.

3.2 Exact Algorithm

3.2.1 Version with No Gaps

We now present an exact algorithm which solves the Search Problem. We first explain it for the simple case where the gap parameters lb and gb are set to 0 and then we show how it can be extended to the general case.

Let M be the motif we seek. A very naive algorithm would consist of systematically testing all sets R of

TABLE 1
Complexity Results for the Motif Search Problem

Motif \ Input Graph		Tree	Arbitrary
Topological motifs		polynomial	NP-complete
Coloured topological motifs	General Case	polynomial	NP-complete
	fixed colours and no repetition	polynomial	polynomial (conjecture)
Coloured motifs (this paper)		NP-complete, FPT in k	NP-complete

k vertices as candidates for being an occurrence, where $k = |M|$. For R to be considered an occurrence of M, the subgraph induced by R must be connected and there must be a perfect matching in the bipartite graph $H(R, M)$ that has an edge between $r \in R$ and $c \in M$ if and only if c is similar to one of the colors at vertex r. The search space of all combinations of k vertices among the n vertices in G is huge. We therefore show two major pruning ideas arising from the two conditions that R has to fulfill to be validated as an occurrence of M.

The connectivity condition can be checked by using a standard method for graph traversal, such as breadth first search (BFS) [19]. In our case, a BFS mixed with a backtracking strategy is performed starting from each vertex in the graph. At each step of the search, a subset of the vertices in the BFS queue is marked as part of the candidate set R. The queue, at each step, contains only marked vertices and neighbors in G of marked vertices. Also, there is a pointer p to the last considered vertex in the queue.

At each step, there are two cases to be analyzed: either there are k vertices marked or not. If there are k vertices marked, we have a candidate set R at hand. We submit R to the test of the coloring condition, described below, and we backtrack to find the next candidate set. If there are less than k vertices marked, then there are two possible cases to be analyzed: either p is pointing to the last vertex in the queue or not. If p is not pointing to the last vertex in the queue, we move p one position ahead in the queue, mark the next vertex, and queue its neighbors that are not yet in the queue (checking the latter can be done in constant time by adding a flag to each vertex in the original graph). Then, we repeat, that is, start a new step. If, on the other hand, p is pointing to the last vertex in the queue, then we move p to the last marked vertex (if no such vertex exists, the search is finished) and backtrack. The backtracking consists of 1) unmarking the vertex pointed to by p, 2) unqueuing its neighbors that were added when it was marked, and 3) starting a new step. Next, we describe the test of the coloring condition.

Given a candidate set R, one can verify the coloring condition by building the graph H and checking whether it has a perfect matching or not. In fact, we can apply a variation of this checking to a partial set R, that is, we can, while constructing a candidate set R, check whether the corresponding graph H does or does not have a complete matching. The latter is a matching that completely covers the partial candidate set R. If there is no such matching, we can move the search ahead to the next candidate set. This verification can be done in constant time using additional data structures that are a constant times the size of the motif.

Extra optimizations can also be added. For instance, instead of using every vertex as a seed for the BFS, we can use only a subset of the vertices: those colored by one of the colors from the motif, preferably the less frequent in the graph.

3.2.2 Allowing for Gaps

Allowing for local but not global gaps (i.e., setting $lb > 0$ and $gb = \infty$) can easily be done by performing the lb-transitive closure of the initial graph G and applying the same algorithm as before to the graph with augmented edge set. The p-transitive closure of a graph G for a positive integer p is the graph obtained from G by adding an edge between any two vertices u and v such that the length l of a shortest path from u to v in the original graph satisfies $1 < l \leq p$. The p-transitive closure can be done at the beginning of the algorithm or on the fly. In the latter case, when the next vertex is added to the queue, instead of queuing its neighbors only, all vertices at distance at most $p+1$ from it are queued (if they are not already in the queue), where, by distance between any two vertices, we mean the number of edges in a shortest path between them.

If we allow for global gaps only (that is, setting $lb = \infty$ and $gb > 0$), the problem can be reformulated as the one of finding a minimum Steiner tree for a set of vertices of a graph (for details on the Steiner tree problem, see [13]). The set of vertices would be any set R of vertices of G that match the colors in the motif, that is, that verify the coloring condition. If a minimum Steiner tree for R has size no more than $|R| + gb - 1$, then any such Steiner tree for R is a minimal occurrence of the motif (the minimality being in relation to the total number of gaps in the occurrence). The Steiner tree problem in general graphs is NP-hard, but there are several exact solvers for it.

Here again, a less naive strategy than testing all possible sets R can be adopted. This strategy can follow the lines of the graph traversal with backtracking method described for the case of exact matches. The situation in this case is,

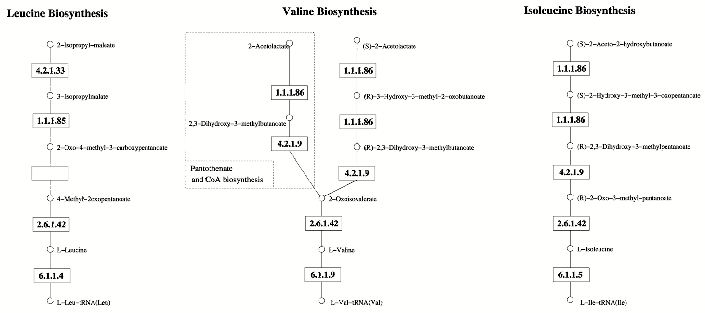


Fig. 3. Bipartite representation of the results obtained when searching for the following motif: $M' = \{1.1.1, 4.2.1, 2.6.1.42, 6.1.1\}$ with local and global gap bounds set to 1. The empty box in the leucine biosynthesis represents a spontaneous reaction.

however, more tricky. The reason is the following: Each vertex v may be considered for incorporation into the queue because it is at a distance d at most gb of different marked vertices. The distance d depends, of course, on the marked vertex. The number of gaps v may add to a candidate set R for a motif thus depends on which marked vertices responsible for the addition of v to the queue are already in R. We must therefore keep track of this information as we add and subtract vertices from the queue (and from R). This information can be kept using a balanced tree of size proportional to $k = |M|$ associated with each queued unmarked vertex v. Each node in the tree corresponds to a marked vertex u that is responsible for the queuing of v and is labeled by the distance from v to u (this distance is at most gb). Keeping, updating, and using the extra information adds a multiplicative term in $O(k \log k)$ to the time complexity of the algorithm, which seems reasonable.

With this system, it now becomes straightforward to take into account both a local and a global gap parameter (that is, to have $lb > 0$ and $gb > 0$). In this case, indeed, one need only to further check that the distance above between v (the unmarked vertex being added to the queue) and u (the marked vertex that led to the queuing of v) is at most $\min\{lb, gb\}$. Another way of proceeding, would consist of 1) finding motifs that verify the local gap parameter and 2) solving a minimum Steiner tree problem in G with the sets R given by the solutions to Step 1.

3.2.3 Time Results

The graph used for the tests has 3,184 vertices and 17,642 edges. For threshold $s = 3$, the number of colors is 171, the mean color frequency is 0.006, and the highest color frequency is 0.089. On average, searching for all occurrences of a motif of size 4 with no gaps takes 8 microseconds of CPU time on a Pentium 4 (CPU 1.70 GHz) with 512 Mb of memory.

4 Application

The approach that we propose and which has been described in the previous sections should enable both generating hypotheses on the evolution of metabolic pathways and analyzing some global features of the whole network.

We start by presenting a case study motivated by trying to understand how metabolic pathways evolve. We do not directly answer this question, which is complex and would be out of the scope of this paper. Instead, we give a first example of the type of evolutionary question people have been asking already and have addressed in different, often semimanual ways in the past [21] and that the algorithm we propose in this paper might help treat in a more systematic fashion.

As in [21], one is often interested in a specific pathway and, for instance, in finding whether this pathway can be considered similar to other pathways in the whole metabolic network, thus suggesting a common evolutionary history. The metabolic pathway we chose as an example is valine biosynthesis. Focusing on the last five steps of the pathway, we derived a motif $M = \{1.1.1.86, 1.1.1.86, 4.2.1.9, 2.6.1.42, 6.1.1.9\}$ and performed the search for this motif initially using a cut-off value s of 4 for the similarity score between two EC numbers (that is, between two reaction labels). With this cut-off value, the motif was found to occur only once (see Fig. 3).

From this strictly defined motif, we relaxed constraints by first lowering the cut-off value s from 4 to 3 and then setting the gap parameters to 1 (motif denoted by M'). Additional occurrences were found. Three of them particularly drew our attention (see Fig. 3).

The first one corresponds to the last five steps of the isoleucine biosynthesis. The second one corresponds to the last five steps of the leucine biosynthesis. Together, they suggest a common evolutionary history for the biosynthesis pathways of valine, leucine, and isoleucine.

An interesting point concerning the second occurrence is the fact that the order of the reactions is not the same as in the other pathways. This occurrence would not have been found if we had used a definition of motif where the order was specified.

Finally, the third occurrence that drew our attention was formed by reactions from both the biosynthesis of valine and a distinct metabolic pathway, namely, the biosynthesis of Panthotenate and CoA. This latter case illustrates a limit of our current general way of thinking about metabolism: Frontiers between metabolic pathways as defined in

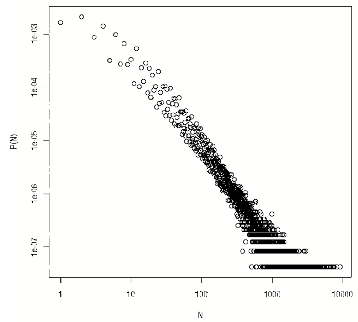


Fig. 4. Log-log representation of the observed number of occurrences for motifs of size $k = 4$ with cut-off value $s = 3$. $P(N)$ is the probability that a motif occurs N times. This probability decays as a power law: $P(N) = N^{-1.6}$.

databases are not tight. If we had taken such frontiers into account, we would not have found this occurrence that overlaps two different pathways. Yet, such an occurrence can be given a biological meaning: It can be seen as a putative alternative path for the biosynthesis of valine.

To complement this analysis, one should add that the results presented in this section hold for 125 organisms in KEGG among which are *S. cerevisiae* and *E. coli*.

4.1 Preliminary Results on Some Global Features of the Network

We now report a few preliminary results concerning some global features of the network. These results are based on the analysis of the number of occurrences found for all possible motifs of fixed size k (we considered $k = 3, 4$) and fixed cut-off value s (we considered $s = 2, 3$). We also report some general properties of the occurrences of the motifs that appear in the network with such parameters.

The number of motifs to be tested is the number of multichoices of k elements among $|C_s|$, where $|C_s|$ is the number of EC numbers with a cut-off value s. For $s = 3$, $|C_s| = 153$ and the number of motifs of size 3 (respectively, 4) is 608,685 (respectively, 23,738,715).

Surprisingly, the distribution of the number of occurrences of motifs of size k is found to be well modeled by a power law (see Fig. 4 for $k = 4$ and $s = 3$). Similar results are obtained for $k = 4$, $s = 2$ and for $k = 3$, $s = 2, 3$ (data not shown). This type of distribution has already been shown to represent a good model for the distribution of the degrees of the vertices in metabolic networks, indicating the existence of a few hubs (highly connected vertices) and a vast majority of poorly connected vertices [6]. What had not been reported before (to the best of our knowledge) is that a similar observation can be made for motifs. Indeed, the distribution of the number of occurrences of motifs of a given size indicates that some motifs are highly repeated in the metabolic network, while many have only a few occurrences. Furthermore, many motifs never occur in the network. Such motifs are naturally absent from a log-log representation, but they correspond to the vast majority of the tested motifs (95 percent in the case of motifs of size 3 and 98 percent for motifs of size 4, even with a cut-off value in both cases of 3).

Observe that, in this paper, we do not discuss over-representation of motifs, which is a difficult theoretical issue in the case of labeled graphs such as metabolic networks. We stress only the fact that, of course, highly represented motifs do not necessarily correspond to over-represented motifs.

Another interesting result that comes out of this systematic study is the number of occurrences that are not included in a single pathway. We call them "interpathway occurrences." As we have shown in the previous application, such occurrences exist. In most cases, they are hard to interpret. but they could correspond to alternative reaction paths that are not visible with the classical representation of metabolism as a collection of disjoint metabolic pathways. We found that, on average, a motif of size 3 (respectively, 4) has 74 percent (respectively, 92 percent) of its occurrences that are interpathway occurrences (always with a cut-off value of 3). All interpathway occurrences may not represent biologically meaningful chemical paths, but the proportions above suggest that a lot of information may be lost when we study pathways and not networks.

5 Conclusion

In this paper, we presented a novel definition of motif, called a "reaction motif," in the context of metabolic networks. Unlike previous works, the definition of motif is focused on reaction labels, while the topology is not specified. Such a novel definition raises original algorithmic issues of which we discuss the complexity in the case of the problem of searching for such motifs in a network. To demonstrate the utility of our definition, we show an example of application to the comparative analysis of different amino-acid biosynthesis pathways. This work represents a first step in the process of exploring the building blocks of metabolic networks. It seems promising in the sense that, with a simple definition of motif, biologically meaningful results are found.

We are currently working on an enriched definition of motif that will take into account information on input and output compounds. The current definition already enables us to discover regularities in the network. Enriched definitions should enable us to test more precise hypotheses.

In this paper, we used a particular formalism for analyzing a metabolic network through the identification of motifs. Other formalisms have been employed or could be considered. As Stelling indicated in his review in 2004 [18], each formalism gives a different perspective and confronting them seems to be a promising way of getting at a deeper understanding of such complex networks.

Availability. A software called MOTUS implementing the work presented in this paper is available upon request to the authors.

Acknowledgments

The authors would like to thank Anne Morgat, Alain Viari, and Eric Tannier for very fruitful discussions. The work presented in this paper was funded in part by the ACI

Nouvelles Interfaces des Mathématiques (project *π-vert*) of the French Ministry of Research, by the ARC (project *IBN*) from INRIA and by the ANR (project *REGLIS*.

REFERENCES

[1] E. Alm and A.P Arkin, "Biological Networks," *Current Opinions on Structural Biology*, vol. 13, no. 2, pp. 193-202, Apr. 2003.
[2] M. Arita, "The Metabolic World of Escherichia Coli Is Not Small," *Proc. Nat'l Academy of Sciences USA*, vol. 101, no. 6, pp. 1543-1547, Feb. 2004.
[3] M.R. Garey and D.S. Johnson, *Computers and Intractability. A Guide to the Theory of NP-Completeness*. Freeman, 1979.
[4] R. Guimerà and L.A. Nunes Amaral, "Functional Cartography of Complex Metabolic Networks," *Nature*, vol. 433, no. 7028, pp. 895-900, Feb. 2005.
[5] L. Hartwell, J. Hopfield, A. Leibler, and A. Murray, "From Molecular to Modular Cell Biology," *Nature*, vol. 402, pp. c47-c52, 1999.
[6] H. Jeong, B. Tombor, R. Albert, Z.N. Oltvai, and A.L. Barabasi, "The Large-Scale Organization of Metabolic Networks," *Nature*, vol. 407, pp. 651-654, 2000.
[7] M. Kanehisa, S. Goto, S. Kawashima, Y. Okuno, and M. Hattori, "The KEGG Resource for Deciphering The Genome," *Nucleic Acids Research*, no. 32, pp. 277-280, 2004.
[8] H. Kitano, "Systems Biology: A Brief Overview," *Science*, vol. 295, pp. 1662-1664, 2002.
[9] H.-W. Ma, X.-M. Zhao, Y.-J. Yuan, and A.-P. Zeng, "Decomposition of Metabolic Network into Functional Modules Based on the Global Connectivity Structure of Reaction Graph." *Bioinformatics*, vol. 20, no. 12, pp. 1870-1876, Aug. 2004.
[10] J.A. Papin, J.L. Reed, and B.O. Palsson, "Hierarchical Thinking in Network Biology: The Unbiased Modularization of Biochemical Networks," *Trends in Biochemical Science*, vol. 29 no. 12, pp. 641-647, Dec. 2004.
[11] J.A. Papin, J. Stelling, N.D. Price, S. Klamt, S. Schuster, and B.O. Palsson, "Comparison of Network-Based Pathway Analysis Methods," *Trends in Biotechnology*, vol. 22, no. 8, pp. 400-405, Aug. 2004.
[12] R.Y. Pinter, O. Rokhlenko, D. Tsur, and M. Ziv-Ukelson, "Approximate Labelled Subtree Homeomorphism," *Proc. 15th Ann. Symp. Combinatorial Pattern Matching (CPM)*, pp. 59-73, 2004.
[13] D. Richards, R. Hwang, and P. Winter, *The Steiner Tree Problem*, 1992.
[14] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar, "Exploring the Pathway Structure of Metabolism: Decomposition into Subnetworks and Application to Mycoplasma Pneumoniae," *Bioinformatics*, vol. 18, no. 2, pp. 351-361, Feb. 2002.
[15] D. Segrè, "The Regulatory Software of Cellular Metabolism," *Trends in Biotechnology*, vol. 22, no. 6, pp. 261-265, June 2004.
[16] P. Shannon, A. Markiel, O. Ozier, N.S. Baliga, J.T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker, "Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks," *Genome Research*, vol. 13, no. 11, pp. 2498-504, 2003.
[17] S.S. Shen-Orr, R. Milo, S. Mangan, and U. Alon, "Network Motifs in the Transcriptional Regulation Network of Escherichia Coli," *Nature Genetics*, vol. 31, no. 1, pp. 64-68, May 2002.
[18] J. Stelling, "Mathematical Models in Microbial Systems Biology," *Current Opinions in Microbiology*, vol. 7, no. 5, pp. 513-518, Oct. 2004.
[19] R.L. Rivest, T.H. Cormen, C.E. Leiserson, and C. Stein, *Introduction to Algorithms*, second ed. MIT Press and McGraw-Hill, 2001.
[20] Y. Tohsato, H. Matsuda, and A. Hashimoto, "A Multiple Alignment Algorithm for Metabolic Pathway Analysis Using Enzyme Hierarchy," *Proc. Int'l Conf. Intelligent Systems and Molecular Biology*, no. 8, pp. 376-383, 2000.
[21] A.M. Velasco, J.I. Leguina, and A. Lazcano, "Molecular Evolution of the Lysine Biosynthetic Pathways," *J. Molecular Evolution*, no. 55, pp. 445-459, 2002.
[22] K. Voss, M. Heiner, and I. Koch, "Steady State Analysis of Metabolic Pathways Using Petri Nets," *Silico Biology*, vol. 3, no. 3, pp. 367-387, 2003.
[23] E.C. Webb, *Recommendations of the Nomenclature Committee of the International Union of Biochemistry and Molecular Biology on the Nomenclature and Classification of Enzymes*. Oxford Univ. Press, 1992.
[24] D.M. Wolf and A.P. Arkin, "Motifs, Modules and Games in Bacteria," *Current Opinions in Microbiology*, vol. 6, no. 2, pp. 125-134, Apr. 2003.

Vincent Lacroix studied at the Institut National des Sciences Appliquées (INSA) in Lyon, France, from which he received the MSc degree in computational biology. He started his PhD work in 2004 at the Laboratoire de Biométrie et Biologie Évolutive (LBBE) under the supervision of Marie-France Sagot. His research subject concerns the combinatorial analysis of metabolic networks.

Cristina G. Fernandes received the BSc and MSc degrees in computer science from the University of São Paulo, Brazil, and the PhD degree in computer science from the Georgia Institute of Technology (Georgia Tech) in 1997. Her research focuses on approximation algorithms, algorithm design and analysis, and complexity theory. Currently, she is an associate professor in the Department of Computer Science at the University of São Paulo, Brazil.

Marie-France Sagot received the BSc degree in computer science from the University of São Paulo, Brazil, in 1991, the PhD degree in theoretical computer science and applications from the University of Marne-la-Vallée, France, in 1996, and the Habilitation from the same university in 2000. From 1997 to 2001, she worked as a research associate at the Pasteur Institute in Paris, France. In 2001, she moved to Lyon, France, as a research associate at INRIA, the French National Institute for Research in Computer Science and Control. Since 2003, she has been Director of Research at INRIA. Her research interests are in computational biology, algorithmics, and combinatorics.

Modes and Cuts in Metabolic Networks: Complexity and Algorithms*

Flavio Chierichetti[1], Vincent Lacroix[2,3], Alberto Marchetti-Spaccamela[1], Marie-France Sagot[2,3], Leen Stougie[4,5]

[1] Universitá di Roma "La Sapienza", Via Eudossiana 18, 00184 Rome, Italy
[2] Équipe BAOBAB ; Université de Lyon ; université Lyon 1 ; CNRS ; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France.
[3] Projet Helix, INRIA Rhône-Alpes, 655 avenue de l'Europe, 38330 Montbonnot Saint-Martin, France
[4] Eindhoven University of Technology, P.O. Box 513, 5600MB Eindhoven, The Netherlands
[5] Centrum voor Wiskunde en Informatica, Kruislaan 413, 1098SJ Amsterdam, The Netherlands

Abstract

Constraint-based approaches recently brought new insight into our understanding of metabolism. By making very simple assumptions such as that the system is at steady-state and some reactions are irreversible, and without requiring kinetic parameters, general properties of the system can be derived. A central concept in this methodology is the notion of an elementary mode (EM for short). The computation of EMs still constitutes a limiting step in metabolic studies and several algorithms have been proposed to address this problem leading to increasingly faster methods. However, although a theoretical upper bound on the number of elementary modes that a network may possess has been established, surprisingly, the complexity of this problem has never been systematically studied.

