Système de recommandation basé sur les réseaux pour l'interprétation de résultats de métabolomique

Clément Frainay

Université Fédérale

Toulouse Midi-Pyrénées

THÈSE

En vue de l'obtention du

DOCTORAT DE L'UNIVERSITÉ DE TOULOUSE

Délivré par :

Université Toulouse 3 Paul Sabatier (UT3 Paul Sabatier)

Présentée et soutenue par :

Clément Frainay

le 26/06/17

Titre :

Système de recommandation basé sur les réseaux pour l'interprétation de résultats de métabolomique.

École doctorale et discipline ou spécialité :

ED SEVAB : Pathologie, Toxicologie, Génétique et Nutrition

Unité de recherche :

INRA TOXALIM

Directeur/trice(s) de Thèse :

Fabien J ourdan, Directeur de Recherche, INRA
Daniel Zalko, Directeur de Recherche, INRA

Jury :

Anne Siegel, Directeur de Recherche, IRISA-CNRS
Philippe Schmitt-Kopplin, Professeur, Helmholtz Zentrum München
Christophe J unot, Cadre scientifique des EPIC, CEA
J ean-Charles Portais, Professeur, UPS

Résumé

La métabolomique permet une étude à large échelle du profil métabolique d'un individu, représentatif de son état physiologique. La comparaison de ces profils conduit à l'identification de métabolites caractéristiques d'une condition donnée. La métabolomique présente un potentiel considérable pour le diagnostic, mais également pour la compréhension des mécanismes associés aux maladies et l'identification de cibles thérapeutiques. Cependant, ces dernières applications nécessitent d'inclure ces métabolites caractéristiques dans un contexte plus large, décrivant l'ensemble des connaissances relatives au métabolisme, afin de formuler des hypothèses sur les mécanismes impliqués. Cette mise en contexte peut être réalisée à l'aide des réseaux métaboliques, qui modélisent l'ensemble des transformations biochimiques opérables par un organisme. L'une des limites de cette approche est que la métabolomique ne permet pas à ce jour de mesurer l'ensemble des métabolites, et ainsi d'offrir une vue complète du métabolome. De plus, dans le contexte plus spécifique de la santé humaine, la métabolomique est usuellement appliquée à des échantillons provenant de biofluides plutôt que des tissus, ce qui n'offre pas une observation directe des mécanismes physiologiques eux-mêmes, mais plutôt de leur résultante. Les travaux présentés dans cette thèse proposent une méthode pour pallier ces limitations, en suggérant des métabolites pertinents pouvant aider à la reconstruction de scénarios mécanistiques. Cette méthode est inspirée des systèmes de recommandations utilisés dans le cadre d'activités en ligne, notamment la suggestion d'individus d'intérêt sur les réseaux sociaux numériques. La méthode a été appliquée à la signature métabolique de patients atteints d'encéphalopathie hépatique. Elle a permis de mettre en avant des métabolites pertinents dont le lien avec la maladie est appuyé par la littérature scientifique, et a conduit à une meilleure compréhension des mécanismes sous-jacents et à la proposition de scénarios alternatifs. Elle a également orienté l'analyse approfondie des données brutes de métabolomique et enrichie par ce biais la signature de la maladie initialement obtenue. La caractérisation des modèles et des données ainsi que les développements techniques nécessaires à la création de la méthode ont également conduit à la définition d'un cadre méthodologique générique pour l'analyse topologique des réseaux métaboliques.

Abstract

Metabolomics allows large-scale studies of the metabolic profile of an individual, which is representative of its physiological state. Metabolic markers characterising a given condition can be obtained through the comparison of those profiles. Therefore, metabolomics reveals a great potential for the diagnosis as well as the comprehension of mechanisms behind metabolic dysregulations, and to a certain extent the identification of therapeutic targets. However, in order to raise new hypotheses, those applications need to put metabolomics results in the light of global metabolism knowledge. This contextualisation of the results can rely on metabolic networks, which gather all biochemical transformations that can be performed by an organism. The major bottleneck preventing this interpretation stems from the fact that, currently, no single metabolomic approach allows monitoring all metabolites, thus leading to a partial representation of the metabolome. Furthermore, in the context of human health related experiments, metabolomics is usually performed on bio-fluid samples. Consequently, those approaches focus on the footprints left by impacted mechanisms rather than the mechanisms themselves. This thesis proposes a new approach to overcome those limitations, through the suggestion of relevant metabolites, which could fill the gaps in a metabolomics signature. This method is inspired by recommender systems used for several on-line activities, and more specifically the recommendation of users to follow on social networks. This approach has been used for the interpretation of the metabolic signature of the hepatic encephalopathy. It allows highlighting some relevant metabolites, closely related to the disease according to the literature, and led to a better comprehension of the impaired mechanisms and as a result the proposition of new hypothetical scenario. It also improved and enriched the original signature by guiding deeper investigation of the raw data, leading to the addition of missed compounds. Models and data characterisation, alongside technical developments presented in this thesis, can also offer generic frameworks and guidelines for metabolic networks topological analysis.

Remerciements

Je tiens tout d'abord à remercier Anne Siegel et Philippe Schmitt-Kopplin d'avoir accepté d'être rapporteurs de cette thèse, ainsi que Christophe Junot et Jean-Charles Portais pour avoir accepté de participer à son évaluation.

Je remercie également Vincent Lacroix, Étienne Thévenot et Jean-Philippe Antignac d'avoir accepté d'être membres de mon comité de pilotage et pour tous leurs précieux conseils.

Je remercie Daniel Zalko de m'avoir accueilli au sein de son équipe et pour avoir soutenu ce projet dès ses débuts.

Je souhaite également adresser un grand merci à Fabien Jourdan qui m'a encadré tout au long de cette thèse, pour la confiance qu'il m'a accordée, son soutien sans failles, sa patience et sa bonne humeur.

Un grand merci à Sandrine Aros et une fois encore à Christophe Junot, pour avoir cru en ce projet et pour leurs nombreuses contributions, de la génération des données à l'interprétation des résultats. Je remercie également l'ensemble de leurs équipes respectives du CEA et de MedDay. Sans l'implication des biologistes et expérimentateurs, cette thèse aurait été vide de sens.

Un grand merci à Marie-France Sagot et Arnaud Mary, ainsi qu'a l'ensemble de leur équipe de l'INRIA, pour leur aide précieuse sur la partie algorithmique et leur accueil chaleureux.

Je tiens également à remercier toutes les personnes avec qui j'ai eu la chance de collaborer au cours de cette thèse sur divers projets, et qui sont venues enrichir cette expérience. Je remercie tout particulièrement Yoann Pitarch, Sandra Therrien-Laperrière, Benedict Yanibada, Pierre Millard ainsi que Franck Giacomoni, Nils Paulhe, Yoann Gloaguen, Karl Burgess pour les éprouvants hackatons.

Je remercie également l'ensemble du groupe MetExplore, Florence, Ludovic, Nathalie, Maxime, Sanu, Benjamin, Florence M., Maxime D., Laurent et Thomas, travailler au sein de cette équipe a été non seulement très enrichissant, mais la très bonne ambiance qui y règne a également largement contribué au bon déroulement de cette thèse. Un grand merci à Nathalie pour les relectures et les discussions

enrichissantes, et un grand merci à Maxime pour l'interfaçage de la méthode et pour nos échanges imagés. Je remercie également chaleureusement Ludovic Cottret pour son aide dans le refacoring de la library.

Un grand merci à Thomas Garcia pour son investissement dans la partie Text-Mining, avec qui cela a été un plaisir de travailler, et dont les qualités scientifiques et humaines ont rendu ma première expérience d'encadrement très aisée.

Je remercie également l'ensemble de l'équipe MeX pour leur accompagnement, en particulier Cyndel, Davy, Élodie, Florence, François, Laura, Laure, Louisa, Nicolas, Marc, Marie, Vincent, ainsi que tous les autres membres de ToxAlim (particulièrement l'équipe d'AXIOM) et de l'école vétérinaire qui ont rendu mon passage au laboratoire très agréable : Caroline, Cécile, Émilien, Jean-Philippe, Laurence, Marie, Marion, Marianne, Thaïs, Vanessa et tous les autres.

Je remercie également tout le personnel de l'INRA ToxAlim, et notamment Marie-Hélène pour son aide cruciale et pour avoir toléré avec autant de patience mon inaptitude face aux questions relatives à l'administration. Je remercie aussi Pierre et Aurore, la cantine de l'INRA ayant joué un rôle non négligeable dans ma volonté de continuer mon parcours professionnel au sein de cet établissement.

Je tiens également à remercier l'équipe pédagogique du master de bioinformatique de Toulouse et les membres de l'école doctorale SEVAB de m'avoir mené jusqu'ici.

Je souhaite également remercier mes parents, ma sœur ainsi que mes amis pour leur soutien, en particulier le groupe de l'Option 4, Élise, Yoann, Alexis ainsi que Antoine et Florent.

Enfin, je souhaite remercier tout particulièrement Camille pour son aide et son soutien permanent, l'influence positive qu'elle a sur moi ainsi que pour sa patience, qui lui a permis de me supporter pendant toute la durée de cette thèse.

Table des matières

Table des figures

Liste des tableaux

Première partie

Contexte : Observer, modéliser et comprendre le métabolisme

Chapitre 1

La métabolomique, comment observer le métabolisme

"To understand the whole one must study the whole."

Henrik Kacser, The Organization of cell metabolism

1.1 Définition

La métabolomique est à la fois une discipline scientifique et un regroupement de techniques, qui visent à l'étude à large échelle des métabolites contenus dans des échantillons biologiques[86]. Cet ensemble de métabolites, appelé métabolome, constitue un profil métabolique qui caractérise l'état physiologique d'un système biologique (organisme, cellule, tissus) à un instant donné[209][238][200]. La comparaison de ces profils offre un outil remarquable pour identifier des marqueurs spécifiques d'une condition environnementale ou génétique. La métabolomique a notamment été employée dans le contexte biomédical pour définir des profils caractéristiques d'une maladie[246]. Ces profils caractéristiques peuvent être utilisés à des fins de diagnostic[74], mais permettent également d'élucider les mécanismes conduisant aux symptômes et signes cliniques[137][14]. Cette dernière application nécessite cependant d'aller au-delà du métabolome, en replaçant les métabolites dans le contexte global du fonctionnement du métabolisme[144][202]. La métabo-

lomique est une discipline récente inspirée d'autres méthodes dites «omiques» qui proposent une approche holistique de la biologie. La métabolomique se heurte par conséquent à une difficulté commune aux approches omiques, celle de l'interprétation des résultats dans un système faisant intervenir des miliers de molécules et macromolécules[144]. La contextualisation des résultats de métabolomique va donc nécessiter l'agrégation des connaissances relatives au métabolisme, ainsi que la recherche des mécanismes et métabolites en lien avec un profil donné.

Le but de la biologie des systèmes est de pouvoir offrir une vue d'ensemble qui combine la totalité des niveaux observés au travers de ces approches omiques, représentés figure 1.1.

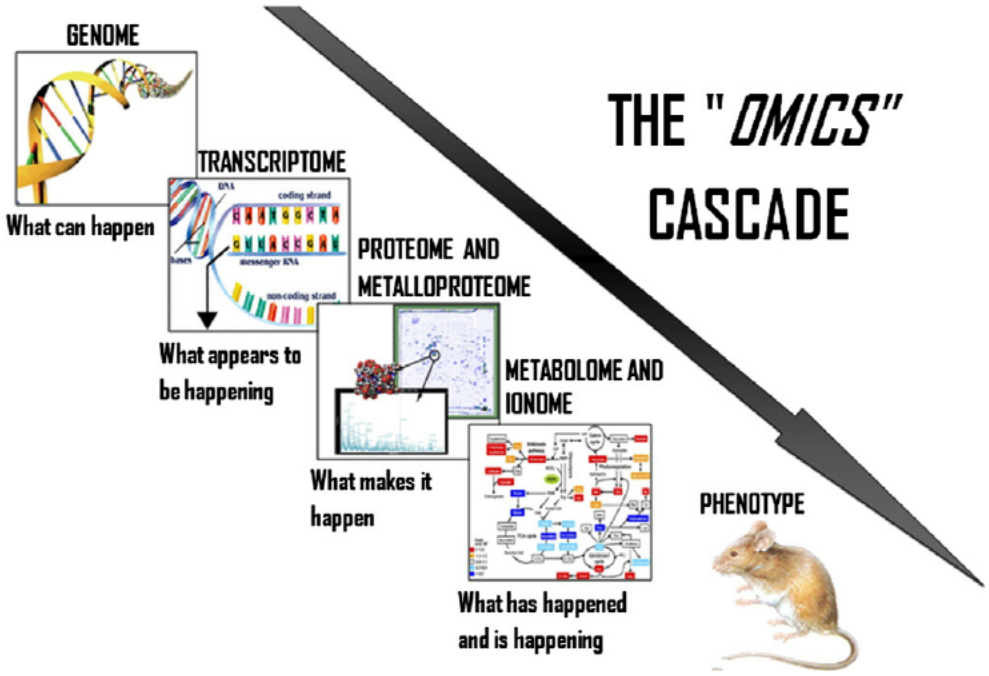


Figure 1.1 – Cascade des différents niveaux de mécanismes biologiques observés par les approches omiques. Illustration issue de García-sevillano *et al.*, 2014[98]

Cependant, chaque niveau est soumis à des contraintes qui lui sont propres, et à des limites méthodologiques spécifiques qui vont impacter sa modélisation. Cette thèse a pour vocation d'identifier ces contraintes et limitations dans le cadre de la métabolomique, et de proposer une méthodologie adaptée pour interpréter

les marqueurs observés par cette approche.

1.2 Acquisition des données

1.2.1 La Spectrométrie de Masse

La spectrométrie de masse est une technique permettant de détecter et identifier des molécules grâce à leur masse et leur charge[102]. Elle repose sur trois étapes : l'ionisation, la séparation en fonction du rapport masse sur charge et le traitement du signal. Différentes méthodes existent pour chacune de ces étapes, chacune pouvant être plus ou moins adaptée à un type d'échantillon, une classe de molécules ou un couplage avec une autre technique. Les bases théoriques dont sont issues les différentes composantes de la spectrométrie de masse, de la préparation d'échantillons au traitement du signal, prennent leurs sources dans différents domaines de la biochimie, la chimie, les mathématiques et de la physique. Les élaborations méthodologiques qui ont marqué son histoire depuis le début de XXe siècle ont conduit à de nombreux prix Nobel en physique et en chimie, et à de nombreuses applications au-delà de l'étude du métabolisme. La section suivante n'a donc pas vocation à proposer une revue exhaustive de ce champ d'études, mais à décrire succinctement le principe fondamental des techniques employées. Ces informations permettent de caractériser les données obtenues en fin d'analyse et ainsi de préciser l'objectif des algorithmes proposés dans cette thèse.

L'étape préliminaire à la spectrométrie de masse est la préparation des échantillons. L'extraction des métabolites d'un échantillon biologique, via par exemple une extraction de contenus cellulaires, va constituer la première déviation entre le métabolome réel et les observations réalisées en métabolomique. Elle est en effet affectée par la dégradation des métabolites inhérente à la technique de préparation des échantillons et à l'incorporation de contaminants chimiques ou biologiques[233]. Elle est également soumise à l'inertie du métabolisme due à l'extraction conjointe des enzymes et des métabolites et de la latence avant leurs séparations. L'échantillonnage même (au sens statistique) peut également être

source de biais. Dans le cas où il serait conduit sur une population cellulaire, les individus qui la composent sont soumis aux effets locaux du micro-environnement, et différents types cellulaires peuvent y être représentés. Dans le cas de l'analyse de biofluides, les métabolites observés peuvent êtres issus de l'excrétion depuis différents tissus ou organes, ou résulter d'une absorption (molécules issus de l'alimentation ou de la prise de médicaments par exemple). De plus, les perturbations affectant des métabolites non circulants ne seront pas observables à partir du biofluide, compliquant la formulation de scénarios mécanistiques. Étant donné que le métabolome est représentatif d'un état physiologique général, une grande variabilité interindividuelle est également observée dans de nombreux cas. Les implications de toutes ces limites sur l'interprétation biologique des résultats n'ont, à notre connaissance, pas été caractérisés. Il est à noter que des analyses statistiques permettent de limiter certains des biais mentionnés dans cette section.

La première phase de la spectrométrie de masse est l'ionisation, qui va consister à attribuer une charge aux molécules. Cette charge est essentielle à la fois pour la séparation des molécules et pour leur détection. Le choix de la méthode d'ionisation va en partie dépendre de la nature de l'échantillon : sa phase (liquide, solide, gazeux), la quantité disponible ou encore sa concentration en sels. Elle va impacter le degré de fragmentation des molécules, qui conditionne la facilité avec laquelle elles seront identifiées en sortie du système. Ce degré de fragmentation joue également un rôle primordial dans l'identification de molécules inconnues. Toutes les molécules ne peuvent être ionisées de la même manière, l'ionisation dépendant des propriétés physico-chimiques de ces dernières. Ainsi, certaines méthodes ne peuvent pas ioniser certains composés qui seront donc omis durant l'analyse. Par exemples, certaines techniques vont omettre les composés non volatils, thermolabiles, de tailles hors d'une certaine gamme, d'une certaine polarité, ou encore les métaux.

La seconde phase est celle de la séparation. Un analyseur va permettre de « trier » les composés par rapport à leur ratio masse sur charge. De manière générale, l'analyseur va séparer les ions en fonction de leurs comportements lorsqu'ils

sont soumis à un champ magnétique ou électrodynamique. En effet, ces comportements vont être caractéristiques de la masse et de la charge des ions. Les techniques diffèrent principalement par les comportements physiques analysés, parmi lesquels la trajectoire des ions (point d'impact, fréquences d'oscillations) ou leur vitesse de transmission. L'impact de ce choix méthodologique sur les données obtenues est principalement lié à la sensibilité et la résolution de ces techniques. Certaines molécules, bien qu'ionisées, ne pourront être séparées. Ceci va conduire à une ambiguïté lors de la phase d'identification, qui consiste à faire le lien entre un signal et la nature du composé à son origine. L'analyseur est couplé à un détecteur afin d'enregistrer le comportement des ions. La détection des ions se fait usuellement par transfert de charge. Les différents détecteurs se distinguent essentiellement par leur capacité à amplifier le signal, leur sensibilité et l'importance du bruit de fond qu'ils génèrent.

La dernière phase est celle du traitement du signal. Elle correspond à l'extraction des données à partir des spectres obtenus. La procédure la plus courante consiste à détecter des pics, intégrer leurs aires sous la courbe, et réaliser les alignements des spectres des différents échantillons pour comparer les aires obtenues. La difficulté de traitement des données obtenues sur des mélanges complexes a conduit à un usage généralisé de méthodes automatiques dédiées à cette étape[188][245]. Une fois encore de nombreuses méthodes existent, et à ce jour aucune d'entre elles n'est suffisamment générique pour être appliquée à tout type de données de spectrométrie de masse tout en présentant un taux d'erreur acceptable. Les principales erreurs induites par ces outils sont la non-détection de pics, l'intégration de pics correspondant à du bruit de fond et les erreurs d'alignements ou d'intégrations. Ces erreurs sont particulièrement fréquentes dans les régions de faibles intensités.

Des analyses multivariées sont ensuite conduites sur ces données extraites des spectres afin de réduire leur dimensionnalité, en se focalisant sur les variables d'importances. Ces variables peuvent être, par exemple, celles qui permettent de classer un échantillon dans un groupe d'intérêt (généralement, un groupe contrôle et un groupe soumis à une perturbation). Ces méthodes apportent également leurs

lots de limitations (le sur-ajustement, ou *overfitting*, étant l'une des préoccupations majeures de la communauté[148]), et le choix d'une méthode appropriée va dépendre essentiellement de la question posée.

La dernière étape du traitement de données correspond à l'identification des métabolites. Dans le cas de la spectrométrie de masse, il peut exister des centaines de molécules différentes qui correspondent à une même masse ou une même composition atomique, et la levée de cette ambiguïté va constituer une étape critique pour l'interprétation biologique des résultats. Le spectromètre de masse peut être couplé à un dispositif en amont permettant une première séparation, basée sur des propriétés physicochimiques autres que la masse (chromatographie liquide ou gazeuse, par exemple), afin de lever des ambiguïtés lors de l'identification. La spectrométrie en tandem est également employée pour étendre l'identification, du fait d'une caractérisation plus fine de la structure des composés. Elle consiste à fragmenter les ions issus d'une première séparation afin de réaliser une deuxième spectrométrie de masse sur les fragments obtenus. De nombreux efforts ont été fournis par la communauté pour faire émerger un consensus sur les standards d'identification en métabolomique[60]. Un système de classification a été proposé pour rapporter les niveaux d'identification[253] :

— identification de niveau 1 : composés identifiés. Ce niveau d'identification est obtenu par comparaison avec le spectre généré par le passage d'un standard, une solution pure de la molécule candidate, dans des conditions expérimentales identiques.
— identification de niveau 2 : annotations putatives des composés. Ce niveau correspond à l'annotation basée sur des propriétés physicochimiques ou une similarité spectrale par rapport aux spectres contenus dans des bases de données ou dans la littérature. Il est à noter que les grandes variabilités des protocoles et des instruments utilisés en métabolomique, rendent la comparaison de spectre difficile, voire équivoque.
— identification de niveau 3 : attribution d'une classe de molécule à un composé. Ce niveau d'identification est usuellement obtenu à l'aide des mêmes

méthodes que le niveau 2, mais diffère de ce dernier par le fait que la levée d'ambiguïté est restreinte à une classe de molécule. Il s'agit donc plus d'une caractérisation du composé que d'une identification.

— identification de niveau 4 : composés ni identifiés ni classés, mais qui peuvent être différenciés depuis les données spectrales.

Les métabolites issus d'identification de niveau 1, voire de niveau 2, sont les seuls qui pourront être pleinement exploités à des fins de reconstruction de scénarios mécanistiques. À cette limitation s'ajoute l'ensemble des métabolites qui ne peuvent être détectés du fait d'une incompatibilité de leurs propriétés physicochimiques avec les techniques employées[11]. Selon l'hypothèse où les métabolites impliqués dans un même processus biologique présenteraient une grande similarité chimique entre eux (et par conséquent des propriétés physicochimiques communes) alors, ce sont des processus entiers qui peuvent être hors de portée.

1.2.2 La Résonance Magnétique Nucléaire

La spectrométrie par Résonance Magnétique Nucléaire (RMN) constitue une autre approche couramment employée en métabolomique[163][147]. Tout comme la spectrométrie de masse, cette technique repose sur des phénomènes physiques et les développements qui ont conduit à son utilisation généralisée, ainsi que les nombreuses applications qui en découlent (couvrant un champ bien plus vaste que la métabolomique) ont fait l'objet de nombreux travaux dont certains récompensés par des prix Nobel. La partie suivante propose une description succincte du principe fondamental et des techniques utilisées, afin d'identifier leurs implications sur l'interprétation des résultats générés par la spectrométrie par RMN.

La détection des métabolites repose sur une propriété physique de certains noyaux des atomes les constituant, la résonance magnétique nucléaire. Le phénomène RMN consiste en une absorption d'une onde électromagnétique par ces atomes qui sera réémise avec une fréquence de résonance caractéristique des propriétés magnétiques de leurs isotopes et de leur environnement (interactions intra et intermoléculaires). La mesure de ces fréquences permet donc de nous renseigner,

entre autres, sur la nature d'un atome et celle de ses voisins, la nature des liaisons chimiques et les conformations moléculaires. La combinaison de ces informations permet une identification des molécules qui s'avère souvent moins ambiguë que lorsqu'elle repose uniquement sur un ratio masse/charge.

Contrairement à la spectrométrie de masse, la spectrométrie par RMN présente une variété de protocoles moindre. Elle dispose d'une grande spécificité et d'une grande reproductibilité, qui permet une comparaison aisée des résultats de différentes analyses. Le principal choix méthodologique est celui du type d'isotope considéré (proton, carbone...).

La spectrométrie par RMN offre une gamme plus large de molécules détectables, qui n'est pas bornée par les propriétés physicochimiques de ses constituants telles que l'hydrophobicité ou leur constante d'acidité. En revanche, elle s'avère être une méthode peu sensible par rapport à la spectrométrie de masse : seuls les métabolites d'une concentration suffisamment élevée seront détectables. Cette limite a conduit à une utilisation plus restreinte de cette méthode pour les analyses non ciblées ayant pour but d'élucider les mécanismes sous-jacents d'une perturbation. En revanche, la reproductibilité de cette technique et son caractère quantitatif (là où la spectrométrie de masse est souvent restreinte à une comparaison relative d'abondance) en font une méthode particulièrement adaptée au profilage.

1.3 Bilan : une vue partielle du métabolome

Il est à noter que les biais induits lors de l'extraction et de la préparation des échantillons, ainsi que les limites liées à l'identification, sont communes aux deux méthodes présentées. Dans l'éventualité où tous les métabolites mesurés (ou au moins les métabolites d'intérêt) seraient identifiés sans ambiguïté, l'interprétation des résultats n'en reste pas moins délicate.

Ainsi, quelle que soit la méthode utilisée, les métabolites observés ne représentent qu'une vue partielle du métabolome[233](Figure 1.2). De plus, cette vue ne permet pas de capturer le dynamisme du métabolisme : l'observation des méta-

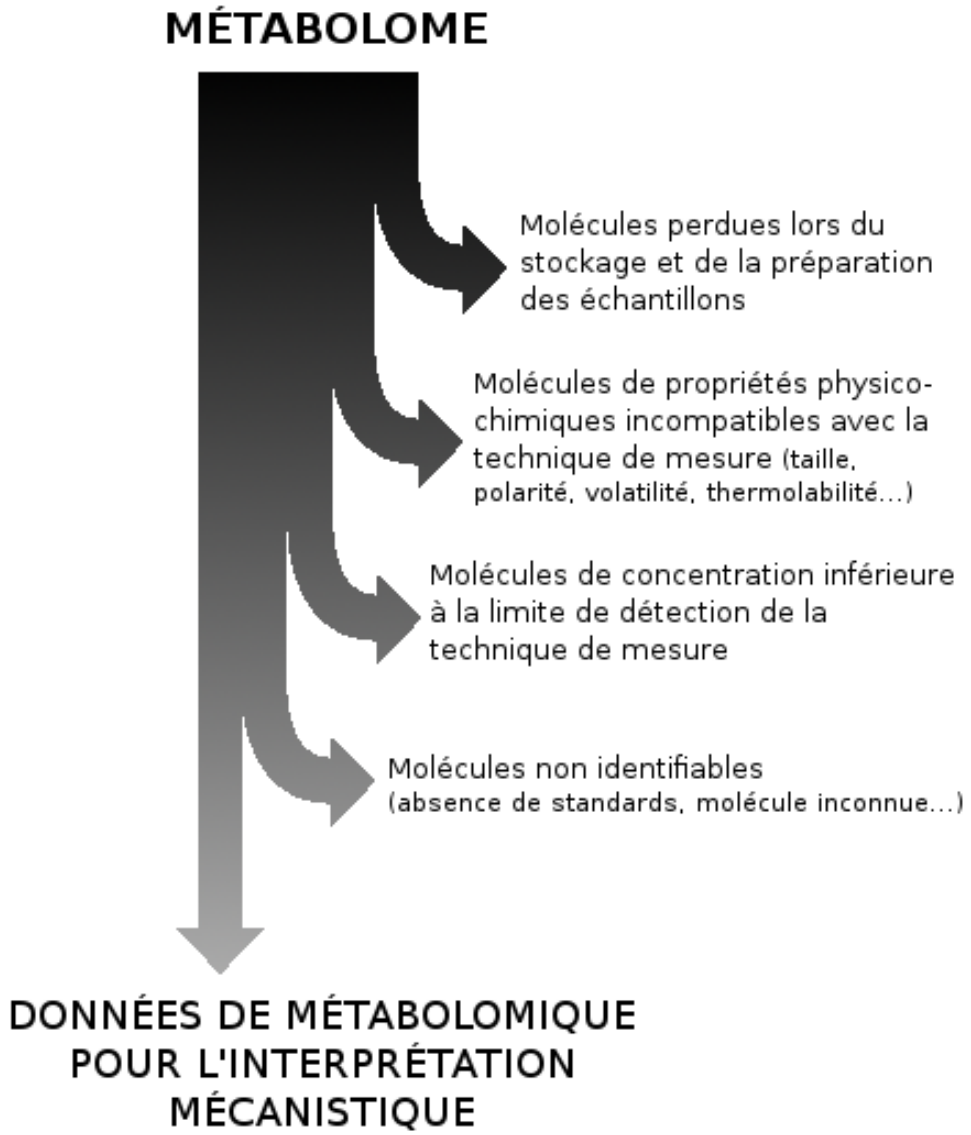


Figure 1.2 – Différence entre le métabolome réel et les données exploitables pour l'interprétation mécanistique

bolites présents de manière transitoire va être peu probable. Enfin, le métabolome est représentatif d'un échantillon biologique, qui par définition ne capture pas l'état physiologique de l'organisme complet ou d'une population dans sa globalité.

Chapitre 2

Les réseaux métaboliques, comment modéliser le métabolisme

2.1 Définitions

"Essentially, all models are wrong, but some are useful."

George E. P. Box, Norman R. Draper, Empirical Model-Building and Response Surfaces.

2.1.1 Notion de modèle

Un modèle peut être défini comme un objet représentatif d'une entité ou d'un phénomène, dont l'étude permet de formuler des hypothèses sur le comportement de l'objet ou du phénomène représenté. De nombreux modèles en sciences, dont les modèles du métabolisme, sont le fruit d'une abstraction par simplification. Cette méthodologie consiste à raffiner les propriétés d'un système en supprimant celles dont l'implication dans le mécanisme à inférer est supposée *a priori* minime, ainsi que celles dont la mesure est difficilement réalisable, voire impossible. Un modèle est donc reconnu approximatif mais néanmoins valide à condition que le produit de la simplification soit suffisant pour garantir la vraisemblance des prédictions

réalisées à partir du modèle. Cette caractéristique ne peut être définie que pour un objectif donné, c'est à dire la question à laquelle on tente de répondre grâce au modèle.

De nombreux modèles existent pour décrire des phénomènes en biologie[178]. Celui utilisé dans cette thèse pour modéliser le métabolisme, le réseau métabolique, est un modèle conceptuel qui vise essentiellement à la compréhension, et peut donc être employé à plusieurs desseins. Contrairement aux modèles dédiés à la prédiction, ce dernier peut également être utilisé pour l'élaboration de nouvelles théories, la découverte de nouvelles relations ou encore la correction de perceptions erronées de la réalité. Il conviendra donc de statuer sur la représentativité du modèle vis à vis des différents phénomènes à étudier, ainsi que sur la qualité des résultats obtenus vis-à-vis de la réalité. En d'autres termes, il s'agira d'identifier les écarts du modèle à la réalité et leurs implications sur la validité des hypothèses produites.

Cet objectif constituera le fil conducteur de la première partie de cette thèse.

L'abstraction faite pour la modélisation du métabolisme dans cette thèse se traduit par sa réduction au potentiel d'un organisme donné en matière de transformations chimiques. Les réductions nécessaires à l'élaboration du modèle seront détaillées dans la partie 2.2 dédiée à leur reconstruction.

2.1.2 Représentation du métabolisme sous forme de réseau

Le formalisme choisi pour modéliser le métabolisme est la représentation sous forme de réseau[160][207]. Un réseau est défini par un ensemble de composants et les relations au sein de cet ensemble. Cette représentation du métabolisme est donc centrée sur les liens qui existent entre les métabolites (au travers des réactions biochimiques), plutôt que sur les métabolites eux-mêmes. De ce fait, les informations relatives à leur structure et leurs propriétés physicochimiques sont rarement intégrées dans ces réseaux. Les liens quant à eux sont issus des informations sur les réactions. Les relations au sein de ces réseaux sont donc orientées, en fonction du sens des réactions.

Certaines réactions étant réversibles, deux orientations d'une même relation peuvent coexister. Théoriquement, toutes les réactions sont réversibles. Néanmoins, certaines directions présentent un coût énergétique prohibitif, ou d'autres encore peuvent impliquer des substrats non disponibles en conditions physiologiques. Le dioxyde de carbone par exemple, ne reste que de manière très transitoire à l'état gazeux dans les cellules, il est donc cantonné à être essentiellement produit, et rarement consommé dans cet état[159].

2.2 Reconstruction des réseaux métaboliques

> *"Le savant doit ordonner ; on fait la science avec des faits comme une maison avec des pierres, mais une accumulation de faits n'est pas plus une science qu'un tas de pierres n'est une maison."*
>
> Henri Poincaré

La création des réseaux métaboliques est couramment nommée « reconstruction » par la communauté. Bien qu'il soit difficile d'attribuer la paternité de cette appellation, l'un des travaux pionniers dans ce domaine est celui de Gaasterland et Selkov, publié en 1995 dans l'article « Reconstruction of Metabolic Networks Using Incomplete Information »[96]. Une reconstruction peut être définie par l'action de « rétablir dans sa forme première », et les reconstructions métaboliques visent à définir le métabolisme en tant que tout, plutôt qu'en une collection de mécanismes isolés. Les réseaux métaboliques sont ainsi créés à partir du regroupement et du réarrangement de fragments d'informations hétérogènes préexistants. Ces informations sont issues de diverses sources, en particulier de l'annotation de génome et de la description d'activités enzymatiques.

Des travaux plus récents ont tenté de créer des réseaux métaboliques à partir d'observations directes du métabolisme. Ces approches seront décrites dans la section suivante, et sont regroupées sous l'appellation de « reconstruction *ab initio* »[44]. Le terme reconstruction souligne le caractère incomplet des réseaux métaboliques. Ils sont le reflet d'une connaissance imparfaite du métabolisme, et il

est fréquent d'observer en métabolomique des métabolites absents de ces réseaux. Le terme de « modèle » est également fréquemment employé au sein de la communauté pour désigner les réseaux « globaux », particulièrement dans le contexte des simulations de flux.

Le terme *genome-scale* est fréquemment apposé à celui de réseau métabolique lorsque ce dernier est réalisé à l'échelle de l'ensemble des réactions d'un organisme (c.-à-d. catalysées par les produits de ses gènes). Ce qualificatif permet de faire la distinction entre ces réseaux et ceux réalisés à l'échelle d'une voie. Pour plus de concision, le terme de réseau métabolique réfèrera dans cette thèse aux réseaux métaboliques « *genome-scale* ». Toute partie de ces réseaux sera référée sous le terme de « sous-réseau ».

2.2.1 Reconstruction à partir du génome

La plupart des réseaux métaboliques d'un organisme sont reconstruits à partir de la séquence complète de son génome[255]. Un prérequis à l'utilisation de ce génome est son annotation, qui permet d'attribuer une fonction aux différents ensembles qui constituent sa séquence. Elle repose en premier lieu sur la détection de ces ensembles, les unités fonctionnelles, qui seront ici réduites aux gènes codants, seuls à être considérés pour la reconstruction des réseaux. Cette détection constitue l'annotation dite « structurelle ». Elle peut être réalisée au travers de l'analyse de fréquence des codons, plus conservés dans le cas d'une partie codante (qui aurait un intérêt pour la survie de l'organisme), et également par la recherche de motifs. En effet, certains motifs sont caractéristiques des « bornes » des gènes, tels que des séquences promotrices où se fixent les facteurs de transcription, ou des sites de fixation des ribosomes.

La seconde étape consiste à réaliser l'annotation fonctionnelle des gènes. Elle repose sur tout un historique de découvertes scientifiques issues de la génétique, de la biochimie ou de la biologie moléculaire, allant de l'association de phénotypes à des mutations aux tests *in vitro* d'activité. Cette méthodologie manuelle ne pouvant être appliquée à l'échelle d'un génome entier, les fonctions sont en règle

générale inférées à partir de l'homologie d'une séquence de fonction inconnue avec celle d'un gène déjà annoté dans un autre organisme.

Une fois l'étape d'annotation fonctionnelle réalisée, le réseau métabolique peut être reconstruit par agrégation de toutes les réactions catalysées par les enzymes dont les gènes ont été identifiés dans le génome de l'organisme.

L'hypothèse que deux enzymes homologues issues de deux organismes différents catalysent les mêmes réactions constitue une hypothèse forte. Il est à noter que la notion d'homologie est une interprétation de la similarité, elle demeure de ce fait une notion partiale. La propagation des annotations entre les génomes peut donc conduire à la propagation des erreurs d'annotations, cumulées à chaque transfert. Ainsi, en l'absence de validation expérimentale, les annotations des génomes restent putatives, et donc la reconstruction du réseau métabolique qui en découle tout autant.

Un réseau obtenu par reconstruction automatique est considéré comme incomplet[100][191], et la présence de faux positifs probable, s'en suivent alors les étapes de raffinements et de validations (Fig 2.1). La partie raffinement va consister en des modifications manuelles du réseau sur la base de résultats expérimentaux ou à partir de la littérature spécialisée. Ces modifications peuvent être des suppressions liées à des annotations erronées, mais également des ajouts. En effet, il est envisageable que des gènes n'aient pas encore été annotés chez l'organisme considéré. Il est également possible que l'organisme considéré puisse effectuer des fonctions métaboliques qui lui sont spécifiques, et qui par définition ne peuvent être propagées depuis les annotations d'un autre organisme. Enfin, certaines réactions s'opèrent de manière spontanée, sans catalyse enzymatique, et ne peuvent donc être inférées depuis le génome. Des méthodes automatiques visant à assister ce processus sont également utilisées, notamment les méthodes de *gap-filling*[23][158][106].

L'étape de validation consiste ensuite à réaliser des simulations à partir du modèle obtenu, afin de vérifier certaines conditions observées expérimentalement ou certaines propriétés théoriques. Ces validations sont généralement réalisées par

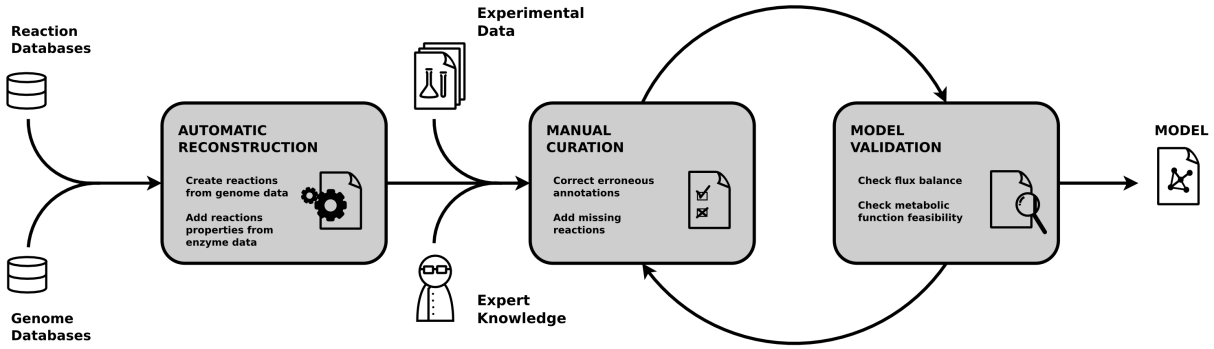


Figure 2.1 – Processus de reconstruction de réseaux métaboliques

simulation de flux grâce à des méthodes d'optimisation linéaire[110]. Vérifier la production de métabolites particuliers sachant la composition du milieu, ou encore l'adéquation entre le taux de production de biomasse théorique et le taux de croissance observé chez des organismes unicellulaires, font partie des options qui peuvent être employées pour la validation.

Les incohérences entre les prédictions du modèle et les observations expérimentales, identifiées lors de l'étape de validation, conduiront à un nouveau cycle de raffinement manuel, et ce processus va continuer de manière itérative. Les nouveaux résultats expérimentaux présentant des écarts au modèle vont également entretenir cette boucle simulation-correction. Ainsi, les réseaux métaboliques sont soumis à une constante évolution. À titre d'exemple, la base de connaissances EcoCyc[146] dédiée à l'organisme *Escherichia coli*, dont la première version date du 4 octobre 1995, continue de voir son contenu modifié (dernière mise à jour en date : 28 avril 2017, version no 21). Un résumé sur la dernière décennie montre que c'est le nombre de réactions qui présente le plus grand différentiel, avec une hausse de plus de 50% (Figure 2.2). Cette évolution dénote une certaine inconstance des réseaux métaboliques. Peu d'attention a été portée sur la validité des résultats inférés depuis des versions antérieures de ces réseaux ni sur la robustesse des méthodes qui leur sont appliquées, vis-à-vis des ajouts et suppressions de nœuds.

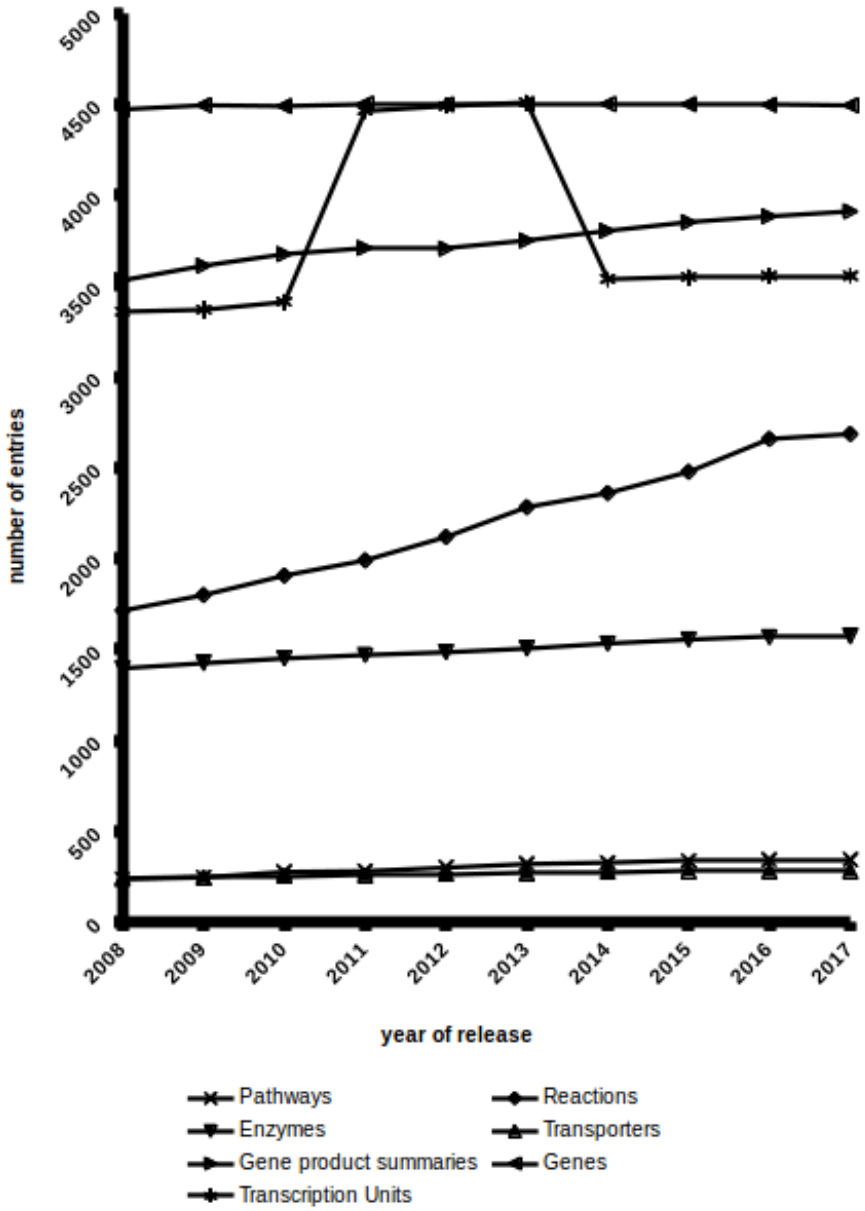


Figure 2.2 – Évolution du nombre d'entrées dans la base de données EcoCyc

2.2.2 Reconstruction ab initio

Une autre approche consiste à construire ces réseaux à partir d'observations du métabolome. La spectrométrie de masse à haute résolution permet une détection des métabolites avec une précision de l'ordre du ppm (partie par million, soit 1 mg/kg). Grâce à cette méthode, il est possible d'obtenir les compositions atomiques des molécules. La reconstruction *ab initio* consiste à inférer les liens qui

existent entre les molécules détectées par ce biais[44]. Les réactions biochimiques peuvent être considérées comme des créations de produits à partir d'événements d'ajouts et de suppressions d'atomes d'un substrat. À partir d'une énumération de ces événements et des différences de masse qu'ils induisent, il est possible de postuler des réactions putatives faisant intervenir des molécules dont la différence de masse correspond à celui de l'un de ces événements, et éventuellement de lever certaines ambiguïtés d'identifications[271] . Par exemple, une différence de masse de 18,01056 Da peut indiquer le gain ou la perte d'une molécule d'eau. Cette méthode souffre des limitations inhérentes aux méthodes de spectrométrie décrites précédemment. Notamment, l'absence de certains intermédiaires va conduire à l'omission de certaines voies, et l'absence d'information de localisation cellulaire peut conduire à des liens aberrants entre molécules qui ne sont jamais présentes simultanément dans un même compartiment. En revanche, contrairement à la reconstruction basée sur le génome, elle permet de proposer des interconversions non enzymatiques. Elle permet également de prendre en compte la promiscuité enzymatique, qui conduit à la transformation par une enzyme de substrats distincts, mais chimiquement proches, des substrats pour lesquelles elle est le plus spécifique. L'usage des réseaux issus de reconstruction purement *ab initio* est bien moins répandu que celui des réseaux issus du génome.

2.3 Utilisation des réseaux métaboliques

2.3.1 Caractérisation topologique du métabolisme

"Welcome to the real world, Neo"

Morpheus, The Matrix

Deux grandes approches se dégagent de ces réseaux. La première est l'approche qui sera ici nommée « macro », qui permet de postuler des caractéristiques fondamentales du métabolisme par rapport aux propriétés générales du réseau[136]. La seconde est l'approche nommée ici « micro », centrée sur les constituants de ces

réseaux, qui permet de définir le « rôle » d'un ou plusieurs métabolites par rapport à leurs liens avec le reste du réseau.

L'aspect macro a été popularisé dans les années 2000, où ce type d'analyses ont été conduites sur de nombreux réseaux dits « real world », terme utilisé pour distinguer les réseaux construits à partir d'observations, par opposition aux réseaux générés aléatoirement. Des travaux ont mesuré des caractéristiques topologiques sur des réseaux *real-world*, tels que le Web, les réseaux de co-citations dans des communautés scientifiques[19] ou des réseaux biologiques[136]. Des propriétés topologiques ont ensuite été proposées par comparaison de ces mesures avec les valeurs obtenues à partir de réseaux construits aléatoirement selon différents modèles. Ces analyses ont tenté de dégager des lois universelles qui régissent la formation des interconnexions dans la nature.

Parmi les plus notables figure la notion d'invariance d'échelle (*scale-freeness*)[19], qui stipule que certaines propriétés du réseau sont conservées lorsque l'on ne considère qu'un sous ensemble de ce réseau, ce qui se rapproche d'une conception « fractale » des réseaux biologiques.

La seconde est la nature supposée « petit monde » (small world) des réseaux[85]. Cette propriété traduit le fait qu'en moyenne, les distances qui séparent deux métabolites, exprimées en nombre de réactions enchaînées pour passer de l'un à l'autre, sont courtes.

Ces propriétés, déjà observées dans de nombreux réseaux « real world », ont conduit à l'hypothèse que le métabolisme était robuste aux suppressions aléatoires de nœuds, et sensible à la suppression de quelques nœuds fortement connectés[263].

Ces analyses ont été très critiquées par une partie de la communauté pour la faiblesse de certaines hypothèses, quand soumises à des tests statistiques, et pour l'inadéquation entre les prédictions issus de ces modèles et les observations[172]. Par exemple, la propriété de petit monde décrite dans les réseaux biologiques a été mise à mal par l'article de Arita sobrement intitulé « The world of *Escherichia Coli* is not so small »[13]. En effet, une inspection détaillée des liens représentés dans ces réseaux révèle la présence de nombreux composés ubiquitaires tels

que l'eau, à la fois produite et consommée par de nombreuses réactions, tendant ainsi à réduire de manière peu pertinente les distances entre métabolites. Ainsi les analyses « macro » ont conduit à ce que certains considèrent comme les premiers « mythes» du domaine de l'analyse de réseau en biologie[172]. Ces conceptions erronées ont pu perdurer notamment à cause de l'incertitude présente dans les réseaux, fréquemment évoquée pour expliquer l'écart entre les observations et les caractéristiques attendues. Ainsi les écarts au modèle ont été imputés à la qualité des données plutôt qu'au modèle. Il est également à noter que la confusion a pu être confortée par le fait que certaines de ces propriétés « macro » ont effectivement été corroborées expérimentalement dans d'autres réseaux. C'est le cas par exemple de l'étude de Jeong et collaborateurs qui a mis en évidence le fait que les protéines « *hubs* » dans les réseaux d'interaction protéines-protéines étaient codées par des gènes essentiels[135]. Cette observation est en adéquation avec la théorie d'invariance d'échelle, qui suppose une robustesse générale des réseaux face aux délétions aléatoires, mais une vulnérabilité face aux délétions ciblant les quelques *hubs*.

Au-delà du piège de la sacralisation des modèles au détriment des données expérimentales, ces événements ont souligné les limites de l'analyse des réseaux biologiques à un niveau purement abstrait, c'est-à-dire lorsque seules leurs topologies sont considérées. La prise en compte de critères biologiques dans l'analyse de ces réseaux a permis de remettre en cause certains de ces mythes, soulignant ainsi l'importance de la contextualisation des réseaux. Ainsi, les méthodologies qui s'appliquent à des réseaux tels que les réseaux d'interactions de protéines ne s'appliquent pas nécessairement de manière directe aux réseaux métaboliques. Des règles spécifiques des contextes qu'ils décrivent doivent s'appliquer afin de garantir la pertinence et la validité des résultats obtenus. Les méthodologies proposées dans cette thèse ont été construites sur la base de ce constat, et seront focalisées sur la prise en compte des spécificités des réseaux métaboliques dans leurs analyses, et plus particulièrement l'intégration de critères biologiques et chimiques dans les calculs.

Les analyses « macro » ont donc progressivement laissé la place à des analyses centrées autour des éléments constituants ces réseaux. Dans le contexte de la métabolomique, les réseaux sont essentiellement employés afin de fournir des hypothèses sur les chaînes causales pouvant expliquer les perturbations observées sur certains métabolites. Comme mentionné précédemment, certains de ces derniers ne sont présents que de manière très transitoire et sont consommés presque instantanément après leur production. Les techniques actuelles ne permettant pas de capturer la dynamique du métabolisme à cette échelle de temps, elles offrent une vue statique des processus impliqués. Il convient d'étendre cette vue en proposant des enchaînements d'événements pouvant relier les observations obtenues sur certains métabolites.

2.3.2 Approche par segmentation en voies métaboliques

> *"rationality is bounded when it falls short of omniscience. And the failures of omniscience are largely failures of knowing all the alternatives, uncertainty about relevant exogenous events, and inability to calculate consequences."*
>
> Herbert A. Simon, Rational decision making in business organizations, Nobel Memorial Lecture 1978.

Une approche commune pour représenter le métabolisme consiste à le diviser en voies métaboliques (ou pathways). Ces voies représentent une portion du réseau métabolique, habituellement centrée sur un ou quelques composés. Une définition communément admise de voie métabolique est la succession de réactions biochimiques, pouvant s'opérer dans une cellule vivante, et conduisant à la modification ou la production d'un composé principal (voire d'un groupe de composés). On peut citer par exemple la voie de biosynthèse du mannitol ou la voie de dégradation du benzoate. Ces voies peuvent être principalement catégorisées en voies anaboliques ou cataboliques suivant qu'elles conduisent à la synthèse ou à la dégradation, consommation ou assimilation du composé principal.

Bien que cette définition puisse paraître intuitive, elle ne couvre pas de nom-

breux cas particuliers. Ainsi, d'autres voies métaboliques viennent enrichir cette classification : les voies conduisant à la conjugaison d'un composé pour sa détoxification, les voies métaboliques dédiées au maintien de l'homéostasie, ou encore l'interconversion entre différents métabolites, la bioluminescence ou la production d'énergie. Pour ces derniers cas, il devient moins aisé, voire impossible, de définir un composé principal qui permettrait de borner une voie métabolique. On peut citer par exemple le cycle de Krebs ou la photosynthèse.

Des ensembles de voies métaboliques peuvent également participer communément à l'élaboration de mécanismes plus complexes, et dès lors être regroupées en voies métaboliques plus générales parfois appelées super-voies.

D'un point de vue topologique, les voies ne sont pas nécessairement linéaires, mais peuvent être branchées ou cycliques. Elles peuvent aussi contenir des structures bien plus complexes du fait des nombreuses redondances et *shunts* métaboliques, dont il est laissé aux curateurs le soin de statuer sur leur incorporation aux voies originellement définies ou la création de nouvelle voies alternatives, conduisant à des définitions variables selon les sources.

Il est également à noter que ces voies métaboliques sont fortement interconnectées par leurs métabolites d'entrées et de sorties et peuvent parfois partager des réactions et des métabolites intermédiaires et donc se chevaucher partiellement. C'est le cas par exemple de la glycolyse avec la *Rubisco shunt*, la voie d'Entner-Doudoroff et la voie des pentoses phosphates, ou encore le chevauchement entre le cycle de Krebs et le cycle du glyoxylate.

La notion de voie métabolique est motivée par la complexité du métabolisme, impliquant plusieurs milliers de composés et de réactions, qui constitue un frein à l'interprétation. Le partitionnement du métabolisme en voies métaboliques, centrées autour de composés ou de processus, offre l'avantage d'une convention de nommage qui permet de se situer dans ce réseau, et ainsi contextualiser les résultats de métabolomique.

Le désavantage principal est la difficulté pour la communauté de s'accorder sur les bornes de ces voies ainsi que sur leur centrage[49], conduisant à l'absence de

consensus. Par conséquent, les principales bases de connaissances qui compilent des voies métaboliques diffèrent grandement dans leurs contenus[249], y compris quant à des voies très étudiées comme le cycle de Krebs[250].

La base de données KEGG[141][142], préférant la notion de map et de module à celle de voie, définit ses entrées par l'ensemble des réactions en lien avec des concepts généraux tels que le métabolisme des acides aminés ou la synthèse des acides gras. La notion de disponibilité enzymatique est également omise puisque l'ensemble de ces réactions peuvent s'opérer dans différents organismes. [1] De ces choix résultent des maps de tailles importantes comparées aux voies contenues dans d'autres bases de connaissances.

C'est le cas de la base de données MetaCyc[48], qui propose des ressources spécifiques à chaque organisme, ainsi qu'un niveau de granularité plus fin dans sa définition des voies métaboliques. Ainsi, la map KEGG de la glycolyse et de la gluconéogenèse (considérée ici comme un seul et même processus) regroupe les pathways HumanCyc de la glycolyse, la gluconéogenèse, la décarboxylation du pyruvate en acétyl-CoA, la dégradation de l'éthanol en acetyl-CoA, la fermentation du pyruvate en lactate et la shunt glycolytique du 3—phospho glycerate (Figure 2.3). Un autre cas extrême est celui de la base de données UniPathway[193] qui référence de courtes séquences linéaires de réactions nommées « linear subpathway ».

1. En revanche, il est possible de mettre en évidence les réactions dont le gène associé est présent dans le génome d'un organisme sélectionné

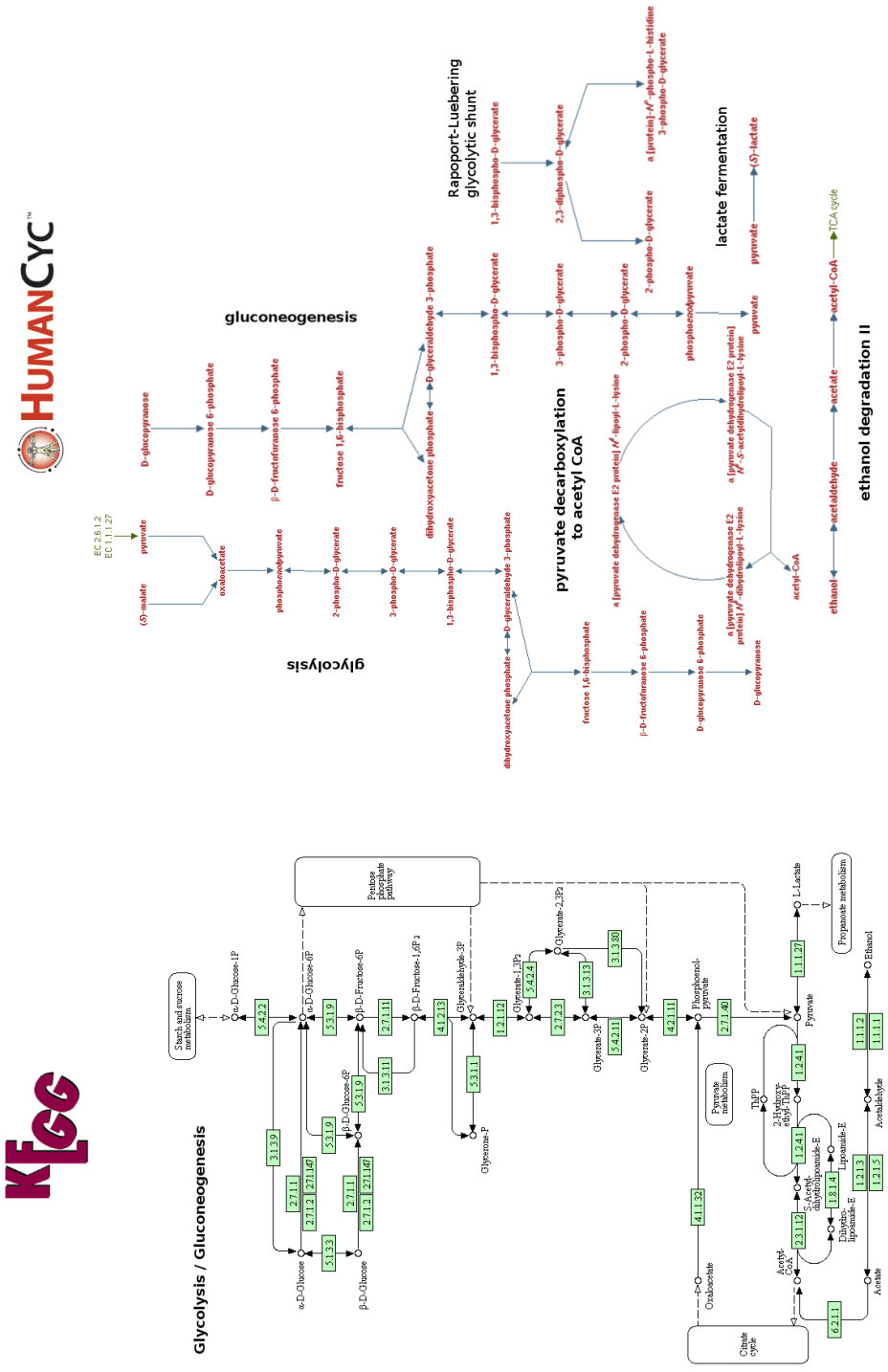


Figure 2.3 – Un même processus biologique représenté dans KEGG et dans HumanCyc

L'interprétation des résultats de métabolomique est fréquemment réalisée à l'aide de méthodes statistiques d'enrichissement[51][50], historiquement utilisées pour les données de transcriptomique[124][125]. Ces méthodes, telles que le test exact de Fisher par exemple, permettent d'identifier les voies métaboliques dont les composants métaboliques sont surreprésentés dans l'échantillon de métabolites obtenus en métabolomique. Les voies y sont représentées en tant qu'ensembles bornés et leur cardinalité est prise en compte dans le calcul. Ainsi, le choix d'une base de données de référence et de leurs conceptions sous-jacentes de la notion de voie impacte considérablement les résultats obtenus et les interprétations qui en découlent.

L'enrichissement est une pratique héritée de la transcriptomique. Les transcrits sont associés à des gènes dont les annotations sont utilisées pour définir des groupes fonctionnels. En revanche, la nature des gènes, qui peuvent être considérés comme des vecteurs d'informations (par exemple l'encodage de structures protéiques, associées à une fonction spécifique), ne peut être transposée aux métabolites. Ceci rend l'annotation fonctionnelle des métabolites et par extension la constitution des groupes pour l'enrichissement (c.-à-d. les voies métaboliques) bien moins consensuelle.

Des efforts ont néanmoins été fournis par la communauté pour définir des critères objectifs pour fragmenter les réseaux métaboliques en modules fonctionnels[237][235]. On peut citer par exemple les modes élémentaires qui se définissent par des sous-réseaux de tailles minimales pouvant fonctionner à un état d'équilibre[278]. Cependant, ils se démarquent des voies métaboliques traditionnelles dans leurs natures et leurs usages. Il n'offrent pas de convention de nommage, et, bien qu'il existe un ensemble fini et unique de ces modes élémentaires, leur nombre est en règle générale très élevé (dépassant les centaines de millions pour des reconstructions partielles) et ils présentent d'importants chevauchements. Leur énumération présente également un coût calculatoire important, limitant leurs applications, notamment vis-à-vis des réseaux à l'échelle du génome.

Au-delà du choix d'une définition appropriée des voies métaboliques, l'utilisation même des voies pour l'interprétation implique différents biais et limitations[180].

La vision segmentée du métabolisme nuit à la reconstruction de scénarios métaboliques impliquant plusieurs voies. Pour des raisons principalement liées à la lisibilité, certaines réactions impliquant des composés intermédiaires sont omises des voies. Ces omissions éclipsent les interconnexions entre différentes voies, d'ordinaire explicitées uniquement entre leurs composés initiaux et terminaux.

Il est également à noter que la représentation du métabolisme sous forme de voies est intrinsèquement liée à notre connaissance du métabolisme. Certaines parties, notamment dans le domaine du métabolisme secondaire, restent à ce jour bien moins connues[191]. Ceci conduit à l'existence de nombreuses réactions attribuées à aucune voie, ou à des voies génériques peu détaillées, dont l'implication biologique demeure floue. Par exemple, à ce jour sur les 177 réactions produisant de l'acetyl-CoA contenues dans MetaCyc, 33 ne sont attribuées à aucune voie métabolique.

L'utilisation des voies métaboliques engendre également des biais à l'interprétation d'ordre cognitifs. Herbert A. Simon, lauréat du prix Nobel d'économie, a proposé la notion de rationalité limitée pour modéliser les prises de décision en économie et en science politique[240][241]. Elle postule que la rationalité des individus dans leurs prises de décision est bornée par leurs limites de capacité cognitive face à des problèmes complexes, les orientant vers un choix raisonnable, réalisable en un temps acceptable, plutôt que vers des choix optimaux. Cette théorie peut également s'appliquer aux choix relatifs à l'interprétation des résultats de métabolomique. Ces choix seraient bornés par la complexité du métabolisme qui conduit les individus à fonder leurs hypothèses sur une à deux voies surreprésentées, quitte à omettre une partie des résultats. Des scénarios plus complexes qui permettraient de rationaliser l'ensemble des résultats peuvent alors être occultés.

