Prédiction de réseaux d'interactions biomoléculaires à partir de données de la génomique comparée

Florian Iragne

N° d'ordre : Numéro d'ordre

THÈSE

PRÉSENTÉE À

L'UNIVERSITÉ BORDEAUX I

ÉCOLE DOCTORALE DE MATHÉMATIQUES ET D'INFORMATIQUE

Par **Florian Iragne**

POUR OBTENIR LE GRADE DE

DOCTEUR

SPÉCIALITÉ : Informatique

Prédiction de réseaux d'interactions biomoléculaires à partir des données de la génomique comparée

---

Soutenue le :

Après avis des rapporteurs :

Anastasia Nikolskaya	Professeur associée
Serge Potier	Professeur

Devant la commission d'examen composée de :

Maylis Delest	Professeur	Rôle
Anastasia Nikolskaya	Professeur associée	Examinateur
Serge Potier	Professeur	Examinateur
David James Sherman	Maître de conférences HDR	Examinateur
Robert Strandh ...	Professeur	Rôle
Prénom Nom	Grade	Rôle

– 2007 –

Les systèmes biologiques présentent de nombreux phénomènes complexes, aux interactions encore plus complexes. La modélisation a pour but de faciliter l'étude et la compréhension des systèmes biologiques, par l'observation ou la simulation des modèles créés. Les réseaux d'interactions biomoléculaires sous-tendent ces modèles. Les travaux de thèse que nous présentons portent sur la prédiction systématique de réseaux d'interactions biomoléculaires, afin de fournir les éléments d'entrée nécessaires au processus de modélisation. Les deux thèmes centraux seront la prédiction de réseaux d'interactions protéine-protéine et l'extrapolation de voies métaboliques.

Nous définissons tout d'abord un cadre formel d'extraction de graphes d'interactions dictée par des politiques, qui permet de créer des résumés intelligents à partir de jeux de données hétérogènes. Une séparation claire des tâches d'extraction et de visualisation de l'information nous permet d'exprimer différents algorithmes existants, estimant par exemple la qualité des réseaux d'interactions biomoléculaires prédits. Nous avons mis en oeuvre ce cadre formel dans le logiciel ProViz [Iragne et al., 2005]. Nous présentons par la suite, des méthodes informatiques d'extrapolation de voies métaboliques, inspirées du précédent formalisme et basées sur l'utilisation de voies de référence et sur une identification robuste d'équivalents fonctionnels. Ces méthodes nous permettent de prédire un ensemble de voies métaboliques centrales, formant la base de modèles pour des organismes dont seules les données génomiques sont disponibles. Les différents résultats, disponibles en ligne[1] [Sherman et al., 2006] ou en cours de publication, nous permettent de valider notre approche.

Biological systems are complex systems, in that they cannot be fully understood through sole study of individual components. Network representations of the relations between individual and groups of functional elements are essential for studying these systems. These networks underly mathematical models of cell processes and their construction is a prerequisite for the modelling step. In this thesis, we present our work on systematic prediction of biomolecular interactions using comparative genomics data. The main two axis of this thesis are the prediction of protein-protein interaction networks, and the extrapolation of metabolic pathways.

We will first present a formal framework, called "graph extraction policies", that enables the construction of biomolecular interaction networks from heterogeneous datasets, using an strategy based on neighbourhood exploration. This framework enables us to include third-part algorithms, to compute network prediction quality or any other information. We have implemented this framework in ProViz [Iragne et al., 2005]. We then introduce two methods to extrapolate metabolic pathways, based on reference pathways and on the robust identification of functional homologs. These methods enable us to provide core metabolic pathways as a base of models for organisms that are only sequenced, and are available online[1] [Sherman et al., 2006] or under publication.

[1] http ://cbi.labri.fr/Genolevures/path

Table des matières

Liste des tableaux

Table des figures

Introduction

Les systèmes biologiques sont des systèmes complexes, dans la mesure où leur pleine compréhension nécessite d'étudier à la fois leurs composants et les interactions qu'ils partagent. Les réseaux d'interactions entre biomolécules, tels les réseaux d'interactions protéine-protéine et les voies métaboliques, sont des vues dynamiques représentant des relations entre composants cellulaires qui modélisent différents aspects du fonctionnement de la cellule. L'étude de ces réseaux est appelée *Systems Biology* ([Onami et al., 2001], [Westerhoff and Palsson, 2004]). L'émergence de cette discipline, qui a aussi pour but de modéliser les objets biologiques [Kitano, 2002], a entraîné le développement de modèles mathématiques pour certaines espèces d'intérêt. Un modèle est une abstraction, une représentation simplifiée ou au contraire très fine d'un phénomène ou d'un ensemble de phénomènes. Les réseaux de relations entre biomolécules sous-tendent ces modèles, qu'ils soient sous la forme de systèmes d'équations différentielles ou de modèles hiérarchiques de systèmes et sous-systèmes discrets.

Les relations d'interaction entre biomolécules sont des représentations dynamiques des systèmes biologiques. Ces relations font appel à la notion de voisinage, défini comme un ensemble d'objets partageant une même relation [Danchin, 1998]. Deux biomolécules sont donc considérées comme voisines s'il existe entre elles une relation, que celle-ci soit basée sur une interaction explicite ou sur le partage d'une même propriété. Les réseaux d'interactions protéine-protéine sont donc des voisinages, de même que le sont les relations de cooccurrence de protéines et métabolites dans les voies métaboliques ou dans les profils d'expressions.

Les réseaux d'interactions biomoléculaires sont obtenus par expérimentation à moyenne ou grande échelle ([Giot et al., 2003], [Ito et al., 2001]), ou par prédiction *in silico* ([Jansen et al., 2003], [Iossifov et al., 2004]). Les méthodes expérimentales permettent généralement d'obtenir des réseaux de qualité, mais sont coûteuses et donc restreintes à certains organismes d'intérêt. Les méthodes informatiques fournissent des réseaux supposés de moindre confiance, mais sont en théorie applicables à tout organisme dont les données génomiques sont disponibles.

Les travaux que nous présentons dans cette thèse portent sur la prédiction systématique de réseaux d'interactions biomoléculaires, en utilisant notamment des données de génomique comparée, afin de fournir les éléments d'entrée nécessaires au

processus de modélisation. Ces travaux s'articulent autour de deux thèmes centraux que sont la prédiction de réseaux d'interactions protéine-protéine et l'extrapolation de voies métaboliques.

Les réseaux d'interactions protéine-protéine La caractérisation expérimentale des interactions physiques entre protéines requiert une grande variété de techniques expérimentales (Yeast Two-Hybrid [Fields and Song, 1989], TAP-TAG [Rigaut et al., 1999], blue native PAGE [Camacho-Carvajal et al., 2004], etc). Ces techniques identifient des interactions plus ou moins stables, à plus ou moins grande échelle. Cependant, aucune n'est capable d'identifier simultanément tous les types d'interactions [Hoffmann and Valencia, 2003]. De plus, ces techniques sont très coûteuses en temps, en argent, et sont complexes à mettre en place. Pour pallier ces inconvénients, de nombreuses techniques de prédiction *in silico* ont été développées. La plupart se basent sur des techniques d'apprentissage ([Jansen et al., 2003], [Zhang et al., 2004]) et sur la réalisation de modèles statistiques, pour prédire les relations d'interaction entre les protéines (voir section 2.1). Pour obtenir des réseaux d'interactions biomoléculaires complets, il est donc nécessaire de pouvoir combiner les différentes expériences et méthodes expérimentales ou prédictives.

Nous avons développé dans cette optique une méthode appelée "Extraction de Graphes dictée par des politiques" (section 3.2). Cette méthode permet de construire des résumés intelligents et intelligibles, consistant en des réseaux d'interactions biomoléculaires obtenus par extensions successives de voisinages. À partir d'un ensemble de biomolécules de départ, d'une distance d'extension de voisinage, d'un jeu de données à fouiller et d'une requête de l'utilisateur, l'algorithme construit un graphe où les sommets sont des biomolécules et où les arcs représentent les interactions biomoléculaires. Chaque biomolécule présente dans le graphe final est caractérisée par un ensemble de relations de voisinage avec une ou plusieurs biomolécules de départ. L'algorithme modifie le graphe en construction en fonction des nouvelles relations de voisinage découvertes à chaque nouvelle itération sur le jeu de données. Des prédicats peuvent être configurés, afin de filtrer le graphe selon divers attributs des biomolécules ou de leurs relations, et ainsi prendre en compte, par exemple, les différences de qualité des différents jeux de données (section 2.1.4).

Dans le logiciel ProViz [Iragne et al., 2005], nous avons mis en oeuvre une version de notre formalisme, adaptée aux réseaux d'interactions protéine-protéine. ProViz a été développé dans le cadre du projet européen IntAct [Hermjakob et al., 2004a]), qui développe une base de données fédérative d'interactions protéiques et de leurs annotations (section 2.1.2). Ce projet est piloté par l'Institut Européen de Bioinformatique (EBI[2]) et implique huit partenaires institutionnels et industriels en

[2]http ://www.ebi.ac.uk/

Europe. Le réseau IntAct a développé une représentation standardisée des interactions protéiques, incluant la définition d'un jeu de données minimal permettant d'identifier une interaction de manière unique. Cette représentation formelle a été transcrite dans un modèle de base de donnée objet, disponible en ligne[3] et navigable à l'aide d'outils de recherche, d'analyse et d'annotation.

Les voies métaboliques Les voies métaboliques sont des réseaux d'interactions biomoléculaires. Ici, les relations impliquent deux types de biomolécules : les enzymes et les métabolites. Les enzymes sont des protéines, voire des complexes protéiques, et les métabolites sont le plus souvent de petites molécules telles que le glucose. Les voies métaboliques sont un point central dans l'étude des organismes biologiques puisqu'elles conditionnent en grande partie leur phénotype. Les relations enzyme-substrat sont des réactions biochimiques bien définies, qui peuvent être caractérisées par un éventail de méthodes expérimentales. Différentes bases de données, telles celle du KEGG [Ogata et al., 1999], ou encore MetaCyc [Caspi et al., 2006] et Brenda [Schomburg et al., 2002] (section 2.2.2), fournissent un catalogue de réactions enzymatiques, organisées en voies métaboliques définies par expertise biologique. Selon la source, les voies métaboliques sont définies indépendamment d'une espèce ou sont espèce-spécifiques. Dans tous les cas, l'étude expérimentale des voies métaboliques d'une espèce nécessite la mise en place de moyens expérimentaux lourds et coûteux. Par ce fait, elle reste limitée à un petit nombre d'espèces d'intérêt. La disponibilité de données de référence pour des organismes modèles, ainsi que de données de génomique comparée pour un nombre croissant d'espèces, nous permet de rechercher des méthodes de prédiction *in silico* de voies métaboliques pour des espèces non-étudiées expérimentalement. Dans le cadre de nos travaux de thèse, nous avons développé des méthodes de modélisation soustractive, permettant d'extrapoler les voies métaboliques d'un organisme de référence à un autre organisme, en nous basant sur une identification fiable et robuste d'équivalents fonctionnels.

Dans le cadre du projet Génolevures [Dujon et al., 2004], nous avons appliqué nos méthodes à l'extrapolation des voies métaboliques de l'organisme modèle *Saccharomyces cerevisiae* à quatre espèces appartenant au même phylum. Dans cette étude, nous avons pu montrer que de 60% à 80 % des voies métaboliques de *S. cerevisiae* sont conservées dans ce phylum, ce qui pourrait nous permettre de définir un ensemble minimal de fonctions communes aux eucaryotes, étant donné que le phylum des hémiascomycètes contient des espèces relativement distantes [Dujon, 2006]. Parmi ces voies du métabolisme central, nous trouvons le cycle TCA, la voie des pentoses phosphates, la glycolyse, les métabolismes du pyruvate, des purines, des pyrimidines et des acides aminés (tables 10 et 11, annexe A). Notre approche nous permet aussi de détecter la perte d'une voie métabolique due à l'adaptation d'un organisme à une niche écologique, comme la perte du métabolisme du galactose chez *C. glabrata*. L'étude des graphes représentant les voies métaboliques met en évidence

[3]http ://www.ebi.ac.uk/intact

des caractéristiques particulières. Elles semblent notamment présenter, pour la plupart, un degré sortant moyen supérieur aux autres voies métaboliques, ce qui suggère que ces voies centrales comportent des chemins alternatifs alors que les autres voies sont plus linéaires. Cette observation est compatible avec l'hypothèse selon laquelle les voies métaboliques universellement conservées font partie du métabolisme central, qui gère le fonctionnement basal de la cellule et qui devraient donc être plus robuste face à la perte d'une ou plusieurs enzymes, que des voies du métabolisme périphérique, plus linéaires et donc plus facilement déconnectables en cas de perte d'enzyme (section 2.2.1). Enfin, notre étude a montré que plus de 60% des résultats ne nécessitent aucune investigation manuelle quant à la prise de décision de conservation, ce qui nous permet de dire que notre méthode est utilisable dans le cadre de l'extrapolation systématique de voies métaboliques, à partir de voies de références et des données de génomique comparée.

La génomique comparée La génomique comparée étudie l'évolution moléculaire des génomes les uns par rapport aux autres, afin de comprendre les mécanismes de l'évolution. Elle s'intéresse notamment à la conservation, à la perte ou au gain de gènes, aux réarrangements chromosomiques ou bien encore à la vitesse d'évolution des séquences entre espèces ou sous-ensembles d'espèces.

Génolevures [Souciet J., 2000] est un projet de génomique comparée à grande échelle entre *S. cerevisiae* et 14 autres espèces représentatives des différentes branches du phylum des hémiascomycètes. La comparaison de ces génomes nous donne la possibilité d'explorer l'évolution des organismes eucaryotes. Le phylum des hémiascomycètes contient l'organisme modèle *S. cerevisiae* ainsi que de nombreux autres organismes d'intérêt scientifique ou industriel. L'étude de ce phylum révèle les évènements évolutifs ayant eu lieu, ainsi que leurs mécanismes et permet des comparaisons avec d'autres phyli, tel que celui des chordae qui contient la classe des mammifères. Nous portons un intérêt particulier aux phases deux et suivantes de Génolevures. La phase deux a consisté au séquençage et à l'annotation de quatre espèces de levures relativement distantes les unes des autres. *Candida glabrata* est une levure pathogène, seconde responsable des candidoses humaines. *Kluyveromyces lactis* est une levure souvent étudiée en génétique et présente un intérêt industriel. *Debaryomyces hansenii* est une levure halotolérante très proche de levures pathogènes telle *Candida albicans*, principal pathogène fongique chez l'Homme. Enfin, *Yarrowia lipolytica* est une levure consommant entre autres des hydrocarbures et se trouvant sur une branche éloignée du phylum des hémiascomycètes. Ces quatre espèces ont été séquencées et annotées par le consortium. Parmi les différentes études réalisées, les familles de protéines nous intéressent particulièrement. Les familles de protéines sont des partitions issues d'un clustering consensus [Nikolski and Sherman, 2007] de l'ensemble des protéines des 4 espèces plus *S. cerevisiae*. Chaque partition contient des protéines réputées avoir la même fonction dans tous les organismes.

Les familles de protéines nous intéressent à plusieurs titres. Tout d'abord, elles permettent une identification fiable et robuste d'équivalents fonctionnels. Nous les avons utilisées dans notre méthode d'extrapolation de voies métaboliques (chapitre 4). De plus, les familles de protéines permettent de calculer des matrices position-spécifique (PSSM [Altschul et al., 1997]) qui représentent le profil de chaque famille. Dans le cadre des travaux de thèse, nous avons étudié une méthode basée sur l'utilisation des PSSM et permettant une classification rapide des gènes d'une espèce séquencée, en différentes catégories : gènes spécifiques à l'espèce, spécifiques à son clade, communs au phylum. Cette étude a permis de montrer qu'un tel classement est faisable et rapide, ce qui permet de réaliser rapidement d'autre études, le temps que le long processus de calcul de familles de protéines soit réalisé (section 4.3).

Chapitre 1

Notions de biologie élémentaire

Sommaire

1.1 Les modules de base du vivant

1.1.1 La cellule

La cellule est l'unité élémentaire du vivant et son étude est nécessaire à la compréhension des mécanismes qui régissent la vie. Les mécanismes cellulaires sont complexes du fait de la diversité des phénomènes en jeu (réactions chimiques, transports, etc) et des éléments impliqués dans ces phénomènes. Dans le vivant, nous pouvons distinguer deux types différents de cellules. La cellule procaryote (figure 1b) est délimitée par la membrane plasmique et le plus souvent une paroi et contient dans cet unique compartiment le code source (l'ADN) et les librairies et binaires (protéines et complexes protéiques). La cellule eucaryote (figure 1a) est délimitée par la membrane plasmique, qui contient le cytoplasme et des organelles, elles-mêmes délimitée par une membrane. Dans ce type de cellule, l'ADN est isolé dans un compartiment appelé noyau, les autres éléments étant dans le cytoplasme et/ou dans d'autres compartiments appelés organelles.

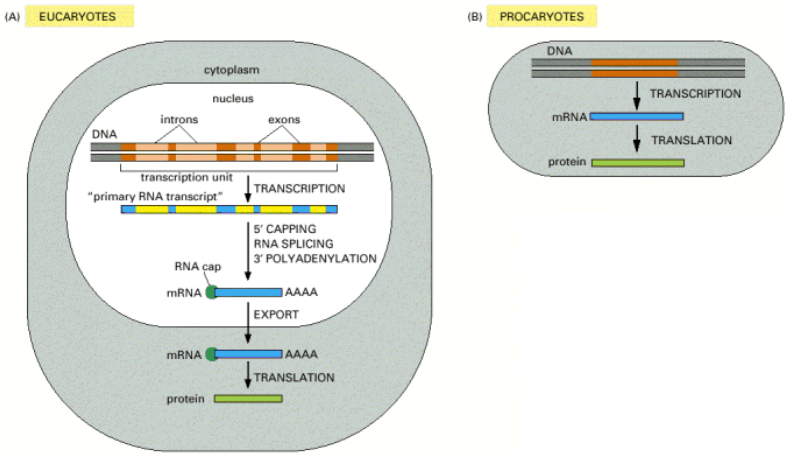


FIG. 1 – Cellules procaryotes et eucaryotes : du gène à la protéine. Les cellules procaryotes ne possèdent pas de noyau, contrairement au cellules eucaryotes, dont le noyau conserve l'ADN. Ce schéma indique aussi le processus permettant de transformer un gène en protéine, par l'intermédiaire de la transcription, puis de la traduction. (d'après [Alberts et al., 2002])

1.1.2 L'ADN, l'ARN, les protéines

L'ADN, pour Acide DésoxyriboNucléique, est le support de l'information constituant le programme de la cellule. L'ADN peut être vu comme une échelle dont les montants sont constitués d'une suite de lettres sur un alphabet de quatre lettres : A, T, G, C (figure 2). Chaque barreau de l'échelle est constitué de l'appariement de deux lettres complémentaires (A s'associe avec T, G avec C).

L'ADN est le constituant principal des chromosomes, qui peuvent être de différentes formes (circulaires ou non) ou tailles. En grossissant le trait, nous pouvons subdiviser l'ADN en deux sous ensembles : les gènes et les zones intergéniques (figure 3). Les gènes sont les portions de l'ADN qui contiennent le programme de la cellule. Ils sont plus ou moins bien caractérisés en fonction de l'espèce étudiée, mais leur rôle de support de l'information est une constante dans le monde du vivant. Les zones intergéniques sont moins facilement caractérisables et sont donc moins connues, bien que l'on sache leur implication dans les phénomènes de régulation de transcription des gènes.

La transcription est le processus biologique qui permet de convertir l'ADN en ARN. L'ARN, pour Acide RiboNucléique, est constitué d'une suite de lettre sur un alphabet de 4 lettres (A, U, G, C), complémentaires deux-à-deux (A avec U, G avec C). Contrairement à l'ADN, l'ARN n'est constitué que d'un seul brin et il existe au moins autant d'ARN différents dans une cellule qu'il existe de gènes différents. L'ARN messager est un intermédiaire entre le code source (le gène) et le binaire (la protéine).

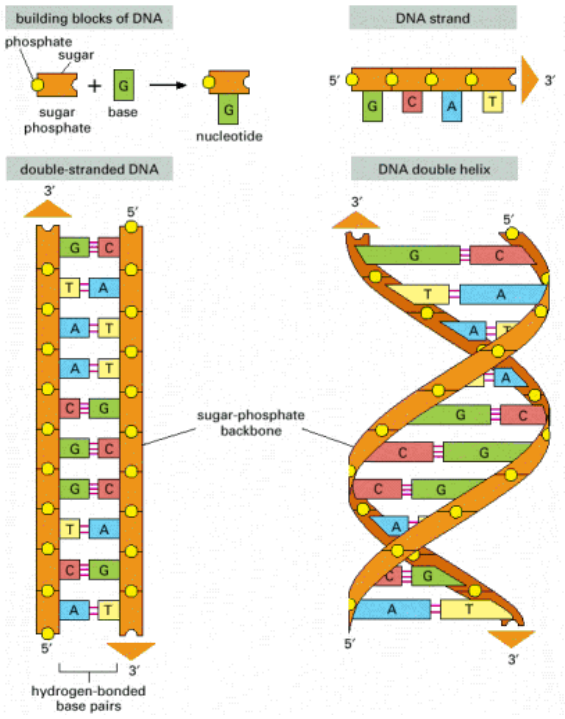


Fig. 2 – L'ADN : structure planaire et double hélice. Ce schéma représente la décomposition de la structure de l'ADN. Chaque brin est formé d'une succession de modules de base (les nucléotides), et deux brins forment une molécules d'ADN qui se structure sous la forme d'une double-hélice. (d'après [Alberts et al., 2002])

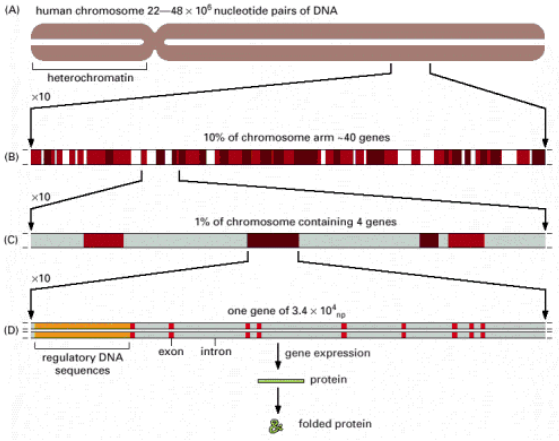


Fig. 3 – L'ADN : du chromosome à la protéine. Ce schéma montre une molécule d'ADN à différentes échelles : tout d'abord au niveau macroscopique (le chromosome), puis en zoomant pour obtenir un brin d'ADN contenant quatre gènes (c), et enfin, un seul gène (d) avec sa structure. (d'après [Alberts et al., 2002])

	A	U	G	C
U	UUU Phe (F)	UCU Ser (S)	UAU Tyr (Y)	UGU Cys (C)
	UUC Phe (F)	UCC Ser (S)	UAC Tyr (Y)	UGC Cys (C)
	UUA Leu (L)	UCA Ser (S)	UAA STOP	UGA STOP
	UUG Leu (L)	UCG Ser (S)	UAG STOP	UGG Trp (W)
C	CUU Leu (L)	CCU Pro (P)	CAU His (H)	CGU Arg (R)
	CUC Leu (L)	CCC Pro (P)	CAC His (H)	CGC Arg (R)
	CUA Leu (L)	CCA Pro (P)	CAA Gln (Q)	CGA Arg (R)
	CUG Leu (L)	CCG Pro (P)	CAG Gln (Q)	CGG Arg (R)
A	AUU Ile (I)	ACU Thr (T)	AAU Asn (N)	AGU Ser (S)
	AUC Ile (I)	ACC Thr (T)	AAC Asn (N)	AGC Ser (S)
	AUA Ile (I)	ACA Thr (T)	AAA Lys (K)	AGA Arg (R)
	AUG Met (M)	ACG Thr (T)	AAG Lys (K)	AGG Arg (R)
G	GUU Val (V)	GCU Ala (A)	GAU Asp (D)	GGU Gly (G)
	GUC Val (V)	GCC Ala (A)	GAC Asp (D)	GGC Gly (G)
	GUA Val (V)	GCA Ala (A)	GAA Glu (E)	GGA Gly (G)
	GUG Val (V)	GCG Ala (A)	GAG Glu (E)	GGG Gly (G)

TAB. 1 – Code Génétique standard : le tableau se lit de gauche à droite, en associant une lettre de la première colonne à une lettre de la première ligne. La troisième position est constituée de l'une des 4 lettres A, U, G, C. Le tout forme une combinaison de 4 lettres sur trois positions, définissant de manière non-ambiguë l'un des vingt acides aminés ou un codon stop

Une protéine est une suite de lettres sur un alphabet à 20 lettres, appelées acides aminés (21 lettres en comptant le STOP). Pour obtenir une protéine, la cellule doit passer par un chemin minimal composé de deux étapes, la transcription suivie de la traduction (figure 1a). Comme vu précédemment, la transcription permet de passer d'une séquence ADN à sa séquence complémentaire en ARN. La traduction permet de traduire le code ARN à quatre lettres en code protéique sur 21 lettres. Pour ce faire, la cellule utilise une structure appelée ribosome, très conservée dans le vivant, et qui permet d'aligner un triplet de lettres ARN sur une lettre protéique, suivant les règles de correspondance établies par le code génétique (table 1).

Sous-jacente à cette traduction, apparaît la notion de dégénérescence du code génétique. En effet, si l'on prend l'ensemble des permutations de 3 lettres ADN ou ARN, nous obtenons 64 combinaisons, ce qui est bien plus que les 21 lettres du code protéique. Chaque lettre du code protéique peut en effet correspondre à plusieurs triplets de lettres du code ADN, respectivement du code ARN.

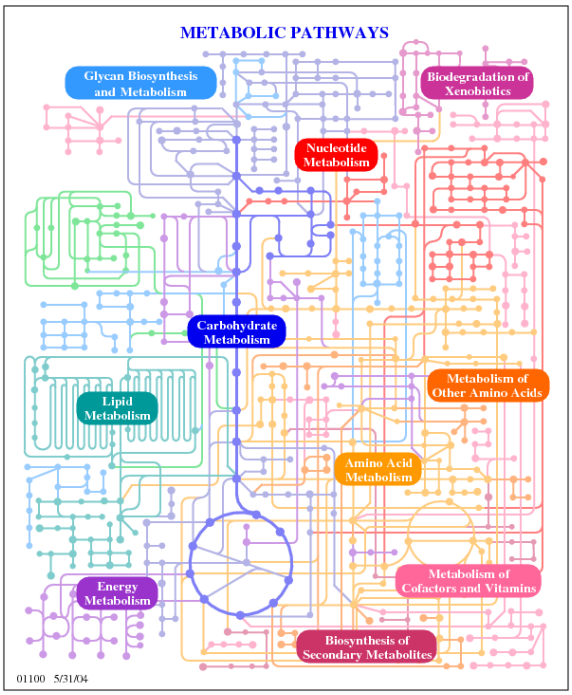


Fig. 4 – Carte des voies métaboliques génériques (source : KEGG). Ce schéma montre l'ensemble des voies métaboliques identifiées et stockées par le KEGG [Ogata et al., 1999].

1.2 La machinerie cellulaire

Nous avons donc vu comment s'articulent les unités de bases qui portent le programme de la cellule, en passant du support (ADN) à l'effecteur (protéine). Les protéines sont capables d'effectuer par elles-mêmes différentes actions au sein de la cellule, mais elles peuvent aussi s'associer en complexes pour effectuer des fonctions différentes, qui ne se résument pas à la somme des fonctions de leurs composantes. La compréhension des mécanismes de formation des complexes protéiques ainsi que la connaissance de la composition de ces complexes et de leurs interactions, comme dans les voies métaboliques (figure 4), sont des éléments essentiels à la compréhension des mécanismes cellulaires.

1.3 La génomique comparée

La génomique comparée est une partie de la biologie qui s'intéresse à l'étude de la comparaison des génomes. Nous l'avons vu précédemment, un génome est constitué de la suite de lettres formant l'ADN. La génomique comparée consiste donc à étudier les parties conservées et divergentes entre plusieurs génomes, cherchant

ainsi à révéler les mécanismes de l'évolution. La génomique comparée se base sur l'hypothèse évolutive que deux espèces données dérivent d'un ancêtre commun. La première étape dans une étude de génomique comparée est donc d'obtenir la séquence complète des génomes à étudier. L'étape suivante est de caractériser ces génomes, en recherchant les différents modules les composant. Ainsi, une première dichotomie peut être effectuée, en déterminant les parties codantes (les gènes) et les parties non-codantes du génome. Il est alors possible d'étudier la conservation des gènes entre génomes, que ce soit en termes qualitatifs (les gènes sont-ils semblables ?), quantitatifs (existe-t-il une correspondance 1-1 entre les gènes de deux espèces ?) ou encore architecturaux (les gènes sont-ils organisés de la même manière sur les chromosomes ?).

1.3.1 Conservation quantitative et qualitative de gènes

Une fois les gènes identifiés pour les génomes d'intérêt, il est possible de comparer leur séquence avec d'autres espèces, afin de déterminer leur pourcentage d'homologie et leurs éventuelles relations d'orthologie ou de paralogie. L'homologie définit à quel point les séquences des gènes étudiés se ressemblent. Pour deux espèces proches, il est probable que la majorité des gènes d'une espèce présente des homologues très similaires dans l'autre espèce. Les relations d'orthologie et de paralogie traduisent une relation plus fine entre séquences.

La figure 5 nous permet d'expliquer ces phénomènes. Dans la sous-figure (A), nous voyons apparaître la notion d'événement de spéciation. Un événement de spéciation se produit lorsqu'une branche de l'arbre du vivant se sépare en deux branches, autrement dit, lorsqu'à partir d'un organisme ancestral, nous "obtenons" deux nouvelles espèces. Dans ce cas, les nouvelles espèces ont des gènes orthologues, tels que le sont les gènes G_A et G_B dans l'exemple : deux gènes identiques dans deux organismes différents. Deux gènes orthologues sont supposés partager une même fonction dans les deux espèces. La sous-figure (B) nous permet de mettre en évidence la notion de paralogie. Deux gènes sont dits paralogues s'ils sont issus d'un événement de duplication à l'intérieur d'une même espèce. En l'occurrence, les gènes G_1 et G_2 sont des paralogues. La sous-figure (C) fait la synthèse de ces deux événements. Nous observons tout d'abord une duplication du gène G dans l'organisme ancestral, duplication suivie d'un événement de spéciation. Ainsi, nous dirons que les gènes $G_{1A}, G_{1B}, G_{2A}, G_{2B}$ sont homologues, que G_1A et G_2B sont des paralogues, et que G_{1A} et G_{1B} sont des orthologues.

L'identification d'homologues est un problème bien étudié, auquel des réponses ont été apportées, notamment par le développement de méthodes et logiciels appropriés (e.g., BLAST [Altschul et al., 1990], FASTA [Pearson and Lipman, 1988]). La distinction entre paralogues et orthologues est cependant plus délicate, mais revêt un aspect fondamental pour les études de génomiques comparée, étant donné que les

orthologues sont supposés être des équivalents fonctionnels, alors que les paralogues peuvent avoir des fonctions différentes des gènes originaux.

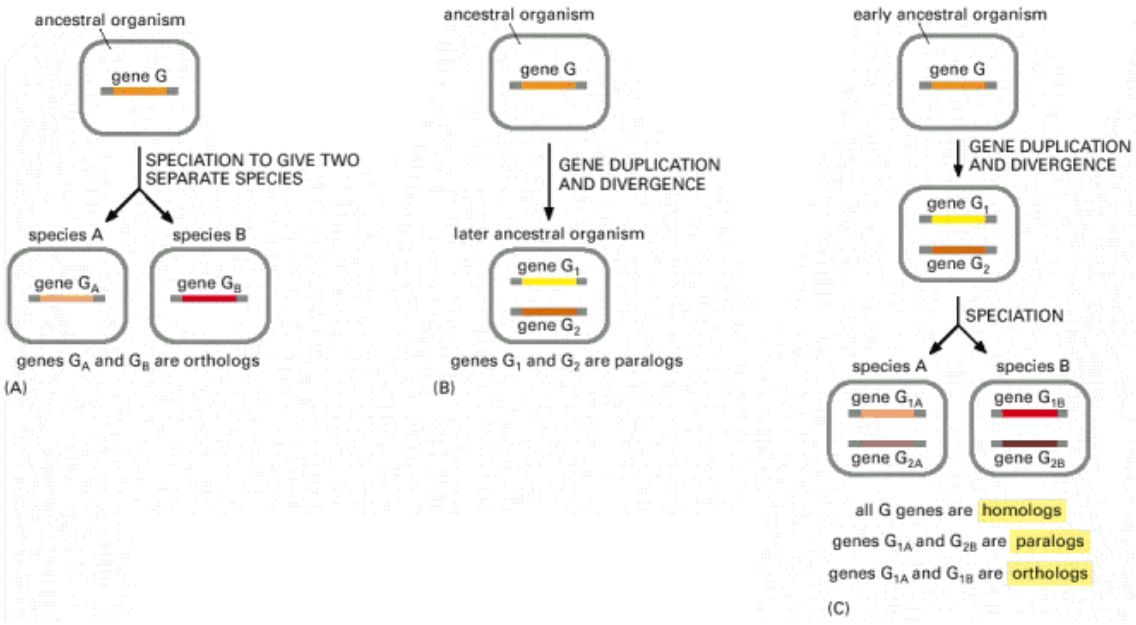


Fig. 5 – Schéma montrant la différence entre gènes orthologues (a) issus d'un événement de spéciation, et gènes paralogues (b) issus d'un événement de duplication post-spéciation. Dans la sous-figure (c), des événements plus complexes se produisent : tous les gènes G sont dits homologues, les gènes G1a et G2b sont paralogues alors que G1a et G1b sont orthologues. (d'après [Alberts et al., 2002])

1.3.2 Phylogénie des espèces

Un autre résultat classique de la génomique comparée est la réalisation d'arbres phylogénétiques, représentant les relations de parenté entre espèces. Dans un tel arbre (figure 6), chaque espèce connue est présente sur une feuille et la longueur des branches séparant deux espèces indique leur distance phylogénétique, que l'ont pourrait traduire comme étant la distance qui les sépare de leur plus proche ancêtre commun. Pour établir de tels arbres, des mesures de distances sont effectuées entre séquences orthologues, très conservées au sein de la plupart des espèces. En grossissant le trait, la stratégie globale consiste à trouver des régions d'homologies au sein de séquences très conservées dans les domaines du vivant (e.g., protéines ribosomiques), puis à définir une distance à partir des pourcentages d'homologie. À l'aide de ces mesures, un arbre comme celui de la figure 6 peut être construit

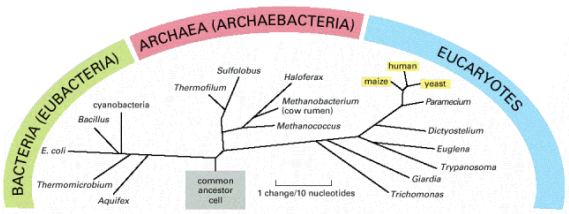


FIG. 6 – Arbre phylogénétique des principales branches de l'arbre de la vie. Les principales branches sont regroupées sous les trois domaines du vivant : la bactéries, les archae et les eucaryotes. (d'après [Alberts et al., 2002])

Chapitre 2

État de l'art

Sommaire

2.1 Les réseaux d'interactions protéine-protéine

Les réseaux d'interactions protéine-protéine sont impliqués dans tous les phénomènes biologiques, certains auteurs avançant même l'hypothèse d'un monde d'interactions protéiques aux débuts de la vie [Andras and Andras, 2005]. Par ce fait, la découverte et l'étude de ces réseaux revêtent un caractère fondamental pour la compréhension des systèmes biologiques. La caractérisation et l'étude des réseaux d'interactions est un domaine d'intense activité, notamment depuis l'émergence des études expérimentales à grande échelle au début des années 2000 ([Uetz et al., 2000], [Ito et al., 2001]). La caractérisation et l'étude des réseaux d'interactions sont des domaines d'étude séparés. La caractérisation consiste à découvrir les relations d'interaction entre les différentes protéines composant le protéome d'une espèce, que ce soit de manière expérimentale (e.g., TAP-TAG [Rigaut et al., 1999], Yeast Two-Hybrid [Fields and Song, 1989]) ou *in silico* [Jansen et al., 2003]. L'étude

des réseaux d'interactions protéine-protéine s'intéresse à différentes questions, dont les réponses sont données par analyse des informations contenues dans les réseaux mêmes. Ces questions peuvent être l'observation de la coévolution de réseaux entre espèces [Fraser et al., 2004], la caractérisation fonctionnelle des gènes [Nabieva et al., 2005], ou encore l'étude des caractéristiques topologiques et structurales des réseaux [Luo et al., 2007]. Dans ces études, l'informatique tient une place de choix, en ce sens que les méthodes informatiques sont une absolue nécessité dans le stockage, le traitement et l'exploitation des données expérimentales, à fins d'analyses et de prédictions [Salwinski and Eisenberg, 2003].

2.1.1 Définition et caractéristiques

Un réseau d'interactions protéine-protéine est composé de protéines, reliées entre elles si elles partagent une interaction physique, déterminée expérimentalement ou par prédiction. Dans le cadre de la détermination des interactions entre protéines, il est important de distinguer les méthodes ayant pour but de déterminer les interactions entre paires de protéines, des méthodes permettant l'étude de la cooccurrence de protéines au sein de complexes. Si la détermination d'interactions binaires est un domaine bien étudié, avec des méthodes expérimentales et informatiques performantes, la caractérisation précise de complexes protéiques reste difficile, tant expérimentalement qu'*in silico*.

Grande échelle et petits mondes Les caractéristiques topologiques des réseaux d'interactions entre protéines sont un sujet d'études nombreuses et variées dans leurs buts. Les principales études sur la topologie et la structure des réseaux d'interactions ont été menées chez *S. cerevisiae*, du fait que cet organisme fait partie des organismes modèles les plus étudiés et que de nombreuses données d'expériences à grande échelle sont disponibles ([Uetz et al., 2000], [Gavin et al., 2002]). Parmi ces études, [Jeong et al., 2001] et [Wagner, 2001] ont montré que les réseaux d'interactions font partie de la catégorie des graphes petits mondes, ou *scale free*. Les réseaux petits mondes sont un phénomène bien connu en sociologie et décrivent le fait que chaque individu d'une population se trouve connecté à n'importe quel autre individu par une courte chaîne de relations sociales. Un exemple très connu est le nombre d'Erdõs, en référence à la distance qui sépare chaque mathématicien du prolixe Paul Erdõs dans le graphe mettant en relation ce dernier avec ses coauteurs, eux-mêmes en relation avec leurs coauteurs, etc. Un autre exemple connu est le graphe de Kevin Bacon, qui met en relation cet acteur américain avec ses différents partenaires au cinéma, ainsi que les partenaires de ses partenaires, et ainsi de suite. Les premiers modèles mathématiques concernant les réseaux petits mondes semblent avoir été publiés par [Watts and Strogatz, 1998]. Depuis, ce modèle a trouvé de nombreuses applications, dans les réseaux électriques, les réseaux neuronaux de *Caenorhabitis elegans*, en biochimie, en protéomique, en sociologie, économie, etc.

La principale caractéristique d'un graphe *scale free* est que la distribution des degrés de connexion des noeuds du graphe suit une loi de puissance. Un tel graphe contient un faible nombre de protéines très connectées, appelées *hubs*, et un grand nombre de protéines faiblement connectées (e.g., figure 7). Cette structure est supposée procurer une certaine robustesse au réseau face à la suppression aléatoire de noeuds dans le réseau. Pour les graphes d'interactions protéiques, cela revient à dire que le réseau est robuste car la probabilité qu'une délétion ou mutation aléatoire touche une protéine peu connectée est très supérieur à la probabilité du même événement pour une protéine très connectée. Cette capacité à supporter la perte d'un nombre relativement important de protéines, due à la structure même du réseau, est cohérente avec les expériences de mutagenèse classique menées chez *S. cerevisiae* ([Winzeler et al., 1999], [Ross-Macdonald et al., 1999]). La topologie *scale free* du réseaux d'interactions de *S. cerevisiae* se retrouve chez d'autres espèces, comme *Drosophila melanogaster* [Giot et al., 2003] ou *Caenorhabitis elegans* [Li et al., 2004]. Cependant, d'autres auteurs remettent en cause cette classification.

Dans leur étude, [Han et al., 2005] mettent en avant le fait que les réseaux d'interactions actuellement connus ne sont que très partiels et qu'il peut ainsi être aventureux de déduire la topologie globale du réseau à partir d'une faible portion de ce dernier. En effet, chacune des expériences de caractérisation à grande échelle des interactions couvre une faible partie du réseau global et le recouvrement entre expériences est limité [von Mering et al., 2002]. Pour définir la topologie du réseau global à partir d'un échantillon de taille réduite, il faut donc faire l'hypothèse que cette couverture limitée ne biaise pas les mesures. [Stumpf et al., 2005] montrent dans leur étude que des échantillons aléatoires de graphes *scale free* ne sont pas eux-mêmes *scale free*, alors que d'autres types de graphes (e.g., graphes aléatoires) vérifient cette hypothèse. Les résultats de l'étude de [Przulj et al., 2004] montrent de même que le modèle de graphe petit monde échoue à représenter certaines caractéristiques des données disponibles, alors qu'un graphe aléatoire présentant une distribution géométrique des degrés est plus en accord avec les données expérimentales.

Quelle est la question ? Graphe petit monde ou graphe aléatoire (figure 8), telle n'est pas la question. Certains auteurs s'accordent sur la nature *scale free* des réseaux d'interactions protéiques ([Wagner, 2001], [Giot et al., 2003]), d'autres contestent cette hypothèse et apportent des arguments intéressants ([Han et al., 2005], [Stumpf et al., 2005]). Dans tous les cas, la question la plus importante n'est pas de savoir quel modèle de graphe s'adapte le mieux aux données réelles, mais plutôt de savoir s'il est raisonnable de choisir un modèle et d'utiliser ses propriétés afin de découvrir les informations contenues dans les réseaux d'interactions. À cette question, nous pouvons apporter plusieurs éléments de réponse.

Tout d'abord d'un point de vue théorique, voire philosophique, la propriété de robustesse du graphe petit monde face à des pertes ou gains aléatoires de protéines est un élément intéressant, conforme à ce que l'on peut penser raisonnable en

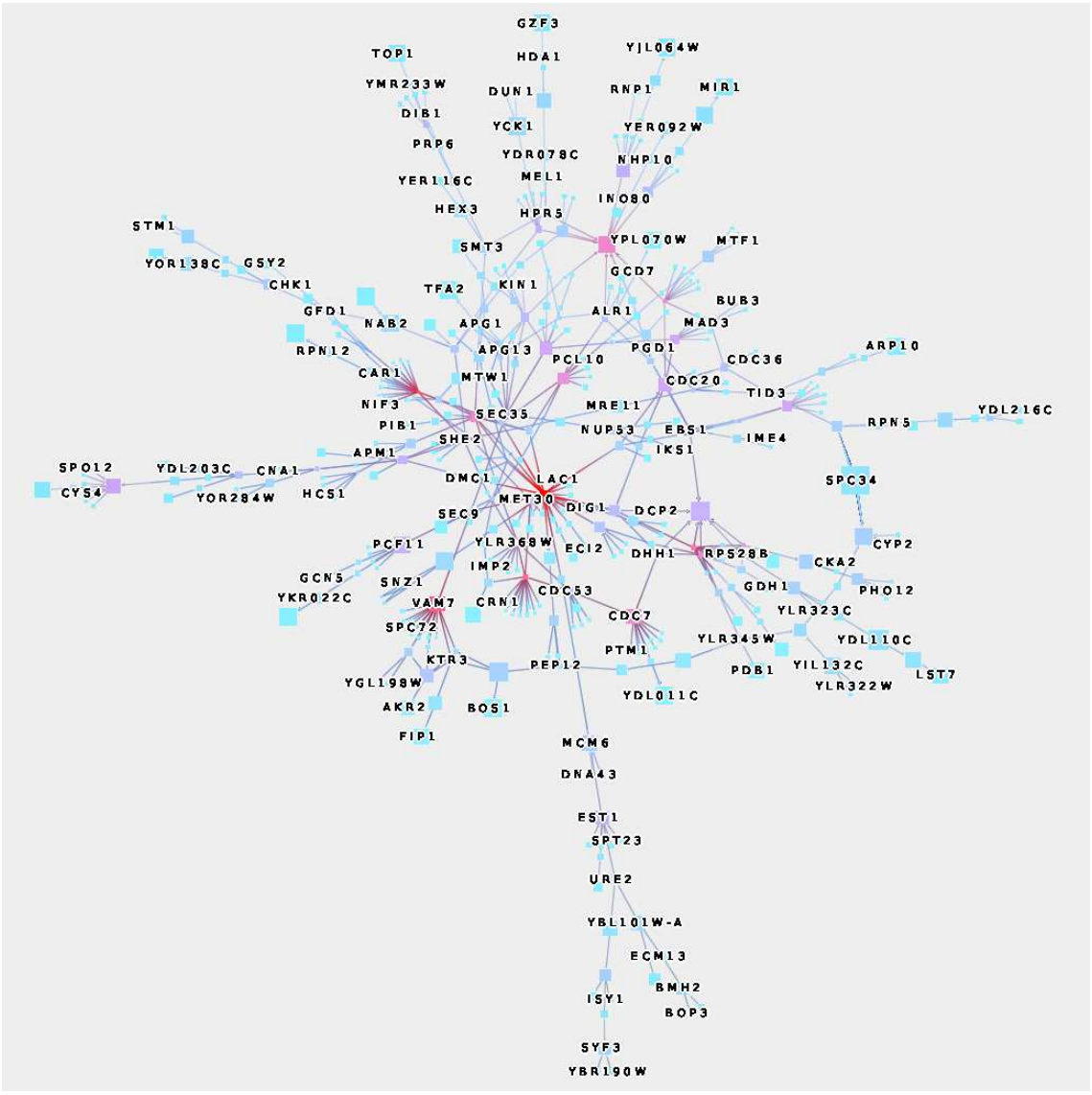


FIG. 7 – Exemple de graphe petit monde. Ce graphe, obtenu grâce au logiciel Tulip [Auber, 2001], représente des données d'interactions protéine-protéine obtenues par la méthodes Yeast Two-Hybrid chez la levure *S. cerevisiae*. Ce graphe fait partie de la classe des réseaux petits mondes, notamment caractérisés par un faible nombre de sommets très connectés et un grand nombre de sommets peu connectés. Cette caractéristique structurale confère une grande robustesse à ce type de réseau.

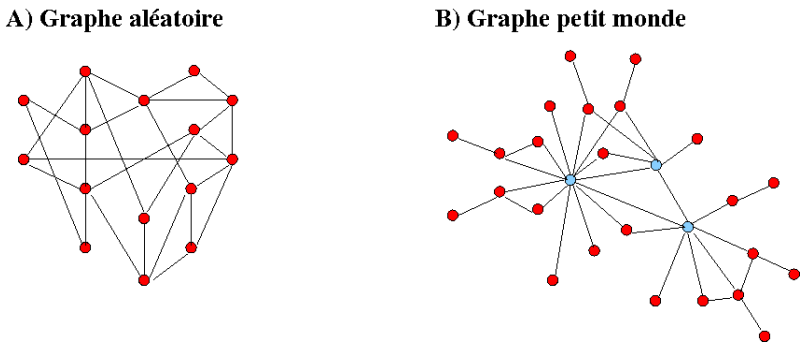


Fig. 8 – Comparaison des modèles de graphe aléatoire et *scale free*. Le graphe aléatoire ou graphe d'Erdös-Rényi contient N sommets, chaque paire étant connectée avec une probabilité p, avec environ pN(N-1)/2 arêtes. la distribution des degrés suit une loi de Poisson. Le graphe *scale free* est caractérisé par un faible nombre de sommets très connectés, appelés *hubs*. La distribution des degrés des sommets suit une loi de puissance. (d'après [Barabasi and Oltvai, 2004])

termes de mécanismes d'évolution. La topologie *scale free* procurant un certain avantage évolutif, il est plausible que cette structure ait été positivement sélectionnée. Ensuite, si la couverture du réseau global par les données expérimentales semble faible, il apparaît que ce sont surtout les interactions transitoires qui sont manquantes, car difficiles à caractériser expérimentalement. De plus, de nombreuses études ([Jeong et al., 2001], [Wagner, 2001], [Hahn and Kern, 2005]) ont modélisé les réseaux d'interactions par des graphes petits mondes et en ont utilisé les propriétés pour découvrir des informations biologiquement pertinentes. Enfin, l'analyse de [Yook et al., 2004], portant sur les bases de données MIPS [Mewes et al., 2002] et DIP [Xenarios et al., 2000], ainsi que sur les données de [Uetz et al., 2000] et de [Ito et al., 2001], a entre autres permis de montrer que la topologie *scale free* permet de caractériser ces jeux de données.