In this paper, we give a systematic overview of the complexity of optimisation problems related to modes. We first establish results regarding network consistency. Most consistency problems are easy, i.e., they can be solved in polynomial time. We then establish the complexity of finding and counting elementary modes. We show in particular that finding one elementary mode is easy but that this task becomes hard when a specific EM (*i.e.* an EM containing some specified reactions) is sought. We also examine a number of EM related problems and establish their complexity. We emphasize that the easy problems can be solved using currently existing software packages.

We then analyse the complexity of a closely related task which is the computation of so-called minimum reaction cut sets and we show that this problem is hard. We then present two positive results which both allow to avoid computing EMs as a prior to the computation of reaction cuts. The first one is a polynomial approximation algorithm for finding a minimum cut set. The second one is a test for verifying if a set of reactions constitutes a reaction cut; this test could be readily included in existing algorithms to improve their performance. Finally, we discuss the complexity of other cut-related problems.

1 Introduction

Metabolism is usually defined as the union of two processes: anabolism (synthesis of molecules through the use of energy and reducing power) and catabolism (degradation of molecules yielding energy and reducing power). From a modeller's perspective, metabolism can be seen as a network of interconnected reactions, each reaction corresponding to the transformation of metabolites into other metabolites. This network can then be studied either from a structural perspective, or from a dynamic perspective.

Studying the dynamics of metabolic networks is usually performed using models based on differential equations whereas structural analyses are mainly based on graph-related formalisms or, as far as metabolism

*Research was partially supported by the Dutch BSIK-BRICKS project, by the Dutch mathematics cluster DIAMANT, the MRT Network ADONET of the European Community (MRTN-CT-2003-504438), the Project II-Vert of the ACI Nouvelles Interfaces des Mathématiques of the French Ministry of Research, the Project IBN of the Actions Concertées Coopératives of the INRIA, the Project REGLIS of the French Agence Nationale de la Recherche, the EU FET Integrated Project AEOLUS, IST- 15964, and the EU COST 293 project GRAAL

is concerned, on a constraint-based modelling. The latter term is commonly employed in the bioinformatics community following two papers by Palsson [25] [7]. In the constraint-based framework, the network may still be modelled as a graph (an edge-labelled hypergraph), but several types of constraints (stoichiometric, thermodynamic and in some cases regulatory) are added to restrict the possible fluxes through the network. The choice of a particular model heavily depends on the type of question one wishes to address (structural or dynamic) but also on the type of data that is available (qualitative or quantitative). Another type of criterion that may be taken into account is the computational cost of a given analysis, and therefore its scalability to large datasets (such as genome-scale metabolic networks).

In a constraint-based approach, only admissible flux distributions are of interest. An admissible flux distribution corresponds to a set of reactions, which, when taken together in given proportions, perform the transformation of available substrates into removable products with the special property that all intermediate compounds are balanced (steady-state assumption) and irreversible reactions are taken in the appropriate direction (thermodynamic constraint). Such an admissible flux distribution is called a mode.

Even though each mode is potentially interesting, not all of them are generally considered. Classically, two major sub-problems have been introduced. The first one is known as flux balance analysis. It consists in searching for a mode that optimises a given objective function. Examples of objective functions include biomass (usually represented as a pseudo-reaction of the network, in general determined from experimental data) or ATP production. This optimisation problem has several applications [10, 11] and can be solved using linear programming (LP).

The second sub-problem is the one we discuss in this paper. In the case where no particular function is to be optimised, all modes are equally interesting. A sensible strategy is then to try to find a set that could generate them all. Such a generating set has been proposed and called the set of elementary modes [30], EM for short. Intuitively, an elementary mode is a special mode that has the property of not containing any other mode.

Elementary modes have been said to represent a formalised definition of a biological pathway. Indeed, a biological interpretation can be given to such flux vectors: a mode is a set of enzymes that operate together at steady state [29] and a mode is elementary when the removal of one enzyme causes it to fail.

The related concept of extreme pathway has also been introduced in the field [27]. Extreme pathways are actually a subset of elementary modes. Both notions coincide in the case where all exchange reactions (reactions connecting some metabolite with the surrounding of the model) are irreversible. For a detailed comparison of both approaches, see [22].

As outlined in [32], the concept of minimal T-invariant used in Petri Nets is also closely related to the concept of elementary mode. Both notions coincide in the case where all reactions are irreversible. For completeness sake, we can also mention that the extreme currents defined by Bruce Clarke [5] also coincide with elementary modes in the irreversible case. Unlike extreme pathways and elementary modes, minimal T-invariants and extreme currents have only been defined in the case of a network of irreversible reactions. Clearly, there are links between the algorithms for enumerating elementary modes and the ones for minimal T-invariants since, as we shall see, they all boil down to enumerating the extreme rays of a convex cone. We will not discuss the techniques in detail here. Interested readers may refer to [6] for algorithms for enumerating minimal T-invariants and to [31, 35, 13] for enumerating elementary modes. More generally, the usefulness of Petri-Net approaches to the study of metabolic pathways is presented in [37].

Another concept we study here is closely related to the notion of elementary mode. This is the concept of a reaction cut set, recently introduced in [20]. In order to avoid any confusion with other types of cuts in graphs or hypergraphs that may be found in the literature (see e.g.[33]), we explicitly choose here to use the term *reaction* cut. An elementary mode may be seen as a set of reactions that, when used together, perform a given task while a minimal reaction cut set is a set of reactions one needs to inhibit to prevent a given task, also called *target reaction*, from being performed. As mentioned in [18], the task to be silenced can be a combination of reactions. Reaction cut sets have been operationally defined as corresponding to a set of reactions whose deletion from the network stops each elementary mode that contains the target reaction(s).

The main contribution of this paper is in giving a systematic overview of the complexity of optimisation problems related to modes. We first establish results regarding network consistency (Section 2.1). Most consistency problems can be solved in polynomial time (are easy). Most, if not all, of these results have been stated before in the literature. It is in fact easy to formulate these problems as LP-problems, which has the side advantage that computer packages are available to solve them.

We then establish the complexity of finding and enumerating elementary modes (Sections 3.1 and 3.2). We show in particular that finding one elementary mode is easy but that this task becomes hard when a specific EM (*i.e.* an EM containing some specified reactions) is sought. We also examine a number of EM related problems and establish their complexity. We emphasize that the easy problems can be solved by existing software.

We then analyse the computational complexity of problems concerning reaction cuts. We prove that finding a *minimum* reaction cut set, one that contains a minimum number of reactions, is hard (Sections 4.1) We then present two positive results which both allow to avoid to compute EMs as a prior to the computation of reaction cuts. The first one (Section 4.2) is a polynomial approximation algorithm for finding a minimum cut set. The second one (Section 4.3 using a result of Section 4.1) is a test for verifying if a set of reactions constitutes a reaction cut; this test could be readily included in existing algorithms for enumerating *minimal* reaction cuts to improve their performance.

2 Modes

In the following, we define more precisely several objects, classically used in constraint-based modelling of metabolic networks.

The **stoichiometric matrix** S of a network is a matrix with n rows and m columns, n being the number of internal metabolites and m the number of reactions. Entry $S(i,j)$ of the matrix takes value k if reaction j produces k units of metabolite i and $-k$ if reaction j consumes k units of metabolite i; otherwise, it takes value 0. The value k corresponds to the stoichiometric coefficient of metabolite i in reaction j. The stoichiometric matrix summarises the structure of the metabolic network.

The set of reactions is partitioned into two subsets: Rev and $Irrev$, the set of, respectively, reversible and irreversible reactions.

A **mode** is a flux vector $v \in \mathbb{R}^m$ such that:

1. $Sv = 0$
2. $v_j \geq 0 \; \forall j \in Irrev$

In [19] it is already observed that standard linear algebra teaches us how to check that $Sv = 0$ in order to decide if $v \geq 0$ is a mode.

We introduce the support of the solution v, denoted by $R(v) = \{j \mid v_j \neq 0\}$, *i.e.*, the set of reactions participating (with non-zero flux) in v.

An **elementary mode** is a vector v that satisfies conditions 1 and 2 and

3. there is no non-trivial admissible flux vector w such that: $R(w) \subset R(v)$.

Modes and elementary modes can be given a geometrical interpretation. Indeed, the set of vectors $\{v \geq 0 \mid Sv = 0\}$ defines a convex cone in the flux space. When all reactions are irreversible, the elementary modes exactly correspond to the extreme rays of this cone. A proof of this fact relies on basic linear algebra. We give it here for completeness sake, since we are not aware that this has ever been proved formally.

Lemma 1. *If all reactions are irreversible, then the set of EMs corresponds one-to-one to the set of extreme rays of the cone* $\{v \geq 0 \mid Sv = 0\}$.

Proof. Clearly the definition of an EM implies that it is an extreme ray of the cone. To see that it is also the other way round, suppose we have two vectors $v \geq 0$ and $v' \geq 0$ with $Sv = 0$ and $Sv' = 0$ such that $R(v') \subset R(v)$. Then clearly v is not an EM and we show that v is neither an extreme ray of the cone.

Let k^* be such that

$$\frac{v'_{k^*}}{v_{k^*}} = \max_{\{k \mid v^k > 0\}} \frac{v'_k}{v_k}.$$

Then

$$v'' = \frac{v'_{k^*}}{v_{k^*}} v - v' \geq 0 \text{ and } Sv'' = 0.$$

Thus v'' is a mode. Moreover,

$$R(v'') \subset R(v) \text{ and } v''_{k^*} = 0.$$

Most importantly,

$$\frac{v'_{k^*}}{v_{k^*}} v = v' + v'',$$

proving the claim. □

Klamt and Gagneur observed [13] that when some reactions are reversible, one can define a pointed cone in a higher dimensional space by representing each reversible reaction by two irreversible reactions, in the obvious way: suppose the reaction r is represented in S by the column s_r then we add the column $s_{\overline{r}} = -s_r$ to S, yielding S^+, and require both v_r and $v_{\overline{r}}$ to be non-negative. The matrix S^+ has extreme rays that consist of those of S and the vectors v with $v_r = v_{\overline{r}} = 1$ and $v_i = 0$ otherwise, corresponding to length-2 cycles consisting of the two reaction making up for a reversible reaction. We can easily detect and simply ignore these length-2 cycles. A consequence of this observation is that we can analyse the complexity and propose algorithms in the irreversible case without loss of generality.

In the other extreme case in which all reactions are reversible ($Irrev = \emptyset$), an elementary mode corresponds to a minimally dependent set of columns of the stoichiometric matrix. Hence the elementary modes are exactly the *circuits* of a *linear matroid* (for definitions of matroids and circuits we refer to [24] or [28]).

From now on we assume that all reactions are irreversible unless explicitly stated otherwise.

2.1 Consistency of the stoichiometric matrix

One of the applications of constraint-based modelling is in checking the consistency of reconstructed metabolic networks [29]. A network is said to be consistent if all its reactions belong to at least some mode, or equivalently, in terms of Petri-net terminology, if the network is covered by T-invariants [16]. When a network is consistent, we say equivalently that its stoichiometric matrix is consistent: the stoichiometric matrix S is *consistent* if $Sv = 0$ has a solution $v_j > 0 \, \forall j$, or equivalently, each reaction is part of some mode (elementary mode).

We give an overview of some problems related to the consistency of stoichiometric matrices. If a matrix S is not consistent, this may indicate a case of incomplete modelling of the metabolic network. In that sense, detecting inconsistency is a valuable tool for finding deficiencies in the metabolic network description.

Theorem 2. *Given a stoichiometric matrix S, checking the consistency of S can be done using LP.*

Proof. Consider the following LP, where we insert a bound on the sum of the values of the v_j's to avoid unboundedness of the problem.

$$\begin{array}{ll} \max & z \\ \text{s.t.} & v_j \geq z \; \forall j \\ & Sv = 0 \\ & \sum_j v_j \leq 1 \end{array}$$

S is consistent if the optimal value is strictly positive, otherwise it is not. □

In case of inconsistency, it is also easy to find a consistent submatrix containing a maximum number of reactions.

Theorem 3. *Given a stoichiometric matrix S, detecting a minimum number of reactions to be deleted to make S consistent can be done using LP.*

Proof. For each reaction h, solve the LP

$$\begin{array}{ll} \max & v_h \\ \text{s.t.} & Sv = 0 \\ & \sum_j v_j \leq 1 \\ & v \geq 0 \end{array}$$

If for reaction h, the optimal value is strictly positive, then h is part of some mode, and one such a mode is given by the optimal solution. Otherwise there is no mode in which reaction h appears. □

Unfortunately, a very practical question, complementary to the previous one, is hard.

Theorem 4. *Given a stoichiometric matrix S, and some other set of reactions represented by a stoichiometric matrix S', find a subset of reactions of S' of minimum cardinality such that the corresponding submatrix added to S yields a consistent matrix is NP-hard.*

Proof. Taking for S an empty matrix and for S' the stoichiometric matrix of the network, the problem is a special case of finding an elementary mode with a minimum number of reactions in its support. NP-hardness of the latter problem will be established in Theorem 7. □

2.2 Difference between hypergraph and stoichiometric matrix

The stoichiometric matrix enables to represent the structure of a metabolic network. In some cases, particularly for visualisation, hypergraphs may also be used. A hypergraph representation of a metabolic network can be done as follows: metabolites are represented as nodes and there is a (directed) hyperedge for each reaction going from its substrates to its products. In fact, this hypergraph can on its turn be represented by its vertex-edge incidence matrix, which is very similar to the stoichiometric matrix; the former matrix has a 1 at each entry where the latter has a positive integer, a -1 where the latter has a negative integer, and their 0 entries coincide.

The hypergraph description does not take into account all parameters of the stoichiometric matrix as can be seen by the following toy example in which two different networks are presented having the same hypergraph description.

Network 1	**Network 2**
input: a,b	input: a,b
output: f	output: f
Reaction 1: a+b $\rightarrow$ c+d	Reaction 1: a+b $\rightarrow$ c+2d
Reaction 2: c+ d $\rightarrow$ f	Reaction 2: c+ 3d $\rightarrow$ f

Observe that the first network is consistent while the second one is not. Therefore, consistency of a network cannot be checked using a hypergraph (regardless of the stoichiometry). Thus, hypergraphs need to be supplemented with weights on vertices-edge combinations if one wants to use them interchangeably with the stoichiometric matrix.

3 Elementary modes

As mentioned above in Section 2, we may see an elementary mode as an extreme ray of the cone $\{v \geq 0 \mid Sv = 0\}$. The solution methods for the easy problems related to finding EMs rely on this equivalence. It is consistent with the observation in [13] that an elementary mode is characterised completely by its set of reactions, *i.e.*, given S and the support $R(v)$ of an elementary mode v, up to scalar multiplication, v is uniquely determined. In this section, we assume consistency of the stoichiometric matrices of the problem instances we consider.

3.1 Finding elementary modes

Surprisingly few results have been established on the complexity of problems concerning detection, counting and enumeration of elementary modes. In their paper, Klamt and Stelling [21] mainly focus on finding an upper bound on the number of elementary modes.

In fact, as mentioned in [12], the complexity of the general problem, given a description of a cone (or polytope) in terms of its facets (inequalities), find a description in terms of (enumerate all) its extreme rays (vertices), as a function of the length of the output (number of rays or vertices) is a long-standing open question in computational geometry.

In this section, we show some difficult aspects of computing elementary modes. In particular, we try to show where the hardness comes from when enumerating elementary modes. We show that the following tasks are easy: finding an EM and finding an EM that contains one specified reaction. However, the following task is hard: finding an EM that contains a specified set of reactions.

As observed already in [19], standard linear algebra teaches us how to check that $Sv = 0$ in order to decide if $v \geq 0$ is a mode. It is also easy to decide if a given mode $v \geq 0$ is an elementary mode by calculating the rank of the submatrix of S consisting of the reaction in the support of v. If this is equal to the rank of S minus 1 the vector v represents an elementary mode [19]. But also finding some EM is easy.

Theorem 5. *Given a stoichiometric matrix S, an elementary mode can be found in polynomial time.*

Proof. We "slice" the cone $\{v \geq 0 \mid Sv = 0\}$ by the inequality $\sum_j v_j \leq 1$ and solve the LP:

$$\begin{array}{rc} \max & z \\ \text{s.t.} & v_h \geq z \\ & Sv = 0 \\ & \sum_j v_j \leq 1 \\ & v \geq 0. \end{array} \tag{1}$$

In case of a consistent matrix, there is an optimal solution which is a non-all-0 vertex of the polytope $\{v \geq 0 \mid Sv = 0, \sum_j v_j \leq 1\}$ satisfying the inequality $\sum_j v_j \leq 1$ with equality. Thus, since any simplex method-based LP-package will indeed output one non-all-0 vertex of the polytope as an optimal solution, let us call this solution v^*, then $\{\lambda v^* \mid \lambda \geq 0\}$ is an extreme ray of the cone $\{v \geq 0 \mid Sv = 0\}$.

□

The optimal solution of the LP in the proof of the lemma gives an elementary mode that contains reaction h. In general, it is easy to detect if there exists a *mode* whose support contains a given set of reactions T_{IN}, and does not contain any of the reactions of another set T_{OUT}: simply add the restrictions:

$$v_j = 0 \; \forall j \in T_{OUT} \tag{2}$$

to LP (1), replace the first restriction of LP (1) by:

$$v_j \geq z \; \forall j \in T_{IN},$$

and check if the optimal solution is positive or 0.

The existence of an *elementary mode* with the same properties for any set T_{IN} is NP-hard in general, which may (partly) explain the difficulties we encounter in enumerating elementary modes.

Theorem 6. *Given a stoichiometric matrix S, sets of reactions T_{IN} and T_{OUT}, deciding if an elementary mode v exists that has positive value in all its coordinates corresponding to T_{IN}, and has value 0 in all its coordinates corresponding to the set T_{OUT} is*
(i) polynomial solvable if $|T_{IN}| = 1$,
(ii) NP-complete in the general case.

Proof. If $|T_{IN}| = 1$ the proof follows from selecting h in the LP (1) as the only reaction in T_{IN}.
NP hardness in the general case is proved by a reduction from Hamiltonian Circuit. Given a directed graph G, for each vertex u in G, create two compounds u_1, u_2 and create a reaction from u_1 to u_2. For each edge (u, w) of G, create a reaction from u_2 to w_1. Choose T_{IN} to be the set of all reactions corresponding to (derived from) vertices in G and $T_{OUT} = \emptyset$. The only elementary mode that contains all the reactions in T_{IN} corresponds to a Hamiltonian circuit and vice versa. □

As we have seen, the problem is easy if $|T_{IN}| = 1$. We can observe that it becomes trivial when $|T_{IN}| \geq rank(S) + 1$. Indeed, according to Lemma 4 in [31], no elementary mode can have as many non-zero elements as that. This leaves an interesting and rather fundamental open problem:

Open problem: What is the complexity of the problem if $|T_{IN}| = k$ for any fixed $k, 1 < k < rank(S) + 1$.

In fact, we conjecture that it is hard already if $|T_{IN}| = 2$. Also the proof of the above theorem leaves open the complexity of the problem if the hypergraph underlying the stoichiometry is acyclic, or if it is known that each elementary mode describes a path in the hypergraph.

Theorem 7. *Given a matrix S and a number k, deciding the existence of an elementary mode with at most k reactions in its support is NP-complete.*

Proof. The proof is a reduction from the NP-complete 3-DIMENSIONAL MATCHING problem (3DM) (see [15]): Given a set of elements $X = \{x_1, \ldots, x_{3n}\}$ and given a collection of 3-element-subsets $\mathcal{S} = \{S_1, \ldots, S_m\}$, does there exist a subcollection of $\mathcal{S}$ of n subsets that covers all elements of X?

For each element and each 3-element set of the 3DM instance, a compound vertex is created. The first reaction is an input reaction that has as output all elements of the 3DM instance; *i.e.*, the first column of the stoichiometric matrix has 1-entries at all element compounds and 0 at all element set compounds. For each 3-element set of the 3DM instance a reaction is created with input the compounds corresponding to the three elements of the set and output the compound corresponding to the 3-element set; *i.e.*, a column in the stoichiometric matrix with -1-entries at the three element compounds, 1 at the element set compound and 0's elsewhere. For each 3-element set there is also an output reaction that has the 3-element set compound as its only input. Finally we choose $k = 2n + 1$.

The vector of reactions which has a 1 at the positions of the first reaction and the two reactions corresponding to each 3-element set of any 3-dimensional matching and 0's elsewhere, clearly forms an EM with $2n + 1$ reactions in its support. On the other hand, any mode must contain the first reaction. Hence, any EM must have a positive value in the first position, and therefore has as output exactly one copy of each element, all of which must have the same value. For every 3-element-set-reaction that we choose, we have to add the corresponding output reaction. Thus to cover all 3n element from the first reaction, we have to choose exactly n reactions that correspond to 3-element sets. Such a set of reactions corresponds to a 3-dimensional matching. □

This theorem shows that finding the *shortest* elementary mode (the one with a minumum number of reactions) is NP-hard. Note that in the theorem, k is considered to be part of the input. For fixed values of k, the problem is trivially solvable in polynomial time by complete enumeration.

As a final example to illustrate the intricacies in detecting elementary modes, we define the notion of a *simple elementary mode* as an elementary mode v such that $\forall j\ v_j \in \{0, 1\}$. The reduction in the proof of Theorem 7 shows that it is hard to find simple elementary modes. Though it is unlikely that any biological relevance will ever be found for the notion of simple elementary mode, the result shows again the subtlety of EM computations, even more so, since the hardness can be extended to any fixed interval of integers.