La notion de rationalité limitée a conduit à la thématique des heuristiques de jugement, raccourcis cognitifs utilisés par les individus pour pallier leurs limites de

raisonnement[258]. L'une d'entre elles, l'heuristique de disponibilité, induirait des biais lors de la prise de décision notamment en fonction de la facilité de rappel des souvenirs. Or, les voies métaboliques ne disposent pas toutes de la même notoriété au sein de la communauté, certaines étant bien plus représentées dans les enseignements par exemple. Cela pourrait orienter l'interprétation d'une élévation du niveau de fumarate vers une dérégulation du bien connu cycle de Krebs, omettant les autres voies métaboliques l'impliquant (69 selon la définition de MetaCyc), telles que la dégradation du 5—nitroanthranilate ou la synthèse de la canavanine par exemple.

2.3.3 Vers une approche holistique : utilisation des réseaux globaux

"A system is more than the sum of its parts."

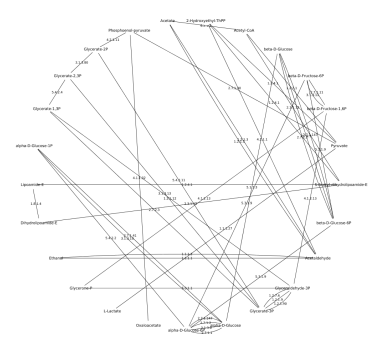
Walter F. Buckley, Sociology and modern systems theory

L'utilisation des modèles métaboliques à l'échelle du génome permet de considérer l'ensemble des connaissances liées au métabolisme. S'affranchir de la partition en voies métaboliques s'avère approprié pour l'étude des mécanismes impliquant plusieurs processus biologiques. En effet, l'interdépendance entre les métabolites au travers des réactions qui les consomment et les produisent, et qui conduit de facto à la représentation du métabolisme sous forme d'un ou plusieurs réseaux, implique la propagation des perturbations. Ainsi, la carence d'un métabolite peut conduire à la perturbation de l'abondance des métabolites produits à partir de ce dernier. Ces perturbations vont à leur tour affecter l'abondance des métabolites produits à partir des précédents et ainsi de suite, par effet domino de substrats en produits. La raréfaction d'un métabolite peut également impliquer l'accumulation d'un co-substrat faute de disponibilité de l'ensemble des substrats nécessaires à la réaction enzymatique le consommant, entraînant une propagation « longitudinale » de la perturbation, de co-substrat en co-substrat. Il est aisé de se figurer que les propagations de ces perturbations ne sont pas bornées à une voie

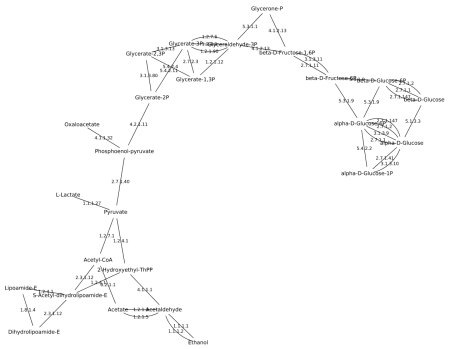
métabolique unique. Par conséquent, l'utilisation des réseaux à grande échelle s'avère être un outil puissant pour l'étude de la propagation des perturbations observées expérimentalement.

En revanche, leur exploitation est bien moins aisée. Elle se fait communément par l'intermédiaire de leurs représentations graphiques sous forme de réseau. Une des principales limites de leur utilisation est liée à leur taille (généralement plusieurs milliers de composés et de réactions) qui rend une exploration visuelle difficile[229]. Il est important de noter à ce stade la distinction qui existe entre la topologie d'un réseau et son dessin. L'analyse visuelle d'un réseau se fait au travers de sa représentation graphique (dessin), qui peut être créée de différentes manières à partir d'une même topologie (Figure 2.4), et dont le choix va conditionner sa facilité d'utilisation.

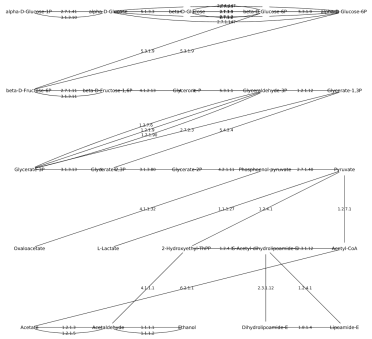
La question de la représentation des réseaux de grande taille est une problématique de longue date dans le domaine de la visualisation d'information[126][162]. De nombreuses manières de les représenter ont été proposées, tentant de faciliter l'accès visuel de l'information. Parmi les plus notoires on peut citer l'approche « Force-directed » de Fruchterman et Reingold[94], qui est basée sur la simulation d'un système physique avec répulsion/attraction des nœuds en fonction de la présence de lien entre eux. Bien que moins rependues, l'approche hiérarchique de Sugiyama[252] ou encore la représentation orthogonale de Fößmeier et Kaufmann[88] sont également utilisées.



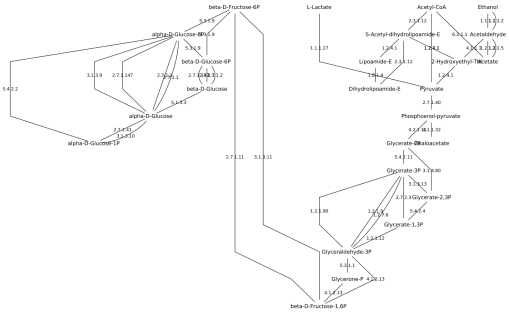
(a) Circular



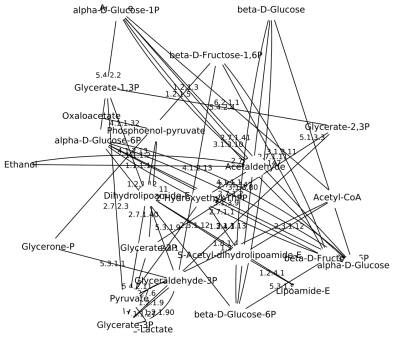
(b) Force Directed



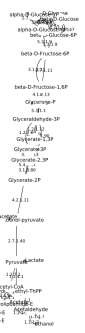
(c) Grid



(d) Hierarchical



(e) Random



(f) Spring Embedded

Figure 2.4 – Sous-réseau de la glycolyse représenté par différents dessins de graphe

Certains critères esthétiques, comme la proportion de chevauchements entre éléments ou le nombre de croisements entre liens, ont été identifiés comme cruciaux dans l'exploitation des réseaux[72][265].

Il a également été mis en évidence que le choix d'une représentation conduisait à des biais d'interprétations. Par exemple, le fait que la distance entre les positions des éléments dans le plan ne représente pas nécessairement leurs distances topologiques dans le réseau peut conduire à une estimation faussée de la relation qui unit deux éléments. Ainsi, la résolution de nombreuses tâches est influencée par les choix de représentations, tels que l'utilisation de flèches droites ou courbées pour symboliser les liens ou encore les angles entre les arcs sortants et entrants[274]. Parmi ces tâches, on peut citer la recherche manuelle d'un chemin de distance optimale dans un réseau[127] ou l'estimation du nombre de voisins en commun.

L'utilisation des réseaux métaboliques est donc limitée par la lisibilité de leur représentation graphique, et nécessite de réduire l'ensemble à considérer. Outre la difficulté de choisir un critère pertinent pour réaliser cette sélection, la lecture même d'un sous-réseau de taille accessible reste biaisée par des critères esthétiques.

Ces limites sont cependant cantonnées à l'exploitation visuelle des réseaux par un opérateur humain[195]. L'utilisation d'outils informatiques et de méthodes mathématiques permet d'assister l'exploitation des réseaux métaboliques et de pallier ces limites. La recherche de chemins optimaux, l'extraction de voisinage, le calcul de divers critères topologiques ou l'identification de groupes fortement connectés sont des tâches classiquement réalisées par des programmes informatiques. Une branche entière des mathématiques est dédiée à la résolution de ces classes de problèmes, dont les bases seront introduites dans le chapitre suivant.

Chapitre 3

La théorie des graphes, comment exploiter les réseaux métaboliques

3.1 La théorie des graphes

> *"In an extreme view, the world can be seen as only connections, nothing else. We think of a dictionary as the repository of meaning, but it defines words only in terms of other words."*
>
> Tim Berners-Lee, Weaving The Web : The Original Design and Ultimate Destiny of the World Wide Web

3.1.1 Introduction

La théorie des graphes est une sous branche des mathématiques discrètes, c'est-à-dire l'étude des ensembles dénombrables. Cette théorie est dédiée à l'étude des relations entre éléments d'un ensemble au travers des graphes, structure modélisant ces relations[140]. Les termes de réseau et de graphe sont souvent utilisés de manière interchangeable en fonction des domaines d'application. Dans cette thèse, le terme de réseau référera au concept défini dans le chapitre précédent, quand le terme de graphe référera à l'objet mathématique. Ce chapitre a pour objectif de définir les bases de cette théorie et ses principales applications aux réseaux métaboliques.

L'origine de la théorie des graphes est souvent attribuée au mathématicien Leonhard Euler et sa formulation du problème des sept ponts de Königsberg. Ce problème fait office de préambule à de nombreux manuscrits traitant de la théorie des graphes. Cette thèse ne dérogera pas à la règle, bien que cela trahisse un manque flagrant d'originalité. La ville de Königsberg de 1736, située en Prusse orientale, est traversée par le fleuve Pregel, dont les bras délimitent 2 îles. Ces îles sont reliées à la ville par 6 ponts, et 1 autre pont permet de relier les 2 îles entre elles (Figure 3.1).

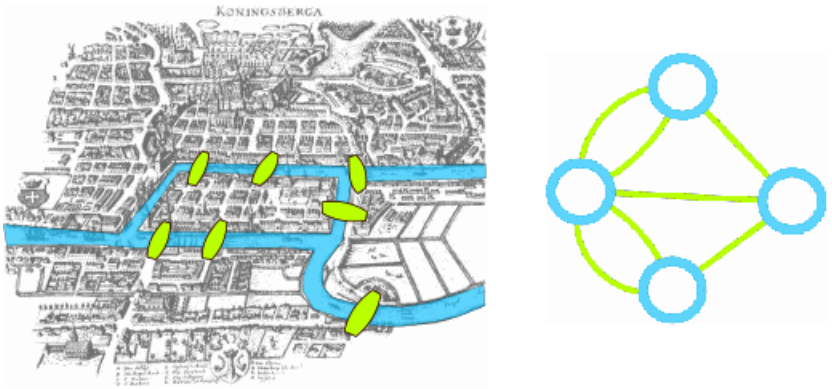


Figure 3.1 – Illustration du problème des ponts de Königsberg. Gauche : plan de la ville, Droite : Abstraction sous forme de graphe, avec les îles et rives représentées par des sommets, et les ponts par des arêtes.

Leonhard Euler pose le problème suivant : déterminer s'il existe un itinéraire dans Königsberg permettant, à partir d'un point de départ quelconque, de rejoindre un autre point en passant par tous les ponts sans traverser 2 fois le même pont. Leonhard Euler formula la solution en utilisant une abstraction du problème, et montra que le nombre de connexions autour d'un élément (îles ou rives) permet à lui seul de résoudre le problème. En effet, un tel itinéraire n'est possible que si le graphe ainsi construit ne possède que 2 ou 0 élément connecté à un nombre impair d'éléments. Dans le cas où le point d'arrivée serait le même que celui de départ, tout élément doit présenter un nombre de connexions pair. Dans le cas contraire, le point de départ et le point d'arrivée doivent présenter un nombre de connexions impair. Par cette preuve, Euler démontra que calculs et mesures concernant la

position des îles et des ponts n'étaient pas nécessaires à la résolution du problème, qui peut être généralisé à n'importe quelle ville sans autre connaissance que le nombre de ponts par rive. Leonhard Euler ouvrit ainsi la voie à la définition de la notion de topologie, étude des propriétés invariantes, formulée près d'un siècle plus tard.

Cette théorie sera appliquée par la suite et jusqu'à nos jours à de très nombreuses problématiques concrètes, des réseaux routiers aux réseaux sociaux, circuits électriques, réseaux biologiques[5][58][207] (interaction de protéines, régulations de gènes, réseaux métaboliques), jusqu'au Web[22], et même l'étude de la littérature anglaise[244].

3.1.2 Définitions et Notations

La sous-section suivante va définir les notions de base et notations usuelles utilisées en théorie des graphes, qui seront employées dans cette thèse.

Un **graphe**, que l'on notera $G(V, E)$, est défini par un ensemble de **nœuds** ou **sommets** (*vertex*) V et un ensemble d'**arêtes** (*edge*) E. Dans le cas du problème des ponts de Königsberg, les nœuds représentent les îles ou rives, et les arêtes représentent les ponts. Une arête $e = \{a, b\}$ définit une relation entre deux nœuds a et b. Ces nœuds sont dits **adjacents**, et e est dite **incidente** à a et b. Le nombre d'arêtes incidentes d'un nœud, utilisé dans la démonstration de Euler, constitue son **degré** (*degree*). L'ensemble des nœuds adjacents à a, contenant b, constitue le **voisinage** de a. Un nœud de degré égal à 0, donc sans voisinage, est dit **isolé**.

Si l'on reformule le problème des ponts de Königsberg dans une version plus contemporaine, où l'on recherche un itinéraire empruntable en voiture, sachant que certains ponts sont à sens unique, il devient alors nécessaire de représenter des connexions unilatérales entre les îles/rives. Les relations unilatérales dans un graphe peuvent être représentées en ajoutant un sens aux arêtes, nommées alors **arc**. Un graphe contenant des arcs est dit **orienté** (*directed graph*). Un arc $e = (a, b)$ est défini par une **origine** a et une **fin** b, définissant sa **direction**. On dit alors que a est un **prédécesseur** de b, et b est un **successeur** de a. On

distingue pour un nœud ses arcs **sortants** de ses arcs **entrants** en fonction de leurs directions : e est un arc sortant de a et un arc entrant de b. On parle alors également de **degré entrant** et de **degré sortant** en fonction du type d'arc incident considéré.

On peut également reformuler le problème précédent des 7 ponts de Königsberg en prenant en compte des propriétés des éléments de ce réseau, par exemple en considérant uniquement les îles ou rives possédant des parkings comme point de départ et d'arrivée valides. Il devient dès lors nécessaire d'associer différentes d'informations aux éléments du réseau. Les graphes possédant de telles propriétés sont dits **attribués**. Les réseaux « real world » possèdent la plus part du temps des **labels** (ou étiquettes) associés aux noeuds ou aux arêtes, permettant de référencer l'entité « réelle » représentée. Des attributs discrets peuvent également être associés à ces éléments, et l'on utilise fréquemment le terme de **couleur** pour désigner ces attributs. Des valeurs continues sont également fréquemment associées aux nœuds ou aux arêtes, et sont fréquemment mentionnées sous le terme de **poids**.

Un **chemin** (*path*) est une suite consécutive d'arcs permettant de relier 2 nœuds, une **source** et une **cible**, via des nœuds intermédiaires (on parle également de **chaîne** dans le cas des graphes non orientés). On parle de **chemin élémentaire** lorsque ce dernier ne passe pas deux fois par le même nœud, et de **chemin simple** lorsqu'il ne passe pas deux fois par la même arête (à noter qu'un chemin élémentaire est dès lors également simple). Le terme de chemin est couramment employé pour désigner implicitement un chemin élémentaire, et l'utilisation de ce terme dans cette thèse y référera également. Les chemins pouvant emprunter un même nœud plusieurs fois sont nommés **marches** (*walk*). Un chemin simple passant une seule fois par l'ensemble des arêtes d'un graphe est dit Eulérien, en référence au problème précédemment cité. Un chemin dont la source et la cible sont un seul et même nœud est nommé **cycle** (ou **circuit** dans le cas orienté, la mention cycle y est tout de même fréquemment employée).

Un graphe est dit **connexe** s'il existe un chemin entre toutes les paires de sommets $u, v \in V$ (dans le cas orienté, si cette condition est remplie en prenant

en compte l'orientation des arcs, alors le graphe est dit **fortement connexe**). Dans le cas contraire, il est dit déconnecté et par conséquent composé de plusieurs **composantes connexes**. De manière générale, un graphe dont les nœuds et arcs constituent des sous-ensembles des nœuds et arcs d'un autre graphe, soit $V' \in V$ et $E' \in E$, alors ce graphe $G'(V', E')$ est un **sous-graphe** de $G(V, E)$.

3.2 Les graphes métaboliques

La section suivante définit les différents moyens de représenter le réseau métabolique sous forme de graphe.

3.2.1 Graphe des composés

Le graphe des composés est un graphe orienté dans lequel les nœuds représentent les métabolites et un arc relie 2 métabolites s'il existe une réaction qui consomme l'un et produit l'autre (3.2 A)[68]. Il est possible de référencer explicitement cette réaction en attribuant des **labels** aux arcs. Étant donné que plusieurs réactions peuvent partager des couples substrats-produits, il peut exister dans ce graphe des arcs reliant les mêmes nœuds. Un tel graphe est nommé **multigraphe**, et ces arcs sont dits **parallèles**. L'ensemble des arcs reliant les mêmes nœuds constituent un **multi arc**. Il est à noter que ce graphe des composés, tout comme les autres graphes métaboliques, n'est pas nécessairement connexe.

3.2.2 Graphe des réactions

Le graphe des réactions est un graphe orienté dans lequel les nœuds représentent des réactions, et un arc relie 2 réactions si l'une produit un métabolite qui est consommé par l'autre (3.2 B)[68]. Ses propriétés sont similaires à celles du graphe des composés, les métabolites reliant les réactions peuvent être explicités via les labels des arcs, auquel cas il peut alors devenir nécessaire de construire un multigraphe si plusieurs produits d'une réaction sont consommés par une même réaction. Étant donnée la nature des résultats de métabolomique, cette représen-

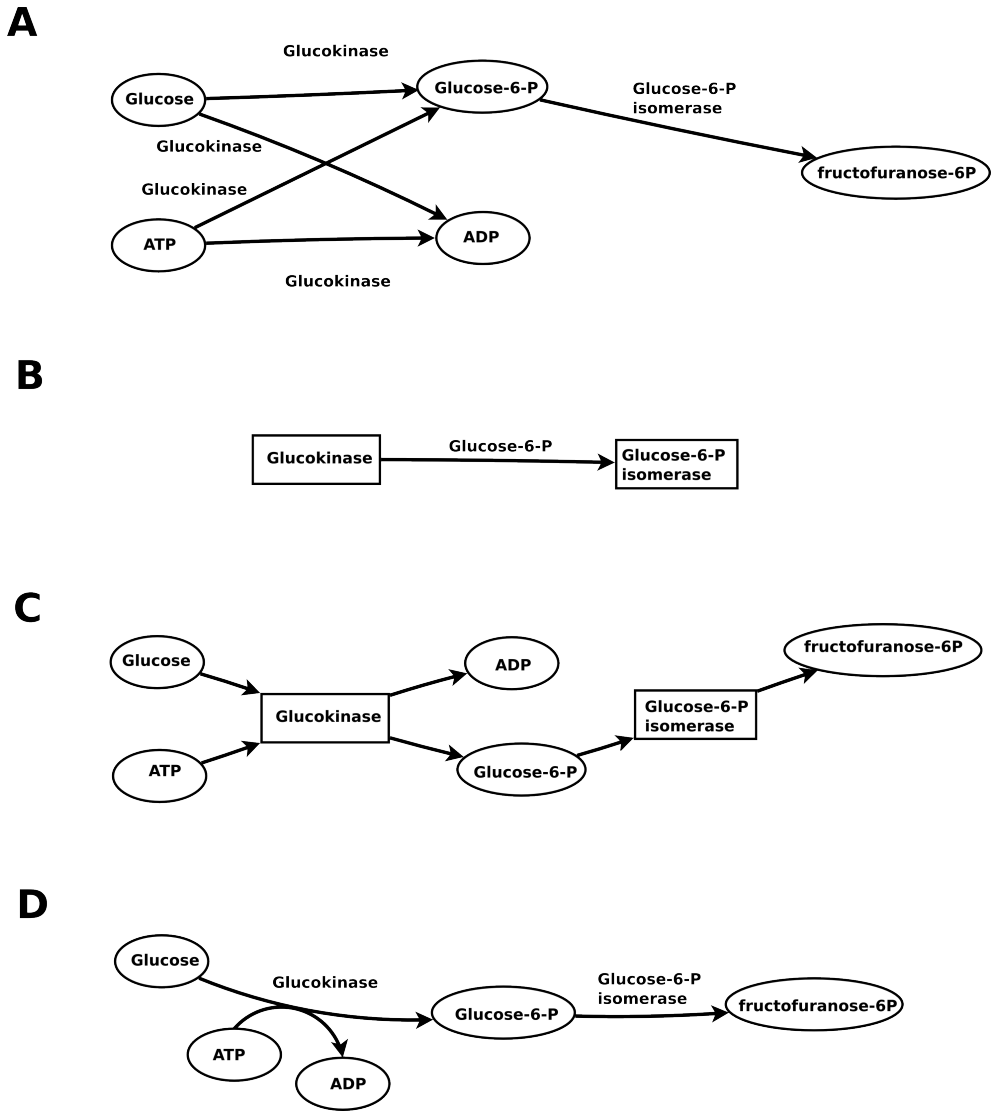


Figure 3.2 – Les différents graphes métaboliques. A : graphe des composés, B : graphe des réactions, C : graphe biparti, D : hypergraphe

tation où les métabolites figurent de manière implicite est peu utilisée pour leur interprétation. Les implications des choix méthodologiques décrits dans les parties suivantes sur l'utilisation du graphe des réactions ne seront donc pas discutées.

3.2.3 Graphe biparti

Un graphe **biparti** (**bipartite graph**) est un graphe dans lequel les nœuds peuvent être séparés en 2 groupes, tel que toutes les arêtes du graphe relient des

nœuds appartenant à des groupes distincts. Le graphe biparti permet de représenter à la fois les métabolites et les réactions sous forme de nœuds. Un nœud métabolite est prédécesseur d'un nœud réaction s'il est substrat de cette réaction, et il est successeur d'un nœud réaction s'il est produit par cette dernière (3.2 C)[68]. Ainsi il n'existe pas d'arcs reliant 2 métabolites ni d'arcs reliant 2 réactions, créant ainsi un graphe biparti. Outre cette propriété, ce qui le distingue du graphe des composés est qu'un graphe métabolique biparti n'est pas un multigraphe, et qu'un chemin simple dans le graphe biparti ne peut emprunter deux fois la même réaction. On remarque également qu'il est impossible de mapper un attribut relatif à un couple substrat-produits dans le graphe biparti, contrairement au graphe des composés. En effet, dans ce dernier des liens entre couples de métabolites sont représentés explicitement, ce qui permet de faire figurer des informations relatives à une transition substrat-produit (par exemple le nombre d'atomes échangés) en tant qu'attribut des arcs.

L'**ordre** du graphe, qui correspond au nombre de ses nœuds, est bien plus important pour le graphe bipartie que pour le graphe des composés, puisqu'il correspond à la somme du nombre de métabolites et de réactions dans le réseau, contre le nombre de métabolites seul dans le graphe des composés. En revanche, sa **taille**, qui correspond au nombre d'arcs, est bien plus réduite. Pour l'ajout d'une réaction irréversible donnée, le nombre d'arcs correspondant est la somme du nombre de substrats et de produits, alors que, dans le cas du graphe des composés, c'est le produit de ces deux nombres. Cette différence d'ordre et de taille va avoir des conséquences sur le nombre d'étapes nécessaires pour résoudre certains problèmes de manière algorithmique.

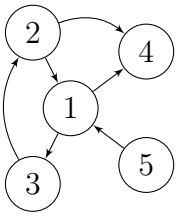
En théorie de la complexité, cette notion de nombre d'étapes nécessaires à la résolution d'un problème est appelée complexité en temps. Elle est définie pour un problème donné, en fonction de la taille de la donnée d'entrée. Résoudre une tâche donnée dans le graphe des composés et dans le graphe biparti ne constitue pas nécessairement des problèmes différents (par exemple, une recherche de plus court chemin), on va plutôt distinguer deux instances du même problème,

ayant chacune une taille différente. Pour les problèmes de théorie des graphes, la taille du problème est généralement exprimée en termes de nombre de nœuds (ordre du graphe, usuellement noté n). Le nombre d'arêtes y est parfois mentionné explicitement, et parfois exprimé en termes de nombre de nœuds lorsque l'on se place dans le pire cas. Le pire cas dépend du problème considéré, mais il correspond généralement au cas où toutes les arêtes possibles existent (le graphe est dès lors un **graphe complet**), ramenant le nombre d'arêtes au nombre de nœuds au carré. Il est à noter qu'en pratique les instances considérées, quel que soit le type de graphe choisi, sont bien loin de ce cas. L'ensemble des graphes métaboliques sont dits **creux** (*sparse*), ce qui signifie, par opposition aux graphes **denses**, que leurs nombres d'arêtes sont relativement faibles par rapport au nombre maximal d'arêtes possibles.

3.2.4 Représentation sous forme matricielle

Depuis le début de cette thèse, les graphes mentionnés sont représentés de manière graphique sous forme **sagittale**, diagrammes composés de flèches. Le graphe à proprement parlé est la structure de donnée sous-jacente à cette représentation graphique, qui correspond usuellement à des **listes d'adjacence**, c'est à dire une table où chaque index correspond à un sommet, et où l'élément correspondant contient la table des sommets adjacents. Il est toutefois également possible de les représenter sous forme matricielle. La représentation la plus utilisée est la **matrice d'adjacence**. Pour un graphe $G(V, E)$ avec un nombre de sommets $n = |V|$, tel que chaque sommet de V soit numéroté de 1 à n, la matrice d'adjacence est une matrice carrée $A = (a_{ij})$ de taille n x n, tel que $a_{ij} = 1$ si $(i, j) \in E$, et $a_{ij} = 0$ dans le cas contraire (exemple Figure 3.3). Ainsi dans le cas non orienté, cette matrice est symétrique. Les graphes métaboliques étant à de rares exceptions près orientés, ce ne sera donc pas le cas des matrices d'adjacences traitées dans cette thèse. Dans le cas des multigraphes, la matrice n'est plus binaire et la valeur de a_{ij} correspond au nombre d'arcs reliant i à j. Étant donné que les arcs reliant un composé à lui-même, nommés **boucles** (*loop*), sont généralement omis dans

les réseaux métaboliques, les éléments de la diagonale sont égaux à 0. Le caractère creux des graphes métaboliques signifie que la majorité des éléments de la matrice ont pour valeur 0. Cette caractéristique permet d'exploiter des encodages minimisant l'espace nécessaire à leur stockage.



(a) représentation sagittale

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 & 0 \\ 1 & 0 & 0 & 1 & 0 \\ 0 & 1 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \end{bmatrix}$$

(b) matrice d'adjacence

Figure 3.3 – Exemple de représentation matricielle d'un graphe

L'utilisation de la représentation matricielle permet d'appliquer des méthodes mathématiques issues de l'algèbre linéaire, en exploitant les propriétés de ces matrices. De nombreux travaux ont notamment été conduits sur les relations entre les propriétés des graphes et les valeurs et vecteurs propres de ces matrices, regroupés sous le terme de théorie spectrale des graphes.

Il est à noter que A^n, le produit matriciel de n copies de la matrice A du graphe G, permet d'obtenir le nombre de marches de longueur n entre deux sommets. En transformant la matrice d'adjacence en matrice stochastique de transitions (ou la somme de chaque ligne est égale à 1 ou 0) il devient possible de modéliser le comportement de processus markoviens tels que les marches aléatoires dans un réseau. Cette propriété sera exploitée au cours de cette thèse, et détaillée dans la partie III.

3.2.5 Hypergraphes métaboliques

Une **hyper-arête** est une arête reliant plus de 2 nœuds (également nommée hyper-arc dans le cas orienté). Un graphe contenant des hyper-arêtes est appelé **hypergraphe**. Dans les réseaux métaboliques sous forme d'hypergraphes, les nœuds représentent les métabolites, et les hyper-arcs représentent les réactions. Leurs

origines correspondent aux substrats de la réaction, et leurs terminaisons correspondent aux produits (3.2 D)[68][208][149].

La représentation sous forme d'hypergraphe est la représentation privilégiée dans de nombreux manuels de biochimie et dans les représentations graphiques des bases de données de voies métaboliques. En revanche, la nature plus complexe des relations représentées sous forme d'hyper-arêtes (pouvant impliquer plus de 2 nœuds) empêche la représentation du graphe sous forme de matrice d'adjacence « classique », réduisant ainsi l'utilisation des très nombreux développements qui reposent sur ce formalisme. De manière générale, l'usage de la majorité des algorithmes et des mesures issues de la théorie de graphes sont restreints aux graphes simples[208]. De plus, nombreuses sont celles dont la formulation dans l'hypergraphe revient à résoudre le problème dans la version simple ou biparti, privilégiant de ce fait ces représentations par la communauté. Néanmoins, certaines problématiques ont été transposées spécifiquement à ces graphes[97][33], dont certaines appliquées aux réseaux métaboliques, offrant la prise en compte de critères biologiques qui ne sont pas transposables dans les 2 représentations précédentes. Ces critères seront détaillés dans la partie suivante.

Il est à noter que si ce type de graphe offre d'ordinaire la possibilité d'identifier des associations substrats-produits particulières au sein d'une même réaction, c'est du fait de choix de représentation graphique, et cette distinction ne peut être considérée dans les calculs. Par exemple, figure 3.2.D, deux associations : Glucose $\rightarrow$ Glucose-6-P au centre et ADP $\rightarrow$ ATP sur le côté.

3.2.6 Bilan : Différents formalismes, différents traitements

Différents formalismes sont donc disponibles pour représenter le réseau métabolique sous forme de graphe. Tous ces graphes métaboliques ont des propriétés communes : ils sont orientés, attribués, creux, et ne possèdent pas de boucles. Ils présentent également des propriétés qui leur sont spécifiques : biparti, multigraphe ou hypergraphe, avec des ordres et des tailles différents, qui vont conditionner leurs utilisations en fonction des limites des méthodes appliquées. La deuxième partie

de cette thèse proposera ainsi des éléments pour motiver la sélection de l'une de ces représentations qui soit appropriée à notre problématique.

3.3 Objectif : Déchiffrer les relations indirectes dans les réseaux

La théorie des graphes offre de nombreux outils pour l'étude des réseaux biologiques[5] dont les réseaux métaboliques. Elle a notamment été employée pour l'identification de motifs répétés au sein d'un réseau ou conservés entre des réseaux issus de différents organismes[222][185][161]. L'autre aspect, qui sera détaillé dans cette thèse, est l'utilisation de la théorie des graphes pour élucider les relations entre les métabolites, notamment au travers de la notion de proximité. Considérer uniquement les interactions directes entre les éléments du réseau revient à sous-estimer la propagation de certains phénomènes dans ce réseau, tels qu'un changement de concentration dans un réseau métabolique. Il convient dès lors d'étendre la notion de connectivité à celle de **proximité**, qui peut s'apprécier au travers de connexions indirectes. Cette proximité permet d'extraire des sous-réseaux de tailles interprétables dont les membres partagent des liens forts (une proximité) avec un ou plusieurs métabolites d'intérêt, tels que des métabolites discriminants deux conditions, obtenus par analyse métabolomique. Elle permet également d'identifier des regroupements de métabolites « proches », nommés **communauté**, qui peuvent s'apparenter à la notion de cluster appliquée aux réseaux. La proximité, quand elle est généralisée à l'ensemble du réseau, ouvre sur la notion d'importance d'un sommet. Un sommet « important » peut alors être caractérisé par une proximité générale à tout autre sommet du réseau, ou à un ensemble de sommets d'intérêt. De nombreuses interprétations biologiques peuvent être postulées à partir de ces notions de groupes, d'importance et plus généralement de proximité dans les réseaux métaboliques.

Les sections suivantes ont pour but de formaliser ces notions au travers de mesures issues de la théorie des graphes. Certaines de ces mesures seront détaillées

dans les parties suivantes lors de leur application aux réseaux métaboliques.

3.3.1 Distances et problème du plus court chemin

La proximité repose sur l'interprétation d'une distance entre des éléments. La représentation des réseaux sous forme de graphes offre une définition intuitive de la distance au travers de la longueur du chemin le plus court entre deux sommets[69]. Ce problème est très étudié en théorie des graphes, avec de nombreuses applications telles que la recherche d'itinéraire routier ou la planification. Pour son application aux graphes « *real-world* » il est souvent nécessaire de prendre en compte différents paramètres afin d'obtenir une distance réaliste. Il est par exemple utile dans le cas de la recherche de chemins dans les réseaux routiers de prendre en compte la longueur des routes ou une estimation du temps nécessaire pour les traverser en voiture. Ces paramètres peuvent être formulés sous forme de poids dans un graphe attribué. Ce poids peut être vu comme un « coût de passage » pour l'emprunt d'un arc ou d'un sommet, la problématique de recherche de chemin avec un nombre « d'étapes » minimal devient alors une problématique de recherche de chemin de poids cumulé minimal. On parle parfois du **chemin le plus léger** (*lightest path*).

3.3.2 Notions de centralité et de métriques d'influence

La **centralité** est un indicateur permettant de classer les éléments d'un graphe (usuellement les sommets), en fonction de leur importance dans le réseau. L'importance est un concept relatif qui se définit par rapport à un rôle donné, ainsi la centralité peut être définie de nombreuses manières. La centralité peut par exemple estimer l'influence d'un sommet sur le reste du réseau, en mettant en exergue ceux possédant un plus grand nombre de connexions (**centralité de degrés**) ou ceux dont la distance moyenne avec tout autre sommet est la plus faible (**centralité de proximité**, ou *closeness*). Ce type de centralité a été notamment appliqué aux réseaux sociaux pour identifier des personnes d'influence au sein d'un groupe, ou en épidémiologie pour identifier des vecteurs de transmission. Les

indices de centralité permettent de classer les nœuds d'un réseau pour identifier les éléments « exceptionnels », en revanche elles ne permettent pas de comparer les nœuds deux à deux ni de capturer les rôles des nœuds qui ne sont pas particulièrement importants[165]. Ce constat a conduit à la création de **métriques d'influences**. L'une des plus connues, intrinsèquement liée à la notion de proximité, est celle **d'accessibilité** qui mesure le nombre de nœuds atteignables depuis le nœud considéré, étant donnée une distance maximale (exprimée en longueur de chemin). De manière générale, la centralité peut être vue comme une extension du concept de proximité entre deux nœuds à celui de proximité relative d'un nœud avec un ensemble d'autres nœuds, voir avec la totalité des nœuds du réseau.

3.3.3 Notion de *Network flow*

Les relations indirectes peuvent donc être représentées sous forme de chemins (au sens large de successions de nœuds adjacents). Il a été postulé que les qualités représentatives des concepts précédemment cités, en particulier la centralité, reposent sur une définition adéquate des échanges qui conduisent à ces relations indirectes. En d'autres termes, l'importance d'un acteur dans un réseau dépend de la manière dont l'information circule dans ce réseau. Il devient donc essentiel de définir des contraintes sur les chemins qui vont servir à définir la centralité, et plus généralement la proximité. Dans ce contexte, Borgatti conceptualise la notion de *flow*[37], au travers d'une série d'exemples : le premier est la livraison d'un colis, qui va transiter d'une personne à une autre jusqu'à atteindre son destinataire. Le parcours du colis en question est soumis à un objectif précis, la réception par le destinataire, et a vocation à remplir cette tâche de manière optimale. Ainsi le colis n'a pas de raison de transiter plusieurs fois par la même personne, et aura tendance à circuler via un nombre d'acteurs le plus petit possible. Il est dans ce cas raisonnable de considérer des chemins élémentaires, et de surcroît **géodésiques**, c'est-à-dire les plus courts chemins, pour définir la centralité des acteurs du réseau.

Un des autres exemples fourni par Borgatti est celui de la diffusion d'une rumeur au sein d'un réseau social. Contrairement au colis, la rumeur n'a pas d'exis-

tence physique propre et peut être présente en plusieurs endroits du réseau à la fois, car elle n'est pas « perdue » par l'émetteur, lorsque partagée avec son voisin. Elle circule ainsi dans le réseau par **duplication** plutôt que par **transmission**. Sa diffusion n'est pas motivée par un destinataire à atteindre, et un acteur peut recevoir plusieurs fois la rumeur, ce qui peut rendre l'utilisation des plus courts chemins moins appropriée dans ce contexte. L'auteur note également que, hors cas de perte de mémoire, un acteur ne diffuse pas la rumeur à une personne qu'il a lui-même informée précédemment, et inversement. Ainsi, la diffusion de la rumeur ne va pas emprunter la même arête plus d'une fois, et sera par conséquent mieux modélisée par des chemins simples (*trials*). Borgatti définit ainsi différentes topologies circulatoires[37] : les processus de transmission via des chemins géodésiques, simples, élémentaires ou non (marches) et les processus de duplication, parallèles ou en série selon que la diffusion aux voisins se fait de manière simultanée ou non, et via des chemins simples, élémentaires ou des marches.

3.4 Conclusion et objectifs de la thèse

Cette partie a permis d'introduire la nature des données utilisées dans cette thèse, les réseaux métaboliques et les données de métabolomique, ainsi que les concepts de bases de la théorie des graphes, dont certaines méthodes présentent un fort potentiel pour l'interprétations de ces données. Il a été proposé que ces méthodes soient intimement liées à la notion de proximité dans les réseaux métaboliques, et que les erreurs d'interprétations qui ont marqué les débuts de leur analyse topologique ont montré l'importance d'intégrer des informations domaine-spécifiques afin de proposer des distances réalistes prenant en compte la nature des relations entre éléments d'un réseau. Cette première partie met en avant deux axes principaux pour garantir cette pertinence : une pondération porteuse de sens et une topologie adéquate des chemins considérés. La partie suivante est dédiée aux travaux de cette thèse portant sur ces deux axes dans le contexte des réseaux métaboliques. Elle s'appuie notamment sur des caractéristiques fondamentales de ces réseaux et leurs implications sur la pertinence des résultats obtenus via

leurs usages. Ces travaux permettent l'émergence de recommandations génériques sur l'utilisation des réseaux métaboliques, au travers de la notion de proximité, sous-jacente à de nombreuses méthodes. En revanche, comme mentionnée précédemment, une application pertinente des méthodes de la théorie des graphes aux réseaux métaboliques va être dépendante de la question posée. La partie III propose une nouvelle méthode d'interprétation de résultats de métabolomique. Elle est basée sur la notion de centralité et sur les ajustements proposés dans cette thèse pour prendre en compte au mieux la nature des données présentées, en particulier les limites de l'observation et de la modélisation du métabolisme.

Deuxième partie

Garantir la pertinence des applications de la théorie des graphes aux réseaux métaboliques

Chapitre 4

Problème des composés auxiliaires

La partie suivante a pour but de proposer des ajustements méthodologiques qui visent à garantir la pertinence biologique des approches de théorie des graphes dans leurs applications aux réseaux métaboliques. Elle sera focalisée sur les applications portant sur l'analyse des relations au sein de ces réseaux, telles que définies précédemment au travers de la notion de proximité métabolique, et tout particulièrement dans le cadre d'analyses de résultats de métabolomique. Ces ajustements peuvent prendre trois formes : pondération des relations, contraintes topologiques sur les chemins considérés, mais également la modification des réseaux d'entrée.

4.1 Introduction

Le problème des composés auxiliaires (*side compounds*) constitue le principal obstacle à la définition de distance pertinente dans les réseaux métaboliques. C'est en effet celui qui a reçu le plus d'attention de la part de la communauté[261]. Il est notamment la cause des erreurs d'interprétation qui ont accompagnées les théories des propriétés small-world et scale-free des réseaux métaboliques[172], présentées en première partie. Ce problème fondamental concerne le rôle des molécules auxiliaires telles que l'ADP, l'eau ou le dioxyde de carbone dans les réseaux métabo-

liques, et par extension la nature même des liens dans les réseaux métaboliques. Si le chemin passant directement du glucose au pyruvate par l'intermédiaire de l'ADP n'est pas pertinent du point de vue des biochimistes, qu'est-ce qui distingue l'ADP des autres nœuds ? L'ADP est un composé ubiquitaire dans la cellule, impliqué dans un très grand nombre de réactions. On peut supposer qu'une perturbation au niveau de l'une d'entre elles affecte de manière négligeable la disponibilité de l'ADP pour les autres réactions. Ainsi, l'ADP serait caractérisée par une plus grande résilience face aux perturbations, et ne saurait constituer un vecteur de leur propagation. D'autres molécules sont fréquemment considérées comme disposant des mêmes propriétés, et présentent une caractéristique commune : un fort degré par rapport aux autres nœuds du réseau[63][64]. Ces nœuds sont communément nommés « *hubs* ». Une des caractéristiques tout à fait particulières des réseaux métaboliques repose sur la manière dont sont considérés ces *hubs*. Ils ont tendance à être omis (ou leurs arcs adjacents) des représentations graphiques, et écartés des interprétations mécanistiques, alors que dans de nombreux réseaux real-world, les *hubs* tendent à être considérés comme des acteurs de premier plan. Ils sont notamment perçus comme des facteurs de la cohésion de l'ensemble du réseau, ce qui n'est pas communément admis dans le cas des *hubs* de réseaux métaboliques. Une autre question fondamentale se rapportant au problème précédent pourrait être formulée de la manière suivante : qu'est-ce qui distingue le lien qui unit l'ADP et le glucose du lien qui existe entre le glucose et le glucose-1-phosphate ? Cette autre approche va généralement conduire à considérer les structures chimiques des composés, en considérant les liens du réseau métabolique comme des échanges de groupements chimiques, voir, à une autre échelle, d'atomes.

L'article suivant présente les différentes méthodes derrière ces deux approches, l'utilisation de critères chimiques ou topologiques, et souligne leurs limites face à certains cas concrets. Il y est également proposé des profils de cas où elles vont respectivement être plus adaptées que l'autre. Cet article fera également office d'introduction à d'autres problématiques inhérentes aux réseaux métaboliques, qui seront détaillées dans les sections suivantes.

4.2 Méthodes de recherche de chemins métaboliques

Briefings in Bioinformatics Advance Access published January 27, 2016

OXFORD

Briefings in Bioinformatics, 2016, 1–14

doi: 10.1093/bib/bbv115
Paper

Computational methods to identify metabolic sub-networks based on metabolomic profiles

Clément Frainay and Fabien Jourdan

Corresponding author. Fabien Jourdan, INRA, Toulouse University, INP, UMR 1331, Toxalim, Research Centre in Food Toxicology, 180 chemin de Tournefeuille, F-31027 Toulouse, France. Tel.: +33 0582-066395; Fax: +33 0561-285244; E-mail: Fabien.Jourdan@toulouse.inra.fr

Abstract

Untargeted metabolomics makes it possible to identify compounds that undergo significant changes in concentration in different experimental conditions. The resulting metabolomic profile characterizes the perturbation concerned, but does not explain the underlying biochemical mechanisms. Bioinformatics methods make it possible to interpret results in light of the whole metabolism. This knowledge is modelled into a network, which can be mined using algorithms that originate in graph theory. These algorithms can extract sub-networks related to the compounds identified. Several attempts have been made to adapt them to obtain more biologically meaningful results. However, there is still no consensus on this kind of analysis of metabolic networks. This review presents the main graph approaches used to interpret metabolomic data using metabolic networks. Their advantages and drawbacks are discussed, and the impacts of their parameters are emphasized. We also provide some guidelines for relevant sub-network extraction and also suggest a range of applications for most methods.

Key words: metabolomics; metabolic network; graph algorithm; path search; sub-network extraction

Metabolic graphs to put metabolomic profiles into context

Monitoring and understanding the metabolic status of an organism under various environmental or genetic conditions is a key challenge in human health and bioengineering [1]. Untargeted metabolomics, combined with multivariate statistics, was developed to achieve this goal by detecting metabolites with significant changes in concentration [2–4]. Once the metabolites of interest are confidently identified (preferentially level 1 in the classification proposed in [5, 6]), the resulting list is a metabolic profile that characterizes the physiological response of the organism to the perturbation being studied. Nevertheless, contrary to gene- and protein-based studies, no single metabolomics technology enables all metabolites to be monitored. This partial view of the metabolome complicates the identification of the metabolic functions that change metabolite concentrations. Moreover, as global metabolomics is applied without preconception, the downstream biological interpretation of the results requires taking into account the whole metabolic knowledge (thousands of metabolic reactions) available for a given organism. This data-mining challenge is of utmost importance in many fields of application such as human health, where understanding the biology behind each disease will be a cornerstone in future personalized medicine [7].

The most commonly used ways to biologically interpret metabolomic profiles consist in searching the metabolic pathways available in public databases, such as KEGG [8, 9] or BioCyc [10], to identify those involving the metabolites of interest. With complementary statistical tests (e.g. enrichment analysis [11, 12]), this approach provides clues about which pathways are more likely to be involved in the response to the perturbation being studied [13–15]. Nevertheless, a single metabolite can be mapped on numerous pathways, and the analysis fails to explain how the pathways are connected, especially if the metabolic perturbation spans several pathways. The other main drawback is the subjective definition of pathways, which differs from one database to another [16], even though some attempts to formally define the term have been made [17].

To avoid this fragmented interpretation, it is useful to consider the whole metabolism as an integrated system before focusing on the parts connecting compounds of interest. To do

Clément Frainay is a bioinformatics PhD candidate with a background in biochemistry, working on developing new graph algorithms to mine metabolomic profiles in the context of genome-scale metabolic networks.
Fabien Jourdan has a PhD in computer science. Currently, he is a bioinformatics researcher at INRA (Toulouse, France) working on potential impacts of food contaminants on human metabolism using metabolomics and network modelling.
Submitted: 6 October 2015; **Received (in revised form):** 16 December 2015

so, all biochemical reactions that have been experimentally identified in a given organism or predicted *in silico* from similar ones can be gathered and interconnected in a single metabolic network [18]. In addition to avoiding division into pathways, this formalism makes it possible to account for reactions that are not attributed to any pathway [19, 20]. This integrative view of metabolism nevertheless implies dealing with genome-scale networks containing thousands of densely connected reactions (7440 in the latest reconstruction of human metabolic network [21]). To make sense of these large networks, they need to be turned into mathematical objects called graphs (nodes connected by edges), after which dedicated algorithms can be used to identify less dense and more informative sub-networks [22].

The main strength of the graph-based methods reviewed in this article is their ability to take into account only network structure, chemical information and an input list of metabolites. Note that these computational solutions are not designed to predict the organism's steady-state metabolic routes (elementary modes [23–25]), metabolic fluxes (which is the scope of constraint-based modelling [26, 27]) or changes in metabolite concentration over time (which is the scope of ordinary differential equation modelling [28]). Indeed, these three more predictive approaches require additional parameters, such as system boundaries or flux constraints, which are not necessarily available for untargeted metabolomics studies.

This article reviews graph-based methods used to extract sub-networks from genome-scale networks, based on the combination of linear paths between metabolites of interest. The application of these methods has been the subject of increasing interest in the field of bioengineering, as it can help in the search for production pathways by proposing reaction steps that can link two compounds, the terminal one being the compound of economic interest [29]. In this article, we focus on their great potential for the biological interpretation of metabolic profiles by extracting the most relevant part of the whole network that connects metabolites in the profile. For example, this approach helped decipher the impact of a toxicological metal (cadmium) on yeast metabolism [30]. We applied several sub-network extraction methods to two well-known biosynthetic processes to highlight the advantages and drawbacks of these approaches and to provide some guidelines on their usage.

Graph modelling of metabolism

To look for connections between identified metabolites, modelling requires turning the metabolic network into a graph. A graph is composed of vertices (also called nodes) linked by edges. The graph is called an oriented graph if the direction of the edges is defined.

Metabolic graphs can be used as a purely descriptive object by representing knowledge in a visual and intuitive way. More importantly, graphs are mathematical objects on which algorithms can be used to find solutions to complex problems and to identify new hypotheses. This active research field in mathematics, known as graph theory, has been successfully used in different fields of application such as road traffic, telecommunications, social networks and cartography [31–34]. It is also increasingly used in biology [35–37], in particular for metabolic networks [38, 39].

Metabolic networks are often drawn as hypergraphs [40–42] in which more than two nodes can be linked by one edge (hyperedge), as shown in Figure 1A. In this case, metabolites are represented by nodes and reactions by hyperedges connecting their sources (reactants) to their targets (products). Hypergraphs are widely used for representation purposes but far less for computation, as many graph algorithms cannot be directly applied to hypergraphs [42]. Metabolic networks are therefore mostly described using simple graphs whose edges only connect two nodes. In bipartite graphs, nodes represent metabolites or reactions, and the edges only connect nodes from two different classes, linking a metabolite to a reaction if it is a substrate or a product of the reaction (see Figure 1B). In a compound graph, each node refers to a metabolite, and an edge links two nodes if they are involved in a reaction as substrate and product (see Figure 1C). A reaction is then represented by several edges, which is less intuitive, but makes the graph easier to process because nodes represent the same kind of entities. Further discussions on ways to model metabolic networks using graphs can be found in [38, 39, 43–45]. In this review, we only consider compound graphs, although most of the methods presented can also be directly applied to bipartite graphs.

Connecting compounds of interest in graphs

The compounds in the metabolic profile can be mapped onto the corresponding nodes in the graph. The challenge is then to find reaction cascades (paths) connecting them because these reactions and intermediary metabolites may be involved in the metabolic response to the perturbation under investigation [30]. In graph theory, a path is defined as a sequence of distinct edges connecting a source node to a target one through distinct intermediate nodes.

Global network connectivity can lead to a huge number of possible paths, thereby complicating the elucidation of links between identified compounds. For instance, Kuffner *et al.* highlighted the existence of more than 500 000 reaction paths (at most nine steps in length) between glucose and pyruvate [46]. Obviously, most of these paths were artefacts and did not make any biochemical sense. For example, Figure 2 shows a path from glucose to pyruvate going through ADP, whereas it should go through glucose-6-P and along the whole glycolysis pathway.

The aim of this article is to provide an overview of methods that can reduce this set of paths to a more meaningful one. The first part of the review focuses on solutions proposed to adapt graph theory algorithms to identify biologically meaningful paths. Because metabolomic profiles generally contain more than two metabolites, the second part of the review focuses on integrative sub-network extraction methods capable of dealing with lists of metabolites of interest. Finally, we show how these methods behave in simple applications.

How to obtain a biologically relevant path?

Finding parsimonious paths

Facing the huge number of possible paths between metabolites, it is possible to focus on the most parsimonious ones by keeping only those with the smallest number of steps, that is, the shortest paths. Shortest path search is a well-documented problem in graph theory [47]. The overall scheme of path search is nevertheless shared by most methods: nodes are recursively added to the path only if going through them is the best option in terms of path length (see Supplementary Material Section S1 for an overview of the algorithm). Their use to link two compounds in metabolic networks was introduced by Arita [48], but, as will be shown in the following, those algorithms cannot be applied directly and require adaptations to obtain biologically relevant results.

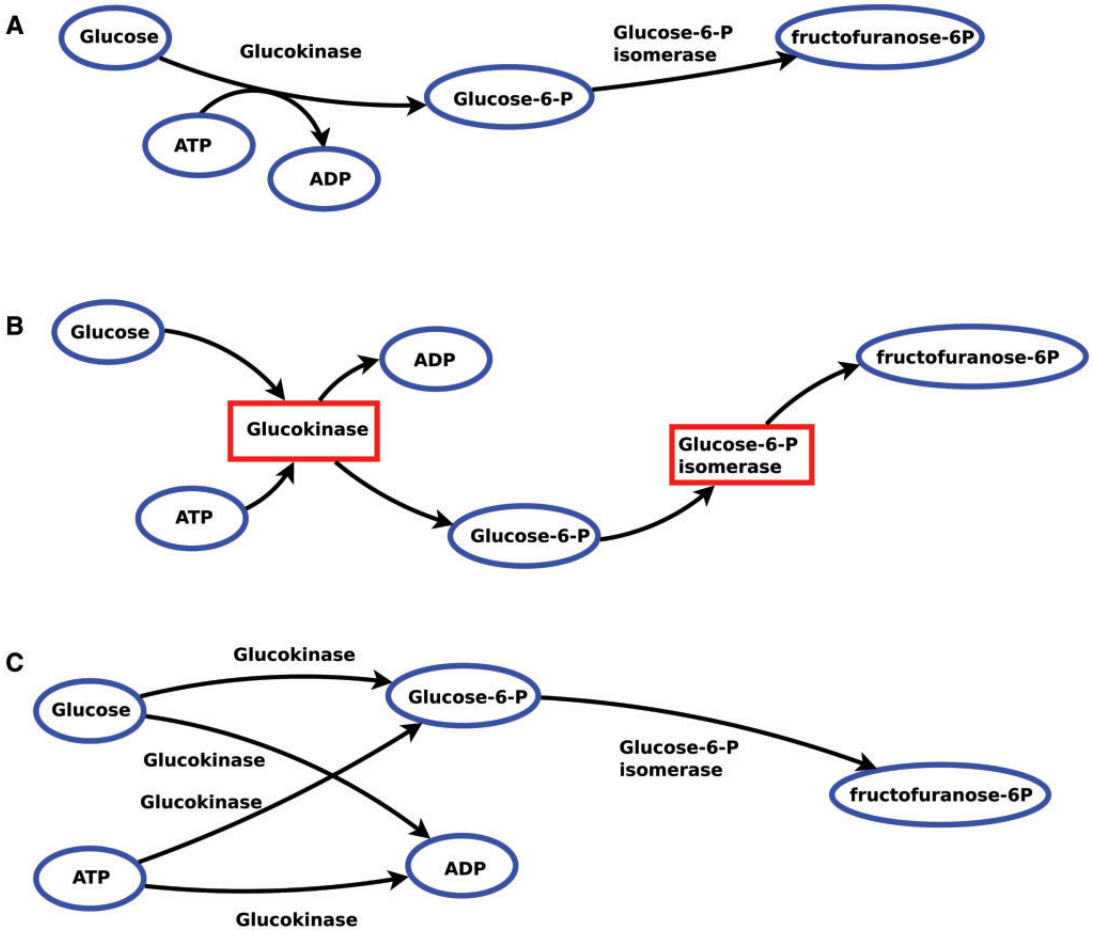


Figure 1. Different types of metabolic graph representations for the simple network made of glucokinase and glucose-6-P-isomerase. (**A**) Hypergraph. (**B**) Bipartite graph. (**C**) Compound graph.

Avoiding including side compounds in paths

The main issue when applying shortest path algorithm to metabolic graphs is the presence of side compounds, which often leads to irrelevant paths (e.g. ADP in Figure 2) [45, 49]. A side compound is a molecule that is used for complementary purposes such as energy carrier, proton donor or acceptor, usually taken from a « pool » in the cell (side compounds are also called pool-metabolites).

Defining side compounds using network topology

Several molecules are usually considered to be side compounds, including ATP (adenosine tri-phosphate), H_2O, CO_2 and NADP (nicotinamide adenine dinucleotide phosphate). Hence, one way to solve the problem is to systematically remove them from the graph [50]. Nevertheless, as the definition of an a priori list can be quite subjective and uncertain, network topology-based definition of side compounds was introduced [50, 51]. The underlying assumption is that side compounds are usually involved in many reactions, meaning that they are connected to a large number of nodes (these highly connected nodes are often called 'hubs' [52]). This number of connections, called degree, can be used as a parameter to decide whether a node represents a side compound to filter them out [53]. However, this requires defining a threshold above which nodes will be removed. Such a definition is also dubious because some highly connected compounds, such as methionine or pyruvate, are mostly considered as « main » compounds. Moreover, some metabolites considered as side compounds in some pathways may be involved in mechanisms as main compounds in other pathways, for instance, ATP in nucleotide biosynthesis (see Figure 2 in Supplementary Material Section S3). In that particular case, filtering ATP out will result in losing a relevant and valid path. On the contrary, some 'main' compounds can occasionally act as side compounds. For example, fumarate is a main compound of the TCA cycle but is involved as a side compound in the dihydroorotate dehydrogenase reaction (EC 1.3.98.1) in the biosynthesis of uridine monophosphate. The definition of side compounds should thus be context-dependent and not defined for the entire network.

To relax the stringency of filtering approaches and work directly on the original network, one solution consists in avoiding potential side compounds when building the path. This decision can be guided by a weight function attributed to each node or edge. The weighting policy defines a cost for the addition of nodes or edges in the path. Using the degree to weight the node during the shortest path search (also called Lightest path search in the weighted case) was claimed by Croes *et al.* [54, 55] to be a way to avoid side compounds in most cases without removing them, allowing the compounds to be taken into consideration if no better way to connect the two compounds is found. The stringency of the weighting can be increased by using the square of the degree or the cube of the degree.

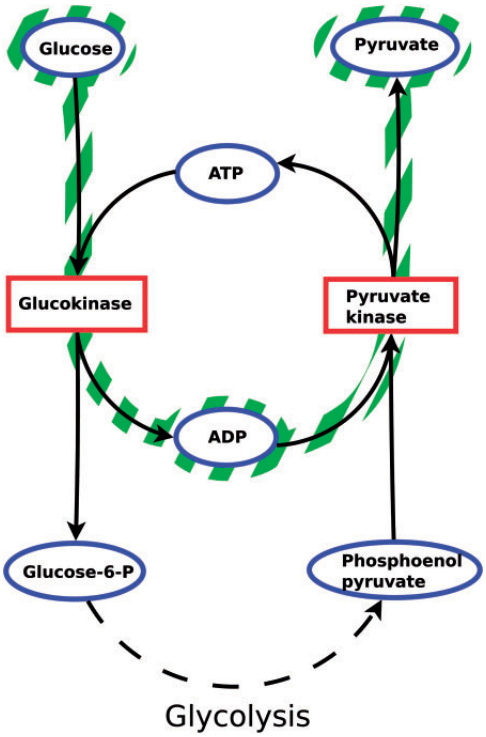


Figure 2. Irrelevant path from glucose to pyruvate. The dashed line represents the glycolysis pathway.

However, degree weighting does not solve the problem of context dependence. The cost of adding a compound is the same whatever the incoming path. One way to create a context-dependent path search is to add weights on edges instead of nodes. The cost of adding a compound then depends on the incoming node.

Defining side compounds using biochemical criterion

Some biochemistry rules can be added to a path search to guarantee paths are meaningful instead of relying on topological criterion such as degree. Some attempts have been made to create context-dependent path-building algorithms using information on chemical structure.

Atom mapping rules. Atom mapping makes it possible to compute molecule alignment by representing a molecule as a graph, with atoms as nodes and bounds as edges. This problem is referenced as a computationally difficult problem (NP-hard problem [56]), roughly meaning that solving this mapping problem on large molecules can become intractable on classical computers. Even if some approximations can be achieved, they will still require long computation time when applied to genome-scale networks (see [57] for a review of current methods for atom mapping computation). To overcome this problem, some databases like KEGG precompute atom mapping and make this information available through dedicated databases (e.g. KEGG's RPAIRS database [58]).

Using atom mapping, Arita [48, 59] and Blum and Kohlbacher [60] proposed a custom shortest path algorithm, which adds a compound to a path under the condition that it contains at least one atom from the source. The underlying idea of this approach is to reproduce *in silico* an isotopic atom tracking experiment.

Atom mapping can also be used as a filtering criterion by removing all the edges in a compound graph that do not represent atomic exchanges or more specifically carbon exchanges between substrate and product. Pey *et al.* showed that filtering to avoid side compounds improved the relevance of a path search result between bicarbonate and cytidine-diphosphate [61].

In silico atom tracking can be done by working on an atomic level network where each metabolite node is split into several atom nodes. Atoms of two different metabolites are linked by an edge if they are predicted to be exchanged between a substrate and a product during a reaction. Finding the shortest path on this network enables the atom conservation constraint to be fulfilled. Previous works also added weighting heuristic to find paths that conserve the maximum number of the source's atoms [62–64].

The lightest path approach, using degree weighting policy, can also be combined with this atom tracking filtering. Blum and Kohlbacher showed that this combination gave a better performance than the same methods used separately when trying to retrieve 137 experimentally verified paths from EcoCyc [65, 66]. Valid atom mapping can nevertheless lead to an irrelevant path, as atoms can be transferred from main to side compounds. Blum and Kohlbacher proposed removing meaningless mapping by filtering reactant pairs with carbon transfer patterns that are not representative of their class of reactions (given by the first three digits of EC numbers).

KEGG's RPAIRS also define a manually curated 'main' substrate–product pair for each reaction [58]. These were used by Faust *et al.* as graph filtering criteria, keeping only main pairs, and also used to define a weighting policy, penalizing pairs that are not classified as 'main' instead of removing them [67]. The main drawback of RPAIRS is that it makes strong assumptions on which transition in a reaction is the main one. The other drawback is that RPAIRS does not cover all compounds and reactions of the KEGG ligand database. More importantly, RPAIRS are only available in KEGG and may not cover all the reactions described in other databases. With the development of resources other than KEGG to access genome-scale metabolic reconstructions [68], such as BioModels [69, 70] and MetExplore [71], it is preferential to have methods that do not rely on specific databases.

Chemical similarity rules. Faced with the time-complexity of atom mapping computation, one alternative, proposed by Rahman *et al.*, consists of using chemical similarity as edge weighting in a compound graph [72, 73]. To compute chemical similarity, each molecule is translated into a bit vector, called a fingerprint, to take advantage of the extremely efficient computation of their comparison. Structural keys are a type of fingerprint in which position represents a kind of atom or a chemical substructure [74]. The bit is set to 1 if the sub-structure is present, otherwise to 0 (See Figure 3B for an example). Other types of fingerprints do not have a predefined set of substructures, but this set is built from the molecule itself using a graph traversal algorithm [75].

Many distances or similarity indexes can be computed [76]. The most commonly used for similarity computation of molecular fingerprints is the Tanimoto coefficient (also known as the Jaccard index), which corresponds to the ratio of the size of the intersection of the two sets to the size of their union.