La réponse à la question est donc qu'il semble raisonnable d'utiliser les graphes petits mondes pour modéliser les réseaux d'interactions, puis d'en exploiter les propriétés intrinsèques. Toutefois, l'utilisation de ce modèle doit se faire en sachant qu'il ne reflète peut-être pas totalement les caractéristiques des données expérimentales, et donc que tout résultat basé uniquement sur l'analyse des propriétés structurales doit être envisagée avec précaution. De plus, la possibilité de modéliser les données expérimentales à l'aide d'autres types de graphes ouvre de nouvelles perspectives d'études basées sur les propriétés intrinsèques de ces graphes, qui devront à leur tour faire preuve de validité et d'utilité.

Particularités des *hubs* Dans leur étude, [Jeong et al., 2001] tentent d'établir une corrélation entre la topologie particulière des réseaux d'interactions et l'essentialité des protéines. Pour ce faire, ils construisent tout d'abord le réseau d'interactions de *S. cerevisiae* à partir des données de [Uetz et al., 2000], puis caractérisent sa topologie et montrent son appartenance à la classe des graphes petits mondes. Par la suite, les auteurs classent les protéines en fonction de leur connectivité (i.e., le nombre d'interactions dans lesquelles elles sont impliquées) et observent que la probabilité qu'une protéine soit essentielle est trois fois plus importante pour les protéines fortement connectées que pour celles faiblement connectées, mettant ainsi en avant le rôle central de ces *hubs*.

Une autre propriété intéressante est que la vitesse d'évolution des protéines fortement connectées est significativement plus lente que celle des autres protéines. Cette propriété a été montrée par les résultats des études de [Hirsh and Fraser, 2001] et [Fraser et al., 2002]. Pour l'expliquer, plusieurs hypothèses sont possibles. L'hypothèse la plus simple et la plus immédiate, est que les protéines très fortement connectées subissent une forte pression de sélection, limitant leur évolution, afin de ne pas entraîner de déconnexion du réseau d'interactions protéiques, pouvant résulter en des conséquences dramatiques pour la cellule. Dans leur étude, [Fraser et al., 2002] montrent plus précisément que les protéines centrales ont une vitesse d'évolution inférieure aux autres protéines, et que de plus, les protéines en interaction avec ces *hubs* évoluent à la même vitesse. D'un point de vue évolutif, cette idée se comprend aisément. En effet, si une protéine centrale subit une mutation, il est probable que sa capacité à interagir va en être changée. Aussi, toujours dans le but de limiter les conséquences néfastes pour la cellule, les mutations compensatoires dans les partenaires des *hubs*, mutations qui permettent de maintenir les interactions, vont être sélectionnées de manière préférentielle. Cette propriété d'évolution lente est cependant nuancée, si ce n'est remise en cause, par [Jordan et al., 2003]. Ces auteurs montrent dans leur étude, que seule une faible partie des *hubs*, concernant les protéines les plus connectées parmi les *hubs*, possède cette propriété. [Fraser and Hirsh, 2004] apportent un éclairage différent sur ce constat. Ces auteurs montrent en effet que la corrélation entre connectivité et vitesse d'évolution dépend grandement du jeu de données sur lequel l'étude est effectuée. Les auteurs montrent que les contradictions observées entre les différents jeux de données sont dues à la faible couverture individuelle par rapport au réseau global, ainsi qu'au faible taux de recouvrement des différentes expériences, et enfin à la nature des interactions détectées (e.g., complexes protéiques pour le TAP-TAG [Rigaut et al., 1999], interactions binaires pour le Y2H [Fields and Song, 1989]). De ces différentes études, nous pouvons garder l'idée que les protéines centrales ont une vitesse d'évolution inférieure aux autres protéines, mais que cette observation n'est pas un critère absolu d'identification de ces *hubs*, puisque cette propriété n'émerge pas de tous les types de données d'interaction.

Une dernière caractéristique intéressante est que les *hubs* peuvent être divisés en deux catégories principales : les *hubs* statiques et les *hubs* dynamiques. Cette distinction a été mise en avant par [Han et al., 2004] dans une étude basée sur l'analyse de profils d'expression. Ces auteurs ont ainsi qualifié certaines protéines centrales de statiques, lorsqu'elles interagissent au sein de complexes protéiques et donc avec la majorité de leurs partenaires au même moment (*party hubs*), et d'autres de dynamiques, pour les protéines interagissant avec leurs partenaires en des moments ou localisations différents (*date hubs*). D'après cette étude, il semble que les *hubs* statiques soient les modules fonctionnels centraux du système, alors que les *hubs* dynamiques permettraient d'établir des liaisons entre ces modules centraux. Cette observation permet de penser que la topologie du réseaux global est principalement due aux *hubs* dynamiques.

D'un point de vue phylogénétique, [Ekman et al., 2006] ont montré qu'une grande partie des *hubs* est conservée au sein du domaine des eucaryotes, par comparaison avec les protéines non-*hubs*. Ces auteurs observent de plus que les *hubs* statiques sont mieux conservés dans d'autres domaines du vivant, notamment chez les procaryotes, que les *hubs* dynamiques. Cette étude, menée chez la levure *S. cerevisiae*, utilise la recherche d'orthologues dans différentes espèces des domaines du vivant. Pour valider leurs résultats, les auteurs ont observé la conservation de différentes classes de domaines protéiques (domaines anciens, spécifiques aux eucaryotes, spécifiques aux levures ou orphelins) dans les différentes classes de protéines (*hubs* ou non, statiques ou dynamiques). Les résultats semblent confirmer les observations basées sur la recherche d'orthologues. En effet, les protéines non-centrales possèdent une plus grande proportion de domaines spécifiques aux levure ou orphelins que les protéines centrales. De même, les *hubs* statiques contiennent une plus grande proportion de domaines anciens et une plus faible proportion de domaines spécifiques aux levure que les *hubs* dynamiques. Bien qu'une règle stricte de corrélation ne puisse être établie, il apparaît que les protéines les plus anciennes soient des *hubs*, et particulièrement des *hubs* statiques, même si certaines protéines peu connectées sont anciennes.

[Ekman et al., 2006] font une autre observation concernant la composition en domaines des *hubs*. Les auteurs observent que les *hubs* sont souvent des protéines multi-domaines, plus longues en moyenne que les protéines moins connectées. Ils établissent de plus un autre critère de distinction entre *hubs* statiques et dynamiques. Les premiers partagent une majorité de domaines avec leurs partenaires d'interaction et ont souvent une structure en domaines très conservée. Les seconds partagent un nombre de domaines moins important avec leurs partenaires et présentent fréquemment des zones peu structurées, pouvant permettre une certaine flexibilité au niveau des interactions possibles [Ward et al., 2004].

2.1.2 Sources et formats de données

Les sources de données d'interactions protéine-protéine sont très nombreuses et différentes. Il faut ici distinguer les sources de données provenant directement d'expérimentations à plus ou moins grande échelle ([Uetz et al., 2000], [Ito et al., 2001], [Gavin et al., 2002], [Uetz and Pankratz, 2004]), des bases de données visant principalement à établir un catalogue des interactions protéiques connues (DIP [Xenarios et al., 2000], MIPS [Mewes et al., 2002], BIND [Bader et al., 2001] ou encore IntAct [Hermjakob et al., 2004b]). Un aspect connexe est l'échange des données d'interactions protéiques.

Il est important, dans le cadre de nos recherche, de bien connaître les spécificités des différentes sources de données, ainsi que les conséquences de ces particularités. Dans cette partie, nous ferons un rapide tour d'horizon des sources de données expérimentales et des bases de données, puis nous nous intéresserons au format PSI-MI [Hermjakob et al., 2004a], standard de représentation et d'échange d'interactions biomoléculaires.

Sources de données Les données expérimentales d'interactions protéiques proviennent majoritairement d'expériences à grande échelle (e.g., [Ho et al., 2002]). Nous avons eu un aperçu des principales caractéristiques de ces données en section 2.1.1, notamment concernant la couverture des expériences individuelles, ainsi que le recouvrement des données qu'elles fournissent. Nous parlerons plus en détail en section 2.1.4 des défauts des différentes méthodes expérimentales et prédictives, ainsi que des méthodes nécessaires à la prise en compte de ces défauts. Les données expérimentales actuellement disponibles concernent majoritairement quelques organismes modèles, tels que la levure *S. cerevisiae* ou encore *Drosophila melanogaster*.

Dans cette partie, nous nous focaliserons sur les bases des données d'interactions protéiques. Comme nous l'avons vu précédemment, les bases de données d'interactions protéiques sont nombreuses (BIND, DIP, IntAct, etc). Si au départ les différences se faisaient tant d'un point de vue du contenu que de la manière d'accéder à celui-ci, la plupart des bases de données disposent aujourd'hui des mêmes informations. En effet, dans le cadre de la mise en place du consortium IMEx (International Molecular Exchange[1]), les principaux acteurs du domaine se sont engagés dans un processus d'échange de leurs données grâce à l'adoption d'un format commun d'échange. Parallèlement à ce processus, un effort d'annotation et de vérification des données a été entrepris.

Parmi les différents acteurs, nous pouvons citer IntAct[2] comme étant l'une des bases de données les plus actives et les plus complètes, tant en termes de données que d'outils de recherche et d'analyse. IntAct contient environ 136 000 interactions

[1] http ://imex.sf.net
[2] http ://ebi.ac.uk/intact

binaires, couvrant 56 000 protéines et 6000 expériences. Dans ces données, plus de 50% des expériences utilisent la technique du double hybride [Fields and Song, 1989] ou une de ses variantes ([Vidal et al., 1996], [Ito et al., 2001]). De plus, trois espèces (*S. cerevisiae*, *Drosophila melanogaster* et *Homo sapiens*) représentent à elles-seules 75% des interactions identifiées. IntAct fournit des moyens de recherche classiques (recherche par mot-clé, identifiant, référence croisée, etc), ainsi que des outils permettant d'analyser les résultats des recherches. Ces outils, utilisables en ligne, permettent de voir le graphe représentant le réseau d'interactions protéiques recherchées, de chercher la séquence d'interactions reliant deux protéines d'intérêt, ou encore de comparer la composition en domaines InterPro [Mulder et al., 2007] des protéines en interactions. D'autres logiciels tels ProViz ([Iragne et al., 2005], section 3.5), sont des logiciels utilisables hors ligne et permettent des analyses dynamiques et interactives des données d'interactions.

Une caractéristique d'IntAct est l'utilisation d'un vocabulaire contrôlé, permettant de décrire de manière unique et non-ambiguë toutes les expériences, et ainsi de faciliter l'utilisation des données stockées. Ce vocabulaire contrôlé, développé de manière interne dans un premier temps, a été mis en commun et enrichi dans le cadre du projet HUPO PSI-MI (Human Proteome Organisation - Proteomic Standards Initiative - Molecular Interactions[3]).

Représentation et échange de données Comme nous venons de le voir, les principales bases de données d'interactions biomoléculaires font actuellement un effort d'annotation, de vérification et d'échange de leurs données d'interactions (consortium IMEx). L'échange de données a été rendu possible par la définition d'un format XML commun, permettant le stockage non-redondant des interactions biomoléculaires (e.g., protéine-protéine, protéine-ADN), ainsi que de toutes les informations nécessaires à l'exploitation de ces données (références externes, méthodes expérimentales, séquences protéiques, publications, etc). Ce format a été développé dans le cadre du projet HUPO PSI [Hermjakob et al., 2004a], qui a pour but la définition de formats standards pour l'échange de données de protéomique, principalement pour la spectrométrie de masse et les interactions protéine-protéine. Une présentation générale de ce format est disponible à l'adresse suivante :
`http://psidev.sourceforge.net/mi/xml/doc/user/#structure`.

La définition de ce format est un élément très important pour favoriser l'échange de données. Cependant, un format unique ne garantit pas à lui seul la compatibilité des différentes sources de données. Pour répondre à ce problème, un vocabulaire contrôlé a été développé. Ce vocabulaire consiste en la définition unique et non-ambiguë de tous les termes nécessaires à la caractérisation d'une interaction (e.g., méthode de caractérisation, méthode de marquage, etc). Parallèlement à l'utilisation de ce vocabulaire contrôlé, l'usage de références externes et d'ontologies (e.g.,

[3]http ://psidev.sourceforge.net/mi/xml/doc/user/

Gene Ontologie [Harris et al., 2004]) est largement encouragé. Le format PSI-MI, associé à l'utilisation d'un vocabulaire contrôlé, permet ainsi de combiner des jeux de données provenant d'expériences différentes et donc de réaliser des études sur des ensembles de données plus importants. De nombreux outils[4] et logiciels (e.g., Proviz [Iragne et al., 2005], PIMWalker[5]) utilisent le format PSI-MI, que ce soit comme source de données ou comme format d'export.

2.1.3 Prédiction d'interactions protéine-protéine

La prédiction de réseaux d'interactions protéine-protéine regroupe des techniques variées, tant dans les approches qu'elles utilisent que dans les buts qu'elles visent. Les méthodes de prédiction d'interactions s'attaquent à deux problèmes distincts que sont la prédiction de complexes protéiques et la prédiction d'interactions binaires. Les techniques employées pour résoudre ces questions utilisent différentes approches, généralement basées sur les propriétés topologiques et structurales des réseaux, ou bien sur l'intégration de différentes données (e.g., fouille de données, réseaux bayésiens).

Approches topologiques

[Yu et al., 2006] présentent dans leur article une méthode permettant de prédire des interactions protéiques en se basant uniquement sur les propriétés structurales du graphe d'interactions. L'idée de base de leur approche vient de la manière dont sont menées les expériences de caractérisation de complexes protéiques [Gavin et al., 2002], où une protéine est utilisée comme appât pour accrocher toutes les protéines pouvant interagir avec elle. Cette expérience est répétée, mais cette fois en utilisant comme appât une protéine caractérisée précédemment, et ainsi de suite jusqu'à ce que toutes les protéines interagissant aient été utilisées comme proie. Le résultat est donc une liste de partenaires pour chaque protéine appât, les listes étant très similaires puisque représentant un même complexe protéique. Cependant, la variabilité inhérente à l'expérimentation biologique fait que ces listes ne sont pas strictement identiques. De manière générale, le graphe d'interactions représentant le complexe caractérisé contient tous les partenaires identifiés, et non seulement le sous-ensemble identifié par toutes les expériences. La méthode de [Yu et al., 2006] cherche donc à mimer cette décision de complètement du graphe d'interactions. Un complexe protéique est représenté par un graphe, appelé clique (figure 9), dans lequel chaque protéine est reliée à toutes les autres (modèle matrice [Bader and Hogue, 2003]). L'algorithme de Yu *et al.* cherche donc à trouver des cliques de taille maximale dans un réseau d'interactions existant, afin de pouvoir compléter les relations manquantes. Dans l'exemple de la figure 9, nous pouvons voir que la clique K est composée de

[4]http ://psidev.sf.net
[5]http ://pim.hybrigenics.com

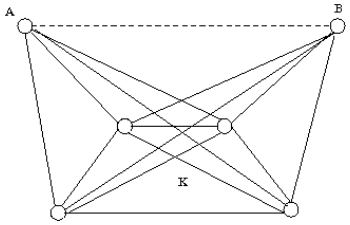


Fig. 9 – Prédiction d'interactions par complètement de clique. [Yu et al., 2006] proposent l'idée que la clique K est incomplète, car le lien entre les sommets A et B est manquant. Leur solution est donc d'ajouter un arc entre les sommets A et B. Les traits pleins indiquent les interactions expérimentalement caractérisées. Le trait en pointillés indique l'interaction prédite.

quatre protéines reliées les unes aux autres. Nous observons que les protéines A et B sont reliées à toutes les protéines de la clique K, mais ne sont pas reliées entre-elles. La méthode de Yu *et al.* permet de corriger cet oubli en prédisant une interaction entre A et B, complétant ainsi la clique K.

Estimant la qualité de leur méthode et des résultat qu'elle produit, [Yu et al., 2006] montrent que les mesures statistiques d'évaluation de qualité sont meilleures pour leurs résultats que pour les données expérimentales à grande échelle (e.g., [Gavin et al., 2002]). D'autres mesures statistiques, ainsi que des exemples biologiquement pertinents leur permettent d'affirmer la qualité des résultats produits. Par comparaison avec l'algorithme *MCODE* de [Bader and Hogue, 2003], Yu *et al.* montrent que leurs résultats ne sont qu'un sous-ensemble des résultats fournis par MCODE, mais de confiance supérieure.

MCODE est un algorithme basé sur les propriétés topologiques des graphes. Le but de cet algorithme est de délimiter des sous-graphes densément connectés dans les réseaux d'interactions protéiques, émettant l'hypothèse que ces sous-graphes représentent des complexes protéiques. Cette hypothèse se base sur l'acceptation de la topologie *scale-free* pour décrire les réseaux d'interactions. MCODE peut fonctionner en mode dirigé, dans lequel il permet d'extraire un seul sous-graphe contenant une protéine d'intérêt, ou en mode automatique, dans lequel il extrait un réseau ne comportant que des régions fortement connexes. Comme la méthode de Yu *et al.*, MCODE se base sur les propriétés topologiques du réseau pour prédire les complexes protéiques. Or, nous avons vu que la classe des graphes petits mondes n'est pas forcément la mieux adaptée pour représenter les réseaux d'interactions (section 2.1.1). [Bader and Hogue, 2003] font cette même remarque dans leur article décrivant MCODE.

S'il est raisonnable de penser que les deux méthodes que nous venons de présenter sont d'une utilité certaine pour la détection de complexes protéiques dans les données d'interactions, il faut aussi garder à l'esprit que cette détection est très dépendante de la qualité et de la nature des données. En effet, certaines méthodes expérimentales,

comme celle utilisée par [Gavin et al., 2002], permettent de caractériser en un seule passe tous les partenaires d'un complexe protéique, sans pour autant fournir la liste exacte des interactions liant chaque partenaire à l'intérieur du complexe. Dans ce cas, plusieurs solutions sont disponibles pour représenter le complexe. Soit nous considérons que toutes les protéines interagissent les unes avec les autres (modèle matrice), soit nous considérons que chaque protéine n'interagit qu'avec la protéine appât de l'expérience (modèle spoke [Bader and Hogue, 2002]). Il est évident que chacun de ces modèles ne peut représenter de manière parfaite les relations internes entre partenaires du complexe (figure 10). Dans le cas du modèle matrice il est très probable que toutes les protéines ne peuvent être en interaction les unes avec les autres. Dans le cas du modèle spoke, il est improbable qu'il n'y ait pas d'interaction entre les protéines périphériques. Bader *et al.* indiquent cependant que le modèle spoke se révèle jusqu'à trois fois plus précis que le modèle matrice. La méthode de [Yu et al., 2006] étant basée sur l'idée du modèle matrice, nous pouvons donc considérer qu'un biais existe dans la prédiction d'interactions binaires. Un dernier fait doit être pris en compte pour l'évaluation des méthodes de Bader *et al.* et Yu *et al.*. Nous avons vu en section 2.1.1 que les *hubs* des réseaux d'interactions protéiques peuvent être séparés en deux catégorie. Les *hubs* statiques sont supposés interagir en même temps avec l'ensemble de leurs partenaire, et ce de manière stable, alors que les *hubs* dynamiques rencontrent leurs partenaires en des endroits et moments différents. Les deux types de *hubs* partagent cependant la même caractéristique topologique, à savoir une forte connexité. Les méthodes de Bader *et al.* et Yu *et al.* ne pouvant faire de distinction entre *hubs* dynamiques et statiques, ils prédisent tous deux des complexes protéiques monoblocs pour des *hubs* dynamiques, complexes qui n'ont pas de réalité biologique. Ainsi, ces méthodes semblent pouvoir être utilisées comme une première approche pour la prédiction de partenaires impliqués dans un même complexe protéique, mais les relations binaires prédites doivent être envisagées avec prudence.

Approches par extrapolation

La prédiction de réseaux d'interactions protéine-protéine est une tâche délicate. Nous avons vu précédemment qu'une relation de voisinage, de cooccurrence dans un complexe protéique, est un problème auquel il est possible de répondre avec une certaine confiance. La relation d'interaction entre deux protéines se place à un niveau de détail plus élevé et est donc par ce fait plus délicate à réaliser.

Dans [Wuchty, 2006], l'auteur présente une méthode de prédiction d'interactions protéine-protéine basée sur l'interaction de domaines protéiques. L'auteur se base sur le résultat de l'étude de [Aloy et al., 2004], qui montrent que 94% des interactions protéiques de la levure *S. cerevisiae* sont gouvernées par l'interaction d'une seule paire de domaines protéiques. Suivant cette idée, Wuchty teste l'hypothèse selon laquelle la qualité d'une interaction protéique est corrélée avec la plus forte

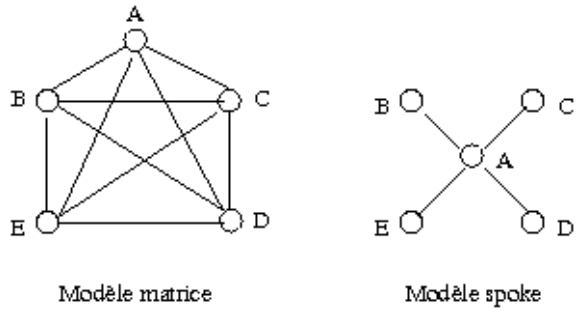


Fig. 10 – Comparaison des modèles *spoke* et matrice. Les deux graphes représentent le même complexe protéique. La protéine A a été utilisée comme appât et les protéines B, C, D et E ont été caractérisées comme partenaires de A. Le modèle matrice représente une interaction entre chaque paire de protéines du complexe, alors que le modèle spoke représente une interaction entre l'appât et ses partenaires identifiés.

probabilité d'interaction de paires de domaines des protéines impliquées (figure 11). Pour ce faire, il utilise les données d'interactions protéiques de *S. cerevisiae* [Bader et al., 2004] et de *Drosophila melanogaster* [Giot et al., 2003], dans lesquelles un score de confiance est attribué à chaque interaction.

Cette expérience permet à l'auteur de montrer qu'il existe une forte corrélation entre le niveau de confiance attribué à l'interaction et l'*expectation value* (probabilité que l'événement considéré soit dû au hasard) de l'interaction de domaines de plus haute confiance. Son hypothèse de départ apparaissant valide, l'auteur applique alors sa méthode à la prédiction d'interactions protéine-protéine chez *Plasmodium falciparum*. Pour ce faire, il annote les protéines de cet organisme en fonction des annotations de la base Integer8 [Kersey et al., 2005] et des domaines de la base de donnée Pfam [Bateman et al., 2004]. Le résultat de cette prédiction est un ensemble de 1428 interactions protéiques, impliquant 361 protéines. L'auteur tente alors d'évaluer la qualité de ses résultats en étudiant la corrélation des interactions prédites avec d'autres informations, telles que les données d'expression, la cooccurrence de termes Gene Ontology [Harris et al., 2004], et recherche enfin les mêmes interactions dans d'autres organismes en utilisant la base de données InParanoid [O'Brien et al., 2005] comme moyen d'identification d'orthologues. Dans ces trois tests, l'auteur montre que les prédictions issues de sa méthode sont de qualité. Un dernier test effectué consiste en la comparaison des prédictions avec un jeu de données expérimentales obtenues chez *Plasmodium falciparum* [Lacount et al., 2005]. Ce test semble échouer, puisque seules deux interactions sont communes aux données expérimentales et prédites. Pour expliquer ce très faible recouvrement, l'auteur avance deux raisons, qui sont la faible annotation des protéines de *Plasmodium falciparum* en domaines Pfam, ainsi que la faible couverture expérimentale du protéome (environ un quart des protéines on été étudiées).

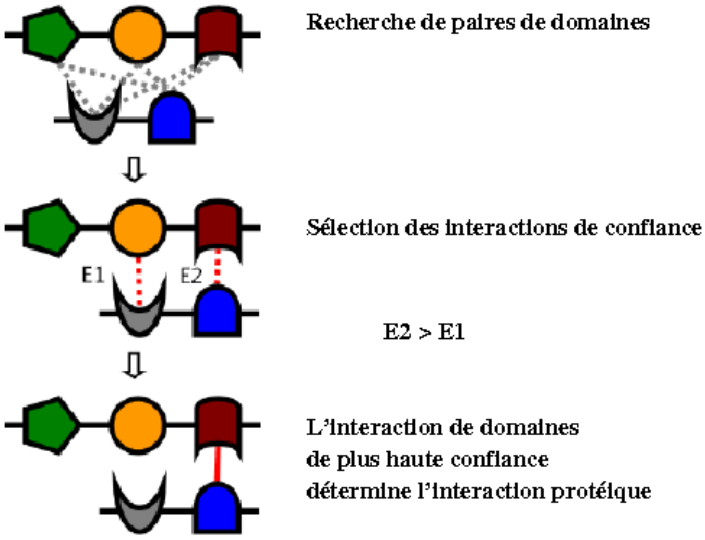


FIG. 11 – Prédiction d'interactions par interaction de paires de domaines. Chaque paire de protéines est l'objet d'une recherche de domaines protéiques. Une sélection est ensuite opérée pour ne retenir que les interactions de grande confiance entre domaines. La dernière étape infère une interaction à partir de l'interaction de domaines de plus grande confiance. (d'après [Wuchty, 2006]).

Cette étude est un bon exemple d'une tactique consistant à étudier un problème plus simple ou mieux connu, pour extrapoler les résultats à un domaine plus complexe. Ici, l'auteur montre la rélation directe qui existe entre interaction protéique et interaction de domaines protéiques. Il applique ensuite sa méthode pour prédire avec succès des interactions protéiques chez *Plasmodium falciparum*. La prédiction de domaines et de leurs interactions étant un problème moins complexe et mieux connu, la tâche de prédiction d'interactions protéiques est donc facilitée. Nous avons nous-même développé de telles méthodes dans le cadre de travaux antérieurs à la thèse [Goffard et al., 2003].

Approches intégratives

Les approches intégratives utilisent des modèles probabilistes dont les valeurs sont obtenues par apprentissage sur des jeux de données en relation avec les interactions protéiques. De nombreuses techniques et variantes de techniques sont développées et utilisées, et peuvent globalement être séparées en deux catégories, selon qu'elles présupposent ou non l'indépendance des variables qu'elles considèrent. Ainsi, les réseaux bayésiens naïfs considèrent que toutes les variables sont indépendantes. Or, de nombreux phénomènes sont interdépendants dans les cellules. Ainsi, nous savons que les interactions de type protéine-ADN sont fortement liées aux niveaux d'expression des gènes impliqués. Cependant,

ces méthodes semblent fournir des résultats intéressants [Jansen et al., 2003]. Au contraire, d'autres techniques tels que les arbres de décisions ne considèrent pas les variables comme indépendantes. Enfin, il faut noter que certaines techniques postulant l'indépendance des variables peuvent être modifiées pour prendre en compte une interdépendance de données, telle la méthode proposée par [Jaimovich et al., 2006]. Nous ne nous intéresserons pas en détail aux différentes méthodes d'intégration, mais présenterons leur principe de fonctionnement au travers de quelques exemples. Pour une étude plus détaillée et comparée des différentes approches, l'article de [Qi et al., 2006], dans lequel les auteurs exposent et comparent six classificateurs pour la prédiction d'interactions protéine-protéine, peut être consulté.

Les jeux de données d'apprentissage Toutes ces techniques font partie du domaine de la fouille de donnée et de l'apprentissage automatique. La stratégie générale de ces méthodes est d'apprendre un modèle statistique en se basant sur un jeu d'apprentissage positif contenant des interactions de haut niveau de confiance, ainsi que sur un jeu d'apprentissage négatif qui ne contient que des interactions qui n'existent pas. La définition de ce jeu de données d'apprentissage (*gold standard* dans la littérature) est donc une étape cruciale dans l'utilisation de ces méthodes. Comme le notent [Jansen and Gerstein, 2004], la définition des jeux d'apprentissage est un problème plus ou moins complexe en fonction de la nature des données, ainsi que de la question posée. En effet, si la définition d'ensembles d'interactions et d'absence d'interactions physiques entre protéines est un problème qui semble simple à résoudre, il n'en est pas de même pour la définition d'ensembles de fonctions biologiques. Pour la prédiction de fonctions biologiques des protéines, il est possible de définir les données positives, par exemple en regardant les termes Gene Ontology attribués à une protéine, mais il est compliqué de définir qu'une protéine n'a pas une fonction, puisque la caractérisation expérimentale des protéines recherche rarement ce que ne fait pas une protéine. Pour la prédiction des interactions protéine-protéine, qui est notre principal centre d'intérêt, le choix se fait entre deux possibilités, à savoir interaction ou absence d'interaction.

Le jeu de données négatives est généralement construit à partir de données de localisation cellulaire ([Kumar et al., 2002] et [Huh et al., 2003]). L'hypothèse est que deux protéines qui ne sont pas dans le même compartiment cellulaire ont une faible chance d'interagir. Il est possible d'émettre plusieurs réserves à l'encontre de cette hypohtèse, en remarquant tout d'abord qu'il existe des mécanismes de translocation cellulaire pouvant transporter deux protéines séparées dans un même compartiment, et que de plus, les données expérimentales de localisation ne sont peut-être pas complètes. Il s'avère cependant, comme le montrent [Jansen and Gerstein, 2004], que ces données sont robustes et permettent la définition d'ensembles négatifs de qualité. À titre de confirmation, nous pouvons noter que les interactions protéiques du MIPS [Mewes et al., 2002], souvent considérées de qualité, ne comportent que 1,5% d'interactions entre protéines issues de compartiments cellulaires différents.

Certains auteurs [Ben-Hur and Noble, 2006] notent tout de même que la création de jeux de données négatives, basés uniquement sur le critère de non-colocalisation, peut entraîner un biais dans les mesures de qualité des interactions prédites.

Le jeu de données positives est généralement construit à partir d'ensembles d'interactions protéiques de la meilleure qualité disponible. Dans leur article, [Qi et al., 2006] utilisent trois jeux de référence, couramment employés :

- DIP [Xenarios et al., 2000] *small-scale subset* pour les interactions physiques
- les complexes protéiques du MIPS pour la cooccurrence dans des complexes
- les voies métaboliques du KEGG [Ogata et al., 1999] pour la cooccurrence dans des voies métaboliques

De manière générale, les données issues d'expériences à grande échelle ne sont pas de bonnes candidates pour la définition du jeu d'apprentissage, à cause de leur faible qualité. Les jeux de données du type DIP *small-scale subset*, qui ne contient que des interactions protéiques déterminées par des expériences individuelles et annotées, présentent des interactions de plus haute confiance et sont donc préférés.

Application à la prédiction de cooccurrence dans les complexes [Jansen et al., 2003] publient les résultats de leur méthode prédisant la cooccurrence de paires de protéines dans un complexe. Cette méthode est basée sur l'utilisation de réseaux bayésiens, permettant l'intégration de données génomiques (e.g., coessentialité, colocalisation) et de données expérimentales. Elle peut-être utilisée pour combiner des jeux de données, contenant éventuellement du bruit, ou bien pour effectuer une prédiction *de novo* d'interactions protéiques à partir des données génomiques.

La stratégie classique de ce genre d'étude est tout d'abord de définir les jeux de données d'apprentissage. Les auteurs ont ici choisi les complexes protéiques du MIPS [Mewes et al., 2002] pour la partie positive et les données de localisation cellulaire pour la partie négative [Kumar et al., 2002]. Ensuite, un ensemble de caractéristiques des protéines doit être défini. Dans cette étude, les auteurs se penchent sur l'expression des gènes correspondant aux protéines en interaction, ainsi que sur les fonctions biologiques (termes GO ou classification MIPS) et l'essentialité de ces protéines. Le jeu d'apprentissage est alors séparé en deux parties, un jeu d'entraînement de l'algorithme et un jeu de test. Cette séparation permet d'évaluer la qualité de prédiction du réseau bayésien généré, en comparant ces prédictions avec les données du jeu de test, donnant ainsi deux mesures très importantes que sont les taux de vrais et faux positifs. Un vrai-positif est une interaction prédite et qui existe réellement, alors qu'un faux-positif est une interaction prédite alors qu'elle n'existe pas. Le ratio de ces deux mesures permet de juger de la qualité de la prédiction.

Le résultat du processus d'apprentissage est donc un réseau bayésien valué, qui permet de réaliser les prédictions d'interactions. Le réseau est constitué d'un suite

de probabilités conditionnelles. Un exemple typique de règle encodée dans un réseau bayésien pourrait être :

1. Soient deux protéines A et B, caractérisées par un vecteur de propriétés
   - A = {Exa = caractéristiques d'expression de A, Fa = fonction biologique de A, Esa essentialité de A}
   - B = {Exb = caractéristiques d'expression de B, Fb = fonction biologique de B, Esb = essentialité de B}
2. Soit Pab, la probabilité que A et B soient dans un même complexe protéique.
3. Quelle est la valeur de Pab sachant que Exa, Exb, Fa, Fb, Esa, Esb ?
4. Le réseau bayésien fourni cette valeur : si celle-ci est au dessus de la limite, une interaction est prédite

La même méthode permet à [Jansen et al., 2003] de combiner différentes sources de données expérimentales sur les complexes protéiques. La combinaison de ces données expérimentales et la prise en compte des prédictions précédentes, permettent de prédire un ensemble de complexes protéiques associés à des mesures de qualité et couvrant l'ensemble du protéome de *S. cerevisiae*. Ces résultats permettent bien sûr de retrouver des complexes connus, mais aussi de mettre en évidence des interactions nouvelles, que les auteurs ont testé expérimentalement par la technique du TAP-TAG [Rigaut et al., 1999]. Utilisant les nouveaux membres de complexes comme appâts, ils parviennent ainsi à vérifier leur résultats, notamment concernant les complexes protéiques des hélicases, du nucléosome et du complexe de réplication.

De nombreuses études ([Lin et al., 2004], [Yamanishi et al., 2004], [Zhang et al., 2004], [Wu et al., 2006], [Yamanishi et al., 2005]) utilisent les réseaux bayésien, mais aussi d'autres méthodes tels les arbres de décision ou les procédures d'élection, à fin de prédictions d'interactions binaires, de cooccurrence dans des complexes protéiques ou voies métaboliques, ou encore de prédiction de fonctions biologiques. Les résultats de ces études ont pour avantage d'être accompagnés par des mesures statistiques évaluant à la fois la qualité globale des réseaux prédits, mais aussi la qualité des prédictions au niveau des protéines.

2.1.4 Évaluation de la qualité des réseaux d'interactions biomoléculaires

L'évaluation de la qualité des réseaux d'interactions protéine-protéine, que ceux-ci soient issus de l'expérimentation ou de la prédiction, est un point clé pour l'analyse de ces réseaux. En effet, nous avons vu en section 2.1.1 que de nombreuses études utilisent les réseaux d'interactions comme base de départ pour prédire les fonctions biologiques des protéines ([Nabieva et al., 2005], [Deng et al., 2004], [Brun et al., 2003], [Huynen et al., 2003]) ou encore pour rechercher des cibles de médicaments [Strong and Eisenberg, 2007]. La fiabilité de ces études repose

en grande partie sur la qualité intrinsèque des données utilisées, même si certaines méthodes tiennent compte des erreurs inhérentes à certaines méthodes expérimentales [Goldberg and Roth, 2003].

De la fiabilité des méthodes expérimentales Dans leur étude, [Sprinzak et al., 2003] évaluent à 50% le taux d'erreur de la méthode du *Yeast Two-Hybrid*. Pour ce faire, les auteurs ont compilé les interactions protéiques de *S. cerevisiae* provenant des bases de données MIPS [Mewes et al., 2002], DIP [Xenarios et al., 2000] et BIND [Bader et al., 2001], ainsi que des études à grande échelle de [Uetz et al., 2000] et [Ito et al., 2001]. Après élimination des doublons, les auteurs obtiennent une base de 9347 interactions binaires, chacune étiquetée avec la méthode expérimentale utilisée. Sur ces données, Sprinzak *et al.* montrent que le taux de vrais-positifs pour des méthodes biochimiques et immunologiques va de 80% à 100%, alors que ce taux chute à 60% pour le double hybride à petite échelle et à 50% pour le double hybride à grande échelle. La méthode de Sprinzak *et al.* est basée sur l'utilisation de données de localisation cellulaire des protéines et sur les fonctions cellulaires de ces dernières. L'hypothèse sous-jacente est que des protéines en interaction doivent se trouver dans le même compartiment cellulaire, au moins au moment de l'interaction, et qu'il est de plus probable que les deux partenaires de l'interaction soient impliqués dans un même processus cellulaire.

Cette évaluation de la qualité des résultats produits par la méthode du double hybride est en adéquation avec l'évaluation faite par [von Mering et al., 2002]. Dans cette étude, les auteurs tentent d'évaluer la qualité des réseaux fournis par différentes techniques expérimentales, ainsi que par des techniques de prédiction *in silico*. La première observation faite par les auteurs est que parmi les 80 000 interactions binaires disponibles pour *S. cerevisiae* en 2002, seules 2400 sont retrouvées par différentes méthodes. Ce faible recouvrement, que nous avons mentionné en section 2.1.1, traduit probablement le fait que les différentes méthodes ne permettent pas de caractériser les mêmes types d'interactions (e.g., transitoire ou stables, processus biologiques différents). Les auteurs montrent ainsi que les méthodes permettant la caractérisation de complexes protéiques [Gavin et al., 2002] trouvent moins d'interactions entre des protéines impliquées dans les mécanismes de transport, probablement du fait de la difficulté à purifier des complexes attachés à la membrane cellulaire. De même, la méthode du double hybride [Uetz et al., 2000] retrouve peu d'interactions entre protéines impliquées dans le processus de traduction. La seconde observation est que les interactions retrouvées par plus d'une expérience, pas obligatoirement par différentes méthodes expérimentales, ont une valeur de confiance très supérieure aux autres. Ce phénomène est aussi confirmé par [Sprinzak et al., 2003] et [Gerstein et al., 2002].

Dans toutes ces études, comme dans celle de [Goldberg and Roth, 2003], les critères utilisés sont la colocalisation cellulaire et le partage d'annotations fonctionnelles concernant les processus cellulaires. D'autres études, telle celle de

[Deng et al., 2003], se focalisent sur les relation de coexpression des ARN messagers codant pour les protéines en interaction. Si ces études semblent fournir des résultats intéressants, il a cependant été montré que le type de donnée apportant le plus grand enrichissement en information est le partage d'annotations fonctionnelles ([Jansen and Gerstein, 2004], [Lin et al., 2004]).

Quelles solutions envisager ? Nous venons de voir que de nombreuses études se sont penchées sur la question de la qualité des interactions protéines-protéines, que celles-ci soient issues d'expérimentations ou de prédictions. Plus qu'une simple évaluation, les méthodes proposées sont aussi un outil de travail. En effet, chacune de ces méthodes propose une mesure, calculable pour le réseau dans sa globalité ou pour les interactions binaires. Ainsi, il est tout à fait possible d'envisager une méthode permettant d'établir une valeur de confiance seuil, permettant ainsi d'élaguer le réseau d'interactions et de ne retenir que les relations de plus ou moins haute qualité. Cette approche a notamment été développée dans les articles [Bader et al., 2004] et [Bader, 2003]. Dans le premier article, les auteurs utilisent un modèle statistique permettant de mesurer la qualité des interactions protéine-protéine. Dans le second article, [Bader, 2003] proposent un outil appelé SEEDY, utilisant la mesure de qualité précédemment développée, et qui permet de définir des complexes protéiques à partir d'un ensemble de relations d'interactions et d'un ensemble de protéines d'intérêt.

2.1.5 Conclusion

Dans cette partie, nous avons tout d'abord eu un aperçu des principales caractéristiques des réseaux d'interactions protéine-protéine. Bien que des études récentes [Han et al., 2005] semblent nuancer l'opinion générale [Barabasi and Oltvai, 2004], cette dernière veut que les réseaux d'interactions fassent partie de la classe des réseaux petits mondes ou *scale free*. Les caractéristiques principales de ces réseaux sont que la distribution des degrés (i.e., le nombre de connexions) des sommets suit une loi de puissance, avec un faible nombre de noeuds très connectés (les *hubs*) et un grand nombre de noeuds faiblement connectés. Cette topologie procure une certaine robustesse au réseau qui peut supporter un nombre conséquent de délétions de protéines peu connectées sans pour autant provoquer de changements importants dans la structure globale. D'un point de vue évolutif, ces structures seraient en grande partie dues au phénomène de duplication de gènes ([Hughes and Friedman, 2005], [Ispolatov et al., 2005]). À un niveau de détail supérieur, des études ont montré que les *hubs* sont souvent des protéines essentielles pour la cellule [Jeong et al., 2001]. De plus, ces protéines centrales peuvent être séparées en deux sous-classes [Han et al., 2004], selon qu'elles soient statiques, en formant des complexes avec l'ensemble de leurs partenaires en même temps, ou

dynamiques, en formant des interactions avec différents sous-ensembles de partenaires et à différents moments ou endroits.

Nous avons ensuite fait un tour d'horizon des principales sources de données d'interactions protéiques. L'information la plus importante est que les différentes bases de données sont aujourd'hui assez similaires du point de vue de leur contenu, grâce à l'adhésion au consortium IMEx[6]. La création de ce consortium, visant à la synchronisation du contenu des bases de données d'interactions, est la suite logique du développement du format d'échange PSI-MI [Hermjakob et al., 2004a] qui permet une représentation standardisée des interactions biomoléculaires. Les principales différences entre les différents acteurs se font donc au niveau des moyens et outils disponibles pour l'accès aux données. À ce titre, il nous semble que la base de donnée fournissant les outils les plus aboutis est IntAct [Hermjakob et al., 2004b].

Enfin, nous avons abordé la question des méthodes informatiques appliquées aux interactions protéine-protéine. Ces méthodes ont toutes pour but de prédire des interactions directes ou indirectes entre protéines et se différencient majoritairement par leur stratégie. Les premières réalisent des prédictions *de novo* [Jansen et al., 2003], les secondes cherchent à tirer partie des résultats existant en les améliorant, notamment par l'utilisation de mesures de confiance [Sprinzak et al., 2003]. Bien que certaines méthodes utilisent les propriétés structurales des réseaux d'interactions, ou encore certaines hypothèses biologiques comme la composition en domaines, une grande majorité des méthodes se basent sur des modèles statistiques, permettant l'intégration de sources de données diverses et peu corrélées, pour réaliser des prédictions d'interactions. Deux difficultés majeures se retrouvent dans toutes ces études. D'une part, la détermination de jeux de données d'apprentissage et d'autre part, la réunion d'ensembles de données d'interactions hétérogènes, car provenant de sources très diverses (techniques expérimentales différentes, expérimentation ou prédiction, etc). Au chapitre 3, nous proposons des éléments de réponses à ces problèmes.

2.2 Les voies métaboliques

L'étude des voies métaboliques est un domaine en pleine expansion, et ce pour diverses raisons. Du point de vue de l'intérêt biologique, la détermination des voies métaboliques constitue un enjeu central pour la compréhension des mécanismes cellulaires. En effet, les voies métaboliques regroupent des phénomènes aussi divers et essentiels que la production d'énergie nécessaire à la cellule, la synthèse de lipides, impliqués notamment dans la formation des membranes cellulaires, le métabolisme des acides aminés, ainsi que les mécanismes de dégradation et de recyclage de toutes les biomolécules. L'étude des voies métaboliques connaît un essor dû à l'apport des méthodes informatiques, celles-ci étant bien adaptées au traitement

[6] http ://imex.sf.net

des grandes quantités de données issues de l'expérimentation, structurées au sein de bases de données comme celles du KEGG [Ogata et al., 1999] ou encore WIT [Overbeek et al., 2000]. Les voies métaboliques représentent, de plus, des données qui sont intrinsèquement structurées, donc adaptées à l'application de méthodes informatiques.

2.2.1 Définition et caractéristiques

Une voie métabolique est un réseau de réactions biochimiques. Ces réactions mettent en jeu des substrats, qui sont les molécules sources, des produits, qui sont les molécules finales, et enfin dans la plupart des cas des enzymes qui catalysent la réaction. Chaque réaction biochimique peut être vue comme un module qui peut s'articuler avec d'autres modules, tout substrat d'une réaction pouvant être le produit d'une autre. Cette modularité représente théoriquement la source d'une importante variété de voies métaboliques distinctes. Cependant, seule une partie de ces combinaisons relève d'une réalité biologique, les études expérimentales ou *in silico* ayant pour but de découvrir et caractériser les réseaux et sous réseaux qui existent réellement chez un organisme particulier, ou au sein d'un ensemble d'organismes.

Le petit monde des voies métaboliques De nombreuses études ont été menées pour déterminer les caractéristiques des voies métaboliques chez différentes espèces. De manière assez surprenante, on découvre que les voies métaboliques, comme les réseaux d'interactions protéine-protéine (section 2.1.1), font partie d'une classe de graphes appelés graphes petits mondes ou *scale free*. La principale caractéristique d'un graphe petit monde est que la distribution des degrés de connexion des noeuds du graphe suit une loi de puissance. Ainsi, nous trouvons un faible nombre de molécules très connectées, appelées *hubs*, et un grand nombre de molécules faiblement connectées (e.g., figure 7, page 18). Cette configuration du réseau confère aux voies métaboliques une certaine robustesse face à la disparition soudaine de fonctions enzymatiques, au cours de l'évolution. En effet, il est aisément compréhensible que la perte d'un *hub*, suite à une mutation par exemple, peut entraîner la perte d'un grand nombre de réactions biochimiques. Hors, les réseaux *scale free* sont caractérisés par un faible nombre de hub et un grand nombre de sommets faiblement connectés. La probabilité qu'une mutation perturbe fortement le réseau est donc faible, compte-tenu des propriétés structurales de ce dernier. Un autre moyen de caractériser un tel réseau est de calculer la longueur moyenne du plus court chemin qui relie deux sommets quelconques (mesure LM) et le diamètre du graphe, qui représente la longueur du plus long chemin parmi les plus courts chemins entre deux sommets du graphe. Concernant les voies métaboliques, l'appartenance aux réseaux petits mondes a été confirmée par différents auteurs, tels que [Ma and Zeng, 2003] ou [Jeong et al., 2000].

Le monde du vivant sur une seule échelle ? Nous avons vu que les différentes études sur les caractéristiques structurales des voies métaboliques s'accordent sur la nature petit monde de ces réseaux. Cependant, elles ne s'accordent pas sur d'autres caractéristiques. Ainsi, [Jeong et al., 2000] montrent que la mesure LM est constante parmi un ensemble de 43 espèces couvrant les trois domaines du vivant (procaryotes, archae et eucaryotes). De plus la valeur de cette mesure est de 3.2, ce qui se traduit par le fait qu'en moyenne, tout métabolite est transformable en n'importe quel autre en seulement trois étapes. Ce résultat est surprenant à deux égards. D'une part, il est philosophiquement assez étrange que des organismes très simples comme des parasites ou des bactéries montrent la même complexité dans leur processus métaboliques, que des organismes plus complexes tels les eucaryotes supérieurs. D'autre part, les voies métaboliques sont souvent longues, comme la glycolyse ou le cycle TCA, et il semble donc curieux que la mesure LM soit seulement de 3 étapes. Les recherches de [Ma and Zeng, 2003] montrent que ce résultat peut être mis en cause. En effet, l'étude de Jeong *et al.* place tous les métabolites sur un même plan. Or, certains métabolites sont très fréquents dans les réactions biochimiques et sont par exemple en charge des transferts d'électrons (e.g., NADH) ou de groupements fonctionnels (e.g., phosphate, groupe méthyle), ou bien encore de petites molécules (e.g., H2O, CO2).