Corollary 8. *Given a matrix S, deciding the existence of a simple elementary mode is NP-complete.*

3.2 Counting elementary modes

System biologists are interested in enumerating all elementary modes of a metabolic network. Before turning to that problem, we show that merely counting elementary modes is hard. In [21] the authors show that the number of elementary modes can be bounded by $\binom{m}{n+1}$, but they did not give the complexity of computing the exact number.

Counting elementary modes is essentially a problem of counting the rays of a polyhedral cone, which in its turn is equivalent to a problem of counting vertices of a polytope, which is known to be ♯P-complete [9][1] in general. Therefore not surprisingly, counting elementary modes turns out to be ♯P-complete. ♯P-complete is a class of computationally hard counting problems (for precise definitions we refer to [26]).

Theorem 9. *Given a matrix S counting the number of elementary modes is ♯P-complete.*

[1] In fact, [9] only claims NP-hardness, but the proof establishes #P-completeness.

Proof. The proof follows by a reduction from the ♯P-hard problem COUNTING PERFECT MATCHINGS IN A BIPARTITE GRAPH [36]. Given a bipartite graph $G = (U, V, E)$ with two color classes U and V, each of size n, we construct the following hypergraph H. First, we create an input compound vertex s, which we connect to each vertex in U by an ordinary edge, which we direct from s to the U-vertex. We direct all edges of E from U to V. Finally, we create an output compound vertex t which we connect with one hyperedge to all vertices of V, and direct this hyperedge from V into t. This relates in the obvious way to a $\{-1, 0, +1\}$-stoichiometric matrix. It is easy to see that an EM corresponds one-to-one to a perfect matching in G. □

3.3 Enumerating elementary modes

Listing all feasible solutions of a combinatorial problem is a fundamental problem in combinatorics. Typical cases of interest that have been considered in the literature are enumerating the spanning trees of a graph, enumerating the vertices and the facets of a convex polyhedron or an arrangement of hyperplanes given by a system of linear inequalities.

Since the number of feasible solutions to be enumerated may be exponential in the size of the input description the efficiency of an enumeration algorithm is measured in both the input and output sizes (see e.g., [23]). Namely, an enumeration problem is said to be solvable in polynomial total time if the output can be generated in time polynomial in the input and output size. Usually the stricter requirement of *polynomial delay* is required. In this case we require that, given a feasible set of solutions S, the time required for generating a new feasible solution (not in S) is polynomial in the input size. Clearly, if an enumeration problem can be solved with polynomial delay then it is also solvable in polynomial total time.

In case all reactions are reversible, an elementary mode corresponds to a minimally dependent set of columns of the stoichiometric matrix. Hence the elementary modes are exactly the *circuits* of a *linear matroid* (for definitions of matroids and circuits we refer to [24] or [28]). In [4] it has been shown how to enumerate circuits of matroids *with polynomial delay*, *i.e.*, the time needed between the consecutive generation of any two circuits is polynomial in the number of elements in the ground set of the matroid; in our case the number of reactions, columns of the stoichiometric matrix. As a result, circuits of a matroid, hence elementary modes of a completely reversible network, can be enumerated in time polynomial in their number. In fact, the modes of the cone form a linear subspace.

Theorem 10. *In case all reactions in a metabolic network are reversible, the elementary modes can be enumerated with polynomial delay.*

The enumeration task becomes dramatically more difficult if the reactions are irreversible. In this case, the modes of the network form a cone, and the elementary modes are the rays of the cone.

Open question: Can elementary modes can be enumerated with polynomial delay if $Irrev \neq \emptyset$.

Indeed, this touches a basic open problem in computational geometry (see *e.g.* [12]): given a polyhedral description of a cone, can the rays be enumerated with polynomial delay, or the even the weaker question if the description in terms of its rays can be found in time polynomial in the number of rays. The enumeration methods proposed in the literature are all based on the double description method introduced in [12]. The fastest one at this moment is by Terzer and Stelling [34].

4 Reaction cuts

In this section, we focus on Reaction Cut Sets. The notion of minimal cut sets in a reaction graph was first introduced by Klamt and Gilles [20]. The motivation is to study so-called "failure modes" that render the functioning of a given target reaction r^o impossible. A minimal cut set is a set of reactions that cut reaction r^o. Operationally, this has been defined as a set of reactions whose deletion from the network stops each elementary mode that contains r^o.

Before proceeding we mention that the notion of s, t-cut of an hypergraph, i.e., a cut that separates nodes s and t, has been proposed and studied for directed hypergraphs. In [14] it has been observed that finding s, t-cuts in unweighted directed hypergraphs can be done in polynomial time if all hyperdges are defined by

a subset of input nodes and a single destination node; in the context of metabolic networks this would model the situation in which each reaction is irreversible and produces a single metabolite. We also refer to [2] for a survey presentation of related results on directed hypergraphs.

In what follows, we study two problems: finding a reaction cut of minimum cardinality, which we call Min Reaction Cut, and enumerating all minimal reaction cuts. We prove that Min Reaction Cut is APX-hard. For definition of this complexity class we refer to [1]: we observe that APX-hardness implies that there exists a constant c such that finding a solution that is at most a factor c away from the optimum is a NP-hard problem.

Building on results obtained in the previous section, we propose an approximation algorithm. The algorithm runs in polynomial time as it does not require enumeration of all elementary modes containing the target reaction to be cut.

We then notice how as a consequence of Theorem 11 an easy improvement over existing algorithms for enumerating all minimal reaction sets can be obtained.

4.1 Finding minimal reaction cuts

The first basic problem about reaction cuts is recognising them.

Theorem 11. *Given a stoichiometric matrix S, some target reaction r^o, and a subset F of reactions, deciding if F is a reaction cut of r^o can be done using LP.*

Proof. Consider the following LP:

$$\begin{array}{ll} \max & v_{r^o} \\ \text{s.t.} & Sv = 0 \\ & v_j = 0 \ \forall j \in F \\ & \sum_j v_j \leq 1 \\ & v_j \geq 0 \ \forall j \notin F \cup r^o. \end{array}$$

The optimal solution value is positive if and only if F is not a reaction cut of r^o. □

Finding the optimal cut is a lot more difficult.

Theorem 12. Min Reaction Cut *is APX-hard.*

Proof. We first show a reduction from the APX-hard problem Hitting set (see [1]): Given a set of elements $X = \{x_1, \ldots, x_n\}$ and given a collection of subsets $\mathcal{S} = \{S_1, \ldots, S_m\}$, find a minimum cardinality subset of elements $Y \subset X$ such that $S_i \bigcap Y \neq \emptyset \ \forall i = 1, \ldots, m$.

For each element x_j and for each set S_i, we create a compound vertex, which we also denote by x_j and S_i, respectively. To facilitate the exposition we create three additional compounds s, t and t'. As reactions we create the making of s as an input reaction, the consumption of t' as an output reaction and the reaction $1t \to 1t'$ as the target reaction r^o. For each x_j, we create a $1s \to 1x_j$. Similarly, for each S_i, we create a $1S_i \to 1t$. For each set $S_i = \{x_{i_1}, \ldots, x_{i_k}\}$, we create a *composed reaction* with input compounds $1x_{i_1}, \ldots, 1x_{i_k}$ and output compound $1S_i$. Thus, the stoichiometric matrix contains only entries with value -1,0, or $+1$.

To each set $S_i = \{x_{i_1}, \ldots, x_{i_k}\}$ corresponds an elementary mode consisting of the reactions $(s \to x_{i_1}), \ldots, (s \to x_{i_k}), (x_{i_1}, \ldots, x_{i_k} \to S_i), (S_i \to t), (t \to t')$. Indeed, it is easy to check that the vector that assigns a 1 to each of these reactions and a 0 otherwise is indeed a mode. Removing any reaction from this set gives a submatrix which does not have any mode.

Moreover, suppose that some elementary mode would have $v(S_i \to t) = a_i > 0$ and $v(S_j \to t) = a_j > 0$, respectively. Then this mode should also have $v(x_{i_1}, \ldots, x_{i_k} \to S_i) = a_i$ and $v(x_{j_1}, \ldots, x_{j_l} \to S_j) = a_j$, and also $v(t \to t') = a_i + a_j$ and $v(s \to x_\ell) = a_i \ \forall x_\ell \in (S_i \backslash S_j)$, $v(s \to x_\ell) = a_j \ \forall x_\ell \in (S_j \backslash S_i)$, $v(s \to x_\ell) = a_i + a_j$ $\forall x_\ell \in (S_i \cap S_j)$, and $v(s \to x_\ell) = 0$ otherwise. Hence this is the linear combination of two elementary modes of the above type, and therefore by itself not an elementary mode.

Thus, the elementary modes corresponding to the sets of $\mathcal{S}$ are exactly all the elementary modes, and from each of them some reaction must be selected in the reaction cut. Selecting $(s \to x_\ell)$ cuts all the

elementary modes whose corresponding set contains x_ℓ. Thus, the reactions from s to the x's of a hitting set cut all elementary modes. On the other hand, any reaction $(x_{i_1}, \ldots, x_{i_k} \to S_i)$ or $(S_i \to t)$ reaction in the reaction cut can be replaced by one reaction $(s \to x_j)$ (with $x_j \in S_i$), giving another reaction cut. Thus, there exists a minimum reaction cut consisting only of reactions of type $(s \to x_j)$, hence corresponding to a hitting set. This completes the proof of NP-hardness.

To prove APX-hardness, we observe that the reduction is approximation preserving: minimal reaction cuts and hitting sets have a one-to-one correspondence, and that the reduction is indeed an AP-reduction (see for the definition *e.g.* [1]). Since HITTING SET is an APX-hard optimisation problem, this observation implies that MIN REACTION CUT is APX-hard. □

4.2 Approximation algorithm for finding a minimum reaction cut

On the positive side, we design an approximation algorithm for finding minimum reaction cuts, even for a weighted version of the problem. We assume that a weight function w associates to each reaction r a positive weight $w(r)$. Given a stoichiometric matrix S and a weight function w, we are interested in finding a reaction cut F^* of minimum total weight.

The algorithm consists of two phases: in the first phase, a set F of reactions is constructed by starting from the empty set and adding reactions until a reaction cut of the target reaction r^o is obtained. The set F is not necessarily a minimal reaction cut. In the second phase, minimality is obtained by removing reactions from F. The algorithm *Reaction Cut* (RC) is described below.

Given a stoichiometric matrix S and a set of reactions F, we denote by S_F the stoichiometric matrix obtained from S by removing the columns corresponding to all reactions in F; with a slight abuse of notation, we denote the sum of the weights of reactions in a set G by $w(G)$.

Algorithm RC (Reaction Cut)
input:
a stoichiometric matrix S, a weight function w, a reaction r^o to be cut;
phase 1
$F = \emptyset$;
while F is not a reaction cut of r^o
do begin
let C be the set of reactions defining an elementary mode in S_F that includes r^o
let $\bar{w} = \min_{r \in C} w(r)$
for each reaction r in C
do begin
$w(r) = w(r) - \bar{w}$
if $w(r) = 0$ **then** $F = F \bigcup \{r\}$
end
end
phase 2
let $r_1, r_2, \ldots, r_k$ be the reaction in F
for $j = 1$ **to** k **do**
if $F - r_j$ is a reaction cut of r^o **then** $F = F - r_j$
output: F

For the performance analysis of the solution found by the algorithm we exploit the local ratio technique, a general technique for proving performance ratios of approximation algorithms devised in [3]. Translated into terms of weighted reaction cut, it is based on decomposing the weight function associated to each reaction.

Lemma 13. *Let S be a stoichiometric matrix, and let F^*, F_1^* and F_2^* be the minimum reaction cuts of r^o with respect to three different weight functions w, w_1 and w_2, respectively, such that $w(r) = w_1(r) + w_2(r)$ for each reaction r. Then*

$$w(F^*) \geq w_1(F_1^*) + w_2(F_2^*)$$

Proof.

$$w(F^*) = w_1(F^*) + w_2(F^*) \geq w_1(F_1^*) + w_2(F_2^*)$$

□

The local ratio technique has been applied to a number of combinatorial optimisation problems arising in several areas (scheduling, graph, packing, etc.). Inspired by [8], we prove the following theorem.

Theorem 14. *Given a stoichiometric matrix S and a target reaction r^o, Algorithm* Reaction Cut *runs in polynomial time and returns a reaction cut F of r^o such that $w(F) \leq \lambda w(F^*)$, where F^* is the minimum reaction cut of r^o and λ is the maximum number of reactions in an elementary mode in S including r^o.*

Proof. Assume that S contains n reactions. In Phase 1, the algorithm performs the test of checking whether a set of reactions is a reaction cut of x at most n times. It also computes an elementary mode including reaction x for n times at most. Analogously, Phase 2 of the algorithm performs at most n times the test of deciding whether a set is a reaction cut of x. By Theorems 6(i) and 11, it follows that the running time of the algorithm is polynomial.

The proof proceeds by induction on the number of reactions, with the basis of a stoichiometric matrix of only 1 reaction clearly being true. Suppose it is true for n reactions and consider a stoichiometric matrix S with $n+1$ reactions. Let F be the reaction cut set returned by RC.

Let C be the elementary mode detected in the first call on Phase 1 and $\delta = \min_{r \in C} w(r)$.

We define two new weight functions w_1 and w_2:

$$\begin{aligned} w_1(r) &= \delta \text{ if } r \text{ belongs to } C \text{ and } w_1(r) = 0 \text{ otherwise} \\ w_2(r) &= w(r) - w_1(r). \end{aligned}$$

Let F_1^* and F_2^* be minimum reaction cut sets under weight functions w_1 and w_2, respectively. Since $w(r) \geq w_1(r) \geq 0$, we have $0 \leq w_2(r) \leq w(r)$ and, therefore, the conditions of Lemma 13 apply.

Claim 1. $w_1(F) \leq \lambda w_1(F_1^*)$
Observe that $w_1(F_1^*) = \delta$, because for cutting elementary mode C, one reaction of C is sufficient and necessary, while for any other elementary mode, a reaction with weight 0 can be selected in the reaction cut. Moreover, the weight of $w_1(F) \leq m\delta$, where m denotes the number of reactions in C, because all the reactions not belonging to C have cost 0. This together with $m \leq \lambda$ proves the claim.

Claim 2. $w_2(F) \leq \lambda w_2(F_2^*)$
Let F_1 be the set of reactions selected after passing Phase 2 for the first time, *i.e.* the set of reactions that cut C. Notice that in fact F_1 contains one reaction with weight δ only. Let $F_2 = F \setminus F_1$, which is, by definition of the algorithm, the RC solution for the problem with stoichiometric matrix S_{F_1} obtained by deleting the columns of reaction set F_1 from S and weight function w_2. Let $\mathcal{F}_2^*$ be the optimal solution to the latter problem.

Since $w_2(F_1) = 0$, any reaction cut for S_{F_1} w.r.t. w_2 can be supplemented to a reaction cut for S w.r.t. w_2, by adding F_1 at no extra cost, if necessary. In particular, $w_2(F) = w_2(F_1) + w_2(F_2) = w_2(F_2)$, and $w_2(F_2^*) = w_2(\mathcal{F}_2^*)$. Application of the induction hypothesis to the performance of RC to S_{F_1} with weight function w_2 proves that $w_2(F_2) \leq \lambda w_2(\mathcal{F}_2^*)$ and therefore $w_2(F) \leq \lambda w_2(F_2^*)$.

Both claims together with Lemma 13 yields:

$$w(F) = w_1(F) + w_2(F) \leq \lambda w_1(F_1^*) + \lambda w_2(F_2^*) \leq \lambda w(F^*).$$

□

We finally observe that the above result can be easily extended to the case when more than one reaction should be cut. Given S, assume we are interested in finding a cut of reactions $x_1, x_2, \ldots$; two problems arise: we might be interested in either the problem of cutting all reactions $x_1, x_2, \ldots$ or in cutting at least one.

The result of Theorem 14 can be easily extended to both problems above, by adding compounds and reactions to the stoichiometric matrix. Namely, if we are interested in cutting all reactions in $x_1, x_2, \ldots$ we may add one compound y to the output of each reaction x_i, $i = 1, 2, \ldots$ and add a new reaction $\bar{r}$ that

transforms y in an output z. Clearly, cutting $\bar{r}$ requires to cut each reaction in $x_1, x_2, \ldots$. Note that the above transformation might not be feasible because it is not mass balanced; however a slight modification ensures the mass balance and feasibility properties. A similar transformation applies to the problem in which we are interested in cutting at least one reaction in $x_1, x_2, \ldots$.

4.3 Enumerating reaction cuts

Beyond the question of finding a reaction cut, or a minimum reaction cut, the question of enumerating all reaction cuts may also be interesting. As for modes, one can concentrate on minimal sets [20].

Minimality refers to reaction cuts from which no reaction can be removed without destroying the cutting property. Klamt and Gilles [20] propose an algorithm based on enumerating all possible subsets of reactions starting from singleton sets, then all pair sets, then all triples, and so on. For each candidate set F, they propose to test whether all elementary modes are cut by F. Clearly this test is theoretically, and many times also practically, very inefficient. We propose as an alternative to use Theorem 11.

It remains an intriguing open problem if we can do essentially better in case of irreversible reactions. In case all reactions are reversible, a minimal reaction cut is a co-basis of the linear matroid constituted by the columns of the stoichiometric matrix. Bases of matroids, and therefore co-bases of matroids can be enumerated with polynomial delay [17].

5 Conclusion

Elementary modes and minimal reaction cuts are common tools in metabolic network analysis. Their computation is not trivial and poses a computational challenge. Several algorithms have been proposed to bring solutions to this problem but no systematic complexity analysis had been carried out.

We show here that some problems, like checking the consistency of a network, finding one elementary mode or checking that a set of reactions constitutes a cut, are easy problems and we emphasise that "easy" also means that they can readily be solved using existing LP software. It also implies that many problems in flux balance analysis can be done using LP software.

We also prove the hardness of central problems like finding an elementary mode containing a specified set of reactions, counting elementary modes or finding a minimum reaction cut.

Furthermore, we propose an approximation algorithm for computing the minimum reaction cut as well as an improvement for enumerating minimal cut sets. Both results are based on the idea of avoiding to compute the elementary modes for obtaining the reaction cuts.

One may argue that a reaction cut that disables too many elementary modes is not desirable. As an alternative one may therefore be interested in finding a reaction cut which cuts the target reaction but leaves as many elementary modes intact as possible or a reaction cut that leaves some prespecified set of reactions intact. Almost every variation of the minimum reaction cut that emerges in this way is NP-hard.

At present, pathway analysis is still confronted with a problem of scalability to genome-wide models. This paper provides a basis on the complexity of the underlying computational tasks. Such an analysis should help in deciding which tasks can be tackled.

References

[1] G. Ausiello, P. Crescenzi, G. Gambosi, V. Kann, A. Marchetti-Spaccamela, and M. Protasi. *Complexity and approximation – Combinatorial optimization problems and their approximability properties*. Springer-Verlag, Berlin, 1999.

[2] Giorgio Ausiello, Paolo Giulio Franciosa, and Daniele Frigioni. Directed hypergraphs: Problems, algorithmic results, and a novel decremental approach. In Antonio Restivo, Simona Ronchi Della Rocca, and Luca Roversi, editors, *ICTCS*, volume 2202 of *Lecture Notes in Computer Science*, pages 312–327. Springer, 2001.

[3] A. Bar-Noy, R. Bar-Yehuda, A. Freund, J. Naor, and B. Schieber. A unified approach to approximating resource allocation and scheduling. *J. ACM*, 48(5):1069–1090, 2001.

[4] E. Boros, K. M. Elbassioni, V. Gurvich, and L. Khachiyan. Algorithms for enumerating circuits in matroids. In *ISAAC*, pages 485–494, 2003.

[5] B. L. Clarke. Complete set of steady states for the general stoichiometric dynamical system. *J. Chem. Phys.*, 75:4970–4979, November 1981.

[6] J. M. Colom and M. Silva. Convex geometry and semiflows in p/t nets. a comparative study of algorithms for computation of minimal p-semiflows. In *Proceedings of the 10th International Conference on Applications and Theory of Petri Nets*, pages 79–112, London, UK, 1991. Springer-Verlag.

[7] M. W. Covert and B. O. Palsson. Constraints-based models: Regulation of gene expression reduces the steady-state solution space. *J. Theor. Biol.*, 221:309–325, 2003.

[8] C. Demetrescu and I. Finocchi. Combinatorial algorithms for feedback problems in directed graphs. *Inf. Process. Lett.*, 86(3):129–136, 2003.

[9] M. E. Dyer. The complexity of vertex enumeration methods. *Mathematics of Operations Research*, 8:381–402, 1983.

[10] J. S. Edwards, R. U. Ibarra, and B. O. Palsson. In silico predictions of Escherichia coli metabolic capabilities are consistent with experimental data. *Nat Biotechnol*, 19(2):125–130, Feb 2001.

[11] Stephen S Fong and Bernhard Palsson. Metabolic gene-deletion strains of Escherichia coli evolve to computationally predicted growth phenotypes. *Nat Genet*, 36(10):1056–1058, 2004.

[12] K. Fukuda and A. Prodon. Double description method revisited. In *Combinatorics and Computer Science*, number 1120 in Lecture Notes in Computer Science, pages 91–111, 1996.

[13] J. Gagneur and S. Klamt. Computation of elementary modes: a unifying framework and the new binary approach. *BMC Bioinformatics*, 5:175, 2004.

[14] Giorgio Gallo, Claudio Gentile, Daniele Pretolani, and Gabriella Rago. Max horn sat and the minimum cut problem in directed hypergraphs. *Math. Program.*, 80:213–237, 1998.

[15] M. R. Garey and D. S. Johnson. *Computers and Intractability. A Guide to the Theory of NP-Completeness*. Freeman, 1979.