Finding the lightest path on the graph weighted using chemical similarity makes it possible to find a path that minimizes the structural differences between substrate and product at each step, thereby avoiding side compounds. In addition to this

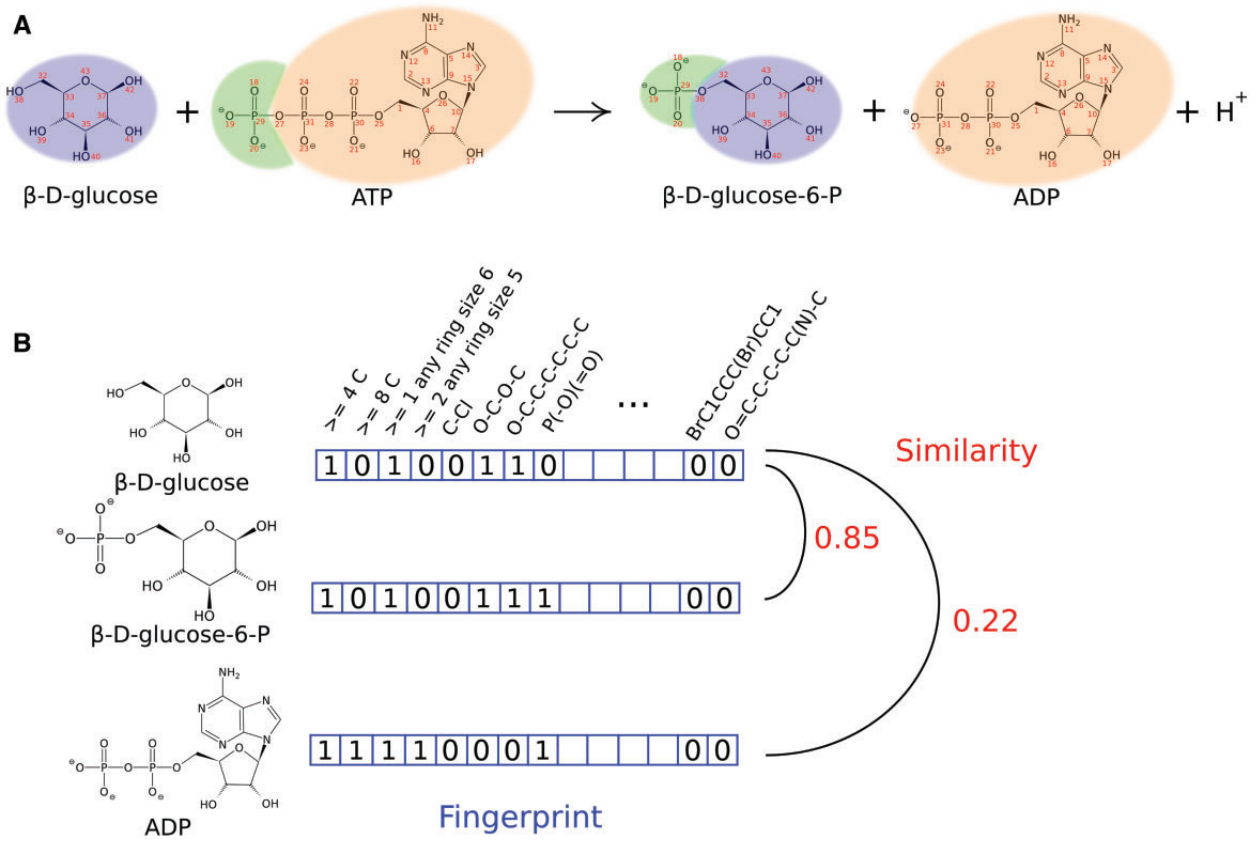


Figure 3. Atom mapping and chemical similarity computed on the substrates and products of the glucokinase reaction. (**A**) Atom mapping. (**B**) Chemical fingerprint. Examples of sub-structures extracted from PubChem Fingerprint. Similarity computed using the Tanimoto index on PubChem Fingerprints.

local similarity defined for each biochemical transition, a global similarity criterion computed from the source metabolite and the current node can be used to guarantee the relevance of the path. Pertusi *et al.* also proposed avoiding the addition of a node, which would reduce the similarity between the current node and the target node [73].

Unlike atom mapping, fingerprints neither represent explicitly the number of occurrences of a given structure, nor how each substructure is linked to other ones in the molecule. As shown in Figure 3A, atom mapping shows that no atoms are exchanged between glucose and ATP during the glucokinase reaction. However, as they share some common structural features (a six-atom ring for example), their chemical similarity is not null (Figure 3B). The chemical similarity weight can thus allow some paths without atomic continuity. However, fingerprints have the advantage of being computable on the fly, and can also focus on sub-structures associated with biological activity [74]. One implementation of the shortest path algorithm based on chemical similarity is provided by Pathway Hunter Tool web server [72].

This method aims to minimize structural dissimilarity at each step and can discard a low conservation transition even if the produced molecule is similar to the target one. An alternative proposed by McShan *et al.* is to use the A* algorithm [77]. Informally, at each step of the search, a function is used to 'guess' the remaining distance to the end node, avoiding assessing parts of the graph that are less likely to be involved in the path. This algorithm does not necessarily find the shortest path, but the best path according to the 'guessing' criterion used. McShan *et al.* used the chemical similarity between the current compounds and the target compound to define the best choice at each step.

Unlike the degree weighting scheme, both atom conservation and chemical similarity are defined for a specific substrate–product pair. This makes it possible to discard a side compound depending on the context of the reaction producing it, and could consider a mechanism involving high-degree compounds as main intermediaries.

Compartmentalized network case

Chemistry-based weighting policy has limitations when applied to a compartmentalized network owing to the presence of transport reactions. Transport edges have the same molecule as source and target (leading to a perfect chemical match between them). Thus, transport edges carry underestimated weights and are more likely to be added to paths. This may drain paths in other compartments than those of the input compounds. Furthermore, untargeted metabolomics is not yet able to distinguish in which organelles metabolites are detected, forcing the user to make an arbitrary choice regarding location of the compounds in the cell when searching for paths in compartmentalized networks. However, if information on location has been gathered from other experiments, or if a particular interest in a given transporter has been shown, these data can be used to compute a custom weighting or filtering policy. Otherwise, the analysis can be performed on a non-compartmentalized network by merging all compartments. However, the paths may then involve reactions that are specific to

different compartments without guaranteeing that there is a transport reaction that makes this path possible.

Co-substrate availability

The methods presented here do not take into account the availability of all substrates required by a reaction. This can lead to a path using inaccessible co-substrates. Other methods based on flux distribution, such as elementary modes and extreme pathways, can handle this kind of issue [78]. However, they usually require some extra level of information to guarantee the stationary state of the sub-graph found (such as classified internal and external compounds by defining the boundaries of the system under study). Those data may be lacking in metabolic networks and are not directly provided by metabolomics methods.

Some attempts have been made to add constraints on co-substrates in path search using hypergraph structure. For example, Mithani *et al.* proposed forbidding the addition of a compound to a path if it is already used as co-substrate [79].

From linear paths to branched sub-networks

Finding relevant alternatives to the 'optimal' path

The aim of the previously described algorithms is to find an optimal linear route. Optimality relies on a parsimonious assumption. The scope of the parsimonious assumption is fixed by the choice of the weighting policy (number of steps, chemical similarity, etc.). But this assumption is not always checked, for example, the citric acid pathway contains a lot of 'by-passes' (see pathway in Supplementary Material Section S3), which are optimal in terms of the number of steps, but which rarely occur in physiological state. Hence, the shortest path-based algorithms focus on these shortcuts and fail to recover the whole citric acid pathway. Furthermore, this approach can only partially retrieve mechanisms that involve cycles, as a path search forbids going through a node twice. In general, these kinds of approaches lead to masking of relevant alternative paths.

One way to overcome the 'optimal only' path is to use the K-shortest paths union to retrieve the k-best ranked paths [80, 81], thereby making it possible to identify possibly relevant alternatives [59].

Handling metabolic profiles composed of more than two compounds

Most of the work described so far focuses on retrieving paths between pairs of compounds. In metabolomics analysis, the list of compounds of interest generally contains more than two metabolites. The simplest way to handle this list using a path-based approach is to process each possible pair independently and then to merge all the paths found. Nevertheless, this problem breakdown may mask relevant parts of the network. In Figure 4, the part of the graph with striped background corresponds to the union of the shortest paths between nodes of interest (in black). Each of these shortest paths is composed of three steps and is completely independent; thus, the overall number of edges in the final sub-graph is six. The union of shortest paths thus fails to emphasize the part of the graph highlighted with solid background, linking all the nodes of interest with an overall number of edges of five. This core sub-network potentially highlights a metabolic backbone connecting the three compounds. The following methods aim to simultaneously account for all compounds to find these kinds of central sub-networks.

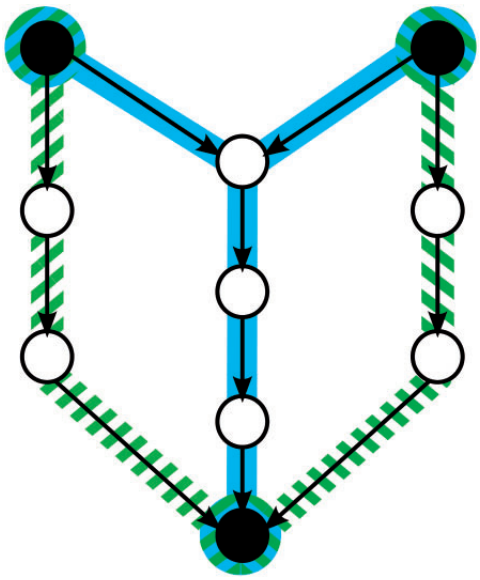


Figure 4. The by-pair path search bias. The part of the graph with striped background corresponds to the union of all shortest paths between the black nodes, with an overall weight of 6. The part of the graph highlighted with a solid background corresponds to the Steiner Tree, with an overall weight of 5.

Steiner tree

Instead of relying on a pairwise path search, Faust *et al.* propose computing the Steiner tree between all the compounds of interest [82]. Formally, a tree is defined as a graph without a cycle where any nodes can be connected by a unique path [47] (see striped and solid sub-networks in Figure 4). A Steiner tree is the minimum-cost tree (where cost is defined by the sum of the weight of all the edges it contains) connecting all nodes of interest [83, 84], making it possible to compute branching pathways. Because solving the Steiner tree problem is not feasible on large graphs such as genome-scale networks [84, 85] (NP-Hard problem), most of the implemented methods rely on approximations (see [86] for an example).

Metabolic stories

Using Steiner trees, two nodes can only be linked by one path, and so alternative solutions that would better model metabolism plasticity may be missed. An alternative method, introduced by Milreu *et al.*, enumerates all maximum directed acyclic sub-graphs (each called a 'metabolic story') connecting compounds of interest [30, 87]. Sub-network extraction is not performed on the whole network but on the union of all the lightest paths between compounds of interest. In practice, the method generates hundreds of alternative stories. The challenge is then to reduce this number to focus on the most relevant scenarios. The method proposed by Milreu *et al.* consists in scoring metabolic stories by preferentially considering stories starting with compounds of interest with decreasing concentrations to ones with increasing concentrations. This makes it possible to focus on metabolic stories in agreement with a selected biological interpretation: a 'domino effect' of changes in concentration or enzyme activations or inhibitions. In addition to ranking, the first step of the method implies that the story search is not applied to the whole network. Finally, avoiding cycles may discard some relevant alternative paths.

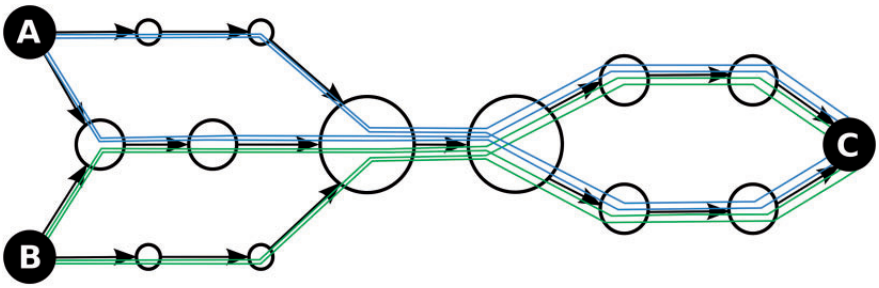


Figure 5. Betweenness centrality defined from a set of nodes of interest. The nodes of interest are in black. There are eight possible 'shortest paths' between the black nodes. For a given node, geodesic centrality represents the proportion of paths that contain it. The size of the nodes is proportional to their centrality.

Centrality-based sub-network extraction

Path-based sub-network extraction keeps all nodes belonging to at least one path between two nodes of interest. But this approach does not differentiate between nodes that belong to only a few paths and those involved in many. These overrepresented nodes can be considered as more relevant, as they may constitute bottlenecks between nodes of interest. An alternative to path-based extraction is computing a relevance score for each element in the network and then keeping only elements with high scores. One of the graph measures reflecting this property is centrality [35, 88, 89], which can be defined in several ways as presented in the following.

Betweenness centrality [90] (or betweenness) can be seen as the propensity of a node (or an edge) to be on a path between any nodes in the network. The 'importance' of a node according to this definition represents their control over the overall flow in the network. The use of the term 'betweenness' usually refers to one particular way to compute it, geodesic betweenness, which considers only the shortest paths between each pair of nodes in the graph. This measure can be computed for a set of nodes of interest: betweenness then becomes the propensity of a node to be on the shortest path between nodes in a particular set (see Figure 5).

The geodesic method based on shortest paths may still disregard relevant alternatives by never taking certain alternative paths into account. Newman *et al.* proposed a method to compute betweenness centrality without focusing on the shortest path by using Random Walks [91]. The idea is to randomly traverse a graph starting from each node of interest until reaching another one. The betweenness of a node will correspond to its overall traversal occurrences. Random Walk betweenness can be estimated through a walk simulation or computed using Markov chain properties. Application to metabolic networks was proposed by Callut *et al.* and Faust *et al.* [82, 92, 93], and is available on the NeAT web server [94]. One prerequisite of this approach is that each node has to be connected to at least one of the target nodes of interest [91]. This assumption is not necessarily true in genome-scale metabolic networks. It thus requires a mandatory pruning step to remove all the nodes that do not satisfy this constraint. Random Walk centrality can also be approximated using bounded length walks [92].

An alternative, which had been intensively used in media and social networks, consists in measuring the likelihood of encountering a node during any network traversal starting from nodes of interest, regardless of whether the traversal will end at a node of interest. This Eigenvector centrality was popularized by the PageRank algorithm developed by Google [95]. PageRank ranks web pages according to their importance on the whole World Wide Web. It can also be applied to other kinds of networks. Many custom PageRank measures, such as PageRank with priors, make it possible to compute relevance according to a set of nodes of interest. Recently, PageRank also attracted interest in a biological network study [96]. Zhang *et al.* proposed a customized implementation of the PageRank algorithm specific to metabolic networks; this implementation is available in SubNet Software [97].

A relevant sub-network can then be extracted by removing all elements with a centrality below a given threshold or by adding the most central edges until all compounds of interest are connected. Both methods have their drawbacks: the first approach can lead to disconnected sub-graphs, and the second can add some low-centrality nodes to the sub-graph.

As path-based methods compute paths with distinct nodes and edges, mechanisms involving cycle or 'loop reaction' such as fatty acid elongation are not covered by those methods. Other methods such as metabolic stories or Steiner Tree are also unable to handle cycles. Sub-network extraction from relevance scoring has the advantage of not being limited to a restrictive topology such as a tree or acyclic graphs. However, in practice, these methods often lead to large networks compared with path-search merging. Although the obtained sub-networks are meaningful, they are more difficult to interpret and often lack specificity compared with traditional human-defined pathways [82].

Centrality itself can also be used as a weight in metabolic graphs. Path or branched path search algorithms can then be applied on the centrality weighted graph. Faust *et al.* showed that the best pathway coverage was obtained by applying Steiner Tree approximation on a Random Walk betweenness weighted network [82].

Other graph-based methods compute the relative importance of different network elements. Some measure the impact of a given node or edge removal (e.g. whether its removal disconnects the graph) on graph connectivity, while others measure the average distance between a given node and any other nodes in the network. Nevertheless, betweenness centrality fits sub-network extraction better because it highlights compounds that are more likely to be encountered on a path between metabolites of interest.

Remaining challenges

Dealing with network incompleteness

Currently, most metabolic networks contain erroneous reactions and gaps [16]. This can lead to biased interpretation of experimental data, regardless of the graph method and associated parameters chosen. Compared with path search methods, centrality measures can be more robust to this kind of error. They highlight compounds or reactions that are more likely to be

related to the compounds of interest (considering all possible paths), instead of providing one or a few paths that could fall into an erroneous part of the network. Another alternative is to infer missing reactions [73, 98, 99], for example, by adding putative reactions based on the chemical similarity of the substrate and the product [73], as many enzymes show promiscuous activities.

Metabolite structural information (usually provided as InChI [100] or SMILES [101]) is also often lacking in genome-scale metabolic networks [72]. This incompleteness means that, when chemistry-based methods are used, the path search could be conducted on a partial network. Nevertheless, cross-references allow some of these data to be retrieved using various compound databases [102, 103] and their related web services (ChEBI [104], PubChem [105], KEGG [9]). Finally, metabolic networks can involve large molecules (several thousand Daltons) whose structure cannot be explicitly described as a single character sequence and generic compounds ('alcohol', 'fatty acid', etc.).

Handling reaction reversibility

Metabolic networks are usually represented as directed graphs, meaning that each edge carries a direction that distinguishes substrates from products. For reversible reactions, two edges pointing in opposite directions can be added. This particular configuration creates some problems for path search algorithms in metabolic networks. For instance, the same reaction can be traversed twice, once in each direction, creating meaningless shortcuts by linking two metabolites from the same side of the reaction's biochemical equation. Path search algorithms can be adapted to keep track of reactions already added during path building to avoid going through a reversible reaction twice. However, to our knowledge, the combinatorial implications of this modification have not been discussed, and there is no formal proof that the resulting path is the shortest. In fact, the work of Fekete *et al.* [106] shows that this constraint on the shortest path leads to an NP-complete problem, proving that this is not a trivial task. It has to be noticed that storing used reactions is not applicable to centrality methods that do not explicitly compute paths such as Random Walk or Eigenvector centrality.

As commented by Croes *et al.* [55], reaction reversibility in databases is defined under particular physiological conditions (pH, substrate concentrations, etc.). It can lead to irrelevant paths if the experimental conditions differ from those used to build the model. One solution consists in considering all reactions as reversible, by defining the metabolic graph as undirected. However, it is difficult to guarantee the validity of this assumption under the experimental condition for each reaction, and thus to interpret the validity of paths obtained. Experimental or *in silico* flux predictions are one way to define these directions. However, experimental data are not always available, and flux computation through constraint-based modelling requires additional parameters (system boundaries, constraints on fluxes and, in most cases, an objective function).

Handling reaction reversibility is a major challenge for the biological relevance of results. It is still an open problem.

Why has no consensus been reached on an appropriate method?

Assessing the quality of sub-networks

First, it should be noted that assessing the quality of the different algorithms and associated parameter sets is challenging. Some comparisons between sub-network extraction methods [55, 82, 107] compare extracted sub-networks and existing pathways. Comparing sub-networks with pathways is a good indicator of the relevance of the result because it automatically links extracted sub-networks to known biological functions. However, this assessment fails to emphasize the main advantage of the sub-network extraction approach, namely, highlighting novel pathways or mechanisms that span multiple pathways. Pathways also represent a global view of a biological function, including side compounds and multiple entries, while sub-network extraction focuses on the succession of biochemical reactions linking several compounds. Therefore, the fact that a sub-network only partially matches a pathway does not imply that it is meaningless or irrelevant.

Defining an appropriate criterion to assess the quality of a sub-network is still an ongoing topic and a key challenge in graph theory applied to metabolic networks.

Comparison of weighting policies on two test cases

We applied several methods alone or in combination on two non-trivial use-cases likely to raise specific issues. The aim was to illustrate the main implications of the chosen methods, and the results cannot be extrapolated to every use-case. Therefore, this analysis should not be interpreted as a quality assessment of the methods presented here, but as an indicator of the advantages and drawbacks of topology-based and biochemistry-based strategies, the two main strategies used.

The network used is a compound graph built from genome-scale metabolic network Recon2 from Thiele *et al.* [21]. Because global metabolomics cannot distinguish the location of compounds in a cell, the original network is transformed into a flat network with no compartments. The shortest paths were computed using the Dijkstra algorithm [108] implemented in JAVA. During the path search, a new edge was added only if the current best path did not already contain the associated reaction. Graph manipulations and basic operations were performed using the java JGraphT library. The chemical fingerprints were computed using CDK Java Library [109], and their similarity was computed using the Tanimoto coefficient. For shortest path-based extraction, similarity was used to compute edge weight. For the degree weighting policy, the valuation of an edge is the degree of the target node. Atom mapping was extracted from KEGG RPAIRS database. In Figures 6 and 7, it can be seen to facilitate interpretation because atom mapping can be a good indicator for manually checking the relevance of a path; however, it does not necessarily represent the path that would be obtained using shortest path combined with atom mapping weighting or filtering. It should be noted that the shortest paths presented are not necessarily the only possible solution, as many paths can share the same weight.

A test case involving large cofactors: propanoyl-CoA degradation

We applied the shortest path algorithm between propanoyl-CoA and succinate using different weighting policies (see Figure 6). Without any weighting, the shortest paths found are irrelevant shortcuts going through phosphate or ADP. Degree, chemical similarity or atom tracking is able to discard those side compounds. Most of the fingerprint types used are able to retrieve the path as described in HumanCyc [110] (see Supplementary Material Section S3). However, the chemical similarity computed from the MACCS fingerprint goes from propanoyl-CoA to succinyl-CoA through coenzyme A. Because the coenzyme represents the major part of the two molecules, the similarity between them and coenzyme A is

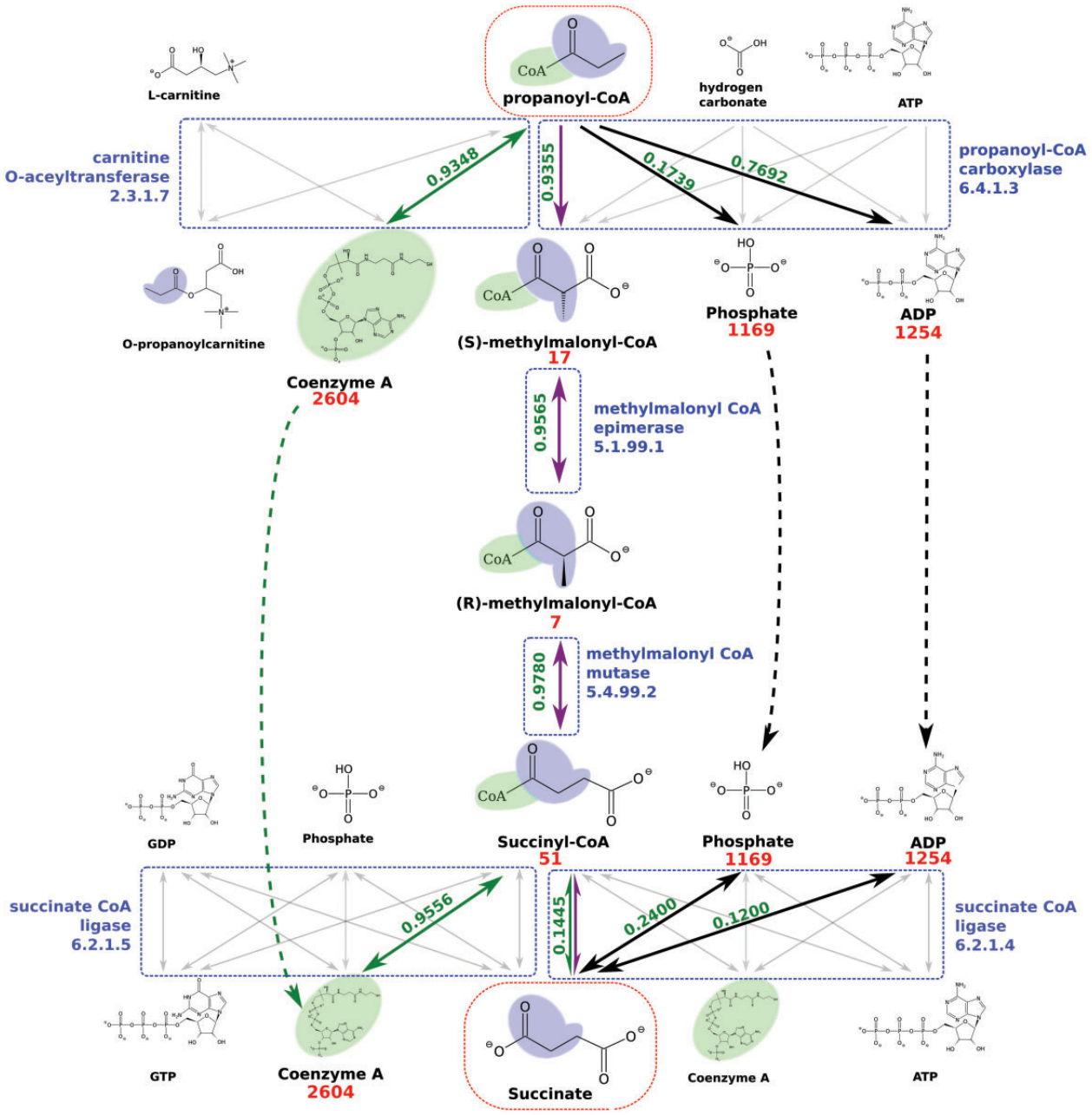


Figure 6. Path searches applied to the propanoyl-CoA to succinate degradation pathway using different weighting policies. Edges involved in at least one path are in bold. MACCS fingerprint similarities are displayed on edges, compound degrees are displayed below nodes. The part of the molecules highlighted in the same colour represents atom mapping. The pathway as described in HumanCyc, shown here in the middle, goes through (S)-methylmalonyl-CoA, (R)-methylmalonyl-CoA and succinyl-CoA. The same path was retrieved using Dijkstra shortest path algorithm on Recon2 human global network weighted by degree, degree square, degree cube and chemical similarity using PubChem, Klekota-Roth, EState and CDK's extended fingerprints. The shortest paths computed on unweighted graph only use ADP or phosphate as intermediary compounds (paths on the right); the shortest path computed using MACCS and CDK's substructure fingerprint (path on the left) goes from propanoyl-CoA to succinate through coenzyme A and succinyl-CoA.

close to 1. This sub-path is also valid from the point of view of atom tracking.

It should be noted that results will vary depending on the type of fingerprint used (see Supplementary Material Section S2). Unfortunately, this choice is not clear and, to our knowledge, the respective qualities of different types of fingerprints have not yet been assessed for use in a metabolic path search.

By disabling the use of the same reaction twice, the succinate cannot be reached from coenzyme A using both directions of the ATP-dependent succinate-CoA ligase reaction. However, it may be accomplished by successively using GTP-dependent and ATP-dependent ligase reactions, as they are considered as different reactions. This result can be recognized as being irrelevant, as the production of succinate would require it as a substrate for an intermediary reaction.

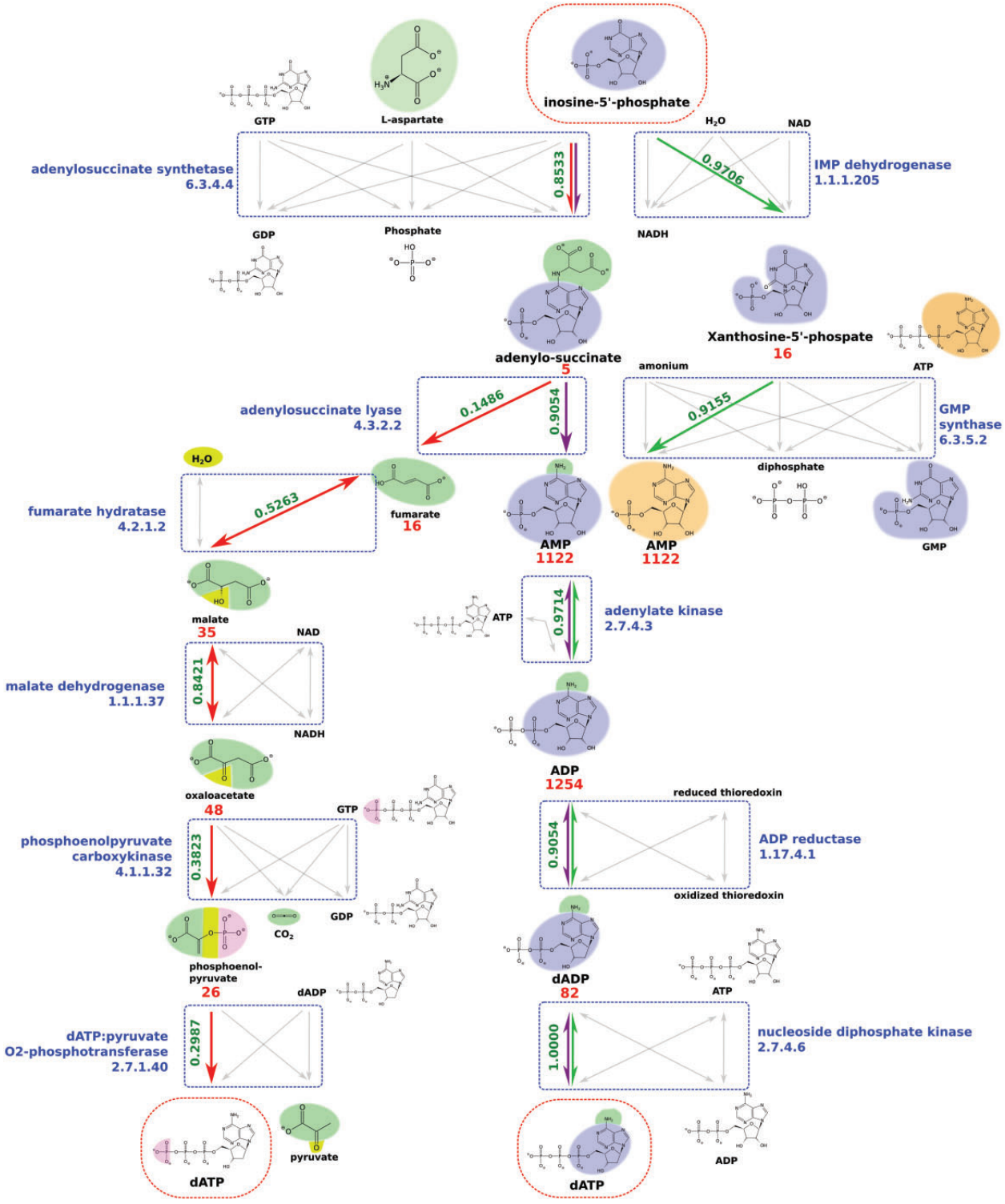


Figure 7. Path searches applied to the adenosine deoxyribonucleotide biosynthesis from inosine-5'-phosphate using different weighting policies. Edges involved in at least one path are in bold. MACCS fingerprint's similarities are displayed on edges, compound degrees are displayed below nodes. The part of molecules highlighted in the same colour represents atom mapping. The pathway as described in HumanCyc, shown here in the middle, goes through adenyl-succinate, adenosine monophosphate (AMP), adenosine diphosphate (ADP) and deoxyadenosine diphosphate (dADP). The path retrieved using the Dijkstra shortest path algorithm on Recon2 human global network weighted by degree goes (path on the left) through fumarate, malate, oxaloacetate and phosphoenolpyruvate. Chemical similarity using MACCS, PubChem, Klekota-Roth, EState and CDK's extended fingerprints (path on the right) goes through xanthosine-5'-monophosphate, adenosine monophosphate (AMP), adenosine diphosphate and deoxyadenosine diphosphate.

A test case involving high-degree compounds: de novo biosynthesis of adenosine deoxyribonucleotides

The shortest path was computed on the same network between inosine-5'-phosphate (IMP) and adenosine deoxyribonucleotide (dATP) (see Figure 7). As *de novo* biosynthesis of dATP involves high-degree compounds such as AMP and ADP (see pathway in Supplementary Material Section S3), using the degree as weighting policy failed to reconstruct it automatically. In practice, degree-based weighting fails to reconstruct 'central' pathways linked to many other pathways (such as glycolysis or the citric acid cycle), as a lot of their intermediary compounds have high degrees [66].

Chemical similarity makes it possible to retrieve most parts of this pathway, except that AMP is reached from xanthosine-5'-phosphate (XMP) because it is more similar to both AMP and IMP than adenylosuccinate. However, a closer look at the reaction linking XMP to AMP shows that, despite high similarity, they share no atoms. This bias may be encountered when two substrates or two products share high similarity (in this case XMP and ATP), making elucidation of main transitions more difficult. In the test case presented here, a simple filtering step using RPAIRs annotation, which removes all edges where no atoms are shared, makes it possible to retrieve the most relevant path, if combined with the chemical similarity method.

Conclusion and suggestions on how to efficiently find sub-networks

The methods described in this review have great potential for the interpretation of metabolic profiles containing metabolites with significant concentration changes between experimental conditions. In fact, by extracting potentially affected part of the global network, these methods provide some insight into the mechanisms associated with perturbations [111].

Several sub-network extraction methods rely on metabolic path search (see Section S4 in Supplementary Material for a non-exhaustive overview of tools implementing these methods). As mentioned by Planes *et al.* [107], the main challenge in metabolic path search is to define appropriate constraints to discard meaningless paths. For example, constraints on nodes' degree will produce paths avoiding too generic molecules, constraints on atom mapping will produce paths with atomic continuity and constraints on chemical similarity will produce paths without links between too chemically different molecules.

This review shows that choosing among these methods is not trivial because each one may be the most efficient for a particular biochemical case. Degree-based methods can lead to biased conclusions when dealing with mechanisms involving high-degree compounds. On the contrary, chemical-based methods can fail when dealing with mechanisms involving large cofactor attachments. Chemical-based methods are also limited to mechanisms that do not involve macromolecules or generic compounds.

This discrepancy explains why there is currently no consensus method to answer the relevant path search problem. A good approach would be to test several methods and combine them (filter and weighting for example), especially if the molecules involved in the paths are susceptible to fall into the special cases described earlier. This observation highlights the need for toolboxes including a large range of approaches, to easily combine them or compare their respective results.

Beyond the choice of the algorithm and of the appropriate parameters, the quality of the results relies, to a great extent, on reconstruction of the underlying network. Moreover, chemical-based methods would benefit from increasing the information available in genome-scale networks on chemical structure (such as InChI or SMILE) and on reactant pairs for each reaction (such as RPAIR annotation or mapping). Other remaining challenges include handling reversible reactions, using compartment information and dealing with generic compounds.

Here we present graph-theory methods in the context of metabolomic studies for downstream interpretation of metabolomic profiles. However, they are not limited to metabolic profiles and can be used for bioengineering purposes by arising some potential production routes between compounds of interest, see for example [29, 112]. As these methods are not necessarily restricted to compound-to-compound paths, they also make it possible to explore relations between metabolites and reactions of interest pointed out by transcriptomic or proteomic results [113]. Another popular analysis based on path search and centrality is load point computation [114]. This index describes the number of paths that use a given reaction and enables ranking according to choke points, for example. It can be used to estimate the potential lethality of a given reaction and help identify potential drug targets [115–117]. Path search methods can also be used earlier in the data processing pipeline to facilitate the identification of metabolites [118–121]. Finally, path search can also be used in the network reconstruction process and to predict a new synthetic path by inferring missing compounds and reactions. This is the purpose of the algorithms implemented in PathPred [99, 122] [or MetaMapp [123].

Supplementary data

Supplementary data are available online at http://bib.oxfordjournals.org/.

Key Points

- Reviews the different sub-network extraction approaches that can be used for the interpretation of metabolomic profiles.
- Describes methods that take into account network topology and the chemical structure of the metabolites.
- Exemplifies the difficulty of assessing the quality of these approaches.
- Suggests testing several methods and compiling results to generate complementary hypotheses.

Acknowledgements

The authors thank Nathalie Poupin, Helen Purchase, Sanu Shameer and Ludovic Cottret for their valuable comments and remarks.

Funding

This work was supported by the French Ministry of Research and National Research Agency (ANR) as part of the French MetaboHUB, the national metabolomics and fluxomics infrastructure (Grant ANR-INBS-0010).

References

1. Sévin DC, Kuehne A, Zamboni N, *et al.* Biological insights through nontargeted metabolomics. *Curr Opin Biotechnol* 2015;**34**:1–8.

2. Nicholson JK, Lindon JC, Holmes E. 'Metabonomics': understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica* 1999;**29**:1181–9.
3. Raamsdonk LM, Teusink B, Broadhurst D, *et al.* A functional genomics strategy that uses metabolome data to reveal the phenotype of silent mutations. *Nat Biotechnol* 2001;**19**:45–50.
4. Fiehn O, Kopka J, Dörmann P, *et al.* Metabolite profiling for plant functional genomics. *Nat Biotechnol* 2000;**18**:1157–61.
5. Sumner LW, Amberg A, Barrett D, *et al.* Proposed minimum reporting standards for chemical analysis Chemical Analysis Working Group (CAWG) Metabolomics Standards Initiative (MSI). *Metabolomics* 2007;**3**:211–21.
6. Creek DJ, Dunn WB, Fiehn O, *et al.* Metabolite identification: are you sure? And how do your peers gauge your confidence?. *Metabolomics* 2014;**10**.
7. Pearson H. Meet the human metabolome. *Nature* 2007;**446**:8.
8. Ogata H, Fujibuchi W, Goto S, *et al.* A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res* 2000;**28**:4021–8.
9. Kanehisa M, Goto S, Sato Y, *et al.* Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res* 2014;**42**:D199–205.
10. Caspi R, Altman T, Billington R, *et al.* The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of pathway/genome databases. *Nucleic Acids Res* 2014;**42**:D459–71.
11. Xia J, Wishart DS. MetPA: a web-based metabolomics tool for pathway analysis and visualization. *Bioinformatics* 2010;**26**:2342–4.
12. Xia J, Wishart DS. MSEA: a web-based tool to identify biologically meaningful patterns in quantitative metabolomic data. *Nucleic Acids Res* 2010;**38**:71–7.
13. Droste P, Miebach S, Niedenführ S, *et al.* Visualizing multi-omics data in metabolic networks with the software Omix case study. *BioSystems*2011;**105**:154–61.
14. Houtkooper RH, Argmann C, Houten SM, *et al.* The metabolic footprint of aging in mice. *Sci Rep* 2011;**1**:134.
15. Mu C, Yang Y, Luo Z, *et al.* Metabolomic analysis reveals distinct profiles in the plasma and urine of rats fed a high-protein diet. *Amino Acids* 2015;**47**:1225–38.
16. Ginsburg H. Caveat emptor: limitations of the automated reconstruction of metabolic pathways in Plasmodium. *Trends Parasitol* 2009;**25**:37–43.
17. Schilling CH, Schuster S, Palsson BO, *et al.* Metabolic pathway analysis: basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog* 1999;**15**:296–303.
18. Thiele I, Palsson BØ. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat Protoc* 2010;**5**:93–121.
19. Faust K, Croes D, van Helden J. Prediction of metabolic pathways from genome-scale metabolic networks. *Biosystems* 2011;**105**:109–21.
20. Faust K, van Helden J. Predicting metabolic pathways by sub-network extraction. *Methods Mol Biol* 2012;**804**:107–30.
21. Thiele I, Swainston N, Fleming RMT, *et al.* A community-driven global reconstruction of human metabolism. *Nat Biotechnol* 2013;**31**:419–25.
22. Kell DB. Metabolomics and systems biology: making sense of the soup. *Curr Opin Microbiol* 2004;**7**:296–307.
23. Schuster S, Hilgetag C. On elementary flux modes in biochemical reaction systems at steady state. *J Biol Syst* 1994;**2**:165–82.
24. Pfeiffer T, Sánchez-Valdenebro I, Nuño JC, *et al.* METATOOL: for studying metabolic networks. *Bioinformatics* 1999;**15**:251–7.
25. Von Kamp A, Schuster S. Metatool 5.0: fast and flexible elementary modes analysis. *Bioinformatics* 2006;**22**:1930–1.
26. Raman K, Chandra N. Flux balance analysis of biological systems: applications and challenges. *Brief Bioinform* 2009;**10**:435–49.
27. Bordbar A, Monk JM, King ZA, *et al.* Constraint-based models predict metabolic and associated cellular functions. *Nat Rev Genet* 2014;**15**:107–20.
28. Burrage K, Hood L, Ragan MA. Advanced computing for systems biology. *Brief Bioinform* 2006;**7**:390–8.
29. Planson A-G, Carbonell P, Grigoras I, *et al.* Engineering antibiotic production and overcoming bacterial resistance. *Biotechnol J* 2011;**6**:812–25.
30. Milreu PV, Klein CC, Cottret L, *et al.* Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure. *Bioinformatics* 2014;**30**:61–70.
31. Burt RS. The social structure of competition. *Netw Organ Struct Form Action* 1992;**57**:91.
32. Mackaness W, Beard M. Use of graph theory to support map generalization. *Cartogr Geogr Inf Syst* 1993;**20**:210–21.
33. Barabási A-L, Albert R. Emergence of scaling in random networks. *Science* 1999;**286**:509–12.
34. Albert R, Jeong H, Barabási AL. Error and attack tolerance of complex networks. *Nature* 2000;**406**:378–82.
35. Aittokallio T, Schwikowski B. Graph-based methods for analysing networks in cell biology. *Brief Bioinform* 2006;**7**:243–55.
36. Zhu X, Gerstein M, Snyder M. Getting connected: analysis and principles of biological networks. *Genes Dev* 2007;**21**:1010–24.
37. Pavlopoulos GA, Secrier M, Moschopoulos CN, *et al.* Using graph theory to analyze biological networks. *BioData Min* 2011;**4**:10.
38. Lacroix V, Cottret L, Thébault P, *et al.* An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans Comput Biol Bioinform* 2008;**5**:594–617.
39. Cottret L, Jourdan F. Graph methods for the investigation of metabolic networks in parasitology. *Parasitology* 2010;**137**:1393–407.
40. Yeung M, Thiele I, Palsson BØ. Estimation of the number of extreme pathways for metabolic networks. *BMC Bioinformatics* 2007;**8**:363.
41. Klamt S, Haus U-U, Theis F. Hypergraphs and cellular networks. *PLoS Comput Biol* 2009;**5**:e1000385.
42. Pearcy N, Crofts JJ, Chuzhanova N. Hypergraph models of metabolism. *Int J Biol Vet Agric Food Eng* 2014;**8**:812–16.
43. Deville Y, Gilbert D, van Helden J, *et al.* An overview of data models for the analysis of biochemical pathways. *Brief Bioinform* 2003;**4**:246–59.
44. Arita M. The metabolic world of *Escherichia coli* is not small. *Proc Natl Acad Sci USA* 2004;**101**:1543–7.
45. Ma H, Zeng AP. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics* 2003;**19**:270–7.
46. Küffner R, Zimmer R, Lengauer T. Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics* 2000;**16**:825–36.
47. Jungnickel D. *Graphs, Networks and Algorithms*. 2007, New York, USA: Springer Publishing Company.
48. Arita M. Metabolic reconstruction using shortest paths. *Simul Pract Theory* 2000;**8**:109–25.
49. Holme P. Model validation of simple-graph representations of metabolism. *J R Soc Interface* 2009;**6**:1027–34.

50. Van Helden J, Wernisch L, Gilbert D, *et al*. Graph-based analysis of metabolic networks. *Ernst Schering Res Found Workshop* 2002;**38**:245–74.
51. Fell D a, Wagner a. The small world of metabolism. *Nat Biotechnol* 2000;**18**:1121–2.
52. Jeong H, Tombor B, Albert R, *et al*. The large-scale organization of metabolic networks. *Nature* 2000;**407**:651–4.
53. Gerlee P, Lizana L, Sneppen K. Pathway identification by network pruning in the metabolic network of *Escherichia coli*. *Bioinformatics* 2009;**25**:3282–8.
54. Croes D, Couche F, Wodak SJ, *et al*. Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res* 2005;**33**:W326–30.
55. Croes D, Couche F, Wodak SJ, *et al*. Inferring meaningful pathways in weighted metabolic networks. *J Mol Biol* 2006; **356**:222–36.
56. Garey MR, Johnson DS. *Computers and Intractability: A Guide to the Theory of NP-Completeness*. 1979, New York, USA: W. H. Freeman & Co.
57. Chen WL, Chen DZ, Taylor KT. Automatic reaction mapping and reaction center detection. Wiley Interdiscip. *Rev Comput Mol Sci* 2013;**3**:560–93.
58. Kotera M, Hattori M, Oh MA, *et al*. RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Inf* 2004;**15**:62.
59. Arita M. In silico atomic tracing by substrate-product relationships in Escherichia coli intermediary metabolism. *Genome Res* 2003;**13**:2455–66.
60. Blum T, Kohlbacher O. MetaRoute: fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics* 2008;**24**:2108–9.
61. Pey J, Prada J, Beasley JE, *et al*. Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biol* 2011;**12**:R49.
62. Boyer F, Viari A. *Ab initio* reconstruction of metabolic pathways. *Bioinformatics* 2003;**19**:ii26–34.
63. Pitkänen E, Jouhten P, Rousu J. Inferring branching pathways in genome-scale metabolic networks. *BMC Syst Biol* 2009;**3**:103.
64. Latendresse M, Krummenacker M, Karp PD. Optimal metabolic route search based on atom mappings. *Bioinformatics* 2014;**30**:2043–50.
65. Keseler IM, Mackie A, Peralta-Gil M, *et al*. EcoCyc: fusing model organism databases with systems biology. *Nucleic Acids Res* 2013;**41**:D605–12.
66. Blum T, Kohlbacher O. Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J Comput Biol* 2008;**15**:565–76.
67. Faust K, Croes D, van Helden J. Metabolic pathfinding using RPAIR annotation. *J Mol Biol* 2009;**388**:390–414.
68. Oberhardt MA, Papin JA. Applications of genome-scale metabolic reconstructions. *Mol Syst Biol* 2009;**5**:320.
69. Li C, Donizelli M, Rodriguez N, *et al*. BioModels database: an enhanced, curated and annotated resource for published quantitative kinetic models. *BMC Syst Biol* 2010;**4**:92.
70. Wimalaratne SM, Grenon P, Hermjakob H, *et al*. BioModels linked dataset. *BMC Syst Biol* 2014;**8**:91.
71. Cottret L, Wildridge D, Vinson F, *et al*. MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res* 2010;**38**:W132–7.
72. Rahman SA, Advani P, Schunk R, *et al*. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics* 2005;**21**:1189–93.
73. Pertusi DA, Stine AE, Broadbelt LJ, *et al*. Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics* 2014;1–9.
74. Klekota J, Roth FP. Chemical substructures that enrich for biological activity. *Bioinformatics* 2008;**24**:2518–25.
75. Brown N, McKay B, Gasteiger J. Fingal: a novel approach to geometric fingerprinting and a comparative study of its application to 3D-QSAR modelling. *QSAR Comb Sci* 2005; **24**:480–4.
76. Willett P. Similarity methods in chemoinformatics. *Annu Rev Inf Sci Technol* 2005;**43**:1–117.
77. McShan DC, Rao S, Shah I. PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics* 2003;**19**:1692–8.
78. Papin J a, Stelling J, Price ND, *et al*. Comparison of network-based pathway analysis methods. *Trends Biotechnol* 2004;**22**: 400–5.
79. Mithani A, Preston GM, Hein J. Rahnuma: hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics* 2009;**25**:1831–2.
80. Yen JY. Finding the K Shortest loopless paths in a network. *Manage Sci* 1971;**17**:712–16.
81. Eppstein D. Finding the k shortest paths. *SIAM* 1997;**28**: 652–73.
82. Faust K, Dupont P, Callut J, *et al*. Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics* 2010;**26**:1211–8.
83. Gilbert EN, Pollak HO. Steiner minimal trees. *SIAM J Appl Math* 1968;**16**:1–29.
84. Hwang FK, Richards DS, Winter P. *The Steiner Tree Problem*. 1992, Amsterdam, Netherlands: Elsevier.
85. Karp RM. Reducibility among combinatorial problems. In: *Complexity of Computer Computations*. 1972; 85–103.
86. Takahashi H, Matsuyama A. An approximate solution for the Steiner problem in graphs. *Math Jpn* 1980;**24**:573–7.
87. Acuña V, Birmelé E, Cottret L, *et al*. Telling stories: enumerating maximal directed acyclic graphs with a constrained set of sources and targets. *Theor Comput Sci* 2012;**457**:1–9.
88. Borgatti SP, Everett MG. A Graph-theoretic perspective on centrality. *Soc Networks* 2006;**28**:466–84.
89. Klein DJ. Centrality measure in graphs. *J Math Chem* 2010; **47**:1209–23.
90. Freeman LC, Borgatti SP, White DR. Centrality in valued graphs: a measure of betweenness based on network flow. *Soc Networks* 1991;**13**:141–54.
91. Newman MEJ. A measure of betweenness centrality based on random walks. *Soc Networks* 2005;**27**:39–54.
92. Dupont P, Callut J, Dooms G, *et al*. *Relevant Subgraph Extraction from Random Walks in a Graph*. 2006, Louvain, Belgium: Université catholique de Louvain.
93. Callut J. *First Passage Times Dynamics in Markov Models with Applications to HMM: Induction, Sequence Classification and Graph Mining*. 2007, Louvain, Belgium: Université catholique de Louvain.
94. Brohée S, Faust K, Lima-Mendez G, *et al*. NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res* 2008;**36**:W444–51.
95. Page L, Brin S, Motwani S, *et al*. *The PageRank Citation Ranking: Bringing Order to the Web*. 1999, Stanford, CA, USA: Stanford University.
96. Bánky D, Iván G, Grolmusz V. Equal opportunity for low-degree network nodes: a pagerank-based method for protein target identification in metabolic graphs. *PLoS One* 2013;**8**:1–7.
97. Zhang Q, Zhang ZD. SubNet: a Java application for subnetwork extraction. *Bioinformatics* 2013;**29**:2509–11.

98. Ellis LBM, Gao J, Fenner K, *et al*. The University of Minnesota pathway prediction system: predicting metabolic logic. *Nucleic Acids Res* 2008;**36**:427–32.
99. Moriya Y, Shigemizu D, Hattori M, *et al*. PathPred: an enzyme-catalyzed metabolic pathway prediction server. *Nucleic Acids Res* 2010;**38**:138–43.
100. Heller S, McNaught A, Stein S, *et al*. InChI—the worldwide chemical structure identifier standard. *J Cheminform* 2013;**5**:7.
101. Weininger D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J Chem Inf Model* 1988;**28**:31–36.
102. Wohlgemuth G, Haldiya PK, Willighagen E, *et al*. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics* 2010;**26**:2647–8.
103. Bernard T, Bridge A, Morgat A, *et al*. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Brief Bioinform* 2014;**15**:123–35.
104. Davies M, Nowotka M, Papadatos G, *et al*. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* 2015;**43**:612–20.
105. Bolton EE, Wang Y, Thiessen PA, *et al*. PubChem: integrated platform of small molecules and biological activities. *Annu Rep Comput Chem* 2008;**4**:217–41.
106. Fekete S, Kamphans T, Stelzer M. Shortest paths with pairwise-distinct edge labels: finding biochemical pathways in metabolic networks. *CoRR* 2010;**9**.
107. Planes FJ, Beasley JE. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Brief Bioinform* 2008;**9**:422–36.
108. Dijkstra EW. A note on two problems in connexion with graphs. *Numer Math* 1959;**1**:269–71.
109. Steinbeck C, Han Y, Kuhn S, *et al*. The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J Chem Inf Comput Sci* 2003;**43**:493–500.
110. Romero P, Wagg J, Green ML, *et al*. Computational prediction of human metabolic pathways from the complete human genome. *Genome Biol* 2005;**6**:R2.
111. Mahdavi V, Ghanati F, Ghassempour A. Integrated pathway-based and network-based analysis of GC-MS rice metabolomics data under diazinon stress to infer affected biological pathways. *Anal Biochem* 2015;**494**:31–36.
112. Ranganathan S, Maranas CD. Microbial 1-butanol production: Identification of non-native production routes and *in silico* engineering interventions. *Biotechnol J* 2010;**5**:716–25.
113. Kim T, Dreher K, Nilo-Poyanco R, *et al*. Patterns of metabolite changes identified from large-scale gene perturbations in arabidopsis using a genome-scale metabolic network. *Plant Physiol* 2015;**167**:1685–98.
114. Rahman SA, Schomburg D. Observing local and global properties of metabolic pathways: 'Load points' and 'choke points' in the metabolic networks. *Bioinformatics* 2006;**22**:1767–74.
115. Wei R, Li H, Wang Y, *et al*. Identification of colorectal cancer candidate genes based on subnetwork extraction algorithm. *Adv Intell Comput Theor Appl* 2015;**9227**:706–12.
116. Sharma A, Pan A. Identification of potential drug targets in Yersinia pestis using metabolic pathway analysis: MurE ligase as a case study. *Eur J Med Chem* 2012;**57**:185–95.
117. Perumal D, Lim CS, Sakharkar MK. A comparative study of metabolic network topology between a pathogenic and a non-pathogenic bacterium for potential drug target identification. *Summit on Translat Bioinforma* 2009;**2009**:100–4.
118. Jourdan F, Cottret L, Huc L, *et al*. Use of reconstituted metabolic networks to assist in metabolomic data visualization and mining. *Metabolomics* 2010;**6**:312–21.
119. Silva RR, Jourdan F, Salvanha DM, *et al*. ProbMetab: an R package for Bayesian probabilistic annotation of LC-MS based metabolomics. *Bioinformatics* 2014;**30**:1336–7.
120. Rogers S, Scheltema RA, Girolami M, *et al*. Probabilistic assignment of formulas to mass peaks in metabolomics experiments. *Bioinformatics* 2009;**25**:512–18.
121. Jourdan F, Breitling R, Barrett MP, *et al*. MetaNetter: inference and visualization of high-resolution metabolomic networks. *Bioinformatics* 2008;**24**:143–5.
122. Kotera M, Tabei Y, Yamanishi Y, *et al*. Supervised de novo reconstruction of metabolic pathways from metabolome-scale compound sets. *Bioinformatics* 2013;**29**:i135–44.
123. Barupal DK, Haldiya PK, Wohlgemuth G, *et al*. MetaMapp: mapping and visualizing metabolomic data by integrating information from biochemical pathways and chemical and mass spectral similarity. *BMC Bioinformatics* 2012;**13**:99.

4.3 Discussion

L'article présenté propose plusieurs pistes pour générer des chemins porteurs de sens dans les réseaux métaboliques, notamment en évitant l'incorporation de métabolites auxiliaires. Comme mentionné précédemment, le degré peut être une option, bien que certains composés auxiliaires ne soient pas des *hubs*, par exemple l'ammoniac, et inversement, par exemple le pyruvate. Il a été proposé que la définition même de composé auxiliaire demeure arbitraire et dépendante d'un contexte. Ainsi a-t-il été mentionné l'exemple de l'implication de l'ADP dans la voie de synthèse des nucléotides, où il constitue l'un des intermédiaires principaux. Il a donc été proposé que ce soit les transitions substrats-produits qui soient considérées comme auxiliaires plutôt que les composés eux-mêmes. Cette approche s'intègre aisément lorsque le réseau métabolique est représenté sous forme de graphe des composés. Ces transitions sont alors explicitement représentées par des arcs, et une pondération peut par conséquent être intégrée pour moduler leur incorporation à des chemins.

Il est également possible de supprimer ces arcs du réseau, communément en établissant un seuil sur le critère utilisé pour définir les transitions auxiliaires. Dans le cas des graphes bipartis, aucune relation directe n'existe entre un couple substrat-produit, il n'est donc pas possible d'intégrer d'informations au niveau des transitions entre ces couples, telles que la similarité chimique[268][269] ou l'échange d'atomes par exemple. Cette limite nous a conduit à considérer le graphe des composés pour les travaux présentés dans la suite de cette thèse.

La similarité chimique et les proportions d'atomes échangés peuvent offrir des critères de pondérations et de filtres pertinents pour traiter les transitions auxiliaires. Néanmoins, différentes méthodes existent pour obtenir ces valeurs[71], et leur choix peut induire des différences notables. La similarité chimique est classiquement définie par la distance de Tanimoto entre des *fingerprints*[257], vecteurs binaires dont chaque position correspond à des groupements chimiques ou des propriétés structurelles, et la valeur de ces éléments correspond à leur présence ou leur absence[152][251]. Les différentes méthodes portent principalement sur les

groupements et propriétés considérées, et nos résultats suggèrent que ce choix peut considérablement influencer les chemins obtenus[89]. De plus, il a été montré que les méthodes basées sur les vecteurs binaires peuvent conduire à des résultats contre-intuitifs lorsque des similarités relativement faibles sont considérées[87], ce qui peut être le cas lors des comparaisons des substrats et produits d'une réactions. Les méthodes d'*atom mapping* sont quant a elles usuellement basées sur des alignements de graphes moléculaires (dont les nœuds correspondent aux atomes et les arêtes aux liaisons chimiques), et la comparaison de ces algorithmes a révélé une adéquation avec des alignements manuels supérieure à 91%, ce qui suggère une certaine maturité de ces techniques[217]. Par conséquent, notre attention s'est portée sur ces méthodes pour la pondération des transitions, bien qu'elles impliquent un temps de calcul bien plus long. Ces calculs peuvent néanmoins être effectués lors d'un pré-traitement du réseau, et par conséquent être réalisés une seule fois par réseau. La comparaison de ces méthodes par cette étude suggère néanmoins que ces méthodes varient en fonction du type de réactions considéré, et qu'en moyenne, leur adéquation avec les alignements manuels est plus importante pour les réactions catalysées par des oxidoreductases que pour celles catalysées par des ligases[217].

L'article présenté constitue également la première comparaison fondamentale des différentes méthodes basées sur des pondérations. Bien qu'il soit difficile de définir précisément la notion de pertinence de chemin, cet article dégage néanmoins plusieurs comportements, dont certains peuvent être considérés comme peu pertinents suivant le contexte. Par exemple, les méthodes de pondération basées sur le transfert d'atome[28][27][12][41][115][164][213] ou la similarité chimique[219][183] vont conduire à des chemins qui tendent à « suivre » les coenzymes de grande taille lorsqu'un complexe impliquant l'attachement d'une petite molécule à un cofacteur est atteint [1]. Ces cofacteurs étant généralement impliqués dans un nombre important de réactions, les pondérations basées sur le degré[63][64] conduiront

1. Il a cependant été suggéré par Pertusi *et al.* que la représentation des cofacteurs sous forme de *feature* unique dans les *fingerprints* pourrait pallier les écarts de similarités induit par l'attachement de ces cofacteurs[210]

à leur évitement si un autre chemin est possible. En revanche, ces méthodes ne prennent pas en compte les cas où des *hubs* sont impliqués en tant que composé principal dans des réactions, ainsi que les composés auxiliaires impliqués dans un faible nombre de réactions. Il n'est généralement pas possible de savoir *a priori* si les mécanismes en lien avec une liste de métabolites d'intérêt impliquent les cas présentés. Ainsi, tout comme il n'existe pas de méthode capable d'observer l'ensemble du métabolome, il n'existe pas de distance générale qui représente de manière adéquate l'ensemble des relations entre métabolites et une analyse combinant les différentes approches peut être nécessaire.

Les tags RPAIR[156], proposés par KEGG[141], constituent cependant une alternative intéressante puisque les transitions auxiliaires y sont définies manuellement au cas par cas pour chaque réaction, sur la base des motifs d'échanges d'atomes également validés manuellement. Bien que limité aux réseaux issus de la base de données KEGG, l'utilisation de ces données permet de filtrer un réseau métabolique pour créer un réseau de transitions principales[81]. Plusieurs outils, dont de nombreux sont mis à disposition par la plateforme NeAT[45], utilisent ces réseaux et permettent entre autres la recherche de chemin pertinent. Cependant, l'usage des approches basées sur ces données a été limité, voir rendu obsolète, par l'abandon de la base de données RPAIR par KEGG en octobre 2016. L'attribution de tags et les atom-mapping pour chaque transition ne sont donc plus disponibles, et l'information fournie par KEGG est désormais limitée aux seules transitions principales et leurs motifs d'échanges atomiques communs, utilisés pour la définition de classes de réactions dans la base de données RCLASS. Pour l'instant, peu d'outils pouvant se contenter de ces informations ont été adaptés pour prendre en compte ce changement d'accès aux données. Outre le challenge du maintien sur le long terme des outils issus de la bioinformatique, cet évènement met en évidence la nécessité d'utiliser de données « *open* » pour garantir la pérennité des méthodes proposées par la communauté.

Chapitre 5

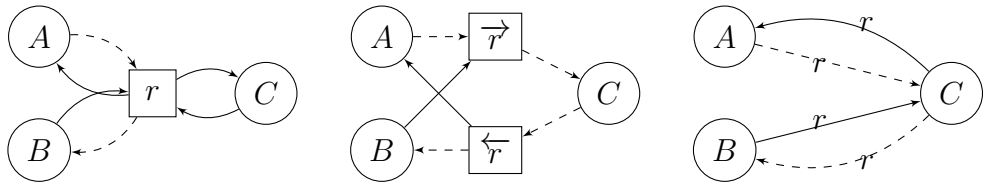
Gestion des réactions réversibles dans les graphes métaboliques

5.1 Introduction

Certaines réactions biochimiques peuvent, sous certaines conditions, s'opérer dans un sens ou dans un autre. Les métabolites impliqués devenant dans un cas substrats et dans l'autre produits. En revanche, les conditions qui vont déterminer l'orientation dans un sens ou l'autre ne peuvent être rencontrées simultanément[261]. En termes de graphe, cela signifie que des ensembles d'arcs vont être mutuellement exclusifs[83]. Ainsi, pour une réaction donnée, le graphe peut être dans 2 états différents, l'un possédant uniquement les arcs de l'une des orientations, l'autre les arcs d'orientation opposée. La proportion de réactions réversibles dans les réseaux métaboliques étant relativement importante (50,2% de réaction réversibles pour Recon 2[256]), le nombre de combinaisons d'états possibles devient prohibitif pour la genèse de l'ensemble des graphes valides. De plus, il est difficile de définir *a priori* une direction réactionnelle sur la base de critères biochimiques. Les ensembles mutuellement exclusifs cohabitent donc dans un même réseau, ce qui va potentiellement conduire à des distances sous-estimées entre les métabolites lorsque le chemin sous-jacent va emprunter ces arcs incompatibles (exemple Figure 5.1). Cette limite, notamment mise en évidence

par Van Helden et al.[261], implique dès lors de considérer cette contrainte lors de la recherche de chemins métaboliques. En effet, à titre d'exemple, nous avons calculé les plus courts chemins entre toutes les paires de métabolites du réseau métabolique d'*Escherichia coli* de la base KEGG. Sur les 428 400 chemins obtenus, 152 633, soit plus de 35%, présentent des transitions incompatibles.

Les chemins sont généralement construits de manière itérative par ajouts successifs d'arcs. Il semble donc aisé d'éviter l'ajout d'un arc si son homologue incompatible est déjà présent dans le « protochemin » en construction. C'est d'ailleurs l'approche qui avait été initialement proposée[83][82][261][63][64] pour gérer le problème des directions mutuellement exclusives. En revanche, garantir que le nombre d'arcs ou le poids du chemin est minimal en maintenant cette contrainte n'est pas trivial, ce qui sera démontré dans cette section. L'article suivant illustre cette complexité par un exemple sur lequel l'approche précédente échoue à obtenir un chemin de poids minimal, ce qui remet en cause son usage pour la définition d'une distance, qui est basée sur ce critère de minimalité. Théoriquement, la méthode peut également échouer à trouver un chemin valide (qui remplit la contrainte d'exclusion mutuelle) alors qu'il en existe un.