Ma et Zeng ont dressé une liste de ces composés, qui correspondent à la notion de métabolite externe développée par [Schuster et al., 2002]. Les auteurs de cette étude définissent les métabolites internes comme les métabolites qui sont consommés et produits lors des réactions biochimiques, par opposition aux métabolites externes dont la concentration est supposée inchangée au cours des différentes réactions. En supprimant ces métabolites externe, [Ma and Zeng, 2003] observent une variation importante de la mesure LM. Les auteurs montrent dans leur étude menée sur les 80 espèces regroupées au KEGG, que cette mesure varie de 7.22 pour les bactéries à 8.50 pour les archae et 9.57 pour les eucaryotes. La deuxième mesure relevée par cette étude est le diamètre du réseau métabolique qui traduit l'éclatement maximal d'un réseau. MA *et al.* montrent de même que le diamètre des réseaux semble suivre en moyenne l'évolution de la mesure LM, avec une valeur de 20.6 pour les bactéries, 23.4 pour les archae et 33.1 pour les eucaryotes. Ces différences dans la mesure LM et la valeur du diamètre montrent que, si la structure globale est la même quelque soit le domaine du vivant étudiée, la structure fine différencie les différents règnes. De même, à l'intérieur d'un même domaine du vivant, ces mesures peuvent varier de manière significative en fonction des différentes adaptations des organismes à leur environnement. Ma *et al.* citent à ce propos l'exemple des organismes parasites qui possèdent un réseau métabolique de petite échelle (peu de métabolites) et qui présentent des mesures LM et des diamètres très inférieurs aux moyennes, ce qui peut s'expliquer par l'adaptation à leur niche écologique, à savoir les cellules de leur hôte [Podani et al., 2001]. En tirant partie de l'équipement biochimique de leur hôte, les parasites auraient donc évolué vers une réduction d'échelle de leur métabolisme, tout en conservant les propriétés de robustesse d'un réseau petit monde.

2.2.2 Sources de données

Les centres de ressources sur les voies métaboliques sont nombreux et variés, tant dans leur type de contenu que dans la manière d'accéder à ces contenus. Les bases de données comme BRENDA [Schomburg et al., 2002], EMP [Selkov et al., 1996] ou encore ExPASy-ENZYME [Bairoch, 2000] se focalisent sur la création d'un répertoire enzymatique. D'autres, telle ChemFinder[7] sont orientées vers la création d'un catalogue de composés impliqués dans les réactions biochimiques. Enfin, d'autres bases de données telles KEGG [Ogata et al., 1999], WIT [Overbeek et al., 2000] ou MetaCyc [Caspi et al., 2006], effectuent un travail de synthèse et proposent des cartes de voies métaboliques, associées à des informations sur les enzymes, les substrats et les produits, ainsi que des données spécifiques pour certaines espèces.

Dans le cadre de nos recherches, nous nous intéressons surtout à ce dernier type de ressources. En effet, dans le cadre de la prédiction de voies métaboliques basée sur la disponibilité d'informations génomiques, nous ne pouvons exploiter les ressources ne fournissant que des informations génériques sur les voies métaboliques. Aussi, nous nous attardons ici sur la description et la comparaison de ces différentes sources, plus particulièrement le KEGG et les voies métaboliques de SGD.

Kyoto Encyclopedia of Genes and Genomes La base de données du KEGG est une source de données de grande qualité pour les micro-organismes tels que les levures. Cette ressource contient l'ensemble des voies métaboliques connues, ainsi qu'un nombre croissant de voies de régulation et de transport. Les voies métaboliques du KEGG sont manuellement générées à partir d'un ensemble de réactions issues des deux ouvrages de référence que sont "Biological Pathways" [Gerhard, 1992] et "Metabolic Maps" [Nishizuka, 1980]. Pour chaque espèce, KEGG maintient un catalogue de gènes, associés notamment à des numéros enzymatiques EC [Hoffmann-Ostenhof and Thompson, 1958].

Les données peuvent être consultées de différentes manières. KEGG fournit une interface de requête complète (DBGET) permettant d'interroger les données disponibles, tant au niveau des voies métaboliques que des gènes et des métabolites, et qui fournit de nombreux liens vers des données externes. Les différentes ressources sont aussi directement accessibles. Ainsi, pour les voies métaboliques, une page présente la liste des noms des différentes voies enregistrées, chacune associée à un lien hypertexte menant au dessin de la voie générique. Depuis cette page, il est possible d'afficher le dessin de la voie spécifique à une espèce, et chaque composant présent possède un lien hypertexte vers une page détaillée.

Les voies métaboliques du KEGG sont caractérisées par le fait qu'elles sont issues de la littérature biochimique. Cependant, elles ne présentent pas l'ensemble des

[7] http ://chemfinder.com

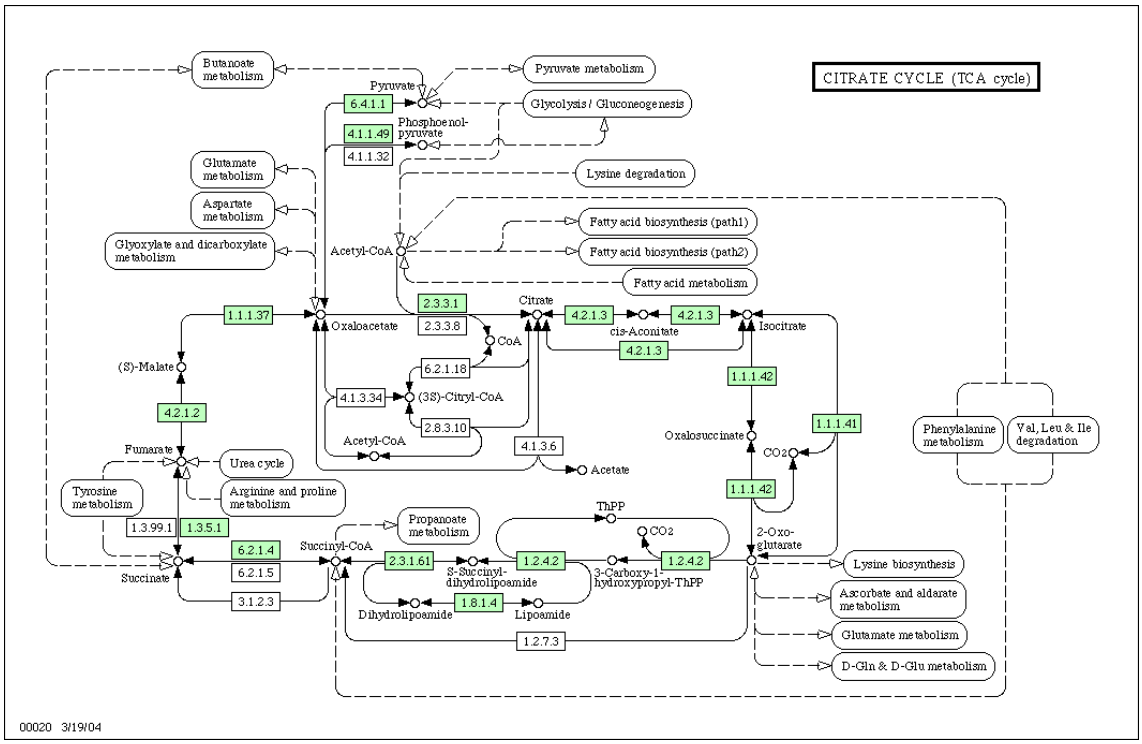


FIG. 12 – Carte du cycle TCA chez *S. cerevisiae* (données KEGG). Les rectangles verts correspondent à des enzymes ayant au moins un homologue identifié chez *S. cerevisiae*. Les flèches en pointillés indiquent les interconnexions avec d'autres voies métaboliques. Nous pouvons observer que les métabolites énergétiques et cofacteurs ne sont pas représentés.

réactions et des réactifs décrits par les équations biochimiques. En effet, les voies métaboliques du KEGG ne tiennent compte, majoritairement, que des réactions impliquant des enzymes identifiées (e.g., le cycle TCA figure 12). Dans le cadre de nos recherches, ce fait nous intéresse particulièrement puisque nous disposons de données génomiques que nous souhaitons utiliser, notamment les informations de conservation des gènes, incidemment des enzymes issues de l'expression de ces gènes. Une autre caractéristique est que les métabolites jouant un rôle de cofacteur ou d'apport d'énergie ne sont pas systématiquement représentés dans les voies métaboliques, ce qui peut être intéressant dans le cadre d'analyses structurales (cf., section 2.2.1).

Les voies métaboliques espèce-spécifiques sont automatiquement générées à partir d'un coloriage des voies métaboliques génériques (figure 13). Une réaction enzymatique est conservée s'il existe au moins un gène étiqueté avec le numéro EC correspondant à celui de l'enzyme générique. Les voies métaboliques spécifiques à une espèce sont donc le résultat d'une extrapolation, en prenant comme réseau de référence le réseau enzymatique générique et en faisant correspondre les numéros EC de l'enzyme générique et de l'enzyme espèce-spécifique.

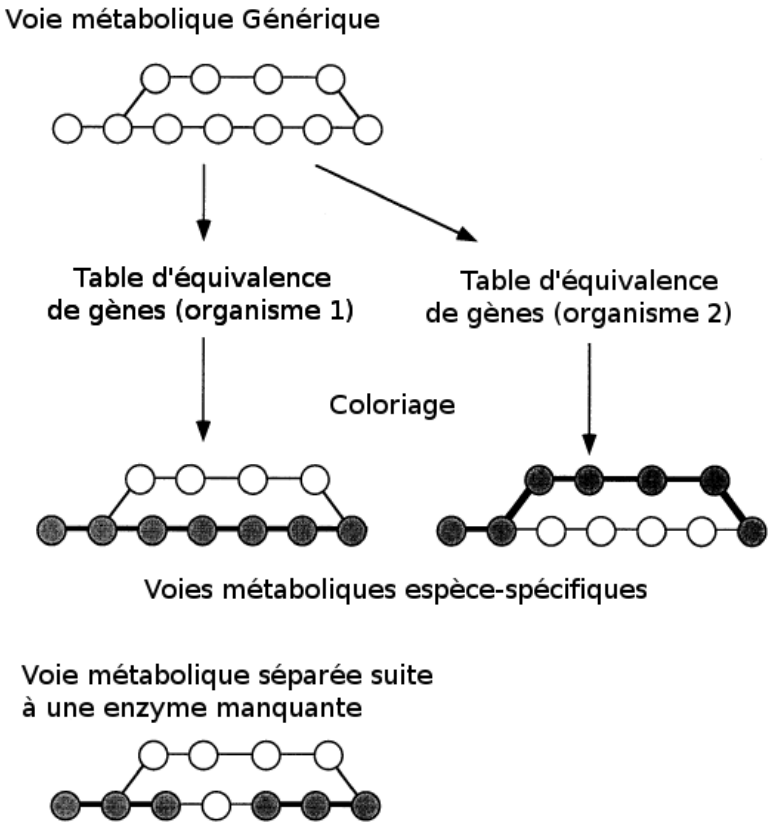


Fig. 13 – Schéma du principe de génération des voies métaboliques espèce-spécifiques sur les données KEGG. La voie métabolique générique est coloriée en fonction des données génomiques disponibles pour l'organisme : si un équivalent fonctionnel existe pour une enzyme, celle-ci est marquée. Le résultat final est un réseau composé d'enzymes marquées, définissant la voie métabolique spécifique à cet organisme. Les enzymes manquantes peuvent éventuellement séparer la voie métabolique générique en plusieurs voies pour un organisme. (d'après [Ogata et al., 1998])

Les données fournies par le KEGG sont généralement considérées comme de grande qualité. Différentes études s'attachent à comparer tant la qualité et la quantité de contenu, que la manière dont ce dernier est accessible. Dans l'article [Wittig and De Beuckelaer, 2001], six bases de données de voies métaboliques sont comparées, dont celle du KEGG. Parmi les observations faites sur la qualité de cette dernière, les auteurs notent qu'elle contient des incohérences, notamment au niveau des métabolites. Certains métabolites sont désignés par plusieurs identifiants dans la base de données KEGG LIGAND, et à l'inverse, certains identifiants désignent plusieurs métabolites différents. De plus, les données brutes et les résumés qui en ont été tirés sont parfois divergents. Ainsi, la liste des substrats désignés pour une enzyme particulière ne correspond pas toujours à la liste que l'on peut construire manuellement à partir des différentes équations de réactions biochimiques. Ceci peut être problématique lorsque ces données sont traitées automatiquement par des méthodes informatiques.

La fiabilité de la détection des gènes homologues au sein des espèces présentes dans les données du KEGG peut également être mise en question. Le KEGG a développé un outil nommé GFIT [Bono et al., 1998], pour Gene Function Identification Tool, qui permet d'identifier les gènes codant pour des enzymes et de leur assigner un numéro EC. GFIT recherche des gènes orthologues dans un organisme d'intérêt à partir d'un ensemble de gènes déjà annotés, la notion d'orthologie du KEGG regroupant à la fois les notions d'orthologues et de paralogues, telles que ces notions sont habituellement interprétées (voir section 1.3.1). L'identification des orthologues se fait par une recherche du meilleur alignement réciproque (RBH [Rivera et al., 1998]) entre la séquence du gène requête traduite en acide aminée, et la banque de données de séquence du KEGG, en utilisant le logiciel FASTA [Pearson and Lipman, 1988]. Une fois cet orthologue trouvé, son numéro EC est reporté sur la nouvelle séquence. Le KEGG construit ainsi un catalogue de gènes orthologues. S'il manque des enzymes, qui n'ont donc pas été identifiées lors de la recherche d'orthologues, une deuxième passe est réalisée par le logiciel GFIT en utilisant des critères moins stringents. En dernier ressort, le logiciel PATHCOMP est utilisé pour rechercher des réactions enzymatiques alternatives [Goto et al., 1997], permettant d'assurer la continuité de la voie métabolique.

Nous pouvons observer que cette procédure est confrontée à un problème majeur qui est l'existence de gènes paralogues. En effet, de nombreux gènes se sont dupliqués dans les organismes au cours de l'évolution, subissant parfois des modifications très différentes. Lors de la procédure d'identification du RBH, il est possible que l'orthologue identifié soit en fait un paralogue. Ainsi, le numéro EC reporté sur la nouvelle séquence peut ne pas être le bon et l'erreur peut se propager lors des cycles d'identifications suivants. Pour tenter de remédier à ce problème, le catalogue de gènes fait l'objet d'une annotation manuelle, vérifiant à la fois la qualité des alignements et la qualité des voies métaboliques générées suite à l'identification des

homologues. L'hypothèse qui guide cette vérification est que si l'identification d'orthologues est de bonne qualité, alors les voies métaboliques prédites seront le plus souvent complètes. On peut noter que cette hypothèse est un argument circulaire, étant donné que l'on pose comme prémisse que l'identification est correcte si les voies sont complètes, puis que l'on observe la complétude de ces mêmes voies pour décider de la qualité de l'identification, sans avoir jamais prouvé la prémisse. En suivant ce raisonnement, il est possible d'obtenir une mauvaise identification d'orthologues, donnant une forte proportion de faux-positifs, tout en conservant une bonne couverture des voies métaboliques. Cependant, cette absence de preuve *a priori* peut être compensée *a posteriori* par le fait qu'un effort d'annotation manuelle est mené pour vérifier les données d'orthologie, ainsi que par le fait que les voies métaboliques concernent majoritairement des phénomènes bien conservés au sein du vivant.

De ces différentes observations sur la qualité des données fournies par le KEGG, nous pouvons tirer plusieurs conclusions. Tout d'abord, les voies métaboliques génériques sont de bonne qualité, issues de la littérature et manuellement vérifiées. Même si certaines incohérences peuvent apparaître, la grande majorité des informations est correcte. Ensuite, le codage de ces voies, ne conservant que les métabolites les plus importants en omettant par exemple les métabolites énergétiques dans les voies où il ne sont pas les produits finaux, permet de ne pas biaiser les analyses structurales et topologiques (cf. section 2.2.1). Cependant, la génération de voies métaboliques espèce-spécifiques doit être utilisée en connaissance de cause et avec précaution. Dans le cadre de nos travaux de recherche, nous considérons que les données provenant du KEGG sont d'une qualité suffisante pour être utilisées.

Saccharomyces Genome Database est un centre de ressources s'intéressant à la levure *S. cerevisiae*. SGD propose notamment une base de données de voies métaboliques générées grâce au logiciel "Pathway tools", développé par Peter Karp [Karp et al., 2002] au SRI[8]. Cet ensemble d'outils permet de générer des voies métaboliques spécifiques à une espèce en prenant comme référence la base de données MetaCyc et un catalogue de gènes annotés. MetaCyc est une base de données de voies métaboliques expérimentalement décrites et manuellement annotées par rapport à la littérature. Limitée au départ à la bactérie *Escherichia coli*, MetaCyc propose aujourd'hui des voies métaboliques pour environ 900 espèces. Ce centre de ressource fournit aussi un ensemble de voies métaboliques de référence pouvant servir à prédire des voies espèces spécifiques, comme le fait SGD pour *S. cerevisiae*. Ainsi, en comparant les annotations de gènes et les voies métaboliques de référence, un premier ensemble de voies est généré. Celles-ci sont ensuite annotées et corrigées manuellement en fonction des données publiées sur *S. cerevisiae*. De plus, des voies spécifiques, non disponibles dans les données de référence, sont ajoutées lorsqu'un support expérimental est disponible.

[8]http ://www.sri.com/

Les données de voies métaboliques fournies par SGD sont accessibles par différents moyens. Une première interface permet d'effectuer des recherches à partir de noms de gènes, de numéros EC, de termes d'ontologie ou directement de la liste des voies métaboliques disponibles. Pour chaque voie métabolique, une image cliquable représentant la voie métabolique s'affiche. La vue par défaut montre uniquement l'enchaînement de réactions entre les différents métabolites, mais il est également possible d'afficher les enzymes mises en jeu ou la formule chimique des composés (e.g., cycle TCA figure 14). Pour chaque élément de l'image, des liens hypertexte sont disponibles et donnent de plus amples informations. Enfin, un outil permet de faire des comparaisons de voies métaboliques entre *E. coli* et *S. cerevisiae*, en présentant le nombre de voies métaboliques pour chaque organisme dans chaque grande classe de voie métabolique (dégradation, biosynthèse, etc). Un dernier logiciel, appelé *Expression viewer*, permet d'effectuer une mise en correspondance de données d'expression, de métabolomique ou de protéomique, avec les voies métaboliques.

De même que pour les données du KEGG, les voies métaboliques fournies par SGD sont issues d'un processus d'extrapolation. Cependant, un processus d'annotation manuelle complète s'ajoute à l'étape de génération automatique et seules les réactions confirmées par la littérature sont conservées. De ce fait, les voies métaboliques fournies par SGD peuvent être considérées comme des données de haute qualité. La contrepartie de cette haute qualité est que la couverture en nombre de gènes impliqués est plus faible que pour KEGG, à savoir 775 gènes impliqués dans les voies métaboliques de SGD, soit environ un tiers de moins que pour KEGG. Une autre caractéristique de ces voies métaboliques est qu'elles sont très découpées. Pour une couverture assez similaire en terme de classes de voies métaboliques (dégradation, biosynthèse des acides aminés, métabolisme des sucres, etc), KEGG propose 87 voies métaboliques là où SGD en propose 154, soit environ le double de KEGG. Une dernière caractéristique opposant SGD et KEGG est que les voies métaboliques fournies par SGD tiennent compte de tous les métabolites présents dans les équations des réactions biochimiques, que ce soit l'eau, l'oxygène ou encore des molécules énergétiques telles le NADH. De ce fait, il est possible que les analyses informatiques sur la structure des voies métaboliques soient perturbées par ces informations additionnelles.

Différentes sources pour différentes observations Après étude des sources de données que nous désirons utiliser et qui sont parmi les plus utilisées, nous pouvons tirer quelques enseignements. Tout d'abord, d'un point de vue de la pertinence biologique et de la fiabilité des données, il semble que les données de SGD soient les plus appropriées. Ces données ont cependant les inconvénients de leurs avantages. Pour être plus fiables, les données ont une couverture moindre en terme de nombre de gènes impliqués, mais une couverture similaire en terme de classes de voies métaboliques. La deuxième différence majeure entre SGD et KEGG concerne la

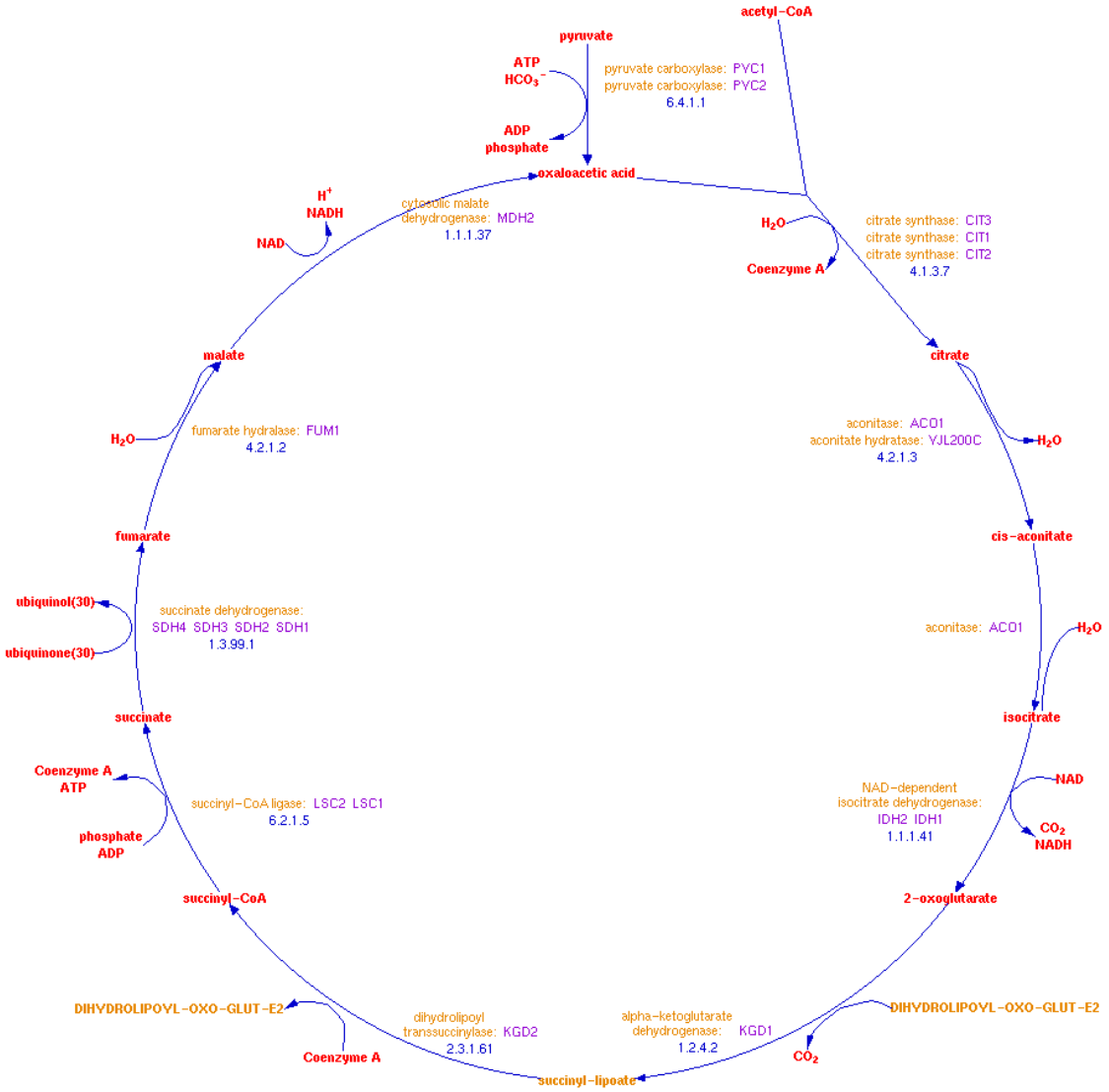


Fig. 14 – Carte du cycle TCA chez *S. cerevisiae* (données SGD). Les noms des enzymes identifiées chez *S. cerevisiae* sont inscrit en violet, à côté du nom de la fonction enzymatique. Nous pouvons observer que les métabolites énergétiques et cofacteurs sont représentés.

prise en compte des métabolites que l'on pourrait dire "annexes", tels que l'oxygène, l'ATP, le NADH, etc. En effet, ces composés sont très répandus dans les réactions biochimiques et de ce fait interagissent avec de nombreuses enzymes, altérant ainsi de manière artificielle la structure sous-jacente du réseau représentant les voies métaboliques. Aussi, dans le cadre d'analyses informatiques de la structure de ce réseau, la présence ou l'absence de ces composés doit influer sur les résultats. Il sera donc intéressant de prendre en compte ce paramètre.

2.2.3 Analyse et prédiction de voies métaboliques

L'analyse et la prédiction de voies métaboliques est un domaine de recherche en pleine expansion ([Horne et al., 2004], [Yamanishi et al., 2005]). Les voies métaboliques font l'objet de deux types de modélisation qu'il faut bien distinguer [Deville et al., 2003]. La modélisation informatique des voies métaboliques est une approche théorique, fondée sur les mathématiques, et qui a pour but la simulation de la dynamique cellulaire afin d'en expliquer le fonctionnement. Cette modélisation s'intéresse à la fois aux aspects qualitatifs et quantitatifs du métabolisme. Les modèles de données visent quant à eux à établir la façon la plus adaptée de représenter des données du métabolisme en fonction du problème étudié. Ce type de modélisation est plus particulièrement axé sur une étude qualitative des réseaux métaboliques, en étudiant notamment les caractéristiques structurales des réseaux.

Encodage des voies métaboliques

Le graphe de composé est un modèle très utilisé ([Fell and Wagner, 2000], [Wagner and Fell, 2001], [Ma and Zeng, 2003]). Dans ce type de graphe, les sommets sont les composés et les arêtes représentent les réactions de transformation d'un composé en un autre (figure 17(b), page 48). Ces graphes peuvent être dirigés ou non. Dans un graphe non dirigé, deux composés sont reliés si ils apparaissent dans une même équation de réaction biochimique. Dans la version dirigée, un arc va de la molécule 1 à la molécule 2 si la première molécule est un substrat et la deuxième un produit, selon l'équation de la réaction. Ce type de modèle de données présente l'avantage d'être simple à construire, notamment dans sa version non dirigée, et de nombreuses méthodes d'analyse informatique peuvent lui être appliquées. Dans l'article [Fell and Wagner, 2000], les auteurs encodent les voies métaboliques principales de *E. coli* dans un graphe de composés non dirigé, équivalent à une matrice d'adjacence. Ce codage permet d'analyser plusieurs types d'informations, telles que la distribution des degrés de connexion des sommets ou la longueur moyenne des chemins entre deux composés. Cette étude leur a permis d'observer que les réseaux métaboliques, comme de nombreux autres réseaux, peuvent être associés à la classe des réseaux petits mondes (*scale-free*) et que la distribution des degrés des sommets suit une loi de puissance. Une version dirigée de ce même graphe

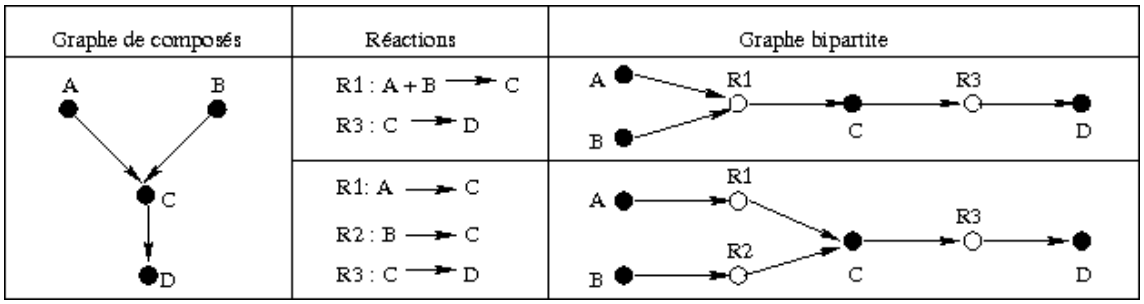


Fig. 15 – Exemple d'ambiguïté avec le graphe de composés. Nous voyons dans cet exemple que deux ensembles de réactions distincts peuvent être représentés par un seul graphe de composé, ce qui introduit une ambiguïté dans l'interprétation : il est impossible de savoir si C est obtenu par la réaction A+B ou par transformation de A ou de B. Le graphe bipartite lève cette ambiguïté (d'après [Deville et al., 2003])

[Ma and Zeng, 2003] donne des résultats similaires, à savoir que la distribution des degrés moyens sortants et entrants suit une loi de puissance, donc que les réseaux métaboliques appartiennent à la classe des réseaux petits mondes. Bien que cet encodage soit adapté à ce type d'étude, il ne permet cependant pas de traiter toutes les questions. Comme le remarquent [Deville et al., 2003], ce modèle de données couvre une faible partie des données disponibles dans les voies métaboliques, puisque les enzymes n'apparaissent pas. De plus, le pouvoir descriptif des graphes de composés est relativement faible. En effet, dans un graphe de composés, qu'il soit dirigé ou non, il est impossible de distinguer les ensembles de substrats et produits impliqués dans une seule réaction, la structure de la réaction biochimique étant alors perdue (figure 15).

Le graphe de réactions Pour pallier ces défauts, certains auteurs ont tenté d'utiliser un graphe de réactions, qui est la forme duale du graphe de composés. Dans ce type de graphe, les sommets du graphes sont les réactions biochimiques ou plus simplement, les enzymes catalysant ces réactions (figure 17(c)). Une arête relie deux sommets si ceux-ci partagent un métabolite, qu'il soit un substrat ou un produit. Les graphes de réactions ont été utilisés par [Wagner and Fell, 2001], en parallèle des graphes de composés, et par [Ogata et al., 2000]. Dans ce dernier article, les auteurs comparent le graphe de réactions du métabolisme de *E. coli* à un graphe de son génome, où les noeuds sont les gènes et les arêtes relient les gènes adjacents sur le génome. Cette comparaison permet aux auteurs de découvrir des clusters de gènes correspondant à des clusters d'enzymes dans les voies métaboliques, donc une variante des opérons. Comme pour les graphes de composés, la couverture d'information des graphes de réaction est assez faible, puisqu'ici ce sont les composés qui ne sont pas représentés. Il est donc difficile de savoir, dans une suite de réactions, ce qui est produit et ce qui est consommé. Comme le montrent [Deville et al., 2003], il est possible d'obtenir le même graphe de réaction pour deux ensembles de réactions

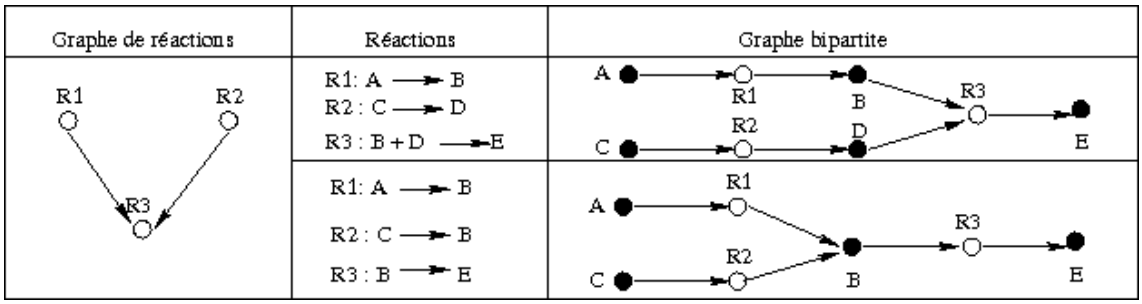


FIG. 16 – Exemple d'ambiguïté avec le graphe de réactions. Nous voyons dans cet exemple que deux ensembles de réactions distincts peuvent être représentés par un seul graphe de réaction, ce qui introduit une ambiguïté dans l'interprétation : la notion de substrat et produit est totalement perdue dans ce graphe. Le graphe bipartite lève cette ambiguïté (d'après [Deville et al., 2003])

différents (figure 16). Cependant, comme pour le graphe de composés, le graphe de réactions peut être suffisant pour l'étude de certaines propriétés structurales ou topologiques des voies métaboliques, de même que pour certaines études de comparaison de réseaux. Pour mener d'autres types d'études, telles que des synthèses ou prédictions de voies métaboliques, il est nécessaire d'utiliser d'autres types de codage des réseaux.

Le graphe bipartite est un type de graphe dans lequel on peut distinguer deux classes de sommets. Concernant les voies métaboliques, la première classe contient les enzymes alors que la seconde regroupe les métabolites (figure 17(d)). Il ne peut y avoir d'arête qu'entre des sommets d'ensembles différents. Ici, une enzyme ne pourra être reliée qu'à un métabolite et *vice versa*. Les graphes bipartites peuvent aussi être utilisés en version dirigée ou non dirigée. Dans le cadre de la version dirigée, un arc part d'un métabolite et rejoint une enzyme si le métabolite est un substrat, et part d'une enzyme pour rejoindre un métabolite si ce dernier est un produit. Dans [Jeong et al., 2000], les auteurs utilisent un graphe bipartite pour encoder les voies du métabolisme central de 43 organismes et comparent les propriétés structurales et topologiques des différents réseaux. De manière assez étonnante, ces réseaux semblent conserver les mêmes propriétés structurales, caractéristiques des réseaux *scale free*, et ce quelque soit la branche étudiée dans l'arbre phylogénétique du vivant. D'autres auteurs [Krishnamurthy et al., 2003] utilisent des hypergraphes (figure 17(e)) pour modéliser les voies métaboliques et proposent à la fois un système de stockage, de modélisation et d'analyse des voies métaboliques. Les hypergraphes présentent les mêmes avantages en terme de codage des voies métaboliques que les graphes bipartites, mais leur visualisation est peut-être un peu moins intuitive. Les hypergraphes sont, de plus, facilement transformables en graphes bipartites et réciproquement. Un autre avantage des graphes bipartites est de pouvoir encoder d'autres réseaux que des voies métaboliques, notamment des voies de signalisation

[Fukuda and Takagi, 2001] ou des processus cellulaires [Demir et al., 2002], avec cependant la nécessité d'une extension du modèle. Les graphes bipartites n'assurent pas une couverture totale de l'information des voies métaboliques, puisque les relations de régulation ne peuvent y être codées. Ils restent cependant adaptés à un large éventail d'études, allant de l'étude topologique et structurale à la synthèse et à la prédiction de voies métaboliques.

Le modèle objet Le but de ce type de modèle est de pouvoir représenter la totalité de l'information associée aux voies métaboliques (figure 17(f)). Dans un modèle objet, une voie métabolique est un objet lui-même composé d'autres objets (e.g., les réactions biochimiques) et caractérisé par des propriétés. De même, chaque réaction contient d'autres objets, tels les composés et enzymes mis en jeu, ainsi que différentes propriétés caractérisant la réaction. Ces modèles sont beaucoup plus complexes à construire et à manipuler que ceux basés sur les graphes, mais leur couverture en information est en revanche excellente. De plus, les modèles objets permettent théoriquement de représenter à la fois les voies métaboliques, les voies de signalisation, de régulation, ainsi que les processus cellulaires. Ces modèles objets sont utilisés dans de nombreux projets dédiés à la représentation du métabolisme, comme BioPax [Luciano, 2005] ou MetaCyc [Caspi et al., 2006], ou dédiés à d'autres sources de données telles les interactions protéiques, avec le projet IntAct [Hermjakob et al., 2004a].

Conclusion Cette étude des différents modèles de représentation de voies métaboliques nous permet de tirer plusieurs conclusions. La conclusion la plus importante est qu'il n'existe pas de modèle à la fois simple et assurant une couverture totale des données. Le corollaire de cette conclusion est qu'il est nécessaire de définir le modèle le mieux adapté aux besoins avant de commencer toute étude portant sur les voies métaboliques. Des études ponctuelles sur la structure ou la topologie d'une voie ou d'un ensemble de voies métaboliques peuvent facilement être réalisées sur des graphes de composés ou de réactions. Des études plus poussées sur la structure des réseaux, ou bien la synthèse ou la prédiction de voies métaboliques nécessiteront probablement l'utilisation de graphes bipartites ou d'hypergraphes. Enfin, des études fines sur les relations de régulation ou sur des relations plus complexes que les réactions enzyme-substrat demanderont le développement de modèles objets.

Méthodes d'étude des voies métaboliques

Nous avons vu dans les sections précédentes que l'étude des voies métaboliques est un domaine en plein essor. De manière générale, nous pouvons distinguer deux types d'études, en fonction de leur vision des voies métaboliques. Certaines études délèguent tout [Podani et al., 2001] ou partie ([Ma and Zeng, 2003],

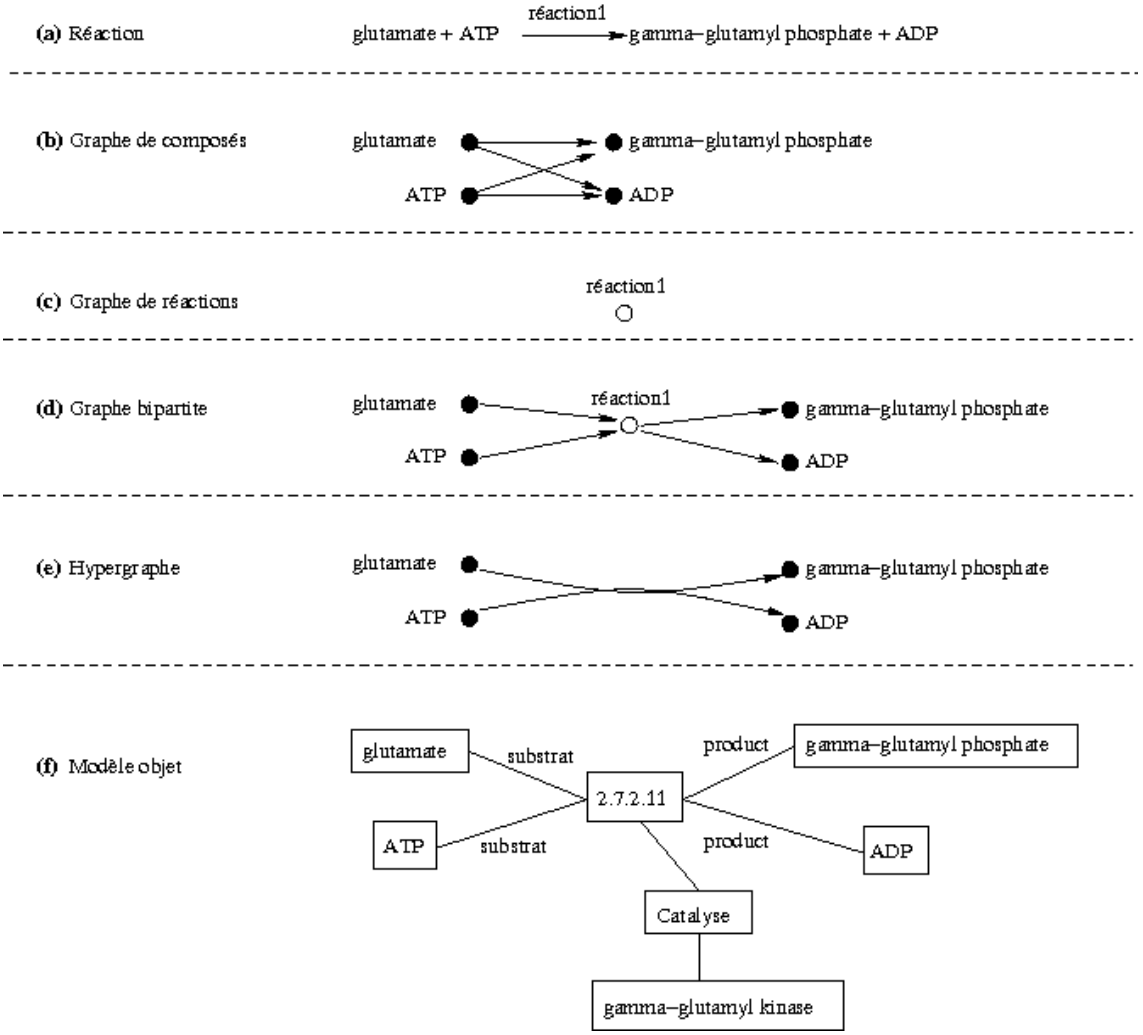


FIG. 17 – Modèles d'encodage des voies métaboliques. Ce schéma représente les différents graphes obtenus pour l'encodage de la réaction 1 (a), selon que le modèle choisi soit le graphe de composés (b), le graphe de réactions (c), le graphe bipartite (d), l'hypergraphe (e) ou le modèle objet (f). (d'après [Deville et al., 2003])

[Horne et al., 2004]) de la construction des voies métaboliques à des entités externes, et utilisent donc comme base de travail les cartes métaboliques établies par les bases de données spécialisées (e.g., KEGG, WIT, MetaCyc). Ces études concernent généralement l'analyse des propriétés structurales et topologiques du réseau métabolique, ainsi que la comparaison des réseaux métaboliques de différentes espèces. D'autres études ([Wagner and Fell, 2001], [Schuster et al., 2002], [Yamanishi et al., 2005]) prennent comme point de départ les matrices stoechiométriques et ont pour but de développer des méthodes de synthèse de voies métaboliques. Ces études ont souvent un but de caractérisation fonctionnelle, voire d'ingénierie cellulaire. En effet, les modèles utilisés sont supposés permettre de simuler le comportement des voies métaboliques face à l'inactivation d'une ou plusieurs enzymes (e.g., par un médicament), ou bien encore d'étudier le meilleur moyen de produire un métabolites à partir de certains métabolites de départ (e.g., les usines cellulaires).

Analyse de voies métaboliques Comme nous l'avons vu à la section 2.2.1, les analyses de voies métaboliques consistent en la description des propriétés structurales et topologiques de ces réseaux, ainsi qu'en la comparaison inter-espèce, révélant de cette manière une partie de l'histoire évolutive des organismes. Les études de [Jeong et al., 2000], confirmées par [Ma and Zeng, 2003], ont montré que le réseau métabolique d'une espèce fait partie de la classe des réseaux petits mondes, présentant la particularité d'avoir un faible nombre de sommets très connectés et un grand nombre de sommets peu connectés, donnant ainsi une certaine robustesse au réseau face à la perte d'un sommet. Robustesse d'autant plus importante que les chemins alternatifs sont plus nombreux du fait de la présence de molécules centrales fortement connectées, les *hubs*. Les résultats de Ma *et al.* montrent aussi que la longueur moyenne du plus court chemin entre deux sommets quelconques est assez caractéristique du domaine du vivant étudié (bactéries, archae, eucaryotes). Ces résultats semblent aussi indiquer que les eucaryotes sont plus proches des archae concernant ces caractéristiques topologiques. Cette dernière hypothèse semble confirmée par [Podani et al., 2001].

L'histoire évolutive du vivant a été étudiée du point de vue des séquences d'ARN ribosomaux [Woese et al., 1990], ainsi que des conservations de gènes [Snel et al., 1999], aboutissant à une meilleure définition des domaines du vivant. D'autres études, telle celle de Podani *et al.*, tentent de comparer les espèces en étudiant des organisations de plus haut niveau, telles les voies métaboliques. Dans cette étude, les auteurs comparent les voies métaboliques de 43 espèces appartenant aux domaines des archae, des bactéries et des eucaryotes. Les données utilisées sont fournies par la base de données WIT [Overbeek et al., 2000] et sont représentées en utilisant le modèle de données développé dans un article précédent [Jeong et al., 2000], qui correspond au graphe de composés. Podani *et al.* convertissent leurs graphes de composés en matrices et en extraient deux principaux

jeux de données, l'un concernant l'organisation du système métabolique, l'autre concernant les réaction biochimiques. Deux sous-ensembles sont générés pour chacun de ces jeux de données, correspondant à la traduction des relations entre métabolites d'une part, et aux relations entre enzymes d'autre part. Pour chaque espèce, les auteurs disposent donc de quatre jeux de données distincts. Une analyse statistique multivariée est effectuée pour toutes les espèces, sur chaque classe de jeu de données. Les différentes approches utilisées sont le *Neighbor Joining (NJ)* [Saitou and Nei, 1987], le *unweighted group average clustering (UPGMA)* [Sneath and Sokal, 1973], le *ordinal clustering (OC)* [Agresti, 1999] et le *nonmetric multidimensional scaling (NMDS)* [Cox et al., 2003]. Les résultats de ces analyses semblent confirmer les résultats de [Ma and Zeng, 2003], en ce qu'ils montrent que pour les différents jeux de données, tant du point de vue des métabolites que des enzymes, les bactéries sont clairement séparées des archae et eucaryotes, alors que ces derniers sont inséparables. Ces résultats sont aussi en accord avec les résultats de cladistique classique ([Woese et al., 1990], [Snel et al., 1999]).

Synthèse de voies métaboliques La synthèse de voies métabolique se place dans une perspective d'étude fonctionnelle. En effet, les travaux que nous avons précédemment présentés s'intéressent à la caractérisation des réseaux métaboliques, à la succession de réactions biochimiques, alors que les travaux basés sur la synthèse de voies métaboliques se focalisent les aspects fonctionnels et quantitatifs, tels que la recherche de production optimale de métabolites, l'étude des différentes voies aboutissant à la synthèse des mêmes composés, etc. Ce type d'étude requiert un fondement mathématique clairement défini, permettant des études comparatives, des prédictions et simulations. Ces approches ont émergées au début des années 1990 ([Mavrovouniotis and Stephanopoulos, 1990], [Schuster and Hilgetag, 1994]), en réponse à cette demande de caractérisation fonctionnelle. Comme le notent [Papin et al., 2003], les applications des méthodes de synthèse de voies métaboliques sont nombreuses. Nous pouvons notamment citer la création d'organismes modifiés par ingénierie cellulaire, en vue d'améliorer le rendement de production de certains métabolites, la génération et le test d'hypothèses sur la structure et la fonction de réseaux métaboliques, ou encore l'étude de certaines propriétés fonctionnelles des réseaux, telles que la robustesse ou l'adaptabilité au milieu. Avec l'avènement de la génomique et de la *systems biology*, la synthèse de voies métaboliques devient un domaine important pour la compréhension des systèmes biologiques.

Différentes approches ont été développées dans le cadre de la synthèse de voies métaboliques, certaines étant basées sur la théorie des graphes (e.g., réseaux de petri [Oliveira et al., 2001]) ou sur des méthodes heuristiques [Mavrovouniotis and Stephanopoulos, 1990], qui tentent de synthétiser des voies métaboliques en partant d'un ensemble d'enzymes, de métabolites et de règles d'association. Les méthodes de synthèse de voies métaboliques que nous allons

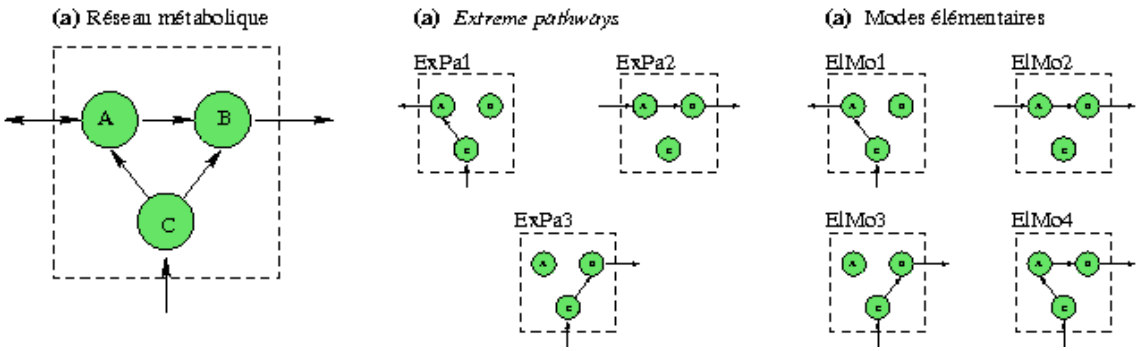


Fig. 18 – Modes élémentaires et *extreme pathways*. Cette figure montre la décomposition d'un réseau métabolique simple en trois *extreme pathways* ou quatre modes élémentaires. Les voies générées sont uniques, non-décomposables et indépendantes. Ce dessin ne représente que des réactions simples et ne prend pas en compte les cofacteurs éventuels, qui sont deux phénomènes augmentant très fortement la complexité des analyses. (source [Papin et al., 2003])

présenter dans la suite sont basées sur les flux métaboliques, déduits de la topologies du réseau métabolique global, sans découpage *a priori*. La méthode des modes élémentaires (*elementary modes* [Schuster and Hilgetag, 1994]) et celle des *extreme pathways* [Schilling et al., 1999], permettent de définir un ensemble de voies métaboliques uniques par l'étude de ces flux (figure 18). Ces deux méthodes partagent la même stratégie globale, avec comme point de départ un réseau métabolique existant, duquel est dérivée une matrice stoechiométrique servant de jeu de données pour l'analyse par les algorithmes. La matrice stoechiométrique met en relation les réactions biochimiques avec les différents métabolites qu'elles impliquent. Le résultat de l'analyse est un ensemble de voies métaboliques non redondantes, qui couvrent l'ensemble des flux possibles définis par la matrice. D'un point de vue mathématique, les voies métaboliques ainsi définies sont des vecteurs de valeurs dans un espace à n dimensions où n est le nombre de réactions biochimiques existant dans le système. Cet espace de haute dimension défini l'ensemble des flux possibles, compte-tenu des réactions du système. Cependant, tous les flux ne représentent pas des réactions réelles du fait des stoechiométries particulières. Ainsi, seule une sous-partie de cet espace défini des flux réels que l'on peut visualiser sous la forme d'un cône dans l'espace à n dimensions dont les bords représentent les *extreme pathways*.