[16] M. Heiner and I. Koch. Petri net based system validation in systems biology. In *Proc. ICATPN*, volume 3099 of *LNCS*, pages 216–237, Bologna, June 2004.

[17] L. G. Khachiyan, E. Boros, K. M. Elbassioni, V. Gurvich, and K. Makino. On the complexity of some enumeration problems for matroids. *SIAM J. Discrete Math.*, 19(4):966–984, 2005.

[18] S. Klamt. Generalized concept of minimal cut sets in biochemical networks. *Biosystems*, 83(2-3):233–247, 2006.

[19] S. Klamt, J. Gagneur, and A. von Kamp. Algorithmic approaches for computing elementary modes in large biochemical reaction networks. *IEE Proc.-Syst. Biol.*, 152(4):249–255, Dec 2005.

[20] S. Klamt and E. D. Gilles. Minimal cut sets in biochemical reaction networks. *Bioinformatics*, 20(2):226–234, Jan 2004.

[21] S. Klamt and J. Stelling. Combinatorial complexity of pathway analysis in metabolic networks. *Mol Biol Rep*, 29(1-2):233–236, 2002.

[22] S. Klamt and J. Stelling. Two approaches for metabolic pathway analysis? *Trends Biotechnol*, 21(2):64–69, Feb 2003.

[23] Eugene L. Lawler, Jan Karel Lenstra, and A. H. G. Rinnooy Kan. Generating all maximal independent sets: Np-hardness and polynomial-time algorithms. *SIAM J. Comput.*, 9(3):558–565, 1980.

[24] J. G. Oxley. *Matroid theory*. Oxford Science Publications. The Clarendon Press Oxford University Press, New York, 1992.

[25] B. O. Palsson. The challenges of *in silico* biology. *Nat. Biotechnol.*, 18:1147–1150, 2000.

[26] C. H. Papadimitriou. *Computational complexity*. Addison-Wesley, 1994.

[27] C. H. Schilling, D. Letscher, and B. O. Palsson. Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, 203(3):229–248, 2000.

[28] A. Schrijver. *Combinatorial optimization: Polyhedra and Efficiency*, volume 24 of *Algorithms and Combinatorics*. Springer-Verlag, Berlin, 2003.

[29] S. Schuster, D. A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol*, 18(3):326–332, 2000.

[30] S. Schuster and C. Hilgetag. On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, 2:165182, 1994.

[31] S. Schuster, C. Hilgetag, J. H. Woods, and D. A. Fell. Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism. *J Math Biol*, 45(2):153–181, 2002.

[32] S. Schuster, T. Pfeiffer, F. Moldenhauer, I. Koch, and T. Dandekar. Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, 18(2):351–361, Feb 2002.

[33] P.D. Seymour. The matroids with the max-flow min-cut property. *J. Comb. Theory Ser. B*, 23(7):189–222, 1977.

[34] Marco Terzer and Jörg Stelling. Accelerating the computation of elementary modes using pattern trees. In *WABI*, pages 333–343, 2006.

[35] R. Urbanczik and C. Wagner. An improved algorithm for stoichiometric network analysis: theory and applications. *Bioinformatics*, 21(7):1203–1210, 2005.

[36] L. G. Valiant. The complexity of computing the permanent. *Theoretical Computer Science*, 8:189–201, 1979.

[37] K. Voss, M. Heiner, and I. Koch. Steady state analysis of metabolic pathways using Petri nets. *In Silico Biol*, 3(3):367–387, 2003.

BMC Systems Biology

Software

Open Access

Metabolic network visualization eliminating node redundance and preserving metabolic pathways

Romain Bourqui[1], Ludovic Cottret[2], Vincent Lacroix[2], David Auber[1], Patrick Mary[1], Marie-France Sagot[2] and Fabien Jourdan*[3]

Address: [1]LaBRI, Université Bordeaux I, 351 Cours de la libération, 33405 Talence CEDEX, France, [2]BAOBAB Team, Inria Rhône-Alpes, Projet HELIX, Université de Lyon ; université Lyon 1 ; CNRS ; UMR 5558, Laboratoire de Biométrie et Biologie Evolutive, 43 boulevard du 11 novembre 1918, Villeurbanne F-69622, France and [3]UMR1089 Xénobiotiques INRA-ENVT, 180 chemin de Tournefeuille – St-Martin-du-Touch, BP 3, 31931 Toulouse CEDEX, France

Email: Romain Bourqui - bourqui@labri.fr; Ludovic Cottret - cottret@biomserv.univ-lyon1.fr; Vincent Lacroix - lacroix@biomserv.univ-lyon1.fr; David Auber - david.auber@labri.fr; Patrick Mary - mary@labri.fr; Marie-France Sagot - sagot@biomserv.univ-lyon1.fr; Fabien Jourdan* - Fabien.Jourdan@toulouse.inra.fr

* Corresponding author

Published: 3 July 2007

BMC Systems Biology 2007, **1**:29 doi:10.1186/1752-0509-1-29

Received: 17 January 2007
Accepted: 3 July 2007

This article is available from: http://www.biomedcentral.com/1752-0509/1/29

Abstract

Background: The tools that are available to draw and to manipulate the representations of metabolism are usually restricted to metabolic pathways. This limitation becomes problematic when studying processes that span several pathways. The various attempts that have been made to draw genome-scale metabolic networks are confronted with two shortcomings: 1- they do not use contextual information which leads to dense, hard to interpret drawings, 2- they impose to fit to very constrained standards, which implies, in particular, duplicating nodes making topological analysis considerably more difficult.

Results: We propose a method, called MetaViz, which enables to draw a genome-scale metabolic network and that also takes into account its structuration into pathways. This method consists in two steps: a clustering step which addresses the pathway overlapping problem and a drawing step which consists in drawing the clustered graph and each cluster.

Conclusion: The method we propose is original and addresses new drawing issues arising from the no-duplication constraint. We do not propose a single drawing but rather several alternative ways of presenting metabolism depending on the pathway on which one wishes to focus. We believe that this provides a valuable tool to explore the pathway structure of metabolism.

Background

Metabolism visualization for systems biology studies

The scale of metabolic studies varies according to the data and to the biological questions. For instance, toxicologists often follow the degradation of a given molecule; in that case they focus only on a very small number of reactions. At a larger scale, biologists studying glycolysis will focus on this particular metabolic pathway. Most of the work on metabolism visualization has been done at this level of detail [1-12]. However, in order to investigate an organism's metabolic response to stress, it is relevant to study all the pathways simultaneously. For instance, this will be useful for treating the results of high throughput experiments such as transcriptomic data where relevant gene

products are identified in many pathways. Visualization is a suitable and obvious solution to achieve this kind of study, for instance by representing all the metabolic pathways in one drawing and by coloring relevant enzymes and metabolites [13-15]. In [16], the authors use this approach to analyze simultaneously transcriptomic and metabolomic data (they used Biocyc *omics viewer* [14]). Based on this representation, they managed to identify at once perturbations in the Calvin cycle, glycolysis and TCA cycle. Such kinds of studies emphasize the necessity to develop methods that allow to visualize the entire metabolic network in a single drawing.

Highlighting pathways according to experimental data provides some clues on metabolic processes. However, to integrate these conclusions in a systems biology approach, it is necessary to understand how these pathways are linked and how processes span over them.

The issue of analyzing biological processes spanning several metabolic pathways appears in many contexts. As we already mentioned, it appears when analyzing metabolomic or transcriptomic experiments, which are generally not pathway-focused. This issue also arises for topological analyses based on motif detection [17]. A motif (defined as a set of reaction types) may occur in different parts of the network (which illustrates the need to visualize the whole network in a single picture), and each occurrence may be composed of reactions belonging to different pathways (which examplifies the need to explicitly visualize the links between the pathways).

Therefore, pathway visualization is not suitable for such tasks but neither is network visualization without pathway information. Indeed, to be useful for mapping experiments, it is necessary to represent the entire network structure while keeping the contextual information provided by its division into metabolic pathways. Note that this is one of the requirements for biological network visualization proposed in [18]. Recently, in addition to the studies that use the network as a background, great efforts have been devoted to the analysis of the topological properties of metabolic networks [19,20]. Indeed topology could, for instance, give clues on the evolution of the organisms they are related to. More generally, topological features like shortest path, connectivity, node degrees and node/edge metrics have become common investigation tools. To visually retrieve topological information, it is necessary that the drawing provides a faithful image of the network structure. This is a challenging problem which has not been addressed by current metabolic network visualization tools [13,14] which choose to allow node duplication and therefore do not face this issue.

In the case where nodes are not duplicated, pathways which share reactions and compounds cannot all be drawn equally well (a well-drawn pathway being a pathway having all its nodes drawn next to each other). Therefore, choices have to be made on which pathways will be drawn well in priority. We propose both an automatic way of making this choice and possibilities for the user to define his own priorities. This last option adds an interesting feature to the tool: depending on the choices made, the backbone of metabolism (the set of well-drawn pathways) can be adjusted to the pathways one is interested in. This backbone can either include the glycolysis and the TCA cycle as it is traditionnally the case in most drawings or, alternatively, it can include pathways that share compounds or reactions with glycolysis and the TCA cycle and which would, if not chosen, be drawn in the background. Playing around with this option enables to get a grip on the interdependence of the pathways.

The aim of this paper is to propose an algorithm to draw the entire metabolic network. The produced representation will have to follow textbook drawing conventions (see the following section), display information on the metabolic pathways and keep the topology of the network by avoiding node duplication.

Metabolic network drawing and visualization

Drawing metabolic pathways

A metabolic pathway (also called a metabolic map) is a subnetwork of the metabolic network. The decomposition of the entire network into metabolic pathways is generally done according to biological functions: molecule degradation (catabolism), molecule synthesis (anabolism) or energy transfer [21]. Until recently, these pathways have been manually drawn, for instance for teaching purposes, or to exchange results [22,23]. Then, numerical versions of these manual drawings were proposed and used on web servers such as KEGG [3,24].

In the last few years, automatic drawing algorithms have been designed, mainly for two reasons. First the number of organisms for which a metabolic network is described is increasing quickly. Indeed, *in silico* methods have been designed to reconstruct metabolic pathways from annotated genomes [25] which are more and more numerous. Second, these putative networks follow a regular curating process implying many changes in their structures. In this section, we describe the algorithms that have been proposed for drawing metabolic pathways since they could be extended to the entire network.

Because biologists are used to textbook representations, most of the automatic methods consist in following the drawing habits of these representations [22]. Even if there is no standard for these conventions, it is possible to iden-

tify the most commonly used ones. Some of the aesthetic criteria are also used in graph drawing [26-28]: lowering the number of edge crossings and lowering the number of bends on edges. Moreover, the biological nature of pathways implies some conventions. The notion of reaction cascade is central since generally metabolic pathways describe the transformation of input metabolites into output ones. Most automatic drawing algorithms have been designed to emphasize this structure. The algorithm proposed in [5] and implemented in Biominer uses a hierarchical drawing algorithm which embeds nodes on regular horizontal layers [29]. Others propose adapted versions of classical hierarchical drawing algorithms, like in [6] (implemented in BIOPATH [30]) or in [9] (implemented in Wilmascope).

However, these algorithms do not emphasize cyclic patterns which are also relevant (see for instance the TCA cycle). Thus, other methods were designed to take into account these two configurations. The first one was proposed in [4] where the authors introduce a compound graph layout algorithm, that is, they first detect cycles then treat them as metanodes creating a Directed Acyclic Graph (DAG) and applying a hierarchical drawing algorithm on this DAG. In [10], the authors refine the approach by detecting nodes shared by two cycles thus providing two cyclic representations instead of one. Finally, [11] proposed the same kind of approach for signaling pathways, adding the ability to manually constrain the drawing. However, all these algorithms were initially designed to draw pathways and are not well adapted to draw networks. For instance, we tried to use the software SimWiz which implements the algorithm proposed in [4] to draw the metabolic network of *Escherichia coli* but the program failed because the network was too large. We were nevertheless able to draw the metabolic network of *Mus musculus*, which is smaller. The result is shown in figure 1. In this case, the main problem is due to the cycle detection which is applied on the whole network thus highlighting cycles that span over different pathways.

Scaling to the whole metabolic network

In the Graph Drawing community, efficient drawing algorithms have been designed to draw large networks. Among them, force-based layouts [31,32] are commonly used. Such layouts mimic physical systems, that is, nodes are considered as masses (or particles) and edges behave as springs (or magnetic forces). This system evolves from a random embedding to one corresponding to an equilibrium, providing a suitable layout. These algorithms generate quite good drawings since they generally emphasize dense subgraphs and spread low degree nodes on the screen space. They are used in Cytoscape [33] or in the online SBML viewer [34] for instance. However, as mentioned in [18], such drawings are not satisfying for biologists. The first reason is that they do not follow textbook drawing conventions, and the second is that they emphasize topological clusters which generally do not correspond to a metabolic pathway decomposition. To overcome this last problem, force-based methods could be used in a compound graph layout as it is done in [8] (implemented in PatikaWeb [12]). However, this tool is not dedicated to metabolic pathway visualization and thus does not follow all textbook drawing conventions.

The two main efforts for automatically drawing metabolic networks while keeping metabolic pathway information and respecting drawing conventions are: Reactome [13] and the Pathway Tools cellular overview diagram [14]. As it was mentioned before, in both tools nodes are duplicated thus the only drawing problem is to embed metabolic maps. Both achieve it by grouping maps according to their common functions. The latter assumes that a hierarchy on the pathways is given as input to the algorithm and is then used to display pathways close to each other when they are close to each other in the hierarchy. This functionality is not included in the current implementation of our algorithm. Nevertheless, it is still possible to circumvent this problem by redefining coarse-grained pathways (corresponding to groups of pathways of common functions) in the input data.

In the following sections, we first describe our metabolic network drawing algorithm. Then we discuss our approach and compare it to other published methods using the metabolic network of *Escherichia coli* (*E. coli*) as benchmark.

Implementation

Using a mixed bipartite graph to model metabolic networks

A graph provides an intuitive way of organizing large amounts of relational data. The general definition of a graph $G = (V, E)$ is simple. It consists of a set V of n vertices ($|V| = n$) and a set E of m edges, each of which corresponds to a pair-wise relationship between two of the nodes ($E \subseteq V \times V$). Modeling the metabolic network consists in choosing which biological objects are associated to nodes and edges. It is necessary to do this model description before introducing the graph drawing algorithm, since it will constrain the representation. For instance, a model may imply that some nodes have a high degree, thus complicating a planarization process.

Bipartite graph

A metabolic network is a set of biochemical reactions (*i.e.* reactions that convert one or more compounds into one or more other compounds). Different models could be used (for a detailed discussion, see [35]). Here, we consider that there are two kinds of nodes: reactions and sub-

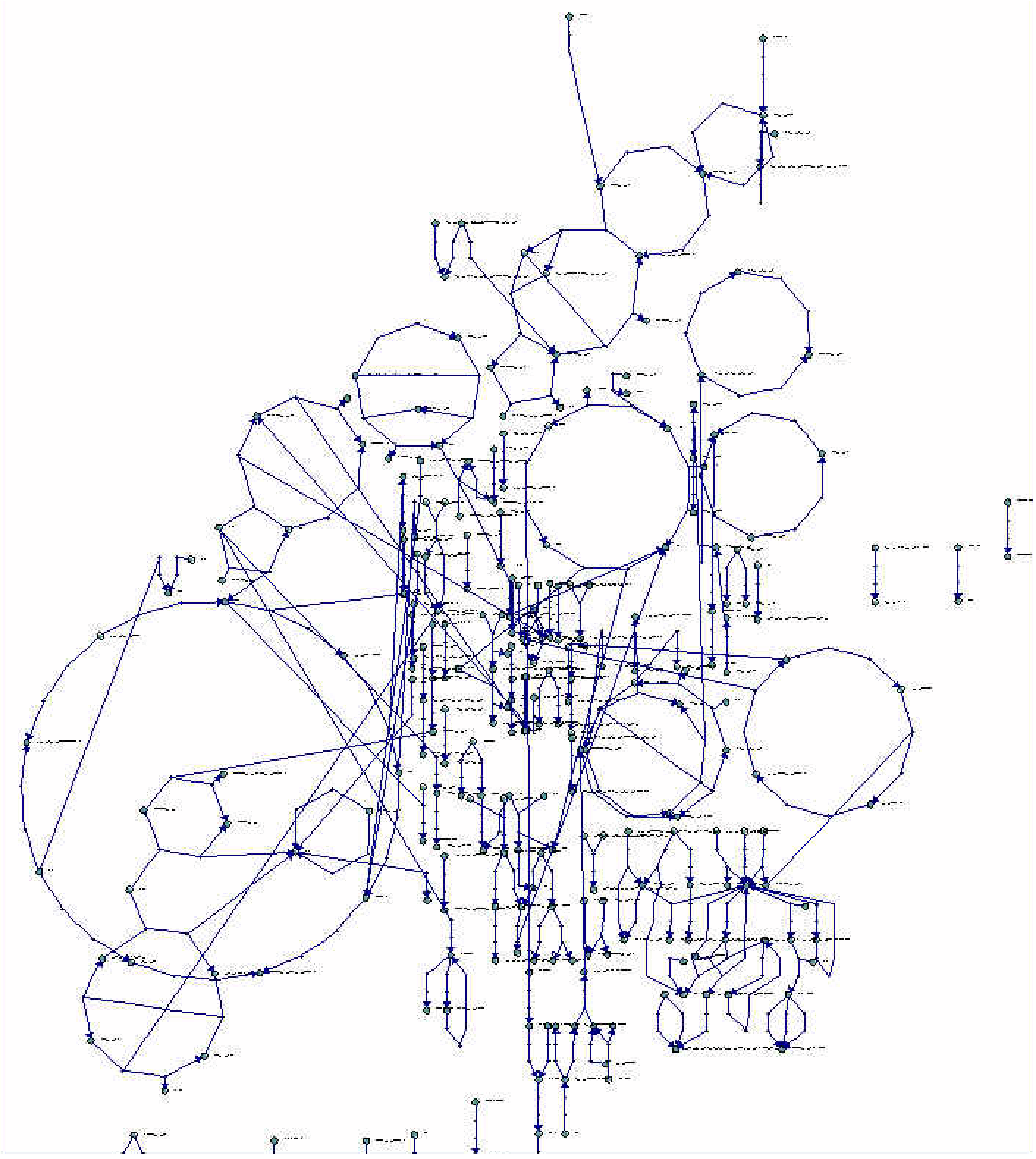


Figure 1
Mus musculus metabolic network drawn using SimWiz software implementing [10] algorithm. This network appears to be smaller than the one of *E. coli*. This is simply due to the fact that our knowledge of mouse metabolism is very partial.

strates (see Figure 2) and that there is an edge between a reaction and a substrate if the substrate is consumed or produced by the reaction. The discussion of this choice is out of the scope of this paper, but the main motivation is due to the use of this model in many textbook drawings. This graph is generally called a *bipartite* graph since its set of nodes can be split into two subsets where the elements are not linked (no link between reactions and no link between substrates). Thus the set of vertices can be split into two subsets $R = \{v \in V \mid v \text{ is a reaction}\}$ and $S = \{v \in V \mid v \text{ is a substrate}\}$, and $V = R \oplus S$ and $E \subseteq \{(u, v) | u \in R, v \in S\} = R \times S$.

Mixed graph

Metabolic reaction can be either reversible (*i.e.* it can occur in both directions) or irreversible (*i.e.* it can occur in only one direction). This orientation is defined according to the physiological properties of a reaction. SBML descriptions of reactions provide this kind of information. In order to model such a biological phenomenon, we use a *mixed* graph. In a mixed graph, the set E of edges is splitted in two subsets A and E', where A is the set of *arcs* (*i.e.* oriented edges), E' is a set of non-oriented edges and $E = A \oplus E'$.

Thus, for modeling the whole network, we use a *mixed bipartite* graph $G = (R, S, A, E')$.

Graph hierarchy

A metabolic pathway is a subnetwork of the metabolic network. Here, it corresponds to a graph $G_p = (V_p, E_p)$ where $V_p \subset V$ and $E_p = \{(u, v) \in E | u \in V_p \text{ and } v \in V_p\} \subset E$ (*i.e.* E_p is the set of edges and arcs induced by V_p on E). For a given metabolic network G, we note $P_G = \{G_i | 1 \le i \le n_p\}$ its n_p metabolic pathways. One can notice that for each G_i, V_i and E_i can be decomposed in four subsets R_i, S_i, A_i and E_i' (*i.e.* G_i is a mixed bipartite graph).

Taking pathways into account leads to the following graph hierarchy : the graph G representing the whole network and n_p induced subgraphs G_i representing its n_p metabolic pathways.

Drawing algorithm

The algorithm we propose has two main steps: first, a multi-scale clustering is performed creating a quotient graph (strictly speaking, the quotient graph is built by considering isolated nodes as singletons), and second, clusters and quotient graph are drawn using three drawing algorithms. In the next section, we first explain our clustering algorithm and then, we present the drawing algorithms we use.

Multi-scale clustering

One of the main problems is that metabolic pathways often share nodes. For instance, in Figure 3, the yellow, blue and purple regions respectively represent pathways p_1, p_2 and p_3. One can see an overlap between p_1 and p_2 (one node) and between p_2 and p_3 (four nodes). This situation is not rare in real networks: in the *E. coli* metabolic network, 658 nodes (out of a total of 1140) are shared between several pathways, and the average number of pathways per node is more than 2.4. Since we choose not to duplicate nodes, and since vertices of a pathway have to be drawn next to each other, our algorithm has to decide whether a node is embedded next to a pathway or next to another. For example, the shared node between p_1 and p_2 could be drawn near p_1 or near p_2. This is achieved by a two-step process. The first step consists in computing an independent set of pathways (*i. e.* a set of pathways which do not share nodes) and the second one in detecting cycles and paths.