(a) Graphe biparti **(b) Graphe biparti modifié** **(c) Graphe des composés**

Figure 5.1 – Problème d'exclusion mutuelle des directions de réactions réversibles. Avec r la réaction réversible d'équation $A + B \leftrightarrows C$. Des chemins invalides allant de A à B sont représentés par des lignes pointillées

L'article suivant va mettre en évidence les différences de définition du problème selon le type de graphe métabolique choisi. Le problème des réactions réversibles s'apparente à des problèmes théoriques de coloration de graphe, en particulier ceux de recherche de chemins avec des contraintes sur les couleurs des arcs[2][176]. En

effet, les couleurs sont un moyen de représenter tout attribut discret, dans notre cas l'attribution d'un sens réactionnel pour une réaction donnée. Ainsi les arcs des graphes des composés se voient attribuer une couleur correspondant à leur réaction d'origine, et un chemin où il n'existe pas de répétition d'une même couleur satisfait la contrainte d'exclusion mutuelle. Dans le cas du graphe biparti, la recherche de chemin simple empêche l'emploi d'une même réaction de manière répétée. En revanche, afin d'éviter le passage de substrats à substrats ou de produits à produits, les arcs peuvent arborer une couleur correspondant au couple réaction et côté de l'équation. La recherche d'un chemin valide revient alors une fois de plus à la recherche d'un chemin sans répétition de couleurs (exemple Figure 5.2).

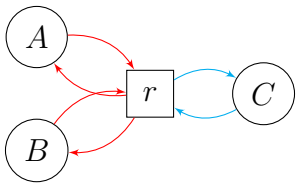


Figure 5.2 – Exemple d'utilisation de couleurs pour la gestion des réactions réversibles dans le graphe biparti. Avec r la réaction réversible d'équation $A + B \leftrightarrows C$. Chaque couleur correspond à un couple réaction-côté de l'équation. Un chemin qui n'emprunte pas plus d'une fois une même couleur respecte la condition d'exclusion mutuelle des directions opposées d'une réaction réversible.

Il a déjà été montré que trouver un plus court chemin dans de telles conditions constitue un problème NP-complet dans le cas général[84][254], qui par conséquent ne peut être résolu en pratique sur des réseaux d'une taille du même ordre de grandeur que celle des réseaux métaboliques genome-scale. En revanche, ces démonstrations ne prennent pas en compte certaines spécificités des réseaux métaboliques, et la preuve de complexité ne couvre pas nécessairement ces instances. Par exemple, dans les réseaux bipartis, les arcs de mêmes couleurs sont toujours incidents à un même noeud, et pour tout arc (a, r) incident à une réaction réversible, il existe un arc (r, a) dans le graphe. L'article suivant va fournir la preuve que le problème est également NP-complet dans le cas des réseaux orientés et dans le

cas particulier des réseaux métaboliques qui ajoutent des contraintes particulières sur l'attribution des couleurs. L'article suivant proposera également un algorithme inspiré de l'algorithme des *k-shortest paths* de Yen[276], capable d'obtenir une solution exacte en un temps acceptable en pratique, à la condition que soit définie une longueur maximale de chemin à considérer.

5.2 Proposition d'un algorithme de recherche de chemins métaboliques valides

Article en préparation

Handling reaction reversibility in metabolic path search

Clément Frainay,[1] Arnaud Mary,[2,3] Marie-France Sagot[2,3*]

[1]Toxalim, University Toulouse, INRA UMR 1331,
Universit de Toulouse 3 Paul Sabatier, F-31027 Toulouse, France
[2]Erable team, INRIA Grenoble Rhne-Alpes,
38330 Montbonnot-Saint-Martin, France
[3]University Lyon 1, CNRS UMR 5558,
F-69622 Villeurbanne, France

*To whom correspondence should be addressed; E-mail: Marie-France.Sagot@inria.fr

Finding relevant paths in a metabolic network is a convenient tool to decipher relationships between metabolites, which can raise new hypotheses regarding complex mechanisms. This potential led to several studies aiming at reaching meaningful paths, primarily focus on the side compounds problem. However, little has been done regarding the proper use of reversible reactions in paths. In this paper, we emphasise the difficulty of such task and propose a new algorithm that can be achieved at a genome-scale level.

Context

The problem of finding a relevant path in metabolic networks is a challenging task which gained a lot of attention during the past decade. Metabolic path extraction from genome-

scale metabolic networks shows great potential for interpreting omics data, predicting novel drug targets, engineering new biosynthesis routes or analysing pathways. The length of such paths has also been very convenient to define distance in metabolic networks, allowing identifying topological characteristics and proposing metabolism evolution theories. However, it has been argued that the lack of biological context in the construction of metabolic paths, thus solely relying on topological features, tend to underestimate distances in the metabolic networks, leading to misguided interpretations of metabolism properties[18].

Starting with the work of Arita in 2002[2], many methods have been proposed to find biologically relevant paths, mainly trying to address the problem of side compounds. Side compounds are ubiquitous metabolites used in many reactions, serving annex purposes such as electron donor or acceptor, phosphate groups donors, etc. Facing the huge number of possible paths between two compounds in a metabolic graph, the first attempts to refine this set of paths was to focus on the shortest ones, as many biological processes tend to be parsimonious. Shortest path search applied to metabolic network often raises paths that use side compounds as intermediaries, leading to irrelevant results[13][25]. As the definition of an a priori list of side compounds can be dubious, many proposed methods used the lightest paths, aiming to find a path minimising their node or edges' weight instead of length. Several weighting schemes were proposed, using topological features such as degree[6][7], or biochemical criteria such as chemical similarity[23][20] or atomic continuity[4][17][3][2][5][14][22].

A second critical point in the surge of relevant paths in metabolic networks is the use of reversible reactions. The substrate and products of many reactions can be interchanged given specific conditions such as substrate/product concentrations. This leads to the modelling of both directions for many reactions in the metabolic networks. However,

direct and reverse directions of a reaction should be considered as mutually exclusive since both conditions required for each direction cannot be met at the same time[11]. This problem was first raised by Van Helden et al. in 2002[25], and very few attention has been given to that constraint and on how to change algorithms to fit it.

It has been suggested to "keep track" of the reactions used during path elongation in order to avoid the addition of a reaction already used[11][10][25][6][7]. We will show that this modification does not guarantee the shortness or lightness of the path found in all cases and that the problem of finding a metabolic path without reaction duplicate is far from trivial.

We will then propose a new algorithm that yields to a valid shortest path and can be performed in an acceptable amount of time in practice.

Problem definition

Different ways to model reaction reversibility in metabolic graphs exist, which have an impact on the problem definition. We do not discuss the hypergraph case[8][21][15], as hyperpaths are nonlinear sub-hypergraphs, which is not convenient for defining distance in metabolic networks. Hypergraphs are a way to handle reaction stoichiometry, a third relevance criterion. However, it is mainly used for proper bioproduction scenario search and pathway extraction contrary to path search aiming to emphasise and measure relationships between compounds, regardless of the availability of other required compounds and without the assumption of steady states nor reaction boundaries.

Compound Graph Case

A compound graph is a representation of the metabolic network as a graph. The vertices of the graph correspond to compounds and edges to chemical transitions (1.a). Each edge

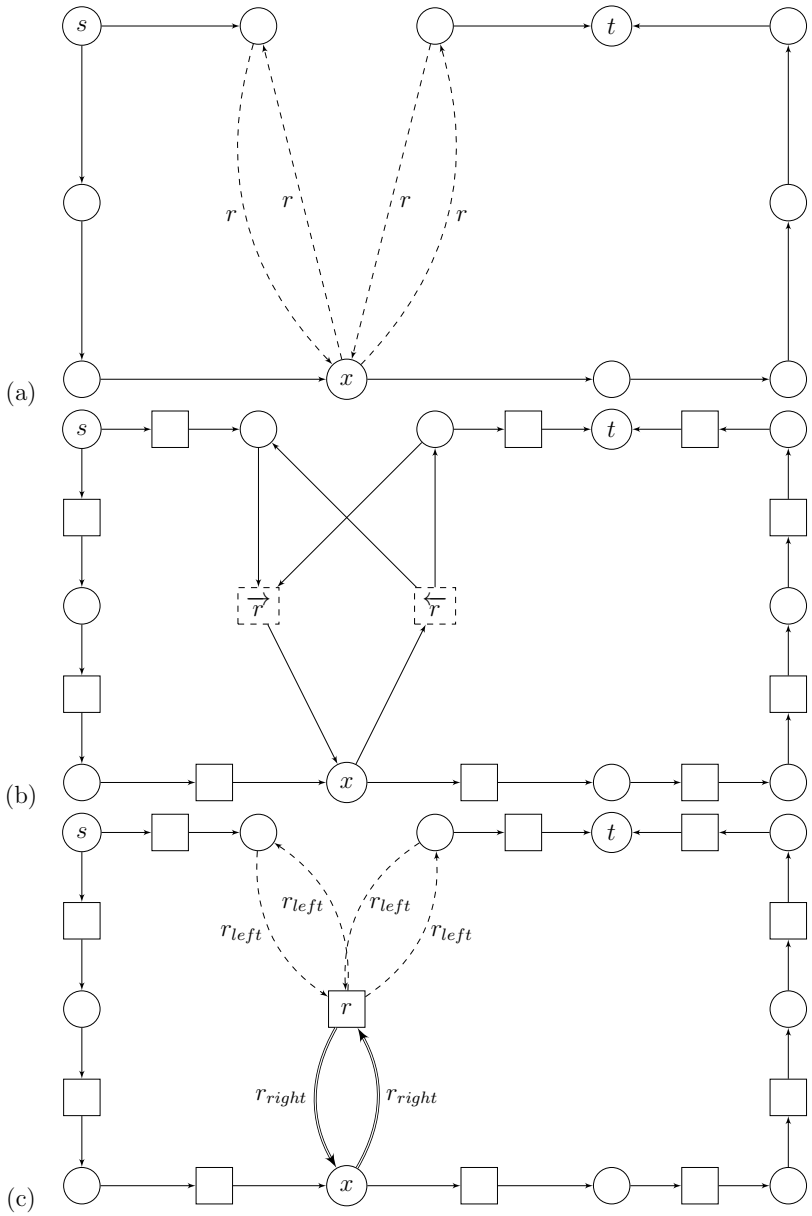


Figure 1: Representation of reversible reactions in metabolic graphs.
Reaction represented as square nodes, compounds represented as circular nodes.
(a) Compound graph with edge label. (b) Bipartite graph with duplicated reactions.
(c) Bipartite graph with edge label.

holds a label which corresponds to the reaction performing the transition. We define a valid path by a list of successive edges starting from a source s and ending at a target t where each corresponding label is represented at most once. A valid shortest path between s and t is the path with the least number of edges among all valid paths between s and t.

Performing path search on a graph with constraints on labels (often referred as colours) is a task that has been assessed by many works from mathematics and computer science[1][19].

The problem of finding a valid path can be defined as follows:

Given a directed multi-graph G, with $V(G)$ its set of n vertices, $E(G)$ its set of m edges, and L a set of labels; a labelling function $l : E(G) \leftarrow L$ that assigns a label in L to every edge, according to the reaction performing the reaction and two vertices, $s \in V(G)$ and $t \in V(G)$:

Find the shortest path P starting in s and ending in t using edges with distinct labels.

In agreement with what has been previously suggested for this problem, namely "keep track of the labels of arcs that were already traversed"[10], we perform a modified version of Dijkstra algorithm with backtracking, used to find the shortest path in a graph with positive edge weights. This modified version store each shortest path, and the addition of a new edge is performed only if its label has not been already added to the path. Since, to our knowledge, no formal definition of a proper algorithm for this task has been published, this algorithm reflects our personal interpretation of what has been previously suggested. This kind of approach guarantees the validity of the obtained path. However, it can fail in finding the shortest one and may also fail to found any solution despite existing ones. For a reversible reaction, the direction that is closest to the source will be added first to a putative shortest path, definitely disallowing the addition of the reverse direction. See Figure 1 as an example.

Let be p_1 the shortest path between the source s and the node x, $x \in V(G)$, containing the label r, $r \in L$.

Let be p_2 the shortest path between the node x and the target t, also containing the label r.

Let be $P = p_1 + p_2$ the shortest path between s and t in G. P is thus invalid since it contains the label r twice.

If the use of a r-labelled edge is disabled after its use in p_1, the obtained path P_2 become $p_1 + p'_2, r \notin p'_2$

Let be p'_1 the shortest path between s and x after the removal of r-labelled edges.

Breadth first algorithm implies that the first visit of a r-labelled edge (and consequently the only allowed) would be the closest to s in P. In the case where $|p'_1| + |p_2| < |p_1| + |p'_2|$, those approaches will fail to retrieve the valid shortest path (See example in 1). □

In fact, the work of Fekete et al. shows that solving the shortest path with pairwise distinct edge labels is an NP-complete problem[12], meaning that solving it becomes intractable for large graphs such as metabolic networks.

Bipartite Graph Case

In the metabolite bipartite graph, both reactions and compounds are represented as nodes in the graph. Compounds can only be linked through reactions, there are no edges directly linking two compounds or two reactions. Path search methods based on DFS or BFS traversal ensure that a node is not encountered twice, preventing to create a metabolic path using the same reaction twice. However, in the reversible case, two substrates can be linked through one reaction, which is still breaking the biochemical rule defined previously.

Two strategies can be used :

- Duplicating reversible reactions in order to depict the two directions as different reactions (1.b). In that case, a valid path becomes a path with no repetitions on node labels instead of edge labels, which intuitively holds the same complexity, and is equivalent to solving the problem on the compound graph.

- Use reaction equation sides as edge label (1.c). In that case, using an edge incident to a substrate of reaction r should prevent using a product of the reaction r. The problem is still finding a path with no edge labels repetition. However, this labelling implies that edges with the same label are always consecutive in a simple path (a node cannot be used twice in a simple path. The term path usually implicitly refers to simple paths, by opposition to walks).

We define a valid path in the bipartite graph by a list of successive edges starting from a source compound s and ending at a target compound t where each reaction node is traversed at most once, and were, for each reaction node r_i, its compound predecessor c_{i-1} and successor c_{i+1} do not belong to the same side of r_i reaction equation.

Given a labelling function $l : E \leftarrow L$ that assigns a label in L to every edge according to the incident reaction node and the side of the equation holding the incident compound node, finding a valid path with distinct labels can be reduced to finding a path without two successive edges with the same label.

Intuitively, we can see that the previous proof regarding the incapacity of classical shortest path search to find a valid path still holds in the bipartite graph. However, finding a path without two successive edges with the same label, also known as the shortest properly coloured path problem, can be solved in polynomial time on a undirected graph[24].

Working on an undirected network would consider every reaction as reversible, which

can be relevant since the direction of a reaction only holds for a particular, not necessarily known, physicochemical context, leading many path search programs to use undirected networks. However, real-life observations show that some reactions have a strongly preferred direction, therefore ignoring reaction directions during path search can lead to very unlikely solutions for biochemical routes extraction.

We will show in the next section that solving the properly coloured path on a directed graph is NP-complete.

Proof of Complexity

Proposition 0.1 *The problem of finding a properly-coloured path in a directed graph (**dPCP**)is NP-complete*

Proof Based on the construction proposed by Stefan Szeider in [24], we show that **3-SAT** reduce polynomially to ***d*PCP**.

$$\textbf{3-SAT} \propto \boldsymbol{d}\textbf{PCP} \tag{1}$$

Let $\varphi = \{C_1, ..., C_n\}$ a collection of clauses. Each clause C_i is a set of 3 distinct literals $\{x_{i,1}, x_{i,2}, x_{i,3}\}$ such as no clause contains both $\overline{x}$ and x.

For each clause C_i, we construct a directed graph G_i with a set of vertices $V(G_i) = \{s_i, v_{i,1}, v_{i,2}, v_{i,3}, t_i\}$ and a set of directed edges $E(G_i) = \cup_{j=1}^{3} \{(s_i, v_{i,j}), (v_{i,j}, t_i)\}$. We denote the graph G as the union of graphs $G_1, ..., G_n$ joined by edges $E^* = \cup_{i=1}^{n-1} \{t_i, s_{i+1}\}$.

Let be $I \subset V(G)$ a set of *incompatible* vertex pairs $\{v_{i,j}, v_{i',j'}\}$ such as $x_{i,j} = \overline{x_{i',j'}}$ and $i \neq i'$ holds for the corresponding literals in φ. Intuitively, finding an oriented path $P(s_1, t_n)$ in G that do not contain a pair of vertices in I allow finding a literal truth assignment that satisfies φ, and for any assignment satisfying φ, there exist a path in G

that do not contain a pair of vertices in I.

In order that any vertex appear in at most one pair in I, we modify G such as for each vertex $v_{i,j}$ contained in $q > 1$ pairs in I, we add new vertices $v_{i,j}^1, ..., v_{i,j}^q$ and create edges $(s_i, v_{i,j}^1)$, $(v_{i,j}^q, t_i)$ and $\cup_{k=1}^{q-1} \left\{(v_{i,j}^k, v_{i,j}^{k+1})\right\}$. $v_{i,j}$ is then removed from G and the q pairs $\{v_{i,j}, w\}$ in I are replaced by $\left\{v_{i,j}^k, w\right\}$ for $k = 1, ..., q$. The resulting graph and pair set are denoted respectively G' and I'.

We duplicate each edge $(v_{i,j}^k, v_{i,j}^{k+1})$ into two edges $\left\{(v_{i,j}^k, w), (w, v_{i,j}^{k+1})\right\}$, adding new dummy nodes w, and define $f(e)$ a function that assigns to an edge $e \in E(G')$ a colour c such as for each pair $\{u, u'\}$ in I', $f(s, u) = f(u', t')$ and $f(s', u') = f(u, t)$, otherwise each edge hold a unique colour.

We then modify G' such as vertices from each pairs in I' are merged into one unique node. The resulting graph is denoted $G*$. We can see that a path in G that do not contain any pair in I correspond to a path in $G*$ without two successive edges with the same colour.

Since our construction can be carried out in polynomial time, (2) holds and ***d*PCP** is NP-complete. □

Proposition 0.2 *The problem of finding a properly coloured path in a bipartite metabolic graph (**mPCP**)is NP-complete*

Proof Based on the previous proof, we show that **3-SAT** reduces polynomially to ***m*PCP**.

$$\textbf{3-SAT} \propto \boldsymbol{m}\textbf{PCP} \tag{2}$$

Metabolic bipartite graph holds the following particularity over directed graphs : For each neighbour n of a reversible reaction vertex r^r, there exist two edges (n, r) and (r, n). Since the n cannot be on both side of the equation reaction, all edges between n and r

should be of the same colour. Since a reaction can hold 2 and only 2 equation sides, the number of colours of all the adjacent edges of a reaction vertex equals 2.

We build a metabolic graph M following the construction of G : Each clause i is represented by a source and a target compounds s_i and t_i. Each literal x_{ij} is represented as a reaction consuming s_i and producing t_i. Successive clauses are connected through a clause reaction node $r_{i,i+1}$ consuming t_i and producing s_{i+1}. As in G, finding a path going from s_1 to t_q without using two incompatible reactions in I allow finding a literal truth assignment that satisfies φ, and for any assignment satisfying φ there exists a path in M that do not contain a pair of vertices in M. Following proof 1, we create a bipartite metabolic graph M', by replacing any reaction node r that is present in n pairs in I ($n > 1$) by a path of n reactions noted $r^1, ..., r^n$. r is then removed from the n pairs $\{r, w\}$ in I and replaced by $\{r^k, w\}$ for $k = 1, ..., n$.

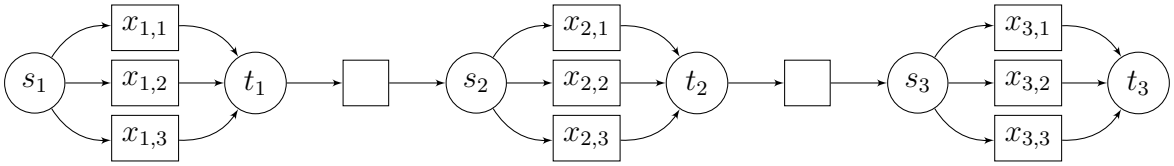


Figure 2: Bipartite metabolic graph M built from 3 clauses

We define a gadget $W_{u,v}$ as the following metabolic bipartite graph : a reversible reaction w, with two left reactant t_u^w and s_u^w and two right reactant t_v^w and s_v^w. Each reactant is connected to w through two edges of opposite directions. Edges connecting w to a left reactant hold a colour α, and edges connecting w to a right reactant hold a colour β. s_u^w and s_v^w are respectively consumed by reactions r_u^w and r_v^w.

For each incompatible reaction pairs $\{r, r'\}, r \in C_i, r' \in C_j, i < j$ we create a gadget $W_{r,r'}$. The outgoing edge of r, noted (r, t), is replaced by an edge (r, t_r^w) and the outgoing edge of r', noted (r', t'), is replaced by an edge $(r', t_{r'}^w)$. t is added as a product of r_r^w and

t' is added as a product of $r^w_{t'}$. The constructed metabolic bipartite graph is noted $G*$.

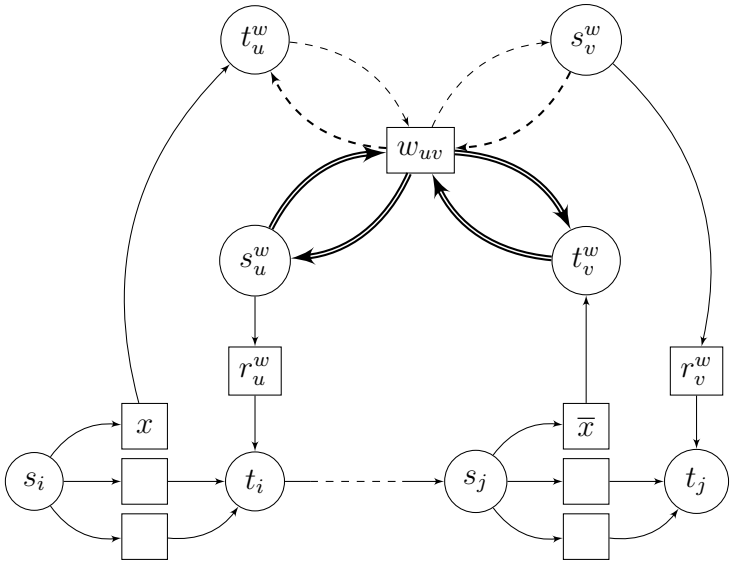


Figure 3: Construction between two incompatible literals, used for the construction of $M*$. Dashed edges and double edges respectively hold the same colour. All other edges are assumed to hold a unique colour.

For a simple path p from s_1 to t_n in $G*$, if $r \in p, (r,t) \in E(G')$ and $(r,r') \in I$ then $r' \notin p$ and $t \in p$ if the path is a properly coloured path.

Therefore, a path in M that do not contain any pair in I correspond to a proper metabolic path in the metabolic bipartite graph $M*$, i.e a simple path without two successive edges with the same colour.

Since our construction can be carried out in polynomial time, (2) holds and ***m*PCP** is NP-complete. □

A practical solution bounded by path length

We propose an efficient algorithm based on Yen's k-shortest paths algorithm[26] to produce meaningful metabolic paths. Instead of enumerating paths in order of length up to

the k^{th}, we enumerate paths until a valid one is found. It means that in the trivial case where there are no reversible reactions in the graph, the computation cost would remain roughly the same as a classic path search. In order to reduce complexity, only paths below a given length are considered, avoiding the enumeration of all possible paths in the graph. This can be done by a slight modification of the Dijkstra algorithm[9], skiping the update of the distance between two nodes when it exceed the maximum length.

The original algorithm by Yen compute the k-shortest paths by creating deviations of shortest one at each node, recomputing the shortest path between spur nodes and the target.

The time complexity of Yen's algorithm depends on the underlying shortest path algorithm. The number of calls to the shortest path algorithm is equal to Kl, where K is the number of paths considered and l the length of those paths. In our case, l is bounded by the given parameter. Furthermore, since the valid shortest path can only be derived from a valid root path from the i-shortest one, we do not perform rewiring from all spur nodes in the path, focusing on the valid portion of the path starting from its source. This modification lower the number of calls to the shortest path as well as the number of candidate paths considered.

The pseudo-code of a basic implementation is provided in Algorithm 1, using two additional functions: $BoundedShortestPath$ referring to the modification of Dijkstra algorithm suggested previously, and $MaxValidRoot$ that returns the longest subpath of a given path starting from the same source and with each edge label present at most once.

Discussion

We proposed a practical solution for valid path search, circumventing the complexity of the original problem by bounded it by a maximum length. Path length is generally

Algorithm 1 Calculate the shortest valid metabolic path

Input:

G, the compound graph
s, the source node
t, the target node
k, the maximun path length

Output:

The shortest path of length $< k$ with each reaction used at most once.

```
function GetMetabolicPath(G, s, t, k)
    p = BoundedShortestPath(G, s, t, k)
    unseen[ ] ← ∅
    seen[ ] ← ∅
    insert(p, unseen)
    while unseen ≠ ∅ do
        sort(unseen)
        p ← pop(unseen)
        if maxValidRoot(p) == p then return p
        for i from 0 to |maxValidRoot(p)| do
            root = (p_j)_{j≤i}
            G' ← G
            for each p' ∈ seen do
                if root == (p'_j)_{j≤i} then
                    e ← (p'_i, p'_{i+1})
                    G' ← G' \ {e}
            G' ← G' \ {p_j | j < i}
            spurPath ← BoundedShortestPath(G', p_i, t, k − i)
            if spurPath ≠ ∅ then
                candidate ← (root + spurpath)
                insert(candidate, unseen)
        seen.add(p)
    return ∅
```

seen as a good relevance criterion in metabolic network : since each node transition is surrounded by a lack of certainty regarding enzymes and co-substrates availability, as well as other biochemical consideration usually not modelled on this scale, long paths are often regarded as dubious. Furthermore, repositories of manually curated pathways can offer a good estimation of the typical length of relevant paths through their topological properties.

The tractability of the proposed algorithm in practical case is also ensured by the properties of metabolic networks. Metabolic networks are sparse graph, with a power-law-*like* distribution of degree. The few nodes with high degree are usually considered as side compound and therefore avoided using the lightest path or even removed from the graph. One can also filter the edges of the compound graph using extrinsic data such as RPAIRs[16], drastically lowering the average degree in the network, and thus the maximum number of paths between two nodes. It should also be noted that metabolic network usually contains many irreversible reactions, consequently, in practice the shortest path between two nodes would be valid in numerous cases.

Conclusion

We show that, in contrary to what is assumed in the literature, the problem of mutual exclusion of reversible reactions' directions in metabolic path search is far from trivial. We prove that the search for such path constitutes an NP-complete problem. Classical BFS approach, even combined with backtracking as it has been suggested, cannot guarantee an optimal solution, and can even fail to found a valid path while one exists. However, finding an optimal and valid path on the undirected bipartite graph (considering every reaction as reversible) is feasible in a practical way using matching algorithms. However, it might yield irrelevant results due to the use of reaction directions that would be very

unlikely, such as one consuming carbon dioxide in a gaseous state, rarely available in the cell. Furthermore, in contrarily to compound graph, the bipartite graph does not allow mapping compounds-pair attributes, such as chemical similarity or RPAIRs tags. Those attributes have been used for computing weighting schemes in order to ensure better meaningfulness of metabolic path, mainly by avoiding side compounds. This shows the difficulty to combine both consistency regarding reversible reaction and side compounds avoiding using the lightest path.

In order to produce valid metabolic paths in a directed graph, including compound graph, one can enumerate paths until a valid one is found. We proposed an algorithm allowing reducing the complexity of such approach by limiting the length of considered paths.

References

[1] A. Abouelaoualim, K.Ch. Das, L. Faria, Y. Manoussakis, C. Martinhon, and R. Saad. Paths and trails in edge-colored graphs. *Theoretical Computer Science*, 409(3):497–510, 2008.

[2] Masanori Arita. Metabolic reconstruction using shortest paths. *Simulation Practice and Theory*, 8(1):109–125, 2000.

[3] Torsten Blum and Oliver Kohlbacher. MetaRoute: Fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics*, 24(18):2108–2109, 2008.

[4] Torsten Blum and Oliver Kohlbacher. Using Atom Mapping Rules for an Improved Detection of Relevant Routes in Weighted Metabolic Networks. *Journal of Computational Biology*, 15(6):565–576, jan 2008.

[5] F. Boyer and A. Viari. Ab initio reconstruction of metabolic pathways. *Bioinformatics*, 19(Suppl 2):ii26–ii34, oct 2003.

[6] Didier Croes, Fabian Couche, Shoshana J Wodak, and Jacques van Helden. Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic acids research*, 33(Web Server issue):W326—-W330, jul 2005.

[7] Didier Croes, Fabian Couche, Shoshana J. Wodak, and Jacques Van Helden. Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356(1):222–236, 2006.

[8] Yves Deville, David Gilbert, Jacques van Helden, and Shoshana J Wodak. An overview of data models for the analysis of biochemical pathways. *Briefings in bioinformatics*, 4(3):246–259, 2003.

[9] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1):269–271, 1959.

[10] Karoline Faust, Didier Croes, and Jacques van Helden. Prediction of metabolic pathways from genome-scale metabolic networks. *Bio Systems*, 105(2):109–21, aug 2011.

[11] Karoline Faust, Pierre Dupont, Jérôme Callut, and Jacques van Helden. Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics (Oxford, England)*, 26(9):1211–8, may 2010.

[12] Sandor Fekete, Tom Kamphans, and Michael Stelzer. Shortest Paths with Pairwise-Distinct Edge Labels: Finding Biochemical Pathways in Metabolic Networks. *arXiv preprint arXiv:1012.5024*, page 9, 2010.

[13] Clément Frainay and Fabien Jourdan. Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Briefings in Bioinformatics*, 18(1):43–56, 2016.

[14] Allison P. Heath, George N. Bennett, and Lydia E. Kavraki. Finding metabolic pathways using atom tracking. *Bioinformatics*, 26(12):1548–1555, jun 2010.

[15] Steffen Klamt, Utz-Uwe Haus, and Fabian Theis. Hypergraphs and Cellular Networks. *PLoS Computational Biology*, 5(5):e1000385, 2009.

[16] M Kotera, M Hattori, MA Oh, and R Yamamoto. RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Informatics*, 2004.

[17] Mario Latendresse, Markus Krummenacker, and Peter D Karp. Optimal metabolic route search based on atom mappings. *Bioinformatics (Oxford, England)*, 30(14):2043–50, jul 2014.

[18] Gipsi Lima-Mendez and Jacques van Helden. The powerful law of the power law and other myths in network biology. *Molecular bioSystems*, 5(12):1482–1493, 2009.

[19] Adria Lyra and Carlos A Martinhon. On paths, trails and closed trails in edge-colored graphs. *Discrete Mathematics and Theoretical Computer Science*, 14(2):57–74, 2012.

[20] D.C. McShan, S. Rao, and I. Shah. PathMiner: predicting metabolic pathways by heuristic search. *Bioinformatics*, 19(13):1692–1698, 2003.

[21] Nicole Pearcy, Jonathan J Crofts, and Nadia Chuzhanova. Hypergraph Models of Metabolism. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering*, 8(8):812–816, 2014.

[22] Esa Pitkänen, Paula Jouhten, and Juho Rousu. Inferring branching pathways in genome-scale metabolic networks. *BMC systems biology*, 3(1):103, jan 2009.

[23] S A Rahman, P Advani, R Schunk, R Schrader, and Dietmar Schomburg. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics (Oxford, England)*, 21(7):1189–93, apr 2005.

[24] Stefan Szeider. Finding paths in graphs avoiding forbidden transitions. *Discrete Applied Mathematics*, 126(2-3):261–273, 2003.

[25] J van Helden, L Wernisch, D Gilbert, and S J Wodak. *Graph-based analysis of metabolic networks.* Number 38. Springer Berlin Heidelberg, 2002.

[26] Jin Y Yen. Finding the k shortest loopless paths in a network. *management Science*, 17(11):712–716, 1971.

5.3 Discussion

L'article proposé suggère que la prise en compte de l'exclusion mutuelle des directions opposées est un problème complexe. Les solutions précédemment proposées[83][82][261][63][64], bien qu'elles garantissent le respect de cette contrainte, ne peuvent assurer la minimalité du chemin obtenu et peuvent également échouer à trouver certains chemins valides, ce qui peut conduire à faussement considérer des métabolites comme déconnectés. Il est possible d'obtenir une solution exacte dans le cas du graphe biparti non orienté, ce qui revient à considérer toutes les réactions comme réversibles. Bien que théoriquement possibles, en conditions physiologiques de nombreuses réactions n'ont qu'une direction possible, ou du moins présentent des directions peu vraisemblables. Il est également à noter que cette solution implique l'utilisation du graphe biparti, car les arcs incompatibles y sont toujours consécutifs contrairement aux autres graphes métaboliques, ce qui réduit la complexité du problème. L'utilisation du biparti est incompatible avec certaines méthodes proposées précédemment pour gérer le problème des composés auxiliaires, qui exploite des attributs définis pour des couples de composés (similarité chimique[219][183], atom-mapping[28][164][27][12][41][115][213] ou tag RPAIR[81]). Cette limite vient corroborer l'hypothèse précédente selon laquelle il n'existe pas de méthode idéale pour prendre en compte toutes les contraintes biochimiques dans la construction de chemin dans les réseaux métaboliques. Bien que le problème s'avère trop complexe pour être résolu en pratique sur le graphe des composés, il est cependant possible de construire des heuristiques pour réaliser des choix « raisonnables » face à des problèmes de ce type. Une stratégie répandue est de limiter l'espace de recherche. La partie précédente a exposé les limites des reconstructions et de l'utilisation des réseaux métaboliques. Lors de la construction d'un chemin métabolique, chaque étape est entourée d'une zone d'ombre concernant la faisabilité de la transition impliquée (présence des enzymes, disponibilité des co-substrats, *etc.*), voire son existence même. Plus un chemin va être long, plus il va être difficile d'y accorder du crédit du fait de l'accumulation d'événements extrinsèques qui

peuvent remettre en cause chacune de ses étapes. Par conséquent, la longueur des chemins peut être un paramètre adéquat pour borner l'espace de recherche, en fixant une longueur maximale au-delà de laquelle les chemins ne seront plus considérés. Par ailleurs, cette approche avait notamment été suggérée par Croes et al.[63], dans leurs travaux à l'origine des premiers algorithmes prenant en compte des contraintes sur les réactions réversibles.

Chapitre 6

Discussion sur la pertinence des chemins métaboliques

Les chemins métaboliques permettent de définir un lien entre deux métabolites, et peuvent servir de norme pour définir une distance mesurable entre ces derniers au travers de leurs longueurs. En revanche, l'exploitation de leur composition pour proposer des scénarios d'enchaînements d'évènements expliquant les corrélations qui peuvent exister entre les variations de concentration des métabolites reste très limitée.

6.1 Limite des plus courts chemins

> *"What is found in biology is mechanisms, mechanisms built with chemical components and that are often modified by other, later, mechanisms added to the earlier ones. While Occam's razor is a useful tool in the physical sciences, it can be a very dangerous implement in biology."*
>
> Francis Crick (1988) What Mad Pursuit

La première limite est que ces chemins répondent à un critère d'optimalité, en termes de longueur ou de poids, qui ne reflète pas nécessairement la réalité des processus biologiques[214]. Pour reprendre les illustrations de Borgatti sur

la circulation dans les réseaux[37], il est peu vraisemblable de considérer que les échanges de matière dans les réseaux métaboliques suivent la même logique que les livraisons de colis, ou tout du moins pas pour l'ensemble des couples de métabolites. S'il est raisonnable de considérer que la pression de sélection naturelle tend à rendre le métabolisme efficace d'un point de vue énergétique, et, sachant que de nombreuses réactions métaboliques ont un coût énergétique non négligeable, l'hypothèse de parcimonie en termes de nombre de réactions pour modéliser ces processus peut être une bonne approche. En revanche, elle doit être raisonnée à l'échelle de l'ensemble du réseau ou de ses modules, et non à travers le prisme d'un seul couple de métabolites. Ainsi, la comparaison des observations de transitions métaboliques par marquage isotopique et du potentiel métabolique inféré à partir du réseau révèle que certaines voies, comme le cycle de Krebs, ne sont pas minimales en termes de transitions[201]. En effet, il existe plusieurs « raccourcis » pour rejoindre des métabolites de cette voie. Cependant, les composés intermédiaires produits au cours de ce processus peuvent être réutilisés ou détournés vers d'autres processus, ce qui pourrait tendre à rendre l'ensemble plus efficace en termes de coût énergétique global, plus résilient face aux perturbations, ou encore d'enrichir l'éventail des composés qui peuvent être synthétisés par l'organisme. De nombreux travaux ont tenté de théoriser l'évolution du métabolisme[212][131], qui ne seront pas détaillés ici, cette sous-section n'ayant que pour but de souligner le fait que les plus courts chemins ne représentent pas nécessairement les enchaînements d'événements qui s'opèrent en conditions réelles. Face à cette limite, plusieurs méthodes ont opté pour l'usage des k-plus courts chemins[76] afin de ne pas se limiter au seul plus court. Il est également à noter que le plus court chemin n'est pas nécessairement unique, et se restreindre à un seul d'entre eux peut conduire à l'omission de chemins pertinents.

6.2 Limite topologique des chemins

La linéarité des chemins, bien que commode pour la définition d'une distance, ne représente pas nécessairement la complexité des processus métaboliques qui

peuvent être mis en jeu lors d'une perturbation. En effet, les scénarios pertinents ne sont pas restreints à cette propriété topologique et peuvent impliquer des cycles ou des embranchements, ce qui est fréquemment observé dans les voies métaboliques issues de différentes bases de données. L'une des principales limites inhérentes aux chemins est qu'ils sont définis pour des couples de métabolites uniquement. Les profils métaboliques obtenus en métabolomique révèlent généralement des ensembles plus larges de métabolites discriminant deux conditions. Raisonner sur de telles listes en considérant la somme des interactions entre chaque paire de ses constituants, indépendamment les unes des autres, peut conduire à l'omission de scénarios qui peuvent être plus parcimonieux. Face à cette limite, il a été proposé de considérer des sous-réseaux de taille minimale, ou optimaux vis à vis d'un critère donné, et qui soit définis à partir de l'ensemble des métabolites d'intérêt considérés simultanément[3][186][10]. Ces méthodes répondent à une autre problématique que celle de la mesure de distance, et ne seront pas approfondies dans cette partie.

6.3 Disponibilités enzymatiques

Il est également à souligner que les réseaux métaboliques représentent le potentiel métabolique complet d'un organisme, qui n'est pas accessible dans sa totalité en conditions physiologiques. De nombreuses réactions peuvent ne pas être disponibles, dues entre autres aux nombreuses régulations qui s'opèrent en particulier aux niveaux transcriptionnel et traductionnel. L'intégration de tous ces niveaux de régulation permettrait une modélisation fidèle du métabolisme, et la poursuite de cette approche holistique constitue un des principaux challenges de la biologie des systèmes. Difficiles à mettre en oeuvre du fait des limites inhérentes à chacun de ces modèles et à la difficulté d'acquisition simultanée de ces données, plusieurs approches vont néanmoins dans ce sens.

6.4 Disponibilités des co-substrats

Outre la disponibilité des enzymes, un autre facteur nécessaire à la réalisation d'une réaction est la disponibilité simultanée de l'ensemble des substrats. Les méthodes d'extractions de chemins métaboliques présentées ne considèrent pas les autres co-substrats dans la réalisation d'une transition entre un substrat et un produit. Certaines méthodes ont été proposées pour intégrer ces informations dans la recherche de chemins[190]. Pour considérer la disponibilité des co-substrats, ces méthodes tiennent plus de l'extraction de sous-réseau « faisable » que de chemin. Il est possible par exemple de définir les **scope compounds**[58][112] : les métabolites virtuellement productibles à partir de liste de métabolites et de composés auxiliaires. Pitkanen *et al.* ont également proposé une approche similaire sans être explicitement basée sur les *scope compounds*[214] : un substrat est disponible que si au moins l'une des réactions qui le produit est disponible, et une réaction est disponible si l'ensemble de ses substrats sont disponibles. Dans leur article, Pitkänen rebaptise alors le réseau bipartie en réseau « AND/OR ». La disponibilité d'un nœud y est définie par sa visite lors d'un **parcours de graphe** (exploration des nœuds de proche en proche) intégrant cette contrainte. Le sous-réseau ainsi obtenu constitue une proposition de « voie métabolique » représentant les relations entre des métabolites d'intérêt. Il est à noter que l'obtention d'un tel réseau qui soit de taille minimale constitue également un problème NP-complet.

En revanche, ces méthodes ne prennent pas en compte la stœchiométrie des réactions[211][215]. La prise en compte de la stœchiométrie conduit à l'utilisation d'autres types de modélisations qui offrent la possibilité d'intégrer des informations quantitatives. Là où la théorie des graphes se focalise sur les interactions, la modélisation du métabolisme par une représentation sous forme de matrice de stœchiométrie, couplée à des approches d'optimisation linéaire, offre la possibilité d'une approche quantitative.

6.4.1 Analyse sous contrainte : une approche alternative

Ces approches permettent de réduire l'espace des flux possibles par l'application de différentes contraintes[168]. Il est ainsi possible d'inférer si un métabolite peut être produit étant données certaines contraintes[221], à partir de la matrice de stœchiométrie. De manière générale, ces méthodes se basent sur l'hypothèse que le système est à l'état d'équilibre, c.-à-d. qu'il ne peut y avoir accumulation d'un composé, afin de réduire l'espace des distributions de flux possibles. Cependant, par définition les conditions expérimentales induisant une perturbation de l'état non physiologique peuvent conduire à des déséquilibres. Il convient alors de définir des points d'entrées et de sortie dans le système permettant de garantir l'état d'équilibre, intégrant des métabolites « externes » non soumis à la contrainte d'équilibre. Ces points d'entrées et de sortie sont fréquemment mentionnés sous le terme de réactions d'échange. Ces réactions d'échange permettent également de modéliser la production et la consommation de métabolites par le système, même en l'absence de perturbations. Leurs ajouts se font à la discrétion du modélisateur, et parfois à partir de données d'exométabolome, la fraction extracellulaire du métabolome, combinées à des informations sur les transporteurs membranaires. D'autres contraintes peuvent ensuite s'ajouter, comme la production ou la consommation d'un certain métabolite par exemple.

Bien que ces méthodes permettent d'écarter des scénarios prédits comme infaisables étant données les conditions établies au préalable, elles ne permettent pas la prédiction du processus dont les effets ont été observés, offrant plutôt un espace de solutions. Il est cependant possible de trouver une solution optimale au regard d'un critère donné afin de réduire l'ensemble des solutions de manière plus stringente : la FBA (*flux balance analysis*)[168][110][204]. Cette approche est particulièrement utilisée en bio-ingénierie pour proposer des voies métaboliques permettant la production optimale d'une molécule d'intérêt, tout en minimisant par exemple la production de métabolites secondaires toxiques. Dans le cadre d'applications plus fondamentales sur le comportement d'organismes en réponse à des perturbations, c'est la production de biomasse qui est fréquemment utilisée

comme critère à optimiser chez les procaryotes[34]. En revanche, il est beaucoup plus difficile de définir *a priori* une fonction à optimiser au niveau des fonctions cellulaires d'un tissu par exemple. De plus, dans le cadre de ce type d'application, les limites liées à l'hypothèse de parcimonie sont également à prendre en compte[34][228]. Le postulat de ces méthodes peut être formulé de la manière suivante : l'organisme considéré a, sous l'influence d'une pression de sélection, évolué pour arborer un comportement optimal vis-à-vis d'un processus biologique donné, qui est connu et représentable de manière mathématique. Des méthodes ont cependant été proposées pour contourner le besoin de fonction objective, en considérant des distributions de flux par échantillonnages aléatoires dans l'espace de solutions obtenu[234][218].

Les approches basées sur les matrices de stoechiométrie et sur la théorie des graphes ont souvent été comparées, conduisant à de vifs débats quant à leur utilité pour l'inférence de voies métaboliques[215][66][80][67]. Les méthodes de modélisations par contraintes permettent une approche quantitative qui s'avère très appropriée pour étudier la faisabilité d'un scénario[168]. Elles nécessitent en revanche une connaissance plus poussée du système étudié, en y définissant notamment ces bornes (métabolites internes et externes). Elles présentent également un coût calculatoire important et ont par conséquent été principalement appliquée à des sous-réseaux, bien que réalisables sur des réseaux génomes-scale. Il est également à noter que les méthodes d'optimisation peuvent conduire à l'obtention de très nombreux scénarios optimaux ou suboptimaux, rendant difficile leur interprétation.

Pour ce qui est des limites présentées dans cette partie, les méthodes de flux sont souvent considérées comme non affectées par les problématiques des réactions réversibles, car elles pourraient permettre de définir un sens pour chacune d'entre elles. Pour être exact, ces réactions sont tout de même considérées comme réversibles, avec leurs deux directions pouvant se dérouler simultanément, mais il est possible de définir que le flux net est positif pour l'une d'entre elles. Pour la problématique des composés auxiliaires, en considérant la faisabilité des réactions

et donc la disponibilité de l'ensemble de leurs substrats, elles ne conduisent pas aux raccourcis aberrants observables avec certaines méthodes topologiques.

Les approches basées sur la théorie des graphes conduisent à des scénarios dont la vraisemblance reste très limitée du fait d'une importante simplification du système étudié. En revanche, elles permettent de formuler des hypothèses dans le cas général, car elles ne nécessitent la définition de peu voire d'aucun paramètre. Elles ont généralement un coup calculatoire faible et sont donc tout à fait adaptées à leur application sur des réseaux à l'échelle du génome.

6.4.2 Applicabilité du modèle choisi

> *"Use the right level of description to catch the phenomena of interest. Don't model bulldozers with quarks"*
>
> Leo Kadanoff

Les approches basées sur les chemins sont purement qualitatives, et n'ont pas vocation à prédire la production nette d'un métabolite. De nombreuses autres méthodes de modélisation du métabolisme, pas nécessairement basées sur la théorie des graphes, existent pour répondre à ces questions. On peut par exemple citer les réseaux de Petri, les réseaux bayésiens ou les systèmes d'équations différentielles. Cette section a pour but de clarifier le domaine d'application du modèle employé, et non de dresser une liste exhaustive des avantages et inconvénients des différents modèles existants, ni de les comparer aux méthodes de recherches de chemins. Cependant, un intérêt particulier a été apporté aux approches *constraint based*[168], du fait de leur popularité au sein de la communauté et de l'apparente concurrence qui les opposerait selon certains auteurs[66].

Le choix d'un modèle repose sur un équilibre entre sa simplicité et sa représentativité. Les approches topologiques pour l'interprétation de phénomènes biologiques appartiennent à la frange « simple » du spectre des modèles disponibles. De ce fait, des limitations leur sont imputables dues à l'omission de nombreuses contraintes biologiques. En revanche, les modèles simples offrent l'avantage de nécessiter moins de paramètres et de données expérimentales pour leur construc-

tion. Cette caractéristique peut les rendre plus appropriés que des modèles plus étoffés, mais dont l'estimation des paramètres est ardue et les données d'entrées incertaines.

Des exemples peuvent illustrer ces différences d'applicabilité :

— L'étude d'une voie métabolique spécifique, bien décrite dans la littérature, pour interpréter des résultats quantitatifs ciblés, obtenus sur une lignée d'organismes procaryotes modèle cultivée en conditions contrôlées, avec un milieu de culture de composition connue. Ce type de problèmes est adapté à une modélisation fine des processus mis en jeu, en utilisant des méthodes de flux ou même des modèles cinétiques basés sur des équations différentielles.

— L'étude du métabolisme à l'échelle du génome, pour interpréter des résultats obtenus en métabolomique à partir de prélèvement de biofluides conduits sur des patients présentant des différences d'âges, de sexe, d'habitude alimentaire et de nombreuses autres sources de variabilités.

Le dernier cas semblerait plus difficile à modéliser, étant donné les limites inhérentes à la métabolomique, la variabilité interindividuelle ainsi que l'observation indirecte du processus au travers de la fraction excrétée du métabolome. Ces contraintes sont symptomatiques de nombreuses expériences de métabolomique conduites chez l'homme. Le reste de la thèse se focalisera sur les données générées par ce type d'expérimentations.

6.5 Dépendance par rapport à la qualité des données

Comme mentionné précédemment, les données de métabolomique sont caractérisées par leur incomplétude par rapport au métabolome réel[11].

L'une des principales causes de cette incomplétude est l'absence de spectres standards pour réaliser une identification appropriée des composés. Les composés issus des banques de spectres publiques HMDB[270], MassBank[121], ReSpect[232] et GNPS[264] ont été extraits et mappés sur plusieurs réseaux métaboliques d'or-

ganismes modèles (Figure 6.1). Cette étude révèle qu'une forte proportion de métabolites dans les réseaux ne peuvent être identifiés de façon certaine à partir des banques de spectres publiques, et seront dès lors omis lors de l'interprétation. Elle montre également que ces réseaux présentent un nombre important de métabolites sans identifiant standard, nuisant à la recherche de correspondance entre les données de métaboliques et les entrées des réseaux. Cette autre limitation sera discutée plus en détails dans la partie III.

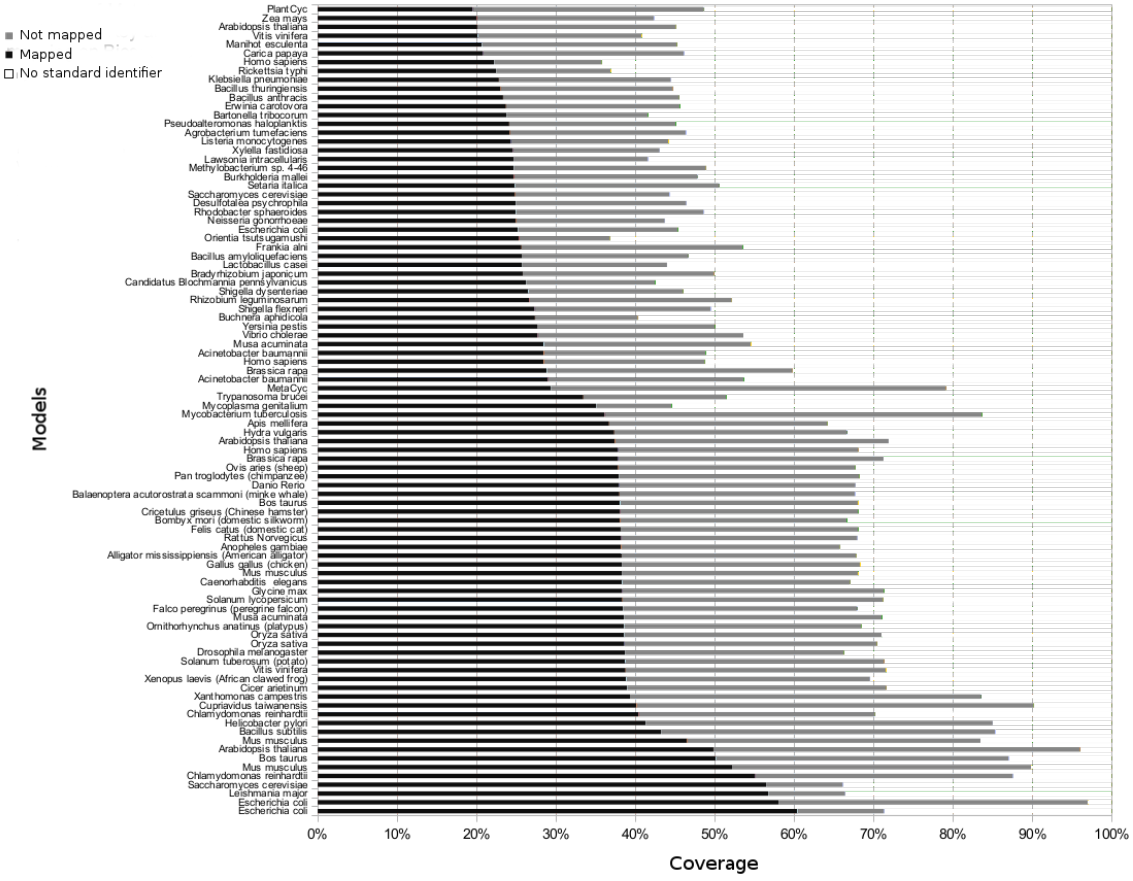


Figure 6.1 – Proportion de métabolites dans les réseaux métaboliques qui possèdent un spectre standard pour l'identification dans les bases de données HMDB, MassBank, ReSpect et GNPS

Les méthodes visant la simulation de processus dans les réseaux métaboliques sont très sensibles à la qualité des données d'entrées, ce qui limite leur utilisation à grande échelle à partir de données de métabolomique. Les approches topologiques visant à quantifier la proximité qui existe entre métabolites d'intérêt, est en re-

vanche moins sensible. En effet, l'ajout ou le retrait *a posteriori* d'un métabolite d'une liste d'intérêt n'affecte pas les distances précédemment calculées entre les autres membres de cette liste. En revanche elles sont également particulièrement tributaires de la qualité du réseau. Certaines petites variations peuvent avoir un impact sur la composition des chemins, sans particulièrement affecter de manière extrême leurs longueurs. *A contrario*, le retrait de certains métabolites ou certaines réactions peut affecter de manière drastique les distances dans le réseau, voir déconnecter certains de ses éléments. Des mesures topologiques telles que les centralités, qui seront détaillées dans le prochain chapitre, permettent d'identifier ces points critiques. L'absence de tels acteurs dans un chemin, où un faible différentiel de longueurs entre les k-shortest paths, pourrait être utilisés comme gages de stabilité de la distance entre deux métabolites face aux suppressions dans le réseau.

L'une des limites des méthodes de recherche de chemins est qu'elles ne considèrent qu'une partie du réseau bornée par les métabolites d'intérêt, c'est-à-dire qu'elles vont proposer des métabolites intermédiaires et des cascades réactionnelles permettant de joindre deux métabolites d'intérêt, et ne vont pas considérer les métabolites précurseurs de ces derniers. Les interprétations dérivées des sous-réseaux issus de la compilation des chemins obtenus supposent donc que l'ensemble des métabolites qui « bornent » les processus en lien avec la perturbation sont connus. Cette hypothèse n'est vraisemblablement pas vérifiée dans la plupart des cas étant donnée la nature des données, et tout particulièrement dans le cas des études sur des échantillons de biofluides où seule la fraction circulante du métabolome est considérée.

La section suivante va proposer une méthode qui tentera de répondre à cette limite, en intégrant les pistes mentionnées dans cette partie pour la prise en compte de chemins pertinents dans les réseaux métaboliques, notamment la gestion des composés auxiliaires.

Troisième partie

Interpréter des résultats de métabolomique grâce aux réseaux

Chapitre 7

Systèmes de recommandation et centralité dans les réseaux

7.1 Introduction

> *"A large shift is going on in our connected society : we are leaving the Information Age and entering the Recommendation Age. Today information is ridiculously easy to get ; you practically trip over it on the street. Information gathering is no longer the issue -making smart decisions based on the information is now the trick... So recommendations act as shortcuts through the information mass, getting us to the right, or « right enough » answer."*
>
> Adam Richardson

Comme décrit en introduction, la métabolomique offre une vue partielle du métabolome. Par conséquent, statuer sur les mécanismes affectés par une condition s'avère ardu étant donné les pièces manquantes du puzzle. Dans l'optique de proposer des métabolites potentiellement affectés par la perturbation, mais absents de la liste de molécules discriminantes, la centralité apparaît comme un outil approprié. En effet, en mettant en exergue les métabolites partageant une plus grande proximité avec les métabolites discriminants, de nouvelles hypothèses

peuvent émerger.

Ce type d'approches s'apparente à la famille des systèmes de recommandation. Les systèmes de recommandation ont pour vocation de suggérer à un utilisateur des éléments (items) pertinents, étant donné par exemple les éléments pour lesquels l'utilisateur a déjà manifesté un intérêt. Ces systèmes ont acquis une grande popularité ces dernières années, notamment dans le domaine des activités en lignes. Ils ont été utilisés pour suggérer de nouveaux articles sur les sites de vente en ligne, des films[29], des vidéos ou des musiques sur les plateformes de visionnage ou d'écoute à la demande[17] ou encore des personnes à suivre sur les réseaux sociaux numériques[107][169]. Les systèmes de recommandation répondent à un nouveau besoin, dans un monde où l'accès à l'information est devenu aisée, mais où le défi repose dans la manière de traiter intelligemment cette masse d'information. Les systèmes de recommandation permettent de « dégrossir » cette masse de données pour obtenir une information la plus pertinente et la plus raffinée possible. Ces développements récents offrent également une opportunité pour l'analyse de données de métabolomique étant donné l'analogie entre les problématiques qu'ils résolvent et les besoins en métabolomique. La métabolomique met en évidence un ensemble de métabolites d'intérêt, en lien avec une condition biologique donnée. Cet ensemble étant incomplet[11], un système de recommandation proposant des métabolites d'intérêt potentiel, étant donné un ensemble initial, semble constituer une opportunité intéressante pour tirer parti au maximum des résultats obtenus en métabolomique. De plus, les réseaux métaboliques fournissent une structure de données qui a été largement exploitée pour les systèmes de recommandation.

7.2 Types de systèmes de recommandation

Deux grandes catégories de systèmes de recommandation existent (ainsi qu'une catégorie tierce de méthodes hybrides les combinant). La première est le *collaborative filtering*[196][231], où les marqueurs d'intérêt des autres utilisateurs sont utilisés pour formuler les recommandations. Un exemple d'application vise à identifier des utilisateurs au profil d'intérêt similaire à celui de l'utilisateur cible pour

lui proposer du contenu globalement apprécié par ces utilisateurs d'intérêt similaires. Cette stratégie ne s'applique pas, ou du moins pas de manière évidente, à la problématique d'extension de profils métaboliques.

La seconde catégorie est dite du *content-based*. Ces systèmes, plutôt que de s'appuyer sur le comportement des utilisateurs similaires, suggèrent du contenu similaire à celui pour lequel l'utilisateur a exprimé un intérêt[175][4][99]. Les recommandations sont indépendantes des choix des autres utilisateurs, et reposent uniquement sur une similarité entre éléments[139]. Elles peuvent être, par exemple, réalisées sur la base de comparaison de *feature vectors*. Chaque position de ces vecteurs représente une propriété ou une caractéristique, et la valeur qui leur est attribuée indique si cette propriété est portée par l'item considéré.

Dans le cas des recommandations sur les réseaux sociaux numériques, cette similarité peut être exprimée au travers d'une proximité dans le réseau[107][134][16][170]. Pour les réseaux sociaux numériques représentant des liens d'amitié ou plus largement de connaissances, une hypothèse peut être faite sur la transitivité de ces relations. Pour un utilisateur donné, les amis de ses amis vont avoir tendance à entrer dans son cercle de connaissance si ces liens d'amitié sont suffisamment forts et récurrents. Cette notion de transitivité peut par analogie être transposée aux réseaux métaboliques. Suggérer des métabolites potentiellement atteints par une perturbation métabolique, en fonction de leur proximité avec les métabolites d'abondance anormale, constitue une problématique similaire. La centralité offrant un moyen d'estimer une proximité relative à un ensemble de nœuds d'un réseau, elle apparait comme appropriée pour répondre à ce besoin. Elle a été notamment exploitée par le système de recommandation d'utilisateur à suivre de la plateforme de microblogage Twitter[103].

7.3 Centralités

La centralité peut être définie comme le critère permettant d'identifier les acteurs importants d'un réseau[267][43]. La section suivante a pour but d'expliciter cette notion au travers des exemples les plus emblématiques de centralité. Comme

la définition générale peut le suggérer, il peut exister autant de mesures de centralité qu'il existe de définition de l'importance. La section proposée est par conséquent loin d'être une revue exhaustive des travaux relatifs à cette problématique, qui constitue une thématique prolifique, notamment en sociologie.

Face à la grande variété de mesures, il convient de définir les critères permettant de rationaliser un choix parmi ces dernières. Les quelques mesures proposées ici sont représentatives de différentes notions sous-jacentes à la définition de centralité. Ces notions portent notamment sur la définition d'importance, la définition du rôle d'un acteur dans un réseau, la manière dont circule l'information ou les ressources dans un réseau, mais également l'objectif derrière l'identification d'acteurs clés. Cette section fera ainsi office d'introduction à la partie qui traitera du choix d'une mesure de centralité adaptée à la recommandation de métabolite, qui reposera sur l'analyse des notions intrinsèques précédemment mentionnées.

7.3.1 Centralités de proximité

La première est celle de **closeness** (centralité de proximité). La *closeness* permet de mettre en exergue les nœuds dont les distances les séparant des autres nœuds du graphe sont relativement faibles[154][39][150]. La mesure la plus utilisée est celle de *closeness* géodésique de Freeman[92] qui utilise la longueur du plus court chemin comme mesure de distance. Elle est généralement définie comme l'inverse de la somme des distances entre un nœud donné et tout autre nœud du graphe. Soit :

$$C_C(i) = \frac{1}{\sum_{j \in V} d(i,j)}$$

Avec $d(i,j)$ la distance entre le nœud i et le nœud j, et V l'ensemble des sommets du graphe.

La figure 7.1 illustre ce concept au travers du réseau d'interaction au sein d'une colonie de babouins observée par Dunbar et Dunbar, tel que décrit par Stephenson et Zelen[248]. Les individus mâles et femelles de plus haute centralité sont ici caractérisés par une forte proximité avec l'ensemble du groupe, et correspondent

au mâle leader et à la femelle dominante de la colonie, identifiés par observation comportementale. [1]

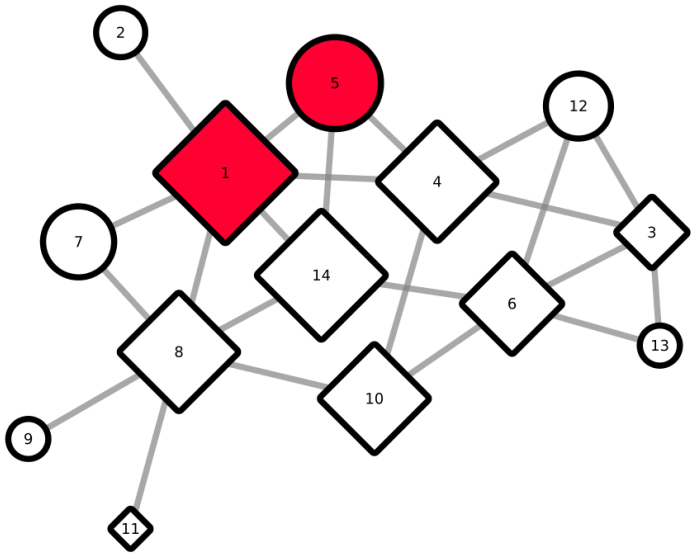


Figure 7.1 – Illustration de la centralité de proximité : exemple du réseau d'interactions sociales d'une colonie de babouins. Les losanges représentent les individus femelles, et les ronds les individus mâles. Les arêtes représentent les interactions sociales (essentiellement les événements de toilettage mutuels). La taille des nœuds est proportionnelle à la *closeness*. Les nœuds rouges représentent le mâle leader et la femelle dominante du groupe. Les numéros des nœuds correspondent aux identifiants des individus.