Ces voies métaboliques décrivent la conversion de substrats en produits, selon des réactions qui obéissent à la loi de conservation de masse, en considérant que les cofacteurs sont à l'équilibre. Ainsi, contrairement à la vision classique, la synthèse de voies métaboliques tient compte de la disponibilité des substrats et de la consommation des produits. La combinaison de modes élémentaires ou d'*extreme pathways* conduit à la formation de voies métaboliques cohérentes avec la réalité des lois de la biochimie. Cette caractéristique permet d'envisager des applications intéressantes pour

l'étude d'organismes, puisque l'on peut ainsi étudier les flux métaboliques en fonction des substrats disponibles, ou bien définir les différents enchaînements permettant de produire un métabolite à partir d'un milieu de culture défini. Ces méthodes ont permis de définir *a priori* les milieux minimum pour les organismes *Helicobacter pylori* [Schilling et al., 2002] et *Haemophilus influenzae* [Schilling and Palsson, 2000], données compatibles avec les résultats expérimentaux connus pour ces deux espèces. D'autres auteurs ont utilisé ces méthodes pour étudier la redondance dans les voies métaboliques de ces mêmes organismes ([Papin et al., 2002], [Price et al., 2002]). Enfin, ces méthodes de synthèse ont été appliquées au métabolisme central de *E. coli* afin de trouver les rendements optimaux et sous-optimaux de production d'acides aminés aromatiques à partir de sources d'hydrocarbones. Les résultats ont été utilisés pour fabriquer une souche de *E. coli* possédant la voie métabolique ayant le meilleur rendement théorique, rendement validé expérimentalement. L'utilisation des modes élémentaires permet aussi d'étudier une classe de problèmes voisins, à savoir la caractérisation de voies métaboliques ayant un rendement quasi-optimal et possédant des propriétés intéressantes (e.g., produits intermédiaires utiles) [Schuster et al., 1999].

La synthèse de voies métaboliques se révèle donc très intéressante pour l'étude fonctionnelle et structurale des voies métaboliques, que ce soit par la méthode des modes élémentaires ou par les *extreme pathways*. Cependant, un problème majeur se pose quant à la généralisation de ces méthodes à l'étude du métabolisme complet des organismes. En effet, la génération des vecteurs à partir de la matrice est un problème NP-complet [Samatova et al., 2002] et il n'existe pas encore d'algorithme permettant ce calcul en temps polynomial. Comme le notent [Papin et al., 2003], différentes perspectives et solutions sont envisageables pour résoudre cette difficulté. Une solution est de subdiviser le réseau global en sous-réseaux, que ce soit sur des critères biologiques (e.g., glycolyse, cycle TCA) ou de manière algorithmique (e.g., connectivité des métabolites) comme le fait le logiciel METATOOL [Pfeiffer et al., 1999]. Les perspectives intéressantes sont la formulation des modes élémentaires par des réseaux de petri [Oliveira et al., 2001] ou encore la parallélisation des algorithmes calculant les *extreme pathways* [Samatova et al., 2002].

Conclusion Nous avons vu dans cette partie que les méthodes de reconstruction et d'analyse de voies métaboliques peuvent globalement se classer en deux catégories, selon qu'elles utilisent ou non la définition mathématique des flux. Les méthodes basées sur les données de voies métaboliques stockées dans les bases de données dédiées telles que KEGG ou MetaCyc, permettent d'étudier leurs caractéristiques topologiques et structurales [Jeong et al., 2000], que ce soit pour un organisme particulier ou pour comparer les réseaux d'un ensemble d'organismes [Podani et al., 2001]. Le principal avantage de ces méthodes est qu'elles sont simples à mettre en oeuvre, que les données disponibles sont nombreuses et que les analyses réalisables permettent une bonne caractérisation des réseaux. Le principal désavantage de ces

méthodes est qu'elles ne permettent pas facilement de réaliser des simulations de comportement ou de traiter des questions d'ingénierie cellulaire, du fait d'une absence de modèle mathématique sous-jacent [Papin et al., 2003] et d'outils adaptés.

Les méthodes de synthèse de voies métaboliques, que ce soient les modes élémentaires [Schuster and Hilgetag, 1994] ou les *extreme pathways* [Schilling et al., 1999], permettent de réaliser des études structurales des voies métaboliques, des simulations et prédictions, ou encore des modèles pour l'ingénierie cellulaire. Ces méthodes présentent cependant un désavantage important, à savoir que le calcul des vecteurs de flux est un problème NP-complet, pour lequel des algorithmes polynomiaux n'existent pas encore. Des solutions sont cependant envisagées pour contourner ce problème de complexité, en subdivisant par exemple le réseau global en sous-réseaux où le calcul est faisable. Les avancées théoriques sur les modes élémentaires, ainsi que les avancées dans l'implémentation des algorithmes laissent espérer que ces méthodes seront dans le futur plus facilement applicables à des réseaux métaboliques complets et de taille importantes.

Chapitre 3

Extraction de réseaux d'interactions biomoléculaires dictée par des politiques

Sommaire

3.1 Introduction

La compréhension des systèmes biologiques nécessite l'analyse de grands volumes de données, venant de sources diverses et complémentaires, nécessitant un processus d'intégration de l'information. De plus, ces systèmes présentent un haut niveau de complexité, tant dans le nombre d'éléments impliqués dans leur fonctionnement, que dans les relations existant entre ces éléments. Cette complexité intrinsèque implique de n'étudier que des sous-parties du système complet, en d'autres termes, d'envisager le système selon des points de vue multiples pour en appréhender la complexité. Nous définissons ici la notion de vue comme étant un ensemble de données cohérentes, décrivant des relations de voisinage entre entités appartenant au même champ sémantique. Les voies métaboliques et les réseaux d'interactions protéine-protéine sont des exemples de vues sur les systèmes biologiques. Les vues que nous

construisons sont donc des résumés intelligents et intelligibles des données disponibles.

De nombreuses méthodes ont été développées pour exploiter ces mêmes données ([Jansen et al., 2003], [Nanni and Lumini, 2006], [Wang et al., 2006], [Wu et al., 2006], [Espadaler et al., 2005]). Bien que partant de stratégies similaires (cf. section 2.1.3), ces méthodes présentent souvent l'inconvénient d'être restreintes à une classe de problèmes ou à une étude particulière, si bien qu'elles ne sont que peu ou pas réutilisables. La plupart des études utilisent un jeu de données particulier (e.g., [Uetz et al., 2000]) ou au contraire tentent de fusionner différents ensembles [Sprinzak et al., 2003], mais sans jamais fournir de méthode générale permettant cette fusion. Si ces études apportent souvent des résultats intéressants, il n'en est pas moins difficile de les comparer et de les réutiliser, étant donné l'abscence de base commune à leur développement.

Dans cette partie, nous présentons un cadre formel permettant l'extraction d'informations pertinentes à partir d'ensembles de données hétérogènes, ainsi que leur intégration en un résumé structuré, définissant ainsi des vues sur les systèmes biologiques. Ce formalisme doit pouvoir extraire des relations de voisinage entre biomolécules, de même qu'il doit fournir des moyens de visualisation des données extraites. Les politiques d'extraction de résumés doivent permettre de définir le type de voisinage pris en compte (e.g., interactions protéine-protéine, cooccurrence dans les voies métaboliques, profils d'expression), tandis que les politiques de visualisation doivent permettre la mise en évidence des propriétés émergentes du résumé. Enfin, ce cadre de travail doit être à-même d'intégrer des méthodes existantes, qu'elles fassent parties des politiques d'extraction ou de visualisation.

Nous pensons donc qu'une stratégie appropriée doit clairement établir une frontière entre l'extraction et la représentation de l'information. Cette séparation permet de s'assurer de la correction du résumé fourni en regard des politiques définies, puisque la première étape extrait le sous-ensemble de données pertinentes correspondant à la question posée par l'utilisateur (politique d'extraction), et que la deuxième étape construit la réponse à partir de ce sous-ensemble de données (politique de visualisation). Cette stratégie en deux étapes distinctes est assez intuitive et nous assure modularité, flexibilité et extensibilité.

Le formalisme que nous allons définir est basé sur cette approche. Il fournit des méthodes génériques pour la réalisation de résumés, à partir de politiques d'extraction d'informations et de politiques de représentation de l'information. Ces politiques sont définies par rapport à la requête d'un utilisateur. Une politique d'extraction correspond à la définition du champ sémantique des relations de voisinage qui vont être extraites des données, alors qu'une politique de visualisation défini la manière de matérialiser les relations, pour en faire émerger les propriétés intéressantes. Notre formalisme permet en outre de tirer parti des méthodes de prédiction et de visualisation existantes. Dans cette partie, nous donnerons tout d'abord une définition formelle de notre cadre de travail, puis nous définirons les algorithmes d'extraction

et de représentation de l'information. Enfin, nous montrerons diverses applications de ce cadre formel à des problèmes concrets, ainsi que la mise en oeuvre de notre méthode dans le logiciel ProViz.

3.2 Définition du cadre formel

Le cadre formel que nous définissons est basé sur l'extraction de relations de voisinage entre biomolécules. Cette extraction se fait selon un principe d'extensions de voisinages autour de biomolécules d'intérêt. Le résumé obtenu est donc composé d'ensembles de biomolécules, de telle sorte que les relations entre ces entités sont transitives :

$$p \in \{p|\pi_k(p)\},\ \text{où}\ \pi_k = \exists(p_0,...,p_k)\ \ \&\ \ (R_0,...R_k)\ \text{s.t.}\ p_0R_0p_1 \wedge \\ p_1R_1p_2,...,\ p_{k-1}R_{k-1}p_k \wedge \forall i|i <= k, \pi_p(p_i) \wedge \forall i|i < k, \pi_o(R_i) \quad (1)$$

Biomolécules et observations

Lemme 1 *Soit $\{\mathbb{P}\}$ un ensemble de biomolécules et $\{\mathbb{O}\}$ un ensemble d'observations. Les observations $o \in \{\mathbb{O}\}$ sont des k-tuples de biomolécules $p \in \{\mathbb{P}\}$. Pour une observation o, une biomolécule p peut avoir un rôle $r \in R$, où R est l'ensemble des rôles possibles. Nous définissons la fonction rôle :$\{\mathbb{P}\} \times \{\mathbb{O}\} \rightarrow R$, qui retourne le rôle d'une biomolécule p dans o, ainsi que la fonction $e_p : 2^{\{\mathbb{O}\}} \rightarrow 2^{\{\mathbb{P}\}}$, qui extrait les biomolécules associées à une observation o.*

Définition 1 *Pour une protéine p donnée, nous définissons récursivement la relation N_i de i-*voisinage *:*

$$\begin{aligned} N_0(p) &= \{p\} \\ N_{i+1}(p) &= N_i(p) \cup \underbrace{\{p' \in P \mid \exists\ q \in N_i(p), o \in \{\mathbb{O}\}\ s.t.\ p' \in o\ \ \&\ \ q \in o\}}_{frontierF_{i+1}} \end{aligned}$$

où F_i est la i-frontière de N_{i-1} pour $i > 0$.

La généralisation de cette définition aux ensembles de protéines $P_0 \subseteq \{\mathbb{P}\}$ est :

$$\begin{aligned} N_0(P_0) &= \textstyle\bigcup_{p \in P_0} N_0(p) = P_0 \\ N_{i+1}(P_0) &= \textstyle\bigcup_{p \in P_0} N_{i+1}(p) \end{aligned}$$

Voisinages et frontières Nous considérons les voisinages N_i et frontières F_i comme des tuples de biomolécules et observations $\langle\{\mathbb{P}\}, \{\mathbb{O}\}\rangle$.

$$N_i = \langle P_{Ni}, O_{Ni}\rangle\ et\ F_i = \langle P_{Fi}, O_{Fi}\rangle\ \text{où}\ P_{Ni}, P_{Fi} \subseteq \{\mathbb{P}\}\ et\ O_{Ni}, O_{Fi} \subseteq \{\mathbb{O}\}$$

N'ayant aucun préjugé sur les relations entre les biomolécules d'intérêt, tous les tuples N_0 seront notés $\langle P_{N0}, \emptyset \rangle$. Les biomolécules et observations des voisinages N_i et frontières F_i seront notées P_{N_i}, P_{F_i} et O_{N_i}, O_{F_i}.

Filtrage du réseau Nous définissons les filtres comme des prédicats utilisés pour la sélection de biomolécules et observations.

Définition 2 *Un filtre π est une application :*

$$Filtre\ \pi = \begin{cases} \pi_p : \{\mathbb{P}\} \rightarrow \mathbb{B} & \textit{sur les biomolécules} \\ \pi_o : \{\mathbb{O}\} \rightarrow \mathbb{B} & \textit{sur les observations} \end{cases}$$

Lemme 2 *Soit $O^\pi \subseteq \{\mathbb{O}\}$ un ensemble filtré d'observations : $O^\pi = \{o \in \{\mathbb{O}\},\ \pi_o(o)\}$ et $P^\pi \subseteq \{\mathbb{P}\}$ un ensemble filtré de biomolécules : $P^\pi = \{p \in \{\mathbb{P}\},\ \pi_p(p)\}$. Ces prédicats transforment chaque voisinage N_i en un voisinage filtré N_i^π, et chaque frontière F_i en une frontière filtrée F_i^π.*

Étiquetage du réseau Des étiquettes peuvent être ajoutées aux observations, pour associer une information contenue dans les données extraites ou obtenue par calcul (e.g., indice de qualité pour les interactions protéiques). Les fonctions d'étiquetage sont ainsi définies :

Définition 3 *Soit $\mathcal{A}_o$, un alphabet d'étiquetage des observations. La fonction d'étiquetage λ est définie par $\lambda : \{\mathbb{O}\} \rightarrow \mathcal{A}_o$.*

3.3 Définition des algorithmes

À chaque itération, l'algorithme *Core* (page 59) extrait un ensemble d'observations O à partir du jeu de données D, puis appel l'algorithme *Extend* (page 59) qui calcule une frontière filtrée F_{i+1}^π, en utilisant N_i et O. Par construction, *Extend* associe un ensemble d'observations et un ensemble de biomolécules. Le voisinage N_{i+1}^π est obtenu par l'union et le filtrage de N_i et F_{i+1}^π.

3.3.1 Extraction des données

Algorithmes L'algorithme *Core* calcule des voisinages successifs de biomolécules à partir d'un jeu de données D, d'une fonction *Find* qui sélectionne les observations, d'un voisinage N_0 contenant les biomolécules d'intérêt, et d'un ensemble de filtres. Il peut être réduit à la série suivante :

$$N_o \rightarrow N_0^\pi \rightarrow N_1^\pi = N_0 \cup F_1^\pi \rightarrow N_2\pi = N_1^\pi \cup F_2^\pi \rightarrow \ldots$$
où $N_i \cup F_i = \langle P_{Ni} \cup P_{Fi}, O_{Ni} \cup O_{Fi} \rangle$ et $i \in [1...I]$

La fonction *Find* prend en charge la sélection des observations du jeu de données D. Elle est définie par $\{\mathbb{O}\} \times \{\mathbb{P}\} \rightarrow \{\mathbb{O}\}$. *Find* requiert un jeu de données D et un ensemble de biomolécules $P_{F_{i-1}}$. En retour, elle fournit un ensemble d'observations O. La fonction *Find* peut être implémentée de différentes manières, une version simple étant de retourner les observations de D qui contiennent au moins une biomolécule $p \in P_{N_i}$.

Algorithme 1 *Core*

Entrées: $Find : \{\mathbb{O}\} \times \{\mathbb{P}\} \rightarrow \{\mathbb{O}\}$, $D \subseteq \{\mathbb{O}\}$, $N \in \langle\{\mathbb{P}\}, \{\mathbb{O}\}\rangle$, $\pi = \langle\pi_o, \pi_p\rangle$, $Build : \langle\{\mathbb{P}\}, \{\mathbb{O}\}\rangle \times \langle\{\mathbb{P}\}, \{\mathbb{O}\}\rangle \times \langle V, E\rangle \rightarrow \langle V, E\rangle$ et $i \in [1...I]$

Soit $P_F = P_N$
Soit $G = \langle\emptyset, \emptyset\rangle$
pour tout $i \in [1...I]$ **faire**
  Soit $O_F = Find(D, P_F)$
  Soit $F = Extend(O_F, F, N, \pi)$
  Soit $N = \langle\pi_p(P_N \cup P_F), \pi_o(O_N \cup O_F)\rangle$
  Soit $G = build(N, F, G)$
fin pour
Retourner G

L'algorithme *Extend* prend en charge les extensions successives du voisinage N_0. Il calcule la frontière F_i en fonction du voisinage N_{i-1}. Chaque observation $o \in O$ contenant une biomolécule $p \in P_{N_{i-1}}$ est ajoutée aux observations de F_i.

Algorithme 2 *Extend*

Entrées: $O \in \{\mathbb{O}\}$, $F \in \langle\{\mathbb{P}\}, \{\mathbb{O}\}\rangle$, $N \in \langle\{\mathbb{P}\}, \{\mathbb{O}\}\rangle$ et $\pi = \langle\pi_o, \pi_p\rangle$

Soit $O = O_F$
Soit $F = \langle\emptyset, \emptyset\rangle$
pour tout $o \in O$ **faire**
  Soit $n = e_p(o) \setminus P_N$
  si $n \neq \emptyset$ **alors**
    Soit $P_F = P_F \cup n$
    Soit $O_F = O_F \cup \{o\}$
  finsi
fin pour
Soit $F = \langle\pi_p(P_F), \pi_o(O_F)\rangle$
Retourner F

Propriété 1 *Le dernier voisinage retourné par l'algorithme* Core-Extend *vérifie la formule (1).*

Preuve : la preuve est réalisée par induction mathématique. Soit D le jeu de données. Soit π_p et π_o les filtres appropriés. Le cas de base $i = 0$ est immédiat,

puisque $N_0 = \pi_p(D)$. Supposons que l'équation (1) est valide pour N_i. Pour la construction de N_{i+1}, nous ne faisons qu'ajouter des biomolécules qui vérifient π_p et qui participent dans des observations vérifiant π_o. En outre, La longueur du chemin maximal dans le jeu de données originel est au plus i, par construction, et ceci seulement s'il n'existe pas de relation réflexive R sur ce chemin. Ainsi, le voisinage N_{i+1} vérifie l'équation (1).

Exemple Soit D un jeu de données expérimentales. La figure 19 (page 61) décrit les observations contenues dans D. Chaque observation est identifiée par un numéro et est associée à la liste des biomolécules impliquées. Ces observations sont issues d'une expérience visant à caractériser les interactions protéine-protéine. Nous supposons par ailleurs que les protéines C et K ont des rôles non liés au cycle cellulaire. Nous proposons donc une instanciation de notre formalisme, dans laquelle nous omettons volontairement la spécification d'un algorithme de construction, puisque nous nous focalisons ici sur l'extraction et non sur la représentation de l'information. Cette instance possède un prédicat de filtrage π_p sur le rôle des biomolécules, afin de ne retenir que les protéines impliquées dans le cycle cellulaire. Le voisinage N_0 contient les biomolécules d'intérêt A, B et C. La description et le déroulement de l'instance *Exemple* proposée sont représentés sur la figure 19.

Sur cette figure, nous pouvons noter que l'observation 4 n'est pas présente dans le résumé final. En effet, l'observation 4 est composée des protéines I et K qui ne sont pas à distance 1 des protéines de départ. Aussi, la fonction *Find* ne retourne-t-elle pas cette observation. L'instance *Exemple* n'ayant qu'une seule itération, l'observation 4 ne sera pas prise en compte. Nous voyons donc ici que les informations extraites par une instance de notre cadre formel peuvent ne correspondre qu'à un sous-ensemble des informations disponibles dans le jeu de données D.

3.3.2 Construction de graphes

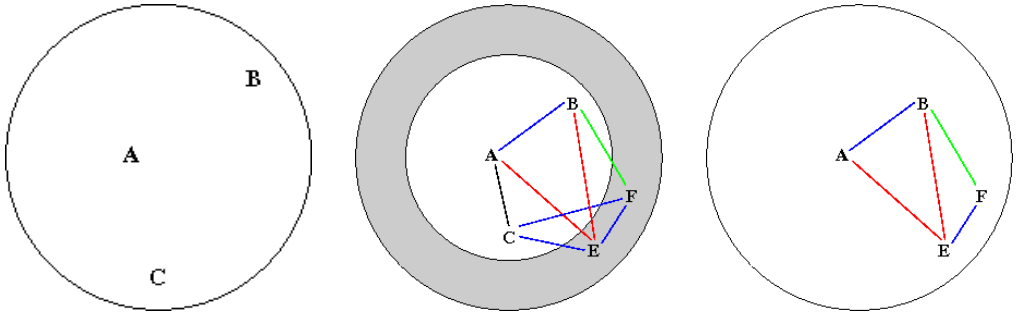
Le second composant de notre cadre formel est la visualisation des informations extraites. Cette matérialisation du résumé réalisé à l'étape précédente doit permettre à l'utilisateur de mettre en avant différentes propriétés des relations extraites. L'utilisateur peut donc définir différentes politiques de visualisation, qui vont correspondre à autant d'instances de l'algorithme *Build* (section 3.4). Le résultat est un graphe d'interactions biomoléculaires (définition 4).

Définition 4 *Un* graphe d'interactions biomoléculaires *est un graphe $G = \langle V, E\rangle$ où V est un ensemble de sommets et E est un ensemble d'arcs. Un sommet $v \in V$ représente une biomolécule et un arc $e \in E, e = \langle v_1, v_2\rangle$ entre deux sommets v_1 et v_2, représente une relation de voisinage entre les biomolécules correspondantes.*

$$Exemple = \begin{cases} Find(D,\ P_F) \ = \ \{o \ \in \ D,\ \exists\ p \ \in \ P_F \text{ s.t. } p \ \in \ o\} \\ \pi_p(p) \ = \ \{p \ \in \ P_F,\ \exists\ o \ \in \ O_F \text{ s.t. rôle}(p, o) \ = \ \text{“}cycle\ cellulaire''\} \\ \pi_o(o) \ = \ vrai \\ N_0 \ = \ \langle\ \{A, B, C\},\ \{\emptyset\}\ \rangle \\ I = 1 \end{cases}$$

Observations :
1 : A, B, E | 2 : C, F, E | 3 : B,F | 4 : I,K | 5 : A,C

Décomposition de l'algorithme :



(1) Voisinage N_0 (2) Frontière F_1 (3) Voisinage N_1^π

(1) $P_{N0} = \{A, B, C\}$ et *Find* retourne l'ensemble $O_1 = \{1, 2, 3\}$
(2) $P_{F_1} = \{A, B, C, E, F\}$ et $O_{F_1} = \{1, 2, 3\}$
(3) $P_{N_1^\pi} = \{A, B, E, F\}$ et $O_{N_1^\pi} = \{1, 2, 3\}$

Fig. 19 – Instanciation et exécution de l'exemple d'extraction d'informations sur le jeu de données D. Chaque observation n est représentée par un ensemble d'arcs étiquetés par une couleur. L'observation 1 est représentée par des arcs rouges, la 2 par des arcs bleus, la 3 par un arc vert, la 4 par aucun arc, la 5 par un arc noir. Chaque observation est caractérisée par un identifiant et par la liste des protéines en interaction.

L'algorithme *Build* est en charge de la construction du graphe. À chaque étape, *Build* modifie le graphe en cours de construction en utilisant P_{N_i} et F_i pour décider quels sommets et arêtes doivent être ajoutés ou modifiés dans le graphe G. Dans tous les cas, deux sommets représentent deux biomolécules et sont connectés dans le graphe si et seulement si une observation décrivant cette relation est présente dans le jeu de données.

Propriété 2 *Pour chaque algorithme de construction de graphe, V et E croissent de manière monotone, en fonction des voisinages et frontières filtrées N_i^π et F_i^π.*

Propriété 3 *Build construit une graphe à partir d'informations filtrées, si bien qu'il ne contient pas nécessairement toute l'information contenue dans* D.

3.4 Applications

Notre cadre formel pour l'extraction de graphes d'interactions biomoléculaires dictée par des politiques est générique, en ce sens que seule la stratégie globale est définie (notion d'extensions de voisinages, séparation de l'extraction et de la représentation). Ainsi, s'application de notre méthode nécessite de définir des instances adaptées à chaque problème (instance *Exemple*, figure 19). La flexibilité est à ce niveau assurée par la possibilité de définir la fonction *Find*, les prédicats de filtrages sur les biomolécules et observations, ainsi que le nombre d'extensions de voisinage souhaitées. Enfin, la séparation de l'étape d'extraction et de l'étape de visualisation nous permet d'ajouter de la modularité, en permettant différentes représentations des mêmes données, mettant en exergue différentes propriétés. Dans les travaux de Master de J-P Soularue [Soularue, 2005], nous avons fait la preuve qu'il est possible de définir des instances de notre formalisme utilisant les méthodes de [Bader and Hogue, 2003] et [Bader, 2003], que nous avons vu en section 2.1.3.

Topologie des graphes Nous proposons trois algorithmes concrets de construction de graphes :

- *Build_SBG* construit un graphe d'interactions binaires
- *Build_GOPC* construit un graphe de complexes protéiques
- *Build_MetaboPath* permet la reconstruction de voies métaboliques

De nombreux autres sont possibles, dont certains ont été décrits dans les travaux de master précédemment mentionnés.

Graphes d'interactions binaires La topologie d'un *graphe d'interactions binaires* (algorithme *Build_SBG*) est très proche de la notion de *spoke model* introduite dans [Bader and Hogue, 2003]. Cette topologie est adaptée aux données

ne contenant que des interactions binaires. En effet, les interactions n-aires, telles que dans les complexes protéiques, ne peuvent être correctement représentées. Nous avons vu en section 2.1.3, que certains auteurs représentent les complexes protéiques sous forme de clique ou graphe complet, dans lequel chaque sommet est en relation avec tous les les autres. Dans le cadre des complexes protéiques, cela reviendrait à induire une interaction pour toute paire de protéines membres du complexe. Le modèle *spoke* ou le modèle matrice ne peuvent représenter de manière fidèle les relations entre protéines au sein d'un complexe, même si [Bader and Hogue, 2003] ont montré que le modèle *spoke* est plus précis. Comme ces deux modèles, notre graphe d'interactions binaires n'est pas adapté à la représentation de complexes protéiques, à moins que les interactions précises entre les partenaires ne soient disponibles. Pour résoudre ce problème, nous proposons le graphe de complexes, présenté ci-après.

Algorithme 3 *Build_SBG*

Entrées: $N \in \langle\{\mathbb{P}\},\{\mathbb{O}\}\rangle$, $F \in \langle\{\mathbb{P}\},\{\mathbb{O}\}\rangle$, $ref\ G = \langle V, E\rangle$
Soit $V = biomols(N) \cup V$
Soit $E = E \cup \{\langle v_n, v_m\rangle\}, \exists o \in obs(F)\text{s.t.}v_n, v_m \in o$

Graphes de complexes protéiques Afin de résoudre le problème de représentation des complexes protéiques dont la topologie exacte n'est pas connue, nous avons défini le *graphe de complexes protéiques* (algorithme *Build_GOPC*). Cette politique de représentation des données introduit des sommets additionnels représentant l'entité virtuelle du complexe protéique. Pour un complexe protéique c, composé des protéines $P = \{p_1, ..., p_n\}$, l'ensemble de sommets correspondant sera $V_c = \{p_1, ..., p_n, v_c\}$ et l'ensemble d'arcs, $E = \{(p_1, v_c), ..., (p_n, v_c)\}$ où v_c est le sommet qui représente c. Les interactions binaires, que ce soient entre protéines du complexe ou entre d'autres protéines du réseau, sont représentées de la même manière que pour la politique *Build_SBG*. Nous utilisons une fonction *is_complex* pour détecter les complexes protéiques présents dans le jeu de données : $is\_complex : \{\mathbb{O}\} \rightarrow \{\mathbb{P}\}$. Nous supposons aussi que nous pouvons ajouter des sommets au graphe courant par un appel à une fonction *Create_new_vertex*, qui retourne le sommet créé. Cette politique de visualisation est particulièrement adaptée à la représentation de données d'interactions provenant de sources et méthodes hétérogènes (e.g., Y2H, TAP-TAG). Cependant, des sommets sont arbitrairement ajoutés au réseau, sommets qui sont par définitions très connectés. Aussi, les études ultérieures basées sur ce graphe doivent prendre en compte les éventuelles modifications de la topologie globale du réseau, amenées par cet ajout de sommets virtuels.

Reconstruction de voies métaboliques Une visualisation assez intuitive des voies métaboliques est l'utilisation d'un graphe bipartite (voir section 2.2.3), entre enzymes et métabolites. Nous utilisons la fonction *is_biochem_reaction* pour savoir si

Algorithme 4 *Build_GOPC*

Entrées: $N \in \langle\{\mathbb{P}\},\{\mathbb{O}\}\rangle$, $F \in \langle\{\mathbb{P}\},\{\mathbb{O}\}\rangle$, $ref\ G = \langle V, E\rangle$
  pour tout $o \in obs(f)$ **faire**
    si $is\_complex(o)$ **alors**
      Soit $v_c = Create\_new\_vertex(o)$
      Soit $V = \{v_c\} \cup ep(o) \cup V$
      Soit $E = \{\langle v_c, v\rangle \mid v \in ep(o)\}$
    sinon
      Soit $V = ep(o) \cup V$
      Soit $E = \{\langle v_1, v_2\rangle \mid v_1, v_2 \in biomols(o)\} \cup E$
    finsi
  fin pour

une observation est une interaction de type enzyme-substrat ($is\_biochem\_reaction$: $\{\mathbb{O}\} \rightarrow \{\mathbb{P}\}$). Nous utilisons aussi la fonction *exists_enzyme_activity* , pour savoir s'il existe déjà un sommet dans le graphe représentant l'activité enzymatique associée à l'enzyme en cours d'étude. Cette politique de visualisation défini un réseau métabolique général, en ce sens qu'il ne représente pas un ensemble d'enzymes associées à des métabolites, mais plutôt un ensemble d'activités enzymatiques associées à leur substrats et produits. Ainsi, il est possible de croiser des informations provenant de différentes espèces, permettant par là-même des études sur la conservation des voies métaboliques entre espèces (voir chapitre 4).

Un tel graphe représente la synthèse des réactions biochimiques espèce-spécifiques, dessinant ainsi un réseau métabolique général, avec ses possibles voies alternatives. En utilisant les données de génomique comparée (e.g., conservation de gènes) pour filtrer le graphe, l'utilisateur peut avoir une vue immédiate des voies métaboliques d'une espèce considérée, et donc avoir des indices sur la possible conservation de ces voies métaboliques au sein de son espèce d'intérêt. Au chapitre 4 section 4.5, nous pouvons voir la mise en application d'une politique similaire.

Exemple Dans l'exemple suivant, nous allons illustrer les algorithmes 3 et 5. Nous définissons le jeu de données D composé de 5 observations (figures 20 et 21). La seule information supplémentaire est que l'observation 2 est le résultat d'une méthode expérimentale permettant l'identification de complexes protéiques, dans laquelle la protéine C a été utilisée comme appât. α désigne le complexe protéique composé des protéines D et E, caractérisé par l'observation 3. Nous définissons l'ensemble $N_0 = \langle P_o, \emptyset\rangle$, où $P_o = \langle C\rangle$, ainsi que les prédicats de filtrage π_p = vrai et π_o = vrai. Nous définissons de même un nombre d'itérations arbitraire k, supposé suffisant pour exploiter l'ensemble des relations du jeu de données.

Algorithme 5 *Build_MetaboPath*

Entrées: $N \in \langle\{\mathbb{P}\},\{\mathbb{O}\}\rangle$, $F \in \langle\{\mathbb{P}\},\{\mathbb{O}\}\rangle$, $ref\ G = \langle V, E\rangle$
  pour tout $o \in obs(f)$ **faire**
    si $is\_enzyme\_ligand(o)$ **alors**
      si $exists\_enzyme\_activity$ **alors**
        Soit $V = v_m \cup V$, s.t.V_m is not the enzyme
      sinon
        Soit $V = biomols(o) \cup V$
      finsi
      Soit $E = E \cup \{\langle v_n, v_m\rangle\}$
    finsi
  fin pour

Jeu de données D : 1 = D, C | 2 = C, F, E | 3 = α : D, E | 4 = A, α | 5 = A, E

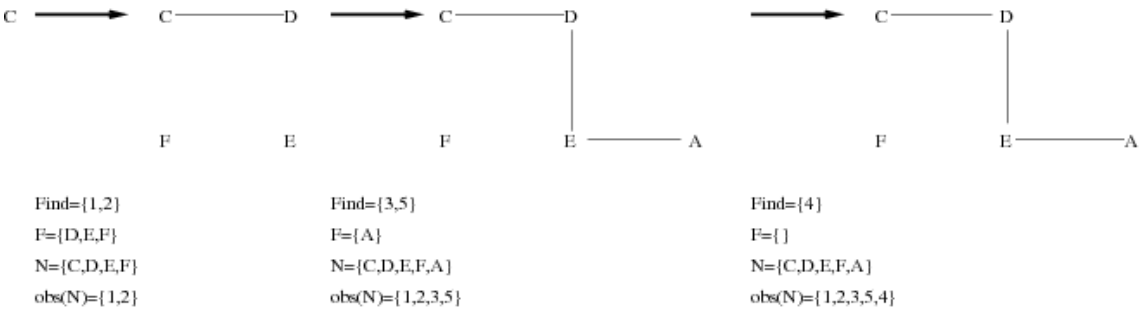


Fig. 20 – Illustration des itérations de l'algorithme *Build_SBG* sur le jeu de données *D*.

Les itérations de notre algorithme, utilisant *Build_SBG* comme politique de visualisation, sont représentées par la figure 20. Celles de la version utilisant l'algorithme *Build_GOPC* comme politique de visualisation sont présentées sur la figure 21.

Dans la première itération de l'algorithme 3, nous pouvons voir qu'il n'existe qu'une arête entre C et D, puisque nous ne connaissons pas les relations d'interaction à l'intérieur du complexe identifié par l'observation 2. Ainsi, les protéines sont ajoutées au réseau, mais nous ne pouvons ajouter aucune arête. L'algorithme trouve alors la protéine A et le complexe protéique α, mais nous ne pouvons ajouter aucune arête entre A, D et E puisque nous ne connaissons pas le détail des interactions de la protéine A avec le complexe α. Nous voyons ici un exemple des limitations dont nous avons précédemment parlées.

Dans la figure 21, nous voyons que l'utilisation de la politique de visualisation *graphe de complexes* permet de représenter les complexes protéiques et leurs relations sans aucune ambiguité. De manière générale, nous voyons donc que cette

Jeu de données D : 1 = D, C | 2 = C, F, E | 3 = α : D, E | 4 = A, α | 5 = A, E

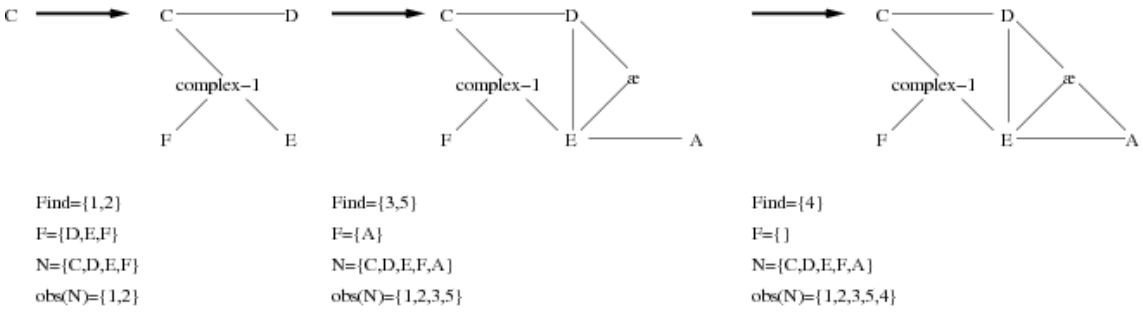


FIG. 21 – Illustration des itérations de l'algorithme *Build_GOPC* sur le jeu de données *D*.

politique de visualisation est bien adaptée à la matérialisation de réseaux d'interactions protéiques contenant des complexes protéiques. Utilisée dans le cadre d'un logiciel de visualisation de graphes, cette politique est bien adaptée à la mise en évidence des complexes protéiques, en définissant par exemple un filtre ne conservant que les arêtes entre les sommets représentant les complexes et leurs partenaires protéiques, ainsi qu'entre ces derniers. Le graphe résultant étant alors composé de composantes connexes représentant les complexes protéiques.

3.5 Mise en oeuvre dans le logiciel ProViz

Nous avons défini dans les sections précedentes, un formalisme permettant l'extraction de données à partir de sources de données hétérogènes, puis la matérialisation des relations de voisinages extraites, sous la forme de graphes. L'analyse des réseaux d'interactions biomoléculaires requière une combinaison d'outils algorithmiques et de visualisation d'information, intégrés dans une plateforme logicielle intégrant elle-même des accès à des bases de données locales et distantes. Dans cette partie, nous présentons le logiciel appelé ProViz [Iragne et al., 2005], qui permet une visualisation hautement interactive de grands réseaux d'interactions protéine-protéine et qui intègre un accès à la base de données d'interactions biomoléculaires IntAct [Hermjakob et al., 2004b].

Le dessin et l'exploration interactive de graphes sont des domaines bien étudiés et très actifs dans la recherche en informatique, et de nombreux logiciels sont disponibles pour cette tâche (PIMRider [Legrain et al., 2001], Osprey du projet GRID [Breitkreutz et al., 2003], Cytoscape [Shannon et al., 2003]). L'adaptation de ces logiciels et techniques aux besoins spécifiques des biologistes pour l'exploration des réseaux d'interactions biomoléculaires est un défi permanent pour la bioinformatique. Plus que l'aspect graphique, l'enjeu est ici d'ajouter de l'information et des

fonctions adaptées, permettant à l'utilisateur de découvrir les relations intéressantes noyées au sein des bases de données. Dans ce cadre, nous avons mis en oeuvre l'extraction de graphe dictée par des politiques, que nous venons de définir, et implémenté les politiques de visualisation de *graphes d'interactions binaires* et de *graphes de complexes*.

Utilisations de ProViz L'intégration d'un vocabulaire contrôlé, de même que la prise en compte des habitudes de travail des biologistes, sont des atouts essentiels pour ProViz. Il peut être utilisé pour explorer de grands graphes d'interactions protéiques, afin d'identifier des éléments ou des interactions d'intérêt, que se soit par la recherche par mots-clé ou par l'analyse de la structure du graphe. ProViz peut aussi être utilisé afin d'extraire des vues du graphe d'interactions, que celles-ci soient générées par filtrage, clustering ou encore par sélection manuelle. ProViz peut enfin permettre de comparer les graphes d'interactions de différentes espèces en utilisant des ensembles de gènes orthologues. Ce logiciel est hautement interactif et permet des mises à jour de l'image de l'ordre de 50ms lors de la manipulation de graphes d'un million d'éléments sur des ordinateurs standards.

Interface utilisateur L'interface utilisateur de ProViz est volontairement simple et épurée, afin de permettre une prise en main quasi-immédiate (figure 22). La moitié droite de l'écran affiche la vue courante du graphe en cours d'étude. Différentes vues sont disponibles et accessibles par l'utilisation d'onglets. Une barre d'outil surmontant la fenêtre principale permet un accès rapide aux principales fonctions du logiciel. La sélection d'éléments peut se faire à la souris ou par recherche à l'aide de mots-clés. Chaque élément peut être déplacé par l'utilisation de la souris, et la molette de cette dernière permet d'effectuer différentes opérations de zoom et de déplacement de l'image.

Algorithmes de dessin de graphes Les algorithmes de dessin de graphes choisis pour ProViz l'ont été à la fois pour leurs performances et pour leur capacité à mettre en évidence des informations ayant une pertinence biologique. GEM [Frick et al., 1994] est une version efficace d'un algorithme de dessin de graphes dirigés, basé sur un calcul de la force des liens entre sommets. Il regroupe les sommets en étroite relation et éloigne les sommets moins liés. Il peut être utilisé pour rapidement identifier les protéines ayant un rôle particulier ou pour la visualisation de complexes protéiques. L'algorithme de dessin hiérarchique [Messinger et al., 1991] tend à mettre en évidence les relations de parenté entre les sommets, de manière claire et non-ambiguë. Cette approche peut se révéler utile dans le cadre d'études de graphes composés de cascades d'interactions ou dans le cas de comparaisons de graphes d'interactions protéine-protéine avec des graphes de voies métaboliques. Enfin, l'algorithme de dessin de graphe circulaire place tous les sommets sur la périphérie d'un unique cercle [Mäkinen, 1988]. C'est un algorithme de dessin neutre,

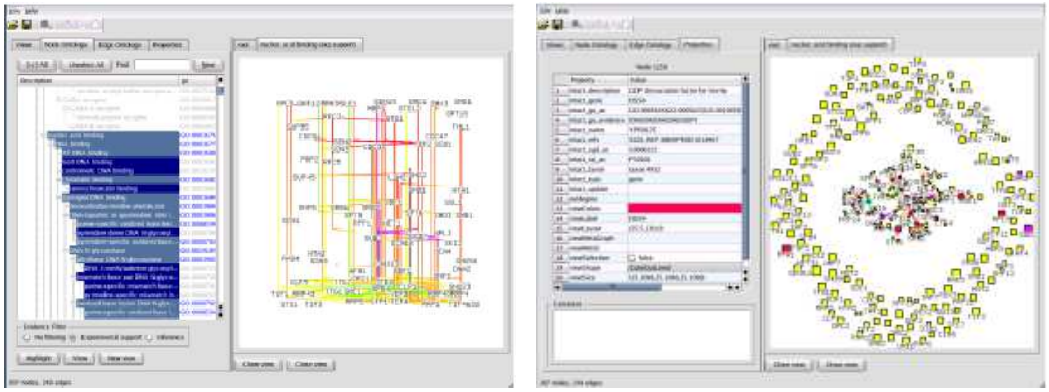


FIG. 22 – Deux vues d'un graphe d'interactions protéiques mettant en évidence la protéine de levure Dss4. L'image de gauche est une vue obtenue par filtrage du graphe par termes de Gene Ontology. L'image de droite représente le même sous-graphe après application du layout GEM

dans le sens où aucune valeur sémantique ni aucun classement ne sont appliqués aux sommets ou aux arcs.

Intégration d'attributs biologiques La moitié gauche de l'écran affiche les informations sur les éléments du graphe en cours d'étude. Quatre onglets sont disponibles : **Views**, pour voir les propriétés des vues disponibles ; **Node Ontology**, qui permet de sélectionner des sommets à partir de terme Gene Ontology [Harris et al., 2004] ; **Edge Ontology**, permettant de sélectionner des interactions à partir de termes d'un vocabulaire contrôlé, défini dans le cadre du projet IntAct (voir section 2.1.2) ; **Properties**, permettant d'accéder à toutes les propriétés d'un sommet ou d'un arc. La figure 22 montre la liste des propriétés associées au sommet représentant la protéine Dss4 de *S. cerevisiae* (GDP dissociation factor), avec notamment le nom du gène, des liens vers des ressources externes, ainsi que des informations de la Gene Ontology.

Accès aux bases de données L'accès et l'intégration de bases de données dans ProViz est un bon exemple de l'ajout de fonctionnalités par le mécanisme de plugins fourni par la plateforme Tulip [Auber, 2001]. La plupart des programmes manipulant des réseaux d'interactions protéine-protéine utilisent préférentiellement le format XML standard PSI Molecular-Interaction [Hermjakob et al., 2004a], développé par le groupe HUPO PSI [Orchard et al., 2003]. ProViz peut utiliser ce format. Les fichiers PSI-MI sont globalement organisés en quatre parties : (1) les protéines, accompagnées de toutes les informations nécessaires et intéressantes, telles que les GO termes et des références vers d'autres bases de données ; (2) les expériences, décrivant les méthodes et procédures utilisées pour la caractérisation des interactions entre protéines ; (3) les interactions, décrivant les relations établies entre les

différentes protéines ; (4) la disponibilité des données, indiquant les publications relatives aux expériences, le copyright ainsi que toutes les informations relatives à la propriété intellectuelle des données représentées.

La plateforme de développement Tulip Le développement de ProViz est basé sur la plateforme logicielle Tulip [Auber, 2001], dédiée à la manipulation et à l'affichage tri-dimensionnel de grands graphes. Tulip fournit un riche ensemble de services de base pour les opérations sur les graphes : calculs de métriques, dessin des sommets et des arcs, sélection et extraction de vues et de sous-graphes, étiquetage des sommets et des arcs à partir d'ensembles arbitraires d'attributs. Les opérations spécifiques au domaine d'application sont fournies par le système intégré de plugins. Ainsi, chaque programme utilisant Tulip comme plateforme de développement peut ajouter aux fonctions de bases toutes les fonctions spécifiques au domaine d'application.

Intégration du cadre formel pour l'extraction de graphes d'interactions biomoléculaires dictée par des politiques Dans le cadre de ProViz, nous avons développée des instances concrètes du formalisme que nous avons présenté ci-avant. Pour ce développement, nous nous sommes basés sur l'utilisation du format XML PSI-MI, adopté par une majorité de sources de données d'interactions protéiques (IntAct, DIP, BIND, voir section 2.1.2). Une interface graphique permet de sélectionner les biomolécules d'intérêt présentes dans le jeu de données, puis de définir le nombre d'extensions de voisinages souhaitées, ainsi que la politique de visualisation et les filtres sur les biomolécules et observations (figure 23). Cette interface graphique calcule en temps réel le nombre de protéines retrouvées par l'algorithme, ainsi que le nombre de relations. Ces mesures peuvent guider l'utilisateur dans sa recherche, pour limiter, le cas échéant, la taille du graphe obtenu. La figure 24 présente un graphe issu de cette extraction dictée par des politiques. Il s'agit en l'occurrence d'un graphe représentant les données d'une expérience de complexome menée chez *S. cerevisiae*, représentées en utilisant la politique de visualisation de *graphe de complexes* (algorithme 4, page 64).

3.6 Conclusion

La construction de vues sur les systèmes biologiques, en utilisant l'information contenue dans de grands volumes de données hétérogènes, est l'un des défis actuels de la biologie. Nous avons défini ici un formalisme permettant l'extraction d'informations à partir de différentes sources de données, ainsi que leur intégration et visualisation par l'utilisation de politiques de construction de graphes. Ce formalisme répond à un problème récurrent d'absence de base commune entre les différentes méthodes existant pour le développement d'analyses sur les graphes d'interactions biomoléculaires, donc sur des graphes représentant des relations de voisinage entre

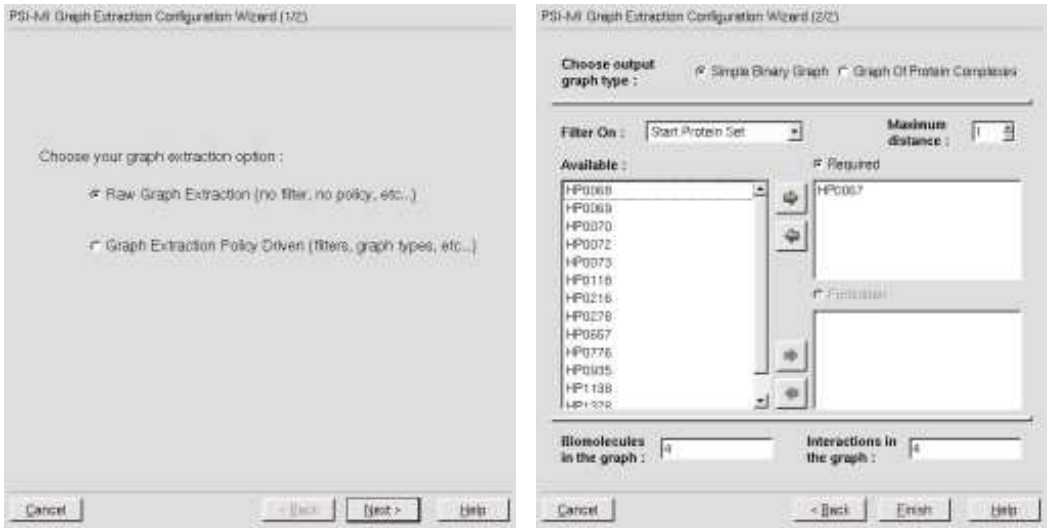


FIG. 23 – Captures d'écran de l'interface de configuration pour l'extraction de graphes dictée par des politiques. La première capture montre l'interface permettant de choisir la politique de visualisation. La deuxième montre l'interface permettant de choisir les biomolécules d'intérêt, le nombre d'extensions de voisinages, ainsi que les filtres sur les biomolécules et observations.