First pass : computation of an independent set of pathways

First of all, the algorithm searches for a subset $P_{ind} = \{p_1, ..., p_{ind}\}$, $ind \ge 1$, $P_{ind} \subseteq P_G$ such that 1. the pathways of P_{ind} are independent and 2. $\sum_{i=1}^{i=ind} |p_i|$ is maximized. For instance, in Figure 3a, $\{p_1, p_3\}$ is the independent set that maximizes this sum among all possible independent sets of pathways ($\{p_1\}$, $\{p_2\}$,$\{p_3\}$, $\{p_4\}$, $\{p_5\}$, $\{p_1, p_3\}$, $\{p_1, p_4\}$, $\{p_1, p_5\}$, $\{p_2, p_4\}$ and $\{p_4, p_5\}$).

The problem of finding a maximum independent set is known to be NP-Hard [36]. This problem can be reduced to a coloration problem (the graph is then the dependence graph, where each pathway corresponds to a node and there is an edge between two nodes when the pathways share nodes in the original graph). To find a solution, we use the Welsh and Powel heuristic [37]. Then, for each color class C, $\sum_{p_i \in C} |p_i|$ is computed, and a maximum one is chosen as our independent set.

Let $P_{Nind} = P_G \backslash P_{ind}$. Then, for all the pathways in P_{Nind}, we exclude nodes that are shared with at least one other pathway in P_G. We denote this reduced set by P'_{Nind}.

Each element of P_{ind} and P'_{Nind} is a set of nodes. These sets define a clustering on the original graph since there is no overlapping between them. This clustering is used by replacing each subgraph induced by an element of P_{ind} or

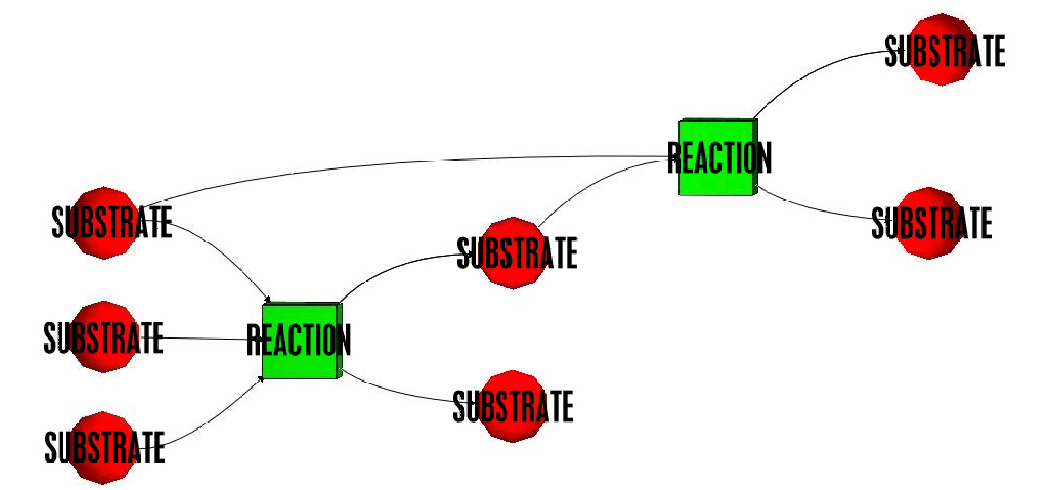


Figure 2
Bipartite graph describing two biochemical reactions.

P'_{Nind} by a metanode representing it (see Figure 3b). We call this first clustered graph G_{clust1}.

For all the pathways in P_{ind} and in P'_{Nind}, we search for the longest independent mixed cycles (Cycles C_1 and C_2 are independent if C_1 and C_2 do not share any node). A mixed cycle is a sequence of nodes $v_1, v_2, ..., v_l$, $l \geq 3$ such that $\forall$ $1 < i \leq l$, $(v_{i-1}, v_i) \in E' \cup A$ and $(v_l, v_1) \in E' \cup A$.

Moreover, $\forall$ $1 < i < l$, if v_i represents a reaction and v_{i-1} a substrate consumed in (resp. produced by) this reaction, then v_{i+1} is produced by (resp. consumed in) v_i. This problem is also NP-Complete even if $A = \varnothing$ [36]. To "solve" it, we use an exact maximum length cycle algorithm and bound the computation time with a threshold. If the threshold is reached, we stop the algorithm and consider that the longest mixed cycle we have already found is a longest one. This allows to have an exact result in the best case and an approximation of a longest mixed cycle otherwise. The technique computes all mixed paths using a *mixed* breadth-first search (BFS). In Figure 3c, one can see the longest independent cycles of each element of P_{ind} and P'_{Nind} highlighted in red. These cycles are clustered into metanodes yielding a multi-scale graph called G_{clust2}. For all the metabolic networks on which we tested our algorithm, the threshold was not reached (*i.e.* we found an exact solution).

Second pass : detection of cycles and paths
The next step of the algorithm consists in computing the longest independent mixed cycles in G_{clust2}, excluding metanodes. At each iteration, we cluster a longest cycle into a metanode and exclude it for the next search. We then compute the longest mixed paths, i.e. the longest sequences of nodes of degree less or equal to two $v_1, v_2, ..., v_l$, $l \geq 2$, where $\forall 1 < i \leq l$, $(v_{i-1}, v_i) \in E' \cup A$.

In figure 3d, one can see the two new metanodes, the left one is a path and the other one is a cycle. The result of this clustering is the quotient graph that will be the input of the drawing algorithm.

Drawing algorithm
To draw the metabolic network, we use three drawing algorithms: one for the quotient graph and two for the metanodes.

Drawing metanodes
To draw subgraphs represented by metanodes, we use a recursive drawing algorithm. This algorithm draws all the subgraphs from the most nested to the least nested. According to our clustering method, a subgraph is either a cycle or an acyclic graph. In the first case, we use a circular drawing algorithm (see figure 4); in the second case, we use the hierarchical drawing algorithm presented in [38].

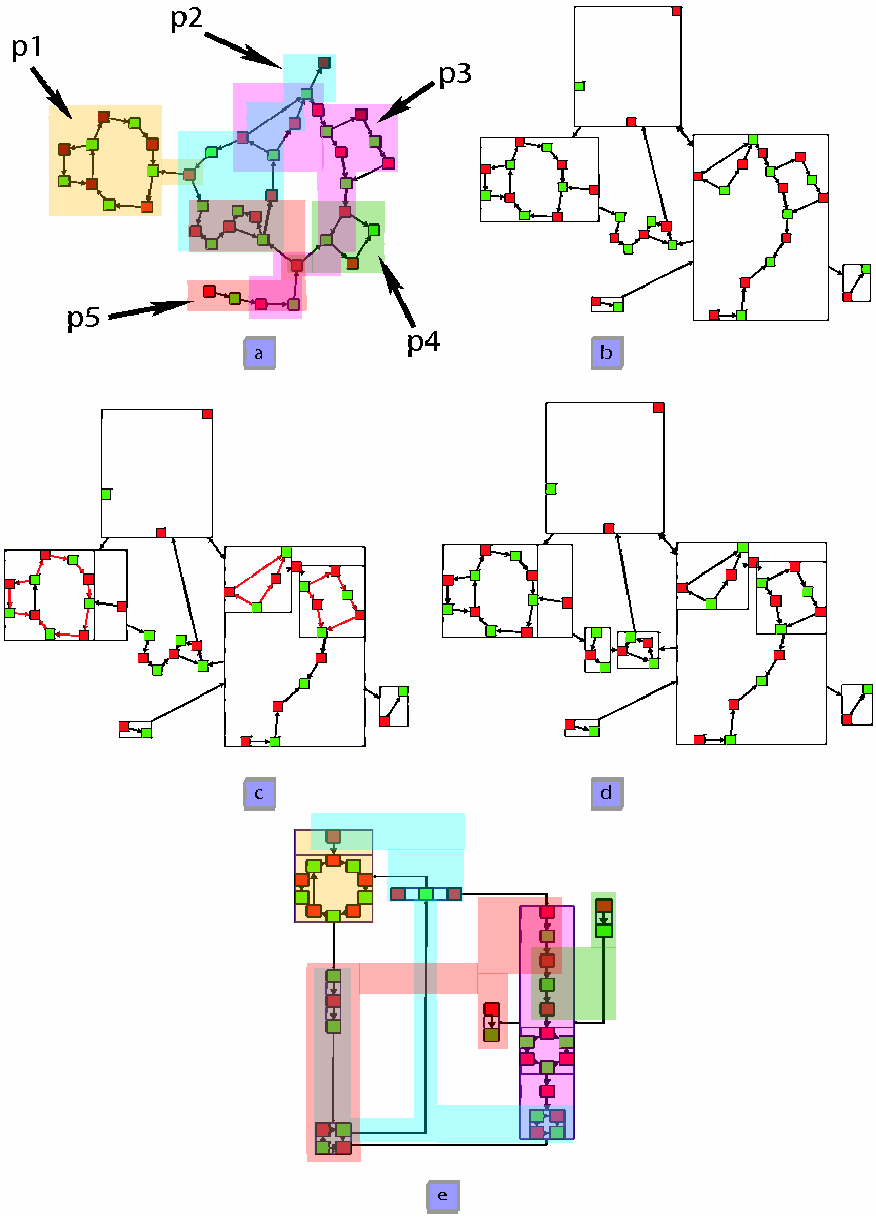


Figure 3
Algorithm overview. (a) a network where each pathway is depicted by a color (b) clustering according to metabolic pathways overlapping (c) cycles detection in metanodes (d) cycles and paths detection (e) final representation

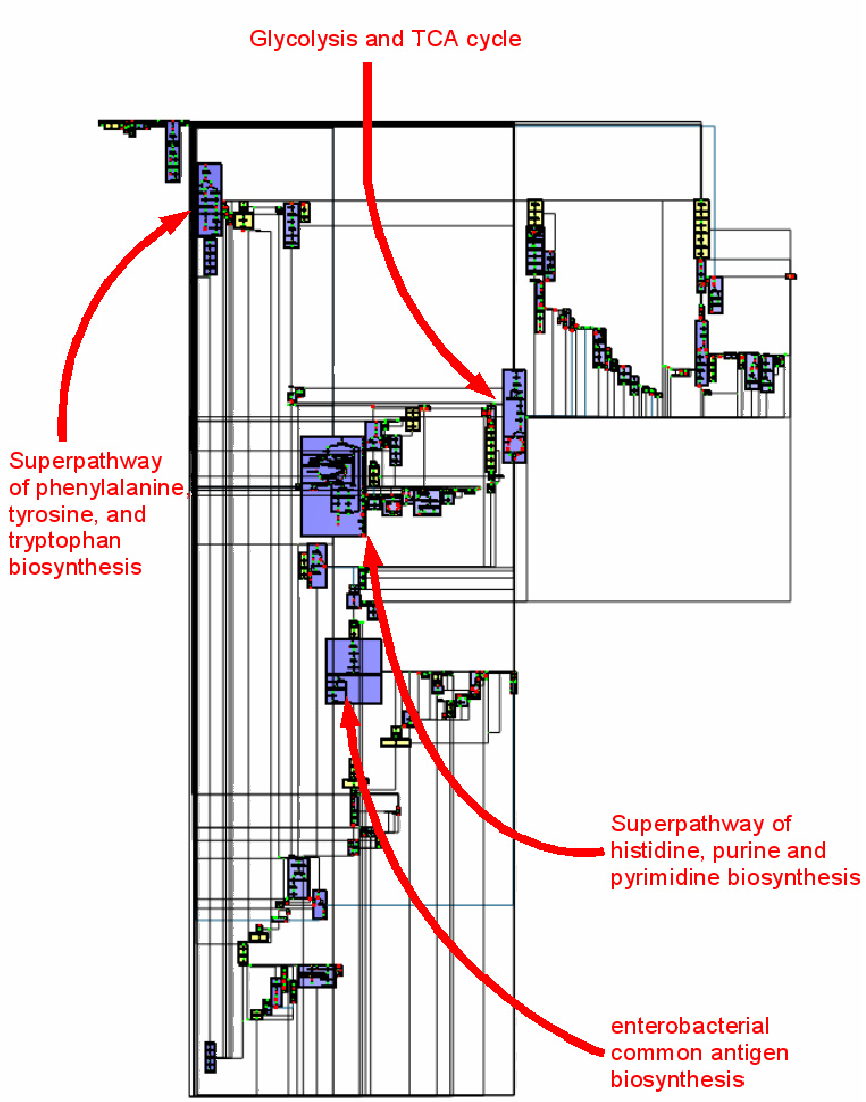


Figure 4
Whole metabolic network of E. coli drawn by MetaViz. The metanodes in purple represent metabolic pathways completely drawn. The metanodes in yellow correspond to specific structural schemes (chains or cycles) found by MetaViz.

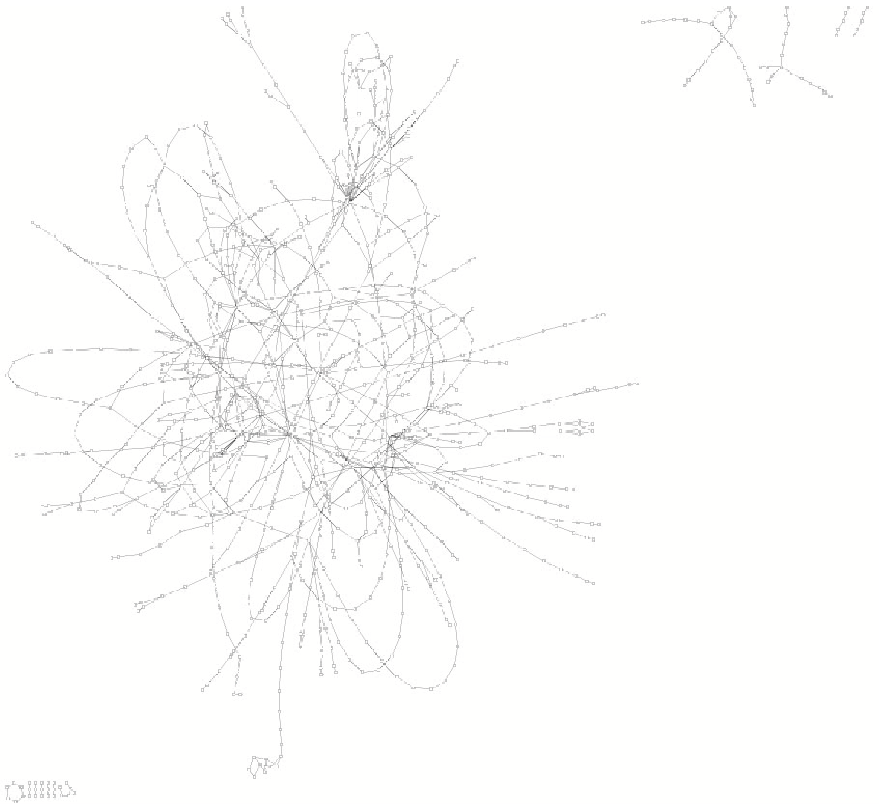


Figure 5
Whole metabolic network of E. coli drawn by Cytoscape.

Drawing the quotient graph

We want a drawing that optimizes the angular resolution and the number of bends to obtain a better visibility. The Mixed-Model algorithm of C. Gutwenger and P. Mutzel [39] is a trade-off between all these aesthetic criteria. Moreover, drawings produced by this algorithm are similar to manually drawn metabolic networks.

To use the Mixed-Model algorithm, we need to make modifications on the quotient graph. Indeed, it can only be applied to planar graphs; therefore, we have to planarize (*i.e.* make it planar) the quotient graph. This problem is well-known and is NP-Hard [40]. Many techniques exist that do it either by augmentation or by deletion of edges (or nodes). For a survey on this topic, one can refer to [41]. The drawback of an augmentation based technique is that it may add up to $|V|^4$ nodes, thus the drawing becomes difficult to understand. That is why we use our own heuristic: vertices of higher degree are removed one by one until the graph becomes planar. All removed nodes are then re-inserted. Removed edges are re-added one by one as long as the graph is planar.

The re-insertion of edges for each node is done with no prior order, using a greedy approach. The edges that have been removed and not re-inserted during the planarization step will be re-inserted after the planar subgraph is drawn.

The obtained planar subgraph of the quotient graph is drawn by the Mixed-Model algorithm [39]. To summarize, this algorithm has two steps :

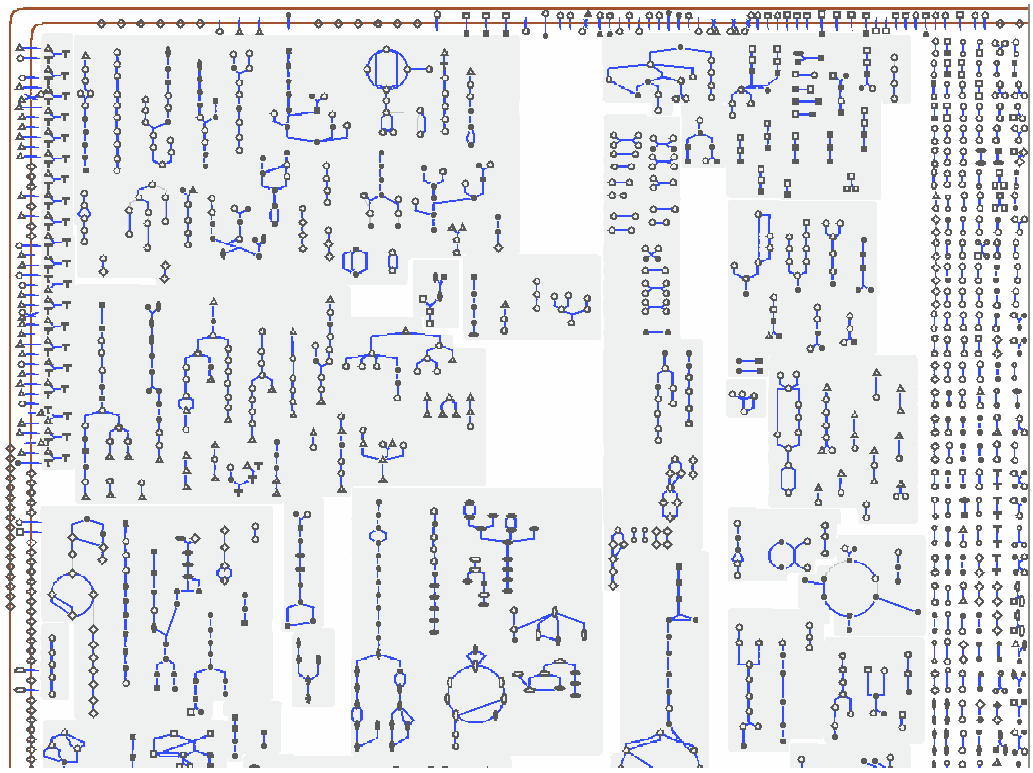


Figure 6
Whole metabolic network of E. coli drawn by the Pathway Tools cellular overview diagram.

• The first step builds an ordered partition of the set of nodes. This partition is called shelling ordering. The principle is to remove successively nodes that are on the external face of the graph.

• The second one is the "recomposition" of the graph according to the shelling ordering. To guarantee that there is neither edge-edge crossing nor node-edge overlapping, the ordering is traversed in reverse order.

As described in the background section, if a vertex is in a pathway, it has to be drawn close to the other vertices of the pathway. Taking into account such a constraint in the Mixed-Model algorithm can be done during the decomposition phase. Let $SO = \{V_1, V_2, ..., V_r\}$ be the shelling ordering. When a vertex n is added to a set V_i, $1 \leq i < r$, we add in priority vertices which have a constraint with n into the next V_j, $j > i$. Those nodes will be more likely to be drawn next to each other.

The last step of our drawing algorithm is to draw edges removed during the planarization step. These edges are routed on the external face, using an orthogonal drawing with three bends per edge. Figure 4 shows the drawing obtained by our algorithm on the metabolic network of *E. coli*. This is an organism which has been widely studied, its metabolism is composed of 198 pathways, 1140 substrates and reactions (*i.e.* nodes) and 1321 links (*i.e.* edges) between them.

Parameter: focus pathways

The algorithm allows to focus on several pathways, *i.e.* one can choose pathways to be entirely clustered. Users constrain the independent set algorithm by giving an ordered list of pathways that are clustered if possible. Indeed, such a list may not be represented by an independent set in the dependence graph (*i.e.* one or more nodes are shared by pathways of the list). In this case, the order of the list gives the priority associated to each path-

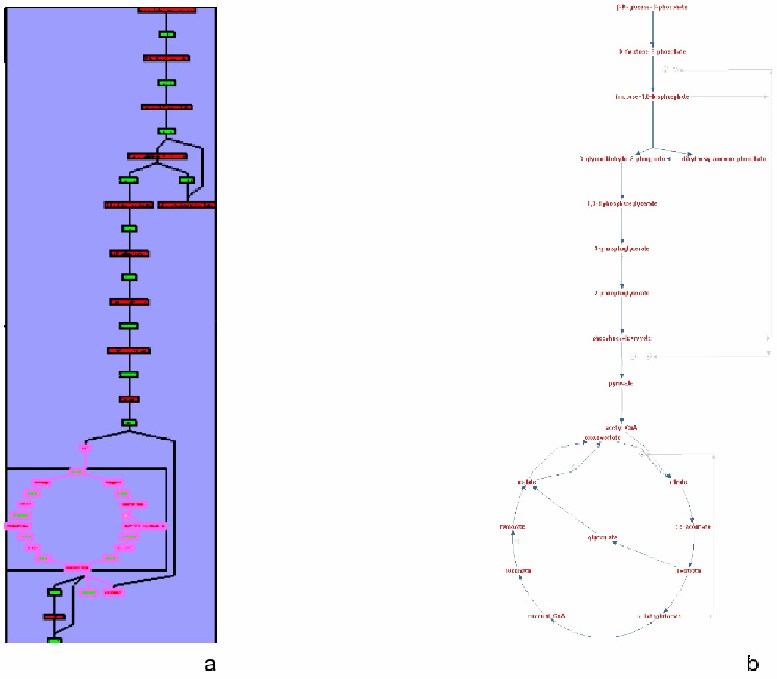


Figure 7
The superpathway of glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass. (a) In MetaViz. The nodes corresponding to the TCA cycle are surrounded in pink. (b) In BioCyc.

way and helps to extract an independent set of pathways from the list. Nodes representing those pathways and their neighbors are removed from the dependence graph. An independent set is then computed in the resulting dependence graph. The final independent set is obtained by adding this independent set and those computed in the list.