Parmi les nombreuses applications de la *closeness*, on peut citer son lien avec les problèmes d'emplacement d'installations (*facility location problem*) qui visent par exemple à identifier l'emplacement de construction d'un hôpital de manière à minimiser la distance à parcourir pour le rejoindre depuis un point de la ville quelconque[273]. En biologie, elle a par exemple été utilisée pour mettre en évidence la centralisation du métabolisme d'*Escherichia Coli* autour de la glycolyse

1. Le réseau présenté est la résultante de l'intégration d'une femelle et d'un jeune mâle dans la colonie. Dans le réseau initial, la femelle dominante n'occupe pas la place la plus centrale en termes de proximité. Il est à noter que les interactions correspondent essentiellement à des comportements de toilettage mutuels, et la fréquence de ces événements n'est ici pas considérée.

et du cycle de l'acide citrique[177].

D'autres implémentations de la *closeness* proposée ici existent. Certaines considèrent l'inverse de l'excentricité, qui pour un nœud donné correspond à la distance maximale (longueur du plus court chemin) qui le sépare d'un autre nœud du graphe. Des variantes existent, comme la centroïde *closeness*, qui considère la distance à un (voire plusieurs) centroïde(s) plutôt que l'ensemble des distances entre paires de nœuds. Ce centroïde peut être défini en fonction des propriétés structurales du graphe, en choisissant par exemple le centre du graphe (nœud de plus faible excentricité). Le centroïde peut également être défini en fonction de propriétés extrinsèques au graphe (en sélectionnant par exemple le mâle dominant dans le réseaux figure 7.1). Il a également été proposé de considérer les distances relativement au **diamètre** du graphe, qui correspond à la longueur maximale du plus court chemin observable dans un graphe, afin de pouvoir plus facilement comparer les mesures obtenues sur différents graphes[260]. Une des limites des mesures basées sur la *closeness* est que, dans le cas de graphes non connexes, il existe pour chaque nœud au moins une paire le contenant où il n'existe pas de chemin, conduisant à une distance égale à l'infinie. Usuellement, c'est la somme des distances réciproques qui est considéré, où l'on substitue les membres $\frac{1}{\infty}$ par 0.

7.3.2 Centralités d'intermédiarité

La **betweenness** (centralité d'intermédiarité) correspond à la proportion de plus courts chemins entre deux nœuds qui passe par le nœud considéré. La *betweenness* peut donc être définie de la manière suivante :

$$C_B(i) = \sum_{s \neq i \in V} \sum_{t \neq i \in V} \frac{\sigma_{st}(i)}{\sigma_{st}}$$

Avec σ_{st} l'ensemble des plus courts chemins entre s et t, et $\sigma_{st}(i)$ le sous-ensemble de ces chemins qui contiennent le sommet i.

En considérant la proportion de chemins qui traversent un nœud, la *betweenness* va mettre en évidence les acteurs du réseau qui présentent un fort potentiel

de **contrôle** sur la transmission des biens ou de l'information dans le réseau. Cette mesure peut également mettre en avant des nœuds qui vont constituer des « ponts » entre différentes communautés (cette mesure est d'ailleurs corrélée à la modularité, fréquemment utilisée pour définir des communautés). Dans l'article de Freeman à l'origine de cette mesure, elle est résumée par le nombre de fois qu'un acteur donné a été mis à contribution pour qu'un acteur en joigne un autre[90]. Ainsi, le retrait des nœuds de haute betweeness va avoir tendance à perturber les échanges au sein du réseau, voir à le déconnecter.

La centralité de *betweenness* a notamment été utilisée par Padgett et Ansell pour illustrer la centralisation du pouvoir politique florentin du début de la Renaissance autour de la famille Medicis[205]. Sur la figure suivante (Figure 7.2A) est représenté le réseau des mariages et des relations économiques entre les principales familles florentines de l'époque. La famille Medicis se démarque clairement des autres par sa *betweenness* élevée. Ceci traduit le fait que son parti, représenté par les nœuds jaunes, est fortement centralisé autour de cette dernière. Ainsi non seulement il existe très peu de liens directs entre ses partisans, mais leurs connexions au reste de l'élite florentine ne se fait quasi uniquement que par l'intermédiaire de la famille Médicis. Cette caractéristique du réseau n'est en revanche pas capturée par la centralité de *closeness* (Figure 7.2B), dont la distribution des scores ne discrimine pas particulièrement la famille Médicis, bien qu'elle la place tout de même en 2^{e} postion *ex aequo* avec la famille Guasconi, et 1re position au sein de son parti.

En biologie, la *betweenness* a notamment été utilisée pour théoriser l'organisation modulaire du réseau d'interactions protéiques chez la levure, et y identifier des protéines potentiellement essentielles[138][282]. La corrélation entre l'essentialité des protéines et leur *betweenness* dans les réseaux d'interactions protéiques, nommée la *centrality-lethality rule*, a également conduit à son utilisation pour l'identification de cibles thérapeutiques[120].

La *betweenness* partage une définition similaire avec la centralité de stress, qui prend en compte le nombre absolu de plus courts chemins plutôt que la proportion

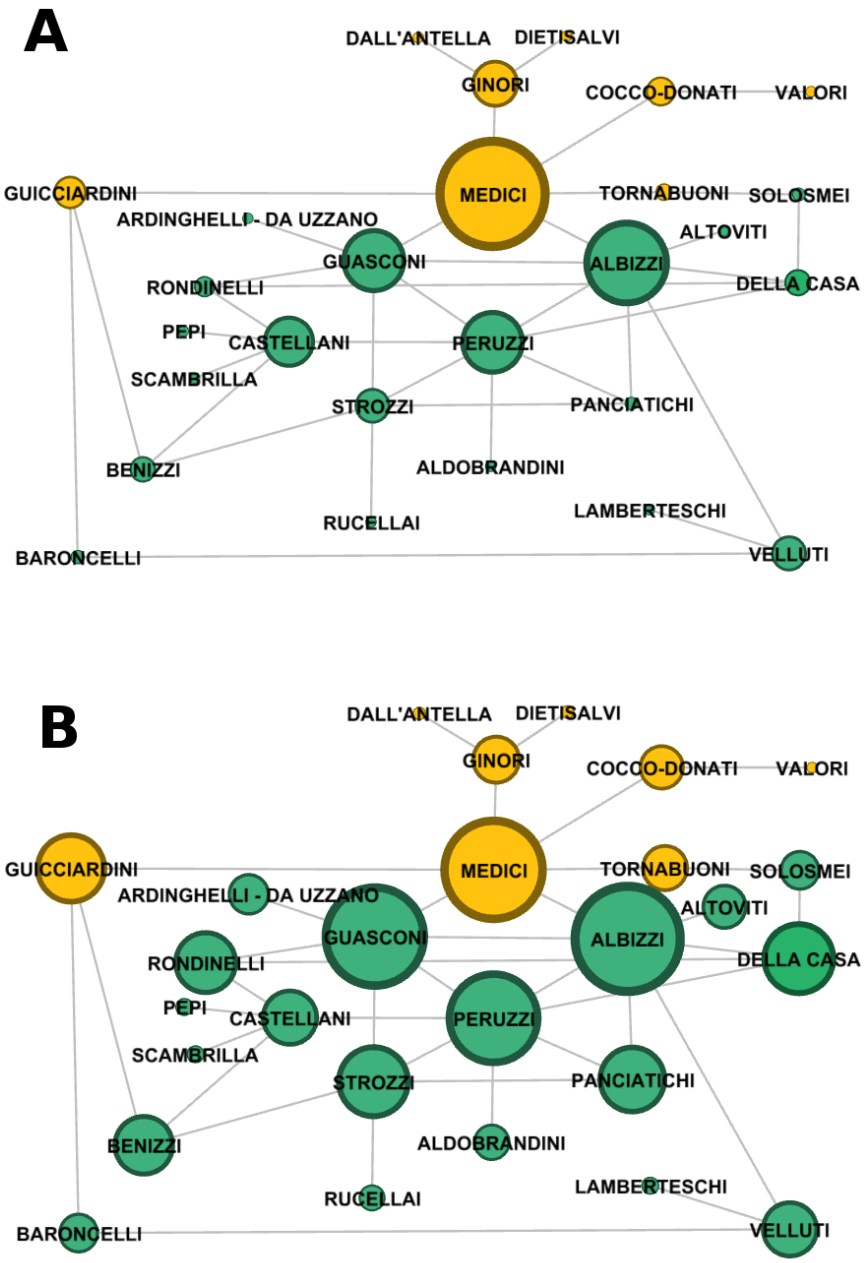


Figure 7.2 – Illustration de la centralité d'intermédiarité : exemple du réseaux de mariages et de liens économiques au sein des principales familles florentines du début de la Renaissance. A. Centralité de *betweenness* B. Centralité de *closeness*. La taille des nœuds est proportionnelle à leur valeur de centralité

pour une paire donnée. D'autres variations existent, ajoutant des contraintes au niveau des chemins considérés[42]. On peut par exemple ne considérer que les chemins ne partageant pas d'arêtes (*edge disjoint paths*, on parle également dans ce cas de *flow betweenness*[91]) ; ou ne partageant pas de nœuds (*vertex-disjoint paths*) pour éviter de surestimer la *betweenness* lorsque des chemins sont trop semblables. Il est également possible de considérer les k plus courts chemins (on parle alors de *k-betweenness*), fixer une longueur maximale de chemin, ou pénaliser les chemins par leurs longueurs[79] afin de rendre la mesure plus robuste aux additions et suppressions d'arêtes. Tout comme pour le cas de *closeness*, les chemins considérés peuvent être calculés à partir d'un sous-ensemble des paires de nœuds du graphe.

Il est à noter que cette mesure reste applicable en l'état sur un graphe déconnecté. En revanche, contrairement à la *closeness* qui prend en compte la longueur du chemin, cette dernière considère une énumération de chemins. Dès lors, il convient de prendre en compte la totalité des plus courts chemins pour une paire de nœuds, étant donné que plusieurs solutions peuvent exister pour une même paire.

7.3.3 Mesures de vitalité

> *"Aussi, lui absent, on se trouvait en face du néant pur et simple, et tout cet édifice formidable s'écroulait comme un château de cartes."*
>
> Jules Verne, les Cinq Cents Millions de la Bégum

Les mesures de **vitalité** constituent un cas particulier de centralité où l'importance d'un nœud est déterminée à partir d'une fonction calculée sur l'ensemble du graphe. C'est la différence entre les résultats de cette fonction observés avant et après retrait du nœud qui constitue la mesure de centralité[154]. Par exemple, la *closeness vitality* va être calculée par l'impact du retrait d'un nœud sur la somme de toutes les distances du graphe (l'**index de Wiener**). La mesure de fragmentation pondérée par la distance (DF-measure), proche de la *closeness vitality*, correspond à la moyenne de l'inverse des distances après retrait d'un nœud,

aussi appelé *overall cohesion*.

La centralité de stress peut également être vue comme une mesure de vitalité, car elle correspond au nombre de plus courts chemins perdus après retrait d'un nœud. Dans un même registre, la fragmentation (F-measure) consiste en une mesure de vitalité qui correspond à la proportion de paires déconnectées après retrait d'un nœud[36].

Il peut ainsi exister autant de mesures de vitalité qu'il existe de mesures topologiques globales. Ce type de centralité est principalement utilisé dans des contextes de perturbation d'un réseau, pour en identifier ses vulnérabilités. À titre d'exemple, elles ont été utilisées pour analyser les vulnérabilités du réseau électrique des États du nord et de l'ouest des États-Unis[119], identifier les individus pouvant fragmenter un réseau criminel[181], ou encore mettre en évidence les régions cérébrales dont les liaisons seraient particulièrement à risque[117][132]. La figure 7.3 montre un exemple de l'effet du retrait d'un nœud de haute vitalité de stress sur le réseau social du club de Karaté suivi par le sociologue Wayne Zachary[277]. Zachary s'est intéressé aux phénomènes de fission dans les groupes d'individus, et a suivi un groupe de personnes au sein d'un club de Karaté, où une querelle entre l'instructeur Mr Hi et l'administrateur John A a conduit à la fission du club, plusieurs membres ayant rejoint le club nouvellement formé par Mr Hi après son départ. Bien que Zachary se soit intéressé au partitionnement du réseau pour prédire les groupes issus de la fission et non la centralité de ces acteurs, on peut remarquer que Mr. Hi dispose de la plus grande centralité de stress, ce qui aurait pu suggérer l'importance qu'a eu son départ sur le club de Karaté. On peut également noter qu'il est le seul dont le retrait induit une déconnexion du graphe (propriété qui pourrait être capturée par la vitalité F-measure), ce qui illustre également son importance dans le maintien de la cohésion du club.

L'une des limites des mesures de vitalité/vulnérabilité est qu'elles impliquent généralement un coût calculatoire qui peut être prohibitif pour des réseaux de grande taille. En effet, elles nécessitent de répéter n fois (n étant le nombre de nœuds du graphe) un calcul de mesure globale, impliquant dans de nombreux cas

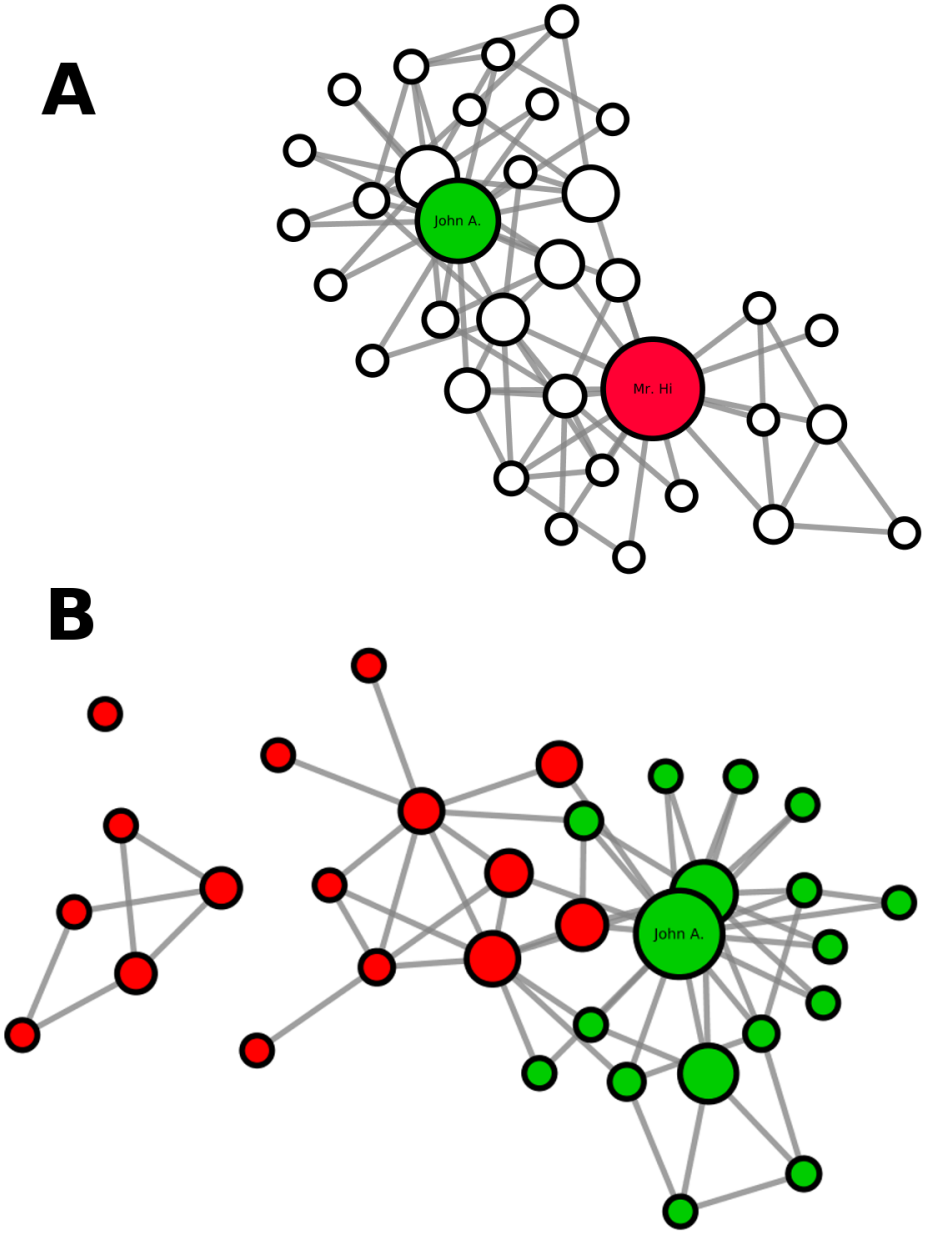


Figure 7.3 – Représentation de la vitalité des individus au sein du réseau des liens d'amitiés entre les membres d'un club de Karaté. A. Centralité de stress. La taille des nœuds est proportionnelle au nombre de plus courts chemins qui seraient affectés par leur retrait. B. Effet du retrait de Mr. Hi sur le réseau. La couleur des nœuds correspond à la répartition des groupes après la fission.

le calcul de l'ensemble des plus courts chemins.

7.3.4 Centralités de feedback

"The richest man is the one with the most powerful friends"

Don Altobello, The Godfather : Part III

Les centralités de *feedback* ont la particularité d'accorder plus d'importance aux nœuds dont les voisins sont importants. Elles sont particulièrement adaptées aux processus de diffusion, ou le rôle d'un individu va dépendre du rôle de ses voisins. Ce type de centralité a notamment été employée pour identifier des personnes à risque dans le contexte de transmission du VIH au sein des réseaux de rapports sexuels ou de partages de seringues[21]. De manière triviale, il est possible d'identifier les individus à risque en se focalisant sur ceux impliqués dans le plus grand nombre de ces relations[227], c'est ce qu'on appelle la **centralité de degré**. Les individus de fort degré sont fréquemment désignés sous le terme de *hub*. En revanche, la centralité de degré échoue à mettre en exergue les individus de faible degré qui sont impliqués dans une relation avec un *hub*, alors que cette seule interaction suffirait à les considérer légitimement comme « à risque »[35]. Les centralités de *feedback* permettent de considérer ce facteur, et peuvent être perçues comme une généralisation de la centralité de degré (qui ne considère que les nœuds à une distance de 1), en prenant en compte un voisinage « étendu », défini par les nœuds atteignables par des marches d'une longueur donnée.

La centralité de Katz[143], historiquement une des plus anciennes centralités de *feedback*, parfois mentionnées sous le terme de statuts plutôt que de centralité, est basée sur cette idée. Elle a été utilisée dans de nombreuses applications, dont l'identification de gènes candidats à partir des réseaux d'interaction protéique[280]. La mesure de Katz est une somme pondérée du nombre de marches qui émanent (ou se terminent) en un sommet donné. La pondération diminue l'impact d'une marche dans sa contribution à la somme totale, en fonction de sa longueur (le paramètre qui régit l'impact de cette pénalité doit être défini par l'utilisateur).

Concrètement, cette mesure illustre le fait qu'un individu peut influencer indirectement un autre, au travers d'une chaîne d'individus s'influençant mutuellement. Plus ces chaînes entre les deux individus sont nombreuses, plus l'influence du premier sur le second est certaine, mais plus elles sont longues, plus l'influence sera faible. La centralité de Katz peut être formulée de la manière suivante :

$$C_{katz}(i) = \sum_{k=1}^{\infty} \sum_{j=1}^{n} \alpha^k (A^k)_{ji}$$

Où k correspond à la longueur des marches, n au nombre de nœuds dans le réseau, α le facteur d'atténuation et A la matrice d'adjacence. Le facteur d'atténuation α doit être inférieur à 1, ainsi quand k tend vers l'infini, $\alpha^k = 0$. Il est à noter qu'en représentation matricielle, le nombre de marches de longueur k entre deux nœuds est donné par la matrice d'adjacence à la puissance k. Le vecteur de centralité $\vec{C}_{katz}$ peut être exprimé de la manière suivante :

$$\vec{C}_{katz} = (\sum_{k=1}^{\infty} (\alpha A)^k))1$$

Où 1 est un vecteur de taille n dont tous les éléments sont égaux à 1.

Hubbell proposa une formulation générale où le vecteur 1 est remplacé par un vecteur de statut attribuant un score à chaque nœud, permettant une pondération *a priori* des éléments du réseau sur un critère extrinsèque[128]. Plusieurs alternatives à la centralité de Katz ont également été proposées, comme la centralité de Hoede qui remplace le facteur d'atténuation uniforme par une fonction de pondération[118], où celle de Bonacich qui permet l'usage de poids négatifs[32].

Bonacich proposa également la centralité de vecteur propre (*eigenvector centrality*), en remarquant la similitude entre la formulation des centralités de *feedback* précédentes et la définition du vecteur propre d'une matrice[31]. Il proposa alors de considérer le principal vecteur propre comme mesure de centralité. Ce vecteur propre correspond à la limite de la mesure de Katz quand α tend vers $1/\lambda$, où λ correspond à la plus grande valeur propre de A. Les liens entre cette mesure de centralité et celle de Hubbell, ainsi que celle de Katz et Hoede qui correspondent à des cas particuliers de cette dernière, font que la centralité de vecteur propre fut

considérée par certains comme un « résumé élégant » de ces 3 mesures[39].

7.3.5 PageRank

> *"The road to wisdom ? Well it's plain and simple to express : err and err and err again but less and less and less"*
>
> Piet Hein

La centralité de PageRank est une centralité de *feedback* apparentée à la centralité de vecteur propre, à la différence près qu'elle considère une version modifiée de la matrice d'adjacence, sous forme de matrice stochastique[206]. Une matrice stochastique est une matrice carrée dont la somme des lignes est égale à 1, et où les éléments sont compris entre 0 et 1. Cette matrice, ici notée S représente ainsi les probabilités de transitions entre les nœuds, où les marches sont modélisées par des chaines de Markov[174]. Les processus de Markov représentent des évènements de changement d'état, où les prédictions de l'état futur ne dépendent que de l'état présent, on parle de processus « *memoryless* ». Les marches aléatoires, qui ne sont pas contraintes par l'interdiction de visiter plus d'une fois un même nœud ou un même arc, répondent à ce critère. L'élément i, j de la matrice S^k représente alors la probabilité que le nœud j soit atteint par une marche aléatoire de longueur k, émanant de i. Une centralité peut alors être définie à partir de l'espérance qu'un nœud soit visité par une marche aléatoire émanant de n'importe quel nœud du réseau.

Une mesure de centralité, la *Random-Walk betweenness*, proposée par Newman, est basée sur un concept similaire, et a notamment été appliquée aux réseaux métaboliques pour l'extraction de sous-réseaux et de métabolites d'intérêt[198][262].

Le PageRank exploite également ces propriétés. Il fut créé par Lawrence Page et Sergey Brin, cofondateurs de la société Google, et fut utilisé pour la création de leur moteur de recherche. Le PageRank permettait d'ordonner les résultats d'une requête pour proposer en priorité des pages importantes. L'importance y est définie par le nombre de liens pointant vers cette page, pondérée par l'importance des pages dont sont issus ces liens, ce qui correspond bien à la définition des

centralités de *feedback*.

L'idée derrière le PageRank est de simuler le comportement des utilisateurs lors d'une navigation sur le web par des chaînes de Markov. Le terme de *random surfer* est ainsi souvent utilisé pour décrire le principe du PageRank. Le *random surfer* part d'une page choisie aléatoirement, et navigue de page en page en empruntant des liens au hasard. À chaque transition, il existe une certaine probabilité que sa marche s'arrête pour reprendre depuis un nouveau point de départ choisi aléatoirement. Cette probabilité est calculée à partir du *damping factor*, et constitue la principale particularité du PageRank.

Ainsi, à partir de la matrice d'adjacence est construit une matrice de transition T, avec $T_{ij} = A_{ij}/(\sum_k A_{ik})$, pour tout élément i, j tel que $A_i j > 0$. On construit ensuite la matrice transformée G, parfois nommée « Google matrix », de la manière suivante :

$$G_{ij} = \begin{cases} \alpha T_{ij} + (1-\alpha)\frac{1}{N} & \text{si } \sum_k T_{ik} = 1 \\ \frac{1}{N} & \text{si } \sum_k T_{ik} = 0 \end{cases}$$

avec N le nombre de nœuds représentés dans la matrice d'adjacence, et α le *damping factor*. Le PageRank correspond aux entrées du principal vecteur propre de cette matrice, qui peuvent dans ce contexte se traduire par la probabilité que le *random surfer* visite un nœud.

Sous forme matricielle la définition de la matrice considérée peut être définie de la manière suivante :

$$G = \alpha T + (1-\alpha)E$$

Où E est une matrice $N \times N$ dont toutes les entrées valent $\frac{1}{N}$. Il est à noter que, contrairement à la définition précédente, cette définition ne prend pas en compte le cas des « *dangling nodes* », c'est-à-dire les nœuds puits qui ne pointent vers aucun autre nœuds. Dans l'article original de Page et Brin, les nœuds puits sont simplement retirés, rendant la matrice T stochastique. Il est également possible de remplacer les lignes de T dont la somme est égale à 0 par un vecteur w, dont tous les éléments sont égales à $\frac{1}{N}$ dans notre cas, pour rendre la matrice stochastique.

Le *damping factor* permet, comme dans le cas du facteur d'atténuation de l'indice de Katz, de limiter l'importance des marches longues, mais permet également de « sortir » des nœuds puits ou de petites composantes fortement connexes grâce aux réinitialisations qu'il induit (évitant ce qui est parfois appelé l'*Hotel California effect*[198], qui piège le *random surfer*)[24].

Ces réinitialisations offrent un autre intérêt, celui d'assurer la convergence de la **méthode de la puissance itérée**, principale méthode de calcul du PageRank, vers le principal vecteur propre de la matrice[26]. Cette convergence implique qu'il existe une distribution de probabilité stationnaire pour la visite des nœuds. En d'autres termes : quand la longueur des marches tend vers l'infini, le système atteint un équilibre. D'autres méthodes existent pour obtenir la distribution stationnaire, impliquant généralement de couteuses inversions de matrices. Néanmoins, la méthode de la puissance itérée est l'une des plus plébiscitées du fait qu'elle est applicable à de très grandes matrices creuses, qu'elle requiert un minimum de stockage mémoire et qu'elle est facile à implémenter.

La méthode de la puissance itérée consiste à monter en puissance une matrice de manière itérative jusqu'à convergence. Cette convergence est garantie si la matrice est **stochastique**, **irréductible** et **apériodique**. Par définition, la matrice de transition considérée est stochastique. En revanche, une matrice de transition est irréductible si et seulement si le graphe est *fortement connexe*, c'est-à-dire qu'il existe un chemin orienté entre n'importe quelle paire de nœuds. Ordinairement non vérifiée dans les réseaux *real-world*, dont les réseaux métaboliques, cette caractéristique est garantie par la réinitialisation aléatoire des marches[24]. En effet, la modification de la matrice de transition conduit à l'affectation d'une valeur non nulle à tous les éléments, ce qui revient à rendre le graphe **complet** (il existe un arc entre tout couple de nœuds). Cette caractéristique suffit également à rendre la matrice apériodique. La périodicité d'un graphe (et de sa matrice d'adjacente) correspond au plus grand diviseur commun de la longueur de ses cycles, et il est dit apériodique si cette périodicité est égale à 1. Dans un graphe orienté complet d'ordre > 2, il existe pour tout nœud i une **marche fermée** (marche allant de i

à i) de longueur 2, et une de longueur 3. Le plus grand dénominateur commun est alors de 1. [2]

Originellement, une distribution de probabilité uniforme est utilisée pour régir les réinitialisations : tous les nœuds ont la même probabilité d'être choisi comme point d'arrivée d'une réinitialisation : $(1-\alpha)\frac{1}{N}$. Les auteurs notent qu'il est néanmoins possible de biaiser ces réinitialisations afin de favoriser certain nœuds, en définissant une probabilité pour chaque nœud, créant ainsi une matrice E qui serait représentative des goûts de l'utilisateur et définie depuis des paramètres extrinsèques[114]. Cette procédure est baptisée **Personalized Page Rank**, et constitue la principale adaptation du PageRank dans son exploitation pour les systèmes de recommandation.

L'une des principales limites du PageRank dans le contexte du moteur de recherche est qu'il est aisé de créer des structures de pages web qui augmentent artificiellement le PageRank des pages[24]. On peut citer par exemple les « *Tightly Knit Community* », groupes de pages avec de nombreux liens entre elles, mais aucun lien sortant du groupe. Motivé par les implications économiques de la visibilité d'un site web sur les moteurs de recherche, la création de ces structures, appelées *link farms* s'est généralisée et a conduit à l'utilisation de nouvelles stratégies par les moteurs de recherches[108]. En revanche, le PageRank a été utilisé et adapté dans de nombreuses autres applications[101]. Il a par exemple été employé pour étudier le connectome en neuroscience[46], prédire le trafic sur les réseaux routiers, classer des équipes sportives ou encore identifier des espèces animales en danger dans un réseau trophique[7]. Plusieurs systèmes de recommandation sont également basés sur le PageRank personnalisé ou utilisent des principes similaires impliquant des marches aléatoires. Par exemple, le système de recommandation de la plate-forme de microblogage Twitter pour la suggestion d'utilisateurs à suivre[103], le système

2. Note : les graphes bipartis qui contiennent des cycles sont toujours périodiques : le nombre de pas pour former un cycle est nécessairement pair, donc avec un dénominateur commun de 2 au minimum. Dans le cas des graphes métaboliques, il devient alors nécessaire que la transformation du graphe forme un graphe non bipartie, en autorisant les téléportations du *random surfer* de métabolite à métabolite, où de réaction à réaction.

de recommandation par mot-clés *FolkRank* utilisés par BibSonomy[122], ou encore d'autres systèmes de recommandation d'œuvres culturelles[29].

Dans le cadre de la biologie, il a été utilisé pour la priorisation de gènes candidats et l'analyse de réseaux d'interactions[153][224], dont un exemple notable est celui du *GeneRank*[194]. Il a également été appliqué aux réseaux d'interaction protéine-protéine[133] pour identifier des protéines cibles et inférer des annotations fonctionnelles, donnant notamment naissance au *Protein Rank*[93].

7.4 Bilan

Il existe donc un large éventail de solutions pour identifier des acteurs important dans un réseau. Ces solutions diffèrent par la manière de modéliser les échanges dans ces réseaux, mais également par différentes manières de conceptualiser l'importance. Certaines d'entre elles ont reçu un attrait particulier dans le contexte des systèmes de recommandation, et certaines ont également été utilisées pour formuler des hypothèses sur les rôles de gènes ou de protéines. Le chapitre qui suit propose une extension de ces applications à l'interprétation et la suggestion de métabolites d'intérêt, en présentant des arguments motivant le choix d'une mesure de centralité qui soit appropriée.

Chapitre 8

Application aux réseaux métaboliques

8.1 Choix d'une mesure appropriée

Comme mentionné dans la première partie, la qualité descriptive d'une mesure de centralité va dépendre de la topologie des chemins considérés, et dans une autre mesure de la pondération qui va être appliquée. Ces deux critères permettent de capter les caractéristiques des relations étudiées. Dans le cas des réseaux métaboliques, nous avons proposé différents axes pour établir des relations indirectes entre métabolites, qui tentent de représenter des phénomènes biologiques de manière pertinente.

En revanche, la définition d'un indice de centralité adéquat ne dépend pas uniquement d'une définition valide des chemins considérés, mais également d'une définition adéquate de l'importance. Borgatti *et al.*, dans des travaux postérieurs à ceux présentés sur les *flows*[37], proposent de nouvelles catégories de centralité[39]. Les exemples mentionnés dans la partie précédente décrivent une classe de centralité dérivée de la *closeness*. Ce type de méthode va proposer une définition de l'importance basée sur la distance d'un nœud avec le reste du réseau. Ces méthodes sont dites mesures de longueur, puisqu'elles considèrent la longueur des chemins comme facteur de l'importance. D'autres méthodes vont considérer le nombre de

chemins plutôt que leurs longueurs, on parle de mesures de volume. Une des représentantes les plus connues de cette classe est la mesure de *betweenness* (centralité d'intermédiarité) présentée précédemment.

La différence entre centralité de longueurs et de volumes repose sur des conceptions différentes de la proximité : deux nœuds peuvent être considérés comme proche du fait qu'un chemin court les relie, ou parce que de nombreux chemins les relient. Les auteurs suggèrent que, dès lors qu'il s'agit d'évaluer le risque de recevoir en un temps opportun quelque chose circulant dans le réseau, les centralités de longueur paraissent adaptées, car la longueur offre une estimation des temps d'arrivée. En revanche, s'il s'agit de statuer sur la certitude que quelque chose circulant soit reçu, alors les mesures de volume sont plus appropriées, en prenant en compte le nombre de chemins alternatifs possibles.

Il est à noter que ces différents indices ne sont pas incompatibles. Ainsi, les éléments centraux en accord avec une conception de la centralité en matière de volume peuvent également être considérés centraux dans le cas d'une conception axée sur la distance.

L'autre catégorisation des indices de centralité est basée sur la nature « radiale » ou « mediale » des chemins considérés. Les mesures *closeness-like* considèrent des chemins qui émanent ou se terminent sur le nœud évalué, et sont donc nommées « radiales ». En revanche pour les mesures *betweenness-like*, ce sont les chemins où le nœud évalué intervient en tant qu'intermédiaire qui sont considérés, d'où l'appellation « mediale ».

Dans le contexte de l'interprétation de résultats de métabolomique, la centralité est définie relativement à un groupe de nœuds d'intérêt (les métabolites discriminants) et non par rapport à l'ensemble des nœuds du réseau. Ainsi, une mesure radiale peut considérer les chemins partant d'un nœud donné et rejoignant l'un des nœuds d'intérêt, et inversement. Dans le cas d'une mesure médiale telle que la *betweenness*, une définition intuitive serait de considérer uniquement les chemins qui relient les nœuds d'intérêt entre eux.

Dans le cas des réseaux métaboliques, et plus particulièrement dans le cas

de la métabolomique, on observe des perturbations de l'abondance de certains métabolites. Comme décrit précédemment, une large partie des métabolites n'est pas mesurable. On veut dès lors tenter d'identifier des métabolites qui seraient atteints par la propagation de ces perturbations, ou des métabolites dont la perturbation de leur abondance aurait pu atteindre les métabolites d'intérêt identifiés en métabolomique. De manière intuitive, cette définition tend à nous orienter vers une mesure radiale de la centralité, en considérant des chemins qui prennent pour source ou cible des métabolites d'intérêt. Étant donné que l'on souhaite considérer la vraisemblance qu'une propagation de perturbation atteigne un nœud, il semble acceptable de considérer le nombre de chemins reliant ce nœud à l'origine de la perturbation plutôt que la distance qui les sépare, orientant le choix de la centralité vers une mesure de volume. De plus, comme mentionnées dans la partie précédente, les limites inhérentes à la construction de chemins métaboliques, ainsi que la faiblesse de l'hypothèse de parcimonie nécessaire à la définition d'une distance, suggèrent que la considération de plusieurs chemins soit particulièrement adaptée. Bien que nous nous concentrerons sur une mesure de volume radiale, il est cependant à noter que les autres types de centralité offrent des approches complémentaires pour raffiner l'élucidation des rôles des métabolites centraux qui seront identifiés[236][155].

Ainsi, la mesure de centralité adéquate devra :

— capturer le volume de chemins circulant radiaux (émanant ou atteignant les métabolites discriminants)
— considérer une propagation des perturbations par duplication parallèle par des marches. La propagation d'une perturbation pouvant amplifier l'écart d'abondance d'un métabolite déjà atteint, le chemin circulant peut emprunter plusieurs fois un même nœud ou un même arc, et sera donc mieux représenté par des marches
— considérer que les perturbations ne peuvent pas se propager d'un composé auxiliaire à un composé principal ou que les métabolites issus d'un pool suffisamment large sont insensibles aux perturbations, leur disponibilité

n'étant pas remise en cause.

— étant donné la nature des réseaux considérés, la mesure doit être robuste face aux ajouts/délétions de nœuds, et doit être adaptée à l'usage de réseaux non connexes.

Ces critères nous ont conduit à la sélection du PageRank comme mesure de centralité, qui sera combinée à une définition des probabilités de transitions basée sur des critères d'échange d'atomes, présentée en partie II. Cependant, l'une des limitations du PageRank dans le contexte des réseaux métaboliques, est qu'il ne permet pas de suggérer de potentiels « précurseurs » des nœuds d'intérêt. En effet, ces nœuds sont considérés comme l'origine de la propagation des perturbations observées, autrement dit la source des marches aléatoires. Les nœuds d'intérêt peuvent cependant être des intermédiaires dans cette propagation, voir les produits terminaux des mécanismes impliqués. Afin de combler ce manque, les résultats du PageRank vont être complétés par ceux du CheiRank[77][78][281]. Dans un graphe orienté, le calcul du CheiRank revient à calculer le PageRank sur ce même réseau après inversion de la direction de chaque arc. Le CheiRank va ainsi considérer des marches aléatoires « ascendantes », considérant les prédécesseurs des nœuds d'intérêt, pouvant ainsi potentiellement mettre en évidence des précurseurs importants. Les ajustements du PageRank conduisant à sa personnalisation seront transposés au CheiRank, et les suggestions de métabolites dépendront de la combinaison de ces deux scores, une approche parfois mentionnée sous le terme de 2Drank [78][281]. La section suivante présente l'application de cette approche à l'interprétation d'une signature métabolique caractéristique d'un syndrome neuropsychiatrique : l'**encéphalopathie hépatique**[75][239][6][47] (EH).

8.2 Application à la signature métabolique de l'encéphalopathie hépatique

Table 8.1 – Liste des métabolites d'intérêts obtenus à partir de la signature métabolique de l'encéphalopathie hépatique

3-(4-hydroxyphenyl)lactate	L-cystine
4-acetamidobutanoate	L-glutamate
4-pyridoxate	L-glutamine
5,6-dihydrothymine	L-kynurenine
5-Hydroxyindoleacetate	L-methionine
5-hydroxy-L-tryptophan	L-Octanoylcarnitine
5-oxoprolinate	L-phenylalanine
citruline	L-tryptophan
cortisol	L-tyrosine
D-gluconate	N-acetyl-D-glucosamine
D-glycerate	N-acetyl-L-alanine
glycocholate	O-acetylcarnitine
indole-3-acetate	O-Propanoylcarnitine
L-asparagine	taurocholic acid

Bioinformatics, YYYY, 0–0
doi: 10.1093/bioinformatics/xxxxx
Advance Access Publication Date: DD Month YYYY
Original paper

Subject Section

MetaboRank, metabolites you might be interested in: network based recommendation system to interpret and enrich metabolomics results

Clément Frainay [1], Sandrine Aros [2], Maxime Chazalviel [2], Thomas Garcia [1], Florence Vinson [1], Nicolas Weiss [3, 4], Benoit Colsch [5], Frédéric Sedel [2], Dominique Thabut [4, 6], Christophe Junot [5] and Fabien Jourdan [1,*]

[1] Toxalim, Université de Toulouse, INRA, Université de Toulouse 3 Paul Sabatier, Toulouse, FR
[2] Medday Pharmaceuticals, Paris, FR
[3] Unité de réanimation neurologique, département de neurologie, pôle des maladies du système nerveux central, Groupement Hospitalier Pitié-Salpêtrière Charles Foix, Assistance Publique – Hôpitaux de Paris, Paris, FR
[4] Brain Liver Pitié-Salpêtrière (BLIPS) Study Group, Groupement Hospitalier Pitié-Salpêtrière-Charles Foix, Assistance Publique – Hôpitaux de Paris & INSERM UMR_S 938, CDR Saint-Antoine Maladies métaboliques, biliaires et fibro-inflammatoire du foie & Institut de Cardiométabolisme et Nutrition, ICAN, Paris, FR
[5] DRF/JOLIOT/DMTS/SPI/LEMM, MetaboHUB-Paris, CEA-Saclay, Gif-sur-Yvette, FR
[6] Unité de Soins Intensifs d'Hépato-gastroentérologie, Groupement Hospitalier Pitié-Salpêtrière-Charles Foix, Assistance Publique - Hôpitaux de Paris et Université Pierre et Marie Curie Paris 6, Paris, FR

*To whom correspondence should be addressed.

Associate Editor: XXXXXXX

Received on XXXXX; revised on XXXXX; accepted on XXXXX

Abstract

Motivation: Metabolomics has shown great potential for the understanding of complex diseases, potentially leading to therapeutic target identification. However, no single analytical method allows monitoring all metabolites in a sample, resulting in incomplete metabolic fingerprints. This incompleteness constitutes a stumbling block to interpretation, raising the need for methods which can enrich those fingerprints. We propose MetaboRank, a new solution inspired by social network recommendation systems for the identification of metabolites potentially related to a metabolic fingerprint.

Results: MetaboRank method had been used to enrich metabolomics data obtained on cerebrospinal fluid samples from patients suffering from hepatic encephalopathy. MetaboRank successfully recommended metabolites missing from the original fingerprint. Quality of recommendations was evaluated by using literature automatic search to check that metabolites could be related to the disease. Complementary mass spectrometry experiments and raw data analysis were performed to confirm these predictions. In particular, MetaboRank predicted α-Ketoglutaramate as a metabolite which should be added to the fingerprint of hepatic encephalopathy, thus suggesting that getting confidence in metabolic fingerprints can provide new insight on complex diseases.

Availability: Method is implemented in the MetExplore server and is available at www.metexplore.fr. A tutorial is available at https://metexplore.toulouse.inra.fr/com/tutorials/MetaboRank/2017-MetaboRank.pdf

Contact: contact-metexplore@inra.fr

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

1 Introduction

Untargeted metabolomics studies allow monitoring a large range of small molecules (metabolome) in a tissue, an organism or a biofluid (Fiehn, 2002). When applied to human health research, a subset of this metabolome is considered as a metabolic fingerprint of a given pathology if it is statistically shared by a homogeneous group of patients in comparison to control subjects or another group of patients not affected by the pathology under study. These fingerprints constitute valuable supplementary knowledge that can be used for instance for patient stratification (Sreekumar *et al.*, 2009). Metabolic fingerprints also provide important clues for drug discovery and precision medicine since they reflect the biochemical modulations of human metabolism induced by a pathology (Hocher and Adamski, 2017).

A commonly used approach to establish a link between metabolic fingerprints and endogenous metabolism consists in applying enrichment analysis (Chagoyen and Pazos, 2011; Persicke *et al.*, 2012) on biochemical pathway collections provided by publicly available databases such as KEGG (Kanehisa *et al.*, 2014). This technique aims at finding which metabolic pathways contain a significant number of metabolites that belong to the fingerprint. The notion of metabolic pathway is informative since it assigns functions to a set of reactions. Nevertheless, the definition of their functional boundaries (input and output compounds) often varies from one database to another (Altman *et al.*, 2013) potentially leading to different interpretation when analyzing the same fingerprint in various databases. Fragmented view of metabolism offered by metabolic pathways is also a major limitation in global interpretation of fingerprints, especially when the systemic biochemical modulation associated with a disease is spanning several pathways.

To overcome these limits and take into account the full complexity of metabolism, metabolic fingerprints can be interpreted by considering them in the context of genome scale metabolic network which gathers all the biochemical reactions that can occur in a given organism (Mo and Palsson, 2009). Regarding human metabolism, networks reconstructed from genome annotation and manual curation had been made available to the community (Thiele *et al.*, 2013; Swainston *et al.*, 2016). However, the size of these networks (e.g. Recon2 human metabolic network contains 7440 reactions) makes the visual interpretation difficult and time-consuming. It thus requires the development of algorithms which reduce this complexity by finding the subset of reactions and metabolites (sub-network) that are related to the metabolic fingerprint. Several of these methods have been proposed (Frainay and Jourdan, 2016) and applied to interpret metabolic fingerprints (Milreu *et al.*, 2014) showing the strong added value of using network topology combined with graph-theory algorithms to give biological insight from a list of metabolites.

Most of these methods are based on path search between pairs of compounds in the fingerprint, assuming that the sources and end products involved in the mechanism are known. However, in contrary to gene or protein studies, no single metabolomic technology allows monitoring every small molecule in a sample leading to potentially incomplete fingerprints. Moreover, due to technological or biological artifacts, the annotation of detected molecules is a challenging task which may discard some parts of the fingerprint before downstream interpretation (Neumann and Böcker, 2010; Creek *et al.*, 2014). Fingerprint incompleteness can also be due to the fact that the matrix sampled to decipher metabolic modulations is not the tissue where the biochemical perturbations are occurring (e.g. metabolites are measured in cerebrospinal fluids to study brain afflictions). This gap may overshadow potential metabolic shifts and consequently important metabolites may be missing in the fingerprint conducting to misleading interpretations. The network approach presented in this article will thus not only help in biological interpretation but will also recommend candidate metabolites to enrich metabolic fingerprints.

Recommendation method development is an active research field in information retrieval community. The proposed algorithms have been intensively and successfully used for many applications such as content recommendation in social networks like Twitter™ (Backstrom and Leskovec, 2011; Liben-Nowell and Kleinberg, 2007; Gupta *et al.*, 2013; Liang *et al.*, 2014). Those algorithms typically suggest new people we might be interested in, regarding their connections with people already present in our personal list of interest. We propose to extend this concept to metabolic networks, thus suggesting new metabolites of interest by taking into account how they are connected to metabolites already present in the metabolic fingerprint. Our approach, in contrary to the original method borrowed from worldwide web analysis, will not consider that all the edges are equivalent since relationships between metabolites are more complex to interpret than web page links. In fact, it is necessary to ensure the biological and chemical relevance of connections used to compute the recommendations.

We will show how this approach, called MetaboRank, was successful in complementing the cerebrospinal fluid metabolic fingerprint of the hepatic encephalopathy disease described in (Weiss *et al.*, 2016). Hepatic encephalopathy (HE) corresponds to the neurological or neuropsychological symptoms of acute or chronic liver failure and/or portosystemic shunt. The spectrum is going from mild neuropsychological symptoms to impaired level of consciousness, often leading to coma. Even if the physiopathology is still largely unrevealed, the major role of hyperammonemia in conjunction of inflammation is well established (Weiss *et al.*, 2017). As a consequence, glutamine levels increase in the brain. However, the sole abundance of ammonemia doesn't scale with symptoms' severity and it has been shown that associated inflammation, increased levels of TNF-alpha and IL-6, were much better correlated to symptoms' severity.

In order to better decipher EH metabolic alterations, Weiss et al. used metabolomics to describe for the first time impaired metabolic pathways. Nevertheless, like many metabolomic experiments related to human health, the fingerprint was obtained from biofluid samples, overshadowing the importance of molecules that do not pass the blood-brain barrier. Moreover, the various degrees of disease gravity has led to high inter-individual variability at the metabolic level, making many abundance shifts weakly trustworthy. Those limitations emphasize the need of "looking outside the box" when dealing with biological interpretation.

In order to assess the relevance of hypotheses raised by our recommendation system, we applied automatic processing of literature data to associate concepts described in the literature to the recommended metabolites. We also reanalyzed raw data and patient metadata to confirm some of the recommended metabolites.

MetaboRank is implemented in the freely accessible web server MetExplore (Cottret *et al.*, 2010), allowing interactive analysis of the results.

2 System, methods and data

2.1 Centrality and Recommendation Systems

The underlying concept behind many social network recommendation systems is network centrality which aims at measuring the importance of a node. One of the best-known methods is the PageRank (PR) algorithm which was used by Google to rank web pages in a search result accord-

ing to their importance in the World Wide Web (Page *et al.*, 1999; Brin *et al.*, 1998). Since this first application, it has been successfully applied to many fields (Mihalcea *et al.*, 2004; Ma *et al.*, 2008), including biology (Allesina and Pascual, 2009; Iván and Grolmusz, 2011). PR defines the importance of a node as its probability to be encountered during a random walk in the network. In order to guarantee convergence of the random walk, stationary probability is obtained by adding to each node a probability to "jump" to any other target node, restarting the walk from this target node. This "jump" probability is set in the algorithm through a parameter referred as the damping factor. When those jumps are guided to favor some nodes according to a set of preferred ones, the term Personalized PageRank (PPR) is used (Haveliwala and H., 2002). PPR is well suited for fingerprint analysis since it has the ability to identify metabolites that are likely to be reached (produced) from molecules belonging to the fingerprint.

PPR can be considered as a downstream search that will evaluate the scope of a list of metabolites. But metabolites can also belong to a fingerprint because they are the outcomes of modulated metabolic processes. To enrich these results with upstream metabolites (i.e. potential precursors of fingerprint metabolites) we propose to compute the CheiRank (CR). CR and PR principles are similar except that for CR the links in the network are taken in reverse direction (Ermann and Shepelyansky, 2015). We adapted CR to create a Personalized CheiRank (PCR) which takes into account the list of metabolites in the fingerprint.

We propose to consider node centrality as a combination of these two measures. This two-dimension analysis combining PPR and PCR had been successfully used to analyze Wikipedia pages network (Zhirov *et al.*, 2010) or world trade networks (Ermann and Shepelyansky, 2015) but has never been applied to metabolic networks.

Metabolic networks are usually highly centralized around few hubs, such as Coenzyme A or ATP (Jeong *et al.*, 2000). It can thus be expected that these central nodes will always have a high PR or CR regardless of the content of the metabolic fingerprints. In order to limit this bias and emphasize metabolites which centrality is higher for a given fingerprint than in the general case, the scoring function is defined as the ratio between PPR (resp. PCR) and global PR (resp. CR).

Analysis of metabolic networks requires taking into account biochemical properties related to each reaction in order to obtain relevant results (Arita, 2004). To do so, MetaboRank will be computed on a probability matrix encoding biochemical knowledge as described in the next section.

2.2 Adapting human Genome Scale metabolic Network and defining transition probabilities

The last two decades have seen an exponential growth of metabolic network reconstructions which are made available through public databases (Wimalaratne *et al.*, 2014) or alongside articles (Hucka *et al.*, 2003). However, for practical reasons, most methods designed to analyze metabolomic results in the context of those networks are database-dependent, restricting for example their use to KEGG (Kanehisa *et al.*, 2014). In order to apply our analysis to networks coming from various sources as well as home-brewed networks, we propose a generic method which can be applied to any network described in the standard SBML format (Hucka *et al.*, 2003) with sufficient information on metabolites.

We applied our method to Recon2 human genome scale metabolic network (Thiele *et al.*, 2013). In this model, metabolites are assigned to cellular compartments (mitochondria, cytoplasm…). Nevertheless, current global and untargeted metabolomics approaches do not provide information on cellular localization of metabolites. Hence, we created a modified version of Recon2 network by considering a metabolite belonging to several compartments as a single metabolite.

Metabolic networks can be turned into graph mathematical formalism by assigning network elements to nodes connected by edges. Several ways to turn a network into a graph exist (Lacroix *et al.*, 2008). We chose to use the compound graph where two metabolites are connected if they are produced and consumed by the same reaction. This formalism allows the integration of information about substrate-product transition on edges.

One of the main issues when analyzing metabolic graphs is the presence of side compounds which are ubiquitous compounds involved in many biochemical reactions for annex purposes such as energy carrier or proton donor. This leads to create edges between a "main" substrate node and a side product (like water) node. When computing paths, these side compounds may cause an underestimation of distances by creating irrelevant shortcuts (Arita, 2004; Holme, 2009). One way to overcome this issue is to remove a set of side compounds based on expert knowledge or using degree threshold (Croes *et al.*, 2005). However, several metabolites considered as side-compounds in most reactions may be implicated as "main" compounds in other processes (typically their own biosynthesis pathway). Systemically removing those compounds will lead to the loss of relevant parts of the network. A more suitable approach consists of comparing molecular structure of substrates and products by using chemical similarity or atom mapping. This approach allows dissociating side compounds from main ones on a chemical basis (Blum and Kohlbacher, 2008; Rahman *et al.*, 2005). In those cases, side compounds are not defined globally for the entire metabolic network, but in the context of each reaction they are involved in.

Hence we applied a pre-processing step on the Recon2 uncompartmentalized metabolic network to avoid irrelevant transitions by computing atom-atom mapping using the Reaction Decoder Tool (Rahman *et al.*, 2016). For a reaction, atom-atom mapping consists of establishing a one-to-one correspondence between the substrate and product atoms. This method requires structural description of the compounds which is encrypted using the SMILE format (Weininger, 1988). Since this information was not available for all metabolites in the Recon2 network, we automatically retrieved this knowledge form chemical databases using web services (PubChem, ChEBI (Davies *et al.*, 2015), HMDB (Wishart *et al.*, 2013)). When only the InChI identifiers (Heller *et al.*, 2015) were available, we converted them into SMILE using the Chemistry Development Toolkit (Steinbeck *et al.*, 2003). Finally, all substrate-to-product transitions that do not involve at least one carbon atom transfer were removed.

The previously described steps allow building the graph on which PPR and PCR will be computed. But the algorithm used to compute PPR considers every outgoing edges of a node to have the same probability to be traversed in the random walk. This model can induce a bias when applied to compound graphs. In fact, using equiprobable edges implies that a reaction with many products will be favored against a reaction consuming the same compound but producing fewer metabolites. For instance, in Figure 1.A, the network contains two reactions R1 and R2. If we consider a compound centric weighting policy, all corresponding edges in the compound graph will have the same probability of 1/3 as shown in Figure 1.B. A more proper way would be to spread probabilities between reactions and then subdivide the probabilities of compound graph edges as it is shown in Figure 1C.

We define the probability policy for substrate-to-product transition as follows, whit m_1 and m_2 as two connected nodes, $w(m_1 \rightarrow m_2)$ the edge weight, $P_r(r)$ the set of all products of reaction r, and R_{m_1} the set of reaction consuming m_1 .

$$P(m_1 \overset{r}{\rightarrow} m_2) = \frac{w(m_1 \rightarrow m_2)}{\sum_{m_i \in P_{r(r)}} w(m_1 \rightarrow m_i)} \times \frac{1}{\|R_{m_1}\|}$$

By applying this probabilistic approach, random walks will go through edges like if computation was performed on the bipartite graph representation of the network. However, this policy still allows to benefiting from the compound graph capability, namely adding information about substrate-product chemical transitions in the probability computation, such as chemical similarity (Rahman *et al.*, 2005), atom conservation (Blum and Kohlbacher, 2008) or RPAIR tags scoring (Faust *et al.*, 2009). This probability policy also allows considering data obtained on reactions using transcriptomic or proteomic since reaction individual probabilities can be transferred to the compound to compound edges. Finally, this policy can also be applied to multi-graphs where two metabolites can be connected by several edges when a chemical transition can be catalyzed by several enzymes.

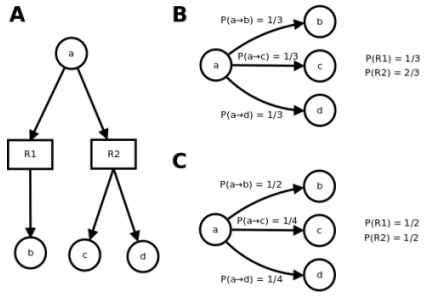


Fig. 1. Effect of the number-of-product bias on compound graph transition probabilities. By overshadowing the reaction levels, seen on the bipartite graph (A), the use of compound graph will favor reactions that involve more product than other consuming reactions (B). The hybrid weighting policy (C) allows suppressing that bias.

The metabolic network adaptation of PPR (resp. PCR) will be called in the following MetaboRank$_{out}$ (resp. MetaboRank$_{in}$), both measures constituting the MetaboRank recommendation system.

2.3 Metabolic fingerprint of Hepatic Encephalopathy

Metabolomics experiments have been conducted on cerebrospinal fluid (CSF) samples from 14 patients suffering from HE against samples from 27 control patients without any proven neurological disease (Weiss *et al.*, 2016). The CSF metabolome was analyzed by LC/MS (Orbitrap-Exactive and Q-Exactive Plus: Positive and negative ESI Scanning from m/z 75 to m/z 1000. Mass resolution: 100 000 FWHM. Data processed using XCMS R package, Filtration according to: Correlation between dilution factor of QC and area. $r^2 > 0.7$, Mean QC/ mean BL>3, CV (QC) < 30%; Feature annotation using public databases, CAMERA R package and ESI-MS and HCD spectral database) and the discriminating fingerprint was built using univariate statistical analyses.

We used the core fingerprint from (Weiss *et al.*, 2016), in order to focus on metabolites confidently identified and presenting the most trustworthy abundance changes. The fingerprint contains metabolites with a relative abundance fold change between patient and control greater than 2 times the standard deviation, a level 1 identification and considered as significant regarding Mann-Withney test (see the full list in supp. table 1).

2.4 Mining literature to corroborate and enrich suggestions

Social-network recommendation system efficiency is commonly evaluated by measuring the number of suggestions followed by a user during future web browsing. Unfortunately this methodology cannot be applied to metabolic recommendation system assessment since recommendations in our case is not part of a decision process but is involved in data interpretation. More generally, assessing the quality of methods providing support to biological interpretation is still a key challenge in the field. In fact, since the disease mechanisms are still partially known, we do not have gold standard datasets (other validated biomarkers) to compare with our recommendation system suggestion.

In order to establish a link between metabolites of interest and HE, we used the Metab2MeSH tool (Sartor *et al.*, 2012). MeSH (Medical Subject Headings) is a controlled vocabulary thesaurus hierarchically structured used to index scientific publications from the MedLine and PubMed databases. Metab2MeSH performs enrichment analysis to identify MeSH terms significantly associated with metabolite names, based on their occurrences in scientific publications. Compounds names from the Recon2 model were converted to PubChem entry names using the Chemical Translation Service web service (Wohlgemuth *et al.*, 2010) and used to retrieve associated MeSH terms using the Metab2MeSH web service.

3 Results

Based on HE core fingerprint, MetaboRank$_{out}$ and MetaboRank$_{in}$ were computed on all metabolites of the metabolic graph resulting in two ranked lists (see supp. table 2). In the following we will focus on compounds ranked in top 50th of MetaboRank$_{out}$ and MetaboRank$_{in}$. The union of these two lists contains 72 compounds and will be called in the following "suggestion list".

To assess the quality of this suggestion list, we compared it to a list obtained by performing an automatic literature search. 38 compounds in recon2 were associated with the MeSH term “hepatic encephalopathy” (MeSH id D006501) (See Figure 2). 10 of them were found in the original metabolic fingerprint. 4 others were found in the MetaboRank$_{in}$ 50th top ranked compounds, and 8 were found in the MetaboRank$_{out}$ 50th top ranked compounds. Overall, the suggestion list allowed enriching the original fingerprint with 10 compounds known in the literature to be related to the disease.

Among the remaining 18 HE-related compounds found in the network, 4 were completely disconnected from the fingerprint, meaning that no single path can be found in the metabolic graph between them and any compound from the fingerprint (see grey lines in Figure 2). The ammonium cannot be reached because we consider only substrate-to-product transitions that involve carbon atom transfers (see method section).

The presence of the D-forms of aspartate, ornithine and arginine in the literature based HE-related list might be due to erroneous compound annotations in literature, as they are very rare in nature. In fact, D and L-forms of those compounds match exactly the same HE-related publications in Pubmed, where the chirality is rarely specified.

The literature based HE-related list may also contain molecules associated with the disease which are not involved in the pathogenicity, but rather mentioned in literature for their intake effect on HE patients. Therefore, the endogenous form of the molecule might not be part of the modulated metabolic mechanism, and won’t be related to the fingerprint. This could be the case for compounds like benzoate and Diazepam present in the Recon2 network and absent from our suggestion list. Sodium benzoate has been used for HE treatment (Misel *et al.*, 2013) in order to activate an alternative pathway of waste nitrogen removal. Diazepam overdose has been shown to induce progressive encephalopathy (Rupasinghe and Jasinarachchi, 2011), and the administration of benzodiazepine medication to cirrhotic patient has been suggested contributing to neurological impairment (Perney *et al.*, 1998) Diazepam has been detected in CSF samples, but has not been included in the HE

metabolic fingerprint (Weiss et al., 2016). The input fingerprint was built from cirrhotic patient, mainly due to alcohol consumption. This drug-induced form of the disease would likely yield a different metabolic fingerprint and could explain why this compound is not suggested by the recommendation system. Finally, HE-related compounds might be missing from the suggestion list because of network incompleteness or too sparse fingerprint.

Recon2 compounds tagged with MeSH term "hepatic encephalopathy"	FP (Fingerprint)	FP + Top MetaboRank$_{in}$	FP + Top MetaboRank$_{out}$	MetaboRank$_{in}$	MetaboRank$_{out}$
5-Hydroxyindoleacetate	✓	✓	✓	Seed	Seed
L-Citruline	✓	✓	✓	Seed	Seed
L-glutamate	✓	✓	✓	Seed	Seed
L-glutamine	✓	✓	✓	Seed	Seed
L-methionine	✓	✓	✓	Seed	Seed
L-phenylalanine	✓	✓	✓	Seed	Seed
L-tryptophan	✓	✓	✓	Seed	Seed
L-tyrosine	✓	✓	✓	Seed	Seed
O-acetylcarnitine	✓	✓	✓	Seed	Seed
octanoyl carnitine	✓	✓	✓	Seed	Seed
L-ornithine		✓		**45**	82
Taurine		✓		**38**	167
5-hydroxytryptophol		✓	✓	**22**	**37**
L-arginine		✓	✓	**50**	**42**
2-oxoglutarate			✓	54	**34**
L-aspartate			✓	56	**27**
L-Carnitine			✓	60	**49**
L-dopa			✓	478	**17**
serotonin			✓	74	**24**
tyraminium			✓	994	**23**
bilirubin				727	1070
choline				299	788
creatinine				NC	776
D-aspartate				216	NC
Diazepam				298	NC
lipoate				806	969
lithocholate				223	559
N-acetyl-L-aspartate				86	85
N-acetyl-L-cysteine				73	117
pantetheine				378	258
phenylacetate				778	139
quinolinate				NC	325
Thiamine diphosphate				748	477
Urea				973	297
D-ornithine				NC	NC
D-arginine				NC	NC
benzoate				NC	NC
ammonium				NC	NC

Fig. 2. 2D-rank of HE related compounds found in Recon2. HE related compounds were found using Metab2Mesh tool. 10 of them were present in the input list obtained from LCR metabolomic profile (blue cells), 10 others were present in the list of recommendations (union of top 50 MetaboRank$_{out}$ and MetaboRank$_{in}$) (orange cells). The light grey cells contain compounds that are disconnected from the input list (NC), dark grey cells contain compounds that have been removed from the network.