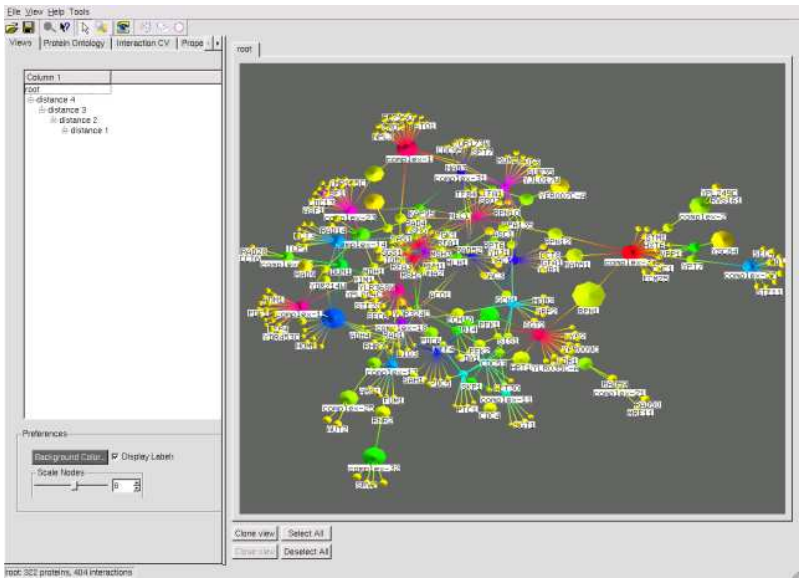


FIG. 24 – Exemple de visualisation d'informations issues de l'extraction de graphes dictée par des politiques. Le résultat présenté est un graphe représentant les complexes et interactions protéiques issus d'une expérience menée chez *S. cerevisiae* et visualisés par la politique de visualisation *graphe de complexes*. Le panneau de gauche (*Views*) permet d'accéder aux différentes distances de voisinage.

des biomolécules (e.g., interactions protéiques, voies métaboliques, profils d'expression).

Notre formalisme se décompose en deux parties, l'une responsable de l'extraction des relations de voisinage entre entités d'intérêt, l'autre responsable de la matérialisation de ces relations. La correction du résultat obtenu est assurée par la définition précise de la sémantique des relations de voisinage que nous extrayons. Ainsi, une interaction protéique ne peut pas être directement ajoutée à une relation de voisinage de voie métabolique, et nécessite la définition d'une relation de voisinage réalisant l'union d'ensembles de relations d'interactions protéiques et de relations dans les voies métaboliques.

Un avantage certain de notre formalisme est de permettre une certaine modularité et flexibilité, par la possibilité de définir différentes instances concrètes d'extraction de relations de voisinage, de visualisation de l'information extraite, ou encore par la définition de prédicats de filtrages plus complexes que de simples filtres sur les étiquettes des entités et de leurs relations. Nous avons notamment démontré que des méthodes existantes ([Bader and Hogue, 2003] et [Bader, 2003]) peuvent être incorporées à notre formalisme [Soularue, 2005]. Cette possibilité d'intégration de méthodes existantes offre de nombreuses perspectives pour l'application de notre formalisme, tant au point de vue de la définition de politiques d'extractions à partir de méthodes intégratives (voir section 2.1.3), que de la définition de méthodes de visualisation [Friedrich and Schreiber, 2003].

Une fois ce cadre de travail défini et validé, l'étape suivante est logiquement l'implémentation des algorithmes concrets. Nous avons présenté le logiciel ProViz [Iragne et al., 2005], dédié à la visualisation hautement interactive de réseaux d'interactions biomoléculaires. Basé sur la plateforme logicielle Tulip, ProViz propose de nombreuses fonctions adaptées aux besoins des biologistes, notamment par l'utilisation du système de plugins intégré à la plateforme de développement. Parmi ces fonctions, nous pouvons noter l'utilisation extensive de vocabulaires contrôlés (Gene Ontology ou PSI-MI), permettant sélections et analyses des graphes visualisés. ProViz propose en outre une première implémentation de notre formalisme, utilisant un jeu de filtres simples et les algorithmes *Build_SBG* ou *Build_GOPC* (pages 63 et 64). Des développements futurs devraient permettre la définition de filtres adaptés à certaines classes de problèmes, ainsi que la définition des politiques de visualisation nécessaires. L'implémentation de l'algorithme de reconstruction de voies métaboliques est aussi un point de grand intérêt.

Chapitre 4

Extrapolation de voies métaboliques à partir de données de la génomique comparée

Sommaire

4.1 Introduction

L'émergence de la *Systems Biology* [Westerhoff and Palsson, 2004] a permis le développement de nombreux modèles mathématiques pour des espèces de références [Kitano, 2002]. Les réseaux de relations entre biomolécules sous-tendent ces modèles, qu'ils soient sous la forme de systèmes d'équations différentielles ou de modèles hiérarchiques de systèmes et sous-systèmes discrets. La construction de ces réseaux est donc un prérequis à la modélisation systématique. Les méthodes expérimentales fournissent des réseaux de haute qualité, mais sont coûteuses et sont donc limitées à un ensemble restreint d'organismes d'intérêt. Les méthodes prédictives fournissent des réseaux de moindre qualité mais sont théoriquement applicables à tout organisme dont les données génomiques sont disponibles. Dans ce chapitre, nous présentons un ensemble de méthodes permettant l'extrapolation de modèles par une analyse, basée sur les graphes, de réseaux métaboliques et par une identification fiable d'équivalents

fonctionnels. Nous présentons aussi les résultats de cette méthode, appliquée à des levures d'intérêt biotechnologique et industriel.

Les levures ont des niches écologiques très variées et partagent cependant de nombreux processus biologiques [Dujon, 2006]. Parmi les levures figure l'organisme modèle *S. cerevisiae*, largement utilisée en biologie moléculaire et cellulaire, ainsi que dans l'industrie en tant qu'usine cellulaire. Cette levure fait partie des petits organismes eucaryotes les mieux annotés et sert donc souvent de référence pour l'annotation d'autres organismes. *S. cerevisiae* fait partie du phylum des hémiascomycètes, qui regroupe un ensemble d'organismes relativement diversifiés d'un point de vue physiologique et écologique. Différentes recherches en génomique comparée au sein de ce phylum ont montré que les résultats obtenus par ces méthodes permettent une bonne compréhension des mécanismes de l'évolution, mis en place lors des événements de spéciation ([Souciet J., 2000], [Kellis et al., 2003]). Les voies métaboliques représentent des relations entre éléments fonctionnels par des réactions enzyme-substrat, qui peuvent être prédites avec plus de confiance que les interactions protéiques ou les relations de régulation (section 2.2.3), et sont par conséquent un bon point de départ pour tester une méthode d'extrapolation de réseaux d'interactions biomoléculaires.

Dans le cadre d'une étude préliminaire, nous avons tenté de caractériser les différences quantitatives entre les ensembles de gènes de deux espèces phylogénétiquement proches, *D. hansenii* et *C. albicans*. Nous avons abordé cette évaluation avec l'idée que deux espèces proches ne se différencient que par un faible nombre de gènes spécifiques à l'une ou l'autre, et que la plus grande partie de leur gènes sont communs avec les autres espèces de leur phylum. Cette étude nous a permis de montrer que ces deux espèces ont 5/6 de gènes en commun avec les autres espèces du phylum hémiascomycète, et que seuls 6% de leur gènes semblent spécifiques à leur clade. Ces résultats nous permettent aussi de penser que, étant donné la forte conservation des gènes au sein du phylum des hémiascomycètes, une approche de prédiction de voies métabolique par extrapolation devrait apporter des résultats intéressants, et notamment montrer une large conservation des voies métaboliques. Pour les méthodes d'extrapolation que nous allons définir, nous utilisons les voies métaboliques du KEGG [Kanehisa, 1997], spécifiques à *S. cerevisiae* et les familles de protéines Génolevures [Nikolski and Sherman, 2007], pour prédire la conservation ou la perte de voies métaboliques chez quatre espèces du phylum hémiascomycète : *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica*. À fins de comparaison, nous utilisons dans un second temps les voies métaboliques fournies par SGD [Cherry et al., 1998]. Les méthodes que nous développons s'inscrivent dans le cadre de la modélisation soustractive, qui s'intéresse à la perte d'éléments par rapport à une référence, par opposition à la modélisation additive, qui se focalise sur le gain de modules. Notre but est de prédire de manière la plus automatique possible, un ensemble de voies métaboliques représentant les fonctions centrales d'une cellule,

donc supposées très conservées, afin de fournir les données d'entrée à un processus de modélisation du fonctionnement de la cellule.

Dans une première approche, nous avons réalisé une prédiction simple de conservation des voies métaboliques, basée sur une technique de coloriage de graphes. Dans cette approche, nous nous sommes limités à l'étude de la conservation des enzymes dans une voie métabolique sans tenir compte de la structure sous-jacente. Pour aider à la décision de conservation ou de perte d'une voie métabolique, nous avons ajouté des informations sur la nature et l'importance supposée des gènes impliqués. Un tableau de résultat présente pour chaque voie métabolique des statistiques sur la conservation des gènes dans les quatre espèces d'intérêt, accompagnées des annotations des gènes de *S. cerevisiae* ainsi que d'informations sur l'essentialité de ces derniers. En regard des résultats obtenus, nous nous sommes aperçus que la prise de décision de conservation ou de perte d'une voie nécessite une étude en profondeur des résultats produits, et que la simple information de conservation ou perte des gènes n'est pas suffisante pour décider de manière automatique de la conservation d'une voie métabolique. Aussi, nous avons cherché à développer une autre méthode permettant une prise de décision automatique et fiable.

Dans cette seconde approche, nous avons considéré que les informations de conservation des enzymes sont importantes, mais que la conservation des connexions au sein d'une voie métabolique revêt un caractère primordial. Nous représentons les voies métaboliques sous la forme de graphes dirigés, ce qui nous permet de définir des composants d'entrée et de sortie, que nous nommerons *ports* dans la suite. Nous définissons en outre la notion de perte corrélée de ports. Dans une voie métabolique, une réaction peut être perdue si l'enzyme impliquée n'a pas d'homologue chez l'espèce étudiée. Les métabolites impliqués dans la réaction peuvent alors être isolés du reste de la voie. Ainsi, la perte d'une enzyme peut entraîner la perte de métabolites. La perte corrélée de ports, qui sont donc des métabolites particuliers, peut nous fournir des indices sur la conservation des voies métaboliques. Dans cette étude, nous montrons que notre méthode permet de prédire de manière systématique et automatique la conservation de voies métaboliques, et ce dans 60 à 80% des cas. Pour les 20 à 40% de cas nécessitant une annotation manuelle, la prise de décision est rapide, notamment grâce à la représentation en graphe, qui permet de cerner immédiatement les portions de voies métaboliques perdues ou conservées.

4.2 Sources de données

Kyoto Encyclopedia of Genes and Genomes Comme nous l'avons vu à la section 2.2.2, la base de données du KEGG est une source de données de qualité pour les micro-organismes tels que les levures. Pour notre étude, nous utilisons les fichiers KGML (KEGG Markup Language) représentant les voies métaboliques de *S. cerevisiae*. Au 04/12/2006 (version 0.6), ces fichiers regroupent 80 voies métaboliques,

couvrant un ensemble de 1154 gènes, soit environ 1/6 des gènes de *S. cerevisiae*. Pour permettre une visualisation plus aisée des graphes extrapolés, nous utilisons la nomenclature fonctionnelle pour les gènes de *S. cerevisiae*, ainsi que les noms biochimiques des biomolécules fournis par les tables de correspondances du KEGG, disponibles à l'adresse `ftp://ftp.genome.jp/pub/kegg/xml/current/sce`.

Saccharomyces Genome Database Nous avons présenté la base de données de voies métaboliques de SGD à la section 2.2.2. Cette base de données propose des informations sur les voies métaboliques de *S. cerevisiae* que nous jugeons plus sûres que celles fournies par le KEGG. Cependant, cette ressource présente deux désavantages. Le premier est que la couverture en terme de nombre de gènes est inférieure dans les données de SGD. En effet, Ces données concernent 775 gènes, soit environ un tiers de moins que KEGG. Néanmoins, en terme de couverture de voies métaboliques, les données de SGD sont équivalentes à celles du KEGG. Le deuxième désavantage des voies métaboliques de SGD est qu'elles prennent en compte tous les métabolites, tels que les cofacteurs ou les molécules énergétiques. Nous avons vu en section 2.2.1 que cela peut se révéler problématique dans le cadre d'études sur la structure des voies métaboliques et nous reviendrons sur ce point lors de la discussion des résultats. Pour notre étude, nous utilisons les fichiers fournis par SGD au format BioPax niveau 1 [Luciano, 2005]. Au 28/02/2007, SGD regroupe 154 voies métaboliques, couvrant donc 775 gènes.

Les familles de protéines Génolevures Les familles de protéines Génolevures [Nikolski and Sherman, 2007] sont un outil puissant pour identifier des gènes homologues chez les levures hémiascomycètes. Ces familles sont calculées par une combinaison de clustering consensus de similarités, homéomorphiques ou non, de séquences protéiques. En tant que telles, elle sont plus fiables que des alignements BLAST filtrés ou réciproques (RBH [Rivera et al., 1998]). Chaque famille de protéine est décrite par une empreinte phylétique, indiquant la présence ou l'absence de membres pour chaque espèce, ainsi que par un profile phylogénétique qui donne le nombre de membres par espèce. Ces mesures permettent de mettre en évidence des profils particuliers de gain ou perte au long des différentes branches du phylum des hémiascomycètes (Table 2). Nous utilisons la dernière version des familles de protéines, délivrées le 31/01/2006 et disponibles à l'adresse http ://cbi.labri.fr/Genolevures/fam/index.html.

Familles de protéines		Nb familles	Copies #1	Nb protéines
Familles universelles	sckdy	2553	1649	18437
Familles non-universelles		1832	1268	8210
e.g. :				
Spécifiques au clade *Saccharomyces*	sck--	395	355	1258
Perte dans le clade *Saccharomyces*	---dy	283	247	652
Perte chez *Y. lipolytica*	sckd-	264	205	1195
Spécifiques à *D. hansenii*	---d-	92	0	312
Spécifiques à *Y. lipolytica*	----y	123	0	402
Total		4385	2917	26647

Tab. 2 – Profils de conservation des familles de protéines de levures. le *pattern phylétique* indique la présence (s, c, k, d, y) ou l'absence (-) de chaque espèce dans une famille. Les nombres totaux de protéines et de familles sont respectivement indiqués dans la première et la dernière colonne. La colonne Copies #1 indique le nombre de familles pour lesquelles chaque espèce n'a qu'un seul membre. Les familles de protéines ont été calculées pour *S. cerevisiae* (s), *C. glabrata* (c), *K. lactis* (k), *D. hansenii* (d) et *Y. lipolytica* (y).

4.3 Étude comparative de *D. hansenii* et *C. albicans*

4.3.1 Méthode

Dans cette étude, notre but est détablir les différents ensembles de gènes (i.e., gènes codant pour des protéines) pour chaque espèces :

1. Gènes spécifiques à une espèce
2. Gènes spécifiques à une espèce et communs au phylum hémiascomycète
3. Gènes commun aux deux espèces
4. Gènes communs aux deux espèces et au phylum hémiascomycète

Pour déterminer ces ensembles, nous utilisons les séquences annotées de *D. hansenii* (Génolevures [Dujon, 2006]) et les séquences de *C. albicans* fournies par le Broad Insitute[1], ainsi que les familles de protéines Génolevures.

L'idée générale de notre méthode consiste à mimer le processus de création de familles de protéines pour les deux espèces que nous étudions. Pour ce faire, nous allons utiliser les PSSM, pour *Postion Specific Scoring Matrices* [Altschul et al., 1997]. Une PSSM représente la signature d'un ensemble d'alignements. Appliquée aux familles de protéines, les PSSM peuvent donc être vues comme la signature de chaque

[1]http ://www.broad.mit.edu/annotation/genome/candida_albicans

famille. Dans notre étude, nous allons donc utiliser les PSSM calculées sur les familles de protéines Génolevures et le logiciel Blastpgp, pour affecter chaque protéine de *D. hansenii* et *C. albicans* à une famille. Le résultat de cette étape permet de séparer deux premiers ensembles de gènes, ceux spécifiques à chaque espèce et ceux communs aux hémiascomycètes (i.e., gènes affectés à une famille). Il est important de noter ici une différence majeure entre l'affectation d'une protéine à une famille par les PSSM et le même événement par le calcul de familles de protéines. En effet, si les familles de protéines représentent des partitions de l'ensemble des protéines impliquées, l'utilisation de PSSM peut aboutir à l'affectation d'une protéine dans plusieurs familles (voir définition 5). Le résultat de l'affectation par les PSSM n'est donc pas composé d'un ensemble de partitions. Pour notre étude, ce phénomène ne porte pas à conséquence, car le signal qui nous intéresse est uniquement la présence ou l'absence de protéine dans les familles.

Définition 5 *Soit F l'ensemble des familles de protéines : $F = (F_1, ..., F_n)$ où $\bigcap_1^n F_i = \emptyset$*
Soit P l'ensemble des familles PSSM : $P = (P_1, ..., P_m)$ où $\bigcap_1^m P_i \neq \emptyset$

Nous pouvons subdiviser les deux ensembles obtenus en sous-ensembles. Ainsi, à partir de l'ensemble des gènes communs aux hémiascomycètes, nous obtenons trois ensembles intéressants :

- Gènes communs aux hémiascomycètes et à *D. hansenii* et *C. albicans*
- Gènes communs aux hémiascomycètes et à *D. hansenii*, mais pas à *C. albicans*
- Gènes communs aux hémiascomycètes et à *C. albicans*, mais pas à *D. hansenii*

Pour le second ensemble de gènes (i.e., non-retrouvés chez les hémiascomycètes), nous obtenons les trois sous-ensembles suivants :

- Gènes spécifiques à *D. hansenii*
- Gènes spécifiques à *C. albicans*
- Gènes communs aux deux espèces

Pour obtenir ces trois derniers ensembles, nous réalisons des Blast réciproques (RBH [Rivera et al., 1998]) entre les séquences protéiques correspondantes ches les deux espèces. Les RBH identifiés définissent les gènes communs, les autres étant obtenus par soustraction. À fins de vérification, des alignement multiples (logiciel T-Coffee [Notredame et al., 2000]) sont réalisés entre gènes supposés communs. Enfin, pour chacun de ces sous-ensembles, nous réalisons une recherche d'homologues dans la base de données UniProKB [Apweiler et al., 2004], filtrée pour éliminer les séquences des organismes étudiés (*S. cerevisiae*, *C. glabrata*, *K. lactis*, *D. hansenii*, *Y. lipolytica* et *C. albicans*).

Dans le but de caractériser les gènes contenus dans ces trois derniers ensembles, nous utilisons tout d'abord le logiciel InterProScan [Zdobnov and Apweiler, 2001].

Ce logiciel permet de rechercher des signatures de domaines protéiques InterPro dans les séquences protéiques, ainsi que d'attribuer des termes Gene Ontology. Toujours dans le même but de caractérisation, nous étudions la position de ces gènes sur des Analyses en Composante Principale du biais d'utilisation de codon, de même que nous étudions leur répartition sur les chromosomes. Une dernière question que nous nous sommes posée sur ces protéines, est de savoir si elles représentent chacune des protéines différentes ou si elles peuvents être regroupées en familles. Pour ce faire, nous avons réalisé des alignements de ces protéines à l'aide du logiciel Blast, puis nous avons réalisé un clustering markovien pour construire des familles, suivant une procédure similaire à celle expliquée dans [Nikolski and Sherman, 2007].

4.3.2 Résultats

Le premier résultat que nous obtenons est la subdivision des ensembles de gènes de *D. hansenii* et *C. albicans* dans les six ensembles que nous avons précédemment décrits. Sur la figure 25, nous pouvons voir que pour chacune des deux espèces, 5/6 des protéines sont communes au phylum des hémiascomycètes. Parmi le millier de protéines non-conservées chez les hémiascomycètes, seul 1/3 semble commun aux deux espèces étudiées. Si l'on tient compte des filtres sur les alignements, ce chiffre tombe à 1/20 pour le filtre S50R50, qui signifie que les alignements retenus ont 50% d'homologie portant au moins sur 50% de la longueur de la plus longue séquence. Pour deux espèces supposées très proches, telles que *D. hansenii* et *C. albicans*, ce filtre n'est pas très stringent, et nous aurions pu espérer un plus grand nombre de protéines communes aux deux espèces.

Concernant les alignements Blast réalisés sur les ensembles de protéines non-universelles, par opposition aux protéines universellement conservées chez les hémiascomycètes, nous voyons que très peu de protéines présentent un résultat dans la base UniProtKB. En effet, sans appliquer aucun filtre sur les alignements, seuls 15 à 16% des protéines spécifiques à chaque organisme présentent un résultat (figures 25b et 25c). En observant les résultats filtrés, on se rend compte que ce pourcentage descend à 5% avec le filtre S50R50. L'interprêtation de ce résultat nous amène donc à penser que très peu de protéines spécifiques à chaque espèce présentent des homologues dans d'autres espèces en dehors de leur phylum. Pour expliquer la présence de ces protéines dans chaque espèce, nous pouvons évoquer deux hypothèses. La première consiste en la conservation spécifique de ces protéines chez *D. hansenii* ou *C. albicans* selon le cas, ainsi que la perte concomittante de ces mêmes protéines chez les autres espèces du phylum. La seconde hypothèse consiste en l'éventualité de transferts horizontaux de gènes.

Les diverses tentatives de caractérisation des protéines non-universelles de *D. hansenii* et *C. albicans* se montrent peu fructueuses. Pour la prédiction de domaines InterPro et d'attribution de termes GO, peu de protéines sont impliquées

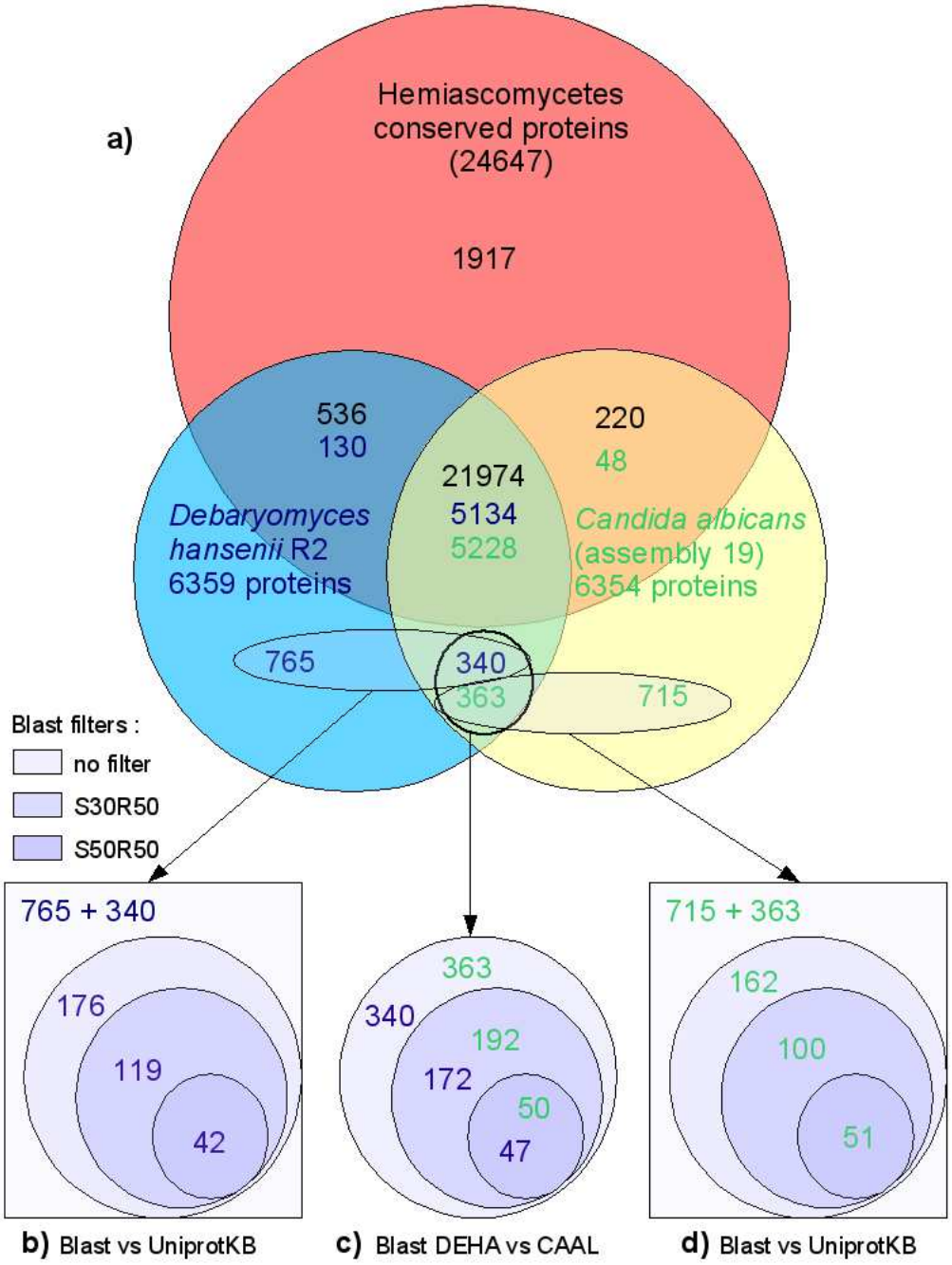


FIG. 25 – Diagramme de Venn montrant a) la répartition des protéines de *D. hansenii* et *C. albicans* parmi les protéines conservées chez les hémiascomycètes, spécifiques à l'espèce ou au clade. Les figures détaillées au bas de l'image indiquent les nombres de résultats Blast distinct pour b) les protéines spécifiques à *D. hansenii* contre UniProtKB, c) les RBH de *D. hansenii* contre *C. albicans* et d) les protéines spécifiques à *C. albicans* contre UniProtKB, le tout en fonction de différents jeux de paramètres de filtrage des alignements. Ces résultats suggèrent qu'un tiers des gènes non-universels de ces deux espèces, sont communs au clade qu'elles représentent, et que deux tiers sont espèce-spécifiques.

dans les résultats et ceux-ci restent très généraux (termes GO de type *catalytic activity* ou encore *metabolism*). De même, la répartition de ces protéines sur des ACP de biais d'utilisation de codons ne montrent pas de caractéristiques particulières à ce niveau. Seule la répartition chromosomique de ces protéines chez *D. hansenii* semble indiquer un biais de représentation aux extrémités des chromosomes (figure 32, annexe A). Un découpage des chromosomes de *C. albicans* n'étant pas disponible, cette étude n'as pas pu être menée.

Le dernier résultat concerne la réalisation de familles des protéines spécifiques à *D. hansenii* et des familles des protéines spécifiques à *C. albicans*. Par cette étude, nous avons pu identifier une centaine de protéines correspondant à une trentaine de familles, tant chez *D. hansenii* que chez *C. albicans*, la plupart de ces familles ne contenant que deux à trois protéines. Pour tenter de caractériser plus avant ces familles, nous avons observé la localisation de leurs membres sur les chromomes. À l'exception notable de la famille 0 chez *D. hansenii* qui forme un cluster à l'extrémité du chromosome E, les autres familles ne présentent pas de biais de localisation chromosomique.

La conclusion que nous pouvons tirer de cette étude est que globalement, plus de 80% des gènes d'une espèce du phylum hémiascomycète sont communs à toutes les espèces de ce phylum. De plus, nous pouvons noter que peu de gènes sont spécifiques à un clade particulier, alors qu'environ 10% des gènes d'une espèce lui sont entièrement spécifiques. Cette comparaison quantitative et qualitative des ensembles de gènes de deux espèces proches nous permet de penser que, dans le cadre de l'extrapolation de voie métabolique chez les hémiascomycètes, nous devrions obtenir un grand nombre de voies conservées, puisque les protéines sont massivement conservées au sein de ce phylum.

4.4 Prédiction de voie métaboliques par coloriage de graphe

4.4.1 Méthode

Dans cette première approche, nous avons développé une méthode simple et rapide, basée sur la coloration de graphes de réactions (cf. section 2.2.3). Pour chaque voie métabolique, nous construisons un graphe non dirigé dans lequel les sommets sont les enzymes et les arcs, les métabolitess. Nous colorions ensuite le graphe en utilisant les familles de protéines. Pour chaque sommet, nous vérifions si un homologue existe dans l'espèce étudiée. Si tel est le cas, nous colorions le sommet pour cette espèce. Une fois ce coloriage réalisé pour chaque espèce, nous pouvons effectuer un calcul du pourcentage de sommets conservés par rapport au nombre de sommets du graphe de référence. Ces informations servent de base à la décision de

conservation de la voie étudiée. Pour faciliter la décision, nous rajoutons différentes informations pouvant se révéler utiles. Ainsi, nous mettons à disposition l'annotation du gène de *S. cerevisiae* et de ses homologues chez les espèces étudiées. Nous présentons aussi les résultats de différentes expériences de délétion systématique de gènes ([Giaever et al., 2002], [Dunn et al., 2004]).

4.4.2 Résultats

Un tableau général disponible en ligne[2] présente une vision globale de la conservation des voies métaboliques, en donnant pour chaque voie un profil de conservation, un lien vers le profil détaillé et un lien vers la voie métabolique de référence. Dans cette vue (figure 26), la couleur d'une espèce indique le pourcentage global de conservation de gènes pour cette espèce dans la voie étudiée. Une page détaillant les résultats de la prédiction de conservation est attachée à chaque voie métabolique. Cette page est divisée en deux parties, la première montrant une vue synthétique de la prédiction (figure 27), la deuxième, une vue détaillée (figure 28). La vue synthétique fournie les informations de conservation pour chaque enzyme dans chaque espèce. Ainsi, pour chaque enzyme, le *pattern phylétique* permet rapidement de voir quelles sont les espèces qui ne présentent pas d'équivalent fonctionnel pour une enzyme donnée. Le profil permet quant à lui d'observer si une espèce montre une expansion ou une contraction au niveau du nombre d'homologues. Une fois ces informations globales prises en compte, le tableau complet des résultats peut permettre d'affiner la décision, en tenant compte par exemple des informations de délétion systématiques de gènes.

Les résultats sur les données du KEGG semblent indiquer que les voies métaboliques étudiées sont globalement bien conservées chez les quatre espèces étudiées. Cette impression est renforcée par les résultats obtenus avec les données de SGD, où il semble que la très grande majorité des voies métaboliques de *S. cerevisiae* soient conservées chez *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica*. Cependant, il est difficile de décider de manière fiable et automatique de la conservation des voies métaboliques. En effet, lorsque toutes les enzymes sont conservées, nous pouvons conclure de manière fiable et automatique que la voie étudiée est conservée. À l'inverse, si aucune enzyme n'est conservée, nous pouvons conclure avec certitude que la voie est perdue. Entre ces cas extrêmes, la décision est délicate. En prenant l'exemple du métabolisme du galactose, connu pour être perdu chez *C. glabrata* ([Bolotin-Fukuhara et al., 2006], [Hittinger et al., 2004]), nous nous apercevons que *C. glabrata* ne conserve qu'environ la moitié des enzymes impliquées dans la voie (2/5 selon SGD, 17/30 selon KEGG). De plus, les gènes perdus n'entraînent pas la létalité de l'organisme selon les informations de délétion systématique des gènes. Il serait donc *a priori* assez raisonnable de conclure à la perte de cette voie métabolique chez *C. glabrata*. Cependant, il est tout à fait envisageable que les enzymes perdues

[2]http ://cbi.labri.fr/Genolevures/path

N.B. : conservation percentage color code 0% 1% to 20% 21% to 40% 41% to 60% 61% to 80% 81% to 90% 91% to 99% 100%

Pathways (link to KEGG)	See conservation details	Synthetic conservation pattern
1,4-Dichlorobenzene degradation	profile	sckdy
ATP synthesis	profile	sckdy
Alanine and aspartate metabolism	profile	sckdy
Aminoacyl-tRNA biosynthesis	profile	sckdy
Aminophosphonate metabolism	profile	sckdy
Aminosugars metabolism	profile	sckdy
Androgen and estrogen metabolism	profile	sckdy
Arginine and proline metabolism	profile	sckdy
Ascorbate and aldarate metabolism	profile	sckdy
Basal transcription factors	profile	sckdy
Benzoate degradation via CoA ligation	profile	sckdy
Benzoate degradation via hydroxylation	profile	sckdy
Bile acid biosynthesis	profile	sckdy
Biosynthesis of steroids	profile	sckdy
Biotin metabolism	profile	sckdy
Blood group glycolipid biosynthesis-neolactoseries	profile	sckdy
Butanoate metabolism	profile	sckdy
C21-Steroid hormone metabolism	profile	sckdy
Carbon fixation	profile	sckdy
Cell cycle	profile	sckdy
Citrate cycle (TCA cycle)	profile	sckdy
Cyanoamino acid metabolism	profile	sckdy
Cysteine metabolism	profile	sckdy
DNA polymerase	profile	sckdy
Fatty acid biosynthesis (path 1)	profile	sckdy
Fatty acid biosynthesis (path 2)	profile	sckdy
Fatty acid metabolism	profile	sckdy
Folate biosynthesis	profile	sckdy
Fructose and mannose metabolism	profile	sckdy
Galactose metabolism	profile	sckdy
Ganglioside biosynthesis	profile	sckdy
Globoside metabolism	profile	sckdy
Glutamate metabolism	profile	sckdy
Glutathione metabolism	profile	sckdy
Glycerolipid metabolism	profile	sckdy

Fig. 26 – Tableau de présentation générale des résultats de la prédiction de voies métaboliques par coloriage de graphe. La première colonne indique le nom de la voie et est associée à un lien donnant accès à la ressource d'origine (voie métabolique KEGG ou SGD selon le cas). Pour chaque voie, un lien permet de consulter le profil de conservation détaillé. La dernière colonne donne le profil de conservation de la voie, où chaque lettre correspond à une espèce et chaque couleur de lettre correspond à un intervalle de pourcentage de conservation d'enzyme. Les espèces ici présentes sont *S. cerevisiae* (s), *C. glabrata* (c), *K. lactis* (k), *D. hansenii* (d) et *Y. lipolytica* (y).

Galactose metabolism

(link to KEGG)

Synthetic pattern of the pathway : sckdy

Reference Protein	EC Number	Genolevures gene families		
		Phyletic pattern	Profile	Family name
YBR184W (YBR184w)	3.2.1.22	s----	1 0 0 0 0	none
PFK1 (YGR240c)	2.7.1.11	sckdy	2 3 2 2 1	GLR.40
PFK2 (YMR205c)	2.7.1.11	sckdy	2 3 2 2 1	GLR.40
ADH7 (YCR105w)	1.1.1.-	sckdy	13 7 11 20 7	GLR.3191
AAD3 (YCR107w)	1.1.1.-	sckdy	8 3 1 4 1	GLR.1984
AAD4 (YDL243c)	1.1.1.-	sckdy	8 3 1 4 1	GLR.1984
YPR1 (YDR368w)	1.1.1.-	sckdy	6 5 7 10 13	GLR.1959
AAD6 (YFL056c)	1.1.1.-	sckdy	8 3 1 4 1	GLR.1984
AAD10 (YJR155w)	1.1.1.-	sckdy	8 3 1 4 1	GLR.1984
FOX2 (YKR009c)		sckdy	2 2 6 16 13	GLC.29
AAD14 (YNL331c)	1.1.1.-	sckdy	8 3 1 4 1	GLR.1984

FIG. 27 – Extrait du tableau de synthèse des résultats de la prédiction de conservation pour la voie métabolique du galactose. La première colonne indique le nom de la protéine, la seconde le numéro enzymatique associé et la troisième, les informations en relation avec les familles de protéines, à savoir le *pattern phylétique*, le profil et le nom de la famille associée. Un rappel du profil de conservation globale des enzymes est présenté au dessus du tableau.

ne fassent partie que de voies alternatives à la voie principale du métabolisme du galactose et que dans ce cas, cette fonction soit préservée chez *C. glabrata*. En regard de ces résultat, nous observons que notre méthode obtient des résultats nous permettant de statuer de manière fiable et automatique sur les cas de conservation ou perte totale, mais pas sur les cas intermédiaires. Une étude manuelle et approfondie permet de prendre une décision dans les situations ambiguës. Notre méthode fournie cependant un moyen simple et efficace de parcourir le contenu de grandes bases de données, telles que celles concernant les voies métaboliques. Ce moyen de navigation est actuellement mis en oeuvre dans le cadre du projet Génolevures, intégré au site web du projet [Sherman et al., 2006].

Nous avons donc constaté que notre méthode de prédiction de voies métaboliques, basée sur un simple calcul de pourcentage de gènes conservés, nous permet d'avoir un premier aperçu de la conservation des voies métaboliques chez les espèces que nous avons étudiées. Cette méthode ne nous permet cependant pas d'exécuter la tâche de prédiction de manière automatique et fiable, afin de fournir une base au processus de modélisation. De ce constat, nous définissons la nécessité de prendre en compte l'information structurale des voies métaboliques, afin de pouvoir prendre une décision fiable et automatique sur leur conservation.

Detailed view

Reference Protein	EC Number	Annotation	Systematic gene deletion	Gene deletion effects (growth rate)					Genolevures gene families				
				20 gen. in YPD	60 gen. in YPD	YNB	YPD + NaCl	YP + lactate	Phyletic pattern	Profile	Family name	Gene Name	Gene Annotation
YBR184W (YBR184W)	3.2.1.22	-	viable	0.1	0.2	Ess	0.3	Ess	s----	1 0 0 0 0	none	-	-
PFK1 (YGR240C)	2.7.1.11	sp|P16861 Saccharomyces cerevisiae YGR240c PFK1 6-phosphofructokinase, alpha subunit P2.193.f2.1	Exhibits growth defect on a fermentable carbon source.	0.4	Ess	Ess	0.5	Ess	sckdy	2 3 2 2 1	GLR.40	YALI0D16357g	sp|P59680 Yarrowia lipolytica 6-phosphofructokinase, identified start
												DEHA0D11132g	highly similar to sp|O94200 Candida albicans 6-phosphofructokinase beta subunit (EC 2.7.1.11) (Phosphofructokinase 1) (Phosphohexokinase) (6PF-1-K beta subunit), start by similarity
												CAGL0L10758g	highly similar to sp|P16862 Saccharomyces cerevisiae YMR205c PFK2 or sp|P16861 Saccharomyces cerevisiae YGR240c PFK1, start by similarity
												DEHA0G22088g	highly similar to sp|O94201 Candida albicans PFK1 6-phosphofructokinase alpha subunit (Phosphofructokinase 1) (Phosphohexokinase) (6PF-1-K alpha subunit), start by similarity
												KLLA0F06248g	gi|417202|sp|Q03216|K6P2_KLULA Kluyveromyces lactis 6-phosphofructokinase beta subunit (Phosphofructokinase 2) (Phosphohexokinase) (6PF-1-K beta subunit), start by similarity
												CAGL0F08041g	highly similar to sp|P16861 Saccharomyces cerevisiae YGR240c PFK1 6-phosphofructokinase, start by similarity
												KLLA0A05544g	gi|417201|sp|Q03215|K6P1_KLULA Kluyveromyces lactis 6-phosphofructokinase alpha subunit (Phosphofructokinase 1) (Phosphohexokinase) (6PF-1-K alpha subunit), start by similarity
												CAGL0I05698g	similar to sp|P16862 Saccharomyces cerevisiae YMR205c PFK2 or sp|P16861 Saccharomyces cerevisiae YGR240c PFK1, hypothetical start

Fig. 28 – Extrait du tableau de synthèse détaillée des résultats de la prédiction de conservation pour la voie métabolique du galactose. La première colonne indique le nom de la protéine, la seconde le numéro enzymatique associé et la dernière, les informations en relation avec les familles de protéines, à savoir le *pattern phylétique*, le profil, le nom de chaque orf pour chaque espèce et le nom de la famille associée. Les colonnes du milieu donnent les résultats des expériences de délétion systématique du gène chez *S. cerevisiae*.

4.5 Prédiction de voies métaboliques par analyse structurale de graphes

4.5.1 Méthode

Dans cette seconde approche, nous développons une méthode de prédiction de voies métaboliques basée sur l'analyse structurale de graphes. Pour chaque voie métabolique de référence, nous construisons un graphe bipartite dirigé où les sommets sont les métabolites et les enzymes, et les arêtes sont les relations entre enzymes et métabolites (voir section 2.2.3). Ces graphes sont notre référence pour la tâche de prédiction. Pour chaque espèce et chaque voie métabolique, nous élaguons le graphe de référence en utilisant les relations d'homologie définies par les familles de protéines Génolevures : pour chaque enzyme, si un équivalent fonctionnel existe dans l'espèce étudiée, le sommet correspondant est conservé. Si cet équivalent n'existe pas, le sommet est supprimé. Nous obtenons alors un graphe contenant tous les métabolites du graphe initial, mais seulement les enzymes ayant un homologue dans l'espèce étudiée. Le déroulement de ces étapes est expliqué par l'algorithme 6.

Nous définissons le graphe G comme graphe de référence, le graphe G_s comme graphe de l'espèce étudiée, et l'ensemble F représentant les familles de protéines, où

chaque familleF_i est composée d'un ensemble de protéines $p_1, ..., p_n$. Nous définissons ensuite les fonctions $sommets(e)$, qui retourne les deux sommets reliés par l'arête e, $est\_enzyme(v)$, qui retourne vrai si v est une enzyme, et enfin $membres(F, v)$ qui retourne vrai si v appartient à une famille de protéines contenant au moins une protéines de *S. cerevisiae*. Le résultat de cet algorithme est un graphe $G_s = \langle V_s, E_s \rangle$ où $V_s \subseteq V$ et $E_s \subseteq E$.

Algorithme 6 *Coloriage de graphe*

Entrées: $G = \langle V, E \rangle$, $G_s = \langle \emptyset, \emptyset \rangle$, $F = [F_1, ..., F_n]$ où $F_i = [p_1, ..., p_n]$

```
pour tout e ∈ E faire
  pour tout v ∈ sommets(e) faire
    si est_enzyme(v) && membres(F, v) alors
      V_Gs = V_Gs ∪ v
      E_Gs = E_Gs ∪ e
    sinon si !est_enzyme(v) alors
      V_Gs = V_Gs ∪ v
    finsi
  fin pour
fin pour
Retourner G_s
```

À partir de ce graphe prédit, nous extrayons différentes mesures. Nous cherchons tout d'abord à évaluer la conservation des métabolites d'entrée et de sortie. Dans la suite, le terme *port* désignera indistinctement les métabolites d'entrée et de sortie. Utilisant un graphe dirigé, nous définissons les ports en termes de théorie des graphes, où les entrées sont les sommets n'ayant aucune arête entrante, alors que les sorties sont les sommets n'ayant aucune arête sortante. Il faut ici noter que dans le cas des voies métaboliques, un port correspond obligatoirement à un métabolite, puisqu'une enzyme est toujours un élément intermédiaire entre deux métabolites et possède à ce titre au moins une arête sortante et une arête entrante. La méthode la plus directe pour calculer la conservation des ports est de vérifier si les chemins qui existent entre eux dans le graphe de référence, sont conservés dans le graphe prédit. Une autre méthode, de complexité très inférieure, est de vérifier si les ports sont toujours dans les mêmes composantes faiblement connexes [Diestel, 2005]. Un graphe dirigé est dit faiblement connexe si pour toute paire de sommets (u,v), il y a un chemin entre u et v, une fois que les arêtes orientées ont été remplacées par des arêtes non-orientées. Si un graphe n'est pas connexe, il est décomposé en différentes composantes, elles-mêmes connexes ou non.

Ainsi dans l'exemple de la figure 29, nous pouvons voir que la composante connexe A est séparée en trois composantes connexes dans le graphe prédit. L'entrée I1 est toujours dans la même composante connexe que la sortie O1, mais est dissociée de la sortie O2. De plus, I2 et O2 sont respectivement isolées de toute sortie ou entrée. Nous pouvons donc conclure pour cette espèce que la voie métabolique

qui produit le métabolite O1 à partir du métabolite I1 est conservée, alors que la voie qui produit O1 à partir de I2 semble perdue, de même que toute réaction transformant I1 ou I2 en O2. La composante connexe B montre que toutes les entrées sont déconnectées des sorties. Dans ce cas, nous pouvons dire que la voie métabolique B semble perdue dans l'espèce étudiée puisqu'il n'y a aucun moyen de transformer un métabolite d'entrée en métabolite de sortie. Dans le cas A, il n'y a pas de conclusion claire et immédiate sur la perte globale, puisqu'une partie de la voie est conservée.

Nous cherchons donc à observer la perte corrélée d'entrées et de sorties, signal fort d'une perte possible de voie métabolique. Afin d'accroître le contenu informatif des résultats, nous avons ajouté une contrainte sur les ports. En effet, les voies métaboliques que nous utilisons sont issues d'un découpage effectué par des experts biologistes. Ces voies métaboliques font cependant partie d'un réseau plus vaste liant les différentes voies, par l'intermédiaire des métabolites communs. À l'échelle de la cellule, les différentes voies métaboliques sont donc interconnectées et il nous est possible de définir les métabolites d'interconnexion, qui sont donc à la fois entrée dans une voie métabolique et sortie dans une autre. Nous émettons l'hypothèse que ces composés présentent un intérêt supplémentaire par rapport aux autres ports, puisqu'ils effectuent la liaison entre plusieurs fonctions distinctes. Nous calculons donc, en plus de la conservation globale des ports, la conservation de ces ports d'interconnexion. Dans ce même but d'enrichissement du contenu informatif, nous calculons le nombre de métabolites isolés dans le graphe prédit.

La prise en compte de ces différents critères et l'étude des résultats nous permettent de définir des règles de bon sens, qui vont permettre une décision fiable et automatique relative à la perte ou la conservations des voies métaboliques, et qui nous donnent les résultats contenus dans les tableaux de synthèse (tables 3 et 5). Comme pour la prédiction par coloriage de graphe, nous considérons que les cas extrêmes sont évidents, à savoir une conservation de la voie métabolique si toutes les enzymes sont conservées, et une perte si toutes les enzymes sont perdues. Entre ces deux zones, nous définissons quatre paliers :

- Conservation possible : moins de 10% de perte de ports
- Cas ambigus : de 10 à 30% de ports perdus
- Perte possible : de 30 à 50% de ports perdus
- Perte probable : plus de 50% de perte de ports

D'autres critères nous sont utiles pour l'analyse manuelle des résultats qu'il est possible d'effectuer après la classification automatique que nous venons de présenter. Ainsi, dans les cas de perte possible ou probable, nous avons souvent conclu à la perte de la voie si la moitié ou plus des métabolites se retrouvent isolés dans la voie prédite.