Results

Data

To test and validate the algorithm, we used data from the version 10.0 of the EcoCyc database. We developed perl scripts using the pathway tools software [42-44] to obtain information on the reactions, compounds and metabolic pathways involved in the metabolism of the *K12* strain of *Escherichia coli*. We chose this organism because it is perhaps the most curated one and we thus avoid most of the data artifacts caused by automatic reconstructions of metabolism.

Several filters are applied on the original data to build our test data. The first one is to withdraw reactions involving large molecules such as proteins. Next, we remove reactions that are involved in no identified metabolic pathway. The last filter has for objective to avoid ubiquitous compounds. Indeed, co-factors such as ATP and NADH participate in many reactions and form hubs in the network which lead to a very fuzzy drawing. One traditional way around this problem is to eliminate the most connected compounds but this implies that metabolic pathways that have these compounds as final products or as precursors become meaningless. We therefore prefer another solution which consists in eliminating the connection between a compound and a reaction if the compound is annotated in EcoCyc as "secondary" in each

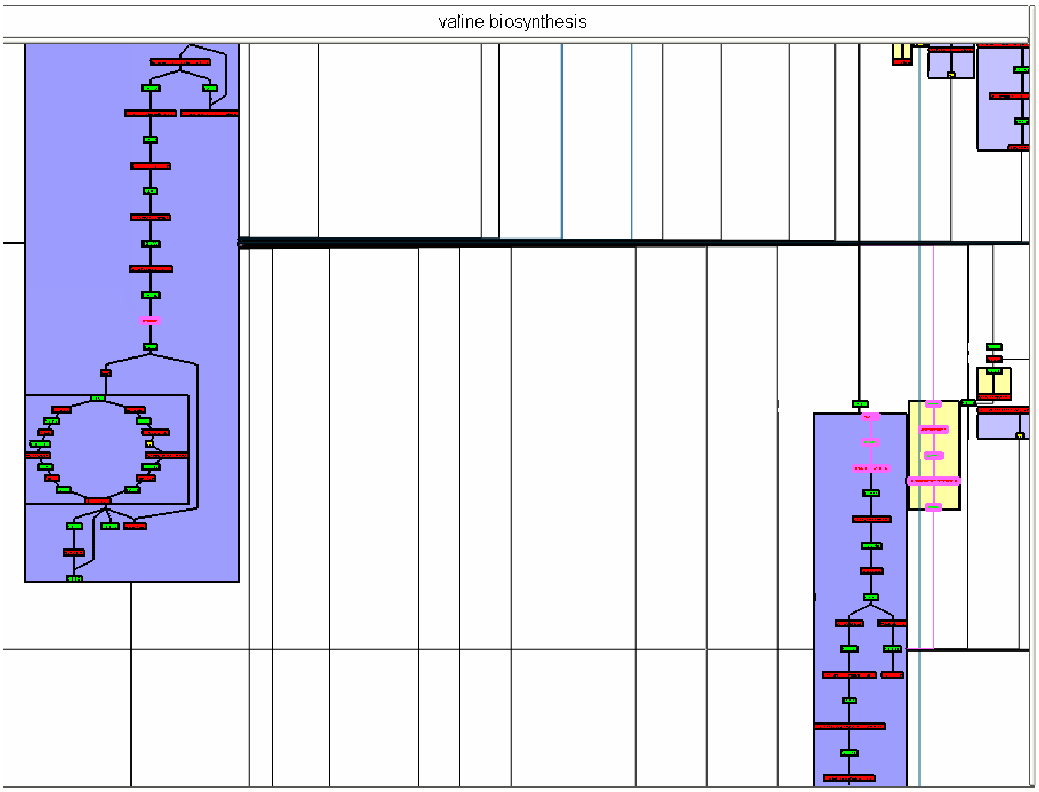


Figure 8
Valine Biosynthesis pathway in MetaViz, without choosing the metabolic pathway to be well drawn. The corresponding nodes are surrounded in pink and we can see that they are shared by 3 metanodes.

metabolic pathway that contains the reaction. A compound is defined as "primary" in a BioCyc metabolic pathway when it is a direct chemical intermediate between the start substrate(s) and the end product(s) and is defined as "secondary" when it is a sub-product or a secondary substrates (e.g cofactors) of the metabolic pathway.

It is important to note that this filter leads to a clearer drawing but any kind of compound filter could be applied. In the same way, the classification of the reactions in the EcoCyc-defined metabolic pathways was an easy way to test our algorithm but other classifications could be used, for instance a decomposition into elementary modes [45] or extreme pathways [46]. A metabolic pathway, as defined in BioCyc, can be either a linear chain of reactions, a branched pathway, a cycle: this topological diversity is interesting for testing our drawing algorithm.

The data is stored in a SBML file [47] and computed by MetaViz. The information about the belonging of each reaction is directly included in the SMBL file as shown below in the entry of one reaction which belongs to three different metabolic pathways:

...

<reaction id="DIHYDROFOLATEREDUCT__45__RXN" name="DIHYDROFOLATEREDUCT-RXN" reversible="true">

<notes>

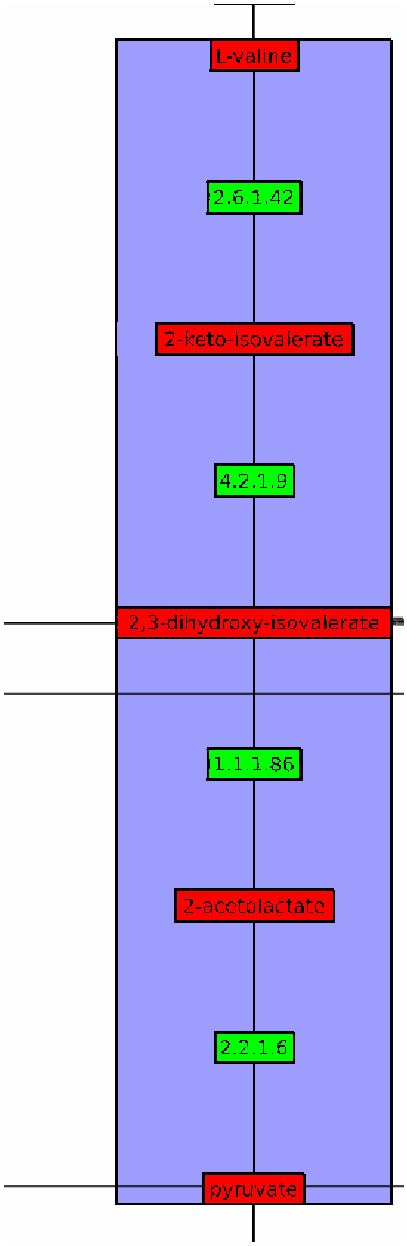


Figure 9
Valine Biosynthesis pathway in MetaViz, after choosing this metabolic pathway to be drawn well.

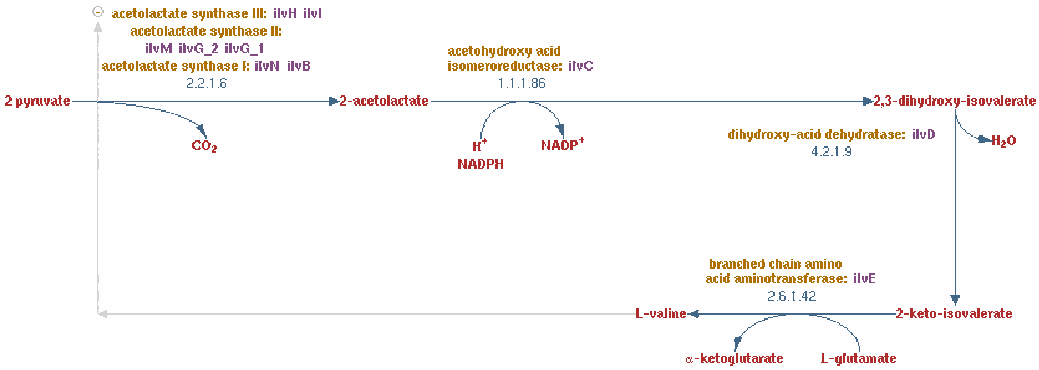


Figure 10
Valine Biosynthesis pathway in BioCyc.

<html:p>SUBSYSTEM: tetrahydrofolate biosynthesis</html:p>

<html:p>SUBSYSTEM: superpathway of chorismate</html:p>

<html:p>SUBSYSTEM: formylTHF biosynthesis I</html:p>

</notes>

<listOfReactants>

<speciesReference species="THF" stoichiometry="1"/>

</listOfReactants>

<listOfProducts>

<speciesReference species="DIHYDROFOLATE" stoichiometry="1"/>

</listOfProducts>

</reaction>

...

After the filtering, the SBML file contains :

• 553 compounds and 597 reactions (the nodes of the network represented in Metaviz)

• 198 metabolic pathways of which 30 are superpathways, *i.e.* pathways which contain other pathways.

Validation

The protocol we adopted for the validation is the following: we systematically compared the behavior of MetaViz to Cytoscape and to the Pathway Tools cellular overview diagram whenever possible. This comparison was carried out for the following tasks:

• Visualization of the whole network;

• Visualization of individual metabolic pathways;

• Visualization of a metabolic pathway in its context.

Visualization of the whole network

Figure 4 shows the whole metabolic network computed by MetaViz from the data described in the previous section. Unlike the drawing obtained by Cytoscape [33] with the same data (Figure 5), the metabolic network is organized into metanodes in MetaViz. The purple metanodes indicate the metabolic pathways selected during the clustering step and which are therefore drawn well (nodes of the pathways are close to each other). These metabolic pathways form the backbone of the drawing, which can be changed by choosing to draw well other metabolic pathways.

The drawing obtained by the Pathway Tools cellular overview diagram (Figure 6) with the same data represents all metabolic pathways but in this case, the layout is fixed. Moreover, it is not possible to zoom further into the drawing.

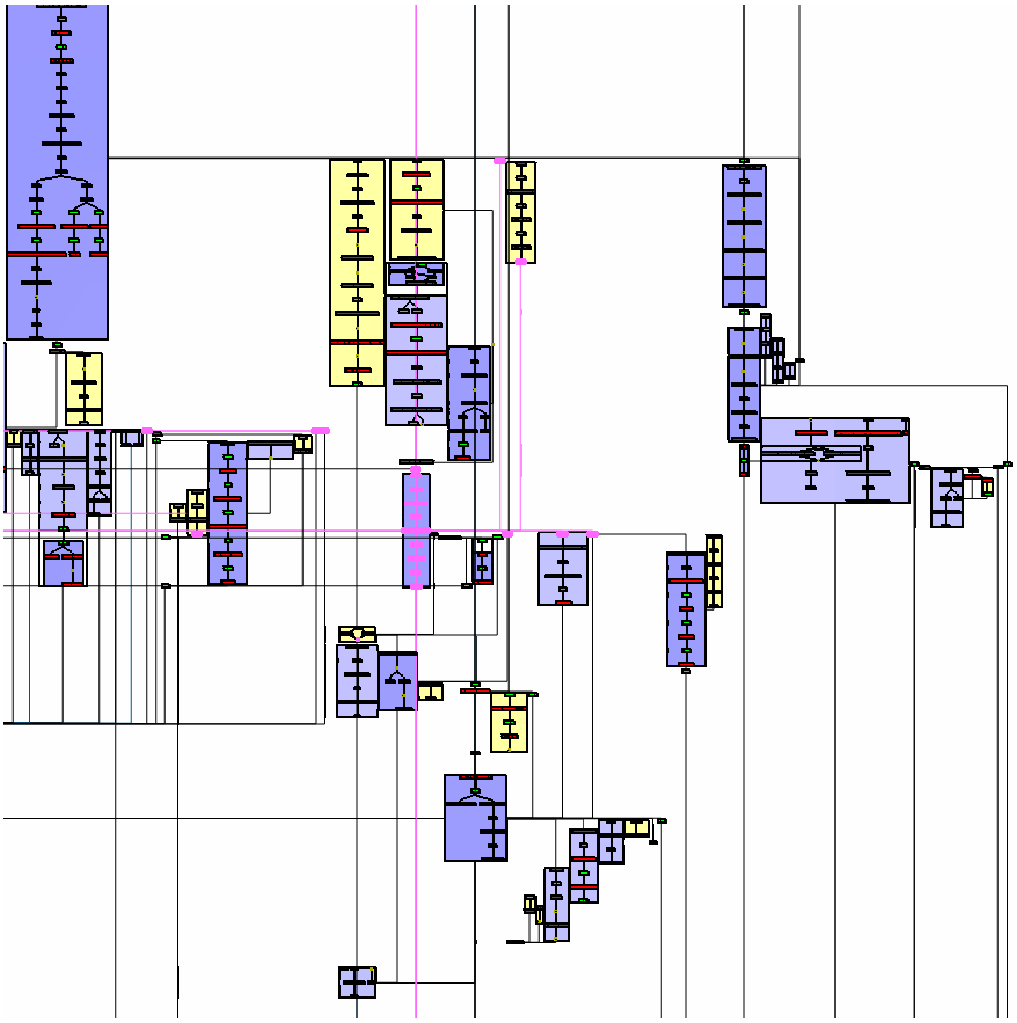


Figure 11
Drawing of the nodes (colored in pink) directly connected to the Valine Biosynthesis Pathway (in the center of the figure).

Unlike the Pathway Tools cellular overview diagram, MetaViz enables to see a metabolic pathway in its context, keeping the same layout. For instance, Figure 7a is merely a zoom of Figure 4.

Drawing of the TCA cycle

We do not compare the results with Cytoscape of which the purpose is not to draw metabolic pathways but only to draw a whole network.

In the data from BioCyc, the TCA cycle is included in the super pathway of "glycolysis, pyruvate dehydrogenase, TCA, and glyoxylate bypass". Because of its great number of nodes, this pathway was chosen by the algorithm to be particularly well drawn: all the nodes (compounds and reactions) involved in this super pathway are grouped together into a same metanode (Figure 7a). The drawing obtained by MetaViz is very similar to the one obtained by the pathway viewer of BioCyc (Figure 7c). The differences

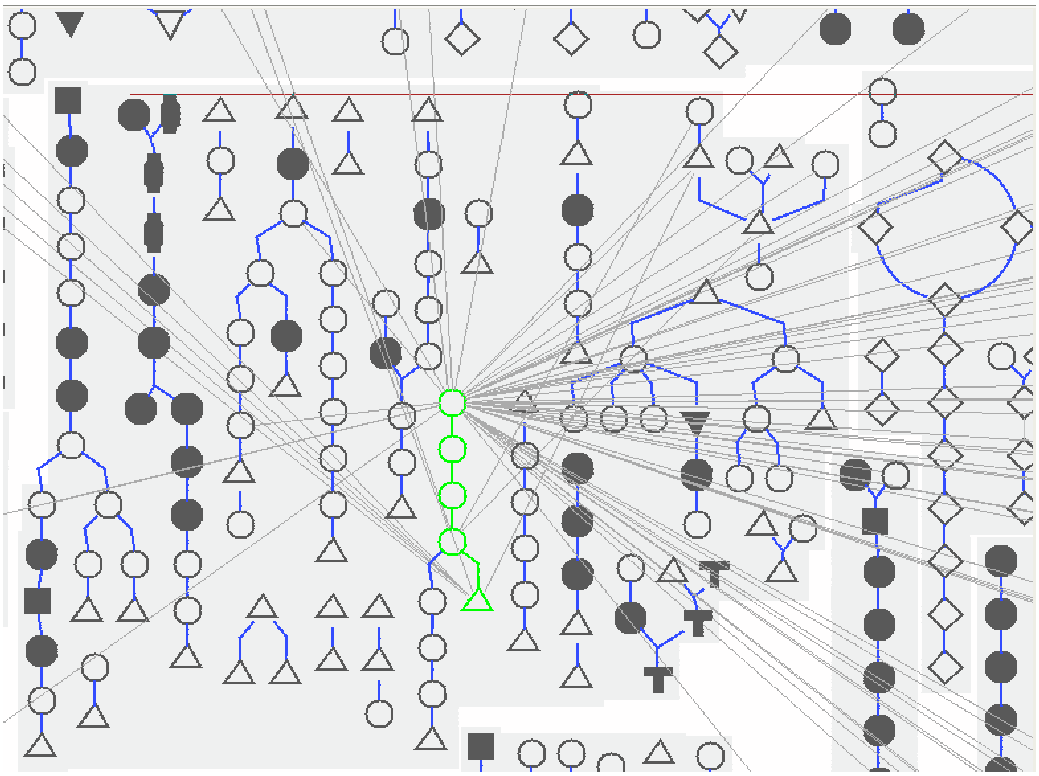


Figure 12
Connections from the valine biosyntheisis pathway in the Pathway Tools cellular overview diagram.

between the two drawings are mostly due to the differences in the types of graph used to model the network: a simple graph in the case of BioCyc, and a bipartite graph in the case of MetaViz.

Drawing of the valine biosynthesis pathway
This pathway is a four-step chain which starts with pyruvate and ends with L-valine.

We present here two cases: 1. the clustering is not guided and 2. the clustering is guided. If the clustering is not guided, this pathway is not selected to be drawn well and is actually split into three parts: one node is drawn in the superpathway of the TCA cycle and glycolysis (because they share the pyruvate), one node corresponds to the superpathway of pantothenate and coenzyme A biosynthesis (because they share L-valine, alpha-keto isovalerate and the 2.6.1.42 reaction) and the third node corresponds to the other reactions (Figure 8).

This metabolic pathway has not been efficiently drawn because some of its elements belong to larger metabolic pathways. Nevertheless, we do not see such a representation as a negative result but instead consider the division of this metabolic pathway into several parts as interesting. Indeed, it means that this metabolic pathway shares several elements with others, showing the interdependence between the pathways. Otherwise, if the clustering is guided and valine biosynthesis is chosen as a focus pathway, MetaViz efficiently represents it (Figure 9). Obviously, this choice leads to the disconnection of the metabolic pathways sharing the same nodes. As mentioned above, we can see here one of the main interests of MetaViz: it is possible to change the backbone of the drawing to center it on specific metabolic pathways. If we

compare this drawing with the one obtained by the pathway viewer of BioCyc (Figure 10), we observe that the order of the nodes is reversed. That is pyruvate is on the left of BioCyc drawing while it is at the bottom of the MetaViz one. Hence pyruvate appears as the input of the pathway. But in BioCyc SBML description these reactions are annotated as reversible. So it is not, in that case, possible to automatically identified pyruvate as the input of the pathway.

Visualization of a metabolic pathway in its context

MetaViz represents explicitly the links between metabolic pathways. These links are ignored when metabolic pathways are separately drawn (as in BioCyc) or when no information about the belonging of the nodes to a metabolic pathway is displayed (as in Cytoscape). The Pathway Tools Cellular Overview diagram proposes to optionally draw these links in superposition to the main drawing. The limit of this approach is that, since these links are not incorporated in the original layout, the final drawing may become very dense and hard to read.

It is possible with MetaViz to highlight the nodes that are neighbors of a selected node. Figure 11 shows the direct neighbors (colored in pink) of the valine biosynthesis pathway. One can then more easily follow each edge to see to which nodes in the network this metabolic pathway is connected.

Figure 12 shows the connections from the valine biosynthesis pathway computed in the Pathway Tools cellular diagram overview. However, because nodes are duplicated and the layout is fixed, a lot of edges are displayed and it is difficult to follow one edge.

Conclusion

In this paper, we present an algorithm to compute the representation of a metabolic network. This method addresses a challenging problem which consists in representing simultaneously the topology and the metabolic pathway information. Indeed, metabolic pathways often share metabolites and reactions, thus to represent them in a single view, previous approaches duplicated these shared elements. However, duplication produces drawings where the depicted connectivity does not fit the real topology of the network. To overcome the problem of shared nodes, we propose a clustering step based both on topology and a metabolic pathway decomposition. During this step, we deal with pathway overlapping by detecting a largest set of independent pathways and sub-pathways. The resulting graph clustering shows the overall organization of the pathways. To follow common drawing conventions, it is drawn using a planar graph drawing algorithm. Finally, each pathway or sub-pathway is drawn using specific drawing algorithms (hierarchical and circular ones). In our collaboration with physiologists, we noticed that they often consider some pathways as being central in their global studies. To respect their habits, the physiologists can provide a set of focus pathways that will be considered as a parameter of the clustering step. Thus our algorithm will generate a drawing where these pathways are entirely and carefully drawn.

This global representation allows the visualization of processes that span over different metabolic pathways. For instance, this approach was successfully used to highlight metabolic processes, especially those traversing different metabolic pathways.