Some metabolites in the suggestion list may not yet be mentioned in literature focusing on HE but they may be present in articles mentioning symptoms or diseases related to HE. To address this issue and thus enlarge the scope of our interpretation, we performed the literature analysis the other way around, starting from compounds in the suggestion list to decipher the ones that are not yet associated to HE in the literature but which could be related to health impacts and symptoms strongly associated with HE. MeSH terms related to HE were extracted using a similarity metric which consider the number of co-occurrences between MeSH terms compared with an expected number of co-occurrences appearing "by chance" (Smalheiser and Bonifield, 2016). Only MeSH terms with an odds ratio above 3 were considered. Figure 3 shows main MeSH terms from categories: diseases, signs and symptoms associated with HE. By overlaying the suggestion list onto this graph (size of nodes in Figure 3) it appears that brain, liver and metabolic diseases are the main categories of diseases related to HE.

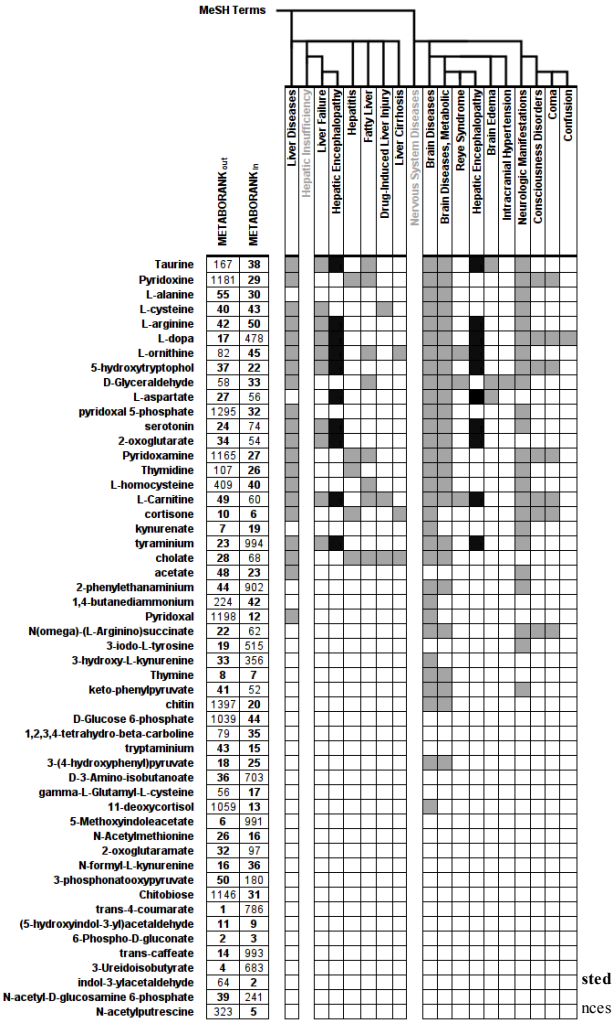
Fig. 3. Suggested compounds mapped onto HE-related disease MeSH subnetwork. Nodes represent Mesh terms. Edges represent tie in the MeSH ontology. The strength of the association with HE is represented as the label font size. Node size represents the number of suggested compounds associated with the corresponding term and/or sub-term. For readability purpose the whole relationships of the MeSH ontology are not represented, only shortest path between each terms are considered.

Figure 4 shows in more details how the 53 metabolites of the suggestion list annotated with at least one MeSH term (see supp. material 3).are related to liver and nervous system diseases and symptoms.

The largest part of the suggestion list is associated with terms belonging to the "brain diseases" category (31) and "liver diseases" category (21) in which HE is classified. 20 compounds were found associated with both liver and brain diseases. Fisher exact test reveals that the high-ranked list is significantly associated with brain and liver disease groups (α=0.01).

By looking to more detailed levels of the MeSH thesaurus in Figure 4, we can see that 4 compounds were associated with MeSH terms related to HE symptoms: coma, confusion and consciousness disorders. One is also associated with intracranial hypertension and brain oedema which often occurs in HE patients.

Few compounds were significantly associated with "liver failure" tagged articles. However, many are overrepresented in corpus related to diseases causing the liver failure, and by extension causing HE: 5 were found associated with hepatitis, 5 with cirrhosis and other alcohol-related diseases.



between the compound name and the MeSH annotation in PubMed, defined accordingly to Smalheiser and Bonifield's metric, with an odds ratio threshold of 3. Only suggested compounds that are found by Metab2Mesh tool are represented.

Besides association with pathological status, the suggestion list is more generally associated with organs and cellular types (astrocytes, neurotransmitters, blood-brain barrier) which play a central role in the HE (additional MeSH terms from chemical and anatomy categories are provided in supp. table 3).

Regarding association with chemicals related MeSH, 15 were associated with MeSH terms Glutamine, Glutamic acid or Ammonia, which are suggested to play a central role in the pathogenicity of the disease. 15 (6 more) were associated with molecules used as treatment (Branched-chain amino-acids, Lactulose and sodium benzoate). One (plus one also associated with HE) is associated with bilirubin which is a marker of liver failure, the main cause of the HE.

Some suggested metabolites are of particular interest in the context of HE. Kynurenic acid, (glutamate receptor antagonist), synthesis is inhibited by hyperammonemia, it has been suggested to exacerbate neuroexcitatory effect of ammonia in HE (Albrecht and Jones, 1999). Kynurenic acid (kynurenate in the metabolic network) was not added to the original HE fingerprint because of mass spectrometry detection limit issues, and consequently a high variability of fold changes between patient and control group. However, a closer look to raw data shows a clear homogeneity in the control group and suggests deregulation specific to the HE patient group, corroborated by Mann & Whitney test (p-value<0.0001) (Fig. 5). Moreover, N-Ω-Hydroxyarginine and N-Ω-L-Arginosuccinate are both involved in the pathway arginine-nitric oxide. It has also been suggested that the N-Ω-Hydroxyarginine inhibit the arginase that produce ornithine and urea from arginine. Finally, Serotonin (PR 24, CR 72) (Knell *et al.*, 1974) and noradrenaline have been shown to increase extracellularly and could be related to the early neuropsychiatric symptom of HE (Shawcross and Jalan, 2005). However, none of these metabolites was detected using our LC/MS methods.

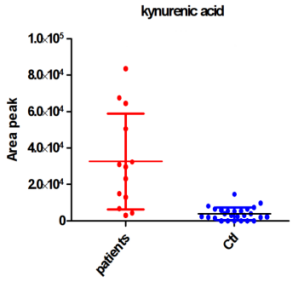


Fig. 5. Kynurenic acid concentrations (arbitrary units) in CSF samples from HE and control patients. Data were obtained by LC/MS using a HILIC column and ESI mass spectrometry detection in the negative mode. Kynurenic acid identification level 1 (*i.e.*, the same chromatographic retention time, accurate measured mass and MS/MS spectrum as those of the reference compound).

Some other suggested compounds appear to be of interest despite no significant association with relevant MeSH terms. For example, α-Ketoglutaramate, which has been found in CSF of patients with hepatic coma, and has been suggested as a biomarker of HE (Halámková *et al.*, 2012). However, very few studies investigated its mechanism, explaining the lack of association with disease MeSH term. This metabolite, as well as the enzyme producing it, have respectively been described as "overlooked" and "underappreciated, but important", regarding HE and other hyperammonemic diseases (Cooper and Kuhara, 2014).

4 Discussion

PR based methods have recently gained much interest for protein-protein interaction networks. For instance, the Protein Rank (Freschi, 2007) is designed for protein functional annotation, using PPR favoring protein with a selected function as random walk seeds. Another example is the SubNet approach, providing a sub-network extraction from interaction network based on PR scoring (Zhang and Zhang, 2013). It uses a "global" PR implementation where random walks can start from any compounds, but favor starts from nodes of interest by adding a constant parameter to bias the damping. Previously, Iván and Grolmusz also proposed to use PPR for protein-protein interaction networks (Iván and Grolmusz, 2011) and successfully retrieved cancer related proteins from proteomics data of melanoma patients. A similar approach, known as GeneRank (Morrison *et al.*, 2005), has been proposed for gene candidate priorization, using gene correlation networks.

In contrary to protein or gene association networks, centrality-based method has been far less applied to metabolic networks. Faust and colleagues also proposed a random walks based approach (available through the NeAT web server (Brohée *et al.*, 2008)) to extract relevant sub-networks from metabolic networks (Faust *et al.*, 2010), using reac-

tant-pair information to avoid side compounds (RPAIR (Kotera *et al.*, 2004)). However this method is mainly focused on KEGG networks and assumes that the list of input metabolites is complete (as it considers only walks linking them) and therefore serves a different purpose than the method proposed here.

The closest implementation was introduced by Bánky et al. who also used PPR (Bánky *et al.*, 2013). However the computation is done on reaction graph, and is dedicated to protein target identification. They avoid overscoring hubs by dividing the PPR by the degree of the node. We chose to divide by the global PR, because the carbon exchange rule drastically changes the topology of the network and makes the degree less straightforward to interpret. Finally, they don't use the PCR thus potentially missing upstream metabolites which could be of interest for the interpretation.

Our method is based on a PPR implementation (also known as PageRank with prior) combined with a PCR for precursor suggestion. To our knowledge this is the first method that allows identifying potential precursors since most previous work was limited to PPR. Since our method is focused on metabolic networks, we also added a network preprocessing and a custom stochastic matrix to avoid metabolic network pitfalls, namely side compounds shortcuts and reactions number of product bias. To our knowledge, MetaboRank is the first recommendation system for interpretation of list of metabolites and the first use of the two-dimensional PPR-PCR computation applied to metabolic networks.

The damping factor parameter used during the computation is usually chosen empirically, and most applications follow the suggestion of 0.85 from the original paper by Brin and Page (Page *et al.*, 1999). Some studies designed to reveal the impact of the damping factor choice on the ranking of web pages, suggest that the algorithm is not excessively sensitive to the variation of the damping factor (at least on web graphs), and that the value of 0.85 seems appropriate when avoiding false negative constitute a priority (Boldi *et al.*, 2005). Unfortunately, it has never been assessed on metabolic networks and there is no clear recommendation for this type of network. Intuitively, we can see that choosing a low damping factor will decrease the likelihood of encountering long walks. The lack of consensus for an appropriate length of a metabolic pathway complicates the definition of a criterion for choosing the most appropriate damping factor. However, we have shown that the default value proposed in the original paper was still sufficient to obtain meaningful suggestions well related to HE.

One criticism against topological methods applied to metabolic networks is the incompleteness and erroneousness of those networks. Metabolic network content is likely to change over time as reactions are continually edited, removed or added during manual curation loops (Thiele and Palsson, 2010). The PageRank seems to be relevant for dealing with this instability since it has been claimed to be more robust to small changes in the network topology (Ng *et al.*, 2001), thanks to the damping process that obfuscates less relevant parts of the network (far from the nodes of interest).

5 Conclusion

MetaboRank is a new method to interpret metabolic fingerprints obtained from metabolomic experiments, in the form of a recommendation system. Several adjustments to the original PageRank approach to ensure the biological relevance of obtained results. MetaboRank suggested metabolites that could be related to the disease, from which several were confirmed by the literature. In particular, MetaboRank predicted α-Ketoglutaramate as a metabolite that should be added to the fingerprint of hepatic encephalopathy, thus suggesting that getting confidence in metabolic fingerprints can provide new insight on complex diseases.

Notably, obtained results show great value for the interpretation of metabolites that were on the edge of significance due to high inter-individual variability. In fact, beside the different level of disease severity between patients, high inter-individual variability may come from pathogenic metabolites involved in highly dynamic processes. This variability makes it difficult to distinguish them from unrelated metabolites, leading to discard them during mechanistic interpretation while still tightly connected to other molecules from the fingerprint. Highly dynamic processes are therefore a key challenge in metabolomic. The recommendation system was able to emphasize two metabolites falling in that case, Taurine and Carnitine, that also appear to play a critical role in the disease according to the literature.

MetaboRank can be applied to metabolomic results from a large range of organisms as it can take any network from a SBML file as input. Furthermore, the proposed mathematical model allows integrating various data at the compound, reaction and reactant pair level. We believe that this method has the potential to facilitate metabolic network exploration by focusing on most relevant metabolites, and could help the elucidation of perturbed metabolic mechanisms and the identification of new drug target. It can also be combined with mechanistic interpretation methods such as pathway enrichment or sub-network extraction. Computed scores can be used as a weighting scheme before subnetwork extraction, such as paths or Steiner Tree computation, which would prior high-scored compounds (Frainay and Jourdan, 2016).

MetaboRank also shows a great potential in the metabolite tedious identification process. In the HE application, some suggested metabolites, like kynurenic acid, had been a posteriori added to the metabolic fingerprint by going back to raw data. Iterative loops between the manual identification from raw data and the suggestion algorithm thus allow refining the metabolic fingerprint and increasing confidence in the mechanistic interpretations inferred from the suggestion list.

This work could be extended by integrating various data at the compound level, reaction level and reactant pair level, using custom transition probabilities based on other omics data or by modifying the topology of the network. The prior vector can also be set to favor some starting nodes among others during the damping phase, based for example on the fold changes obtained from metabolomic results.

Acknowledgements

"PageRank" is a registered trademark of Google. Inc

C.F would like to thank Dr Dima Shepelyansky (Paul Sabatier University, Toulouse, France) and Dr Leonardo Ermann (CNEA, Argentina) for their insightful courses about Google Matrix and fruitful discussions at the Luchon Summer School.

Funding

This work was supported by the French Ministry of Research and National Research Agency as part of the French MetaboHUB, the national metabolomics and fluxomics infrastructure (Grant ANR-INBS-0010) and by PhenoMeNal project, European Commission's Horizon 2020 programme, grant agreement number 654241.

Conflict of Interest: none declared.

References

Albrecht,J. and Jones,E.A. (1999) Hepatic encephalopathy: molecular mechanisms underlying the clinical syndrome. *J. Neurol. Sci.*, **170**, 138–146.

Allesina,S. and Pascual,M. (2009) Googling Food Webs: Can an Eigenvector Measure Species' Importance for Coextinctions? *PLoS Comput. Biol.*, **5**, e1000494.

Altman,T. *et al.* (2013) A systematic comparison of the MetaCyc and KEGG pathway databases. *BMC Bioinformatics*, **14**, 112.

Arita,M. (2004) The metabolic world of Escherichia coli is not small. *Proc. Natl. Acad. Sci.*, **101**, 1543–1547.

Backstrom,L. and Leskovec,J. (2011) Supervised random walks. In, *Proceedings of the fourth ACM international conference on Web search and data mining - WSDM '11*. ACM Press, New York, New York, USA, p. 635.

Bánky,D. *et al.* (2013) Equal Opportunity for Low-Degree Network Nodes: A PageRank-Based Method for Protein Target Identification in Metabolic Graphs. *PLoS One*, **8**, 1–7.

Blum,T. and Kohlbacher,O. (2008) Using atom mapping rules for an improved detection of relevant routes in weighted metabolic networks. *J. Comput. Biol.*, **15**, 565–76.

Boldi,P. *et al.* (2005) PageRank as a function of the damping factor. In, *Proceedings of the 14th international conference on World Wide Web - WWW '05*. ACM Press, New York, New York, USA, p. 557.

Brin,S. *et al.* (1998) The anatomy of a large-scale hypertextual Web search engine. *Comput. Networks ISDN Syst.*, **30**, 107–117.

Brohée,S. *et al.* (2008) NeAT: a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic Acids Res.*, **36**, W444–W451.

Chagoyen,M. and Pazos,F. (2011) MBRole: enrichment analysis of metabolomic data. *Bioinformatics*, **27**, 730–731.

Cooper,A.J.L. and Kuhara,T. (2014) α-Ketoglutaramate: an overlooked metabolite of glutamine and a biomarker for hepatic encephalopathy and inborn errors of the urea cycle. *Metab. Brain Dis.*, **29**, 991–1006.

Cottret,L. *et al.* (2010) MetExplore: a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic Acids Res.*, **38**, W132–7.

Creek,D.J. *et al.* (2014) Metabolite identification: are you sure? And how do your peers gauge your confidence? *Metabolomics*, **10**, 350–353.

Croes,D. *et al.* (2005) Metabolic PathFinding: inferring relevant pathways in biochemical networks. *Nucleic Acids Res.*, **33**, W326–30.

Davies,M. *et al.* (2015) ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res.*, **43**, 612–620.

Ermann,L. and Shepelyansky,D.L. (2015) Google matrix analysis of the multiproduct world trade network. *Eur. Phys. J. B*, **88**, 84.

Faust,K. *et al.* (2009) Metabolic pathfinding using RPAIR annotation. *J. Mol. Biol.*, **388**, 390–414.

Faust,K. *et al.* (2010) Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics*, **26**, 1211–8.

Fiehn,O. (2002) Metabolomics – the link between genotypes and phenotypes. 155–171.

Frainay,C. and Jourdan,F. (2016) Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Brief. Bioinform.*

Freschi,V. (2007) Protein function prediction from interaction networks using a random walk ranking algorithm. In, *2007 IEEE 7th International Symposium on BioInformatics and BioEngineering*. IEEE, pp. 42–48.

Gupta,P. *et al.* (2013) WTF, The Who to Follow Service at Twitter. In, *Proceedings of the 22nd international conference on World Wide Web - WWW '13*. ACM Press, New York, New York, USA, pp. 505–514.

Halámková,L. *et al.* (2012) Enzymatic analysis of α-ketoglutaramate—A biomarker for hyperammonemia. *Talanta*, **100**, 7–11.

Haveliwala,T.H. and H.,T. (2002) Topic-sensitive PageRank. In, *Proceedings of the eleventh international conference on World Wide Web - WWW '02*. ACM Press, New York, New York, USA, p. 517.

Heller,S.R. *et al.* (2015) InChI, the IUPAC International Chemical Identifier. *J. Cheminform.*, **7**, 23.

Hocher,B. and Adamski,J. (2017) Metabolomics for clinical use and research in chronic kidney disease. *Nat. Rev. Nephrol.*, **13**, 269–284.

Holme,P. (2009) Model validation of simple-graph representations of metabolism. *J. R. Soc. Interface*, **6**, 1027–34.

Hucka,M. *et al.* (2003) The systems biology markup language (SBML): a medium for representation and exchange of biochemical network models. *Bioinformatics*, **19**, 524–31.

Iván,G. and Grolmusz,V. (2011) When the web meets the cell: Using personalized PageRank for analyzing protein interaction networks. *Bioinformatics*, **27**, 405–407.

Jeong,H. *et al.* (2000) The large-scale organization of metabolic networks. *Nature*, **407**, 651–4.

Kanehisa,M. *et al.* (2014) Data, information, knowledge and principle: back to metabolism in KEGG. *Nucleic Acids Res.*, **42**, D199–205.

Knell,A.J. *et al.* (1974) Dopamine and serotonin metabolism in hepatic encephalopathy. *Br. Med. J.*, **1**, 549–51.

Kotera,M. *et al.* (2004) RPAIR: a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Informatics*, **15**, 62.

Lacroix,V. *et al.* (2008) An introduction to metabolic networks and their structural analysis. *IEEE/ACM Trans. Comput. Biol. Bioinform.*, **5**, 594–617.

Liang,B. *et al.* (2014) Searching for people to follow in social networks. *Expert Syst. Appl.*, **41**, 7455–7465.

Liben-Nowell,D. and Kleinberg,J. (2007) The link-prediction problem for social networks. *J. Am. Soc. Inf. Sci. Technol.*, **58**, 1019–1031.

Ma,N. *et al.* (2008) Bringing PageRank to the citation analysis. *Inf. Process. Manag.*, **44**, 800–810.

Mihalcea,R. *et al.* (2004) PageRank on semantic networks, with application to word sense disambiguation. In, *Proceedings of the 20th international conference on Computational Linguistics - COLING '04*. Association for Computational Linguistics, Morristown, NJ, USA, p. 1126–es.

Milreu,P.V. *et al.* (2014) Telling metabolic stories to explore metabolomics data: a case study on the yeast response to cadmium exposure. *Bioinformatics*, **30**, 61–70.

Misel,M.L. *et al.* (2013) Sodium benzoate for treatment of hepatic encephalopathy. *Gastroenterol. Hepatol. (N. Y).*, **9**, 219–27.

Mo,M.L. and Palsson,B.Ø. (2009) Understanding human metabolic physiology: a genome-to-systems approach. *Trends Biotechnol.*, **27**, 37–44.

Morrison,J.L. *et al.* (2005) GeneRank: using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, **6**, 233.

Neumann,S. and Böcker,S. (2010) Computational mass spectrometry for metabolomics: Identification of metabolites and small molecules. *Anal. Bioanal. Chem.*, **398**, 2779–2788.

Ng,A.Y. *et al.* (2001) Stable algorithms for link analysis. In, *Proceedings of the 24th annual international ACM SIGIR conference on Research and development in information retrieval - SIGIR '01*. ACM Press, New York, New York, USA, pp. 258–266.

Page,L. *et al.* (1999) The PageRank Citation Ranking: Bringing Order to the Web Stanford, CA, USA.

Perney,P. *et al.* (1998) Plasma and CSF benzodiazepine receptor ligand concentrations in cirrhotic patients with hepatic encephalopathy: relationship to severity of encephalopathy and to pharmaceutical benzodiazepine intake. *Metab. Brain Dis.*, **13**, 201–10.

Persicke,M. *et al.* (2012) MSEA: metabolite set enrichment analysis in the MeltDB metabolomics software platform: metabolic profiling of Corynebacterium glutamicum as an example. *Metabolomics*, **8**, 310–322.

Rahman,S.A. *et al.* (2005) Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics*, **21**, 1189–1193.

Rahman,S.A. *et al.* (2016) Reaction Decoder Tool (RDT): extracting features from chemical reactions. *Bioinformatics*, **32**, 2065–6.

Rupasinghe,J. and Jasinarachchi,M. (2011) Progressive encephalopathy with cerebral oedema and infarctions associated with valproate and diazepam overdose. *J. Clin. Neurosci.*, **18**, 710–711.

Sartor,M.A. *et al.* (2012) Metab2MeSH: annotating compounds with medical subject headings. *Bioinformatics*, **28**, 1408–10.

Shawcross,D. and Jalan,R. (2005) The pathophysiologic basis of hepatic encephalopathy: Central role for ammonia and inflammation. *Cell. Mol. Life Sci.*, **62**, 2295–2304.

Smalheiser,N. and Bonifield,G. (2016) Two Similarity Metrics for Medical Subject Headings (MeSH): An Aid to Biomedical Text Mining and Author Name Disambiguation. *J. Biomed. Discov. Collab.*, **7**, e1.

Sreekumar,A. *et al.* (2009) Metabolomic profiles delineate potential role for sarcosine in prostate cancer progression. *Nature*, **457**, 910–914.

Steinbeck,C. *et al.* (2003) The Chemistry Development Kit (CDK): an open-source Java library for Chemo- and Bioinformatics. *J. Chem. Inf. Comput. Sci.*, **43**, 493–500.

Swainston,N. *et al.* (2016) Recon 2.2: from reconstruction to model of human metabolism. *Metabolomics*, **12**, 109.

Thiele,I. *et al.* (2013) A community-driven global reconstruction of human metabolism. *Nat. Biotechnol.*, **31**, 419–25.

Thiele,I. and Palsson,B.Ø. (2010) A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nat. Protoc.*, **5**, 93–121.

Weininger,D. (1988) SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.*, **28**, 31–36.

Weiss,N. *et al.* (2016) Cerebrospinal fluid metabolomics highlights dysregulation of energy metabolism in overt hepatic encephalopathy. *J. Hepatol.*, **65**, 1120–1130.

Weiss,N. *et al.* (2017) Understanding hepatic encephalopathy. *Intensive Care Med.*

Wimalaratne,S.M. *et al.* (2014) BioModels linked dataset. *BMC Syst. Biol.*, **8**, 91.

Wishart,D.S. *et al.* (2013) HMDB 3.0--The Human Metabolome Database in 2013. *Nucleic Acids Res.*, **41**, D801–7.

Wohlgemuth,G. *et al.* (2010) The Chemical Translation Service--a web-based tool to improve standardization of metabolomic reports. *Bioinformatics*, **26**, 2647–8.

Zhang,Q. and Zhang,Z.D. (2013) SubNet: a Java application for subnetwork extraction. *Bioinformatics*, **29**, 2509–2511.

Zhirov,A.O. *et al.* (2010) Two-dimensional ranking of Wikipedia articles. *Eur. Phys. J. B*, **77**, 523–531.

8.3 Discussion

8.3.1 Pertinence des recommandations

Table 8.2 – Liste des recommandations issues de la signature métabolique de l'encéphalopathie hépatique

Suggestions	$MetaboRank_{out}$	$MetaboRank_{in}$
(2R)-2-hydroxy-3-(phosphonatooxy)propanoate	30	240
(5-hydroxyindol-3-yl)acetaldehyde	11	9
(alpha-D-mannosyl)2-beta-D-mannosyl-N-acetylglucosamine	1399	28
1,2,3,4-tetrahydro-beta-carboline	79	35
1,4-butanediammonium	224	42
1-(1,2,3,4,5-pentahydroxypent-1-yl)-1,2,3,4-tetrahydro-beta-carboline-3-carboxylate	25	11
11beta-Hydroxyandrost-4-ene-3,17-dione	9	992
11-deoxycortisol	1059	13
2,3-diketo-5-methylthio-1-phosphopentane	807	34
2-keto-4-methylthiobutyrate	815	14
2-oxoglutaramate	32	97
2-oxoglutarate	34	54
2-phenylethanamin	44	902
3-(4-hydroxyphenyl)pyruvate	18	25
3-disulfanyl-L-alanine	31	995
3-hydroxy-L-kynurenine	33	356
3-iodo-L-tyrosine	19	515
3-phosphonatooxypyruvate	50	180
3-Ureidoisobutyrate	4	683
4-(2-aminophenyl)-2,4-dioxobutanoate	12	39
4-acetamidobutanal	456	1
4-coumaroyl-CoA	21	585
5-hydroxy-N-formylkynurenine	3	989
5-hydroxytryptophol	37	22
5-L-Glutamyl-L-alanine	132	8
5-Methoxyindoleacetate	6	991
6-O-phosphonato-D-glucono-1,5-lactone	1043	18
6-Phospho-D-gluconate	2	3
acetate	48	23
apelin (1-12)	47	49
apelin-13	38	37
beta-1,4-mannose-N-acetylglucosamine	1151	24
chitin	1397	20
Chitobiose	1146	31
cholate	28	68
choloyl-CoA	20	48
cortisone	10	6
D-3-Amino-isobutanoate	36	703
D-Glucose 6-phosphate	1039	44
D-Glyceraldehyde	58	33
D-Glycerate 2-phosphate	29	243
D-Mannose	131	21
D-ribulose 5-phosphate	45	789
gamma-L-Glutamyl-L-cysteine	56	17
Glycylphenylalanine	46	47
indol-3-ylacetaldehyde	64	2
keto-phenylpyruvate	41	52
kynurenate	7	19
L-alanine	55	30
L-arginine	42	50
L-aspartate	27	56
L-carnitine	49	60

Table 8.2 – Liste des recommandations issues de la signature métabolique de l'encéphalopathie hépatique (suite)

Suggestions	$MetaboRank_{out}$	$MetaboRank_{in}$
L-cysteine	40	43
L-dopa	17	478
L-homocysteine	409	40
methyl indole-3-acetate	5	990
N(omega)-(L-Arginino)succinate	22	62
N-(omega)-Hydroxyarginine	13	4
N-acetyl-D-glucosamine 6-phosphate	39	241
N-acetyl-D-mannosamine	35	76
N-acetyl-L-asparagine	15	10
N-Acetylmethionine	26	16
N-acetylputrescine	323	5
N-formyl-L-kynurenine	16	36
Ornithine	82	45
Pyridoxal	1198	12
pyridoxal 5-phosphate	1295	32
Pyridoxamine	1165	27
Pyridoxamine 5-phosphate	1168	46
Pyridoxine	1181	29
Pyridoxine 5-phosphate	1378	41
serotonin	24	74
Taurine	167	38
Thymidine	107	26
Thymine	8	7
trans-4-coumarate	1	786
trans-caffeate	14	993
tryptamine	43	15
tyramine	23	994

Le système de recommandation proposé dans cette thèse a permis d'identifier 79 métabolites candidats pour assister l'interprétation (Table 8.2), à partir des 2592 métabolites du réseau métabolique considéré. Le voisinage direct des 28 métabolites d'intérêt issus de la signature (Table 8.1) comprend 327 métabolites. Son voisinage étendu à une distance de 2 réactions (en omettant 15 métabolites auxiliaires tels que l'eau ou l'ADP en tant qu'intermédiaires), conduit à un sous-réseau comprenant 1306 métabolites, soit environ la moitié du nombre total de métabolites du réseau initial. Ces résultats mettent en évidence le fait que se limiter à l'ensemble des métabolites à proximité des nœuds d'intérêt ne suffit pas pour obtenir des ensembles interprétables. La méthode alternative proposée dans cette thèse permet en revanche de fournir des ensembles réduits et ordonnés pour conduire l'interprétation.

L'analyse automatique de la littérature a permis d'extraire une liste de 38 métabolites d'importance reconnue, d'après leur surreprésentation significative dans

les articles traitants de l'EH. 10 métabolites étaient présents dans la liste de métabolite d'intérêt générée à partir de la signature, et 10 étaient présents dans la liste des recommandations, doublant ainsi la couverture des métabolites d'importance reconnue. Il est a noter que les 18 métabolites importants non recommandés contiennent 3 potentiels faux-positifs : les formes D de l'aspartate, l'arginine et l'ornithine (rares), dont les formes L (fréquentes) ont été recommandées. Ces faux positifs peuvent résulter du fait que la chiralité de ces molécules est rarement précisée étant donné la rareté des formes D, ce qui conduit les auteurs à référer tacitement la forme L sous une dénomination générale. Malheureusement, à ce jour les méthodes d'analyse automatique de textes peinent à déceler les références sous-entendus. La liste de molécules extraite de littérature contient également 3 métabolites dont la surreprésentation résulte probablement de leur usage en tant que traitement (Diazepam, lipoate et benzoate), et constitue de ce fait un cas particulier.

Les métabolites d'intérêt ont été extraits de la signature métabolique (Table 8.1), et représentent les métabolites identifiés dont l'abondance varie significativement entre le groupe constitué de patients et le groupe contrôle, et dont la variation est jugée suffisamment importante. Ces deux caractéristiques reposent sur des critères de sélection relativement arbitraires : un seuil d'erreur alpha fixé à 0,05 et un fold-change supérieur à 2 fois l'écart-type des mesures. Ces critères stringents permettent de sélectionner confidemment des marqueurs de la maladie, afin d'éviter toute interprétation basée sur des variations dues à des facteurs extrinsèques, et de focaliser l'élucidation des mécanismes à partir des acteurs les plus « importants ». Cette dichotomie, « important » ou non, génère cependant une certaine ambiguïté quant aux traitements des métabolites dont les valeurs sont insuffisantes, mais néanmoins très proches de ces seuils. Le système de recommandation a permis dans une certaine mesure de lever cette ambiguïté pour certains de ces métabolites. En effet, il suggère un lien fort entre les métabolites d'intérêt et l'aspartate, la taurine, la carnitine et l'alanine, tous variants de manière significative entre patients et contrôles, mais dont la variation observée était faible relativement à la

variabilité observée. Excepté l'alanine, tous sont surreprésentés dans la littérature associée à l'EH.

La méthode a également mis en évidence des composés dont l'intérêt est supporté par la littérature, mais qui n'avaient pu être mesurés (intensité sous la limite de détection) ou identifiés faute de standards disponibles (c'est notamment le cas du 5-hydroxytryptophol ou de la tyramine), suggérant ainsi un potentiel de la méthode pour pallier certaines limites techniques.

Une étude approfondie des métabolites recommandés a également permis d'aller au-delà des métabolites surreprésentés dans la littérature relative à l'EH, suggérant le potentiel de la méthode pour orienter vers de nouvelles découvertes ou remettre en lumière des connaissances négligées. Par exemple, l'oxoglutaramate, molécule neurotoxique impliquée dans une voie alternative de conversion de la glutamine en oxoglutarate, a été recommandé par notre approche et est supporté par quelques études suggérant son potentiel en tant que biomarqueur de l'EH et son implication dans d'autres maladies liées à l'hyperammonémie, et dont certaines alertent la communauté sur la nécessité d'accorder plus d'intérêt à ce métabolite et à l'enzyme catalysant sa production[56][55].

Le système de recommandation met également en avant l'importance de l'acide kynurenique, dont l'implication dans l'EH n'a pas été élucidée à ce jour. Cependant, de précédentes études ont révélé que l'induction d'hyperammonémie chez des rats inhibe sa synthèse, interférant avec son activité antagoniste des récepteurs du glutamate, ce qui pourrait amplifier les effets neuro-excitatoire de l'ammoniac[6]. Un retour aux données brutes a révélé que ce métabolite n'avait pas été considéré dans les molécules d'intérêt du fait des limites de détection et d'une forte variabilité des *fold-changes* observés entre patients et contrôles, mais a cependant mis en évidence une remarquable homogénéité des valeurs obtenues à partir du groupe contrôle, contrairement au groupe des patients, appuyant l'hypothèse de son implication dans la physiopathologie de la maladie. Cette recommandation va ainsi inciter la conduite d'expérimentations complémentaires pour expliquer l'origine de cette variabilité propre aux patients atteints d'EH.

En guidant l'exploration du réseau humain à l'aide des suggestions, il a été possible de proposer de nouveaux scénarios mécanistiques, globalement centrés autour de l'ammoniac, mais qui s'éloigne des voies métaboliques « classiques » relatives à son métabolisme. On peut citer par exemple la production d'indole-3-acétate et de 5-hydroxyindoleacetate à partir du tryptophane et du 5-hydroxy-tryptophane (tous présentant des abondances anormales chez les patients) via les monoamines-oxydases impliquées dans la production d'ammoniac et d'amines neuroactives, dont la sérotonine, dont il a été suggéré que son accumulation pourrait être responsable des symptômes neuropsychiatriques précoces de la maladie(Figure 8.1).

Un autre exemple est celui de la production de 4-acetamidobutanoate (également présent dans la liste des molécules d'abondance anormale) par l'acétyle-putrescine, impliquant également les monoamines-oxydases. Les intermédiaires impliqués dans ces scénarios sont pour majeure partie issus des recommandations, et ainsi, bien que les analyses automatiques de la littérature ne révèlent pas d'associations particulières de ces derniers avec des thématiques relatives à l'EH, ils permettent un apport sur le plan mécanistique.

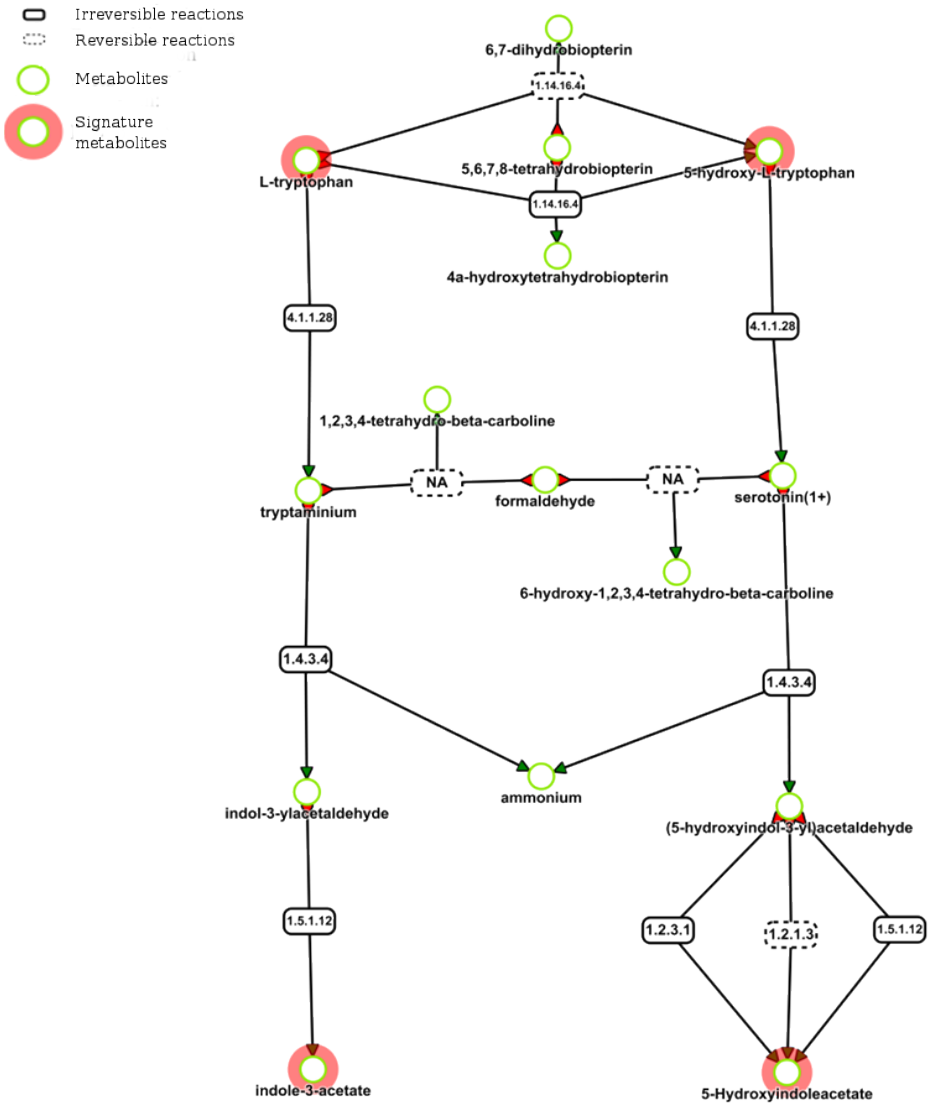


Figure 8.1 – Synthèse de l'indole-3-acétate et du 5-hydroxyindoleacetate à partir du tryptophane et du 5-hydroxy-tryptophane

8.3.2 Spécificité des recommandations

Bien que des méthodes statistiques permettent de mettre en évidence des variations d'abondance de métabolites entre des échantillons soumis à une perturbation et des échantillons contrôles, ils ne permettent pas de statuer sur la spécificité de ces variations à cette perturbation. Certains processus biologiques, comme le métabolisme énergétique ou le métabolisme des acides aminés, semblent être largement représentés dans les signatures métaboliques issues de conditions variées. Il est plausible que certains mécanismes, en raison de leur centralité globale, puissent être affectés par un large ensemble de perturbations dans le réseau. Dès lors, un système de recommandation dont les suggestions seraient focalisées sur ces mécanismes serait d'une pertinence faible, dans la mesure où les mêmes suggestions seraient obtenues, quelle que soit la liste de molécules d'intérêt. Les systèmes de recommandation métaboliques seraient alors également contraints à encourager la sérendipité, en s'écartant de l'évidence pour proposer des suggestions pertinentes. La sérendipité est une notion qui peut être résumée comme le fait de faire des découvertes inattendues. Cet élément représente une composante centrale dans le développement des systèmes de recommandation. En effets certains systèmes de recommandation souffrent de leur incapacité à s'éloigner des suggestions « évidentes » ou trop généralisées pour être pertinentes. On peut par exemple mentionner la limite baptisée « Problème Harry Potter », en référence à la saga de livres dont la grande popularité a conduit des systèmes de recommandation de type *item-based collaborative filtering* (par exemple le système d'Amazon « People who bought this also bought... ») à les proposer à l'ensemble des utilisateurs, quels que soient les livres considérés.

Notre première approche fut d'utiliser la méthode statistique de Monte-Carlo[179] afin de proposer des suggestions qui soient spécifiques de la signature d'entrée. Pour se faire, de nombreux échantillonnages aléatoires de métabolites sont réalisés afin de construire des signatures aléatoires de tailles identiques, qui seront utilisées pour calculer les scores de recommandation. Les recommandations initiales sont alors filtrées pour ne conserver que celles dont le score n'est pas

égalé ou surpassé dans une proportion définie de recommandations issues des signatures aléatoires.

Cette méthode offre l'avantage de calculer une p-valeur, qui constitue un critère aisément interprétable pour la sélection des recommandations. En revanche, cette méthode présente un coût computationnel important du fait de la répétition du calcul de centralité pour un nombre important d'instances. Or, la volatilité des résultats de métabolomique, notamment du fait des ambiguïtés d'identifications, nécessiterait que la méthode puisse être aisément relancée, et les résultats obtenus dans en un temps restreint, afin de pouvoir moduler aisément les données d'entrées pour faciliter la démarche exploratoire. De plus, il est difficile de réaliser un échantillonnage aléatoire qui soit tout à fait représentatif des signatures considérées. En effet, comme énoncé précédemment, certains métabolites du réseau ne sont pas mesurables étant donnée la méthode d'acquisition des données, et devraient donc être exclus des tirages aléatoires. De plus, il est attendu que les éléments de la signature ne soient pas indépendants (c'est d'ailleurs ces dépendances que l'on cherche à mettre en évidence), or, l'emploie de signatures aléatoires réalisées par tirages indépendants pourrait conduire à une estimation faussée de la significativité des recommandations obtenues.

Partant de notre hypothèse selon laquelle la surreprésentation de certains métabolites serait due à leur centralité globale, nous avons opté pour une pénalisation des scores obtenus à partir de la signature, par le score de centralité globale. Cela revient à baser les recommandations sur le rapport entre une centralité considérant les marches émanant des métabolites d'intérêt, par rapport à la même centralité où les marches émanant de n'importe quel métabolite sont considérées. De manière empirique, nous avons observé sur le jeu de données utilisé dans l'article précédent que les métabolites qui ont été exclus de la « top list » par cet ajustement (tels que le pyruvate ou l'acetyl-CoA) correspondent pour la majeure partie aux métabolites dont la p-valeur calculée par approche de Monte-Carlo est supérieure au seuil alpha de 0,05. Cette technique a l'avantage de considérer un score d'ajustement unique et indépendant des signatures, donc calculé une seule fois pour un

même réseau, ce qui la rend très simple à mettre en place et lui confère un temps d'exécution bien moins long.

La problématique de sur-représentation des *hubs* par les méthode de centralité appliquée aux réseaux biologiques a également été identifiée par Banky *et al.*[18] en 2013. Ils préconisent également une pondération du score de PageRank obtenu, mais suggèrent l'utilisation du degré. Ils considèrent cependant une version « brute » du réseau, sans pre-processing lié à la gestion des composés auxiliaires par exemple. Comme mentionné dans leur article, ces modifications affectent de manière substantielle les propriétés du réseau. Notre approche considérant un réseau pondéré ainsi que le retrait des liens ne portant pas d'échanges de carbones, l'interprétation du degré en est profondément affecté et moins triviale. En effet, le degré moyen du réseau Recon2v3 est de 21,125, avec un degré maximal de 3889, alors que notre version sans compartiment, sans molécule non annotée avec un InChI et sans transition non carbonée présente un degré moyen de 5,186 avec un degré maximal de 544.

8.3.3 Stabilité des résultats face aux variations dans le réseau et les données d'entrée

Comme évoqué précédemment la volatilité des résultats de métabolomique et des reconstructions de réseaux nécessitent de la part de la méthode une certaine robustesse du classement vis-à-vis des fluctuations dans ces données.

Les modifications fréquentes des réseaux biologiques peuvent en effet faire échouer la reproduction de résultats obtenus sur des versions antérieures. Cette problématique a par exemple été étudiée par Beber *et al.*[20] (bien que leur étude se focalise principalement sur les réseaux de régulations et la base RegulonDB). La problématique de l'instabilité des réseaux est commune à de nombreux réseaux *real-world*. Par conséquent, la problématique de la robustesse des méthodes de centralité face à des perturbations du type ajout ou délétion de nœuds et d'arêtes a déjà été considérée, notamment par Borgatti *et al.*[38].Le PageRank est particulièrement réputé pour la stabilité des rangs obtenus face à ce type de

perturbations[199][26][24], ce qui a notamment motivé son usage pour les réseaux biologiques.

Concernant la volatilité des données d'entrée et leur impact sur un classement, cette problématique a particulièrement attiré l'attention dans le domaine de la priorisation de gènes candidats[40]. La priorisation de gènes est une problématique très similaire à celle traitée dans cette thèse. Elle consiste à classer les gènes présentant un différentiel d'expression entre plusieurs conditions, identifié en transcriptomique, dans le but d'orienter les analyses complémentaires nécessaires à la compréhension des mécanismes impliqués[192]. Contrairement à notre approche appliquée à la métabolomique, ces méthodes n'ont généralement pas vocation à suggérer des éléments non mesurés pour étendre un profil, mais bien à ordonner les gènes issus d'un profil par ordre de pertinence. Les critères utilisés pour définir cette pertinence reposent généralement sur les données d'expression et les annotations fonctionnelles contenues dans les bases de connaissances. Néanmoins, plusieurs méthodes ont fait usage de la centralité dans les réseaux de régulation de gènes ou d'interactions protéiques pour la priorisation, dont certaines employant le PageRank personnalisé[280][194][224].

Plusieurs approches sont possibles pour mesurer la stabilité des rangs suite aux perturbations. La première est basée sur les ensembles et ne considère que les nœuds les mieux classés. Usuellement, elle est calculée à partir de la taille de l'intersection entre cette top-liste et celle obtenue après perturbation. Cette taille peut également être normalisée par la taille de l'union de ces deux listes. Les limites de ces méthodes sont la définition arbitraire de la taille de la liste à considérer, et le fait que les rangs ne sont utilisés qu'en tant que critère d'appartenance à la « top-liste ». Ainsi, les rangs au sein de la top-liste ne sont pas pris en compte, et si l'on choisit $k = 50$ la taille de la top-liste, le passage d'un élément de la position 50 à 51 affecte la stabilité de la même manière que le passage d'un élément de la position 1 à la dernière position.

La seconde approche est basée sur le calcul de distance entre deux classements, et les mesures les plus classiquement utilisées sont le coefficient de corrélation de

Spearman et l'indice τ de Kendall[145], qui considèrent pour chaque élément la discordance ou la différence des rangs qui leur sont attribués dans deux classements différents. Prendre en compte l'ensemble des éléments ordonnés constitue néanmoins une limite : les nombreux éléments peu pertinents situés en fin de classement, non considérés dans les analyses, présentent des espérances de passage proche de 0 et peuvent être soumis à de grandes variations de rang causées par des infimes variations de score, affectant particulièrement ces mesures. Certains auteurs ont notamment souligné le fait que de nombreuses mesures de centralité sont optimisées pour identifier les acteurs clés d'un réseau, et ne sont pas nécessairement adaptées pour interpréter le rôle des nœuds ne faisant pas partie des plus influents[165]. Cette remarque peut justifier l'usage des méthodes basées sur les ensembles qui se limitent à la liste des nœuds qui vont en pratique être considérés pour les analyses suivantes.

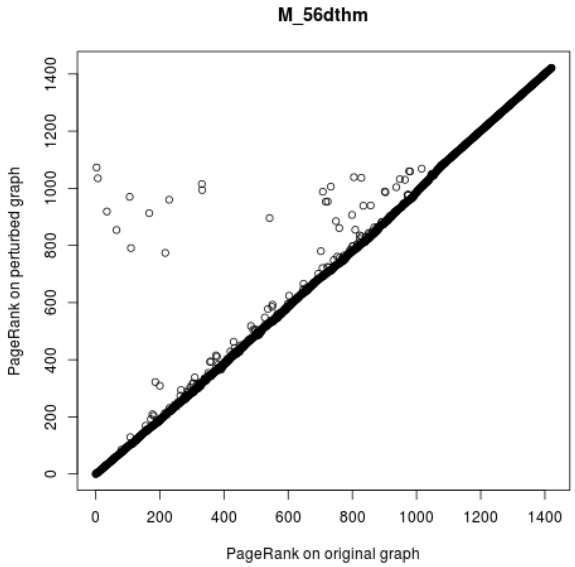
La stabilité des résultats obtenus a été estimée en recalculant les rangs à partir de listes de nœuds d'intérêt où l'un des métabolites de la signature a été retiré. 28 listes de 27 métabolites ont donc été crées à partir de la liste originale de 28 métabolites. En moyenne, la top-liste des 50 premières suggestions est conservée à 96% pour le PageRank et à 95,7% pour le CheiRank après retrait d'un métabolite. La plus basse intersection observée est respectivement de 90% et 88% pour le PageRank et le CheiRank. Comme il peut être observé sur les graphiques suivants, les métabolites de la liste d'intérêt n'impactent pas de la même manière les rangs obtenus après leur retrait (Figure 8.2).

On peut observer que certaines suggestions dépendent essentiellement d'un métabolite particulier de la liste d'intérêt, dont ils sont généralement un voisin direct. Par exemple, la 5,6-dihydrothymine, présentant une abondance anormale chez les patients, n'est impliquée que dans deux réactions : une réaction réversible produisant de la thymine, et une réaction irréversible produisant du 3-Ureidoisobutyrate. Ces deux métabolites se partagent donc la totalité de l'influence transférée depuis la 5,6-dihydrothymine, et apparaissent ainsi dans la liste des recommandations. Le 3-Ureidoisobutyrate (4e position de PageRank) ne peut être produit qu'à partir

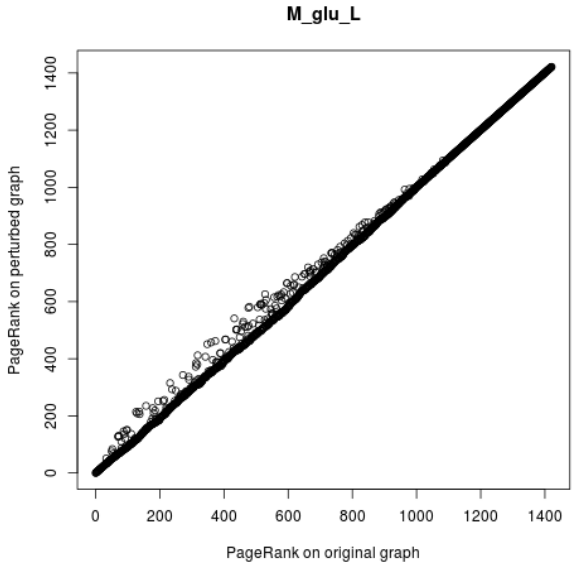
de la 5,6-dihydrothymine, ce qui renforce l'hypothèse selon laquelle son abondance pourrait également être affectée. Cependant, le retrait de la 5,6-dihydrothymine de la liste d'intérêt fait considérablement chuter son rang à la 1074e position, étant donné que la 5,6-dihydrothymine ne reçoit que très peu d'influence de la part des autres molécules de la signature. L'impact de son retrait se distingue ainsi de celui du glutamate, qui, bien que globalement très central et influençant donc une large partie du réseau, reçoit de nombreux « *feedback* » de la part des autres molécules de la signature, compensant son retrait (Figure 8.2).

L'oxaloglutarate quant à lui, conserve sa position dans la liste des recommandations, quel que soit le métabolite d'intérêt retiré. Ce métabolite est en effet en lien avec un grand nombre de métabolites de la liste, eux-mêmes fortement interconnectés, renforçant leurs influences mutuelles. Néanmoins, contrairement au 3-Ureidoisobutyrate, il est impliqué dans un grand nombre de réactions qui pourraient compenser les perturbations affectant ses précurseurs issus de la liste d'intérêt.

L'analyse de la stabilité des rangs offre donc également de précieuses informations pour l'interprétation des recommandations. Elle permet de distinguer deux rôles au sein des métabolites d'importances : ceux appartenant à une communauté impliquant de nombreux métabolites d'abondance anormale, peu affectés par une modification de la liste, et ceux agissant comme des intermédiaires privilégiés d'un métabolite d'intérêt particulier, fortement affectés par le retrait de ce dernier.



(a) Retrait de la 5,6-dihydrothymine



(b) Retrait du glutamate

Figure 8.2 – Effet du retrait d'un métabolite de la signature sur les rangs utilisés pour les recommandations.

8.3.4 Alternatives au PageRank

Une approche concurrente au PageRank dans le domaine des moteurs de recherche est la méthode des « *hubs and authority* », avec laquelle l'approche PageRank/CheiRank partage une certaine analogie[151]. Dans le contexte du Web, un hub est un page qui référence de nombreuses pages d'intérêt, et une autorité est une page référencée par de nombreuses pages d'intérêt. La qualité d'une autorité est tributaire du référencement par de nombreux hubs, et la qualité d'un hub est tributaire du référencement de nombreuses autorités.

Ainsi, dans l'algorithme HITS (*Hyperlink-Induced Topic Search*), les scores sont calculés itérativement, partant de vecteurs initiaux avec des scores de hub et d'autorité de 1 pour l'ensemble des nœuds à classer, puis mis à jour récursivement jusqu'à convergence. Ce calcul revient à calculer par la méthode de la puissance itérée le vecteur principal de la matrice AA^T pour le score de hub et A^TA pour le score d'autorité, ou A correspond à la matrice d'adjacence. La principale variante de l'HITS est l'algorithme SALSA (*Stochastic Approach for Link-Structure Analysis*) qui utilise une matrice stochastique à la place de la matrice d'adjacence, s'appuyant ainsi sur les mêmes fondations théoriques que le PageRank.

La principale différence de ces approches avec le PageRank est qu'elles utilisent une liste d'éléments d'intérêt à ordonner, définie à partir d'informations exogènes (notamment l'occurrence des termes de la requête dans les pages considérées), étendue à leur voisinage direct. Les scores sont donc calculés uniquement pour les éléments de ce *focused subgraph*. Contrairement à notre approche, ces méthodes ont donc vocation à ordonner et catégoriser (hub ou autorité) une liste d'éléments d'intérêt obtenue de manière exogène. L'approche du personalized PageRank/CheiRank quant à elle a vocation à faire émerger de nouveaux éléments par expansion d'une liste d'éléments d'intérêt déjà connus, en calculant un score pour l'ensemble du graphe. Les méthodes HITS et SALSA peuvent néanmoins potentiellement offrir un intérêt pour l'analyse d'une signature métabolique, en proposant une liste de sources (hubs) et de cibles (autorités) au sein de cette dernière, afin d'appliquer des méthodes d'extraction de sous réseaux basés sur la

recherche de chemins. Ces méthodes étant profondément enracinées dans la thématique des moteurs de recherche, la limite qui demeure est la définition d'un *focused subgraph* pertinent d'un point de vue biologique, ainsi que l'interprétation des notions de hub et d'autorité dans ce contexte.

8.3.5 Limite : La nécessité d'informations structurales

La méthode proposée considère une version pondérée du réseau métabolique, où la force d'une interaction entre deux molécules dépend du nombre d'atomes échangés. En revanche, le calcul de ces échanges, réalisé à l'aide d'outils chemo-informatiques d'atom-mapping, nécessite des informations relatives à la structure moléculaire des métabolites. Ces informations sont rarement fournies dans les reconstructions de réseaux, ou de manière partielle, ce qui a nécessité le développement d'outils pour l'annotation automatique des réseaux[25].

Les structures des molécules de petite taille peuvent être représentées sous différents formats. Une première famille de formats, dont le format molfile est peut-être le représentant le plus emblématique, représente les molécules au moyen de tables de connexions, agrémentées d'informations complémentaires comme la nature des liaisons ou les coordonnées atomiques de tous les atomes de la molécule dans l'espace. Ces formats offrent une description précise de la structure d'une molécule et sont relativement simples à traiter informatiquement. En revanche, ces représentations sont très volumineuses. Dans le cas du molfile par exemple, la description d'un atome seul requiert 69 caractères.

L'autre famille de format est constituée des représentations en 2D des molécules, sous forme de chaîne de caractères linéaires. Elles offrent l'avantage d'être bien plus compacte que les représentations précédentes. Grâce à cela, elles ont notamment été utilisées en tant qu'identifiant de molécule afin de faciliter leurs indexations. Ces caractéristiques ont conduit à populariser leur usage dans les bases de données de composés chimiques ainsi que, dans une moindre mesure, leur ajout en tant qu'attribut dans les réseaux métaboliques. Les deux principaux membres de cette famille sont les InChI et les SMILES.

Les SMILES (Simplified Molecular-Input Line-Entry System)[266] offrent l'avantage d'être relativement lisible par un humain. Néanmoins, cette représentation ne permet pas de discriminer tous les composés (notamment les énantiomères). De plus, plusieurs SMILES peuvent exister pour une même molécule. Il existe toutefois des méthodes de calcul de SMILES canoniques permettant de pallier ce dernier point[203]. Les SMILES permettent également d'intégrer des radicaux dans la représentation chimique.

Les InChI (IUPAC International Chemical Identifier)[116]ont quant à eux la particularité de proposer différents niveaux d'information, permettant d'intégrer entre autres la connectivité, les isotopes, la stéréochimie ; et ainsi de mieux discriminer les molécules que les SMILES lorsque tous les niveaux sont renseignés[25]. En revanche, les InChI ne permettent pas d'intégrer des radicaux, ce qui peut être problématique pour des molécules de grande taille. Plus complexe, cette représentation rend également difficile la reconstruction manuelle de la structure contrairement aux SMILES[25].

De nombreuses bases de données de composés chimiques font figurer les InChI ou les SMILES dans leurs entrées, et certaines entrées des réseaux métaboliques sont annotées avec des références vers ces bases. Ces éléments ont conduit à la stratégie représentée figure 8.3 pour l'extraction des InChI/SMILES.

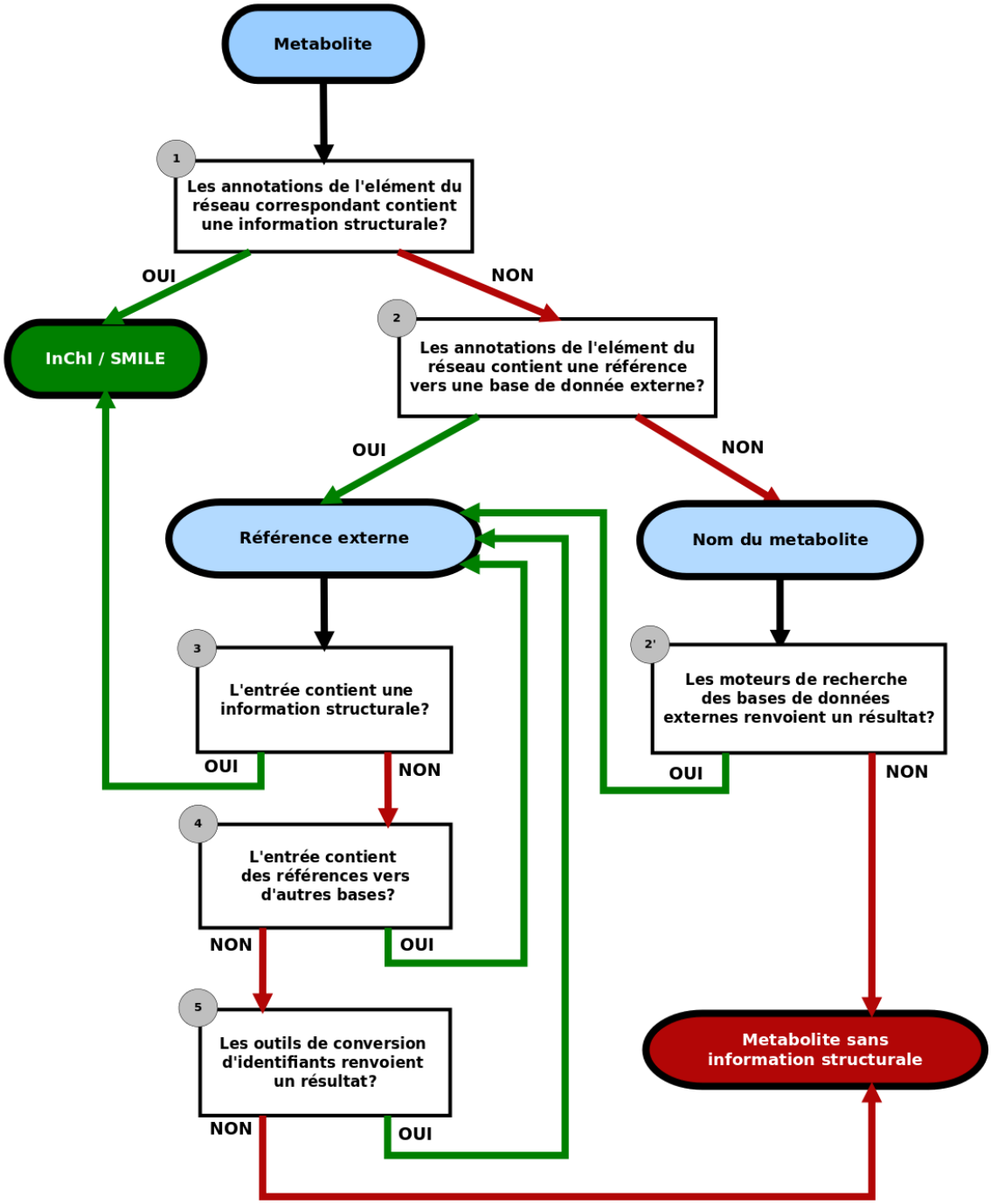


Figure 8.3 – Récupération d'informations structurelles

Les réseaux métaboliques sont fréquemment fournis dans des fichiers au format SBML[129]. Ce standard ne couvrant pas la gestion spécifique des attributs tels que les InChI ou les SMILES, ces informations sont généralement renseignées de manière non standardisée, dans le champ destiné à associer des notes à une entrée,

sous forme de texte brut. De manière générale, ce champ est largement détourné pour l'ajout d'attributs non couvert par le standard. Cette limite a donc nécessité le développement d'un parseur de fichier SBML hautement paramétrable, afin de permettre l'import de ces données. Le développement du parseur spécifique a ainsi permis d'assurer programmatiquement la première et la seconde étape du workflow proposée.

Les étapes 3 et 4 qui consistent à l'extraction de données depuis des bases externes, ont quant à elles nécessité le développement d'une application composite de *Web Scrapping*. Ce type d'application consiste en l'agrégation de données extraites de sources hétérogènes disponibles sur le web, et peut rechercher du contenu en naviguant entre ces sources via les liens hypertextes qui les référencent. Les bases de données couvertes par notre implémentation sont HMDB[270], KEGG[141], CHEBI et PubChem Compounds. Certaines de ces bases proposent des API (interface de programmation), permettant à un programme d'accéder à leur contenu. C'est le cas de CHEBI qui offre une bibliothèque logicielle Java, ainsi que KEGG et PubChem qui proposent, eux, des webservices de type REST (REpresentational State Transfer). Les informations contenues dans HMDB sont elles extraites en parsant les pages web de ses entrées, accessibles au format xml.

Les références entre les différentes bases de données peuvent également être obtenues via le webservice du *Chemical Translation Service*[272] ou le système UniChem[52] permettant la conversion des identifiants d'un métabolite d'une base de données vers une autre, et ainsi compléter la liste de référence externe pour un métabolite (étape 5). Il propose également une fonctionnalité de conversion de noms de molécules vers des identifiants de bases, ce qui a permis une utilisation de ce dernier assimilable à celle d'un moteur de recherche pour l'étape 2'.