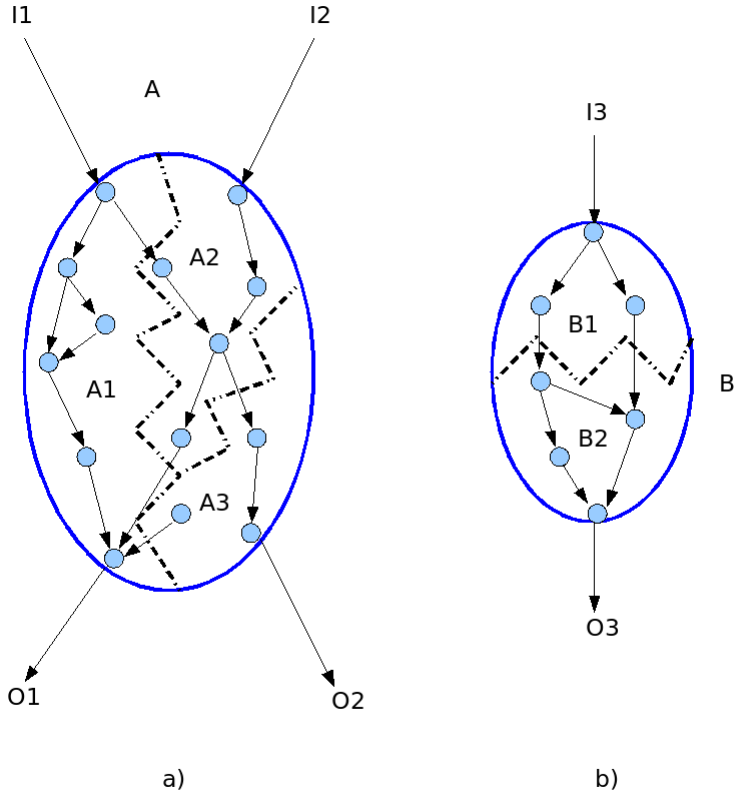


FIG. 29 – Exemple de composantes connexes prédites et de comptage de ports. Les lignes bleues représentent les composantes connexes dans l'organisme de référence *S. cerevisiae*. Les lignes en pointillés représentent les composantes connexes prédites pour l'organisme étudié. Ix et Ox indiquent respectivement les métabolites d'entrée numéro x et les métabolites de sortie numéro x. La composante A1 conserve une relation entrée/sortie pour la voie A. Les composantes A2 et B1 ont des entrées mais n'ont pas de sorties, tandis que les composantes A3 et B2 ont des sorties mais pas d'entrée. Nous concluons que la voie 1 est partiellement conservée (de I1 à O1), alors que la voie B est totalement perdue.

Pour implémenter notre méthode, nous avons développé un parseur XML en C++ qui prend en charge les voies métaboliques au format KGML ou au format BioPax pour les données de SGD. Nous avons aussi développé un script Perl, utilisant la bibliothèque Perl Graph [Orwant et al., 1999], qui prend en charge les informations extraites par le parseur pour construire le graphe de référence, extrapoler les graphes pour les quatre espèces et extraire les informations utiles à la prise de décision de conservation de la voie.

4.5.2 Résultats

Pour chaque voie métabolique nous obtenons un tableau et cinq graphes, un pour *S. cerevisiae* et un pour chaque espèce étudiée. Le tableau contient les informations concernant la conservation des ports et des métabolites connectés pour chaque composante connexe de chaque espèce.

Résultats de la prédiction à partir des données du KEGG Le tableau 3 montre que notre méthode prédit la conservation de plus de la moitié (46 voies conservées sur 80 étudiées) des voies métaboliques de *S. cerevisiae* chez les quatre autres espèces étudiées. Parmi ces voies universellement conservées (table 11, annexe A), nous retrouvons de manière attendue le cycle TCA, la voie des pentoses phosphates, la biosynthèse des acides gras et le métabolisme du pyruvate. Nous ne retrouvons cependant pas la glycolyse. En effet, notre méthode prédit la perte d'un port chez *C. glabrata* et *D. hansenii*, si bien que la voie n'est pas considérée immédiatement comme universellement conservée. Il apparaît que ces deux espèces n'ont pas d'homologue pour le gène de *S. cerevisiae* codant la protéine impliquée dans la transformation de l'éthanol en acétaldéhyde, réaction périphérique de la voie de la glycolyse définie par le KEGG. Nous pouvons donc conclure après analyse que la glycolyse est conservée chez *C. glabrata* et *D. hansenii*, ce qui est conforme aux connaissances biologiques, bien que ces espèces ne pourront peut être pas pousser sur éthanol. La même analyse permet de tirer les mêmes conclusions pour plusieurs autres voies métaboliques qui sont considérées comme non-universellement conservées à la vue des résultats bruts. Parmi ces voies, nous trouvons les métabolismes des purines, des pyrimidines, de l'aspartate, de l'arginine et de la proline.

Un seul cas suggère clairement la perte d'une voie métabolique et trois autres suggèrent une perte possible. L'analyse de ces cas révèle que la voie métabolique perdue est le métabolisme du galactose chez *C. glabrata* et trois des cas de perte possible concernent cette même voie chez les trois autres levures étudiées (tableau 4). Pour *C. glabrata*, deux métabolites sources sont conservés parmi les sept présents chez *S. cerevisiae*, de même que deux sorties sont conservées parmi les neuf de la voie de référence. Nous observons aussi que 18 métabolites parmi les 24 de *S. cerevisiae* se retrouvent isolés dans le graphe prédit. Conformément aux critères de décision

Voies métaboliques	Pattern phylétique	Nombre de voies			
		CAGL	KLLA	DEHA	YALI
Universellement conservées	sckdy	46	46	46	46
Non-universelles	11	25	13	20	
Perte dans 1 espèce					
	sckd-	3	3	3	0
	sck-y	8	8	0	8
	s-kdy	0	4	4	4
Perte dans 2 espèces					
	s-kd-	0	4	4	0
	s-k-y	0	6	0	6
	s--dy	0	0	1	1
Perte dans 3 espèces					
	s--d-	0	0	1	0
	s---y	0	0	0	1
Conservation possible	plusieurs	10	4	12	7
Cas ambigus	plusieurs	7	3	5	4
Perte possible	plusieurs	4	0	2	1
Perte probable	plusieurs	2	2	2	2
Total		80	80	80	80

TAB. 3 – Profils de conservation des voies métaboliques du KEGG suite à la prédiction par analyse structurale de graphes. Le *pattern phylétique* indique la présence (s, c, k, d, y) ou l'absence (-) de chaque espèce dans une voie. Les résultats montrent que la moitié des voies métaboliques sont conservées universellement. La prédiction a été faite pour *C. glabrata* (c, CAGL), *K. lactis* (k, KLLA), *D. hansenii* (d, DEHA) et *Y. lipolytica* (y, YALI). Ce tableau montre que 46 voies métaboliques sont conservées sur les 80 étudiées, soit 57.5%.

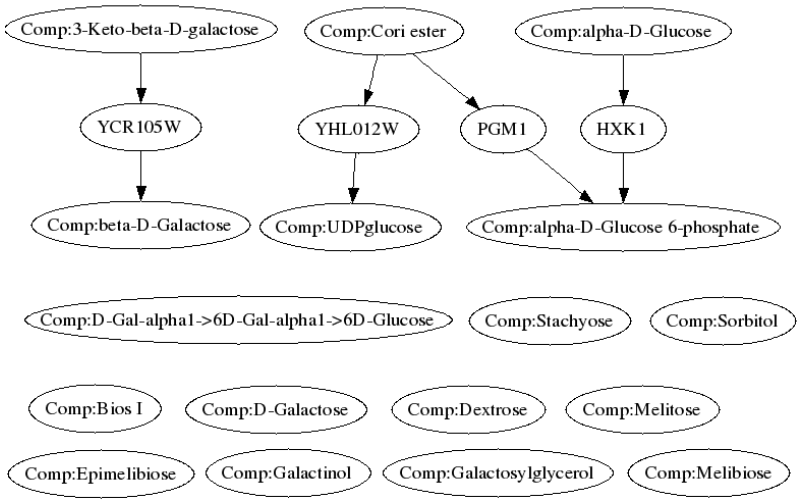


Fig. 30 – Graphe du métabolisme du galactose prédit chez *C. glabrata* (données KEGG). Cet organisme semble conserver le moyen d'entrer dans la voie de la glycolyse en utilisant l'UDPglucose comme métabolite source, mais ne peut transformer le D-Galactose en UDPglucose. Ce graphe nous permet de conclure à la perte du métabolisme du galactose.

que nous avons établis, ces résultats indiquent que la voie métabolique semble perdue chez *C. glabrata*. En analysant ces résultats plus précisément, nous voyons que *C. glabrata* conserve l'enzyme permettant la transformation de l'α-D-Glucose pour entrer dans la glycolyse (figures 30 et 31), mais qu'il n'y a plus aucun moyen de produire ce composé, de même que l'UDPGlucose, à partir du galactose. Nous observons par ailleurs que la sous-voie concernant la conversion du raffinose (mélitose) en D-glucose, D-galactose et D-fructose est complètement perdue (figure 33 annexe A). À la vue de ces résultats, nous pouvons donc confirmer que la voie métabolique du galactose est perdue chez *C. glabrata*. Concernant *K. lactis*, *D. hansenii* et *Y. lipolytica*, un simple regard aux graphes prédits (figures 34, 35 et 36 annexe A) révèle que, malgré la perte de quelques enzymes, le D-galactose peut toujours être utilisé et la conversion du raffinose (mélitose) est maintenue. Ces résultats sont conformes avec des résultats précédemment publiés et qui mentionnent la perte du métabolisme du galactose chez *C. glabrata* [Bolotin-Fukuhara et al., 2006].

Résultats de la prédiction à partir des données de SGD Le tableau 5 montre que notre méthode prédit la conservation de 80% des voies métaboliques de *S. cerevisiae* chez les quatre espèces étudiées. Parmi ces voies universellement conservées (table 10, annexe A), nous retrouvons de manière attendue la glycolyse, la voie des pentoses phosphates, le cycle TCA, la synthèse et la dégradation des acides aminés. Cependant, certaines voies attendues n'apparaissent pas comme universellement conservées à la vue des résultats bruts. Ces voies sont notamment la

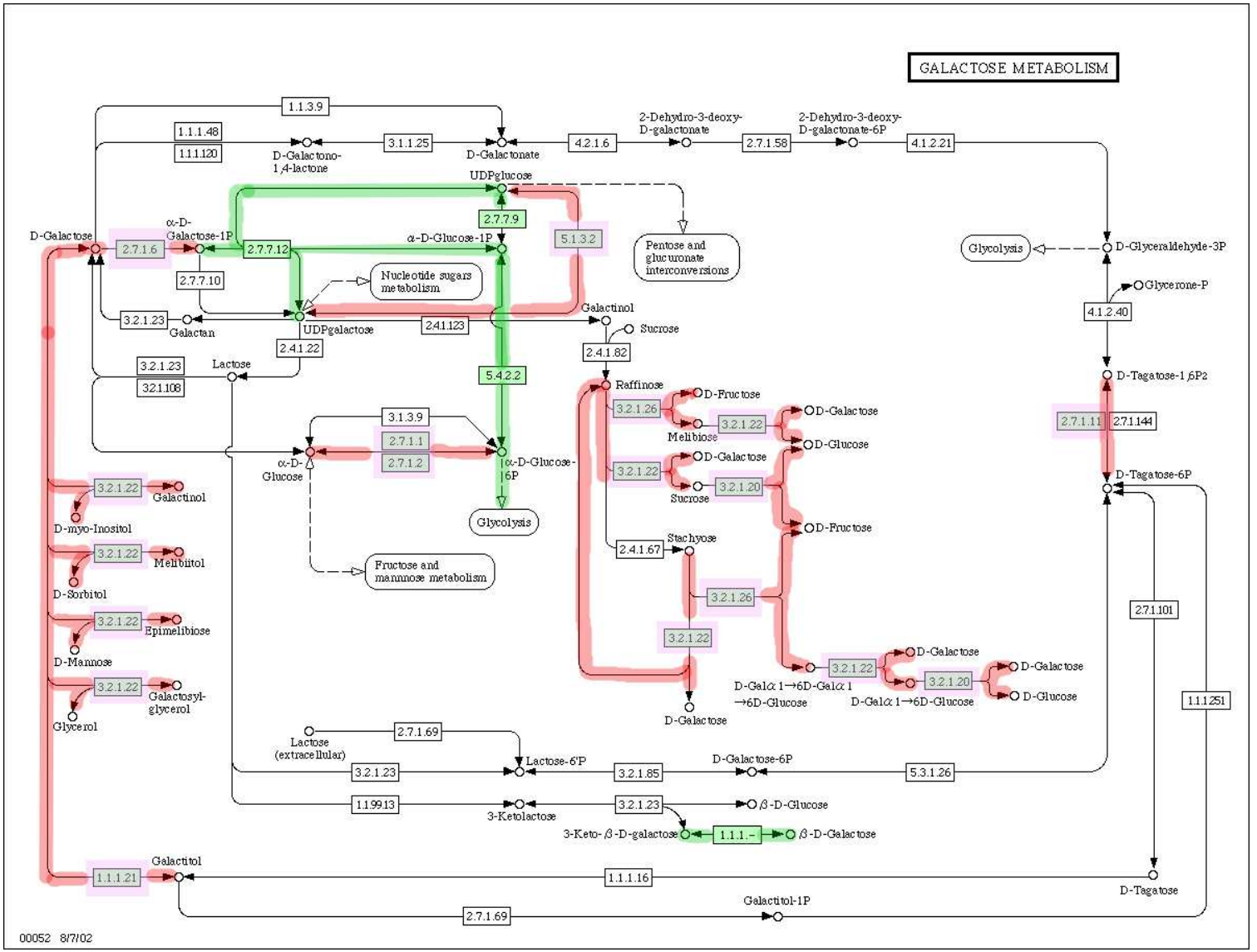


Fig. 31 – Correspondance de la voie métabolique du galactose prédite chez *C. glabrata* avec celle de *S. cerevisiae*. Les traits verts correspondent aux chemins conservés chez *C. glabrata*, alors que les traits rouges correspondent aux chemins perdus par rapport à *S. cerevisiae*. Ce graphe nous permet de conclure à la perte du métabolisme du galactose.

CC Espèce	CC SACE	Entrées Espèce	Entrées SACE	Sorties Espèce	Sortie SACE	Métabolites isolés	% ports perdus
CAGL2	0	1	6	1	8	18/24	75%
CAGL0	1	1	1	1	1	18/24	
KLLA0	0	1	6	2	8	8/24	73%
KLLA2	0	1	6	2	8	8/24	
KLLA1	1	1	1	1	1	8/24	
DEHA0	0	1	6	2	8	8/24	73%
DEHA2	0	1	6	2	8	8/24	
DEHA1	1	1	1	1	1	8/24	
YALI1	0	1	6	1	8	10/24	76.7%
YALI2	0	1	6	2	8	10/24	
YALI0	1	1	1	1	1	10/24	

Tab. 4 – Conservation des ports dans la voie métabolique du galactose (données KEGG) prédite chez les quatre espèces. Chaque ligne représente une composante connexe dans l'espèce indiquée dans la première colonne. CC Espèce indique l'identifiant de la composante connexe dans l'espèce étudiée, correspondant à la composante connexe chez *S. cerevisiae*, désignée dans la colonne CC SACE. Ces résultats suggèrent la perte de la voie métabolique du galactose chez *C. glabrata*, ainsi qu'une perte possible chez les trois autres espèces.

biosynthèse des lipases, du NAD, de l'arginine, de la lysine et du glutamate. Pour le NAD, d'autres voies de biosynthèse existent et sont universellement conservées, de même que pour la biosynthèse et la dégradation des acides gras, ainsi que pour le glutamate.

La voie de biosynthèse de l'arginine semble perdue chez *Y. lipolytica*. En regardant le graphe représentant la voie métabolique prédite, nous nous rendons compte que *Y. lipolytica* n'a pas d'homologue identifié pour le gène ARG1 de *S. cerevisiae* mais conserve le moyen de transformer le L-arginosuccinate en L-arginine. Or, le gène ARG1 est responsable de la transformation du L-aspartate en L-arginosuccinate. On peut donc en déduire que la biosynthèse d'arginine par cette voie est perdue chez *Y. lipolytica*. Sachant que l'arginine est un acide aminé essentiel ou conditionnellement essentiel chez les mammifères [Tapiero et al., 2002], nous sommes étonnés que cette voie soit perdue chez *Y. lipolytica*. En fait, plusieurs voies métaboliques peuvent permettre la production d'arginine et notamment le cycle de l'urée tel qu'il apparaît dans les voies de dégradation de l'arginine, qui sont universellement conservées. Il est donc possible que la voie de biosynthèse de l'arginine soit perdue sans que cela soit létal pour l'organisme. Cette hypothèse est renforcée par le fait que chez *S. cerevisiae*, la délétion du gène ARG1 n'entraîne qu'un retard de croissance à partir de la quinzième génération lorsque la levure est cultivée sur milieu minimum [Giaever et al., 2002].

Voies métaboliques	Pattern phylétique	Nombre de voies			
		CAGL	KLLA	DEHA	YALI
Universellement conservées	sckdy	126	126	126	126
Non-universelles	11	25	15	17	
Perte dans 1 espèce					
	sckd-	4	4	4	0
	sck-y	3	3	0	3
	s-kdy	0	9	9	9
Perte dans 2 espèces					
	s-kd-	0	2	2	0
	s-k-y	0	3	0	3
	sck--	4	4	0	0
Perte dans 3 espèces					
	s---y	0	0	0	2
Conservation possible	plusieurs	1	0	5	3
Cas ambigus	plusieurs	1	0	4	4
Perte possible	plusieurs	6	2	2	1
Perte probable	plusieurs	9	1	2	3
Total		154	154	154	154

TAB. 5 – Profils de conservation des voies métabolique de SGD suite à la prédiction par analyse structurale de graphes. Le *pattern phylétique* indique la présence (s, c, k, d, y) ou l'absence (-) de chaque espèce dans une voie. Les résultats montrent que 80% des voies métaboliques sont conservées universellement. La prédiction a été faite pour *C. glabrata* (c, CAGL), *K. lactis* (k, KLLA), *D. hansenii* (d, DEHA) et *Y. lipolytica* (y, YALI). Nos résultats prédisent en outre la perte probable de 15 voies métaboliques.

La voie de biosynthèse de la lysine semble perdue chez *D. hansenii*. En regardant le graphe représentant la voie métabolique prédite, nous concluons rapidement qu'il n'en est rien. En effet, les familles de protéines Génolevures ne permettent pas d'identifier chez *D. hansenii* un homologue au gène LYS2 de *S. cerevisiae*. Chez *S. cerevisiae*, la protéine codée par ce gène est impliquée dans la dégradation du L-2-aminoadipate-6-semialdéhyde en L-2-aminoadipate. Chez *D. hansenii*, ce gène ne semble pas exister mais la voie métabolique principale est conservée. Nous pouvons donc en conclure que la voie de biosynthèse de la lysine est conservée chez *D. hansenii*.

Le tableau 5 montre que notre méthode prédit la perte probable de 15 voies métaboliques, principalement chez *C. glabrata*. Parmi ces voies, nous trouvons notamment les voies de dégradation du sucrose, du lactose, du tryptophane et le métabolisme du galactose. Concernant la dégradation du sucrose, *C. glabrata* ne présente pas d'homologue au gène SUC2 de *S. cerevisiae* en charge de la dégradation. Cette voie est donc logiquement prédite comme perdue. La voie de dégradation du lactose et la voie du métabolisme du galactose ont en commun la plupart de leurs

CC Espèce	CC SACE	Entrées Espèce	Entrées SACE	Sorties Espèce	Sortie SACE	Métabolites isolés	% ports perdus
CAGL0	0	0	3	1	4	7/9	85.7%
KLLA0	0	3	3	4	4	0/9	0%
DEHA0	0	3	3	4	4	0/9	0%
YALI0	0	3	3	4	4	0/9	0%

Tab. 6 – Conservation des ports dans la voie métabolique du galactose (données SGD) prédite chez les quatre espèces. Chaque ligne représente une composante connexe dans l'espèce indiquée dans la première colonne. CC Espèce indique l'identifiant de la composante connexe dans l'espèce étudiée, correspondant à la composante connexe chez *S. cerevisiae*, désignée dans la colonne CC SACE. Ces résultats suggèrent la perte de la voie métabolique du galactose chez *C. glabrata*.

enzymes et métabolites. Le tableau 6 montre les résultats de la prédiction de la voie métabolique du galactose. On peut voir chez *C. glabrata*que la totalité des métabolites d'entrée est perdue et qu'il ne reste plus qu'un seul métabolite de sortie. Des résultats similaires sont observés pour la voie de dégradation du lactose (table 9 annexe A). Dans les deux cas, le port de sortie principal de la voie est le glucose-6-phosphate et le port principal d'entrée est l'α-D-galactose. Chez *C. glabrata*, les gènes codant les protéines impliquées dans la conversion de l'α-D-galactose ne sont plus présents. Il y a donc une perte totale de la capacité à dégrader le galactose chez *C. glabrata*, et donc incidemment, une perte totale des voies métaboliques de dégradation du lactose et de métabolisme du galactose. Ces résultats, comme ceux obtenus avec les données du KEGG sont conformes aux connaissances issues de la littérature [Hittinger et al., 2004].

4.6 Discussion

4.6.1 De l'intérêt biologique des résultats

Les différentes méthodes de prédiction que nous avons développées montrent, comme nous l'attendions après notre étude comparative de *D. hansenii* et *C. albicans*, que la plupart des voies du métabolisme central sont conservées au sein d'un phylum contenant des espèces relativement distantes [Dujon, 2006]. Nous pensons que ces voies définissent un ensemble de fonctions supposées universelles au sein des eucaryotes, fournissant une base minimale pour le processus de modélisation des organismes. Parmi ces voies du métabolisme central, nous trouvons le cycle TCA, la voie des pentoses phosphates, la glycolyse, les métabolismes du pyruvate, des purines, des pyrimidines et des acides aminés. Notre approche nous permet aussi de détecter la perte d'une voie métabolique due à l'adaptation d'un organisme à une niche écologique, comme la perte du métabolisme du galactose chez *C. glabrata*.

Une spécialisation par gain de modules fonctionnels ? Les résultats complets, que ce soit par la méthode de prédiction par coloriage de graphes[3], ou bien par la méthode de prédiction par analyse structurale de graphes (tables 10 et 11, annexe A), sont en accord avec l'observation qu'il existe peu de différences entre les cinq levures étudiées, bien que celles-ci soient assez distantes d'un point de vue phylogénétique [Dujon et al., 2004]. Cette observation nous permet de suggérer que la différence visible entre les phénotypes pourraient être due au gain de modules fonctionnels et de voies métaboliques. Pour vérifier cette hypothèse, il serait intéressant de pouvoir mener les mêmes analyses en utilisant comme référence un organisme ne faisant pas partie du phylum des hémiascomycètes. Pour ce faire, il faudrait choisir un organisme dont les données génomiques sont disponibles, de même que les données de voies métaboliques. De plus en plus de données de voies métaboliques étant disponibles pour un nombre croissant d'espèces, ce dernier point ne devrait pas poser de difficultés. Le principal obstacle à la réalisation de cette étude se trouve du côté de l'identification des équivalents fonctionnels. En effet, nous prenons comme base départ la confiance que nous pouvons avoir dans les données disponibles. Ainsi, nous avons vu en section 2.2.2 que les différentes sources de voies métaboliques sont plus ou moins fiables. Pour l'identification d'équivalents fonctionnels, nous nous basons de manière exclusive sur les familles de protéines Génolevures [Nikolski and Sherman, 2007], car nous pensons qu'elles sont le moyen le plus fiable d'identification d'homologues, plus robuste que de simple alignements Blast réciproques [Rivera et al., 1998]. Or, ces familles de protéines ne sont disponibles que pour les cinq espèces que nous avons étudiées. Aussi, pour appliquer nos méthodes à d'autres phyli il sera nécessaire, soit d'utiliser d'autres méthodes d'identification d'homologues, soit d'appliquer à ces autres phyli la méthode de calcul des familles de protéines.

Métabolites annexes et robustesse des voies métaboliques Les résultats de l'analyse strucutrale de graphes sur les données du KEGG (tableau 11, annexe A) nous permettent d'observer que les voies centrales, universellement conservées, ont souvent un degré sortant moyen (Définition 6) supérieur à celui des autres voies. Ce résultat suggérerait que les voies centrales sont plus connectées et donc moins enclines à la déconnexion suite à la perte d'une enzyme, que les voies périphériques qui sont plus linéaires. Cette observation est conforme avec la propriété générale des voies métaboliques, qui font parties des réseaux petit monde (cf. section 2.2.1). Cependant, cette caractéristique a trait au réseau métabolique global, non à une voie métabolique particulière ou à toute et chacune voie métabolique. Nous pouvons ainsi poser l'hypothèse que dans le réseau métabolique global, le caractère *scale free* est dû à ces voies du métabolisme central. Pour tester cette hypothèse, il pourrait se révéler intéressant de pratiquer un clustering sur le réseau métabolique global et de rechercher une correspondance entre les différents clusters et les voies métaboliques centrales.

[3] http ://cbi.labri.fr/Genolevures/path

Définition 6 *Le degré sortant d'un sommet* s *est le nombre d'arêtes dont l'extrémité initiale est ce même sommet, en opposition au degré entrant qui est le nombre d'arêtes dont l'extrémité initiale est le sommet* s. *Par voie de conséquence, le degré moyen sortant d'un graphe est la moyenne des degrés sortants des sommets du graphe.*

Le degré moyen sortant supérieur des voies du métabolisme central ne se retrouve pas dans les résultats calculés à partir des données fournies par SGD. Nous pouvons formuler une hypothèse qui peut expliquer cette différence dans les résultats. Comme nous l'avons vu en section 2.2.2, les réactions biochimiques encodées dans les voies métaboliques du KEGG ne tiennent compte que des métabolites principaux et n'incluent pas les cofacteurs, tels que l'oxygène, l'hydrogène, ou bien encore le NADH. À l'opposé, les données fournies par SGD tiennent compte de tous ces cofacteurs. La comparaison des moyennes et médianes des degrés moyens sortants et entrants semble confirmer notre hypothèse. En effet, la médiane des degrés moyens entrants et sortants pour les données du KEGG est de 1.2 alors qu'elle est de 1.6 pour les données SGD, ce qui indique que la moitié des réactions impliquerait trois à quatre métabolites.

Différentes recherche ([Horne et al., 2004] et [Ma and Zeng, 2003]) ont montrées que certains métabolites peuvent être évincés des réactions biochimiques, sans changer la validité de la réaction ni la topologie du graphe, et qu'ils doivent être supprimés des voies métaboliques si des études sont menées sur les propriétés structurales des graphes, telle que la longueur moyenne des chemins ou la connexité moyenne. Nous avons dans un premier temps évincé différents métabolites des données SGD, tels que l'eau et l'oxygène, sans toutefois éliminer des composants tels que l'ATP ou le NAD. Étant donné les résultats obtenus sur la connexité moyenne, il semble qu'il faille donc aller plus loin dans le filtrage des métabolites. [Horne et al., 2004], de même que [Ma and Zeng, 2003] proposent une liste de molécules devant être éliminées des graphes de voies métaboliques avant toute étude globale, ainsi qu'une méthode pour l'utilisation de cette liste. Selon ces auteurs, ces molécules ne doivent pas être complètement supprimées, seules les connexions non utiles devant être éliminées au cas par cas et suivant les réactions. En effet, il semble évident que le NADH peut être éliminé dans les réactions où il ne tient qu'un rôle d'apport d'énergie, alors qu'il semblerait absurde de le supprimer dans sa voie de biosynthèse. Cette idée est aussi voisine de la notion de métabolites internes et externes développée par [Schuster et al., 2002] et que nous avons vu en section 2.2.1.

4.6.2 Méthodes et évolutions futures

Nous observons que l'interprétation automatique des résultats de l'analyse par coloriage reste difficile, alors que 60% (données KEGG, tableau 3) à 80% (données SGD, tableau 5) des résultats de la méthode par analyse structurale ne

nécessitent aucune investigation supplémentaire pour décider de la conservation des voies métaboliques. Nous pouvons donc conclure que notre méthode par analyse strucutrale de graphes peut être appliquée à la prédiction systématique de voies métaboliques, en utilisant des voies de référence et des informations de génomique comparée. Nous pouvons toutefois envisager des ajouts qui pourraient permettre une prédiction encore meilleure.

De la complétude des résultats Nos méthodes ont pour but premier de prédire ce qui est conservé entre les différentes espèces, afin de fournir une base minimale et suffisante pour l'étape de modélisation. Dans les résultats que nous avons présentés (sections 4.4.2 et 4.5.2), nous avons donné des exemples de prédictions annotées manuellement, telle la perte de la voie métabolique du galactose chez *C. glabrata*. Bien que cette annotation puisse fournir des résultats intéressants, notre but n'est pas de fournir une annotation manuelle extensive de toutes les prédictions. Notamment, il est possible que l'identification d'équivalents fonctionnels n'identifie pas tous les homologues valides et qu'une voie soit considérée comme perdue alors qu'elle ne l'est pas. Dans de tels cas et dans le cadre d'une annotation manuelle des résultats, une recherche approfondie d'homologues en utilisant d'autres méthodes telles que [Chen and Vitkup, 2006] et [Kharchenko et al., 2006], peut permettre de combler certains trous dans les voies métaboliques prédites. Parallèlement, la qualité des voie de référence est un point très important. Les connaissances évoluant en permanence, les résultats obtenus dans cette étude pourront être enrichis de concert avec l'enrichissement des voies de référence.

Vers la modélisation additive ? Nos méthodes s'inscrivent dans le cadre de la modélisation soustractive, qui permet de prédire ce qui est perdu ou conservé entre une espèce de référence et des espèces d'intérêt, contrairement à la modélisation additive qui a pour objet d'étude le gain de modules fonctionnels. Nous savons que *Y. lipolytica* peut pousser sur un milieu composé d'hydrocarbones et que *C. glabrata* est une levure pathogène (voir [Dujon et al., 2004]). Nous pouvons donc raisonnablement émettre l'hypothèse que les voies impliquées dans le métabolisme des lipides sont renforcées chez *Y. lipolytica* et que *C. glabrata* a développé des voies métaboliques impliquées dans la pathogénicité. Pour prédire de tels phénomènes, il est possible d'utiliser des méthodes *ab initio* qui combinent des modules enzymatiques pour dessiner de nouvelles voies métaboliques [Horne et al., 2004] ou encore par l'étude de modes élémentaires [Schuster et al., 1999], en conjonction avec l'étude des expansions et contractions au sein des familles de protéines. En effet, pour cause d'adaptation à l'environnement, certaines familles de protéines montrent une expansion ou au contraire une contraction, à savoir une augmentation ou une diminution du nombre d'homologues pour une espèce dans une famille. Un cas bien connu se trouve chez la levure *Y. lipolytica*, où la famille GLS.94 montre une expansion et où les protéines sont homologues à la protéine LIP4 de *S. cerevisiae*, supposément

impliquée dans le métabolisme des lipides (EC 3.1.1.3). Ainsi, il pourrait donc être possible d'utiliser ce type d'information pour prédire des voies qui n'existent pas chez *S. cerevisiae*, ainsi que d'évaluer le renforcement ou le délitement de certaines voies ou sous-voies considérées comme conservées.

Amélioration de la définition des ports Nous pouvons faire deux remarques concernant la définition des ports. Tout d'abord, nous avons pu remarquer dans l'analyse des résultats que l'information de conservation des ports d'interconnexion n'est pas réellement informative. Dans les cas que nous avons étudiés, cette information n'a pas apporté d'indices supplémentaires pour la prise de décision, que ce soit dans le sens de la conservation ou bien de la perte de voie. Nous ne pouvons cependant pas affirmer que cette information est inutile. En effet, il semble raisonnable que les ports d'interconnexion de voies métaboliques soient des points importants. Théoriquement, ces points de jonction représentent à la fois la sortie d'une ou plusieurs voies, et l'entrée d'une ou plusieurs autres voies. Il est donc probable que la perte de ces métabolites, pouvant entraîner une déconnexion globale au niveau du réseau, est un phénomène rare. De plus, nous définissons actuellement les ports en terme de théorie des graphes, nous appuyant donc sur la structure des graphes que nous étudions. Or, nous avons vu en section 2.2.2 que le sens des réactions biochimiques, qui conditionne la structure du graphe et donc la définition des entrées et sorties, est déterminée par des experts. Les graphes des voies métaboliques sont construits à partir des données fournies par le KEGG et SGD, qui stockent les réactions enzymatiques dans une orientation particulière. Les ports sont donc implicitement définis de manière externe à notre méthode et dépendent de la qualité intrinsèque des données. Nous observons ainsi que des métabolites qui peuvent être définis comme biologiquement intéressants ne sont que rarement des ports dans les voies métaboliques du KEGG, tel le galactose qui n'est pas un métabolite d'entrée dans sa propre voie métabolique. Il est donc envisageable d'améliorer la définition des ports, soit par analyse des graphes, soit par expertise biologique.

Vers l'intégration d'autres données ? L'intégration d'autres sources de données, telles les relations d'interactions protéine-protéine, pourrait apporter des informations intéressantes. Pour de nombreuses réactions enzymatiques, différentes enzymes s'associent sous la forme d'un complexe protéique pour réaliser la réaction, si bien que celle-ci peut ne pas avoir lieu si une enzyme disparaît et que le complexe n'est pas formé. Nous pourrions traiter ces cas en utilisant des sommets qui ne représentent plus une enzyme particulière, mais plutôt un complexe ou une fonction enzymatique. En première approche, nous pouvons considérer qu'un complexe est perdu si une enzyme n'a pas d'homologue dans l'espèce étudiée. Dans le cas d'une approche plus avancée, nous pourrions utiliser les données expérimentales sur les complexes protéiques, ainsi que des méthodes informatiques pour la prédiction des interactions protéiques (cf. section 2.1.3). Cette approche fait cependant face à une

difficulté majeure, qui est la disponibilité de données de référence tenant compte à la fois des réactions biochimiques et de la composition des complexes enzymatiques. En effet, à l'étude des données disponibles, nous n'avons pas pu trouver de données concernant la structure ou la composition d'éventuels complexes enzymatiques. À cela, nous pouvons avancer deux hypothèses, soit que ces données ne soient pas disponibles car non connues, soit que les complexes enzymatiques n'existent pas, ce qui semble peu probable eu égard à la prépondérance des complexes protéiques dans les processus cellulaires.

Conclusion

Les travaux que nous avons présentés dans cette thèse ont pour sujet l'étude des réseaux d'interactions biomoléculaires, et plus précisément la prédiction *in silico* de ces réseaux. L'idée principale étant que les réseaux d'interactions peuvent servir de point d'entrée au processus de modélisation, qui s'avère plus que jamais nécessaire pour étudier les systèmes biologiques, systèmes complexes tant dans le nombre et la diversité des éléments qui les composent, que dans les relations qui unissent ces éléments. En ce sens, les approches que nous avons développées se placent clairement dans le domaine de la biologie des systèmes ou *Systems Biology* ([Onami et al., 2001], [Kitano, 2002], [Westerhoff and Palsson, 2004]). Les travaux que nous avons présentés ont pour point commun le but premier de fournir des résultats utiles et utilisables, en cherchant à développer des méthodes innovantes, puis en appliquant ces méthodes dans le cadre de développements logiciels.

Ainsi, nous avons présenté au chapitre 3 un formalisme permettant l'extraction de graphes d'interactions biomoléculaires à partir de données hétérogènes. Ce formalisme apporte au domaine la définition d'une base commune pour le développement de méthodes de prédiction et de visualisation de réseaux d'interactions. Notre approche est basée sur l'extraction de relations de voisinages : deux entités sont dites voisines si elles partagent une relation, que celle-ci soit basée sur une interaction explicite (e.g., interactions protéiques) ou sur le partage d'une même propriété (e.g., profils de coexpression). La deuxième particularité de notre méthode est d'établir une séparation claire entre l'extraction des relations et la matérialisation de celles-ci, sous forme de graphes visualisables. Cette séparation nous permet de définir deux notions, que sont les politiques d'extraction et les politiques de visualisation. Les premières définissent la sémantique des relations de voisinage et la manière de les prendre en charge pour leur extraction, tandis que les dernières définissent la manière de visualiser les relations extraites, afin de mettre en valeur leurs propriétés émergentes. Cette séparation en deux pôles distincts nous procure deux avantages majeurs. Le premier est que la réponse fournie par l'étape d'extraction est correcte par construction, eu-égard à la validité de la politique définie, et que la représentation des relations est aussi correcte par construction, puisque les politiques de visualisations ne font que matérialiser des relations définies comme correctes. Le deuxième avantage est la modularité et la flexibilité que ce découpage nous offre. Nous avons pu montrer que des méthodes existantes ([Bader and Hogue, 2003] et

[Bader, 2003]) peuvent être réutilisées en définissant des instances réelles de notre formalisme [Soularue, 2005]. Cette modularité permet d'envisager l'adaptation des meilleures méthodes disponibles pour la prédiction des différents types de réseaux d'interactions biomoléculaires, et nous fournit une base de développement pour de nouvelles méthodes. Dans cette thèse, nous avons présenté deux applications de ce formalisme. La première a consisté en la définition d'instances permettant d'extraire des relations d'interaction protéine-protéine, la seconde en la définition d'une méthode pour l'extrapolation de voies métaboliques.

La définition et le développement d'une politique d'extraction de graphes d'interactions protéine-protéine se sont faites dans le cadre du développement du logiciel ProViz [Iragne et al., 2005]. Ce logiciel permet une visualisation hautement interactive de grands graphes d'interactions biomoléculaires. Basé sur la plateforme logicielle Tulip, ProViz a pour particularité d'être focalisé sur les analyses biologiques et fourni donc un ensemble d'outils d'analyse et de méthodes de visualisation adaptés. Il propose notamment des analyses tirant partie de différentes ontologies (Gene Ontology pour les protéines, vocabulaire contrôlé PSI-MI pour leurs interactions). La politique d'extraction définie pour ce logiciel consiste à extraire les relations d'interaction physique entre différentes protéines d'intérêt. Deux politiques de visualisation ont été définies et sont utilisables, l'une adaptée à la représentation d'interactions binaires (politique *graphe d'interactions binaires*, algorithme 3), l'autre adaptée à la visualisation de relation binaires et de complexes protéiques (politique *graphe de complexes*, algorithme 4 page 64). L'utilisation de ces politiques est intégrée au logiciel par le biais d'une interface (figure 23, page 23) permettant de définir les paramètres de manière simple (choix des politiques de visualisation, des biomolécules d'intérêt, définition des filtres et du nombre d'extensions de voisinage).

La définition d'une méthode d'extrapolation de voies métaboliques s'est faite dans le cadre du projet Génolevures [Souciet J., 2000]. Une première approche par simple coloriage de graphe a fourni des résultats se prêtant bien à une analyse par expert et est disponible en ligne [Sherman et al., 2006] sur le site du projet Génolevures[4]. Une deuxième méthode, toujours basée sur une approche de modélisation soustractive mais plus orientée sur l'analyse structurale de graphes, a permis de fournir des prédictions de voies métaboliques, plus fiables et automatiques. En regard de notre formalisme d'extraction de graphe, ces deux approches sont identiques, puisque le point central n'est pas l'extraction de relations de voisinage qui sont dans ce cas de simples relations enzyme-métabolite, ni la définition d'une politique de visualisation de l'information, mais plutôt la définition de prédicats de filtrage. Nos approches sont basées sur la politique de reconstruction de voies métaboliques (algorithme 5, page 65). Dans ce cadre, le réseau métabolique générique nous est fourni par les données disponibles chez *S. cerevisiae*, et les données de génomiques comparées proviennent des familles de protéines Génolevures [Nikolski and Sherman, 2007]. Pour chacune des quatre espèces étudiées, les familles

[4]http ://cbi.labri.fr/Genolevures

de protéines nous permettent de réaliser un coloriage du graphe de *S. cerevisiae* afin de ne retenir que les protéines ayant un équivalent fonctionnel identifié, ainsi que les relations entre ces équivalents. Nos deux méthodes diffèrent alors dans les analyses réalisées sur le même graphe. Si la méthode par simple coloriage (section 4.4) tient compte uniquement du pourcentage de conservation d'enzymes impliquées dans chaque voie métabolique, la seconde méthode (section 4.5) effectue une analyse plus fine et tient compte de la structure du réseau prédit et tente de détecter les pertes de voies métaboliques par analyse des pertes d'entrées et de sorties de ces voies.

Les deux méthodes indiquent qu'une grande majorité des voies métaboliques de *S. cerevisiae* sont conservées chez les espèces étudiées, à savoir, *C. glabrata*, *K. lactis*, *D. hansenii* et *Y. lipolytica* (sections 4.4.2 et 4.5.2). Au niveau des résultats, la principale différence entre les deux méthodes est que la méthode par simple coloriage ne permet pas de prendre une décision fiable et quasi-automatique concernant les éventuelles pertes de voies métaboliques, en dehors des cas évidents (perte ou conservation de l'ensemble des enzymes). La seconde méthode facilite une telle décision, par la définition de règles de conservation d'entrées et sorties des voies métaboliques. Cette méthode nous permet de prédire de manière fiable et automatique un ensemble de fonctions métaboliques centrales (tables 10 et 11, annexe A) qui peuvent servir de base au processus de modélisation.

Quel futur pour nos méthodes ? Nous avons vu que notre cadre formel nous a permis de développer des méthodes d'extractions de réseaux d'interactions protéine-protéine, ainsi que d'extrapoler des voies métaboliques. Une application immédiate de ce formalisme est bien sûr l'instanciation de méthodes existantes, permettant de définir de nouvelles politiques d'extraction, de visualisation, ou encore de définir des prédicats de filtrages basés sur des calculs. Par l'utilisation de notre formalisme, la combinaison de ces méthodes que nous avons présentées en section 2.2.3, pourrait permettre une meilleure prédiction des réseaux d'interactions protéiques, des voies métaboliques, mais aussi réaliser des tâches de prédictions plus complexes, comme la prédiction de relations fonctionnelles entre biomolécules (e.g., prédiction de fonctions cellulaires).

Concernant notre méthode d'extrapolation de voies métaboliques, nous avons clairement identifiés différents points pouvant être sujets à amélioration. Tout d'abord, nous avons vu que la nature des données utilisées influe sur les résultats obtenus, notamment au niveau de la prise en compte des métabolites internes [Schuster et al., 2002]. Si les données du KEGG, qui ne représentent pas les métabolites internes, permettent d'observer une connexité moyenne supérieure dans les voies du métabolisme central, indiquant par ce fait une certaine robustesse en accord avec la notion de graphe petit monde (section 2.2.1), les données de SGD ne confirment pas cette observation. Nous pensons que cette absence de confirmation vient du fait que SGD représente à la fois les métabolites internes et les métabolites externes dans ses voies métaboliques. La réalisation d'un catalogue de

ces métabolites externes et leur élimination dans les réactions biochimiques appropriées devrait permettre de confirmer le résultat obtenu sur les données du KEGG. Une autre manière d'approcher le problème serait de réaliser un clustering sur le graphe global des voies métaboliques, afin de voir si une correspondance peut être établie entre les clusters densément connectés et les voies du métabolisme central.

Le but de nos méthodes d'extrapolation de voies métaboliques est de prédire de manière automatique et fiable un ensemble de voies métaboliques centrales, suffisant pour établir une base pour la modélisation des organismes étudiés. Nous pensons pouvoir affirmer que nous avons atteint ce but. Cependant, dans un souci de complétude des résultats, il pourrait être intéressant d'étudier diverses méthodes permettant de trouver des équivalents fonctionnels non-détectés par les familles de protéines. En effet, il est tout à fait envisageable que les familles de protéines soient parfois trop stringentes et ne détectent pas certains homologues. L'utilisation d'autres approches, adaptées au cadre spécifique des voies métaboliques ([Chen and Vitkup, 2006] ou [Kharchenko et al., 2006]), pourrait permettre d'augmenter le nombre de voies métaboliques considérées comme conservées.

Enfin, une dernière perspective est la modélisation additive. Dans nos travaux sur l'extrapolation de voies métaboliques, nous avons opté pour une approche de modélisation soustractive, à savoir la prédiction de ce qui est conservé et perdu entre deux espèces. La modélisation additive consiste en la prédiction de gains de modules fonctionnels, pouvant expliquer au moins en partie les différences phénotypiques entre espèces. Dans ce cadre, nous avons proposé plusieurs pistes à explorer, notamment au travers de l'étude des expansions et contractions de familles de protéines. Une expansion correspond à l'augmentation du nombre d'équivalents fonctionnels caractérisés chez une ou plusieurs espèces dans une famille de protéine. La contraction est le phénomène inverse, à savoir la perte d'homologues. Si l'expansion peut indiquer une diversification dans le processus impliquant ces protéines, la contraction peut, elle, indiquer une perte du processus, ou une place moins importante de ces réactions dans le métabolisme de l'espèce. Nous pouvons citer l'exemple de la famille GLS.94 qui montre une expansion chez *Y. lipolytica* et qui contient des homologues à la protéine LIP4 de *S. cerevisiae*, impliquée dans le métabolisme des lipides. Or, nous savons que *Y. lipolytica* est une levure qui a pour particularité de pouvoir se nourrir d'un milieu riche en hydrocarbones [Dujon et al., 2004]. Il semble donc raisonnable d'émettre l'hypothèse que cette expansion peut correspondre à une diversification dans les lipides utilisables par cette levure, correspondant donc à un gain de modules fonctionnels.

Bibliographie

[Agresti, 1999] Agresti, A. (1999). Modelling ordered categorical data : recent advances and future challenges. *Stat Med*, 18(17-18) :2191–2207.

[Alberts et al., 2002] Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., and Walter, P. (2002). *Molecular Biology of the Cell*. Garland Science.

[Aloy et al., 2004] Aloy, P., Böttcher, B., Ceulemans, H., Leutwein, C., Mellwig, C., Fischer, S., Gavin, A. C., Bork, P., Superti-Furga, G., Serrano, L., and Russell, R. B. (2004). Structure-based assembly of protein complexes in yeast. *Science*, 303(5666) :2026–2029.

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

[Altschul et al., 1997] Altschul, S. F., Madden, T. L., Schäffer, A. A., Zhang, J., Zhang, Z., Miller, W., and Lipman, D. J. (1997). Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res*, 25(17) :3389–3402.

[Andras and Andras, 2005] Andras, P. and Andras, C. (2005). The origins of life – the 'protein interaction world' hypothesis : protein interactions were the first form of self-reproducing life and nucleic acids evolved later as memory molecules. *Med Hypotheses*, 64(4) :678–688.

[Apweiler et al., 2004] Apweiler, R., Bairoch, A., Wu, C. H., Barker, W. C., Boeckmann, B., Ferro, S., Gasteiger, E., Huang, H., Lopez, R., Magrane, M., Martin, M. J., Natale, D. A., O'Donovan, C., Redaschi, N., and Yeh, L. S. (2004). Uniprot : the universal protein knowledgebase. *Nucleic Acids Res*, 32(Database issue).

[Auber, 2001] Auber, D. (2001). Tulip. In Mutzel, P., Jünger, M., and Leipert, S., editors, *9th Symp. Graph Drawing*, volume 2265 of *Lecture Notes in Computer Science*, pages 335–337. Springer-Verlag.

[Bader et al., 2001] Bader, G., Donaldson, I., Wolting, C., Ouellette, B., Pawson, T., and Hogue, C. (2001). BIND–The Biomolecular Interaction Network Database. *Nucleic Acids Res*, 29(1) :242–245.

[Bader and Hogue, 2002] Bader, G. D. and Hogue, C. W. (2002). Analyzing yeast protein-protein interaction data obtained from different sources. *Nat Biotech*, 20(10) :991–997.

[Bader and Hogue, 2003] Bader, G. D. and Hogue, C. W. (2003). An automated method for finding molecular complexes in large protein interaction networks. *BMC Bioinformatics*, 4 :2–2.

[Bader, 2003] Bader, J. (2003). Greedily building protein networks with confidence. *Bioinformatics*, 19(15) :1869–1874.

[Bader et al., 2004] Bader, J., Chaudhuri, A., Rothberg, J., and Chant, J. (2004). Gaining confidence in high-throughput protein interaction networks. *Nat Biotechnol*, 22(1) :78–85.

[Bairoch, 2000] Bairoch, A. (2000). The enzyme database in 2000. *Nucleic Acids Res*, 28(1) :304–305.

[Barabasi and Oltvai, 2004] Barabasi, A.-L. and Oltvai, Z. N. (2004). Network biology : Understanding the cell's functional organization. *Nat Rev Genet*, 5(2) :101–113.