One of the future directions we would like to consider concerns the improvement of the global aspect of our drawing. The drawing conventions that we identified for metabolism are mostly local (emphasizing cycles and reaction cascades). Following them does not ensure to have a global picture that will look like the Boehringer map [23] which may be closer to what biochemists are used to. Indeed, the global picture that we obtain with our method can be puzzling at first glance, and it is only when navigating in the drawing that the user will find more familiar patterns. We believe that we can improve the aspect of the global drawing in considering alternative ways of drawing the quotient graph.

In this paper, we focused on the drawing part of metabolic network visualization. As it was mentioned, drawings are used as a background for high throughput data visualization. Since this algorithm is already implemented in a graph drawing software [38], we plan to develop an input module for omic data. Another issue will be to add more relational information such as signaling processes. We plan to use the third dimension to incorporate the additional edges.

Availability and requirements

Project name: MetaViz

Project home page: http://www.labri.fr/perso/bourqui/software.php

Operating system(s): Currently Linux and Windows. Mac OSX ports is possible.

Programming language: C++

Other requirements: Tulip [38], Qt from Trolltech.

License: GPL

Authors' contributions

FJ initiated this work. RB, VL, LC, DA, MS and FJ defined metabolic network drawing constraints. RB, DA, and FJ established the translation of these constraints into graph drawing ones. RB and DA designed the drawing algorithm. RB and PM implemented the algorithm. LC build the datasets from EcoCyc. VL, LC and MS performed the tests and result analysis. All authors participated in manuscript preparation. All authors have read and approved the final manuscript.

Acknowledgements

The work presented in this paper was funded in part by the ACI Nouvelles Interfaces des Mathématiques (project *p-vert*) of the French Ministry of Research, by the ARC (project *IBN*) from the INRIA and by the ANR (project *REGLIS*).

References

1. Karp PD, Paley SM: **Automated Drawing of Metabolic Pathways.** *Third International Conference on Bioinformatics and Genome Research* 1994.
2. Salamonsen , Yee , Mok , Kolatkar : **BioJAKE: a tool for the creation, visualization and manipulation of metabolic pathways.** *Pacific Symposium on Biocomputing* 1999, **4**:392-400.
3. Kanehisa M: *Post-genome Informatics* Oxford University Press; 2000.
4. Becker M, Rojas I: **A Graph Layout Algorithm for Drawing Metabolic Pathways.** *Bioinformatics* 2001, **17**:461-467.
5. Seo J, Shneiderman B: **Interactively Exploring Hierarchical Clustering Results.** *IEEE Computer* 2002, **35(7)**:80-86.
6. Schreiber F: **Comparison of metabolic pathways using constraint graph drawing.** In *APBC 03: Proceedings of the First Asia-Pacific bioinformatics conference on Bioinformatics* Australian Computer Society, Inc; 2003:105-110.
7. Jourdan F, Melançon G: **A Tool for Metabolic and Regulatory Pathways Visual Analysis.** *Visualization and Data Analysis, VDA* 2003:46-55 [Http://vw.indiana.edu/vda2003/]. Santa Clara Convention Center: SPIE
8. Dogrusoz , Giral , Cetintas , Civril , Demir : **A Compound Graph Layout Algorithm for Biological Pathways.** *GD 2004* 2004, **3383**:442-447.
9. Brandes U, Dwyer T, Schreiber F: **Visualizing Related Metabolic Pathways in Two and Half Dimensions.** *LNCS* 2004, **2912**:11-122.
10. Wegner , Kummer : **A new dynamical layout algoritmh for complex biochemical reaction networks.** *BMC Bioinformatics* 2005, **6**:212.
11. Gabouje SD, Zimányi : **A New Compound Graph Layout Algorithm for Visualizing Biochemical Networks.** *Poster Proceedings Volume of the 4th International Workshop on Efficient and Experimental Algorithms, WEA 05* .
12. Dogrusoz , Erson , Giral , Demir , Babur , Cetintas , Colak : **PATIKAweb: a Web interface for analyzing biological pathways through advanced querying and visualization.** *Bioinformatics* 2005, **22(3)**:374-375.
13. Tope J, Gillespie , Vastrik , DEustachio , Schmidt , de Bono , Jassal , Gopinath , Wu , Matthews , Lewis , Birney , Stein : **Reactome: a knowledgebase of biological pathways.** *Nucleic Acids Research* 2005, **33**:D428-D432.
14. Paley S, Karp P: **The Pathway Tools cellular overview diagram and Omics Viewer.** *Nucleic Acids Research* 2006, **34(13)**:3771-3778.
15. Junker BH, Klukas C, Schreiber F: **VANTED: A System for Advanced Data Analysis and Visualization in the Context of Biological Networks.** *BMC Bioinformatics* 2006, **7**:109. EPub
16. Nikiforova V, Kopka J, Tolstikov V, Fiehn O, Hopkins L, Hawkesford M, Hesse H, Hoefgen R: **Systems Rebalancing of Metabolism in Response to Sulfur Deprivation, as Revealed by Metabolome Analysis of Arabidopsis Plants.** *Plant Physiology* 2005, **138**:304-318.
17. Lacroix V, Fernandes CG, Sagot MF: **Motif search in graphs: application to metabolic networks.** *IEEE/ACM Trans Comput Biol Bioinform* 2006, **3(4)**:360-368.
18. Saraiya P, North C, Duca K: **Visualizing biological pathways: requirements analysis, systems evaluation and research agenda.** *Information Visualization* 2005, **4**:1-15.
19. Jeong H, Tombor B, Albert R, Oltvai Z, Barabasi A: **The Large-Scale Organization of Metabolic Networks.** *Nature* 2000, **407**:651.
20. Pfeiffer T, Dandekar T, Moldenhauer F, Schuster S: **Topological analysis of metabolic networks. Application to the metabolism of Mycoplasma pneumoniae.** *BTK2000: Animating the Cellular Map* 2000, **6**:229-234.
21. Salway JG: *Metabolism at a Glance* Blackwell Science Ltd; 2003.
22. Michal G: **On representation of metabolic pathways.** *BioSystems* 1998, **47**:1-7.
23. Michal G: *Biochemical Pathways (Poster)* Boehringer Mannheim; 1993.
24. BIOCARTA: **Charting pathways of life.** [http://www.biocarta.com].
25. Romero P, Wagg J, Green M, Kaiser D, Krummenacker M, Karp P: **Computational prediction of human metabolic pathways from the complete human genome.** *Genome Biology* 2004:1-17.
26. Purchase H, Cohen RF, James M: **An Experimental Study of the Basis for Graph Drawing Algorithms.** *ACM Journal of Experimental Algorithmics* 1997, **2(4)**:189.
27. Battista GD, Eades P, Tamassia R, Tollis IG: *Graph Drawing: Algorithms for the Visualization of Graphs* Prentice Hall; 1999.
28. Kaufmann M, Wagner D: *Drawing Graphs Springer* 2001.
29. Sugiyama , Misue : **Visualisation of structural information : Automatic drawing of compound digraphs.** *IEEE Transactions on Systems, Man, and Cybernetics* 1991, **21(4)**:876-892.
30. Brandenburg F, Forster M, Pick A, Raitner M, Schreiber F: **Biopath.** *GD'01* 2002.
31. Eades : **A heuristic for graph drawing.** *Congressus Numerantium* 1984, **42**:149-160.
32. Frick , Ludwig , Mehldau : **A fast adaptive layout algorithm for undirected graphs.** *Lecture Notes in Computer Science* 1994, **894**:388-403.
33. Shannon P, Markiel A, Ozierand O, Baliga N, Wang J, Ramage D, Amin N, Schwikowski B, Ideker T: **Cytoscape: A Software Environment for Integrated Models of Biomolecular Interaction Networks.** *Genome Research* 2003, **13**:2498-2504.
34. **SBML viewer** [http://sbw.kgi.edu/layout/]
35. van Helden J, Wernisch L, Gilbert D, Wodak S: **Graph-based analysis of metabolic networks.** *Ernst Schering Research Foundation Workshop* 2002, **38**:245-274.
36. Garey MR, Johnson DS: *Computers and Intractability: A Guide to the Theory of NP-Completeness* New York, NY, USA: W. H. Freeman & Co; 1979.
37. Welsh , Powell : **An upper Bound to the chromatic number of a graph and its application to timetabling problems.** *The Computer journal* 1967, **10**:85-86.
38. Auber D: *Graph Drawing Software* Springer-Verlag 2003 chap. Tulip- A Huge Graph Visualization Framework.
39. Gutwenger C, Mutzel P: **Planar Polyline Drawings with Good Angular Resolution.** In *Graph Drawing '98 (Proc.) Volume 1547*. Springer-Verlag, Lecture Notes in Computer Science; 1998:167-182.
40. Lui P, Geldmacher R: **On the deletion of nonplanar edges of a graph.** *Proceeding on the 10th conf. on Comb., Graph Theory, and Comp* 1977:727-738.
41. Liebers A: **Planarizing Graphs – A Survey and Annotated Bibliography.** *Journal of Graph Algorithms and Applications* 2001, **5**:1-74.
42. Karp P, Riley M, Saier M, Paulsen I: **The EcoCyc and MetaCyc databases.** *Nucleic Acids Research* 2000, **28**:56-59.
43. Karp PD, Paley S, Romero P: **The Pathway Tools software.** *Bioinformatics* 2002, **18(Suppl 1)**:S225-32.
44. Krummenacker M, Paley S, Mueller L, Yan T, Karp PD: **Querying and computing with BioCyc databases.** *Bioinformatics* 2005, **21(16)**:3454-3455.
45. Schuster S, Hilgetag C, Woods JH, Fell DA: **Reaction routes in biochemical reaction systems: algebraic properties, validated calculation procedure and example from nucleotide metabolism.** *J Math Biol* 2002, **45(2)**:153-181.
46. Schilling CH, Letscher D, Palsson BO: **Theory for the systemic definition of metabolic pathways and their use in interpret-**

ing metabolic function from a pathway-oriented perspective. *J Theor Biol* 2000, **203(3):**229-248.
47. Finney AHM: **Systems biology markup language: Level 2 and beyond.** *Biochem Soc Trans* 2003:1472-3.

Uncovering structure in biological networks

Mahendra Mariadassou[1], Jean-Jacques Daudin[1], Vincent Lacroix[2], Vincent Miele[3], Franck Picard[1], Stéphane Robin[1], Marie-France Sagot[2]

[1]UMR INA P-G/ENGREF/INRA MIA 518, Groupe SSB, Paris FRANCE
[mariadas][daudin][picard][robin]@inapg.inra.fr
[2] UMR 5558 Biométrie et Biologie Évolutive, Université Claude Bernard, Lyon I, Villeurbanne FRANCE,
[sagot][lacroix]@biomserv.univ-lyon1.fr
[3] UMR 8071 - INRA 1152 Laboratoire de Statistique et Génome, Université d'Evry, Evry, FRANCE,
miele@genopole.cnrs.fr

Abstract

The Erdös-Rényi model of a network is simple and possesses many explicit expressions for average and asymptotic properties, but it does not fit well to real-word networks. The vertices of these networks are often structured in prior unknown groups (functionally related proteins or social communities) with different connectivity properties. We define a generalization of the Erdös-Rényi model called ERMG for Erdös-Rényi Mixtures for Graphs. This new model is based on mixture distributions. We give some of its properties, an algorithm to estimate its parameters and apply this method to uncover the modular structure of a network of enzymatic reactions.

1 Introduction

The network structure has drained a large interest from the scientific community in many fields in recent years. The two main elements of this renewed interest are the easier access to large datasets structured as networks, especially in social science and in biology, which boosted the development of tools adapted to such data and the evergrowing capacity of computers, which allows one to investigate networks both from a theoretical and an empirical point of view. The appropriate modeling of these networks by a statistical object is an issue to which much attention has been given.

The Erdös-Rényi model of a network is one of the oldest and best studied models and possesses many explicit expressions for average and asymptotic properties such as subgraphs, degree distribution, connectedness and clustering coefficient. However this theoretical model does not fit well to real-world, social, biological or Internet networks. For example the empirical degree distribution may be very different from the Poisson distribution which is implied by this model. Moreover empirical clustering coefficients of real networks are generally higher than the value given by this model. Some generalizations of the Erdös-Rényi model have been recently made in order to correct these shortcomings. For a review of these works see [Albert and Barabási (2002)] or [Newman (2003)]. One of the limits for these studies is that no existing network model seems to be completely satisfying to capture their structure.

One research direction is to incorporate clustering in the model. Assortative mixing or mixing patterns (see [Newman and Girvan (2003)] and [Newman (2004)]) postulate that the vertices may be classified into groups with different connectivity properties. The key element is the mixing matrix which specifies the probability of connection between two groups. [Newman (2003)] gives some theoretical properties of such networks and an algorithm similar to Metropolis-Hasting for simulating networks for a given mixing matrix. The inference of the mixing parameters is quite easy if groups can be defined using external information such as language, race or age. However the inference is more difficult when groups and mixing parameters have to be inferred from the network topology alone. A first step is the greedy optimisation algorithm proposed by [Newman (2004)]. In this article we propose a new statistical method to infer the clustering of vertices and the parameters of the mixing model using a maximum-likelihood approach based only on the network topology. We then apply this method to a network representing the small molecule metabolism of *Escherichia coli*.

2 Erdös-Rényi and scale-free model

NOTATIONS. In this article, we consider an undirected graph with n vertices and define the variable X_{ij} which indicates that vertices i and j are connected :

$$X_{ij} = X_{ji} = \mathbb{I}\{i \leftrightarrow j\},$$

where $\mathbb{I}\{A\}$ equals to one if A is true, and to zero otherwise. Furthermore, we assume that no vertex is connected to itself, meaning that $X_{ii} = 0$. In the following we note K_i the degree of vertex i, *i.e.* the number of edges connecting it to the graph :

$$K_i = \sum_{j \neq i} X_{ij}.$$

ERDÖS-RÉNYI MODEL. This model assumes that edges are independent and occur with the same probability p :

$$\{X_{ij}\} \text{ i.i.d.}, \; X_{ij} \sim \mathcal{B}(p).$$

In this model, the degree of each vertex has a Binomial distribution, which is approximately Poisson for large n and small p. Noting $\lambda = (n-1)p$ we have :

$$K_i \sim \mathcal{B}(n-1, p) \approx \mathcal{P}(\lambda). \tag{1}$$

'SCALE-FREE' NETWORK. In many practical situations, the Erdös-Rényi model turns out to fit the data poorly, mainly because the distribution of the degrees is far from the Poisson distribution (1). The scale-free (or Zipf) distribution has been intensively used as an alternative. The Zipf probability distribution function (pdf) is

$$\Pr\{K_i = k\} = c(\rho) k^{-(\rho+1)}, \tag{2}$$

where k is any positive integer, ρ is positive, $c(\rho) = 1/\sum_{k \geq 1} k^{-(\rho+1)} = 1/\zeta(\rho+1)$ and $\zeta(\rho+1)$ is Riemann's zeta function. Nevertheless, we will show in Section 6 that this distribution may have a poor fit on real datasets as well.

First of all, it is important to notice that the Zipf distribution is used to model the tail of the degree distribution. Consequently it is often better suited for the tail than for the whole distribution. In particular this distribution has a null probability for $k = 0$ whereas some vertices may be unconnected in practice. Moreover the lack-of-fit of the Erdös-Rényi model may be simply due to some heterogeneities between vertices, some being more connected than others. And finally, the scale-free model focuses only on the degree distribution, not at all on the graph topology. A simple way to tackle these three limitations is to consider that the nodes belong to groups which have different connectivity properties.

3 Erdös-Rényi mixture for graphs

3.1 General model

MIXTURE MODEL FOR THE NODES We now propose a mixture model which explicitly describes the way edges connect vertices, accounting for some heterogeneity among vertices. In the following, we denote this model ERMG for Erdös-Rényi Mixture for Graphs. The ERMG model supposes that vertices are structured into Q groups, and that there exists a sequence of independent hidden variables $\{Z_{iq}\}$ which indicate the label of vertices. We note α_q the probability for vertex i to belong to group q, such that :

$$\alpha_q = \Pr\{Z_{iq} = 1\} = \Pr\{i \in q\}, \text{ with } \sum_q \alpha_q = 1.$$

Remark : In the following, we will use two equivalent notations : $\{Z_{iq} = 1\}$ or $\{i \in q\}$ to indicate that vertex i belongs to group q.

We then denote $\pi_{q\ell}$ the probability for a vertex from group q to be connected with a vertex from group ℓ. Because the graph is undirected, these probabilities must be symmetric such that :

$$\pi_{q\ell} = \pi_{\ell q}.$$

We also suppose that edges $\{X_{ij}\}$ are conditionally independent given the groups of vertices i and j :

$$X_{ij} \mid \{i \in q, j \in \ell\} \sim \mathcal{B}(\pi_{q\ell}).$$

More than just describing the degree distribution, our model describes the topology of the network using the connectivity matrix $\mathbf{\Pi} = (\pi_{q\ell})$.

3.2 Examples

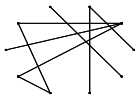
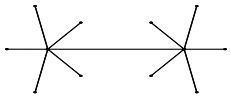
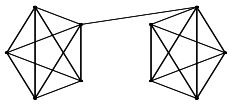
In this section we aim at showing that the ERMG model can be used to generalize many particular structures of random graphs. Table 1 presents some typical network configurations. The first one is the Erdös-Rényi model. We present here some more sophisticated ones.

RANDOM GRAPHS WITH ARBITRARY DEGREE DISTRIBUTIONS. The Erdös-Rényi random graph model is a poor approximation of real-world networks whose degree distribution is highly skewed. A random network having the same degree distribution as the empirical one can be built as follows : 1.n partial edges (with only one starting vertex and no final vertex) are randomly chosen from the empirical degree distribution and 2. these partial edges are randomly joined by pairs to form complete edges (see [Molloy and Reed (1995)]). A permutation algorithm is also proposed in [Shen-Orr *et al.* (2002)].

SCALE FREE NETWORK. The scale-free network proposed by [Barabási and Albert (1999)] is a particular case of random graphs with arbitrary distribution. To this extent, we can propose an analogous model in the ERMG framework. Suppose that the incoming vertices join the network in groups of respective size $n\alpha_q$ ($q = 1..Q$, $n\alpha_1$ being the number of original vertices). Assuming that the elements of a new group connect preferentially to the elements of the oldest groups : $\pi_{q,1} \geq \pi_{q,2} \geq \cdots \geq \pi_{q,q-1}$, we get the same kind of structure as the scale-free model.

STAR PATTERN. Many biological networks contain star patterns, *i.e.* many vertices connected to the same vertex and only to it, see the interaction network of *S. cerevisiae* in [Zhang *et al.* (2005)] for instance. This type of pattern may be modeled by an ERMG with extra-diagonal ones in $\mathbf{\Pi}$.

TAB. 1 – Some typical network configurations and their formulation in the framework of the ERMG model

Description	Network	Q	$\mathbf{\Pi}$	Clustering coef.
Random		1	p	p
Stars		4	$\begin{pmatrix} 0 & 1 & 0 & 0 \\ 1 & 0 & 1 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}$	0
Clusters (affiliation networks)		2	$\begin{pmatrix} 1 & \varepsilon \\ \varepsilon & 1 \end{pmatrix}$	$\frac{1+3\varepsilon^2}{(1+\varepsilon)^2}$

4 Some properties of the ERMG model

4.1 Distribution of the degrees

Conditionally to the belonging of vertices to groups, edges connecting vertex i belonging to group q are independent. The conditional connection probability is :

$$\Pr\{i \leftrightarrow j \mid i \in q\} = \sum_{\ell} \Pr\{i \leftrightarrow j \mid i \in q, j \in \ell\} \Pr\{j \in \ell\} = \sum \alpha_\ell \pi_{q\ell} = \overline{\pi}_q.$$

from which we deduce the following :
Proposition : Given the label of a vertex, the conditional distribution of the degree of this vertex is Binomial (approximately Poisson) : $K_i \mid \{i \in q\} \sim \mathcal{B}(n-1, \overline{\pi}_q) \approx \mathcal{P}(\lambda_q)$, where $\overline{\pi}_q = \sum_\ell \alpha_\ell \pi_{q\ell}$ and $\lambda_q = (n-1)\overline{\pi}_q$.