Le réseau métabolique considéré dans cette étude, Recon 2 (version 3)[256], contenait initialement des annotations structurales pour 49% des composés, notre approche a permis de réduire de près de moitié, la quantité d'annotations structurelles manquantes, et ainsi d'élever la couverture à 77,42% de composés annotés. L'annotation des composés des réseaux métaboliques par agrégation de données

issues des principales bases de composés chimiques limite donc la rareté de ces informations dans ces modèles. Cependant, ces données restent manquantes dans certains cas. Par exemple, la taille importante de certaines molécules telles que les protéines, les peptides ou autres polymères, rend difficile le calcul de représentation 2D en chaîne de caractères (InChI ou SMILES). À titre d'exemple, Recon 2[256] contient ainsi 60 composés comportant la mention « *protein* » dans leur nom, et ne possédant pas de formule brute définie dans le réseau. Parmi les composés possédant une formule exacte dans Recon 2, le composé *psyllium-taurocholic acid complex* est constitué d'après son annotation de 33 738 198 atomes. Bien que rarement présent dans les réseaux métaboliques, principalement axés sur les molécules de petite taille, la modélisation dans Recon 2 de processus impliquant des molécules de grande taille (notamment le keratan sulfate, l'heparan sulfate ou encore le chondroitin sulfate) s'accompagne de la présence d'une centaine de produits de dégradation et d'une centaine de précurseurs de leur biosynthèse. Tous présentent un faible niveau d'annotation, sont référencés sous un terme générique suivi d'une numérotation, et présentent une faible représentation sous cette dénomination dans les bases de données citées précédemment. Par conséquent, leur annotation structurale n'a pu être conduite.

De plus, certaines enzymes, plutôt que d'être limitées à un substrat spécifique, peuvent agir sur une classe entière de composés. Ceci a conduit à la création de composés *génériques* dans les reconstructions afin de représenter ces classes de composés plutôt que l'ensemble de leurs membres, qui aurait conduit à la duplication multiple des réactions concernées. Un exemple issu de Recon 2[256] est celui du *monoacylglycerol 2*, dont la formule contient un radical. Ces composés génériques ne possèdent par conséquent pas de structure absolue qui puisse être représentée par un InChI.

8.3.6 Limite : Correspondance partielle entre données et modèles

L'un des principaux obstacles à la conduite de notre approche, et qui contribue majoritairement à la durée de préparation des données, est la recherche des métabolites de la signature dans le modèle.

Cette limite est due à l'ambiguïté qui réside dans la dénomination des molécules[184][223]. Bien que la nomenclature de l'IUPAC (International Union of Pure and Applied Chemistry) permette une identification précise des molécules, la lourdeur des dénominations ainsi obtenues conduit en pratique à un usage répandu de noms triviaux, généralement hérités de conventions antérieures à la standardisation IUPAC ou basés sur des propriétés ou l'origine des molécules. On peut citer par exemple l'acide méthanoïque, couramment nommé acide formique du fait qu'il fut historiquement isolé à partir de cadavres de fourmis (*formica* en Latin). Un autre exemple, plus symptomatique du faible usage des noms issus de la nomenclature IUPAC, est celui du β-carotène, un pigment qui contribue à la couleur caractéristique des carottes, dont le nom standardisé est 1,3,3-Trimethyl-2-[3,7,12,16-tetramethyl-18-(2,6,6-trimethylcyclohex-1-en-1-yl)octadeca-1,3,5,7,9,11,13,15,17-nonaen-1-yl]cyclohex-1-ene.

Ainsi, les dénominations utilisées dans les modèles et celles utilisées pour l'identification en métabolomique vont différer dans de nombreux cas.

Afin de lever ces ambiguïtés et assurer la concordance entre les noms issus des données et ceux utilisés dans le réseau, les références vers les bases externes vont être utilisées. Ce mode opératoire, présenté figure 8.4, va également bénéficier des développements utilisés pour l'annotation automatique présentée précédemment : l'enrichissement des références externes (étapes 5 et 6) ainsi que la récupération d'attributs depuis ces références (étape 3). Les attributs ici considérés étant les noms et les synonymes, ainsi que les InChI.

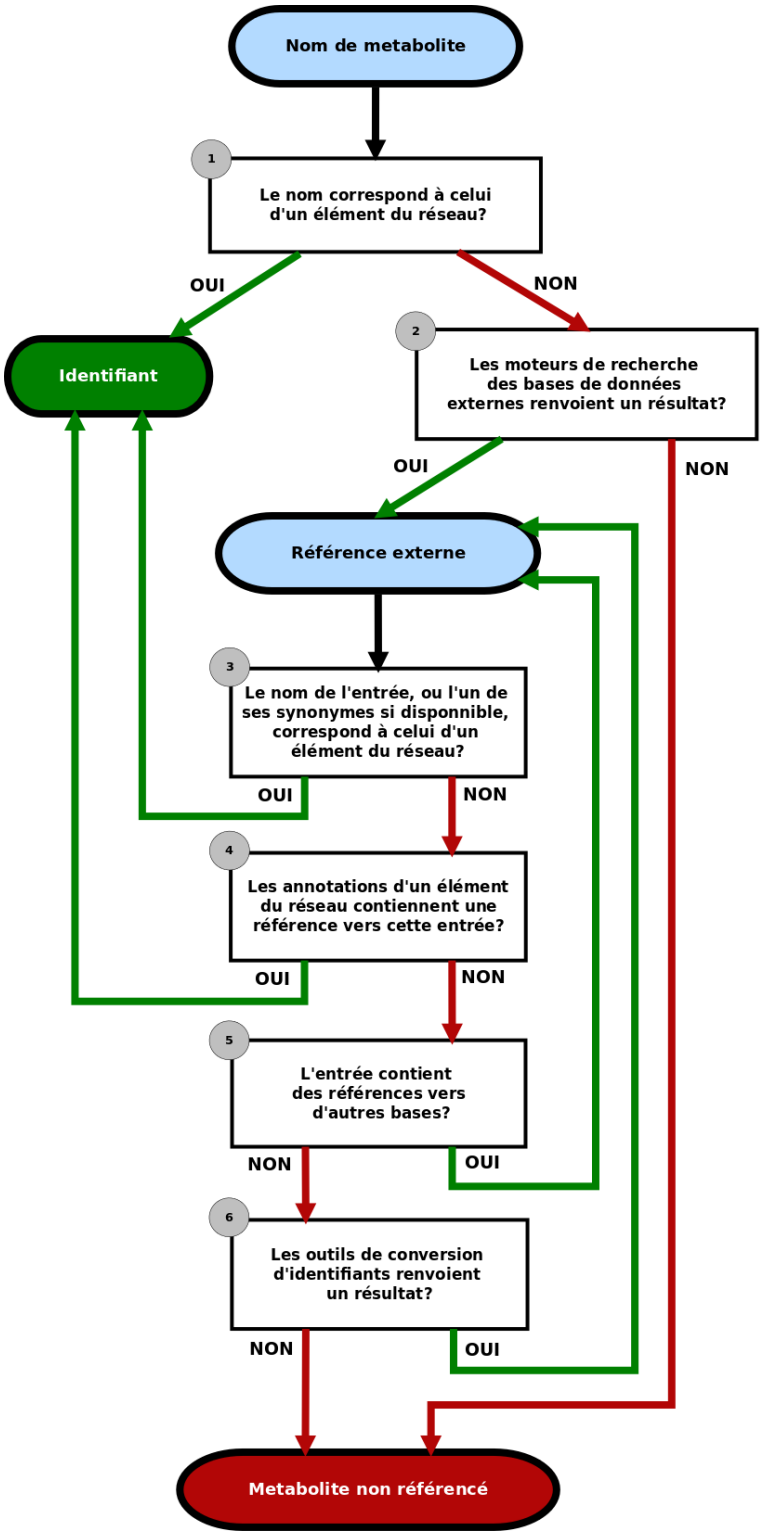


Figure 8.4 – Recherche de métabolites correspondants dans les réseaux

Les InChI constituent un système d'identification unique des métabolites, per-

mettant ainsi la levée d'ambiguïtés. Cependant, dans les modèles du métabolisme il est fréquent de représenter sous une même entrée les différents niveaux de protonation d'une molécule. Par exemple, les formes acide et base d'une molécule ne sont pas distinguées dans les réseaux, bien que leur différence induise une variation de leur dénomination (suffixes -*ate* ou -*ic acid*) ainsi qu'une variation de la portion de leurs InChI relative à la charge. Grâce à la construction en niveaux des InChI, il est cependant possible d'ignorer certains de ces niveaux lors de la recherche de correspondance (charge, stéréochimie)[184][25].

L'utilisation des InChI constitue donc le meilleur moyen d'établir la correspondance entre le système d'identification des données et celui du modèle. Dans le cas où ces InChI ne sont pas renseignés par l'une ou l'autre des parties, ce qui est fréquent dans les réseaux (Figure 6.1), il est toujours possible d'établir une correspondance via les références vers des bases de données externes[113]. Enfin, la recherche manuelle va permettre la validation des correspondances obtenues automatiquement ainsi que de combler certains manques.

Afin de faciliter la recherche des identifiants dans le réseau (étape 1), nous nous sommes inspirés des méthodes de recherche approximative (*fuzzy matching*)[197] utilisées pour les moteurs de recherche ou dans de nombreux logiciels pour réaliser des corrections orthographiques. Certaines de ces méthodes sont basées sur la définition d'une distance entre les mots et vont, faute de correspondance exacte, renvoyer le ou les mots les plus proches de la requête. La mesure de distance la plus emblématique est la distance d'édition, qui considère le nombre d'opérations nécessaires pour convertir un mot en un autre. Ces opérations peuvent être l'ajout, la suppression ou la substitution de caractères. La distance de Levenshtein[167] fut la première définition d'une distance d'édition, basée sur ces 3 opérations. Des variantes ont été proposées, considérant d'autres opérations comme la transposition de caractères adjacents[259]. Il est également possible de faire varier le coût associé à chaque type opérations, et de définir la distance comme la somme des coûts plutôt que le nombre d'opérations. Cette méthode a notamment été utilisée pour définir des distances entre des séquences de nucléotides[8].

Bien que les méthodes de *fuzzy matching* ne puissent pas résoudre les ambigüités telles que la synonymie entre « methanoic acid » et « formic acid », elles peuvent néanmoins résoudre des ambiguïtés liées au simple formatage ou l'omission de préfixes décrivant une conformation particulière. Par exemple « β-**D-Glucose 6-Phosphate** » et « **d-glucose-6-phosphate** » sont a une distance d'édition de 3 opérations (en ignorant la casse) : 2 additions/délétions des caractères « β » et « **-** » en début de mot et 1 substitution d'un espace en un caractère « **-** » entre « **glucose** » et « **6** ». Cependant, l'application de ces distances aux noms de molécules montre certaines limites : Par exemple, « **glucose-1-phosphate** » et « **glucose-6-phosphate** » représente deux entités bien distinctes malgré une distance de 1, alors que « **alpha-D-glucose** » et « α-**D-glucose** » correspond à la même dénomination sous un formatage différent induisant une distance de 5 opérations. Bien qu'il ait été montré que la mesure de similarité de noms de molécules nécessite des méthodes dédiées[275], à notre connaissance, il n'existe aucun outil open source de fuzzy-matching spécifique des noms de composés chimiques, ce qui a motivé le développement de notre propre méthode.

Une liste de 1381 paires de synonymes a été extraite à partir des noms des métabolites de Recon 2 et des noms des entrées PubChem leur correspondant, obtenus grâce aux annotations étendues du modèle. 325 paires comprenaient des noms identiques (sans tenir compte de la casse) et 551 paires présentaient une distance de Levenshtein minimum par rapport aux distances qui les sépare de l'ensemble des autres noms considérés. En d'autres termes, le fuzzy-matching permettait d'établir une correspondance non ambiguë dans environ 40% des cas comparés à 23,5% si restreints aux correspondances exactes.

Afin d'améliorer ce score, une recherche de motifs d'opération a été réalisée pour chaque couple de synonymes afin de caractériser leurs différences. Elle a permis de définir un ensemble de règles pour harmoniser les dénominations à l'aide d'expressions régulières. Ces règles comprennent la suppression des caractères spéciaux et espaces, l'usage d'abréviations courantes et de notations usuelles (« **Coenzyme A** » → « **coa** », « **trans** » → « **E** »), l'encodage en toutes lettres des

caractères grecs (« **alpha** » → « α »), l'usage de la forme basique, non chargée et non cyclique (« **oleic acid** » → « **oleate** », « **tryptaminium** » → « **tryptamine** », « **fructofuranose** » → « **fructose** ») et le réarrangement des formes composées (« **2-nonenal, 4-oxo-** » → « **4-oxo-2-nonenal** »). En complément de ces règles, une matrice de substitutions (régissant les poids attribués à chaque substitution) a été adaptée de manière à pénaliser les substitutions de nombres et les substitutions impliquant des caractères porteurs d'information structurale (« **L** » ⇆ « **D** », $\alpha \leftrightarrows \beta$, « **E** » ⇆ « **Z** »).

Cette méthode a permis sur notre jeu de test de faire monter le nombre de correspondances exactes après harmonisation à 538, et le nombre de correspondances non ambiguës à 791. Elle fait chuter la distance moyenne entre synonymes de 12,4 à 7,7, et la distance médiane de 8 à 2.

Ces développements, combinés à une analyse manuelle, nous ont permis dans de nombreux cas de faire coïncider les dénominations de nos données issues de métabolomique avec celle de nos modèles. Néamoins, une large portion de ces données n'a pu être exploitée faute de représentation dans le modèle Recon 2.

8.3.7 Validation

Méthodes employées pour les systèmes de recommandation du Web et leurs limites

La méthode proposée souffre d'une limite inhérente à de nombreuses approches exploratoires : la difficulté d'évaluer la pertinence des résultats obtenus.

Dans le cas des systèmes de recommandation du web, l'évaluation de la qualité de ces systèmes est une problématique qui a bénéficié d'une grande attention de la part de la communauté et différentes méthodes ont été proposées[15].

La suggestion d'items peut être considérée comme un problème de classification. Une validation croisée peut être utilisée à partir d'une liste d'items pour lesquels un utilisateur a marqué son intérêt (achat, suivi de lien)[139]. Le jeu de données est divisé en un jeu d'apprentissage et un jeu de tests. Une mesure de précision peut ensuite être calculée à partir de la proportion d'items du jeu de test

suggéré à partir du jeu d'apprentissage. Dans notre contexte, cela reviendrait à définir dans quelle mesure des métabolites d'une signature métabolique peuvent être suggérés à partir d'un sous-ensemble de la signature. Le problème de ces approches est l'hypothèse forte selon laquelle le jeu de données initial représente les items pertinents, et le reste, les items non pertinents. Usuellement, les items dont l'intérêt a été relevé ne représentent qu'une infime partie de l'ensemble des items et sont, dans ce scénario, incomplets par définition. Par conséquent, cette approche est critiquée et apparaît peu adaptée à notre problématique, d'autant qu'elle nécessiterait de conduire cette validation sur de nombreuses signatures métaboliques. Il est également à noter que cette validation ne capture pas l'intérêt d'un système de recommandation relatif à la sérendipité.

Une autre méthode couramment utilisée est la conduite de tests A/B, où est mesuré empiriquement le gain induit par l'usage du système sur des indicateurs tels que le nombre de ventes ou la satisfaction utilisateur par exemple[61]. De nombreux autres aspects peuvent être considérés pour estimer empiriquement la qualité d'un système de recommandation : précision (taux de recommandations suivies par l'utilisateur), diversité du contenu proposé, *item space coverage* (qui permet d'estimer l'importance du « *Harry Potter problem* » dans les recommandations), capacité à suggérer des items récemment ajoutés, *etc.* Le choix de ces aspects et les différentes métriques qui les capturent sont principalement motivés par le domaine d'application et l'objectif du système de recommandation. Par conséquent, peu de métriques utilisées par l'industrie du web s'appliquent à notre problématique.

L'une des particularités qui distingue notre problématique de celle des systèmes de recommandation du web est relative aux *feedbacks*. Ce sont ces retours utilisateurs (implicite ou explicite) qui sont fréquemment employés pour évaluer la qualité des recommandations. Dans le cas du web, ces retours peuvent être des achats, des clics, ou encore une notation explicite de l'utilisateur (exemple du « pouce bleu » de Facebook). Ils ont la particularité de requérir une implication très limitée de la part de l'utilisateur.

Dans le contexte de l'interprétation de résultats de métabolomique, l'évaluation de la pertinence d'un métabolite suggéré par rapport à la condition étudiée peut requérir un examen approfondi de la littérature et des données brutes, voire des expérimentations complémentaires coûteuses et chronophages. L'implication considérable de l'utilisateur et des ressources dont il dispose rend l'évaluation de la pertinence des recommandations particulièrement difficile, et constitue l'une des principales limites de cette méthode.

Néanmoins, dans une certaine mesure, lorsque l'utilisateur dispose d'un niveau d'expertise suffisant il peut rendre compte de la pertinence de certaines recommandations au regard de ses connaissances. Les méthodes d'évaluations basées sur les retours utilisateurs explicites pourraient dès lors être employées. Cependant, ce type d'évaluation appliqué aux systèmes de recommandation classiques ne requiert que peu ou aucun prérequis de la part des utilisateurs, qui se contente généralement d'évaluer si les recommandations sont en adéquation avec leurs goûts. Il est dès lors aisé de constituer un panel suffisamment large, et l'hypothèse selon laquelle chaque utilisateur dispose de la même capacité à évaluer sa satisfaction vis-à-vis d'une recommandation facilite l'agrégation des résultats obtenus. En revanche, dans le cadre de notre application des systèmes de recommandation, cette condition n'est pas remplie. L'évaluation de la pertinence requiert un haut niveau d'expertise vis-à-vis de la condition étudiée, qui est non seulement difficile à quantifier pour chaque individu, mais qui restreint également la taille des panels de testeurs.

Application aux systèmes de recommandation métaboliques

Dans le but de contourner cette limite et d'offrir une méthodologie facilement transférable à d'autres études, nous nous sommes orientés vers une confrontation des suggestions à une base de connaissances extraite automatiquement de la littérature scientifique, plutôt qu'aux connaissances d'un panel d'utilisateur.

Ces méthodes tirent parti de l'annotation des articles scientifiques de la base PubMed à l'aide du vocabulaire MeSH (*Medical Subject Heading*) qui permet entre

autres d'apposer des mots-clés aux articles pour faciliter leur indexation. Le logiciel Metab2Mesh permet de recenser l'ensemble des métabolites dont la mention de leur nom est statistiquement surreprésentée dans les articles annotés avec le terme MeSH relatif à la maladie étudiée[230]. Cette approche a permis de mettre en évidence la présence de nombreux métabolites associés à la maladie dans la liste des recommandations. Cette sur représentativité des molécules associées à la maladie dans les recommandations est étayée par le test exact de Fisher.

Outre l'utilité des approches d'*information retrieval* pour la validation des recommandations topologiques basées sur la centralité, elles offrent également une complémentarité vis-à-vis de ces dernières. En effet, les molécules associées à la maladie mais absentes de la signature et des recommandations se sont également avérées utiles pour l'interprétation. Elles ont permis de compléter certains scénarios mécanistiques, mais également de rajouter certaines molécules à la signature après examen des données brutes. C'est le cas de la bilirubine, du phenylacétate et du quinolinate, dont le premier d'entre eux avait été omis dû à une erreur du logiciel de détection de pics. L'approche basée sur la littérature permet donc d'enrichir le système de recommandation basé sur la topologie, et une approche hybride sera donc envisagée pour la suite de ces travaux.

L'une des limites de cette approche pour la validation est qu'elle se focalise sur les mécanismes déjà connus de la maladie, et ne permet pas de mettre en exergue les recommandations qui constitueraient une nouveauté. Afin de mettre en évidence des candidats d'intérêt non associés à la maladie, la même approche a été conduite en étendant les termes MeSH ciblés (initialement celui de l'EH seul) aux MeSH dont la co-occurrence avec celui de la maladie est statistiquement plus fréquente que la moyenne[242]. Cela a permis de capter les métabolites associés aux termes décrivant les principaux symptômes et signes cliniques de la maladie, ou associés à des maladies de physiopathologie proche. C'est par exemple le cas du N-Ω-L-Arginosuccinate, fréquemment mentionné dans les articles traitant de l'altération des niveaux de conscience et du coma, principaux symptômes et conséquences de l'encéphalopathie hépatique. L'approche consistant à prendre en

compte les MeSH connexes au MeSH ciblé permet également de contourner la limite liée à la non-exhaustivité des MeSH. La base de données MeSH est renseignée manuellement, une condition pathologique doit donc acquérir une certaine visibilité avant de voir sa dénomination apparaître dans l'ontologie MeSH. On peut par exemple citer le cas du virus Zika, découvert à la fin des années 50, mais dont l'apparition dans l'ontologie a eu lieu en 2016, après un épisode pandémique et sa mise en lumière par l'Organisation Mondiale de la Santé. Face à cette limite, il est possible de définir *a priori* une liste de MeSH associé à la condition non renseignée pour évaluer et enrichir les recommandations.

En revanche, l'une des principales limites de cette approche est qu'elle ne permet pas de statuer sur la nature des liens entre une molécule et un terme MeSH, et ne permet donc pas d'établir un lien de causalité entre l'abondance de la molécule et les signes cliniques par exemple. C'est notamment le cas du benzoate ou du diazépam, dont une recherche approfondie de la littérature révèle que leur association avec l'EH résulte essentiellement des effets thérapeutiques de leur administration aux patients[189].

Une autre limite est l'absence de dimension temporelle dans les associations MeSH-molécules qui tend notamment à occulter les découvertes récentes au profit des découvertes de notoriétés plus ancrées. Cette limite est particulièrement traduite par la recommandation de l'oxoglutaramate, associé à peu de termes pertinents, malgré des travaux récents suggérant son potentiel de biomarqueur de la maladie et son implication dans les maladies liées à l'hyperammonémie[56][55]. Ces travaux mentionnent tout particulièrement le caractère méconnu et négligé de son implication dans le métabolisme de la glutamine, ce qui traduit l'impact des effets de mode sur cette approche.

L'exemple suivant illustre le type d'erreurs pouvant être induites par ces deux limites : si une molécule a été considérée comme impliquée dans la patogénicité d'une maladie pendant de longues années, mais que cette hypothèse a été réfutée plus récemment, suite à des avancées technologiques par exemple, l'association entre la maladie et la molécule par Metab2MeSH restera forte du fait de l'his-

torique porté par les recherches sur cette maladie. Ceci est d'autant plus vrai si cette idée fausse vient à faire office de cas d'école, et que l'anecdote continue d'être fréquemment mentionnée pour justifier les approches basées sur la technologie émergente. Ce comportement aurait alors comme conséquence d'augmenter l'association entre la maladie et cette molécule d'après Metab2Mesh, postérieurement à la réfutation du lien de causalité qui les aurait unis.

Il est à noter que les conclusions sur la pertinence de la méthode au regard des informations extraites de la littérature ne peuvent être généralisées à toutes les signatures métaboliques. Nous n'avons été capables de mettre en évidence l'intérêt de l'approche que dans le contexte particulier de la signature de l'EH tel que générée par Weiss *et al.* L'analyse de recommandations issues d'autres signatures serait nécessaire pour légitimer ce type d'approche. La difficulté d'obtenir une signature métabolique suffisamment exhaustive et présentant des niveaux d'identification non équivoques, ainsi que l'expertise nécessaire à leur interprétation, a nécessité une mobilisation importante des expérimentateurs et de fréquents échanges. En conséquence, la méthode de recommandation n'a pu être conduite sur d'autres données au cours de cette thèse.

Toutefois, conduire l'analyse sur d'autres signatures ne permet pas de s'affranchir du lien qu'il existe entre la qualité des données d'entrée et la qualité des recommandations. Dans le cadre des recommandations du web, les données d'entrée sont généralement considérées non équivoques et définitives. Il est peu probable qu'un utilisateur achète un objet par accident, où qu'il attribue une note positive a un film qu'il n'a pas apprécié par exemple. En revanche, les données de métabolomique sont sujettes à de nombreuses erreurs et biais, il devient dès lors difficile dans le cas de recommandations aberrantes de distinguer la part imputable au système de recommandation de celle imputable à la faiblesse des données d'entrée.

8.4 Implémentation

La méthode est rendue accessible à la communauté par son implémentation dans la plate-forme d'analyse MetExplore[59]. Elle offre une visualisation des scores de recommandation au travers d'une projection en 2 dimensions[78], facilitant l'identification et la sélection des composés les plus pertinents (exemple Figure 8.5). Les recommandations peuvent être traitées directement depuis la plate-forme, via les fonctionnalités d'enrichissement de voies et la visualisation dynamique des réseaux.

D'autres développements basés sur cette méthode peuvent également être réalisés à l'aide de la bibliothèque logicielle d'analyse topologique de réseaux biologiques Met4J, développée au cours de cette thèse. Cette *library* développée en Java propose différents algorithmes de centralité et de recherche de chemins, ainsi que différentes méthodes de pondérations pour garantir leur pertinence dans le cadre d'applications aux réseaux métaboliques. Elle n'est néanmoins pas limitée à ce seul type de réseau. Grâce à la généricité offerte par le langage Java, une large proportion des fonctionnalités proposées peuvent être conduites sur des réseaux biologiques nouvellement implémentés.

De nombreuses *library* dédiées à l'analyse de graphes sont actuellement disponibles, telles que JGraphT et JUNG en Java, igraph[65] en R ou NetworkX[109] en python. Bien que l'éventail de méthodes proposé par ces bibliothèques logicielles soit très complet, les développements spécifiques aux réseaux biologiques et la gestion des formats de fichier qui leurs sont propres ne sont pas implémentés dans ces dernières. Les outils dédiés à l'analyse de réseaux biologiques, tels que NeAT[45] ou MetExplore[59], proposent certaines méthodes au travers d'interfaces graphiques et n'offrent donc pas par conséquent la modularité nécessaire au développement de nouvelles méthodes.

La *library* Met4J est basée sur la *library* JGraphT pour la gestion de la structure des données et les opérations de base comme l'ajout ou la suppression de nœuds ou d'arc. L'import et l'export des données via les fichiers SBML estréalisée à l'aide de la *library* JSBML[70]. Les fonctionalités relatives à la similarité

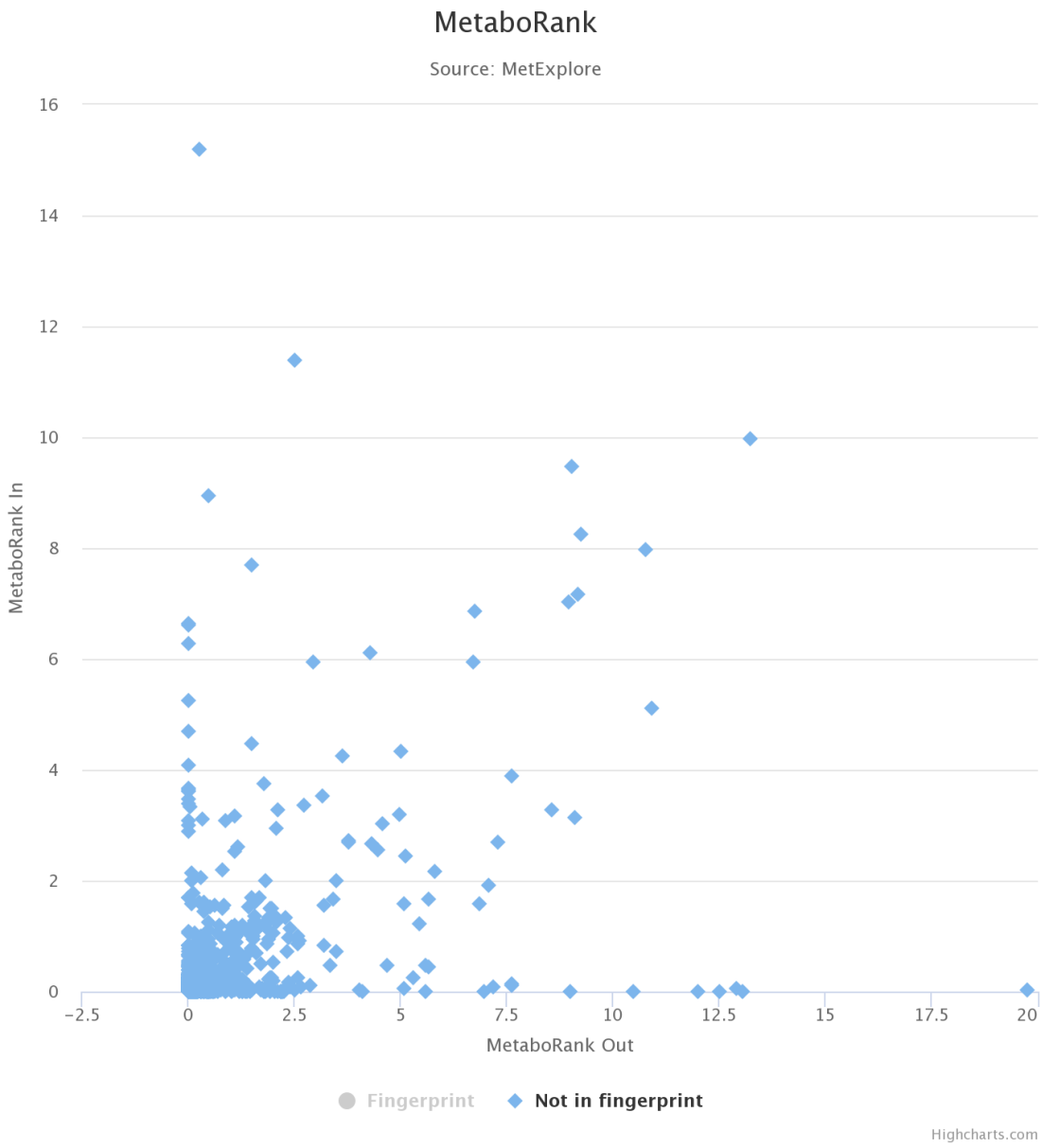


Figure 8.5 – Interface MetExplore pour la visualisation des scores de recommandation

chimique ou l'atom-mapping sont réalisées à l'aide de la *library* de chemoinformatique CDK[247]. Enfin, la *library* Ejml[1] est utilisée pour les calculs matriciels.

Table 8.3 – Fonctionalités de la *library* Met4J

Import/export
Import/export de réseaux métaboliques au format SBML[129][70]
Import/export de réseaux métaboliques au format tabulé
Export de graphes au format .gml
Export de graphes au format cytoscape
Export de matrices au format .csv
Construction de Graphes
Graphes génériques
Graphes compressés (un arc correspond à un chemin)
Graphes métaboliques des composés*
Graphes métaboliques bipartis
Graphes métaboliques des réactions
Graphes de voies
Manipulations de chemins
Extraction de sous-chemins
Concaténation de chemins
Pondérations de graphes
Pondération par le degré[64]
Pondération par les tags RPAIR*[81]
Pondération par la similarité chimique*[219]
Conversion des poids en probabilités
Conversion des poids en probabilités, ajustées par réaction*
Import/export des poids au format tabulé
Opérations de bases : montée en puissance, conversion similarités/distances, normalisation
Recherches de chemins métaboliques et extractions de sous-réseaux
Recherche de plus courts chemins (algorithme de Dijkstra)
Recherche des k-plus courts chemins (algorithme de Yen)
Calcul de matrices de distances (Floyd-Warshall)
Recherche de « *Best paths* », heuristique basée sur la similarité chimique*[183] (algorithme A*)
Construction de *Metric closure graph*
Extraction d'arbres de Steiner[82] (approximation basée sur le minimum spanning tree)
Recherche de chemins métaboliques avec exclusion mutuelles des directions des réactions réversibles*[261]
Centralités
Eigenvector centrality
PageRank[194]
PageRank personalisé[133]
Betweenness centrality[138][282][120]
Random-Walks betweenness centrality[73]
Closeness centrality[177]
Eccentricity
Farness
Neighborhood centrality
Katz centrality[280]
Analyses de voisinages

Extraction de voisinage
Nombre de voisins communs
Adamic-Adar index
Salton index
Coefficient local de clustering
Mesures de graphes
Nombre de composantes connexes
Diamètre
Longueur
index Gamma
index Beta
index Alpha
index Eta
index Pi
Nombre cyclomatique
overall closeness centralization index[279]
Opérations sur les graphes
intersection
union
filtrage basé sur les poids
*Compartments merging**
*Multi-edges merging**
Conversions
Calcul de matrices d'adjacences
Conversion Graphe des composés $\leftrightarrow$ graphe biparti
Similarités de graphes
Tanimoto
Nombre d'arêtes partagées
Similarité de rangs
Kendall τ
Coefficient de correlation de Spearman
Autres analyses
Load Points[220]
Choke Points[220]
Pathway enrichment

* disponibles uniquement pour les graphes des composés

Chapitre 9

Conclusion et perspectives

9.1 Conclusion

"The most exciting phrase to hear in science, the one that heralds new discoveries, is not «Eureka !» but «That's funny.»"

Isaac Asimov

Les travaux présentés dans cette thèse proposent une méthode d'exploitation des réseaux métaboliques pour l'interprétation de résultats de métabolomique. Ils proposent pour ce faire une définition de la proximité dans les réseaux métaboliques qui soit porteuse de sens et précisent les limites et futurs défis relatifs à ces méthodes. Ces propositions ont émergé d'un état de l'art sur l'adaptation des méthodes de recherche de chemins métaboliques et de la caractérisation des limites inhérentes aux modèles du métabolisme et à son observation en métabolomique, présentés dans cette thèse. Ils suggèrent l'utilisation des proportions d'atomes échangés entre substrats et produits pour définir des liens pertinents entre métabolites (définissant ainsi des transitions auxiliaires plutôt que des composés auxiliaires), et démontrent que la prise en compte des réactions réversibles demeure un problème ouvert. Ils soulignent également la limite de la restriction aux plus courts chemins pour caractériser les relations qui unissent deux métabolites, et suggèrent l'utilisation de méthodes moins restrictives telles que l'union de marches aléatoires. Cette clarification du concept de métabolites « unis » et la

formulation de la mesure qui l'accompagne offrent un potentiel intéressant pour l'interprétation de listes de métabolites, qui peut s'appliquer à d'autres approches basées sur les réseaux, comme l'extraction de sous-réseaux pertinents, la recherche de communautés ou de motifs. Enfin, ils soulignent l'intérêt de ces méthodes comparées aux approches traditionnelles basées sur l'enrichissement de voies métaboliques, limitées par la définition arbitraire et non consensuelle de ces dernières, ainsi que par les potentiels biais d'interprétations qu'elles entraînent.

Les travaux présentés proposent également des méthodologies perfectionnant l'exploitation des réseaux métaboliques et leur interopérabilité, en facilitant la recherche de correspondances entre les données de métabolomique et les informations contenues dans les réseaux, ainsi qu'en enrichissant ces dernières par l'intégration de données externes. Ces avancés ont également permis de caractériser plus finement les limites de la métabolomique, en mesurant la représentativité des résultats obtenus vis-à-vis des connaissances actuelles du métabolisme.

Ces résultats soutiennent notre hypothèse selon laquelle la principale limite à l'interprétation des résultats de métabolomique repose sur l'incomplétude du métabolome observé. Cette hypothèse nous a conduit à l'élaboration d'un système de recommandation de métabolites visant à combler ces vides, le premier de ce genre à notre connaissance. La caractérisation des différentes mesures visant à l'identification d'éléments d'importance dans les réseaux (les mesures de centralité), conjointement à la caractérisation en termes de graphe des phénomènes biologiques considérés (la propagation des perturbations d'abondance), a permis de suggérer une mesure de centralité adaptée à notre problématique. Les caractéristiques recommandées sont l'utilisation d'une mesure radiale de volume de marches aléatoires, pondérées et robustes aux modifications du réseau comme des données d'entrée. Elles ont conduit à l'utilisation du PageRank comme mesure de centralité, dont la limite liée à son incapacité à élucider des potentiels précurseurs des métabolites d'intérêt a été comblée par son usage conjoint avec le CheiRank. La problématique relative à l'identification de biomarqueurs qui soient spécifiques de la condition étudiée nous a également conduits à pondérer ces résultats par la cen-

tralité globale, afin de mettre l'accent sur des composés caractéristiques du profil considéré. Un *framework* de validation de ce type de système de recommandation a également été proposé, en exploitant des outils statistiques dédiés à l'analyse de la littérature scientifique, afin d'estimer la concordance entre les recommandations et leur représentation dans la littérature associée à la condition étudiée.

La méthode a été utilisée pour analyser une signature métabolique du liquide céphalo-rachidien de patients atteints d'encéphalopathie hépatique (EH) et a permis d'enrichir les résultats obtenus en métabolomique. Les recommandations ont permis de lever des ambiguïtés quant à la présence de certains métabolites identifiés dans la liste des métabolites à considérer pour l'interprétation, ambiguïtés notamment liées à la variabilité des écarts observés entre patients et contrôles. Les recommandations ont également mis en évidence plusieurs métabolites non mesurés, dont le lien avec l'EH est corroboré par de nombreuses études. Enfin, les recommandations ont permis de mettre en lumière l'acide kynurénique qui n'avait pas été considéré initialement, mais dont les données brutes suggèrent une implication dans la maladie, au niveau de mécanismes précédemment identifiés par hyperammonémie induite chez le rat. Elles ont de plus mis en évidence l'oxoglutaramate, un métabolite peu étudié dans ce contexte bien que des résultats suggèrent son statut de biomarqueur de la maladie. L'analyse des recommandations, couplée à une exploration visuelle du réseau métabolique humain, a également conduit à une reconstruction de scénarios mécanistiques plus poussée, augmentant la couverture des métabolites de la signature.

Cette thèse propose donc une preuve de concepts de l'utilité des systèmes de recommandation pour l'analyse de résultats de métabolomique. La méthode proposée est notamment particulièrement adaptée à l'incomplétude des données et du modèle. Elle permet en effet d'enrichir les données et de limiter cette incomplétude, tout en étant relativement robuste face aux changements de faible envergure du réseau et des données. Cette robustesse est également soutenue par une aisance à l'adaptation à ces changements, du fait d'un temps de calcul de l'ordre de quelques minutes sur un ordinateur personnel pour un réseau à l'échelle du gé-

nome, ainsi qu'un formalisme des résultats sous forme de classement, offrant une interprétation intuitive.

La méthode constitue néanmoins une première étape à la reconstruction de scénarios mécanistiques. Elle vise à fournir une aide à la reconstruction manuelle de ces scénarios plutôt qu'une inférence directe de ces derniers, se démarquant ainsi de la majorité des méthodes précédemment proposées. Ce type d'approche offre l'avantage de permettre d'intégrer les connaissances extrinsèques de l'utilisateur lors de la reconstruction, et garantit une transparence sur l'obtention des scénarios.

9.2 Perspectives

9.2.1 Comprendre les liens qui unissent signatures et recommandations

La méthode proposée permet de mettre en évidence des nœuds particulièrement connectés aux nœuds d'intérêt. Une approche naturelle pour l'interprétation de la liste de nœuds obtenue serait d'analyser la nature de ces connexions. Néanmoins, comme mentionnée au début de cette thèse, la « lecture » des réseaux métaboliques constitue une tâche difficile pour l'œil humain. Des méthodologies spécifiques seraient par conséquent nécessaires pour exploiter pleinement le potentiel des recommandations.

Une approche possible est l'extraction automatique de sous-réseaux de taille réduite. La première approche pourrait consister à extraire l'ensemble des marches considérées par l'algorithme. Un parcours de graphe (en largeur ou en profondeur) permet d'extraire l'ensemble des nœuds atteignables depuis un ou plusieurs nœuds présélectionnés (les métabolites d'intérêts), afin d'extraire le sous-graphe induit contenant l'ensemble de ces nœuds. Étant donné la taille et la connectivité des réseaux considérés, un tel graphe est supposé de grande taille et non utilisable en l'état. Il est possible d'élaguer le réseau obtenu en considérant uniquement les marches qui atteignent un nœud d'intérêt ou un nœud recommandé. Une solution possible est de conduire une approche similaire sur le graphe inversé (orientation

des arcs inversée), et de réduire la liste des nœuds à conserver à ceux remplissant les deux conditions suivantes : atteignable depuis un nœud d'intérêt ou de haut CheiRank (parcours sur le graphe originel), et pouvant atteindre un nœud d'intérêt ou de haut PageRank (parcours sur le graphe inversé). Afin d'éviter de faire figurer des marches de très faible probabilité (étant donné qu'a chaque étape d'élongation la probabilité de terminer la marche est définie par α), il est également possible de contraindre les parcours de graphe à une profondeur donnée au-delà de laquelle les nœuds successeurs ne sont plus visités. Bien qu'un tel sous-graphe puisse être considéré comme représentatif du voisinage des nœuds intérêts induisant la liste de recommandations, il risque en pratique de ne pas être interprétable par un humain de par sa taille.

Faust *et al.* ont suggéré l'utilisation des k-plus courts chemins[76] et de l'arbre de Steiner pour l'extraction de sous-graphes pertinents[82][83]. Bien qu'ils aient montré la pertinence de ces méthodes vis-à-vis de certaines voies métaboliques telles que représentées dans les bases de données, ces méthodes ne permettent pas d'apprécier les caractéristiques topologiques qui conduisent aux recommandations. En effet, ces méthodes réduisent les liens entre métabolites à un unique chemin (ou au maximum k chemins), masquant de nombreux chemins alternatifs. La caractéristique du PageRank qui a principalement motivé son utilisation pour la recommandation repose sur le fait que l'ensemble des chemins alternatifs entre deux métabolites est considéré, ce qui limite l'usage de ces méthodes pour son interprétation. Dupont *et al.* ont proposé une méthode alternative basée sur la *RandomWalk betweenness*[1], qui consiste à estimer pour chaque nœud la probabilité qu'il soit emprunté lors d'une marche aléatoire entre deux nœuds, puis de filtrer le graphe pour éliminer les nœuds pour lesquels cette probabilité était inférieur à un seuil définit *a priori*[73]. Ce type d'approche peut être pertinent pour extraire un sous-réseau interprétable, en utilisant la centralité de PageRank/CheiRank comme critère de filtre ainsi que l'algorithme d'élagage précédemment proposé. La limite des méthodes de filtrage est qu'elles nécessitent la définition d'un seuil difficile à

1. Faust *et al.* ont également utilisé ce critère pour pondérer la recherche d'arbre de Steiner, ce qui conduisait à de meilleurs résultats selon leurs critères

estimer *a priori*. Une autre limite inhérente à l'ensemble des méthodes d'extraction de sous-réseau est qu'elles masquent le degré réel des métabolites qui le composent, or cette information peut être digne d'intérêt pour l'interprétation. Par exemple, un métabolite qui s'avèrent être produit exclusivement à partir d'un métabolite d'abondance anormale ne suscitera pas nécessairement le même intérêt qu'un métabolite pouvant être produit par ce dernier ainsi que par de nombreuses autres voies de biosynthèse. De manière générale, les sous-réseaux obtenus à partir de liste de l'ordre de quelques dizaines de métabolites conduisent à des sous-réseaux de tailles relativement importantes, et il peut exister de nombreux sous-réseaux alternatifs qui peuvent renfermer des éléments clés pour la compréhension des mécanismes étudiés. Ces méthodes peuvent donc conduire, à partir d'un ensemble de métabolites difficilement interprétable, à un ensemble plus grand de métabolites et de réactions encore plus difficiles à interpréter.

Une approche alternative à l'extraction automatique de sous-réseaux est la reconstruction manuelle de scénario métabolique « assistée » par ordinateur. Elle consisterait, au travers de différentes méthodes de visualisation de données, à faciliter l'exploration du réseau métabolique par un utilisateur en favorisant la visibilité de certains métabolites. Ces métabolites seraient sélectionnés selon un critère de pertinence calculé par la machine, tel que la mesure de centralité proposée dans cette thèse. Une telle méthode a été proposée par Van Ham et Perer[111], et appliquée à l'exploration du réseau de citations issu de documents juridiques. Elle s'éloigne de la traditionnelle méthode « *Overview, zoom, details on demand* » utilisée par la plupart des logiciels utilisés pour l'analyse de réseaux métaboliques (dont le serveur web MetExplore[59] dans lequel est implémentée notre méthode), qui consiste à afficher le réseau dans sa totalité, zoomer sur les zones contenant les nœuds d'intérêts, puis afficher les informations contextuelles lorsqu'un élément de cette zone est sélectionné[54]. La taille des réseaux considérés rend cette approche peu fonctionnelle, et nécessite des ressources importantes pour leur affichage. La méthode de Van Ham et Perer repose sur le principe du « Search, Show Context, Expand on Demand », qui a l'inverse va partir d'un élément d'intérêt identifié à

partir d'une requête (*focus point*), puis afficher son voisinage qui sera étendu itérativement à la demande de l'utilisateur par sélection des éléments pertinents de ce voisinage (Exemple figure 9.1).

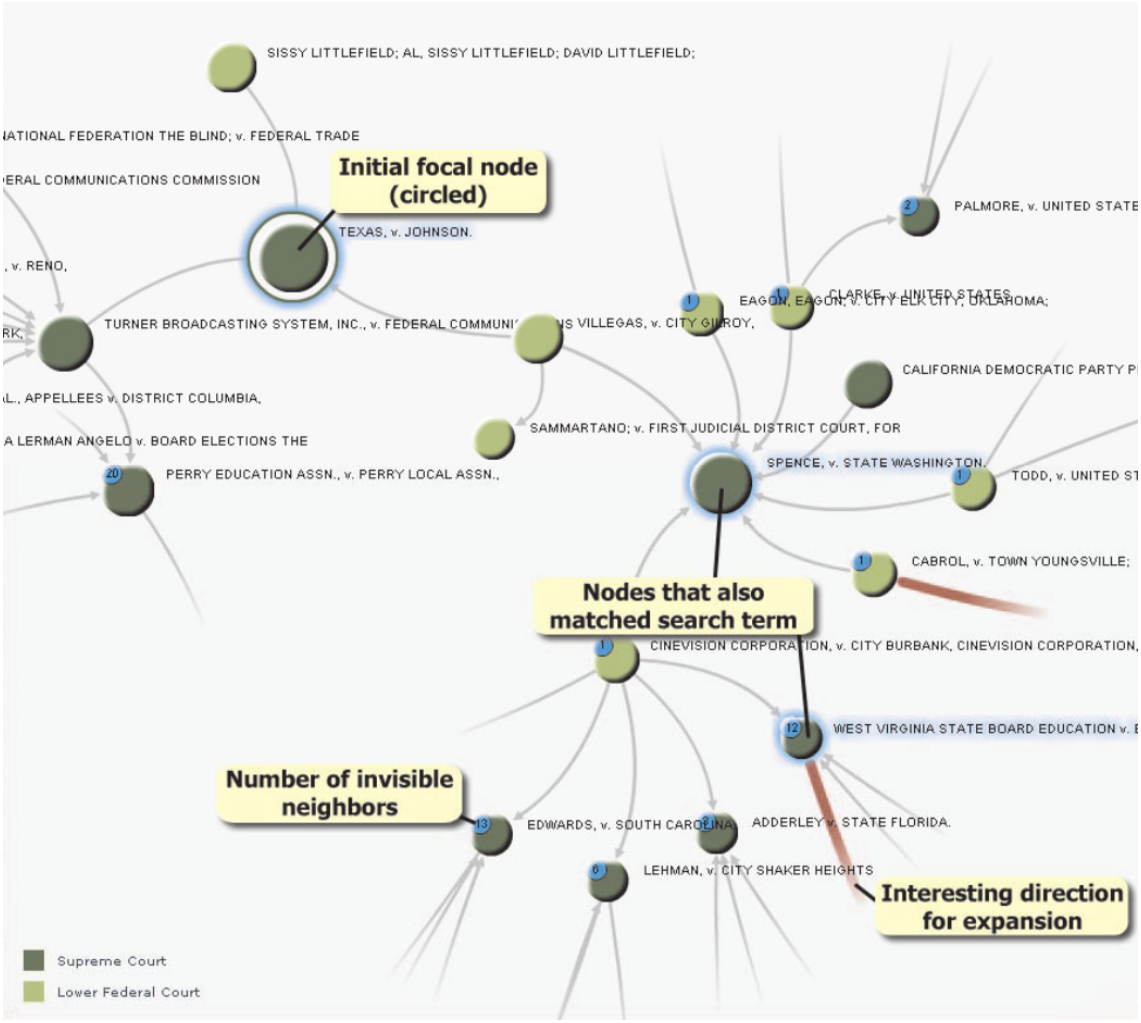


Figure 9.1 – Exemple d'affichage de réseau basé sur le « Search, Show Context, Expand on Demand ». Adapté de Van Ham et Perer

La représentation du voisinage y est clarifiée par son raffinement aux éléments les plus pertinents, définie au travers du **degree of interest** (DOI). Le DOI, proposée par Furnas[95] et adapté par Van Ham et Perer, est une fonction définissant l'importance d'un nœud par une combinaison linéaire de l'intérêt *a priori*, définie à partir des attributs du nœud ou des ses caractéristiques topologiques (typiquement le degré), et de sa distance au *focus point*. Van Ham et Perer y ajoutent un

autre facteur d'intérêt calculé à partir de la pertinence du nœud vis-à-vis de la requête ayant servi à la définition du *focus point*, et modifient la fonction d'intérêt *a priori* pour prendre en compte l'intérêt des voisins du nœud évalué.

D'autres approches similaire existent[62], cependant à notre connaissance ce type de fonctionnalité n'est implémenté dans aucun outil dédié à l'analyse de réseaux métaboliques. Selon notre opinion, la méthode de suggestion proposée dans cette thèse capture les principaux aspects du DOI tels que définis par Van Ham et Perer. Si l'on assimile le résultat d'une requête (essentiellement issue de recherche textuelle dans le contexte de leur étude) aux résultats de métabolomique, et que le *focus point* est sélectionné à partir de ces résultats, le PageRank personnalisé offre un résultat intégrant à la fois la notion de proximité au *focus point*, la notion de *feedback* suggéré par les auteurs, la prise en compte de propriétés topologiques des nœuds et une prise en compte de l'ensemble des résultats de la requête dans la définition d'importance. L'utilisation du système de recommandation proposé dans ce contexte de visualisation de données, plutôt qu'au travers d'une liste de suggestions, semble être une approche prometteuse puisqu'elle facilite l'interprétation des liens entre la signature et les métabolites pertinents, tout en offrant une reconstruction de scénario mécanistique qui peut s'appuyer sur les connaissances de l'utilisateur. Elle permet également de s'affranchir de la définition arbitraire du nombre de recommandations à fournir (fixé, dans notre étude, aux 50 métabolites de plus hauts rangs. La liste de 79 métabolites, présentée Table 8.2 comprenant l'union de ces métabolites issus des deux classements).

9.2.2 Contextualiser les recommandations à partir de la littérature scientifique

L'utilisation dans cette analyse des associations métabolites - termes MeSH[230], initialement employée pour illustrer la pertinence des recommandations, a également montré un potentiel pour interpréter et enrichir ces dernières. Cette approche de contextualisation sémantique des recommandations pourrait être poursuivie en utilisant des méthodes alternatives à l'usage des MeSH. En effet

l'usage des termes MeSH, bien qu'elle fournisse une information non ambiguë et structurée, limite la littérature considérée aux documents annotés. Les politiques d'annotation des articles par leurs auteurs varient selon les éditeurs, et, bien qu'il existe des programmes permettant une annotation automatique, les documents annotés avec des termes MeSH reste essentiellement limités aux articles de la base MEDLINE/PubMed. De plus, ces annotations ont vocation à mettre en avant les principaux thèmes des articles, et ne couvrent pas la totalité et la diversité des informations contenus dans les articles. Enfin, la base même des termes MeSH est particulièrement axée sur le vocabulaire relatif à la médecine et la santé publique, ce qui limite leur usage à ce champ thématique.

Il est néanmoins possible, à partir d'un corpus de publications scientifiques, d'extraire un contexte sémantique entourant la mention d'un métabolite, directement à partir du traitement de ces données textuelles[53][187][173]. Cette approche permet de condenser et raffiner l'information qui peut être contenue dans des milliers d'articles scientifiques, et ainsi offrir une nouvelle caractérisation du rôle d'un métabolite, complémentaire de ses caractéristiques topologiques obtenues par analyse des réseaux métaboliques. Les résultats obtenus pourraient s'avérer utiles pour l'interprétation des recommandations, mais également être implémentés dans l'algorithme de suggestion. Ce système de recommandation hybride permettrait de proposer des métabolites dont l'intérêt est à la fois défini par ses liens avec les métabolites discriminants (inférés depuis le réseau métabolique) et par ses liens avec la condition étudiée (inférés depuis la littérature scientifique).

Il est également possible de considérer, en parallèle de la proximité dans les réseaux métaboliques, des liens entre métabolites qui soient caractéristiques d'une similarité sémantique entre les articles les mentionnant[166][57][104], ou qui représentent leurs cooccurrences dans des articles scientifiques[30]. La méthode de recommandation proposée pourrait être adaptée pour considérer non plus des réseaux de relations « biochimiques » (A et B impliqués dans une réaction C), mais des réseaux de relations « sémantiques » (par exemple A et B mentionnés dans des articles traitants de D).

Cette contextualisation des données est rendue possible notamment par les différents développements en *Information Retrieval* (IR), *Text Mining* et en *Natural Language Processing*[171] (NLP). L'IR est une thématique de recherche qui prend sa source dans le développement des moteurs de recherches, et qui vise à la recherche d'informations, usuellement sous forme de documents textuels, qui soit la plus appropriée vis-à-vis d'une requête. Le *Text Mining* est une discipline plus large encore qui peut être résumé au *Data mining* appliqué aux données textuelles. Les tâches les plus étudiées en *Text Mining* sont la classification et le clustering de documents, l'extraction d'informations d'un document et le résumé automatique de document. Enfin, le NLP, intimement lié aux deux précédents, est un domaine dédié à la compréhension du langage humain par les machines. Il permet entre autres de détecter certaines structures lexicales dans des phrases, comme les noms, verbes et les adjectifs, lever les ambiguïtés liées à l'existence de synonymes, reconnaître des entités nommées (personnes, lieux, organisations), identifier et caractériser des relations entre ces entités ou encore regrouper des mots par thématique.

Ces approches ont notamment été appliquées au traitement de littérature scientifique dans le domaine biomédical[157][105][216][123]. Elles ont notamment permis de détecter des noms de molécules, de gènes ou de protéine dans des documents, extraire les relations entre ces entités, retrouver des séquences associées aux gènes. Elles ont également été appliquées à la priorisation de gènes candidats, problématique proche de celle étudiée dans cette thèse, par extraction des relations gène-fonction à partir de la littérature.

Bien que ces approches soient basées sur l'extraction de découvertes établies, qui par conséquent ne peuvent permettre la considération de mécanismes inconnus, elles offrent cependant des alternatives prometteuses pour la recommandation de métabolites.

9.2.3 Comparer des listes de métabolites au travers de leurs implications mécanistiques

La méthode proposée est focalisée sur l'analyse d'une signature unique. Il peut néanmoins être nécessaire, dans certains contextes, de pouvoir comparer des signatures issues de différentes conditions. La signature considérée dans notre étude résulte de l'agrégation des profils métaboliques de différents patients, présentant des degrés de gravité de la maladie différents et des caractéristiques variées (âge, sexe, environnement...). De plus, l'encéphalopathie hépatique dont ils souffrent, causée par une défaillance hépatique, peut résulter de différents facteurs : alcoolisme, hépatite, empoisonnement. Face à cette variabilité interindividuelle, il peut être pertinent de considérer des signatures patients-spécifiques pour conduire des approches de médecine personnalisée. Une première approche pourrait être de stratifier les patients en sous-groupes, ce qui nécessiterait une comparaison de leurs profils métaboliques.

L'approche triviale serait de quantifier l'intersection des ensembles de métabolites discriminants issus de deux signatures différentes. La limite de cette approche est qu'elle ne considère pas les relations entre les métabolites qui constituent ces ensembles. Ainsi, si deux signatures sont identiques à l'exception d'un couple de métabolites, qu'ils appartiennent à la même « voie métabolique » ou non n'affecte pas la mesure de distance. Par conséquent, il devient difficile de distinguer la part de différences due à de la variabilité biologique de celle liée à une réelle différence de mécanisme.

Il semble donc pertinent de considérer le contexte de chaque métabolite dans cette comparaison. La mesure de pertinence utilisée pour notre système de recommandation pourrait être exploitée pour remplir cet objectif. En attribuant un score à chaque élément du réseau métabolique, elle permet de représenter numériquement l'impact estimé des métabolites discriminants sur le réseau global. Il peut être envisageable de comparer ces impacts plutôt que les signatures elles-mêmes. Le passage de la comparaison d'ensembles d'éléments discrets (éventuellement convertis en variables binaires) à la comparaison de vecteurs numériques permet

d'étendre le panel de méthodologies applicables pour la comparaison. Ce formalisme permet entre autres l'usage de nombreuses méthodes de calcul de distances, la projection dans des espaces multidimensionnels et d'autres approches de clustering. Ce type d'approche a par exemple été utilisée dans le domaine du *Natural language processing*, en définissant la similarité entre deux mots par la distance cosinus des vecteurs de PPR obtenus à partir de ces derniers, dans un réseau de relations sémantiques[130].

Une autre approche serait de calculer directement une distance entre les ensembles de métabolites discriminants par rapport à la topologie du réseau. Le SimRank[134] proposé par Jeh et Widom semble être une approche prometteuse. Inspiré du PageRank et du concept sous-jacent de *Random Surfer*, le SimRank considère, lui, une paire de *Random Surfers*, et définie la distance entre deux nœuds a et b comme la distance attendue pour que les deux surfeurs se rejoignent, partant de a et de b respectivement. Les auteurs étendent ainsi la notion récursive de *feedback* de la centralité à la similarité : Deux éléments sont similaires s' ils sont référencés par des éléments similaires.

Cette approche constitue une bonne alternative aux distances basées sur les chemins les plus courts, en ne se limitant pas à un chemin unique entre deux métabolites. De plus, des adaptations ont été proposées pour prendre en compte des graphes pondérés[9], permettant de bénéficier des méthodes décrites en première partie pour la gestion des composés auxiliaires. Cependant, le SimRank est dédié à la définition de distance entre deux nœuds. Le CoSimRank[226], en revanche, est une adaptation qui permet entre autres de définir une distance entre des ensembles de nœuds, et qui s'avère également bien plus efficace et donc appropriée pour les réseaux de grande taille. Le CoSimRank peut être vu comme une combinaison du PPR et du SimRank. Il considère les différences de PageRank personnalisé (au travers de la distance cosinus ou un simple produit scalaire) obtenues à partir des deux vecteurs de personnalisations, comme proposés précédemment, mais il est calculé à partir de la somme de ces différences obtenues à chaque pas (c.-à-d..

chaque itération de la *power method*. Soit :

$$sim(V, W) = \sum_{k=0}^{\infty} c^k \langle p^{(k)}(V), p^{(k)}(W) \rangle$$

Avec V, W deux ensembles de nœuds, $p_i^{(k)}$ le vecteur de PPR à l'itération k, et c, un facteur d'atténuation qui limite l'influence des différences après un nombre trop important de pas, tels que $0 < c < 1$. [2]. L'approche du CoSimRank permettrait ainsi d'offrir une mesure de similarité qui partage de nombreuses caractéristiques avec notre méthode de recommandations basée sur le PPR, notamment vis-à-vis des chemins considérés, qui selon nous constitue le principal attrait de ces méthodes pour leur application aux réseaux métaboliques.

9.2.4 Vers une approche dynamique de l'étude du métabolisme

Une des limites de notre analyse est que les méthodes de métabolomique utilisées offrent une vue statique, alors que le métabolisme est dynamique par nature. L'une des évolutions de la métabolomique, soutenue par plusieurs développements méthodologiques récents, est la prise en compte de cette dynamique au travers de mesures étalées dans le temps, voire la mesure en temps réel.

Quelques adaptations peuvent permettre la prise en compte de ce type de données par notre système de recommandation. Il a été montré que la convergence du PageRank à une probabilité stationnaire est garantie lorsque le vecteur de priorisation varie. De ce constat a émergé la méthode de téléportation évolutive, proposée par Rossi et Gleich[225]. Cette méthode consiste à faire varier le vecteur de priorisation à chaque itération en fonction de données temporelles. En biaisant la téléportation d'après des données de consultation de page web, les auteurs ont réussi à mettre en évidence des pages Wikipedia dont l'importance est suscitée par des événements externes qui conduisent à des variations de l'intérêt des utilisateurs. Par exemple, la page Wikipedia de l'Échelle de Richter est classée à la 72e position d'importance à la période où a eu lieu un tremblement de terre en

2. valeur suggérée à 0,8 par les auteurs

Australie, alors qu'elle ne figure pas parmi les 200 pages les plus importantes, toutes périodes confondues.

Les auteurs proposent plusieurs interprétations des rangs obtenus :

— Le *Transient Rank*, qui mesure l'importance d'un acteur du réseau à l'instant t

— Le *Summary & Cumulative Rank*, qui permettent de résumer l'importance d'un acteur sur un intervalle de temps, au travers de son *Transient Rank* moyen, minimum ou maximum, ou de son rang cumulé.

— Le *Difference Rank*, qui renseigne sur la différence de rang maximum enregistré sur une période par un acteur. Cette mesure permettrait d'identifier des événements importants et les acteurs qui y sont associés, et ainsi de les distinguer des acteurs globalement importants (caractérisé par un haut *cumulative Rank*).

Ces mesures pourraient permettre une description plus fine du rôle de certains métabolites dans une condition donnée. Les auteurs proposent également l'utilisation de méthode de clustering afin d'identifier des profils d'importances, tels que la diminution ou l'augmentation de la centralité sur la période mesurée ou encore des augmentations périodiques ou ponctuelles. En identifiant des groupes de métabolites de profils similaires, il serait possible d'identifier des mécanismes physiologiques et de caractériser leurs tendances.

D'autres développements ont été conduits autour du PageRank, appliqués aux systèmes dynamiques[182]. Outre l'évolution de la priorisation proposée précédemment, la mesure de centralité sur des réseaux dont la structure même varie au cours du temps a également été proposée. L'EventRank est un exemple de ce type d'approches[243].Dans le contexte des réseaux métaboliques, elles permettraient de faire varier la disponibilité de certaines réactions enzymatiques au cours du temps. Cette disponibilité pourrait être définie, par exemple, à partir de données d'expression.

Bibliographie

[1] P. Abeles. Efficient Java Matrix Library. 2010.

[2] A. Abouelaoualim, K. Das, L. Faria, Y. Manoussakis, C. Martinhon, and R. Saad. Paths and trails in edge-colored graphs. *Theoretical Computer Science*, 409(3) :497–510, 2008.