[Bateman et al., 2004] Bateman, A., Coin, L., Durbin, R., Finn, R. D., Hollich, V., Griffiths-Jones, S., Khanna, A., Marshall, M., Moxon, S., Sonnhammer, E. L., Studholme, D. J., Yeats, C., and Eddy, S. R. (2004). The pfam protein families database. *Nucleic Acids Res*, 32 Database issue.

[Ben-Hur and Noble, 2006] Ben-Hur, A. and Noble, W. S. (2006). Choosing negative examples for the prediction of protein-protein interactions. *BMC Bioinformatics*, 7 Suppl 1.

[Bolotin-Fukuhara et al., 2006] Bolotin-Fukuhara, M., Casaregola, S., and Aigle, M. (2006). *Genome evolution : lessons from Genolevures*, volume 15. Topics in Current Genetics.

[Bono et al., 1998] Bono, H., Ogata, H., Goto, S., and Kanehisa, M. (1998). Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res*, 8(3) :203–210.

[Breitkreutz et al., 2003] Breitkreutz, B.-J., Stark, C., and Tyers, M. (2003). The grid : The general repository for interaction datasets. *Genome Biology*, 4(3).

[Brun et al., 2003] Brun, C., Chevenet, F., Martin, D., Wojcik, J., Guénoche, A., and Jacq, B. (2003). Functional classification of proteins for the prediction of cellular function from a protein-protein interaction network. *Genome Biol*, 5(1).

[Camacho-Carvajal et al., 2004] Camacho-Carvajal, M. M., Wollscheid, B., Aebersold, R., Steimle, V., and Schamel, W. W. (2004). Two-dimensional blue native/sds gel electrophoresis of multi-protein complexes from whole cellular lysates : a proteomics approach. *Mol Cell Proteomics*, 3(2) :176–182.

[Caspi et al., 2006] Caspi, R., Foerster, H., Fulcher, C. A., Hopkinson, R., Ingraham, J., Kaipa, P., Krummenacker, M., Paley, S., Pick, J., Rhee, S. Y., Tissier, C., Zhang, P., and Karp, P. D. (2006). Metacyc : a multiorganism database of metabolic pathways and enzymes. *Nucleic Acids Res*, 34(Database issue) :511–516.

[Chen and Vitkup, 2006] Chen, L. and Vitkup, D. (2006). Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biol*, 7(2).

[Cherry et al., 1998] Cherry, J. M., Adler, C., Ball, C., Chervitz, S. A., Dwight, S. S., Hester, E. T., Jia, Y., Juvik, G., Roe, T., Schroeder, M., Weng, S., and Botstein, D. (1998). Sgd : Saccharomyces genome database. *Nucleic Acids Res*, 26(1) :73–79.

[Cox et al., 2003] Cox, T. F., Cox, M. A. A., and Raton, B. (2003). Multidimensional scaling. *Technometrics*, 45(2) :182–182.

[Danchin, 1998] Danchin, A. (1998). La barque de delphes. *Editions Odile Jacob*.

[Demir et al., 2002] Demir, E., Babur, O., Dogrusoz, U., Gursoy, A., Nisanci, G., Cetin-Atalay, R., and Ozturk, M. (2002). Patika : an integrated visual environment for collaborative construction and analysis of cellular pathways. *Bioinformatics*, 18(7) :996–1003.

[Deng et al., 2003] Deng, M., Sun, F., and Chen, T. (2003). Assessment of the reliability of protein-protein interactions and protein function prediction. *Pac Symp Biocomput*, pages 140–151.

[Deng et al., 2004] Deng, M., Tu, Z., Sun, F., and Chen, T. (2004). Mapping Gene Ontology to proteins based on protein-protein interaction data. *Bioinformatics*, 20(6) :895–902.

[Deville et al., 2003] Deville, Y., Gilbert, D., van Helden, J., and Wodak, S. J. (2003). An overview of data models for the analysis of biochemical pathways. *Brief Bioinform*, 4(3) :246–259.

[Diestel, 2005] Diestel, R. (2005). *Graph Theory*. Springer-Verlag.

[Dujon, 2006] Dujon, B. (2006). Yeasts illustrate the molecular mechanisms of eukaryotic genome evolution. *Trends Genet*, 22(7) :375–387.

[Dujon et al., 2004] Dujon, B., Sherman, D., Fischer, G., Durrens, P., Casaregola, S., Lafontaine, I., De Montigny, J., Marck, C., Neuvéglise, C., Talla, E., Goffard, N., Frangeul, L., Aigle, M., Anthouard, V., Babour, A., Barbe, V., Barnay, S., Blanchin, S., Beckerich, J., Beyne, E., Bleykasten, C., Boisramé, A., Boyer, J., Cattolico, L., Confanioleri, F., De Daruvar, A., Despons, L., Fabre, E., Fairhead, C., Ferry-Dumazet, H., Groppi, A., Hantraye, F., Hennequin, C., Jauniaux, N., Joyet, P., Kachouri, R., Kerrest, A., Koszul, R., Lemaire, M., Lesur, I., Ma, L., Muller, H., Nicaud, J., Nikolski, M., Oztas, S., Ozier-Kalogeropoulos, O., Pellenz, S., Potier, S., Richard, G., Straub, M., Suleau, A., Swennen, D., Tekaia, F., Wésolowski-Louvel, M., Westhof, E., Wirth, B., Zeniou-Meyer, M., Zivanovic, I., Bolotin-Fukuhara, M., Thierry, A., Bouchier, C., Caudron, B., Scarpelli, C., Gaillardin, C., Weissenbach, J., Wincker, P., and Souciet, J. (2004). Genome evolution in yeasts. *Nature*, 430(6995) :35–44.

[Dunn et al., 2004] Dunn, B., Ferea, T., Spellman, P., Schwarz, J., Terraciano, J., Troyanovich, J., Walker, S., Greene, J., Shaw, K., DiDomenico, B., Wang, Q.,

Kaloper, M., Metzner, S., Chung, E., Bondre, C., Venteicher, A., Botstein, D., and Brown, P. (2004). Genetic footprinting : A functional analysis of the s. cerevisiae genome. *Personnal communication to SGD*.

[Ekman et al., 2006] Ekman, D., Light, S., Björklund, A. K., and Elofsson, A. (2006). What properties characterize the hub proteins of the protein-protein interaction network of saccharomyces cerevisiae ? *Genome Biol*, 7(6).

[Espadaler et al., 2005] Espadaler, J., Romero-Isart, O., Jackson, R., and Oliva, B. (2005). Prediction of protein-protein interactions using distant conservation of sequence patterns and structure relationships. *Bioinformatics*, 21(16) :3360–3368.

[Fell and Wagner, 2000] Fell, D. A. and Wagner, A. (2000). The small world of metabolism. *Nat Biotechnol*, 18(11) :1121–1122.

[Fields and Song, 1989] Fields, S. and Song, O. (1989). A novel genetic system to detect protein-protein interactions. *Nature*, 340(6230) :245–246.

[Fraser et al., 2004] Fraser, H., Hirsh, A., Wall, D., and Eisen, M. (2004). Coevolution of gene expression among interacting proteins. *Proc Natl Acad Sci U S A*, 101(24) :9033–9038.

[Fraser and Hirsh, 2004] Fraser, H. B. and Hirsh, A. E. (2004). Evolutionary rate depends on number of protein-protein interactions independently of gene expression level. *BMC Evol Biol*, 4 :13–13.

[Fraser et al., 2002] Fraser, H. B., Hirsh, A. E., Steinmetz, L. M., Scharfe, C., and Feldman, M. W. (2002). Evolutionary rate in the protein interaction network. *Science*, 296(5568) :750–752.

[Frick et al., 1994] Frick, A., Ludwig, A., and Mehldan, H. (1994). A fast adaptive layout algorithm for undirected graphs. In Springer-Verlag, editor, *Proc. Workshop on Graph Drawing 94*, volume LNCS 894, pages 389–403.

[Friedrich and Schreiber, 2003] Friedrich, C. and Schreiber, F. (2003). Visualisation and navigation methods for typed protein-protein interaction networks. *Appl Bioinformatics*, 2(3 Suppl) :19–24.

[Fukuda and Takagi, 2001] Fukuda, K. and Takagi, T. (2001). Knowledge representation of signal transduction pathways. *Bioinformatics*, 17(9) :829–837.

[Gavin et al., 2002] Gavin, A., Bösche, M., Krause, R., Grandi, P., Marzioch, M., Bauer, A., Schultz, J., Rick, J., Michon, A., Cruciat, C., Remor, M., Höfert, C., Schelder, M., Brajenovic, M., Ruffner, H., Merino, A., Klein, K., Hudak, M., Dickson, D., Rudi, T., Gnau, V., Bauch, A., Bastuck, S., Huhse, B., Leutwein, C., Heurtier, M., Copley, R., Edelmann, A., Querfurth, E., Rybin, V., Drewes, G., Raida, M., Bouwmeester, T., Bork, P., Seraphin, B., Kuster, B., Neubauer, G., and Superti-Furga, G. (2002). Functional organization of the yeast proteome by systematic analysis of protein complexes. *Nature*, 415(6868) :141–147.

[Gerhard, 1992] Gerhard, M. (1992). *Biological Pathways, 3rd ed.* Boehringer Mannheim.

[Gerstein et al., 2002] Gerstein, M., Lan, N., and Jansen, R. (2002). Proteomics : Enhanced : Integrating interactomes. *Science*, 295(5553) :284–287.

[Giaever et al., 2002] Giaever, G., Chu, A., Ni, L., Connelly, C., Riles, L., Véronneau, S., Dow, S., Lucau-Danila, A., Anderson, K., André, B., Arkin, A., Astromoff, A., El-Bakkoury, M., Bangham, R., Benito, R., Brachat, S., Campanaro, S., Curtiss, M., Davis, K., Deutschbauer, A., Entian, K., Flaherty, P., Foury, F., Garfinkel, D., Gerstein, M., Gotte, D., Güldener, U., Hegemann, J., Hempel, S., Herman, Z., Jaramillo, D., Kelly, D., Kelly, S., Kötter, P., LaBonte, D., Lamb, D., Lan, N., Liang, H., Liao, H., Liu, L., Luo, C., Lussier, M., Mao, R., Menard, P., Ooi, S., Revuelta, J., Roberts, C., Rose, M., Ross-Macdonald, P., Scherens, B., Schimmack, G., Shafer, B., Shoemaker, D., Sookhai-Mahadeo, S., Storms, R., Strathern, J., Valle, G., Voet, M., Volckaert, G., Wang, C., Ward, T., Wilhelmy, J., Winzeler, E., Yang, Y., Yen, G., Youngman, E., Yu, K., Bussey, H., Boeke, J., Snyder, M., Philippsen, P., Davis, R., and Johnston, M. (2002). Functional profiling of the saccharomyces cerevisiae genome. *Nature*, 418(6896) :387–391.

[Giot et al., 2003] Giot, L., Bader, J. S., Brouwer, C., Chaudhuri, A., Kuang, B., Li, Y., Hao, Y. L., Ooi, C. E., Godwin, B., Vitols, E., Vijayadamodar, G., Pochart, P., Machineni, H., Welsh, M., Kong, Y., Zerhusen, B., Malcolm, R., Varrone, Z., Collis, A., Minto, M., Burgess, S., McDaniel, L., Stimpson, E., Spriggs, F., Williams, J., Neurath, K., Ioime, N., Agee, M., Voss, E., Furtak, K., Renzulli, R., Aanensen, N., Carrolla, S., Bickelhaupt, E., Lazovatsky, Y., DaSilva, A., Zhong, J., Stanyon, C. A., Finley, R. L., White, K. P., Braverman, M., Jarvie, T., Gold, S., Leach, M., Knight, J., Shimkets, R. A., McKenna, M. P., Chant, J., and Rothberg, J. M. (2003). A protein interaction map of drosophila melanogaster. *Science*, 302(5651) :1727–1736.

[Goffard et al., 2003] Goffard, N., Garcia, V., Iragne, F., Groppi, A., and de Daruvar, A. (2003). Ippred : server for proteins interactions inference. *Bioinformatics*, 19(7) :903–904.

[Goldberg and Roth, 2003] Goldberg, D. and Roth, F. (2003). Assessing experimentally derived interactions in a small world. *Proc Natl Acad Sci U S A*, 100(8) :4372–4376.

[Goto et al., 1997] Goto, S., Bono, H., Ogata, H., Fujibuchi, W., Nishioka, T., Sato, K., and Kanehisa, M. (1997). Organizing and computing metabolic pathway data in terms of binary relations. *Pac Symp Biocomput*, pages 175–186.

[Hahn and Kern, 2005] Hahn, M. and Kern, A. (2005). Comparative genomics of centrality and essentiality in three eukaryotic protein-interaction networks. *Mol Biol Evol*, 22(4) :803–806.

[Han et al., 2005] Han, J. D., Dupuy, D., Bertin, N., Cusick, M. E., and Vidal, M. (2005). Effect of sampling on topology predictions of protein-protein interaction networks. *Nat Biotechnol*, 23(7) :839–844.

[Han et al., 2004] Han, J.-D. J., Bertin, N., Hao, T., Goldberg, D. S., Berriz, G. F., Zhang, L. V., Dupuy, D., Walhout, A. J. M., Cusick, M. E., Roth, F. P., and

Vidal, M. (2004). Evidence for dynamically organized modularity in the yeast protein-protein interaction network. *Nature*, 430(6995) :88–93.

[Harris et al., 2004] Harris, M., Clark, J., Ireland, A., Lomax, J., Ashburner, M., Foulger, R., Eilbeck, K., Lewis, S., Marshall, B., Mungall, C., Richter, J., Rubin, G., Blake, J., Bult, C., Dolan, M., Drabkin, H., Eppig, J., Hill, D., Ni, L., Ringwald, M., Balakrishnan, R., Cherry, J., Christie, K., Costanzo, M., Dwight, S., Engel, S., Fisk, D., Hirschman, J., Hong, E., Nash, R., Sethuraman, A., Theesfeld, C., Botstein, D., Dolinski, K., Feierbach, B., Berardini, T., Mundodi, S., Rhee, S., Apweiler, R., Barrell, D., Camon, E., Dimmer, E., Lee, V., Chisholm, R., Gaudet, P., Kibbe, W., Kishore, R., Schwarz, E., Sternberg, P., Gwinn, M., Hannick, L., Wortman, J., Berriman, M., Wood, V., de la Cruz, N., Tonellato, P., Jaiswal, P., Seigfried, T., White, R., and , . (2004). The Gene Ontology (GO) database and informatics resource. *Nucleic Acids Res*, 32(Database issue) :258–261.

[Hermjakob et al., 2004a] Hermjakob, H., Montecchi-Palazzi, L., Bader, G., Wojcik, J., Salwinski, L., Ceol, A., Moore, S., Orchard, S., Sarkans, U., von Mering, C., Roechert, B., Poux, S., Jung, E., Mersch, H., Kersey, P., Lappe, M., Li, Y., Zeng, R., Rana, D., Nikolski, M., Husi, H., Brun, C., Shanker, K., Grant, S., Sander, C., Bork, P., Zhu, W., Pandey, A., Brazma, A., Jacq, B., Vidal, M., Sherman, D., Legrain, P., Cesareni, G., Xenarios, I., Eisenberg, D., Steipe, B., Hogue, C., and Apweiler, R. (2004a). The HUPO PSI's molecular interaction format–a community standard for the representation of protein interaction data. *Nat Biotechnol*, 22(2) :177–183.

[Hermjakob et al., 2004b] Hermjakob, H., Montecchi-Palazzi, L., Lewington, C., Mudali, S., Kerrien, S., Orchard, S., Vingron, M., Roechert, B., Roepstorff, P., Valencia, A., Margalit, H., Armstrong, J., Bairoch, A., Cesareni, G., Sherman, D., and Apweiler, R. (2004b). IntAct : an open source molecular interaction database. *Nucleic Acids Res*, 32(Database issue) :452–455.

[Hirsh and Fraser, 2001] Hirsh, A. E. and Fraser, H. B. (2001). Protein dispensability and rate of evolution. *Nature*, 411(6841) :1046–1049.

[Hittinger et al., 2004] Hittinger, C. T., Rokas, A., and Carroll, S. B. (2004). Parallel inactivation of multiple gal pathway genes and ecological diversification in yeasts. *Proc Natl Acad Sci U S A*, 101(39) :14144–14149.

[Ho et al., 2002] Ho, Y., Gruhler, A., Heilbut, A., Bader, G., Moore, L., Adams, S., Millar, A., Taylor, P., Bennett, K., Boutilier, K., Yang, L., Wolting, C., Donaldson, I., Schandorff, S., Shewnarane, J., Vo, M., Taggart, J., Goudreault, M., Muskat, B., Alfarano, C., Dewar, D., Lin, Z., Michalickova, K., Willems, A., Sassi, H., Nielsen, P., Rasmussen, K., Andersen, J., Johansen, L., Hansen, L., Jespersen, H., Podtelejnikov, A., Nielsen, E., Crawford, J., Poulsen, V., Sorensen, B., Matthiesen, J., Hendrickson, R., Gleeson, F., Pawson, T., Moran, M., Durocher, D., Mann, M., Hogue, C., Figeys, D., and Tyers, M. (2002). Systematic identification of protein complexes in saccharomyces cerevisiae by mass spectrometry. *Nature*, 415(6868) :180–183.

[Hoffmann and Valencia, 2003] Hoffmann, R. and Valencia, A. (2003). Protein interaction : same network, different hubs. *Trends Genet*, 19(12) :681–683.

[Hoffmann-Ostenhof and Thompson, 1958] Hoffmann-Ostenhof, . and Thompson, R. H. S. (1958). International commission on enzymes. *Nature*, 181.

[Horne et al., 2004] Horne, A. B., Hodgman, T. C., Spence, H. D., and Dalby, A. R. (2004). Constructing an enzyme-centric view of metabolism. *Bioinformatics*, 20(13) :2050–2055.

[Hughes and Friedman, 2005] Hughes, A. L. and Friedman, R. (2005). Gene duplication and the properties of biological networks. *J Mol Evol*, 61(6) :758–764.

[Huh et al., 2003] Huh, W., Falvo, J., Gerke, L., Carroll, A., Howson, R., Weissman, J., and O'Shea, E. (2003). Global analysis of protein localization in budding yeast. *Nature*, 425(6959) :686–691.

[Huynen et al., 2003] Huynen, M., Snel, B., von Mering, C., and Bork, P. (2003). Function prediction and protein networks. *Curr Opin Cell Biol*, 15(2) :191–198.

[Iossifov et al., 2004] Iossifov, I., Krauthammer, M., Friedman, C., Hatzivassiloglou, V., Bader, J., White, K., and Rzhetsky, A. (2004). Probabilistic inference of molecular networks from noisy data sources. *Bioinformatics*, 20(8) :1205–1213.

[Iragne et al., 2005] Iragne, F., Nikolski, M., Mathieu, B., Auber, D., and Sherman, D. (2005). ProViz : protein interaction visualization and exploration. *Bioinformatics*, 21(2) :272–274.

[Ispolatov et al., 2005] Ispolatov, I., Krapivsky, P., and Yuryev, A. (2005). Duplication-divergence model of protein interaction network. *Phys Rev E Stat Nonlin Soft Matter Phys*, 71(6 Pt 1) :061911–061911.

[Ito et al., 2001] Ito, T., Chiba, T., Ozawa, R., Yoshida, M., Hattori, M., and Sakaki, Y. (2001). A comprehensive two-hybrid analysis to explore the yeast protein interactome. *Proc Natl Acad Sci U S A*, 98(8) :4569–4574.

[Jaimovich et al., 2006] Jaimovich, A., Elidan, G., Margalit, H., and Friedman, N. (2006). Towards an integrated protein-protein interaction network : a relational markov network approach. *J Comput Biol*, 13(2) :145–164.

[Jansen and Gerstein, 2004] Jansen, R. and Gerstein, M. (2004). Analyzing protein function on a genomic scale : the importance of gold-standard positives and negatives for network prediction. *Curr Opin Microbiol*, 7(5) :535–545.

[Jansen et al., 2003] Jansen, R., Yu, H., Greenbaum, D., Kluger, Y., Krogan, N., Chung, S., Emili, A., Snyder, M., Greenblatt, J., and Gerstein, M. (2003). A Bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, 302(5644) :449–453.

[Jeong et al., 2001] Jeong, H., Mason, S. P., Barabási, A. L., and Oltvai, Z. N. (2001). Lethality and centrality in protein networks. *Nature*, 411(6833) :41–42.

[Jeong et al., 2000] Jeong, H., Tombor, B., Albert, R., Oltvai, Z. N., and Barabási, A. L. (2000). The large-scale organization of metabolic networks. *Nature*, 407(6804) :651–654.

[Jordan et al., 2003] Jordan, I. K., Wolf, Y. I., and Koonin, E. V. (2003). No simple dependence between protein evolution rate and the number of protein-protein interactions : only the most prolific interactors tend to evolve slowly. *BMC Evol Biol*, 3.

[Kanehisa, 1997] Kanehisa, M. (1997). A database for post-genome analysis. *Trends Genet*, 13(9) :375–376.

[Karp et al., 2002] Karp, P. D., Paley, S., and Romero, P. (2002). The pathway tools software. *Bioinformatics*, 18 Suppl 1 :225–232.

[Kellis et al., 2003] Kellis, M., Patterson, N., Endrizzi, M., Birren, B., and Lander, E. S. (2003). Sequencing and comparison of yeast species to identify genes and regulatory elements. *Nature*, 423(6937) :241–254.

[Kersey et al., 2005] Kersey, P., Bower, L., Morris, L., Horne, A., Petryszak, R., Kanz, C., Kanapin, A., Das, U., Michoud, K., Phan, I., Gattiker, A., Kulikova, T., Faruque, N., Duggan, K., Mclaren, P., Reimholz, B., Duret, L., Penel, S., Reuter, I., and Apweiler, R. (2005). Integr8 and genome reviews : integrated views of complete genomes and proteomes. *Nucleic Acids Research*, 33(Supplement 1) :D297+.

[Kharchenko et al., 2006] Kharchenko, P., Chen, L., Freund, Y., Vitkup, D., and Church, G. M. (2006). Identifying metabolic enzymes with multiple types of association evidence. *BMC Bioinformatics*, 7 :177–177.

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

[Krishnamurthy et al., 2003] Krishnamurthy, L., Nadeau, J., Ozsoyoglu, G., Ozsoyoglu, M., Schaeffer, G., Tasan, M., and Xu, W. (2003). Pathways database system : an integrated system for biological pathways. *Bioinformatics*, 19(8) :930–937.

[Kumar et al., 2002] Kumar, A., Agarwal, S., Heyman, J., Matson, S., Heidtman, M., Piccirillo, S., Umansky, L., Drawid, A., Jansen, R., Liu, Y., Cheung, K., Miller, P., Gerstein, M., Roeder, G., and Snyder, M. (2002). Subcellular localization of the yeast proteome. *Genes Dev*, 16(6) :707–719.

[Lacount et al., 2005] Lacount, D. J., Vignali, M., Chettier, R., Phansalkar, A., Bell, R., Hesselberth, J. R., Schoenfeld, L. W., Ota, I., Sahasrabudhe, S., Kurschner, C., Fields, S., and Hughes, R. E. (2005). A protein interaction network of the malaria parasite plasmodium falciparum. *Nature*, 438(7064) :103–107.

[Legrain et al., 2001] Legrain, P., Wojcik, J., and Gauthier, J.-M. (2001). Protein-protein interaction maps : a lead towards cellular functions. *Trends in Genetics*, 17.

[Li et al., 2004] Li, S., Armstrong, C. M., Bertin, N., Ge, H., Milstein, S., Boxem, M., Vidalain, P. O., Han, J. D., Chesneau, A., Hao, T., Goldberg, D. S., Li, N., Martinez, M., Rual, J. F., Lamesch, P., Xu, L., Tewari, M., Wong, S. L., Zhang,

L. V., Berriz, G. F., Jacotot, L., Vaglio, P., Reboul, J., Hirozane-Kishikawa, T., Li, Q., Gabel, H. W., Elewa, A., Baumgartner, B., Rose, D. J., Yu, H., Bosak, S., Sequerra, R., Fraser, A., Mango, S. E., Saxton, W. M., Strome, S., Van Den Heuvel, S., Piano, F., Vandenhaute, J., Sardet, C., Gerstein, M., Doucette-Stamm, L., Gunsalus, K. C., Harper, J. W., Cusick, M. E., Roth, F. P., Hill, D. E., and Vidal, M. (2004). A map of the interactome network of the metazoan c. elegans. *Science*, 303(5657) :540–543.

[Lin et al., 2004] Lin, N., Wu, B., Jansen, R., Gerstein, M., and Zhao, H. (2004). Information assessment on predicting protein-protein interactions. *BMC Bioinformatics*, 5 :154–154.

[Luciano, 2005] Luciano, J. S. (2005). Pax of mind for pathway researchers. *Drug Discov Today*, 10(13) :937–942.

[Luo et al., 2007] Luo, F., Yang, Y., Chen, C. F., Chang, R., Zhou, J., and Scheuermann, R. H. (2007). Modular organization of protein interaction networks. *Bioinformatics*, 23(2) :207–214.

[Ma and Zeng, 2003] Ma, H. and Zeng, A. P. (2003). Reconstruction of metabolic networks from genome data and analysis of their global structure for various organisms. *Bioinformatics*, 19(2) :270–277.

[Mäkinen, 1988] Mäkinen, E. (1988). On circular layouts. *International Journal of Computer Mathematics*, 24 :29–37.

[Mavrovouniotis and Stephanopoulos, 1990] Mavrovouniotis, M.L.and Stephanopoulos, G. and Stephanopoulos, G. (1990). Computer-aided synthesis of biochemical pathways. *Biotechnol. Bioeng.*, 36 :1119–1132.

[Messinger et al., 1991] Messinger, E., Rowe, L., and Henry, R. (1991). A divide and conquer algorithm for the automatic layout of large directed graphs. *IEEE Transactions on Systems, Man and Cybernetics*, SMC-21(1) :1–11.

[Mewes et al., 2002] Mewes, H. W., Frishman, D., Güldener, U., Mannhaupt, G., Mayer, K., Mokrejs, M., Morgenstern, B., Münsterkötter, M., Rudd, S., and Weil, B. (2002). Mips : a database for genomes and protein sequences. *Nucleic Acids Res*, 30(1) :31–34.

[Mulder et al., 2007] Mulder, Nicola, J., Apweiler, Rolf, Attwood, Teresa, K., Bairoch, Amos, Bateman, Alex, Binns, David, Bork, Peer, Buillard, Virginie, Cerutti, Lorenzo, Copley, Richard, Courcelle, Emmanuel, Das, Ujjwal, Daugherty, Louise, Dibley, Mark, Finn, Robert, Fleischmann, Wolfgang, Gough, Julian, Haft, Daniel, Hulo, Nicolas, Hunter, Sarah, Kahn, Daniel, Kanapin, Alexander, Kejariwal, Anish, Labarga, Alberto, Langendijk-Genevaux, Petra, S., Lonsdale, David, Lopez, Rodrigo, Letunic, Ivica, Madera, Martin, Maslen, John, Mcanulla, Craig, Mcdowall, Jennifer, Mistry, Jaina, Mitchell, Alex, Nikolskaya, Anastasia, N., Orchard, Sandra, Orengo, Christine, Petryszak, Robert, Selengut, Jeremy, D., Sigrist, Christian, J. A., Thomas, Paul, D., Valentin, Franck, Wilson, Derek, Wu,

Cathy, H., Yeats, and Corin (2007). New developments in the interpro database. *Nucleic Acids Research*, 35(Supplement 1) :D224–D228.

[Nabieva et al., 2005] Nabieva, E., Jim, K., Agarwal, A., Chazelle, B., and Singh, M. (2005). Whole-proteome prediction of protein function via graph-theoretic analysis of interaction maps. *Bioinformatics*, 21 Suppl 1 :302–302.

[Nanni and Lumini, 2006] Nanni, L. and Lumini, A. (2006). An ensemble of k-local hyperplanes for predicting protein-protein interactions. *Bioinformatics*.

[Nikolski and Sherman, 2007] Nikolski, M. and Sherman, D. J. (2007). Family relationships : should consensus reign ?–consensus clustering for protein families. *Bioinformatics*, 23(2) :71–76.

[Nishizuka, 1980] Nishizuka, T. (1980). *Metabolic Maps*. Biochemical Society of Japan.

[Notredame et al., 2000] Notredame, C., Higgins, D. G., and Heringa, J. (2000). T-coffee : A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, 302(1) :205–217.

[O'Brien et al., 2005] O'Brien, K. P., Remm, M., and Sonnhammer, E. L. (2005). Inparanoid : a comprehensive database of eukaryotic orthologs. *Nucleic Acids Res*, 33(Database issue).

[Ogata et al., 2000] Ogata, H., Fujibuchi, W., Goto, S., and Kanehisa, M. (2000). A heuristic graph comparison algorithm and its application to detect functionally related enzyme clusters. *Nucleic Acids Res*, 28(20) :4021–4028.

[Ogata et al., 1998] Ogata, H., Goto, S., Fujibuchi, W., and Kanehisa, M. (1998). Computation with the kegg pathway database. *Biosystems*, 47(1-2) :119–128.

[Ogata et al., 1999] Ogata, H., Goto, S., Sato, K., Fujibuchi, W., Bono, H., and Kanehisa, M. (1999). Kegg : Kyoto encyclopedia of genes and genomes. *Nucleic Acids Res*, 27(1) :29–34.

[Oliveira et al., 2001] Oliveira, J. S., Bailey, C. G., Jones-Oliveira, J. B., and Dixon, D. A. (2001). An algebraic-combinatorial model for the identification and mapping of biochemical pathways. *Bull Math Biol*, 63(6) :1163–1196.

[Onami et al., 2001] Onami, S., Kyoda, K., Morohashi, M., and Kitano, H. (2001). *Foundations of Systems Biology*. MIT Press.

[Orchard et al., 2003] Orchard, S., Kersey, P., Zhu, W., Montecchi-Palazzi, L., Hermjakob, H., and Apweiler, R. (2003). Progress in establishing common standards for exchanging proteomics data : The 2nd meeting of the hupo psi. *Comp. and Fun. Genomics*, 4(2) :203–206.

[Orwant et al., 1999] Orwant, J., Hietaniemi, J., and Macdonald, J. (1999). *Mastering Algorithms with Perl*. O'Reilly & Associates.

[Overbeek et al., 2000] Overbeek, R., Larsen, N., Pusch, G. D., D'Souza, M., Selkov, E., Kyrpides, N., Fonstein, M., Maltsev, N., and Selkov, E. (2000). Wit :

integrated system for high-throughput genome sequence analysis and metabolic reconstruction. *Nucleic Acids Res*, 28(1) :123–125.

[Papin et al., 2002] Papin, J. A., Price, N. D., Edwards, J. S., and B, B. Ã. P. (2002). The genome-scale metabolic extreme pathway structure in haemophilus influenzae shows significant network redundancy. *J Theor Biol*, 215(1) :67–82.

[Papin et al., 2003] Papin, J. A., Price, N. D., Wiback, S. J., Fell, D. A., and Palsson, B. O. (2003). Metabolic pathways in the post-genome era. *Trends Biochem Sci*, 28(5) :250–258.

[Pearson and Lipman, 1988] Pearson, W. R. and Lipman, D. J. (1988). Improved tools for biological sequence comparison. *Proc Natl Acad Sci U S A*, 85(8) :2444–2448.

[Pfeiffer et al., 1999] Pfeiffer, T., Sanchez-Valdenebro, I., Nuno, J., Montero, F., and S., S. (1999). Metatool : For studying metabolic networks. *Bioinformatics*, 15 :251–257.

[Podani et al., 2001] Podani, J., Oltvai, Z. N., Jeong, H., Tombor, B., Barabási, A. L., and Szathmáry, E. (2001). Comparable system-level organization of archaea and eukaryotes. *Nat Genet*, 29(1) :54–56.

[Price et al., 2002] Price, N. D., Papin, J. A., and Palsson, B. Ã. (2002). Determination of redundancy and systems properties of the metabolic network of helicobacter pylori using genome-scale extreme pathway analysis. *Genome Res*, 12(5) :760–769.

[Przulj et al., 2004] Przulj, N., Corneil, D. G., and Jurisica, I. (2004). Modeling interactome : scale-free or geometric ? *Bioinformatics*, 20(18) :3508–3515.

[Qi et al., 2006] Qi, Y., Bar-Joseph, Z., and Klein-Seetharaman, J. (2006). Evaluation of different biological data and computational classification methods for use in protein interaction prediction. *Proteins*, 63(3) :490–500.

[Rigaut et al., 1999] Rigaut, G., Shevchenko, A., Rutz, B., Wilm, M., Mann, M., and Séraphin, B. (1999). A generic protein purification method for protein complex characterization and proteome exploration. *Nat Biotechnol*, 17(10) :1030–1032.

[Rivera et al., 1998] Rivera, M. C., Jain, R., Moore, J. E., and Lake, J. A. (1998). Genomic evidence for two functionally distinct gene classes. *Proc Natl Acad Sci U S A*, 95(11) :6239–6244.

[Ross-Macdonald et al., 1999] Ross-Macdonald, P., Coelho, P. S., Roemer, T., Agarwal, S., Kumar, A., Jansen, R., Cheung, K. H., Sheehan, A., Symoniatis, D., Umansky, L., Heidtman, M., Nelson, F. K., Iwasaki, H., Hager, K., Gerstein, M., Miller, P., Roeder, G. S., and Snyder, M. (1999). Large-scale analysis of the yeast genome by transposon tagging and gene disruption. *Nature*, 402(6760) :413–418.

[Saitou and Nei, 1987] Saitou, N. and Nei, M. (1987). The neighbor-joining method : a new method for reconstructing phylogenetic trees. *Mol Biol Evol*, 4(4) :406–425.

[Salwinski and Eisenberg, 2003] Salwinski, L. and Eisenberg, D. (2003). Computational methods of analysis of protein-protein interactions. *Curr Opin Struct Biol*, 13(3) :377–382.

[Samatova et al., 2002] Samatova, N., Geist, A., Ostrouchov, G., and Melechko, A. (2002). Parallel out-of-core algorithm for genome-scale enumeration of metabolic systemic pathways. *Proceedings of the first IEEE HiCOMB Workshop*, pages 185–192.

[Schilling et al., 2002] Schilling, C. H., Covert, M. W., Famili, I., Church, G. M., Edwards, J. S., and Palsson, B. O. (2002). Genome-scale metabolic model of helicobacter pylori 26695. *J Bacteriol*, 184(16) :4582–4593.

[Schilling and Palsson, 2000] Schilling, C. H. and Palsson, B. O. (2000). Assessment of the metabolic capabilities of haemophilus influenzae rd through a genome-scale pathway analysis. *J Theor Biol*, 203(3) :249–283.

[Schilling et al., 1999] Schilling, C. H., Schuster, S., Palsson, B. O., and Heinrich, R. (1999). Metabolic pathway analysis : basic concepts and scientific applications in the post-genomic era. *Biotechnol Prog*, 15(3) :296–303.

[Schomburg et al., 2002] Schomburg, I., Chang, A., and Schomburg, D. (2002). Brenda, enzyme data and metabolic information. *Nucleic Acids Res*, 30(1) :47–49.

[Schuster et al., 1999] Schuster, S., Dandekar, T., and Fell, D. A. (1999). Detection of elementary flux modes in biochemical networks : a promising tool for pathway analysis and metabolic engineering. *Trends Biotechnol*, 17(2) :53–60.

[Schuster and Hilgetag, 1994] Schuster, S. and Hilgetag, C. (1994). On elementary flux modes in biochemical reaction systems at steady state. *J. Biol. Syst.*, 2 :165–182.

[Schuster et al., 2002] Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I., and Dandekar, T. (2002). Exploring the pathway structure of metabolism : decomposition into subnetworks and application to mycoplasma pneumoniae. *Bioinformatics*, 18(2) :351–361.

[Selkov et al., 1996] Selkov, E., Basmanova, S., Gaasterland, T., Goryanin, I., Gretchkin, Y., Maltsev, N., Nenashev, V., Overbeek, R., Panyushkina, E., Pronevitch, L., Selkov, E., and Yunus, I. (1996). The metabolic pathway collection from emp : the enzymes and metabolic pathways database. *Nucleic Acids Res*, 24(1) :26–28.

[Shannon et al., 2003] Shannon, P., Markiel, A., Ozier, O., Baliga, N. S., Wang, J. T., Ramage, D., Amin, N., Schwikowski, B., and Ideker, T. (2003). Cytoscape : a software environment for integrated models of biomolecular interaction networks. *Genome Res*, 13(11) :2498–2504.

[Sherman et al., 2006] Sherman, D., Durrens, P., Iragne, F., Beyne, E., Nikolski, M., and Souciet, J. L. (2006). Genolevures complete genomes provide data and tools for comparative genomics of hemiascomycetous yeasts. *Nucleic Acids Res*, 34(Database issue) :432–435.

[Sneath and Sokal, 1973] Sneath, P. and Sokal, R. (1973). *Numerical taxonomy*. W. H. Freeman.

[Snel et al., 1999] Snel, B., Bork, P., and Huynen, M. A. (1999). Genome phylogeny based on gene content. *Nat Genet*, 21(1) :108–110.

[Souciet J., 2000] Souciet J., e. a. (2000). Génolevures : Genomic exploration of the hemiascomycetous yeasts. *FEBS Lett*, 487(1) :1–149.

[Soularue, 2005] Soularue, J.-P. (2005). Inférence de réseaux d'interactions biomoléculaires à partir de sources de données hétérogènes. Master's thesis, Université Bordeaux 1, 351 cours de la libération - 33400 Talence.

[Sprinzak et al., 2003] Sprinzak, E., Sattath, S., and Margalit, H. (2003). How reliable are experimental protein-protein interaction data ? *J Mol Biol*, 327(5) :919–923.

[Strong and Eisenberg, 2007] Strong, M. and Eisenberg, D. (2007). The protein network as a tool for finding novel drug targets. *Prog Drug Res*, 64 :193–215.

[Stumpf et al., 2005] Stumpf, M. P., Wiuf, C., and May, R. M. (2005). Subnets of scale-free networks are not scale-free : sampling properties of networks. *Proc Natl Acad Sci U S A*, 102(12) :4221–4224.

[Tapiero et al., 2002] Tapiero, H., Mathé, G., Couvreur, P., and Tew, K. D. (2002). I. arginine. *Biomed Pharmacother*, 56(9) :439–445.

[Uetz et al., 2000] Uetz, P., Giot, L., Cagney, G., Mansfield, T., Judson, R., Knight, J., Lockshon, D., Narayan, V., Srinivasan, M., Pochart, P., Qureshi-Emili, A., Li, Y., Godwin, B., Conover, D., Kalbfleisch, T., Vijayadamodar, G., Yang, M., Johnston, M., Fields, S., and Rothberg, J. (2000). A comprehensive analysis of protein-protein interactions in saccharomyces cerevisiae. *Nature*, 403(6770) :623–627.

[Uetz and Pankratz, 2004] Uetz, P. and Pankratz, M. (2004). Protein interaction maps on the fly. *Nat Biotechnol*, 22(1) :43–44.

[Vidal et al., 1996] Vidal, M., Brachmann, R. K., Fattaey, A., Harlow, E., and Boeke, J. D. (1996). Reverse two-hybrid and one-hybrid systems to detect dissociation of protein-protein and dna-protein interactions. *Proc Natl Acad Sci U S A*, 93(19) :10315–10320.

[von Mering et al., 2002] von Mering, C., Krause, R., Snel, B., Cornell, M., Oliver, S., Fields, S., and Bork, P. (2002). Comparative assessment of large-scale data sets of protein-protein interactions. *Nature*, 417(6887) :399–403.

[Wagner, 2001] Wagner, A. (2001). The yeast protein interaction network evolves rapidly and contains few redundant duplicate genes. *Mol Biol Evol*, 18(7) :1283–1292.

[Wagner and Fell, 2001] Wagner, A. and Fell, D. A. (2001). The small world inside large metabolic networks. *Proc Biol Sci*, 268(1478) :1803–1810.

[Wang et al., 2006] Wang, B., Chen, P., Huang, D. S., Li, J. J., Lok, T. M., and Lyu, M. R. (2006). Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett*, 580(2) :380–384.

[Ward et al., 2004] Ward, J. J., Sodhi, J. S., McGuffin, L. J., Buxton, B. F., and Jones, D. T. (2004). Prediction and functional analysis of native disorder in proteins from the three kingdoms of life. *J Mol Biol*, 337(3) :635–645.

[Watts and Strogatz, 1998] Watts, D. and Strogatz, S. (1998). Collective dynamics of 'small-world' networks. *Nature*, 393(6684) :440–442.

[Westerhoff and Palsson, 2004] Westerhoff, H. V. and Palsson, B. O. (2004). The evolution of molecular biology into systems biology. *Nat Biotechnol*, 22(10) :1249–1252.

[Winzeler et al., 1999] Winzeler, E., Shoemaker, D., Astromoff, A., Liang, H., Anderson, K., Andre, B., Bangham, R., Benito, R., Boeke, J., Bussey, H., Chu, A., Connelly, C., Davis, K., Dietrich, F., Dow, S., El Bakkoury, M., Foury, F., Friend, S., Gentalen, E., Giaever, G., Hegemann, J., Jones, T., Laub, M., Liao, H., Liebundguth, N., Lockhart, D., Lucau-Danila, A., Lussier, M., M'Rabet, N., Menard, P., Mittmann, M., Pai, C., Rebischung, C., Revuelta, J., Riles, L., Roberts, C., Ross-MacDonald, P., Scherens, B., Snyder, M., Sookhai-Mahadeo, S., Storms, R., Véronneau, S., Voet, M., Volckaert, G., Ward, T., Wysocki, R., Yen, G., Yu, K., Zimmermann, K., Philippsen, P., Johnston, M., and Davis, R. (1999). Functional characterization of the s. cerevisiae genome by gene deletion and parallel analysis. *Science*, 285(5429) :901–906.

[Wittig and De Beuckelaer, 2001] Wittig, U. and De Beuckelaer, A. (2001). Analysis and comparison of metabolic pathway databases. *Brief Bioinform*, 2(2) :126–142.

[Woese et al., 1990] Woese, C. R., Kandler, O., and Wheelis, M. L. (1990). Towards a natural system of organisms : proposal for the domains archaea, bacteria, and eucarya. *Proc Natl Acad Sci U S A*, 87(12) :4576–4579.

[Wu et al., 2006] Wu, X., Zhu, L., Guo, J., Zhang, D. Y., and Lin, K. (2006). Prediction of yeast protein-protein interaction network : insights from the gene ontology and annotations. *Nucleic Acids Res*, 34(7) :2137–2150.

[Wuchty, 2006] Wuchty, S. (2006). Topology and weights in a protein domain interaction network–a novel way to predict protein interactions. *BMC Genomics*, 7 :122–122.

[Xenarios et al., 2000] Xenarios, I., Rice, D., Salwinski, L., Baron, M., Marcotte, E., and Eisenberg, D. (2000). DIP : the database of interacting proteins. *Nucleic Acids Res*, 28(1) :289–291.

[Yamanishi et al., 2004] Yamanishi, Y., Vert, J., and Kanehisa, M. (2004). Protein network inference from multiple genomic data : a supervised approach. *Bioinformatics*, 20 Suppl 1 :363–363.

[Yamanishi et al., 2005] Yamanishi, Y., Vert, J., and Kanehisa, M. (2005). Supervised enzyme network inference from the integration of genomic data and chemical information. *Bioinformatics*, 21 Suppl 1 :468–468.

[Yook et al., 2004] Yook, S., Oltvai, Z., and Barabási, A. (2004). Functional and topological characterization of protein interaction networks. *Proteomics*, 4(4) :928–942.

[Yu et al., 2006] Yu, H., Paccanaro, A., Trifonov, V., and Gerstein, M. (2006). Predicting interactions in protein networks by completing defective cliques. *Bioinformatics*, 22(7) :823–829.

[Zdobnov and Apweiler, 2001] Zdobnov, E. M. and Apweiler, R. (2001). Interproscan–an integration platform for the signature-recognition methods in interpro. *Bioinformatics*, 17(9) :847–848.

[Zhang et al., 2004] Zhang, L., Wong, S., King, O., and Roth, F. (2004). Predicting co-complexed protein pairs using genomic and proteomic data integration. *BMC Bioinformatics*, 5 :38–38.

Annexe A

Extrapolation de voies métaboliques

Fig. 32 – Répartition chromosomique des protéines spécifiques à *D. hansenii* et non-présentes chez les autres levures hémiascomycètes. Chaque chromosome est indentifié par son nom. Les protéines spécifiques sont en rouge, les universellement conservées, en bleu. Un pixel en largeur représente un gène. Cette figure semble indiquer un biais de représentation des gènes spécifiques aux extrémités des chrommosomes.

0	orf19.1099 ; orf19.1208 ; orf19.1234 ; orf19.3490 ; orf19.4485 ; orf19.4712 ; orf19.5191 ; orf19.5315 ; orf19.5775 ; orf19.6896 ; orf19.6960 ; orf19.7005 ; orf19.727 ; orf19.914 ;
1	orf19.1580 ; orf19.4366 ; orf19.4391 ; orf19.4881 ; orf19.4916 ; orf19.4921 ; orf19.5592 ; orf19.6449 ;
2	orf19.2839 ; orf19.3492 ; orf19.3820 ; orf19.4918 ; orf19.4919 ; orf19.6608 ; orf19.7301 ;
3	orf19.4552 ; orf19.6300 ; orf19.6301 ; orf19.6962 ; orf19.6963 ; orf19.6964 ; orf19.6965 ;
4	orf19.3903 ; orf19.3905 ; orf19.3906 ; orf19.3908 ; orf19.4691 ;
5	orf19.2371 ; orf19.2375 ; orf19.5372 ;
6	orf19.4484 ; orf19.6897 ; orf19.7004 ;
7	orf19.1013 ; orf19.5289 ;
8	orf19.1130 ; orf19.4482 ;
9	orf19.1162 ; orf19.993 ;
10	orf19.11 ; orf19.2620 ;
11	orf19.1207 ; orf19.1817 ;
12	orf19.2291 ; orf19.3719 ;
13	orf19.2428.2 ; orf19.5735.3 ;
14	orf19.280 ; orf19.296 ;
15	orf19.2813 ; orf19.70 ;
16	orf19.3375 ; orf19.3378 ;
17	orf19.3510 ; orf19.5284 ;
18	orf19.3726 ; orf19.4068 ;
19	orf19.3923 ; orf19.4689 ;
20	orf19.4402 ; orf19.6115 ;
21	orf19.4486 ; orf19.5190 ;
22	orf19.4652 ; orf19.4653 ;
23	orf19.4702 ; orf19.5057 ;
24	orf19.4880 ; orf19.4917 ;
25	orf19.5205 ; orf19.6219 ;
26	orf19.5283 ; orf19.636 ;
27	orf19.6474 ; orf19.6475 ;
28	orf19.6569 ; orf19.6570 ;
29	orf19.7455 ; orf19.7456 ;

Tab. 7 – Liste des familles de protéines spécifiques à *C. albicans*. Chaque famille est identifiée par un numéro et est associée à la liste des protéines qu'elle regroupe.

0	DEHA1E12111g DEHA1E12122g DEHA1E12133g DEHA1E12144g DEHA1E12166g DEHA1E12177g DEHA1E12232g DEHA1E12243g DEHA1E12265g DEHA1E12276g DEHA1G00374g DEHA1E12287g DEHA1G00352g DEHA1E12298g DEHA1E12309g DEHA1E12320g DEHA1E12353g DEHA1E12364g
1	DEHA1C09053g DEHA1C09152g DEHA1C09218g DEHA1D00473g DEHA1E00583g
2	DEHA1A04070g DEHA1A04081g DEHA1B02233g DEHA1D03443g
3	DEHA1B00121g DEHA1D00594g DEHA1D10230g DEHA1E00187g
4	DEHA1B02013g DEHA1C03377g DEHA1C03388g DEHA1D04642g
5	DEHA1B08129g DEHA1C09020g DEHA1C09141g DEHA1D00836g
6	DEHA1A02739g DEHA1D06875g DEHA1D10021g DEHA1B00143g
7	DEHA1C00209g DEHA1D00330g
8	DEHA1B02101g DEHA1C06226g DEHA1C06908g
9	DEHA1C07392g DEHA1C07403g DEHA1C07414g
10	DEHA1C07887g DEHA1C07920g DEHA1C07931g
11	DEHA1D00242g DEHA1E00220g DEHA1E00264g
12	DEHA1D10219g DEHA1F00198g DEHA1F14179g
13	DEHA1A05720g DEHA1B05786g
14	DEHA1A07073g DEHA1F11121g
15	DEHA1A07359g DEHA1E00308g
16	DEHA1B00660g DEHA1B00671g
17	DEHA1B04961g DEHA1B04972g
18	DEHA1C03201g DEHA1C03476g
19	DEHA1D00517g DEHA1D06820g
20	DEHA1D00693g DEHA1D07106g
21	DEHA1D10098g DEHA1E00253g
22	DEHA1D10252g DEHA1E12782g
23	DEHA1E00242g DEHA1F00176g
24	DEHA1F03333g DEHA1F03344g
25	DEHA1F05379g DEHA1F14135g
26	DEHA1F06039g DEHA1G00418g
27	DEHA1F07381g DEHA1F07392g
28	DEHA1F07953g DEHA1G01639g
29	DEHA1F08668g DEHA1F08679g
30	DEHA1G00341g DEHA1G00363g

TAB. 8 – Liste des familles de protéines spécifiques à *D. hansenii*. Chaque famille est identifiée par un numéro et est associée à la liste des protéines qu'elle regroupe.