4.2 Clustering coefficient.

This coefficient measures the local aggregative trend of a graph. This coefficient C_i is empirically defined, for vertex i, as the numbers of edges between of his neighbors divided by the same number if i and his neighbors formed a clique [Albert and Barabási (2002)]. A first estimator of this empirical clustering coefficient is usually defined as the mean of the C_is : $\widehat{c} = \sum_i C_i/n$.
Denoting ∇ the 'triangle' configuration $(i \leftrightarrow j \leftrightarrow k \leftrightarrow i)$ and V the 'V' configuration $(j \leftrightarrow i \leftrightarrow k)$ for any (i,j,k) in $\{1,\ldots,n\}$, the definition of c can be rephrased as $c = \Pr\{\nabla \mid \mathsf{V}\}$. Because ∇ is a particular case of V, we have :

$$c = \Pr\{\nabla \cap \mathsf{V}\}/\Pr\{\mathsf{V}\} = \Pr\{\nabla\}/\Pr\{\mathsf{V}\}. \tag{3}$$

This property suggests another estimate of c given by : $\widehat{c}' = 3\sum_i \nabla_i / \sum_i V_i$, where V_i is the number of Vs in i : $V_i = \sum_{j>k,(j,k)\neq i} X_{ij}X_{ik}$. In the following we propose the equivalent probabilistic definition of this coefficient.
Definition : The *clustering coefficient* is the probability for two vertices j and k connected to a third vertex i, to be connected, with (i,j,k) uniformly chosen in $\{1,\ldots,n\}$

$$c = \Pr\{X_{ij}X_{jk}X_{ki} = 1 \mid X_{ij}X_{ik} = 1\}.$$

For any triplet (i,j,k), we have

$$\Pr\{\nabla\} = \sum_{q,l,m} \alpha_q\alpha_\ell\alpha_m \Pr\{X_{ij}X_{jk}X_{ki} = 1 \mid i \in q, j \in \ell, k \in m\}, = \sum_{q,l,m} \alpha_q\alpha_\ell\alpha_m\pi_{q\ell}\pi_{qm}\pi_{\ell m}.$$

The same reasoning can be applied to $\Pr\{\mathsf{V}\}$ recalling that the event V in (i,j,k) means that the top of V is i. Combining all this elements leads to the following :
Proposition : In the ERMG model, the clustering coefficient is

$$c = \sum_{q,\ell,m} \alpha_q\alpha_\ell\alpha_m\pi_{q\ell}\pi_{qm}\pi_{\ell m} \Big/ \sum_{q,\ell,m} \alpha_q\alpha_\ell\alpha_m\pi_{q\ell}\pi_{qm} \tag{4}$$

4.3 Likelihoods

In order to define the likelihood of the ERMG model, we use the complete-data framework defined by [Dempster *et al.* (1977)]. Let us denote $\mathcal{X}$ the set of all edges : $\mathcal{X} = \{X_{ij}\}_{i,j=1..n}$, and $\mathcal{Z}$ the set of all indicator variables for vertices : $\mathcal{Z} = \{Z_{iq}\}_{i=1,n}^{q=1,Q}$.
Proposition : The complete-data log-likelihood is

$$\log \mathcal{L}(\mathcal{X},\mathcal{Z}) = \sum_i \sum_q Z_{iq} \log \alpha_q + \sum_i \sum_q \sum_{j>i} \sum_\ell Z_{iq}Z_{j\ell} \log b(X_{ij};\pi_{q\ell}). \tag{5}$$

Proof : This is a direct consequence of the decomposition $\log \mathcal{L}(\mathcal{X},\mathcal{Z}) = \log \mathcal{L}(\mathcal{Z}) + \log \mathcal{L}(\mathcal{X} \mid \mathcal{Z})$ where $b(x;\pi) = \pi^x(1-\pi)^{1-x}$. ■

The log-likelihood of the observed data is obtained by summing the complete-data log-likelihood over all possible values of the unobserved variables $\mathcal{Z}$.

5 Estimation

We propose in this section a algorithm to estimate the parameters of the ERMG model by approximate maximum likelihood. Since the likelihood involves so many terms and is thus untractable, we will maximize an inferior bound of the likelihood instead of the true one. We could have used the EM algorithm, unfortunately the success of the EM algorithm relies on a simple expression of the conditional law of Z knowing X, which is not the case in the ERMG model.

5.1 Lower bound of the likelihood

We consider a functional $\mathcal{J}$ defined for an ERMG parameter θ and a distribution $R_{\mathcal{X}}$ on the Z_{iq}s by :

$$\begin{aligned} \mathcal{J}(R_{\mathcal{X}},\theta) &= \log \mathcal{L}(\mathcal{X},\theta) - KL(R_{\mathcal{X}}, \Pr(\mathcal{Z}|\mathcal{X},\theta)) \\ &= \mathcal{H}(R_{\mathcal{X}}) + \sum_{\mathcal{Z}} R_{\mathcal{X}}(\mathcal{Z}) \log \mathcal{L}(\mathcal{X},\mathcal{Z},\theta) \end{aligned} \quad (6)$$

where KL is the Kullback-Leibler divergence and $\mathcal{H}$ the entropy. When θ is set to a fixed value, we have $\max_{R_{\mathcal{X}}} \mathcal{J}(R_{\mathcal{X}},\theta) = \log \mathcal{L}(\mathcal{X},\theta)$. So the usual maximum likelihood estimator of θ can be written :

$$\widehat{\theta}_{ML} = \arg\max_{\theta} \max_{R_{\mathcal{X}}} \mathcal{J}(R_{\mathcal{X}},\theta)$$

Unfortunately, $\mathcal{J}$ is hard to compute for a generic distribution $R_{\mathcal{X}}$ so we restrict ourselves to a comfortable set of distributions : the completely factorized distributions on the Z_{iq}., for which the computation is tractable. So, our new estimator is given by :

$$\widehat{\theta} = \arg\max_{\theta} \max_{R_{\mathcal{X}} \text{ factorized}} \mathcal{J}(R_{\mathcal{X}},\theta)$$

Like in the EM, we do a step-by-step optimization of $\mathcal{J}$: we first fix θ and seek the optimal $R_{\mathcal{X}}$, then fix $R_{\mathcal{X}}$ to this value and seek the corresponding optimal θ and so on and so forth until the value of $\mathcal{J}$ converges. Note that $\mathcal{J}$ is bound to increase at each step, so that if the maximum likelihood is finite, the algorithm will converge. We note $\mathcal{J}(\mathcal{X})$ the value of $\mathcal{J}$ once convergence has been achieved.

5.2 Optimization of $R_{\mathcal{X}}$ step

Denoting $\mathcal{Z}_i = (Z_{i1}, \ldots, Z_{iQ})$, we notice that $R_{\mathcal{X}}$, being fully factorized, is of the form :

$$R_{\mathcal{X}}(\mathcal{Z}) = \prod_i h(Z_i|\tau_i)$$

where τ_i is a probability vector, and $h(Z|\tau)$ denotes the multinomial density of parameter τ. Injecting this form of $R_{\mathcal{X}}$ in $\mathcal{J}$ and maximizing with respect to the τ_{iq}s under the constraint $\sum_q \tau_{iq} = 1$ gives

$$\widehat{\tau}_{iq} \propto \alpha_q \prod_{j \neq i} \prod_{\ell=1}^{Q} b(X_{ij}, \pi_{q\ell})^{\tau_{j\ell}}$$

REMARK : The computation of a single τ_{iq} requires to know the value of all other $\tau_{j\ell}$s. We thus use a fixed-point algorithm to compute them.

5.3 Optimization of θ step

At this step, we maximize $\mathcal{J}$ with respect to θ under the constraint $\sum_q \alpha_q = 1$. It gives :

$$\widehat{\alpha}_q = \sum_i \widehat{\tau}_{iq}/n, \quad \widehat{\pi}_{q\ell} = \sum_i \sum_j \widehat{\tau}_{iq}\widehat{\tau}_{j\ell} X_{ij} / \sum_i \sum_j \widehat{\tau}_{iq}\widehat{\tau}_{j\ell}$$

5.4 Choice of the number of groups

Our purpose here is not to derive a specific criterion to select the number of groups in the ERMG model. This problem seems difficult to tackle, especially because the log-likelihood of the observed data $\log \mathcal{L}(\mathcal{X})$ is not calculable.

We propose a heuristic criterion based on ICL [Biernacki *et al.* (2000)], the Integrated Completed Likelihood. The ICL criterion uses the same penalty as BIC, but applies it to the complete-data log-likelihood, which is the only likelihood we can calculate in this case. To compute it, we replace the labels Z_{iq}s in (5) by their expected values, the τ_{iq}s. The first term of (5) deals with Q proportions α_qs which involve n data and the second term with $Q(Q+1)/2$ probabilities $\pi_{q\ell}$s which involve $n(n-1)/2$ terms. Hence the Fisher information matrix derived from $\log \mathcal{L}(\mathcal{X}, \widehat{\mathcal{Z}})$ is proportional to n for the α_qs, while it is proportional to $n(n-1)/2$ for the $\pi_{q\ell}$s.

We therefore propose the following heuristic criterion :

$$-2 \log \mathcal{L}(\mathcal{X}, \widehat{\mathcal{Z}}) + (Q-1) \log n + [Q(Q+1)/2] \log[n(n-1)/2]. \tag{7}$$

6 Application to biological networks

6.1 Presentation of the data

The motivation for applying this methodology to biological networks is twofold : (1) obtain a more realistic random graph model for further work on the over-representation of network motifs ([Shen-Orr *et al.* (2002)]) and reaction motifs ([Lacroix *et al.* (2005)]) ; (2) study the properties of such graphs *per se* to get insight on the modular structure of biological networks.

In this section, we show that the ERMG model is more realistic than other models for describing the degree distribution and the clustering coefficient of a metabolic network. We also show that the groups identified by the method can be given a biological meaning. We apply the methodology developed in this paper to the metabolic network of the bacterium *Escherichia coli*. Although the method is generic and could be applied to other types of biological networks (such as protein interaction networks or transcriptional networks), we chose to first focus on metabolic networks because the dataset is more complete and reliable. In this network, vertices are chemical reactions. Two reactions are connected if they share a primary compound. For each reaction, a distinction is made between its primary compounds (main substrate and product) and its secondary compounds (cofactors). Only primary compounds are responsible for edges. Importantly, the same compound may be considered as primary with respect to one reaction and secondary with respect to another reaction. This method is an alternative way to deal with the known bias introduced by ubiquitous compounds (such as water) which artifactually connect a large number of reactions ([Arita (2004)]). Finally, since the information on the reversibility of reactions does not seem to be established (contradictions may be found within a same database), we chose to consider the general case where all reactions are reversible. The data we used was downloaded from `http ://biocyc.org/`. The resulting graph is made up of $n = 605$ vertices and the total number of edges is 1782.

6.2 Erdös-Rényi mixture modeling

NUMBER OF GROUPS AND PARAMETER ESTIMATES. Using the heuristic criterion defined in (7), we select $Q = 21$ groups. Table 2 gives the estimates of proportions α_q and connection probabilities $\pi_{q\ell}$. Among the first 20 groups, 8 are actually cliques ($\pi_{qq} = 1$) and 6 have within probability connectivity greater than 0.5. We also see that the clique structure strongly increases the mean degree λ_q of its elements. More generally, in this example, it turns out that the within connection probabilities π_{qq} are always maximal, although the modeling does not require this. Simulation studies (not shown) prove that it is not an artefact of the method, which can detect a group with no within connection.

The interpretation for the cliques (and pseudo-cliques) is straightforward, each of them corresponds to a single compound involved in all the reactions of the group. Examples of compounds responsible for cliques include chorismate, pyruvate, L-aspartate, L-glutamate, D-glyceraldehyde-3-phosphate and ATP. This illustrates an already established result : the structure of a metabolic network is mainly due to the presence of a few metabolites, called hubs ([Jeong *et al.* (2000)]). The originality of our data is to remove

the secondary metabolites, which ensures that the hubs identified are meaningful, that is, they really form the backbone of the network.

Groups 1 and 16 are quite instructives. The connection probability between them is 1 so they constitute a clique, corresponding to the single pyruvate compound. However, they connect very differently to groups 7 and 10, mainly because CO2 is involved in most reactions of groups 1 and 7 and acetylCoA in those of groups 7 and 10. Interestingly, our method is able to identify this and to "split" the pyruvate hub into groups 1 and 16

Complementary analysis of the groups show that they gather reactions that participate in the same class of metabolic pathways. For instance, group 1 corresponds to the generation of precursor metabolites, group 2 and 3 correspond to amino-acid biosynthesis, and group 4 to cofactor biosynthesis. This indicates that the groups found by our method are coherent in terms of biological processes.

TAB. 2 – Parameter estimates of the ERMG model with $Q = 21$ groups (values smaller than .5 % are masked for readability).

$\alpha(\%)$	0.7	1.0	1.2	1.3	1.3	1.5	1.5	1.6	1.8	1.8	2.0	2.1	2.3	2.6	2.7	2.8	3.0	3.0	3.3	5.8	56.8
	100						64		11	43			2			100					
		100																			
			100									4	7		1			1			
				71																	
					100	28				1						18			16		
					28	100								6							
	64						58		10	4		7	5			5					
								63				5						3			
	11						10		65							1	2	2			
	43				1		4			67			1								
π											62			7			4				
(%)			4				7	5				28	5			5					
	2		7				5			1		5	100			1					
						6					7			25							
			1												40						
	100				18		5		1			5	1			100					
									2		4						100			6	
			1					3	2									21			
					16														19		
																	6			11	
																					1
λ_q	33	7	9	6	17	13	12	7	10	10	10	8	17	6	7	25	21	5	6	5	3

BETWEEN GROUP CONNECTIVITY AND CLUSTERING COEFFICIENT. The graph showing 1782 edges connecting 605 vertices is of course unreadable. Figure 1 presents the graph as a dot-plot where a dot at row i and column j indicates that the edge $i \leftrightarrow j$ is present. To emphasize the connections between the different groups, we reordered the vertices within groups. The limits between groups are obtained using a maximum a posteriori (MAP) classification of vertices : the vertex i is classified into group q for which $\widehat{\tau}_{iq}$ is maximal.

The bottom plot in Figure 1 gives the estimated *posterior* probabilities $\widehat{\tau}_{iq}$. We see that the first groups are quite well defined. The last one (21) has more fuzzy limits : it is actually made of isolated reactions having not much in common.

Finally, we also compare the expected clustering coefficient c given in Equation 4 with the observed one. The expected value for $Q = 21$ groups is 0.544, while the observed one is 0.626. The ERMG model therefore slightly underestimates this coefficient. On the same dataset, the Erdös-Rényi model would give $\widehat{c} = \widehat{\pi} = 0.0098$.

GOODNESS OF FIT OF THE DEGREE DISTRIBUTION Fig. 2 shows that the Zipf pdf (left) fits only very poorly the degree distribution when compared to the Poisson Mixture deriving from our ERMG model (right). This remains true even when we use a threshold to fit only the tail of the distribution, which the Zipf pdf is supposed to model. Both Zipf and Poisson mixture behave poorly for degrees 1 and 2 because of the unexpected high number of vertices with degree 2 : 144, compared to 54 for degree 1 and 12 for degree 0.

We conclude that the ERMG model provides a random graph model which seems to be well adapted to capture the structure of a biological network. This first application of our model to a biological network is promising in the sense that the groups we find are relevant (coherent sets of reactions gathered around central compounds). Future research directions include the study of probabilistic properties of the ERMG model (diameter, probability for a subgraph to be connected) which would give a strong statistical basis to the study of local structural properties.

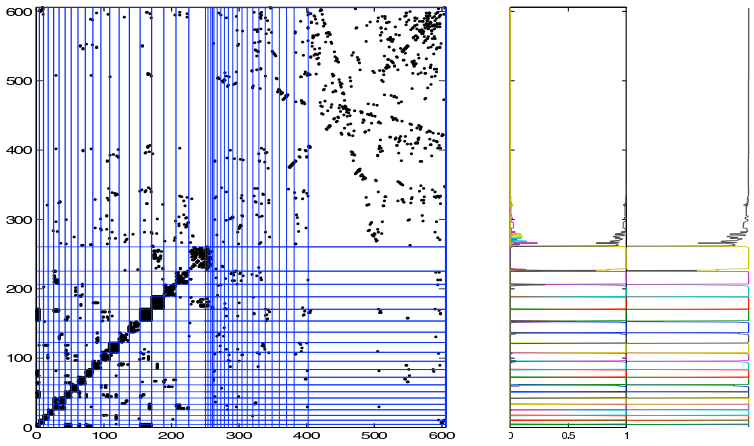


FIG. 1 – Left : Dot plot representation of the graph after classification of the vertices into the 21 groups. Right : Posterior probabilities τ_{iq}.

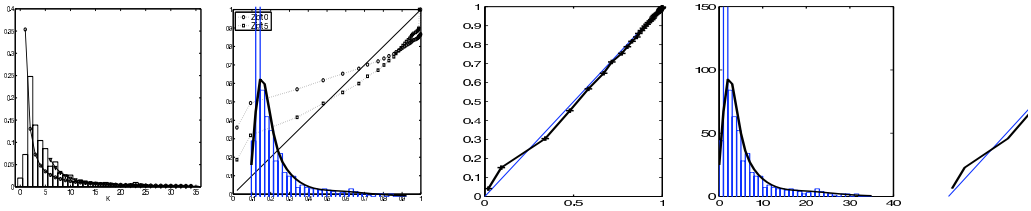


FIG. 2 – Fit of the Zipf (left) and Poisson mixture with $Q = 21$ groups (right) pdf to *E. coli* data. For each, Left : PP plots (Zipf : threshold $1 - \circ -$ and $6 - \triangledown -$). Right : Histogram of degrees with adjusted distributions (top : same thresholds).

Acknowledgements

The authors thank C. Matias, E. Birmelé, (CNRS-Statistic and Genome group, Evry univ.) and S. Schbath (INRA-MIG, Jouy-en-Josas) for all their helpful remarks and suggestions. They also thanks F. Forbes (INRIA Grenoble) for her advices regarding the estimation algorithm.

Références

[Albert and Barabási (2002)] ALBERT, R. and BARABÁSI, A. L. (2002). Statistical mechanics of complex networks. *R. Modern Physics.* **74 (1)** 47–97.

[Arita (2004)] ARITA, M. (2004). The metabolic world of *Escherichia coli* is not small. *PNAS.* **101 (6)** 1543–1547.

[Barabási and Albert (1999)] BARABÁSI, A. L. and ALBERT, R. (1999). Emergence of scaling in random networks. *Science.* **286** 509–512.

[Biernacki *et al.* (2000)] BIERNACKI, C., CELEUX, G. and GOVAERT, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Trans. Pattern Anal. Machine Intel.* **22 (7)** 719–725.

[Dempster *et al.* (1977)] DEMPSTER, A. P., LAIRD, N. M. and RUBIN, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society Series B.* **39** 1–38.

[Jeong *et al.* (2000)] JEONG, H., TOMBOR, B., ALBERT, R., OLTVAI, Z. N. and BARABÁSI, A. L. (2000). The large-scale organization of metabolic networks. *Nature.* **407** 651–654.

[Lacroix *et al.* (2005)] LACROIX, V., GOMES FERNANDES, C. and SAGOT, M. F. (2005). Reaction motifs in metabolic networks. *Proceedings of 5th Workshop on Algorithms for BioInformatics (WABI'05), Lecture Notes in BioInformatics, subseries Lecture Notes in Computer Science.* **3692** 178–191.

[Molloy and Reed (1995)] MOLLOY, M. and REED, B. . (1995). A critical point for random graphs with a given degree sequence. *Rand. Struct. and Algo.* 161–179.

[Newman (2003)] NEWMAN, M. E. J. (2003). *Handbook of Graphs and Networks.* (S. Bornholdt and H. G. Schuster, ed.), chapter Random graphs as models of networks. Wiley-VCH : Berlin.

[Newman (2004)] NEWMAN, M. E. J. (2004). Fast algorithm for detecting community structure in networks. *Phys. Rev.* **E (69)** 066133.

[Newman and Girvan (2003)] NEWMAN, M. E. J. and GIRVAN, M. (2003). *Statistical Mechanics of Complex Networks.* (R. Pastor-Satorras, J. Rubi, and A. Diaz-Guilera, ed.), chapter Mixing patterns and community structure in networks. Springer : Berlin.

[Shen-Orr *et al.* (2002)] SHEN-ORR, S. S., MILO, R., MANGAN, S. and ALON, U. (2002). Networks motifs in the transcriptional regulation network of *escherichia coli*. *Nat. Genetics.* **31** 64–68.

[Zhang *et al.* (2005)] ZHANG, V. L., KING, O. D., WONG, S. L., GOLDBERG, D. S., TONG, A. H. Y., G., L., ANDREWS, B., BUSSEY, H., BOONE, C. and ROTH, F. P. (2005). Motifs, themes and thematic maps of an integrated *saccharomyces cerevisiae* interaction network. *Journal of Biology.* **4 (2)** 1–13.

TITRE en français

Identification de motifs dans les réseaux métaboliques

RÉSUMÉ en français

Cette thèse s'inscrit dans le cadre de l'analyse structurelle des réseaux biologiques. Nous proposons une nouvelle définition de motif dans le contexte des réseaux métaboliques. Un réseau métabolique est modélisé par un graphe coloré et un motif est défini comme un multiensemble de couleurs (une couleur correspond ici à un mécanisme réactionnel). Une occurrence d'un motif est définie comme un ensemble de noeuds connectés et colorés par les couleurs du motif. Nous proposons des algorithmes pour rechercher et inférer de tels motifs, ainsi qu'un critère statistique permettant de décider si un motif est sur-représenté. L'application de nos méthodes au métabolisme d'*Escherichia coli* révèle des structures locales répétées. Nous argumentons que ces structures peuvent être interprétées comme des blocs fontionnels et/ou évolutifs du métabolisme.

MOTS-CLEFS en français

métabolisme, graphe, algorithmique, motif, réseau, évolution, opéron

TITRE en anglais

Motif identification in metabolic networks

RÉSUMÉ en anglais

This thesis lies within the scope of the structural analysis of biological networks. We propose a new definition of motif in the context of metabolic networks. A metabolic network is modelled by a coloured graph and a motif is defined as a multiset of colours (a colour corresponds to a reaction mechanism). An occurrence of a motif is defined as a set of connected nodes coloured by the colours of the motif. We propose algorithms to search for and infer such motifs, as well as a statistical criterion to decide if a motif is over-represented. The application of our methods to the metabolism of *Escherichia coli* reveals repeated local structures. We argue that theses structures can be interpreted as functional and/or evolutionary blocks of metabolism.

DISCIPLINE : Bioinformatique

MOTS-CLEFS en anglais

metabolism, graph, algorithms, network, motif, evolution, operon

INTITULÉ ET ADRESSE DE L'U.F.R. OU DU LABORATOIRE :

Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS
Batiment Gregor Mendel - Université Claude Bernard Lyon 1
43, bv du 11 novembre 1918 - 69622 Villeurbanne cedex