[3] V. Acuña, E. Birmelé, L. Cottret, P. Crescenzi, F. Jourdan, V. Lacroix, A. Marchetti-Spaccamela, A. Marino, P. V. Milreu, M.-F. Sagot, and L. Stougie. Telling stories : Enumerating maximal directed acyclic graphs with a constrained set of sources and targets. *Theoretical Computer Science*, 457 :1–9, 2012.

[4] G. Adomavicius and A. Tuzhilin. Towards the Next Generation of Recommender Systems : A Survey of the State-of-the-Art and Possible Extensions. *IEEE transactions on knowledge and data engineering*, 17(6) :734–749, 2005.

[5] T. Aittokallio and B. Schwikowski. Graph-based methods for analysing networks in cell biology. *Briefings in Bioinformatics*, 7(3) :243–255, 2006.

[6] J. Albrecht and E. A. Jones. Hepatic encephalopathy : molecular mechanisms underlying the clinical syndrome. *Journal of the neurological sciences*, 170(2) :138–146, 1999.

[7] S. Allesina and M. Pascual. Googling Food Webs : Can an Eigenvector Measure Species' Importance for Coextinctions ? *Plos*, 5(9), 2009.

[8] S. F. Altschul, W. Gish, W. Miller, E. W. Myers, and D. J. Lipman. Basic

Local Alignment Search Tool. *Journal of molecular biology*, 215(3) :403–410, 1990.

[9] I. Antonellis, H. Garcia-molina, and C.-c. Chang. Simrank++ : Query Rewriting through Link Analysis of the Click Graph. *Proceedings of the VLDB Endowment*, 1(1) :408–421, 2008.

[10] A. V. Antonov, S. Dietmann, P. Wong, and H. W. Mewes. TICL–a web tool for network-based interpretation of compound lists inferred by high-throughput metabolomics. *The FEBS journal*, 276(7) :2084–94, apr 2009.

[11] I. Aretz and D. Meierhofer. Advantages and Pitfalls of Mass Spectrometry Based Metabolome Profiling in Systems Biology. *International Journal of Molecular Sciences Review*, 17(632), 2016.

[12] M. Arita. Metabolic reconstruction using shortest paths. *Simulation Practice and Theory*, 8(1) :109–125, 2000.

[13] M. Arita. The metabolic world of Escherichia coli is not small. *Proceedings of the National Academy of Sciences of the United States of America*, 101(6) :1543–1547, 2004.

[14] E. G. Armitage and C. Barbas. Metabolomics in cancer biomarker discovery : Current trends and future perspectives. *Journal of Pharmaceutical and Biomedical Analysis*, 2013.

[15] I. Avazpour, T. Pitakrat, L. Grunske, and J. Grundy. Dimensions and Metrics for Evaluating Recommendation Systems. In *Recommendation systems in software engineering*, pages 245–273. Springer Science & Business, 2014.

[16] L. Backstrom and J. Leskovec. Supervised Random Walks : Predicting and Recommending Links in Social Networks. In *Proceedings of the fourth ACM international conference on Web search and data mining*, pages 635–644, 2011.

[17] S. Baluja, R. Seth, Y. Jing, J. Yagnik, S. Kumar, D. Ravichandran, and M. Aly. Video Suggestion and Discovery for YouTube : Taking Random Walks Through the View Graph. In *Proceedings of the 17th international conference on World Wide Web*, pages 895—-904, 2008.

[18] D. Bánky, G. Iván, and V. Grolmusz. Equal Opportunity for Low-Degree Network Nodes : A PageRank-Based Method for Protein Target Identification in Metabolic Graphs. *PLoS ONE*, 8(1), 2013.

[19] A.-L. Barabási and R. Albert. Emergence of Scaling in Random Networks. *Science*, 286(5439) :509–512, 1999.

[20] M. E. Beber, G. Muskhelishvili, and M.-t. Hu. Effect of database drift on network topology and enrichment analyses : a case study for RegulonDB. *Database*, pages 1–12, 2016.

[21] D. C. Bell, J. S. Atkinson, and J. W. Carlson. Centrality measures for disease transmission networks. *Social Networks*, 21 :1–21, 1999.

[22] F. Bellomi and R. Bonato. Network Analysis for Wikipedia. In *Proceedings of Wikimania*, page 81, 2005.

[23] M. N. Benedict, M. B. Mundy, C. S. Henry, N. Chia, and N. D. Price. Likelihood-Based Gene Annotations for Gap Filling and Quality Assessment in Genome-Scale Metabolic Models. *PLoS computational biology*, 10(10), 2014.

[24] P. Berkhin. A Survey on PageRank Computing. *Internet Mathematics*, 2(1) :73–120, 2011.

[25] T. Bernard, A. Bridge, A. Morgat, S. Moretti, I. Xenarios, and M. Pagni. Reconciliation of metabolites and biochemical reactions for metabolic networks. *Briefings in bioinformatics*, 15(1) :123–35, jan 2014.

[26] M. Bianchini, M. Gori, and F. Scarselli. Inside PageRank. *ACM Transactions on Internet Technology (TOIT)*, 5(1) :92–128, 2005.

[27] T. Blum and O. Kohlbacher. MetaRoute : Fast search for relevant metabolic routes for interactive network navigation and visualization. *Bioinformatics*, 24(18) :2108–2109, 2008.

[28] T. Blum and O. Kohlbacher. Using Atom Mapping Rules for an Improved Detection of Relevant Routes in Weighted Metabolic Networks. *Journal of Computational Biology*, 15(6) :565–576, jan 2008.

[29] T. Bogers. Movie Recommendation using Random Walks over the Contextual Graph. *Proc. of the 2nd Intl. Workshop on Context-Aware Recommender Systems*, 2010.

[30] D. Bollegala, Y. Matsuo, and M. Ishizuka. Measuring Semantic Similarity between Words Using Web Search Engines. *www*, 7 :757–766, 2007.

[31] P. Bonacich. Technique for Analyzing Overlapping Memberships. *Sociological Methodology*, 1972(4) :176–185, 1972.

[32] P. Bonacich. Power and Centrality : A Family of Measures. *AJS*, 92(5) :1170–82, 1987.

[33] P. Bonacich, A. Cody Holdren, and M. Johnston. Hyper-edges and multidimensional centrality. *Social Networks*, 26(3) :189–203, 2004.

[34] A. Bordbar, J. M. Monk, Z. A. King, and B. O. Palsson. Constraint-based models predict metabolic and associated cellular functions. *Nature reviews. Genetics*, 15(2) :107–120, 2014.

[35] S. P. Borgatti. Centrality and AIDS. *Connections*, 18(1) :111–113, 1995.

[36] S. P. Borgatti. The Key Player Problem. In *Dynamic Social Network Modeling and Analysis : Workshop Summary and Papers*, 2003.

[37] S. P. Borgatti. Centrality and network flow. *Social networks*, 27(1) :55–71, 2005.

[38] S. P. Borgatti, K. M. Carley, and D. Krackhardt. On the robustness of centrality measures under conditions of imperfect data. *Social networks*, 28(2) :124–136, 2006.

[39] S. P. Borgatti and M. G. Everett. A Graph-theoretic perspective on centrality. *Social Networks*, 28(4) :466–484, oct 2006.

[40] A.-l. Boulesteix and M. Slawski. Stability and aggregation of ranked gene lists. *BRIEFINGS IN BIOINFORMATICS.*, 10(5) :556–568, 2009.

[41] F. Boyer and A. Viari. Ab initio reconstruction of metabolic pathways. *Bioinformatics*, 19(Suppl 2) :ii26–ii34, oct 2003.

[42] U. Brandes. On variants of shortest-path betweenness centrality and their generic computation. *Social Networks*, 30(2) :136–145, 2008.

[43] U. Brandes and T. Erlebach. *Network Analysis : Methodological Foundations*. Springer Science & Business, 2005.

[44] R. Breitling, S. Ritchie, D. Goodenowe, M. L. Stewart, and M. P. Barrett. Ab initio prediction of metabolic networks using Fourier transform mass spectrometry data. *Metabolomics*, 2(3) :155–164, 2006.

[45] S. Brohée, K. Faust, G. Lima-Mendez, O. Sand, R. Janky, G. Vanderstocken, Y. Deville, and J. van Helden. NeAT : a toolbox for the analysis of biological networks, clusters, classes and pathways. *Nucleic acids research*, 36(Web Server issue), 2008.

[46] E. Bullmore and O. Sporns. Complex brain networks : graph theoretical analysis of structural and functional systems. *Nature reviews. Neuroscience*, 10(3) :186–198, 2009.

[47] R. F. Butterworth, J. F. Giguere, J. Michaud, J. Lavoie, and G. P. Layrargues. Ammonia : key factor in the pathogenesis of hepatic encephalopathy. *Neurochem Pathol*, 6(1-2) :1–12, 1987.

[48] R. Caspi, T. Altman, R. Billington, K. Dreher, H. Foerster, C. a. Fulcher, T. a. Holland, I. M. Keseler, A. Kothari, A. Kubo, M. Krummenacker, M. Latendresse, L. a. Mueller, Q. Ong, S. Paley, P. Subhraveti, D. S. Weaver, D. Weerasinghe, P. Zhang, and P. D. Karp. The MetaCyc database of metabolic pathways and enzymes and the BioCyc collection of Pathway/Genome Databases. *Nucleic Acids Research*, 42(D1) :459–471, 2014.

[49] R. Caspi, K. Dreher, and P. D. Karp. The challenge of constructing, classifying and representing metabolic pathways. *FEMS Microbiol Lett.*, 345(2) :85–93, 2013.

[50] M. Chagoyen and F. Pazos. MBRole : enrichment analysis of metabolomic data. *Bioinformatics (Oxford, England)*, 27 :730–731, 2011.

[51] M. Chagoyen and F. Pazos. Tools for the functional interpretation of metabolomic experiments. *Briefings in Bioinformatics*, 14(6) :737–744, 2013.

[52] J. Chambers, M. Davies, A. Gaulton, A. Hersey, S. Velankar, R. Petryszak, J. Hastings, L. Bellis, S. McGlinchey, and J. Overington. UniChem : a unified chemical structure cross-referencing and identifier tracking system. *Journal of Cheminformatics*, 5(1) :3, 2013.

[53] N. Choon-ching and A. Selamat. Text Summarization Review. Technical report, Faculty of Computer Science and Information System, Universiti Teknologi Malaysia, 2005.

[54] A. Cockburn, A. M. Y. Karlson, and B. B. Bederson. A Review of Overview+Detail, Zooming, and Focus+Context Interfaces. *ACM Comput. Surv.*, 41(1) :1–42, 2007.

[55] A. J. L. Cooper and T. Kuhara. α-Ketoglutaramate : An overlooked metabolite of glutamine and a biomarker for hepatic encephalopathy and inborn errors of the urea cycle. *Metabolic brain disease*, 29(4) :991–1006, 2015.

[56] A. J. L. Cooper, Y. I. Shurubor, T. Dorai, J. T. Pinto, E. P. Isakova, Y. I. Deryabina, T. T. Denton, and B. F. Krasnikov. ω-Amidase : an underappreciated, but important enzyme in l-glutamine and l-asparagine metabolism ; relevance to sulfur and nitrogen metabolism, tumor biology and hyperammonemic diseases. *Amino Acids*, 48(1) :1–20, 2016.

[57] C. Corley and R. Mihalcea. Measuring the Semantic Similarity of Texts. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 13–18, 2005.

[58] L. Cottret and F. Jourdan. Graph methods for the investigation of metabolic networks in parasitology. *Parasitology*, 137(9) :1393–1407, 2010.

[59] L. Cottret, D. Wildridge, F. Vinson, M. P. Barrett, H. Charles, M.-F. Sagot, and F. Jourdan. MetExplore : a web server to link metabolomic experiments and genome-scale metabolic networks. *Nucleic acids research*, 38(Web Server issue) :W132–7, jul 2010.

[60] D. J. Creek, W. B. Dunn, O. Fiehn, J. L. Griffin, R. D. Hall, Z. Lei, R. Mistrik, S. Neumann, E. L. Schymanski, L. W. Sumner, R. Trengove, J.-l. Wolfender, and V. D. Hooft. Metabolite identification : are you sure ? And how do your peers gauge your confidence ? *Metabolomics*, 10 :350–353, 2014.

[61] P. Cremonesi, F. Garzotto, S. Negro, A. Papadopoulos, and R. Turrin. Comparative Evaluation of Recommender System Quality. In *CHI'11 Extended Abstracts on Human Factors in Computing Systems*, pages 1927—-1932, 2011.

[62] T. Crnovrsanin, I. Liao, and Y. Wu. Visual Recommendations for Network Navigation. *Eurographics / IEEE Symposium on Visualization*, 30(3), 2011.

[63] D. Croes, F. Couche, S. J. Wodak, and J. van Helden. Metabolic PathFinding : inferring relevant pathways in biochemical networks. *Nucleic acids research*, 33(Web Server issue) :W326—-W330, jul 2005.

[64] D. Croes, F. Couche, S. J. Wodak, and J. Van Helden. Inferring meaningful pathways in weighted metabolic networks. *Journal of Molecular Biology*, 356(1) :222–236, 2006.

[65] G. Csardi and T. Nepusz. The igraph software package for complex network research. *InterJournal, Complex Systems*, 1695(5) :1–9, 2006.

[66] L. F. de Figueiredo, S. Schuster, C. Kaleta, and D. a. Fell. Can sugars be produced from fatty acids ? A test case for pathway analysis tools. *Bioinformatics*, 25(24) :3330–3331, 2009.

[67] L. F. de Figueiredo, S. Schuster, C. Kaleta, and D. a. Fell. Response to comment on 'Can sugars be produced from fatty acids ? A test case for pathway analysis tools'. *Bioinformatics*, 25(24) :3330–3331, 2009.

[68] Y. Deville, D. Gilbert, J. van Helden, and S. J. Wodak. An overview of data models for the analysis of biochemical pathways. *Briefings in bioinformatics*, 4(3) :246–259, 2003.

[69] E. W. Dijkstra. A note on two problems in connexion with graphs. *Numerische Mathematik*, 1(1) :269–271, 1959.

[70] A. Dräger, N. Rodriguez, M. Dumousseau, A. Dörr, C. Wrzodek, N. Le Novère, A. Zell, and M. Hucka. JSBML : a flexible Java library for working with SBML. *Bioinformatics*, 27(15) :2167–2168, 2011.

[71] J. Duan, S. L. Dixon, J. F. Lowrie, and W. Sherman. Analysis and comparison of 2D fingerprints : insights into database screening performance using eight fingerprint methods. *Journal of molecular graphics & modelling*, 29 :157–170, 2010.

[72] C. Dunne and B. Shneiderman. Improving graph drawing readability by incorporating readability metrics : A software tool for network analysts. *University of Maryland, HCIL Tech Report HCIL-2009-13*, 2009.

[73] P. Dupont, J. Callut, G. Dooms, J.-N. Monette, and Y. Deville. Relevant subgraph extraction from random walks in a graph. Technical report, Université catholique de Louvain, 2006.

[74] D. I. Ellis, W. B. Dunn, J. L. Griffin, J. W. Allwood, and R. Goodacre. Metabolic fingerprinting as a diagnostic tool. *Pharmacogenomics*, 8(9) :1243–66, sep 2007.

[75] M. A. Ellul, S. A. Gholkar, and T. J. Cross. Hepatic encephalopathy due to liver cirrhosis. *Bmj*, 351(August) :h4187, 2015.

[76] D. Eppstein. Finding the k Shortest Paths. *SIAM*, 28(2) :652–673, 1997.

[77] L. Ermann, A. D. Chepelianskii, and D. L. Shepelyansky. Towards two-dimensional search engines. *Journal of Physics A : Mathematical and Theoretical*, 45(27) :275101, 2012.

[78] L. Ermann and D. L. Shepelyansky. Google matrix of the world trade network. *European Physical Journal B*, 1 :14, 2011.

[79] E. Estrada, D. J. Higham, and N. Hatano. Communicability betweenness in complex networks. *Physica A : Statistical Mechanics and its Applications*, 388(5) :764–774, mar 2009.

[80] K. Faust, D. Croes, and J. van Helden. In response to 'Can sugars be produced from fatty acids ? A test case for pathway analysis tools'. *Bioinformatics*, 25(23) :3202–3205, 2009.

[81] K. Faust, D. Croes, and J. van Helden. Metabolic pathfinding using RPAIR annotation. *Journal of molecular biology*, 388(2) :390–414, may 2009.

[82] K. Faust, D. Croes, and J. van Helden. Prediction of metabolic pathways from genome-scale metabolic networks. *Bio Systems*, 105(2) :109–21, aug 2011.

[83] K. Faust, P. Dupont, J. Callut, and J. van Helden. Pathway discovery in metabolic networks by subgraph extraction. *Bioinformatics (Oxford, England)*, 26(9) :1211–8, may 2010.

[84] S. Fekete, T. Kamphans, and M. Stelzer. Shortest Paths with Pairwise-Distinct Edge Labels : Finding Biochemical Pathways in Metabolic Networks. *arXiv preprint arXiv :1012.5024*, page 9, 2010.

[85] D. a. Fell and a. Wagner. The small world of metabolism. *Nature biotechnology*, 18(11) :1121–1122, 2000.

[86] O. Fiehn. Metabolomics – the link between genotypes and phenotypes. *Plant Molecular Biology*, 48 :155–171, 2002.

[87] D. Flower. On the Properties of Bit String-Based Measures of Chemical Similarity. *Journal of Chemical Information and Modeling*, 38(3) :379–386, may 1998.

[88] U. Fößmeier and M. Kaufmann. Drawing High Degree Graphs with Low Bend Numbers. In *International Symposium on Graph Drawing*, pages 254–266. Springer Berlin Heidelberg, 1995.

[89] C. Frainay and F. Jourdan. Computational methods to identify metabolic sub-networks based on metabolomic profiles. *Briefings in Bioinformatics*, 18(1) :43–56, 2016.

[90] L. C. Freeman. A Set of Measures of Centrality Based on Betweenness. *Sociometry*, 40(1) :35–41, 1977.

[91] L. C. Freeman, S. P. Borgatti, and D. R. White. Centrality in valued graphs : A measure of betweenness based on network flow. *Social Networks*, 13(2) :141–154, jun 1991.

[92] L. C. Freeman and S. Smith. Centrality in Social Networks Conceptual Clarification. *Social networks*, 1(3) :215–239, 1979.

[93] V. Freschi. Protein function prediction from interaction networks using a random walk ranking algorithm. *Proceedings of the 7th IEEE International Conference on Bioinformatics and Bioengineering*, pages 42–48, 2007.

[94] T. M. Fruchterman and E. M. Reingold. Graph Drawing by Force-directed Placement. *Software : Practice and experience*, 21(11) :1129–1164, 1991.

[95] G. W. Furnas. *Generalized Fisheye Views*. Number 17. ACM, 1986.

[96] T. Gaasterland and E. Selkov. Reconstruction of Metabolic Networks Using Incomplete Information. *Proc. Int. Conf. Intell. Syst. Mol. Siol.*, 3 :127–135, 1995.

[97] G. Gallo, G. Longo, S. Pallottino, and S. Nguyen. Directed hypergraphs applications. *Discrete Applied Mathematics*, 42 :177–201, 1993.

[98] M. Á. García-sevillano, T. García-barrera, N. Abril, C. Pueyo, J. López-barea, and J. L. Gómez-ariza. Omics technologies and their applications to evaluate metal toxicity in mice M. spretus as a bioindicator. *Journal of Proteomics*, 2014.

[99] M. D. Gemmis, P. Lops, C. Musto, F. Narducci, and G. Semeraro. Semantics-aware Content-based Recommender Systems. In *Recommender Systems Handbook*, pages 119–159. Springer US, 2015.

[100] H. Ginsburg. Caveat emptor : limitations of the automated reconstruction of metabolic pathways in Plasmodium. *Trends in Parasitology*, 25(1) :37–43, 2009.

[101] D. F. Gleich. PageRank Beyond the Web. *Society for Industrial and Applied Mathematics*, 57(3) :321–363, 2005.

[102] G. L. Glish and R. W. Vachet. The basics of mass spectrometry in the twenty- first century. *Nature reviews. Drug discovery*, 2(February) :140–150, 2003.

[103] A. Goel, P. Gupta, J. Sirois, D. Wang, A. Sharma, S. Gurumurthy, A. Goel, P. Gupta, J. Sirois, D. Wang, and A. Sharma. The Who-To-Follow System at Twitter : Strategy, Algorithms, and Revenue Impact. *Interfaces*, 45(1) :98–107, 2015.

[104] W. H. Gomaa and A. A. Fahmy. A Survey of Text Similarity Approaches. *International Journal of Computer Applications*, 68(13) :13–18, 2013.

[105] G. H. Gonzalez, T. Tahsin, B. C. Goodale, A. C. Greene, and C. S. Greene. Recent Advances and Emerging Applications in Text and Data Mining for Biomedical Discovery. *Briefings in bioinformatics*, 17(1) :33–42, 2016.

[106] M. L. Green and P. D. Karp. A Bayesian method for identifying missing enzymes in predicted metabolic pathway databases. *BMC Bioinformatics*, 5(76) :1–16, 2004.

[107] P. Gupta, A. Goel, J. Lin, A. Sharma, D. Wang, and R. Zadeh. WTF : The Who to Follow Service at Twitter. In *Proceedings of the 22nd international conference on World Wide Web*, pages 505–514, 2013.

[108] Z. Gyöngyi, H. Garcia-Molina, and J. Pedersen. Combating Web Spam with TrustRank. In *Proceedings of the Thirtieth international conference on Very large data bases-Volume 30*, pages 576—-587, 2004.

[109] A. Hagberg, P. Swart, and D. S Chult. Exploring network structure, dynamics, and function using NetworkX. Technical report, Los Alamos National Laboratory (LANL), 2008.

[110] C. R. Haggart, J. A. Bartell, J. J. Saucerman, and J. A. Papin. Whole-genome metabolic network reconstruction and constraint-based modeling. *Methods in enzymology*, 500 :411, 2011.

[111] F. V. Ham and A. Perer. "Search, Show Context, Expand on Demand" : Supporting Large Graph Exploration with Degree-of-Interest. *IEEE Transactions on Visualization and Computer Graphics*, 15(6) :953–960, 2009.

[112] T. Handorf, O. Ebenhöh, and R. Heinrich. Expanding Metabolic Networks : Scopes of Compounds, Robustness, and Evolution. *Journal of molecular evolution*, 61(4) :498–512, 2005.

[113] H. S. Haraldsdóttir, I. Thiele, and R. M. Fleming. Comparative evaluation of open source software for mapping between metabolite identifiers in metabolic network reconstructions : application to Recon 2. *Journal of cheminformatics*, 6(1) :2, jan 2014.

[114] T. H. Haveliwala. Topic-sensitive PageRank. *Proceedings of the 11th international conference on World Wide Web*, pages 517–526, 2002.

[115] A. P. Heath, G. N. Bennett, and L. E. Kavraki. Finding metabolic pathways using atom tracking. *Bioinformatics*, 26(12) :1548–1555, jun 2010.

[116] S. R. Heller and A. D. McNaught. The IUPAC International Chemical Identifier, InChI. *The ACS Style Guide*, 3 :p. 101–102, jan 2006.

[117] M. P. V. D. Heuvel and O. Sporns. Network hubs in the human brain. *Trends in Cognitive Sciences*, 17(12) :683–696, 2013.

[118] C. Hoede. A new status score for actors in a social network. *Twente University Department of Applied Mathematics (Memorandum no. 243)*, 1978.

[119] Å. J. Holmgren. Using Graph Models to Analyze the Vulnerability of Electric Power Networks. *Risk Analysis*, 26(4), 2006.

[120] A. L. Hopkins. Network pharmacology : the next paradigm in drug discovery. *Nature chemical biology*, 4(11) :682–690, 2008.

[121] H. Horai, M. Arita, S. Kanaya, Y. Nihei, T. Ikeda, K. Suwa, Y. Ojima, K. Tanaka, S. Tanaka, K. Aoshima, Y. Oda, Y. Kakazu, M. Kusano, T. Tohge, F. Matsuda, Y. Sawada, M. Y. Hirai, H. Nakanishi, K. Ikeda, N. Akimoto, T. Maoka, H. Takahashi, T. Ara, N. Sakurai, H. Suzuki, D. Shibata, S. Neumann, T. Iida, K. Tanaka, K. Funatsu, F. Matsuura, T. Soga, R. Taguchi, K. Saito, and T. Nishioka. MassBank : a public repository for sharing mass

spectral data for life sciences. *Journal of mass spectrometry*, 45(7) :703–714, 2010.

[122] A. Hotho, C. Schmitz, and G. Stumme. FolkRank : A Ranking Algorithm for Folksonomies. *LWA*, 1 :111–114, 2006.

[123] C.-c. Huang and Z. Lu. Community challenges in biomedical text mining over 10 years : success, failure and the future. *Briefings in Bioinformatics*, pages 1–13, 2015.

[124] D. W. Huang, B. T. Sherman, and R. A. Lempicki. Bioinformatics enrichment tools : paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1) :1–13, 2009.

[125] D. W. Huang, B. T. Sherman, and R. a. Lempicki. Bioinformatics enrichment tools : Paths toward the comprehensive functional analysis of large gene lists. *Nucleic Acids Research*, 37(1) :1–13, 2009.

[126] W. Huang and P. Eades. How People Read Graphs. *APVis '05 proceedings of the 2005 Asia-Pacific symposium on Information visualisation*, pages 51–58, 2005.

[127] W. Huang, P. Eades, and S. H. Hong. A graph reading behavior : Geodesic-path tendency. *IEEE Pacific Visualization Symposium, PacificVis 2009 - Proceedings*, pages 137–144, 2009.

[128] C. H. Hubbell. An Input-Output Approach to Clique Identification. *Sociometry*, 28(4) :377–399, 1965.

[129] M. Hucka, A. Finney, H. M. Sauro, H. Bolouri, J. C. Doyle, H. Kitano, B. J. Bornstein, D. Bray, A. A. Cuellar, S. Dronov, E. D. Gilles, M. Ginkel, V. Gor, I. I. Goryanin, W. J. Hedley, T. C. Hodgman, P. J. Hunter, N. S. Juty, J. L. Kasberger, A. Kremling, U. Kummer, L. M. Loew, D. Lucio, P. Mendes, E. Minch, E. D. Mjolsness, Y. Nakayama, M. R. Nelson, P. F. Nielsen, T. Sakurada, J. C. Schaff, B. E. Shapiro, T. S. Shimizu, H. D.

Spence, J. Stelling, K. Takahashi, M. Tomita, J. Wagner, and J. Wang. The systems biology markup language (SBML) : a medium for representation and exchange of biochemical network models. *Bioinformatics*, 19(4) :524–531, 2003.

[130] T. Hughes and D. Ramage. Lexical Semantic Relatedness with Random Graph Walks. In *Proceedings of the 2007 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 581–589, 2007.

[131] S. Hummert, K. Bohl, D. Basanta, A. Deutsch, S. Werner, G. Theißen, A. Schroeterc, S. Schuster, and In. Evolutionary game theory : cells as players. *Molecular bioSystems*, 10 :3044–3065, 2014.

[132] Y. Iturria-medina and R. C. Sotero. Studying the human brain anatomical network via diffusion-weighted MRI and Graph Theory. *NeuroImage*, 40 :1064–1076, 2008.

[133] G. Iván and V. Grolmusz. When the web meets the cell : Using personalized PageRank for analyzing protein interaction networks. *Bioinformatics*, 27(3) :405–407, 2011.

[134] G. Jeh and J. Widom. SimRank : A Measure of Structural-Context Similarity. In *Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 538–543, 2002.

[135] H. Jeong, S. P. Mason, A.-L. Barabási, and Z. N. Oltvai. Lethality and centrality in protein networks. *Nature*, 411(May) :41–42, 2001.

[136] H. Jeong, B. Tombor, R. Albert, Z. Oltvai, and A. Barabasi. The large-scale organization of metabolic networks. *Nature*, 407(6804) :651–654, 2000.

[137] C. H. Johnson, J. Ivanisevic, and G. Siuzdak. Metabolomics : beyond biomarkers and towards mechanisms. *Nature reviews Molecular cell biology*, 17(7) :451–459, 2016.

[138] M. P. Joy, A. Brock, D. E. Ingber, and S. Huang. High-Betweenness Proteins in the Yeast Protein Interaction Network. *Journal of Biomedicine and Biotechnology*, 2 :96–103, 2005.

[139] A. Jude. Evaluating Item-Item Similarity Algorithms for Movies. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems*, pages 2141–2147, 2016.

[140] D. Jungnickel. *Graphs, Networks and Algorithms*. Springer Berlin Heidelberg, 3 edition, 2010.

[141] M. Kanehisa and S. Goto. KEGG : Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 27(1) :29–34, 2000.

[142] M. Kanehisa, S. Goto, M. Hattori, K. F. Aoki-Kinoshita, M. Itoh, S. Kawashima, T. Katayama, M. Araki, and M. Hirakawa. From genomics to chemical genomics : new developments in KEGG. *Nucleic acids research*, 34(Database issue) :D354–D357, 2006.

[143] L. Katz. A New Status Index Derived From Sociometric Analysis. *PSYCHOMETRIKA*, 18(1), 1953.

[144] D. B. Kell. Metabolomics and systems biology : making sense of the soup. *Current Opinion in Microbiology*, 7 :296–307, 2004.

[145] M. G. Kendall. A new measure of rank correlation. *Biometrika*, 30(1) :81–93, 1938.

[146] I. M. Keseler, J. Collado-vides, S. Gama-castro, J. Ingraham, P. D. Karp, S. Paley, and I. T. Paulsen. EcoCyc : a comprehensive database resource for Escherichia coli. *Nucleic Acids Research*, 33 :334–337, 2005.

[147] H. K. Kim, Y. H. Choi, and R. Verpoorte. NMR-based metabolomic analysis of plants. *Nature protocols*, 5(3) :536, 2010.

[148] H. L. Kirschenlohr, J. L. Griffin, S. C. Clarke, R. Rhydwen, A. A. Grace, P. M. Schofield, K. M. Brindle, and J. C. Metcalfe. Proton NMR analysis of plasma is a weak predictor of coronary artery disease. *Nature medicine*, 12(6) :705–711, 2006.

[149] S. Klamt, U.-U. Haus, and F. Theis. Hypergraphs and Cellular Networks. *PLoS Computational Biology*, 5(5) :e1000385, 2009.

[150] D. J. Klein. Centrality measure in graphs. *Journal of Mathematical Chemistry*, 47(4) :1209–1223, mar 2010.

[151] J. M. Kleinberg. Authoritative Sources in a Hyperlinked Environment. *Journal of the ACM*, 46(5) :604–632, 1999.

[152] J. Klekota and F. P. Roth. Chemical substructures that enrich for biological activity. *Bioinformatics*, 24(21) :2518–2525, 2008.

[153] K. Komurov, M. A. White, and P. T. Ram. Use of data-biased random walks on graphs for the retrieval of context-specific networks from genomic data. *PLoS computational biology*, 6(8) :10, jan 2010.

[154] D. Koschützki, K. A. Lehmann, L. Peeters, S. Richter, D. Tenfelde-Podehl, and O. Zlotowski. Centrality Indices. *Network analysis*, pages 16–61, 2005.

[155] D. Koschützki and F. Schreiber. Centrality analysis methods for biological networks and their application to gene regulatory networks. *Gene regulation and systems biology*, 2 :193–201, 2008.

[156] M. Kotera, M. Hattori, M. Oh, and R. Yamamoto. RPAIR : a reactant-pair database representing chemical changes in enzymatic reactions. *Genome Informatics*, 2004.

[157] M. Krallinger and A. Valencia. Text-mining and information-retrieval services for molecular biology. *Genome Biology*, 6(224), 2005.

[158] V. S. Kumar, M. S. Dasika, and C. D. Maranas. Optimization based automated curation of metabolic reconstructions. *BMC Bioinformatics*, 8(212) :1–16, 2007.

[159] V. Lacroix. *Identification de motifs dans les réseaux métaboliques.* PhD thesis, Université Claude Bernard - Lyon 1, 2007.

[160] V. Lacroix, L. Cottret, P. Thébault, and M.-F. Sagot. An introduction to metabolic networks and their structural analysis. *IEEE/ACM transactions on computational biology and bioinformatics / IEEE, ACM*, 5(4) :594–617, 2008.

[161] V. Lacroix, C. G. Fernandes, and M.-f. Sagot. Motif Search in Graphs : Application to Metabolic Networks. In *EEE/ACM Transactions on Computational Biology and Bioinformatics*, volume 3, pages 360–368, 2006.

[162] T. V. Landesberger, A. Kuijper, T. Schreck, J. V. Wijk, J.-d. Fekete, and D. Fellner. Visual Analysis of Large Graphs : State-of-the-Art and Future Research Challenges. *Computer graphics forum*, 30(6) :1719–1749, 2012.

[163] M. M. Larive, Cynthia K and Barding Jr, Gregory A and Dinges. NMR Spectroscopy for Metabolomics and Metabolic Profiling. *Analytical chemistry*, 87(1) :133–146, 2014.

[164] M. Latendresse, M. Krummenacker, and P. D. Karp. Optimal metabolic route search based on atom mappings. *Bioinformatics (Oxford, England)*, 30(14) :2043–50, jul 2014.

[165] G. Lawyer. Understanding the influence of all nodes in a network. *Scientific reports*, 5(8665), 2015.

[166] M. D. Lee, B. Pincombe, and M. Welsh. An Empirical Evaluation of Models of Text Document Similarity. *Proceedings of the Cognitive Science Society*, 27(27) :1254–1259, 2005.

[167] V. I. Levenshtein. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet physics doklady*, 10(8), 1966.

[168] N. E. Lewis, H. Nagarajan, and B. O. Palsson. Constraining the metabolic genotype-phenotype relationship using a phylogeny of in silico methods. *Nature reviews. Microbiology*, 10(4) :291–305, 2013.

[169] B. Liang, Y. Liu, M. Zhang, S. Ma, L. Ru, and K. Zhang. Searching for people to follow in social networks. *EXPERT SYSTEMS WITH APPLICATIONS*, 41(16) :7455–7465, 2014.

[170] D. Liben-nowell and J. Kleinberg. The Link-Prediction Problem for Social Networks. *JOURNAL OF THE AMERICAN SOCIETY FOR INFORMATION SCIENCE AND TECHNOLOGY*, 58(7) :1019–1031, 2007.

[171] E. D. Liddy. Natural Language Processing. In *Encyclopedia of Library and Information Science, 2nd Ed.* Marcel Decker, Inc., 2001.

[172] G. Lima-Mendez and J. van Helden. The powerful law of the power law and other myths in network biology. *Molecular bioSystems*, 5(12) :1482–1493, 2009.

[173] E. Lloret and M. Palomar. Text summarisation in progress : a literature review. *Artificial Intelligence Review*, 37(1) :1–41, 2012.

[174] L. Lovasz. Random Walks on Graphs : A Survey. *Combinatorics*, 2 :1–46, 1993.

[175] L. Lü, M. Medo, C. H. Yeung, Y.-c. Zhang, Z.-k. Zhang, and T. Zhou. Recommender systems. *Physics Reports*, 519(1) :1–49, 2012.

[176] A. Lyra and C. A. Martinhon. On paths, trails and closed trails in edge-colored graphs. *Discrete Mathematics and Theoretical Computer Science*, 14(2) :57–74, 2012.

[177] H. Ma and A. P. Zeng. Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2) :270–277, 2003.

[178] D. Machado, R. S. Costa, M. Rocha, E. C. Ferreira, B. Tidor, and I. Rocha. Modeling formalisms in Systems Biology. *AMB Express*, 1(1) :45, 2011.

[179] D. MacKay. Monte Carlo Methods. In *Information Theory, Inference, and Learning Algorithms*, chapter 29, pages 358–386. Cambridge University Press, 2003.

[180] A. Maertens, M. Bouhifd, L. Zhao, S. Odwin-dacosta, A. Kleensang, J. D. Yager, and T. Hartung. Metabolomic network analysis of estrogen-stimulated MCF-7 cells : a comparison of overrepresentation analysis, quantitative enrichment analysis and pathway analysis versus metabolite network analysis. *Archives of Toxicology*, 91(1) :217–320, 2016.

[181] A. Malm and G. Bichler. Networks of Collaborating Criminals : Assessing the Structural Vulnerability of Drug Markets. *Journal of Research in Crime and Delinquency*, 48(2) :271–297, 2011.

[182] M. S. Mariani, M. Medo, and Y.-c. Zhang. Ranking nodes in growing networks : When PageRank fails. *Scientific reports*, 5 :1–10, 2015.

[183] D. McShan, S. Rao, and I. Shah. PathMiner : predicting metabolic pathways by heuristic search. *Bioinformatics*, 19(13) :1692–1698, 2003.

[184] B. Merlet, N. Paulhe, F. Vinson, C. Frainay, M. Chazalviel, N. Poupin, Y. Gloaguen, F. Giacomoni, and F. Jourdan. A Computational Solution to Automatically Map Metabolite Libraries in the Context of Genome Scale Metabolic Networks. *Frontiers in Molecular Biosciences*, 3(2) :1–12, 2016.

[185] R. Milo, S. Shen-Orr, S. Itzkovitz, N. Kashtan, D. Chklovskii, and U. Alon. Network motifs : simple building blocks of complex networks. *Science*, 298(5594) :824–827, 2002.

[186] P. V. Milreu, C. C. Klein, L. Cottret, V. Acuña, E. Birmelé, M. Borassi, C. Junot, A. Marchetti-Spaccamela, A. Marino, L. Stougie, F. Jourdan, P. Crescenzi, V. Lacroix, and M.-F. Sagot. Telling metabolic stories to explore metabolomics data : a case study on the yeast response to cadmium exposure. *Bioinformatics (Oxford, England)*, 30(1) :61–70, jan 2014.

[187] R. Mishra, J. Bian, M. Fiszman, C. R. Weir, S. Jonnalagadda, J. Mostafa, and G. D. Fiol. Text summarization in the biomedical domain : A systematic review of recent research. *Journal of Biomedical Informatics*, 52 :457–467, 2014.

[188] B. B. Misra and J. J. van der Hooft. Updates in metabolomics tools and resources : 2014–2015. *Electrophoresis*, 37(1) :86–110, 2016.

[189] A. Mithani, J. Hein, and G. M. Preston. Comparative analysis of metabolic networks provides insight into the evolution of plant pathogenic and nonpathogenic lifestyles in Pseudomonas. *Molecular Biology and Evolution*, 28(1) :483–499, 2011.

[190] A. Mithani, G. M. Preston, and J. Hein. Rahnuma : hypergraph-based tool for metabolic pathway prediction and network comparison. *Bioinformatics (Oxford, England)*, 25(14) :1831–2, jul 2009.

[191] J. Monk, J. Nogales, and B. O. Palsson. Optimizing genome-scale network reconstructions. *Nature biotechnology*, 32(5) :447–452, 2014.

[192] Y. Moreau and L.-c. Tranchevent. Computational tools for prioritizing candidate genes : boosting disease gene discovery. *Nature Reviews Genetics*, 13(8) :1–14, 2012.

[193] A. Morgat, E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari. UniPathway : a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Research*, 40 :761–769, 2012.

[194] J. L. Morrison, R. Breitling, D. J. Higham, and D. R. Gilbert. GeneRank : Using search engine technology for the analysis of microarray experiments. *BMC Bioinformatics*, 14 :1–14, 2005.

[195] P. Murray, F. Mcgee, and A. G. Forbes. A taxonomy of visualization tasks for the analysis of biological pathway data. *BMC bioinformatics*, 18(21) :1–13, 2016.

[196] M. Najdt, I. Ashrafosman, A. Ali, and A. Afnizanfaizal. Collaborative Filtering : Techniques and Applications. In *International Conference on Communication, Control, Computing and Electronics Engineering (ICCCCEE)*, pages 1–6, 2017.

[197] G. Navarro. A Guided Tour to Approximate String Matching. *ACM computing surveys (CSUR)*, 33(1) :31–88, 2001.

[198] M. J. Newman. A measure of betweenness centrality based on random walks. *Social Networks*, 27(1) :39–54, 2005.

[199] A. Y. Ng, A. X. Zheng, and M. I. Jordan. Stable Algorithms for Link Analysis. In ACM Press, editor, *Proceedings of the 24th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 258–266, New York, New York, USA, 2001.

[200] J. K. Nicholson, J. C. Lindon, and E. Holmes. 'Metabonomics' : understanding the metabolic responses of living systems to pathophysiological stimuli via multivariate statistical analysis of biological NMR spectroscopic data. *Xenobiotica*, 29(11) :1181–1189, 1999.

[201] E. Noor, E. Eden, R. Milo, and U. Alon. Central Carbon Metabolism as a Minimal Biochemical Walk between Precursors for Biomass and Energy. *Molecular Cell*, 39(5) :809–820, 2010.

[202] M. A. Oberhardt and J. A. Papin. Applications of genome-scale metabolic reconstructions. *Molecular Systems Biology*, 5(320) :1–15, 2009.

[203] N. M. O'Boyle. Towards a Universal SMILES representation - A standard method to generate canonical SMILES based on the InChI. *Journal of cheminformatics*, 4(1) :22, 2012.

[204] J. D. Orth, I. Thiele, and B. Ø. Palsson. What is flux balance analysis ? *Nature Publishing Group*, 28(3) :245–248, 2010.

[205] J. F. Padgett and C. K. Ansell. Robust Action and the Rise of the Medici, 1400-1434. *The American Journal of Sociology*, 98(6) :1259–1319, 1993.

[206] L. Page, S. Brin, R. Motwani, and T. Winograd. The PageRank Citation Ranking : Bringing Order to the Web. Technical report, Stanford InfoLab, 1999.

[207] G. a. Pavlopoulos, M. Secrier, C. N. Moschopoulos, T. G. Soldatos, S. Kossida, J. Aerts, R. Schneider, and P. G. Bagos. Using graph theory to analyze biological networks. *BioData mining*, 4(1) :10, 2011.

[208] N. Pearcy, J. J. Crofts, and N. Chuzhanova. Hypergraph Models of Metabolism. *International Journal of Biological, Biomolecular, Agricultural, Food and Biotechnological Engineering*, 8(8) :812–816, 2014.

[209] H. Pearson. Meet the human metabolome. *Nature*, 446(2) :8, 2007.

[210] D. a. Pertusi, a. E. Stine, L. J. Broadbelt, and K. E. J. Tyo. Efficient searching and annotation of metabolic networks using chemical similarity. *Bioinformatics*, 31(7) :1–9, nov 2014.

[211] J. Pey, J. Prada, J. E. Beasley, and F. J. Planes. Path finding methods accounting for stoichiometry in metabolic networks. *Genome Biology*, 12(5) :R49, 2011.

[212] T. Pfeiffer, O. S. Soyer, and S. Bonhoeffer. The evolution of connectivity in metabolic networks. *PLoS biology*, 3 :e228, 2005.

[213] E. Pitkänen, P. Jouhten, and J. Rousu. Inferring branching pathways in genome-scale metabolic networks. *BMC systems biology*, 3(1) :103, jan 2009.

[214] E. Pitkänen, A. Rantanen, J. Rousu, and E. Ukkonen. Finding feasible pathways in metabolic networks. In *Panhellenic Conference on Informatics*, pages 123–133, 2005.

[215] F. J. Planes and J. E. Beasley. A critical examination of stoichiometric and path-finding approaches to metabolic pathways. *Briefings in Bioinformatics*, 9(5) :422–436, 2008.

[216] L. Plaza, A. Díaz, and P. Gervás. A semantic graph-based approach to biomedical summarisation. *Artificial Intelligence In Medicine*, 53(1) :1–14, 2011.

[217] G. A. Preciat Gonzalez, A. Noronha, and I. Thiele. Comparative evaluation of atom mapping algorithms for metabolic reactions : application to Recon 3D. *Journal of Cheminformatics*, 9(1) :39, 2017.

[218] N. D. Price, J. Schellenberger, and B. O. Palsson. Uniform Sampling of Steady-State Flux Spaces : Means to Design Experiments and to Interpret Enzymopathies. *Biophysical Journal*, 87(4) :2172–2186, 2004.

[219] S. A. Rahman, P. Advani, R. Schunk, R. Schrader, and D. Schomburg. Metabolic pathway analysis web service (Pathway Hunter Tool at CUBIC). *Bioinformatics (Oxford, England)*, 21(7) :1189–93, apr 2005.

[220] S. A. Rahman and D. Schomburg. Observing local and global properties of metabolic pathways : 'Load points' and 'choke points' in the metabolic networks. *Bioinformatics*, 22(14) :1767–1774, 2006.

[221] K. Raman and N. Chandra. Flux balance analysis of biological systems : applications and challenges. *Briefings in bioinformatics*, 10(4) :435–449, 2009.

[222] E. Ravasz, A. L. Somera, D. A. Mongru, Z. N. Oltvai, and A. L. Barabási. Hierarchical organization of modularity in metabolic networks. *science*, 5586(2002) :1551—-1555, 2002.

[223] A. Ravikrishnan and K. Raman. Critical assessment of genome-scale metabolic networks : the need for a unified standard. *Briefings in Bioinformatics*, 16(November 2014) :1–12, 2015.

[224] Z. Razaghi-moghadam, R. Abdollahi, S. Goliaei, and M. Ebrahimi. HybridRanker : Integrating network topology and biomedical knowledge to prioritize cancer candidate genes. *Journal of Biomedical Informatics*, 64 :139–146, 2016.

[225] R. Rossi and D. Gleich. Dynamic PageRank using Evolving Teleportation, Algorithms and Models for the Web Graph. *Lecture Notes in Computer Science*, 7323 :126–137, 2012.

[226] S. Rothe and H. Schütze. CoSimRank : A Flexible & Efficient Graph-Theoretic Similarity Measure. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, pages 1392–1402, 2014.

[227] R. B. Rothenberg, J. J. Potterat, D. E. Woodhouse, W. W. Darrow, S. Q. Muth, and A. S. Klovdahl. Choosing a centrality measure : Epidemiologic correlates in the Colorado Springs study of social networks. *Social networks*, 17(February 1994) :273–297, 1995.

[228] E. Ruppin, J. A. Papin, L. F. D. Figueiredo, and S. Schuster. Metabolic reconstruction, constraint-based analysis and game theory to probe genome-scale metabolic networks. *Current Opinion in Biotechnology*, 21(4) :502–510, 2010.

[229] P. Saraiya, C. North, and K. Duca. Visualizing biological pathways : requirements analysis, systems evaluation and research agenda. *Information Visualization*, 4(April) :191–205, 2005.

[230] M. A. Sartor, A. Ade, Z. Wright, D. States, G. S. Omenn, B. Athey, and A. Karnovsky. Metab2MeSH : annotating compounds with medical subject headings. *Bioinformatics (Oxford, England)*, 28(10) :1408–1410, 2012.

[231] B. Sarwar, G. Karypis, J. Konstan, and J. Riedl. Item-Based Collaborative Filtering Recommendation Algorithms. In *Proceedings of the 10th international conference on World Wide Web*, pages 285—-295, 2001.

[232] Y. Sawada, R. Nakabayashi, Y. Yamada, M. Suzuki, M. Sato, A. Sakata, K. Akiyama, T. Sakurai, F. Matsuda, T. Aoki, M. Y. Hirai, and K. Saito. RIKEN tandem mass spectral database (ReSpect) for phytochemicals : A plant-specific MS/MS-based data resource and database. *Phytochemistry*, 82 :38–45, 2012.

[233] A. Scalbert, L. Brennan, O. Fiehn, T. Hankemeier, B. S. Kristal, B. Van Ommen, E. Pujos-Guillot, E. Verheij, D. Wishart, and S. Wopereis. Mass-spectrometry-based metabolomics : limitations and recommendations for future progress with particular focus on nutrition research. *Metabolomics*, 5 :435–458, 2009.

[234] J. Schellenberger and B. Ø. Palsson. Use of Randomized Sampling for Analysis of Metabolic Networks. *The Journal of Biological Chemistry*, 284(9) :5457–5461, 2009.

[235] 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. *Journal of theoretical biology*, 203(3) :229–248, 2000.

[236] F. Schreiber. Comparison of Centralities for Biological Networks. *Proc German Conf Bioinformatics*, 53 :199–206, 2004.

[237] S. Schuster, D. A. Fell, and T. Dandekar. A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nature biotechnology*, 18(March) :2000, 2000.

[238] D. C. Sévin, A. Kuehne, N. Zamboni, and U. Sauer. Biological insights through nontargeted metabolomics. *Current opinion in biotechnology*, 34 :1–8, 2014.

[239] D. Shawcross and R. Jalan. The pathophysiologic basis of hepatic encephalopathy : Central role for ammonia and inflammation. *Cellular and Molecular Life Sciences*, 62(19-20) :2295–2304, 2005.

[240] H. A. Simon. A behavioral model of rational choice. *The quarterly journal of economics*, 69(1) :99–118, 1955.

[241] H. A. Simon. Bounded rationality and organizational learning. *Organization science*, 2(1) :125–134, 1990.

[242] N. R. Smalheiser and G. Bonifield. Two Similarity Metrics for Medical Subject Headings (MeSH) : An Aid to Biomedical Text Mining and Author Name Disambiguation Neil. *Journal of Biomedical Discovery and Collaboration*, pages 1–14, 2016.

[243] I. P. Smyth. EventRank : A Framework for Ranking Time-Varying Networks. In *Proceedings of the 3rd international workshop on Link discovery*, pages 9–16, 2005.

[244] M. Sonderegger. Applications of graph theory to an English rhyming corpus. *Computer Speech and language*, 25 :655–678, 2011.

[245] R. Spicer, R. M. Salek, P. Moreno, D. Cañueto, and C. Steinbeck. Navigating freely-available software tools for metabolomics analysis. *Metabolomics*, 13(9) :106, 2017.

[246] J. L. Spratlin, N. J. Serkova, and S. G. Eckhardt. Clinical Applications of Metabolomics in Oncology : A Review. *Clinical Cancer Research*, 15(2) :431–441, 2009.

[247] C. Steinbeck, Y. Han, S. Kuhn, O. Horlacher, E. Luttmann, and E. Willighagen. The Chemistry Development Kit (CDK) : an open-source Java library for Chemo- and Bioinformatics. *Journal of chemical information and computer sciences*, 43(2) :493–500, jan 2003.

[248] K. STEPHENSON and M. ZELEN. RETHINKING CENTRALITY : METHODS AND EXAMPLES. *Social Networks*, 11 :1–37, 1989.

[249] M. D. Stobbe, S. M. Houten, G. a. Jansen, A. H. van Kampen, and P. D. Moerland. Critical assessment of human metabolic pathway databases : a stepping stone for future integration. *BMC Systems Biology*, 5(1) :165, 2011.

[250] M. D. Stobbe, S. M. Houten, A. H. van Kampen, R. J. Wanders, and P. D. Moerland. Improving the description of metabolic networks : the TCA cycle as example. *The FASEB Journal*, 26(9) :3625–3636, 2012.

[251] D. Stumpfe and J. Bajorath. Similarity searching. *Wiley Interdisciplinary Reviews : Computational Molecular Science*, 1(2) :260–282, mar 2011.

[252] K. Sugiyama. Graph drawing and applications for software and knowledge engineers. Technical report, Institute for Social Information Service, FUJITSU Lab. LTD., 1994.

[253] L. W. Sumner, A. Amberg, D. Barrett, M. H. Beale, R. Beger, C. A. Daykin, T. W.-M. Fan, O. Fiehn, R. Goodacre, J. L. Griffin, T. Hankemeier, N. Hardy, J. Harnly, R. Higashi, J. Kopka, A. N. Lane, J. C. Lindon, P. Marriott, A. W. Nicholls, M. D. Reily, J. J. Thaden, and M. R. Viant. Proposed minimum reporting standards for chemical analysis. *Metabolomics*, 3 :211–221, 2007.

[254] S. Szeider. Finding paths in graphs avoiding forbidden transitions. *Discrete Applied Mathematics*, 126(2-3) :261–273, 2003.

[255] I. Thiele and B. Ø. Palsson. A protocol for generating a high-quality genome-scale metabolic reconstruction. *Nature protocols*, 5(1) :93–121, 2011.

[256] I. Thiele, N. Swainston, R. M. T. Fleming, A. Hoppe, S. Sahoo, M. K. Aurich, H. Haraldsdottir, M. L. Mo, O. Rolfsson, M. D. Stobbe, S. G. Thorleifsson, R. Agren, C. Bölling, S. Bordel, A. K. Chavali, P. Dobson, W. B. Dunn,

L. Endler, D. Hala, M. Hucka, D. Hull, D. Jameson, N. Jamshidi, J. J. Jonsson, N. Juty, S. Keating, I. Nookaew, N. Le Novère, N. Malys, A. Mazein, J. A. Papin, N. D. Price, E. Selkov, M. I. Sigurdsson, E. Simeonidis, N. Sonnenschein, K. Smallbone, A. Sorokin, J. H. G. M. van Beek, D. Weichart, I. Goryanin, J. Nielsen, H. V. Westerhoff, D. B. Kell, P. Mendes, and B. Ø. Palsson. A community-driven global reconstruction of human metabolism. *Nature biotechnology*, 31(5) :419–25, may 2013.

[257] R. Todeschini, V. Consonni, H. Xiang, J. Holliday, M. Buscema, and P. Willett. Similarity coefficients for binary chemoinformatics data : Overview and extended comparison using simulated and real data sets. *Journal of Chemical Information and Modeling*, 52(11) :2884–2901, 2012.

[258] A. Tversky and D. Kahneman. Judgment under uncertainty : Heuristics and biases. *Utility, probability, and human decision making*, pages 141—-162, 1973.

[259] E. Ukkonen. On approximate string matching. *Foundations of Computation Theory*, 1983.

[260] T. W. Valente and R. K. Foreman. Integration and radiality : measuring the extent of an individual's connectedness and reachability in a network. *Social Networks*, 20 :89–105, 1998.

[261] J. van Helden, L. Wernisch, D. Gilbert, and S. J. Wodak. *Graph-based analysis of metabolic networks.* Number 38. Springer Berlin Heidelberg, 2002.

[262] S. Vast, P. Dupont, and Y. Deville. Automatic extraction of relevant nodes in biochemical networks. In *Atelier Apprentissage et BioInformatique, 7e Conférence francophone sur l'apprentissage automatique*, pages 21–31, 2005.

[263] A. Wagner and D. A. Fell. The small world inside large metabolic networks. *Proceedings of the Royal Society of London B : Biological Sciences*, 268(1478) :1803–1810, 2001.

[264] M. Wang, J. J. Carver, V. V. Phelan, L. M. Sanchez, C. A. Garg, Neha and Peng, Yao Nguyen, Don Duy Watrous, Jeramie Kapono, T. Luzzatto-Knaan, and ... Sharing and community curation of mass spectrometry data with GNPS. *Nature biotechnology*, 34(8) :828–837, 2017.

[265] C. Ware, H. Purchase, L. Colpoys, and M. Mcgill. Cognitive measurements of graph aesthetics. *Information Visualization*, 1 :103–110, 2002.

[266] D. Weininger. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *Journal of Chemical Information and Modeling*, 28(1) :31–36, feb 1988.

[267] S. White and P. Smyth. Algorithms for estimating relative importance in networks. *Proceedings of the ninth ACM SIGKDD international conference on Knowledge discovery and data mining - KDD '03*, page 266, 2003.

[268] P. Willett. Similarity Methods in Chemoinformatics. *Annual review of information science and technology*, 43 :3–71, 2009.

[269] P. Willett, J. Barnard, and G. Downs. Chemical Similarity Searching. *Journal of Chemical Information and Modeling*, 38(6) :983–996, nov 1998.

[270] D. S. Wishart, D. Tzur, C. Knox, R. Eisner, A. C. Guo, N. Young, D. Cheng, K. Jewell, D. Arndt, S. Sawhney, C. Fung, L. Nikolai, M. Lewis, M.-A. Coutouly, I. Forsythe, P. Tang, S. Shrivastava, K. Jeroncic, P. Stothard, G. Amegbey, D. Block, D. D. Hau, J. Wagner, J. Miniaci, M. Clements, M. Gebremedhin, N. Guo, Y. Zhang, G. E. Duggan, G. D. Macinnis, A. M. Weljie, R. Dowlatabadi, F. Bamforth, D. Clive, R. Greiner, L. Li, T. Marrie, B. D. Sykes, H. J. Vogel, and L. Querengesser. HMDB : the Human Metabolome Database. *Nucleic acids research*, 35(Database issue) :D521–6, jan 2007.

[271] M. Witting, M. Lucio, D. Tziotis, B. Wägele, K. Suhre, R. Voulhoux, S. Garvis, and P. Schmitt-Kopplin. DI-ICR-FT-MS-based high-throughput deep

metabotyping : a case study of the Caenorhabditis elegans – Pseudomonas aeruginosa infection model. *Analytical and bioanalytical chemistry*, 407(4) :1059–1073, 2014.

[272] G. Wohlgemuth, P. K. Haldiya, E. Willighagen, T. Kind, and O. Fiehn. The Chemical Translation Service–a web-based tool to improve standardization of metabolomic reports. *Bioinformatics (Oxford, England)*, 26(20) :2647–8, oct 2010.

[273] S. Wuchty and P. F. Stadler. Centers of complex networks. *Journal of Theorical Biology*, 223 :45–53, 2003.

[274] K. Xu, C. Rooney, P. Passmore, D. H. Ham, and P. H. Nguyen. A user study on curved edges in graph visualization. *IEEE Transactions on Visualization and Computer Graphics*, 18(12) :2449–2456, 2012.

[275] A. Yamaguchi, Y. Yamamoto, J.-d. Kim, T. Takagi, and A. Yonezawa. Discriminative application of string similarity methods to chemical and non-chemical names for biomedical abbreviation clustering. *BMC Genomics*, 13(Suppl 3) :S8, 2012.

[276] J. Y. Yen. Finding the k shortest loopless paths in a network. *management Science*, 17(11) :712–716, 1971.

[277] W. W. Zachary. An Information Flow Model for Conflict and Fission in Small Groups. *Journal of Anthropological Research*, 33(4) :452–473, 1977.

[278] J. Zanghellini, D. E. Ruckerbauer, M. Hanscho, and C. Jungreuthmayer. Elementary flux modes in a nutshell : Properties, calculation and applications. *Biotechnology journal*, 8(9) :1009–1016, 2013.

[279] A.-p. Zeng, H.-W. Ma, and A.-p. Zeng. The connectivity structure, giant strong component and centrality of metabolic networks. *Bioinformatics*, 19(11) :1423–1430, jul 2003.

[280] J. Zhao, T.-h. Yang, Y. Huang, and P. Holme. Ranking Candidate Disease Genes from Gene Expression and Protein Interaction : A Katz-Centrality Based Approach. *PloS one*, 6(9), 2011.

[281] A. O. Zhirov, O. V. Zhirov, and D. L. Shepelyansky. Two-dimensional ranking of Wikipedia articles. *European Physical Journal B*, 77(4) :523–531, 2010.

[282] E. Zotenko, J. Mestre, and T. M. Przytycka. Why Do Hubs in the Yeast Protein Interaction Network Tend To Be Essential : Reexamining the Connection between the Network Topology and Essentiality. *PLoS Computational Biology*, 4(8), 2008.

Résumé

La métabolomique permet une étude à large échelle du profil métabolique d'un individu, représentatif de son état physiologique. La comparaison de ces profils conduit à l'identification de métabolites caractéristiques d'une condition donnée. La métabolomique présente un potentiel considérable pour le diagnostic, mais également pour la compréhension des mécanismes associés aux maladies et l'identification de cibles thérapeutiques. Cependant, ces dernières applications nécessitent d'inclure ces métabolites caractéristiques dans un contexte plus large, décrivant l'ensemble des connaissances relatives au métabolisme, afin de formuler des hypothèses sur les mécanismes impliqués. Cette mise en contexte peut être réalisée à l'aide des réseaux métaboliques, qui modélisent l'ensemble des transformations biochimiques opérables par un organisme. L'une des limites de cette approche est que la métabolomique ne permet pas à ce jour de mesurer l'ensemble des métabolites, et ainsi d'offrir une vue complète du métabolome. De plus, dans le contexte plus spécifique de la santé humaine, la métabolomique est usuellement appliquée à des échantillons provenant de biofluides plutôt que des tissus, ce qui n'offre pas une observation directe des mécanismes physiologiques eux-mêmes, mais plutôt de leur résultante. Les travaux présentés dans cette thèse proposent une méthode pour pallier ces limitations, en suggérant des métabolites pertinents pouvant aider à la reconstruction de scénarios mécanistiques. Cette méthode est inspirée des systèmes de recommandations utilisés dans le cadre d'activités en ligne, notamment la suggestion d'individus d'intérêt sur les réseaux sociaux numériques. La méthode a été appliquée à la signature métabolique de patients atteints d'encéphalopathie hépatique. Elle a permis de mettre en avant des métabolites pertinents dont le lien avec la maladie est appuyé par la littérature scientifique, et a conduit à une meilleure compréhension des mécanismes sous-jacents et à la proposition de scénarios alternatifs. Elle a également orienté l'analyse approfondie des données brutes de métabolomique et enrichie par ce biais la signature de la maladie initialement obtenue. La caractérisation des modèles et des données ainsi que les développements techniques nécessaires à la création de la méthode ont également conduit à la définition d'un cadre méthodologique générique pour l'analyse topologique des réseaux métaboliques.

Abstract

Metabolomics allows large-scale studies of the metabolic profile of an individual, which is representative of its physiological state. Metabolic markers characterising a given condition can be obtained through the comparison of those profiles. Therefore, metabolomics reveals a great potential for the diagnosis as well as the comprehension of mechanisms behind metabolic dysregulations, and to a certain extent the identification of therapeutic targets. However, in order to raise new hypotheses, those applications need to put metabolomics results in the light of global metabolism knowledge. This contextualisation of the results can rely on metabolic networks, which gather all biochemical transformations that can be performed by an organism. The major bottleneck preventing this interpretation stems from the fact that, currently, no single metabolomic approach allows monitoring all metabolites, thus leading to a partial representation of the metabolome. Furthermore, in the context of human health related experiments, metabolomics is usually performed on bio-fluid samples. Consequently, those approaches focus on the footprints left by impacted mechanisms rather than the mechanisms themselves. This thesis proposes a new approach to overcome those limitations, through the suggestion of relevant metabolites, which could fill the gaps in a metabolomics signature. This method is inspired by recommender systems used for several on-line activities, and more specifically the recommendation of users to follow on social networks. This approach has been used for the interpretation of the metabolic signature of the hepatic encephalopathy. It allows highlighting some relevant metabolites, closely related to the disease according to the literature, and led to a better comprehension of the impaired mechanisms and as a result the proposition of new hypothetical scenario. It also improved and enriched the original signature by guiding deeper investigation of the raw data, leading to the addition of missed compounds. Models and data characterisation, alongside technical developments presented in this thesis, can also offer generic frameworks and guidelines for metabolic networks topological analysis.