CC Espèce	CC SACE	Entrées Espèce	Entrées SACE	Sorties Espèce	Sortie SACE	Métabolites isolés	% ports perdus
CAGL0	0	2	4	2	5	6/11	56%
KLLA0	0	4	4	5	5	0/11	0%
DEHA0	0	4	4	5	5	0/11	0%
YALI0	0	4	4	5	5	0/11	0%

TAB. 9 – Conservation des ports dans la voie de dégradation du lactose (données SGD) prédite pour les quatre espèces étudiées. CC espèce et CC SACE représentent respectivement les identifiants des composantes connexes dans l'organisme étudié et chez *S. cerevisiae*.

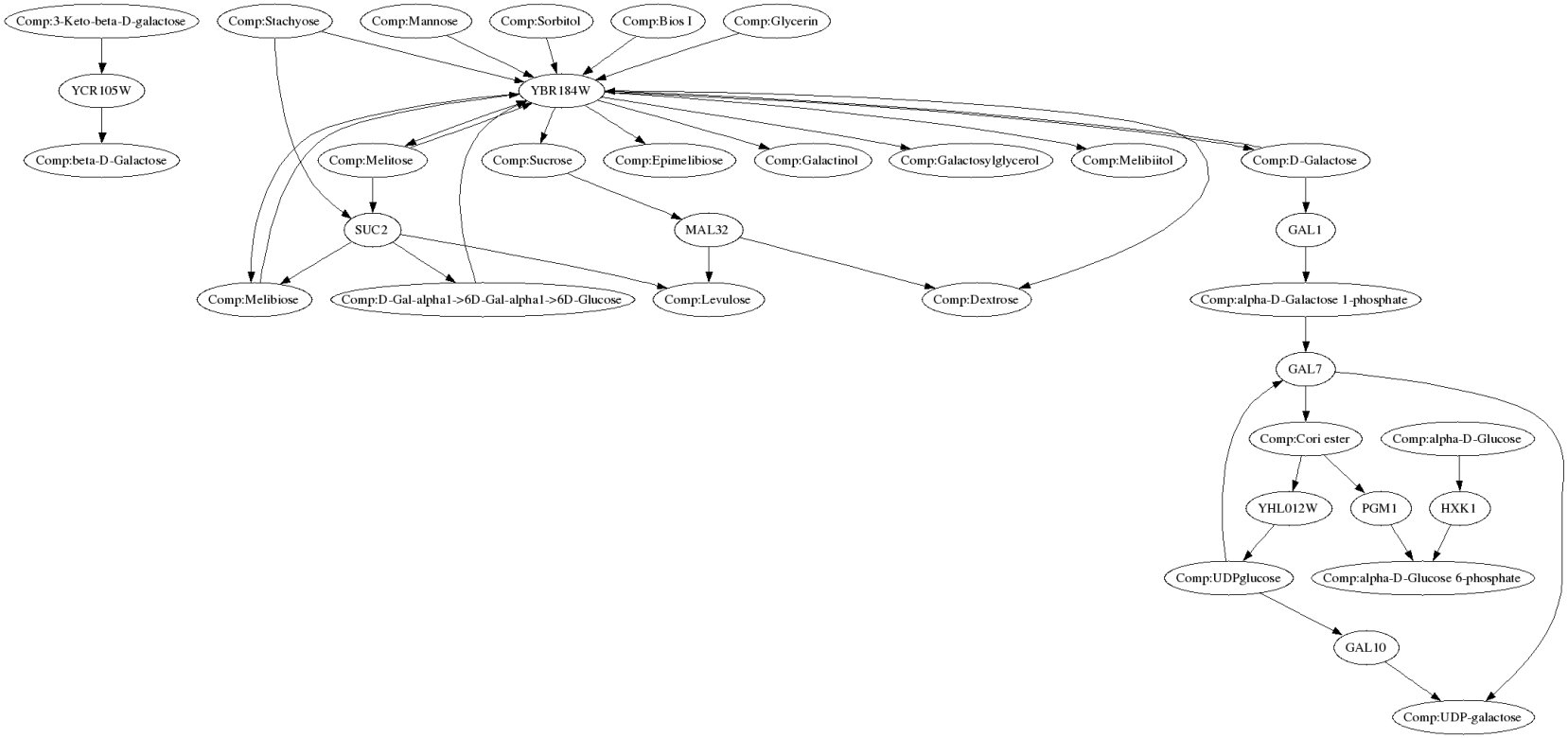


FIG. 33 – Graphe du métabolisme du galactose chez *S. cerevisiae* (données KEGG).

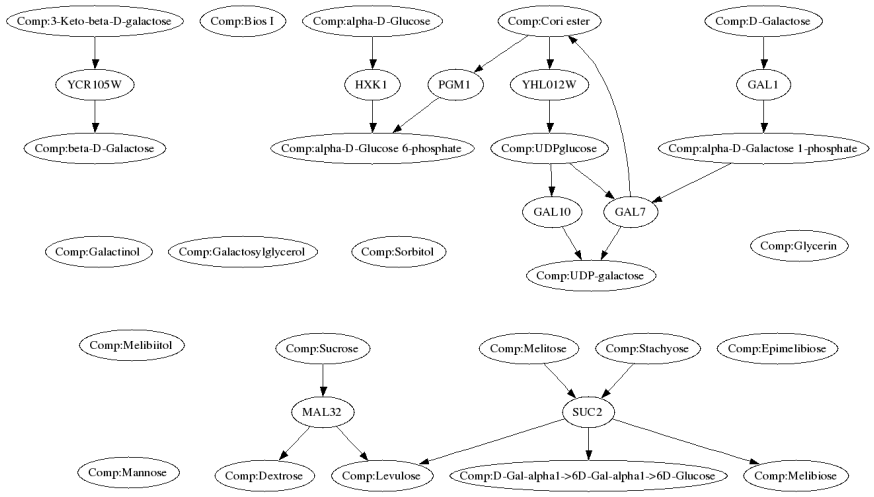


FIG. 34 – Graphe du métabolisme du galactose prédit chez *K. lactis* (données KEGG). Les critères que nous avons établis définissent cette voie comme possiblement perdue. L'observation du graphe nous permet de conclure que la voie semble conservée, puisque le galactose peut toujours être utilisé et que la conversion du raffinose (mélitose) est maintenue.

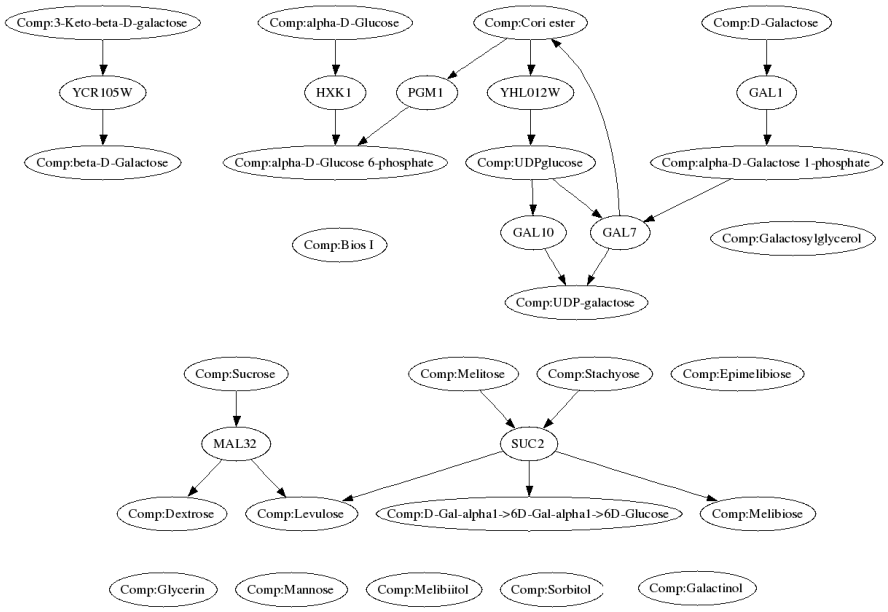


FIG. 35 – Graphe du métabolisme du galactose prédit chez *D. hansenii* (données KEGG). Les critères que nous avons établis définissent cette voie comme possiblement perdue. L'observation du graphe nous permet de conclure que la voie semble conservée, puisque le galactose peut toujours être utilisé et que la conversion du raffinose (mélitose) est maintenue.

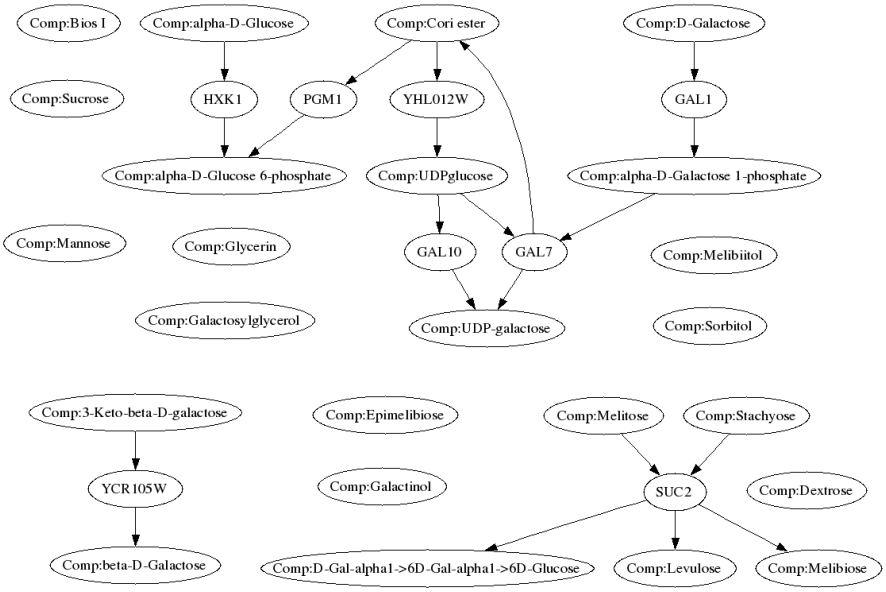


FIG. 36 – Graphe du métabolisme du galactose prédit chez *Y. lipolytica* (données KEGG). Les critères que nous avons établis définissent cette voie comme possiblement perdue. L'observation du graphe nous permet de conclure que la voie semble conservée, puisque le galactose peut toujours être utilisé et que la conversion du raffinose (mélitose) est maintenue.

Pathway Name	Pathway number	Average input degree	Average output degree	CC SACE	CC Species	Input SACE	Input species	Output SACE	Output species	I/O SACE	I/O Species	O/I SACE	O/I Species	Isolated compounds| Total compounds	% of ports lost
phospholipid biosynthesis (Kennedy pathway)	pathway10127	1.3	1.4	SACE0	CAGL0	4	4	5	5	4	4	2	2	0|11	0.0%
				SACE0	KLLA0	4	4	5	5	4	4	2	2		
				SACE0	DEHA0	4	4	5	5	4	4	2	2		
				SACE0	YALI0	4	4	5	5	4	4	2	2		
phospholipid biosynthesis	pathway10132	1.4	1.5	SACE0	CAGL0	1	1	2	2	0	0	1	1	0|18	0.0%
				SACE1	CAGL1	5	5	5	5	4	4	3	3		
				SACE0	KLLA0	1	1	2	2	0	0	1	1		
				SACE1	KLLA1	5	5	5	5	4	4	3	3		
				SACE0	DEHA0	1	1	2	2	0	0	1	1		
				SACE1	DEHA1	5	5	5	5	4	4	3	3		
				SACE0	YALI0	1	1	2	2	0	0	1	1		
				SACE1	YALI1	5	5	5	5	4	4	3	3		
mevalonate pathway	pathway10143	1.6	1.7	SACE0	CAGL0	2	2	4	4	1	1	2	2	0|11	0.0%
				SACE0	KLLA0	2	2	4	4	1	1	2	2		
				SACE0	DEHA0	2	2	4	4	1	1	2	2		
				SACE0	YALI0	2	2	4	4	1	1	2	2		
hexaprenyl diphosphate biosynthesis	pathway10261	1.8	1.8	SACE0	CAGL0	3	3	3	3	1	1	3	3	0|8	0.0%
				SACE0	KLLA0	3	3	3	3	1	1	3	3		
				SACE0	DEHA0	3	3	3	3	1	1	3	3		
				SACE0	YALI0	3	3	3	3	1	1	3	3		
fatty acid biosynthesis, initial steps	pathway10322	1.7	1.7	SACE0	CAGL0	2	2	2	2	2	2	0	0	0|7	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	0	0		
				SACE0	DEHA0	2	2	2	2	2	2	0	0		
				SACE0	YALI0	2	2	2	2	2	2	0	0		
fatty acid elongation, saturated	pathway10400	1.2	1.2	SACE0	CAGL0	1	1	1	1	1	1	1	1	0|4	0.0%
				SACE1	CAGL1	1	1	1	1	1	1	1	1		
				SACE0	KLLA0	1	1	1	1	1	1	1	1		
				SACE1	KLLA1	1	1	1	1	1	1	1	1		
				SACE0	DEHA0	1	1	1	1	1	1	1	1		
				SACE1	DEHA1	1	1	1	1	1	1	1	1		
				SACE0	YALI0	1	1	1	1	1	1	1	1		
				SACE1	YALI1	1	1	1	1	1	1	1	1		
ergosterol biosynthesis	pathway10417	1.5	1.6	SACE0	CAGL0	6	6	10	10	3	3	5	5	0|32	0.0%
				SACE0	KLLA0	6	6	10	10	3	3	5	5		
				SACE0	DEHA0	6	6	10	10	3	3	5	5		
				SACE0	YALI0	6	6	10	10	3	3	5	5		
biotin biosynthesis	pathway10585	1.4	1.2	SACE1	CAGL0	4	2	3	1	2	1	1	0	5|9	57.1%
				SACE0	KLLA0	4	4	3	3	2	2	1	1	0|9	0.0%
				SACE0	DEHA0	4	4	3	3	2	2	1	1	0|9	0.0%
				SACE0	YALI0	4	1	3	1	2	0	1	1	2|9	71.4%
				SACE1	YALI0	4	2	3	1	2	1	1	0	2|9	64.3%
thiamine biosynthesis	pathway10662	1.3	1.3	SACE0	CAGL0	2	2	2	2	0	0	1	1	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	0	0	1	1		
				SACE0	DEHA0	2	2	2	2	0	0	1	1		
				SACE0	YALI0	2	2	2	2	0	0	1	1		
folic acid biosynthesis	pathway10843	2.1	2.2	SACE0	CAGL0	8	8	10	10	7	7	5	5	0|28	0.0%
				SACE0	KLLA0	8	8	10	10	7	7	5	5	0|28	0.0%
				SACE0	DEHA0	8	6	10	9	7	5	5	4	4|28	16.7%
				SACE0	YALI0	8	8	10	9	7	7	5	4	1|28	5.6%
thioredoxin pathway	pathway10860	1.3	1.3	SACE0	CAGL0	1	1	1	1	1	1	1	1	0|2	0.0%
				SACE0	KLLA0	1	1	1	1	1	1	1	1		
				SACE0	DEHA0	1	1	1	1	1	1	1	1		
				SACE0	YALI0	1	1	1	1	1	1	1	1		
glutathione biosynthesis	pathway10883	2	1.4	SACE0	CAGL0	4	4	2	2	4	4	0	0	0|7	0.0%
				SACE0	KLLA0	4	4	2	2	4	4	0	0		
				SACE0	DEHA0	4	4	2	2	4	4	0	0		
				SACE0	YALI0	4	4	2	2	4	4	0	0		
NAD salvage pathway	pathway10906	1.7	1.9	SACE0	CAGL0	4	4	5	5	3	3	2	2	0|13	0.0%
				SACE0	KLLA0	4	4	5	5	3	3	2	2		
				SACE0	DEHA0	4	4	5	5	3	3	2	2		
				SACE0	YALI0	4	4	5	5	3	3	2	2		
NAD biosynthesis	pathway10985	1.9	2.1	SACE0	CAGL0	3	3	4	4	2	2	3	3	0|8	0.0%
				SACE0	KLLA0	3	3	4	4	2	2	3	3		
				SACE0	DEHA0	3	3	4	4	2	2	3	3		
				SACE0	YALI0	3	3	4	4	2	2	3	3		

TAB. 10 – Tableau des résultats complets pour la prédiction de voies métaboliques par analyse structurale de graphes (données SGD). La colonne "CC SACE" indique l'identifiant de la composante connexe chez *S. cerevisiae* et "CC species" indique l'identifiant de son équivalent dans l'espèce étudiée. I/O et O/I représentent les points d'interconnexion des voies. Les lignes en bleu clair représentent les voies métaboliques universellement conservées, en vert les voies conservées dans une ou plusieurs espèces, en bleu les possibles conservations, en violet les possibles pertes et en rouge, les pertes probables. Les lignes non colorées indiquent les cas ambigus. Ce tableau nous montre notamment que 126 voies sont universellement conservées chez les quatre espèces étudiées.

heme biosynthesis	pathway10989	1.1	1	SACE0	CAGL0	1	1	1	1	0	0	0	0	0|11	0.0%
				SACE1	CAGL1	1	1	0	0	0	0	0	0		
				SACE2	CAGL2	2	2	1	1	2	2	0	0		
				SACE0	KLLA0	1	1	1	1	0	0	0	0		
				SACE1	KLLA1	1	1	0	0	0	0	0	0		
				SACE2	KLLA2	2	2	1	1	2	2	0	0		
				SACE0	DEHA0	1	1	1	1	0	0	0	0		
				SACE1	DEHA1	1	1	0	0	0	0	0	0		
				SACE2	DEHA2	2	2	1	1	2	2	0	0		
				SACE0	YALI0	1	1	1	1	0	0	0	0		
				SACE1	YALI1	1	1	0	0	0	0	0	0		
				SACE2	YALI2	2	2	1	1	2	2	0	0		
riboflavin, FMN and FAD biosynthesis	pathway11121	1.3	1.4	SACE0	CAGL0	3	3	6	6	2	2	3	3	0|15	0.0%
				SACE0	KLLA0	3	3	6	6	2	2	3	3		
				SACE0	DEHA0	3	3	6	6	2	2	3	3		
				SACE0	YALI0	3	3	6	6	2	2	3	3		
de novo NAD biosynthesis	pathway11234	1.4	1.6	SACE2	CAGL0	2	0	4	1	2	0	3	1	9|20	50.0%
				SACE1	CAGL1	3	3	5	3	2	2	3	2		
				SACE2	KLLA0	2	0	4	1	2	0	3	1	9|20	50.0%
				SACE1	KLLA1	3	3	5	3	2	2	3	2		
				SACE0	DEHA0	2	2	4	4	2	2	3	3	2|20	14.3%
				SACE1	DEHA1	3	3	5	3	2	2	3	2		
				SACE0	YALI0	2	2	4	4	2	2	3	3	0|20	0.0%
				SACE1	YALI1	3	3	5	5	2	2	3	3		
ubiquinone Q prenylation	pathway11245	1.4	1.4	SACE0	CAGL0	3	3	3	3	1	1	1	1	0|13	0.0%
				SACE1	CAGL1	2	2	2	2	1	1	1	1		
				SACE0	KLLA0	3	3	3	3	1	1	1	1		
				SACE1	KLLA1	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	3	3	3	3	1	1	1	1		
				SACE1	DEHA1	2	2	2	2	1	1	1	1		
				SACE0	YALI0	3	3	3	3	1	1	1	1		
				SACE1	YALI1	2	2	2	2	1	1	1	1		
pantothenate and coenzyme A biosynthesis	pathway11348	1.7	1.7	SACE0	CAGL0	9	9	9	9	8	8	4	4	0|25	0.0%
				SACE0	KLLA0	9	9	9	9	8	8	4	4		
				SACE0	DEHA0	9	9	9	9	8	8	4	4		
				SACE0	YALI0	9	9	9	9	8	8	4	4		
formylTHF biosynthesis	pathway11475	2.7	2.4	SACE0	CAGL0	8	8	6	6	6	6	3	3	0|16	0.0%
				SACE0	KLLA0	8	8	6	6	6	6	3	3		
				SACE0	DEHA0	8	8	6	6	6	6	3	3		
				SACE0	YALI0	8	8	6	6	6	6	3	3		
glycerol teichoic acid biosynthesis	pathway11502	1.7	2	SACE0	CAGL0	2	2	3	3	1	1	3	3	0|5	0.0%
				SACE0	KLLA0	2	2	3	3	1	1	3	3		
				SACE0	DEHA0	2	2	3	3	1	1	3	3		
				SACE0	YALI0	2	2	3	3	1	1	3	3		
mannose and GDP-mannose metabolism	pathway11587	1	1	SACE0	CAGL0	1	1	1	1	0	0	1	1	0|3	0.0%
				SACE0	KLLA0	1	1	1	1	0	0	1	1		
				SACE0	DEHA0	1	1	1	1	0	0	1	1		
				SACE0	YALI0	1	1	1	1	0	0	1	1		
colanic acid building blocks biosynthesis	pathway11610	1.5	1.5	SACE0	CAGL0	3	1	3	1	1	0	1	1	5|12	50.0%
				SACE1	CAGL1	1	1	1	1	0	0	1	1		
				SACE0	KLLA0	3	3	3	3	1	1	1	1	0|12	0.0%
				SACE1	KLLA1	1	1	1	1	0	0	1	1		
				SACE0	DEHA0	3	3	3	3	1	1	1	1	0|12	0.0%
				SACE1	DEHA1	1	1	1	1	0	0	1	1		
				SACE0	YALI0	3	3	3	3	1	1	1	1	0|12	0.0%
				SACE1	YALI1	1	1	1	1	0	0	1	1		
trehalose biosynthesis	pathway11659	1.2	1.2	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|5	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	2	2	1	1		
trehalose catabolism	pathway11695	1	2	SACE0	CAGL0	1	1	0	0	1	1	0	0	0|1	0.0%
				SACE0	KLLA0	1	1	0	0	1	1	0	0		
				SACE0	DEHA0	1	1	0	0	1	1	0	0		
				SACE0	YALI0	1	1	0	0	1	1	0	0		
sucrose biosynthesis	pathway11712	1.6	1.6	SACE0	CAGL0	3	3	3	3	1	1	3	3	0|7	0.0%
				SACE0	KLLA0	3	3	3	3	1	1	3	3		
				SACE0	DEHA0	3	3	3	3	1	1	3	3		
				SACE0	YALI0	3	3	3	3	1	1	3	3		

Core lipid-linked oligosaccharide biosynthesis	pathway11735	1.3	1.3	SACE1	CAGL0	2	2	2	2	0	0	0	0	0|8	0.0%
				SACE0	CAGL1	2	2	2	2	2	2	0	0		
				SACE1	KLLA0	2	2	2	2	0	0	0	0		
				SACE0	KLLA1	2	2	2	2	2	2	0	0		
				SACE1	DEHA0	2	2	2	2	0	0	0	0		
				SACE0	DEHA1	2	2	2	2	2	2	0	0		
				SACE1	YALI0	2	2	2	2	0	0	0	0		
				SACE0	YALI1	2	2	2	2	2	2	0	0		
Lipid-linked oligosaccharide biosynthesis	pathway11777	1.5	1.5	SACE0	CAGL0	6	6	6	6	1	1	0	0	0|20	0.0%
				SACE1	CAGL1	1	1	1	1	0	0	0	0		
				SACE0	KLLA0	6	6	6	6	1	1	0	0		
				SACE1	KLLA1	1	1	1	1	0	0	0	0		
				SACE0	DEHA0	6	6	6	6	1	1	0	0		
				SACE1	DEHA1	1	1	1	1	0	0	0	0		
				SACE0	YALI0	6	6	6	6	1	1	0	0		
				SACE1	YALI1	1	1	1	1	0	0	0	0		
gluconeogenesis	pathway11898	2	2	SACE0	CAGL0	6	6	6	6	3	3	2	2	0|16	0.0%
				SACE0	KLLA0	6	6	6	6	3	3	2	2		
				SACE0	DEHA0	6	6	6	6	3	3	2	2		
				SACE0	YALI0	6	6	6	6	3	3	2	2		
starch and cellulose biosynthesis	pathway11927	1.6	1.8	SACE0	CAGL0	2	2	3	3	1	1	3	3	0|7	0.0%
				SACE1	CAGL1	1	1	1	1	0	0	0	0		
				SACE0	KLLA0	2	2	3	3	1	1	3	3		
				SACE1	KLLA1	1	1	1	1	0	0	0	0		
				SACE0	DEHA0	2	2	3	3	1	1	3	3		
				SACE1	DEHA1	1	1	1	1	0	0	0	0		
				SACE0	YALI0	2	2	3	3	1	1	3	3		
				SACE1	YALI1	1	1	1	1	0	0	0	0		
valine biosynthesis	pathway11978	1.6	1.6	SACE0	CAGL0	3	3	3	3	2	2	2	2	0|8	0.0%
				SACE0	KLLA0	3	3	3	3	2	2	2	2		
				SACE0	DEHA0	3	3	3	3	2	2	2	2		
				SACE0	YALI0	3	3	3	3	2	2	2	2		
tyrosine biosynthesis	pathway12026	1.6	1.6	SACE0	CAGL0	3	3	3	3	3	3	2	2		
				SACE0	KLLA0	3	3	3	3	3	3	2	2		
				SACE0	DEHA0	3	3	3	3	3	3	2	2		
				SACE0	YALI0	3	3	3	3	3	3	2	2		
tryptophan biosynthesis	pathway12032	1.2	1.2	SACE0	CAGL0	2	2	2	2	1	1	1	1	0|9	0.0%
				SACE1	CAGL1	2	2	2	2	1	1	1	1		
				SACE0	KLLA0	2	2	2	2	1	1	1	1		
				SACE1	KLLA1	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	1	1		
				SACE1	DEHA1	2	2	2	2	1	1	1	1		
				SACE0	YALI0	2	2	2	2	1	1	1	1		
				SACE1	YALI1	2	2	2	2	1	1	1	1		
threonine biosynthesis	pathway12041	1.6	1.6	SACE0	CAGL0	4	4	4	4	2	2	2	2	0|10	0.0%
				SACE0	KLLA0	4	4	4	4	2	2	2	2		
				SACE0	DEHA0	4	4	4	4	2	2	2	2		
				SACE0	YALI0	4	4	4	4	2	2	2	2		
threonine biosynthesis from homoserine	pathway12069	1.2	1.2	SACE0	CAGL0	2	2	2	2	1	1	1	1	0|5	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	1	1		
				SACE0	YALI0	2	2	2	2	1	1	1	1		
serine biosynthesis	pathway12075	1.8	1.8	SACE0	CAGL0	4	4	4	4	2	2	3	3	0|8	0.0%
				SACE0	KLLA0	4	4	4	4	2	2	3	3		
				SACE0	DEHA0	4	4	4	4	2	2	3	3		
				SACE0	YALI0	4	4	4	4	2	2	3	3		
proline biosynthesis	pathway12123	1.5	1.3	SACE0	CAGL0	4	4	3	3	2	2	1	1	0|9	0.0%
				SACE0	KLLA0	4	4	3	3	2	2	1	1		
				SACE0	DEHA0	4	4	3	3	2	2	1	1		
				SACE0	YALI0	4	4	3	3	2	2	1	1		
arginine degradation (anaerobic)	pathway12149	1.2	1.4	SACE0	CAGL0	2	2	2	2	1	1	2	2	0|10	0.0%
				SACE1	CAGL1	2	2	3	3	2	2	1	1		
				SACE0	KLLA0	2	2	2	2	1	1	2	2		
				SACE1	KLLA1	2	2	3	3	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	2	2		
				SACE1	DEHA1	2	2	3	3	2	2	1	1		
				SACE0	YALI0	2	2	2	2	1	1	2	2		
				SACE1	YALI1	2	2	3	3	2	2	1	1		
phenylalanine biosynthesis	pathway12156	1.5	1.5	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|6	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	2	2	1	1		

methionine biosynthesis	pathway12163	1.4	1.5	SACE0	CAGL0	6	6	7	7	3	3	2	2	0|16	0.0%
				SACE0	KLLA0	6	6	7	7	3	3	2	2		
				SACE0	DEHA0	6	6	7	7	3	3	2	2		
				SACE0	YALI0	6	6	7	7	3	3	2	2		
homoserine methionine biosynthesis	pathway12178	1.2	1.4	SACE0	CAGL0	3	3	4	4	1	1	1	1	0|9	0.0%
				SACE0	KLLA0	3	3	4	4	1	1	1	1		
				SACE0	DEHA0	3	3	4	4	1	1	1	1		
				SACE0	YALI0	3	3	4	4	1	1	1	1		
S-adenosylhomocysteine catabolism	pathway12183	1	1.5	SACE0	CAGL0	1	1	2	2	1	1	2	2	0|3	0.0%
				SACE0	KLLA0	1	1	2	2	1	1	2	2		
				SACE0	DEHA0	1	1	2	2	1	1	2	2		
				SACE0	YALI0	1	1	2	2	1	1	2	2		
lysine biosynthesis	pathway12194	1.7	1.8	SACE0	CAGL0	6	6	7	7	2	2	4	4	0|16	0.0%
				SACE0	KLLA0	6	6	7	7	2	2	4	4	0|16	0.0%
				SACE0	DEHA0	6	6	7	6	2	2	4	3	1|16	7.7%
				SACE0	YALI0	6	6	7	7	2	2	4	4	0|16	0.0%
leucine biosynthesis	pathway12312	1.7	1.8	SACE0	CAGL0	4	4	5	5	2	2	4	4	0|11	0.0%
				SACE0	KLLA0	4	4	5	5	2	2	4	4		
				SACE0	DEHA0	4	4	5	5	2	2	4	4		
				SACE0	YALI0	4	4	5	5	2	2	4	4		
isoleucine biosynthesis	pathway12358	1.5	1.5	SACE0	CAGL0	3	3	3	3	2	2	2	2	0|10	0.0%
				SACE1	CAGL1	1	1	1	1	1	1	0	0		
				SACE0	KLLA0	3	3	3	3	2	2	2	2		
				SACE1	KLLA1	1	1	1	1	1	1	0	0		
				SACE0	DEHA0	3	3	3	3	2	2	2	2		
				SACE1	DEHA1	1	1	1	1	1	1	0	0		
				SACE0	YALI0	3	3	3	3	2	2	2	2		
				SACE1	YALI1	1	1	1	1	1	1	0	0		
histidine biosynthesis	pathway12395	1.3	1.5	SACE0	CAGL0	3	3	6	6	2	2	4	4	0|19	0.0%
				SACE0	KLLA0	3	3	6	6	2	2	4	4		
				SACE0	DEHA0	3	3	6	6	2	2	4	4		
				SACE0	YALI0	3	3	6	6	2	2	4	4		
glycine biosynthesis from threonine	pathway12409	1	1.5	SACE0	CAGL0	1	1	2	2	1	1	1	1	0|3	0.0%
				SACE0	KLLA0	1	1	2	2	1	1	1	1		
				SACE0	DEHA0	1	1	2	2	1	1	1	1		
				SACE0	YALI0	1	1	2	2	1	1	1	1		
glycine biosynthesis from serine	pathway12420	2	2	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
glycine biosynthesis from alanine	pathway12423	1.3	1.3	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
glutamine biosynthesis	pathway12434	1.3	1.3	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	2	2	1	1		
superpathway of glutamate biosynthesis	pathway12444	2.3	2.3	SACE0	CAGL0	5	5	5	5	5	5	4	4	0|10	0.0%
				SACE0	KLLA0	5	5	5	5	5	5	4	4	0|10	0.0%
				SACE0	DEHA0	5	4	5	4	5	4	4	3	2|10	20.0%
				SACE0	YALI0	5	5	5	5	5	5	4	4	0|10	0.0%
glutamate biosynthesis from ammonia	pathway12488	2	2	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
glutamate biosynthesis from glutamine	pathway12493	1	2	SACE0	CAGL0	1	1	3	3	1	1	3	3	0|4	0.0%
				SACE0	KLLA0	1	1	3	3	1	1	3	3	0|4	0.0%
				SACE0	NA	1	0	3	0	1	0	3	0	4|4	###
				SACE0	YALI0	1	1	3	3	1	1	3	3	0|4	0.0%
homocysteine and cysteine interconversion	pathway12496	1.3	1.4	SACE0	CAGL0	2	2	3	3	1	1	2	2	0|8	0.0%
				SACE0	KLLA0	2	2	3	3	1	1	2	2		
				SACE0	DEHA0	2	2	3	3	1	1	2	2		
				SACE0	YALI0	2	2	3	3	1	1	2	2		
cysteine biosynthesis from homoserine	pathway12509	1.2	1.3	SACE0	CAGL0	3	3	4	4	2	2	2	2	0|10	0.0%
				SACE0	KLLA0	3	3	4	4	2	2	2	2		
				SACE0	DEHA0	3	3	4	4	2	2	2	2		
				SACE0	YALI0	3	3	4	4	2	2	2	2		
aspartate biosynthesis	pathway12516	2.3	2	SACE0	CAGL0	4	4	3	3	3	3	1	1	0|7	0.0%
				SACE0	KLLA0	4	4	3	3	3	3	1	1		
				SACE0	DEHA0	4	4	3	3	3	3	1	1		
				SACE0	YALI0	4	4	3	3	3	3	1	1		

asparagine biosynthesis	pathway12521	2.4	2.8	SACE0	CAGL0	4	4	5	5	3	3	3	3	0|9	0.0%
				SACE0	KLLA0	4	4	5	5	3	3	3	3		
				SACE0	DEHA0	4	4	5	5	3	3	3	3		
				SACE0	YALI0	4	4	5	5	3	3	3	3		
arginine biosynthesis	pathway12537	1.5	1.6	SACE0	CAGL0	7	7	8	8	4	4	4	4	0|21	0.0%
				SACE0	KLLA0	7	7	8	8	4	4	4	4	0|21	0.0%
				SACE0	DEHA0	7	7	8	8	4	4	4	4	0|21	0.0%
				SACE0	YALI0	7	0	8	2	4	0	4	2	3|21	60.0%
				SACE1	YALI0	7	6	8	4	4	4	4	1		
alanine biosynthesis	pathway12623	1.3	1.3	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
serine and glycine biosynthesis	pathway12626	1.8	1.8	SACE0	CAGL0	5	5	5	5	3	3	4	4	0|11	0.0%
				SACE0	KLLA0	5	5	5	5	3	3	4	4		
				SACE0	DEHA0	5	5	5	5	3	3	4	4		
				SACE0	YALI0	5	5	5	5	3	3	4	4		
sulfur amino acid biosynthesis	pathway12632	1.5	1.7	SACE0	CAGL0	6	6	10	10	4	4	5	5	0|23	0.0%
				SACE0	KLLA0	6	6	10	10	4	4	5	5		
				SACE0	DEHA0	6	6	10	10	4	4	5	5		
				SACE0	YALI0	6	6	10	10	4	4	5	5		
methionine and S-adenosylmethionine synthesis	pathway12658	1.7	1.7	SACE0	CAGL0	3	3	3	3	2	2	2	2	0|7	0.0%
				SACE0	KLLA0	3	3	3	3	2	2	2	2		
				SACE0	DEHA0	3	3	3	3	2	2	2	2		
				SACE0	YALI0	3	3	3	3	2	2	2	2		
superpathway of phenylalanine, tyrosine, and tryptophan biosynthesis	pathway12662	1.5	1.6	SACE0	CAGL0	3	3	4	4	3	3	2	2	0|19	0.0%
				SACE1	CAGL1	2	2	2	2	1	1	1	1		
				SACE2	CAGL2	2	2	2	2	1	1	1	1		
				SACE0	KLLA0	3	3	4	4	3	3	2	2		
				SACE1	KLLA1	2	2	2	2	1	1	1	1		
				SACE2	KLLA2	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	3	3	4	4	3	3	2	2		
				SACE1	DEHA1	2	2	2	2	1	1	1	1		
				SACE2	DEHA2	2	2	2	2	1	1	1	1		
				SACE0	YALI0	3	3	4	4	3	3	2	2		
				SACE1	YALI1	2	2	2	2	1	1	1	1		
				SACE2	YALI2	2	2	2	2	1	1	1	1		
superpathway of phenylalanine and tyrosine biosynthesis	pathway12674	1.7	1.8	SACE0	CAGL0	3	3	4	4	3	3	2	2	0|10	0.0%
				SACE0	KLLA0	3	3	4	4	3	3	2	2		
				SACE0	DEHA0	3	3	4	4	3	3	2	2		
				SACE0	YALI0	3	3	4	4	3	3	2	2		
threonine and methionine biosynthesis	pathway12681	1.4	1.6	SACE0	CAGL0	6	6	8	8	3	3	3	3	0|18	0.0%
				SACE0	KLLA0	6	6	8	8	3	3	3	3		
				SACE0	DEHA0	6	6	8	8	3	3	3	3		
				SACE0	YALI0	6	6	8	8	3	3	3	3		
superpathway of isoleucine and valine biosynthesis	pathway12691	1.7	1.7	SACE0	CAGL0	4	4	4	4	2	2	2	2	0|14	0.0%
				SACE1	CAGL1	1	1	1	1	1	1	0	0		
				SACE0	KLLA0	4	4	4	4	2	2	2	2		
				SACE1	KLLA1	1	1	1	1	1	1	0	0		
				SACE0	DEHA0	4	4	4	4	2	2	2	2		
				SACE1	DEHA1	1	1	1	1	1	1	0	0		
				SACE0	YALI0	4	4	4	4	2	2	2	2		
				SACE1	YALI1	1	1	1	1	1	1	0	0		
homoserine biosynthesis	pathway12703	1.7	1.5	SACE0	CAGL0	4	4	3	3	2	2	1	1	0|8	0.0%
				SACE0	KLLA0	4	4	3	3	2	2	1	1		
				SACE0	DEHA0	4	4	3	3	2	2	1	1		
				SACE0	YALI0	4	4	3	3	2	2	1	1		
phenylalanine, tyrosine and tryptophan biosynthesis, complete	pathway12708	1.6	1.7	SACE0	CAGL0	5	5	7	7	3	3	2	2	0|29	0.0%
				SACE1	CAGL1	2	2	2	2	1	1	1	1		
				SACE2	CAGL2	2	2	2	2	1	1	1	1		
				SACE0	KLLA0	5	5	7	7	3	3	2	2		
				SACE1	KLLA1	2	2	2	2	1	1	1	1		
				SACE2	KLLA2	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	5	5	7	7	3	3	2	2		
				SACE1	DEHA1	2	2	2	2	1	1	1	1		
				SACE2	DEHA2	2	2	2	2	1	1	1	1		
				SACE0	YALI0	5	5	7	7	3	3	2	2		
				SACE1	YALI1	2	2	2	2	1	1	1	1		
				SACE2	YALI2	2	2	2	2	1	1	1	1		

superpathway of leucine, isoleucine, and valine biosynthesis	pathway12727	1.8	1.8	SACE0	CAGL0	7	7	7	7	3	3	4	4	0|22	0.0%
				SACE1	CAGL1	1	1	1	1	1	1	0	0		
				SACE0	KLLA0	7	7	7	7	3	3	4	4		
				SACE1	KLLA1	1	1	1	1	1	1	0	0		
				SACE0	DEHA0	7	7	7	7	3	3	4	4		
				SACE1	DEHA1	1	1	1	1	1	1	0	0		
				SACE0	YALI0	7	7	7	7	3	3	4	4		
				SACE1	YALI1	1	1	1	1	1	1	0	0		
chorismate biosynthesis	pathway12742	1.6	1.8	SACE0	CAGL0	4	4	5	5	2	2	2	2	0|13	0.0%
				SACE0	KLLA0	4	4	5	5	2	2	2	2		
				SACE0	DEHA0	4	4	5	5	2	2	2	2		
				SACE0	YALI0	4	4	5	5	2	2	2	2		
ubiquinone biosynthesis	pathway12751	1.6	1.5	SACE1	CAGL0	4	4	3	3	1	1	2	2	0|19	0.0%
				SACE0	CAGL1	3	3	3	3	1	1	1	1		
				SACE1	KLLA0	4	4	3	3	1	1	2	2		
				SACE0	KLLA1	3	3	3	3	1	1	1	1		
				SACE1	DEHA0	4	4	3	3	1	1	2	2		
				SACE0	DEHA1	3	3	3	3	1	1	1	1		
				SACE1	YALI0	4	4	3	3	1	1	2	2		
				SACE0	YALI1	3	3	3	3	1	1	1	1		
superpathway of glycine biosynthesis	pathway12813	1.9	1.7	SACE0	CAGL0	5	5	4	4	5	5	3	3	0|9	0.0%
				SACE0	KLLA0	5	5	4	4	5	5	3	3		
				SACE0	DEHA0	5	5	4	4	5	5	3	3		
				SACE0	YALI0	5	5	4	4	5	5	3	3		
glycogen biosynthesis	pathway12818	2.2	2.2	SACE0	CAGL0	2	2	2	2	1	1	0	0	0|5	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	0	0		
				SACE0	DEHA0	2	2	2	2	1	1	0	0		
				SACE0	YALI0	2	2	2	2	1	1	0	0		
protein modifications	pathway12860	2.4	2.5	SACE0	CAGL0	6	6	7	7	3	3	2	2	0|13	0.0%
				SACE0	KLLA0	6	6	7	7	3	3	2	2		
				SACE0	DEHA0	6	6	7	7	3	3	2	2		
				SACE0	YALI0	6	6	7	7	3	3	2	2		
phosphatidic acid synthesis	pathway12949	1.3	1.1	SACE0	CAGL0	3	3	2	2	1	1	2	2	0|6	0.0%
				SACE0	KLLA0	3	3	2	2	1	1	2	2		
				SACE0	DEHA0	3	3	2	2	1	1	2	2		
				SACE0	YALI0	3	3	2	2	1	1	2	2		
periplasmic NAD degradation	pathway12955	1	1.5	SACE0	CAGL0	1	1	2	2	1	1	0	0	0|3	0.0%
				SACE0	KLLA0	1	1	2	2	1	1	0	0		
				SACE0	DEHA0	1	1	2	2	1	1	0	0		
				SACE0	YALI0	1	1	2	2	1	1	0	0		
citrulline biosynthesis	pathway12961	1.5	1	SACE0	CAGL0	2	2	1	1	0	0	0	0	0|3	0.0%
				SACE0	KLLA0	2	2	1	1	0	0	0	0		
				SACE0	DEHA0	2	2	1	1	0	0	0	0		
				SACE0	YALI0	2	2	1	1	0	0	0	0		
acetate utilization	pathway12979	2.4	2.4	SACE0	CAGL0	3	3	3	3	2	2	2	2	0|6	0.0%
				SACE0	KLLA0	3	3	3	3	2	2	2	2		
				SACE0	DEHA0	3	3	3	3	2	2	2	2		
				SACE0	YALI0	3	3	3	3	2	2	2	2		
superpathway of histidine, purine, and pyrimidine biosynthesis	pathway4104	2.2	2.2	SACE0	CAGL0	14	14	14	14	5	5	5	5	0|56	0.0%
				SACE0	KLLA0	14	14	14	14	5	5	5	5		
				SACE0	DEHA0	14	14	14	14	5	5	5	5		
				SACE0	YALI0	14	14	14	14	5	5	5	5		
phosphatidic acid and phospholipid biosynthesis	pathway4922	1.5	1.5	SACE0	CAGL0	9	9	8	8	7	7	4	4	0|29	0.0%
				SACE1	CAGL1	1	1	2	2	0	0	1	1		
				SACE0	KLLA0	9	9	8	8	7	7	4	4		
				SACE1	KLLA1	1	1	2	2	0	0	1	1		
				SACE0	DEHA0	9	9	8	8	7	7	4	4		
				SACE1	DEHA1	1	1	2	2	0	0	1	1		
				SACE0	YALI0	9	9	8	8	7	7	4	4		
				SACE1	YALI1	1	1	2	2	0	0	1	1		
pathways of chorismate	pathway5208	1.9	2.1	SACE0	CAGL0	12	12	18	18	7	7	7	7	0|49	0.0%
				SACE0	KLLA0	12	12	18	18	7	7	7	7	0|49	0.0%
				SACE0	DEHA0	12	11	18	17	7	6	7	6	3|49	6.7%
				SACE0	YALI0	12	12	18	17	7	7	7	6	1|49	3.3%
TCA cycle, aerobic respiration	pathway5704	1.8	1.7	SACE0	CAGL0	5	5	4	4	5	5	3	3	0|15	0.0%
				SACE1	CAGL1	1	1	0	0	1	1	0	0		
				SACE0	KLLA0	5	5	4	4	5	5	3	3		
				SACE1	KLLA1	1	1	0	0	1	1	0	0		
				SACE0	DEHA0	5	5	4	4	5	5	3	3		
				SACE1	DEHA1	1	1	0	0	1	1	0	0		
				SACE0	YALI0	5	5	4	4	5	5	3	3		
				SACE1	YALI1	1	1	0	0	1	1	0	0		

pyruvate dehydrogenase	pathway5933	1	1.2	SACE0	CAGL0	1	1	1	1	1	1	1	1	0|5	0.0%
				SACE1	CAGL1	1	1	1	1	1	1	0	0		
				SACE2	CAGL2	1	1	0	0	1	1	0	0		
				SACE0	KLLA0	1	1	1	1	1	1	1	1		
				SACE1	KLLA1	1	1	1	1	1	1	0	0		
				SACE2	KLLA2	1	1	0	0	1	1	0	0		
				SACE0	DEHA0	1	1	1	1	1	1	1	1		
				SACE1	DEHA1	1	1	1	1	1	1	0	0		
				SACE2	DEHA2	1	1	0	0	1	1	0	0		
				SACE0	YALI0	1	1	1	1	1	1	1	1		
				SACE1	YALI1	1	1	1	1	1	1	0	0		
				SACE2	YALI2	1	1	0	0	1	1	0	0		
pentose phosphate pathway	pathway5977	1.7	1.7	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|12	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
oxidative branch of the pentose phosphate pathway	pathway6118	1.8	1.8	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|6	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
non-oxidative branch of the pentose phosphate pathway	pathway6123	1.6	1.6	SACE0	CAGL0	1	1	1	1	1	1	1	1	0|7	0.0%
				SACE0	KLLA0	1	1	1	1	1	1	1	1		
				SACE0	DEHA0	1	1	1	1	1	1	1	1		
				SACE0	YALI0	1	1	1	1	1	1	1	1		
glycolysis	pathway6132	2	2	SACE0	CAGL0	5	5	5	5	4	4	1	1	0|14	0.0%
				SACE0	KLLA0	5	5	5	5	4	4	1	1		
				SACE0	DEHA0	5	5	5	5	4	4	1	1		
				SACE0	YALI0	5	5	5	5	4	4	1	1		
purine fermentation	pathway6277	2.1	1.7	SACE0	CAGL0	1	1	1	1	0	0	0	0	0|10	0.0%
				SACE1	CAGL1	5	5	3	3	4	4	2	2		
				SACE0	KLLA0	1	1	1	1	0	0	0	0	0|10	0.0%
				SACE1	KLLA1	5	5	3	3	4	4	2	2		
				SACE0	DEHA1	5	5	3	3	4	4	2	2	2|10	20.0%
				SACE0	YALI1	5	5	3	3	4	4	2	2	2|10	20.0%
glutathione-glutaredoxin redox reactions	pathway6413	2	1.4	SACE2	CAGL1	1	1	1	1	1	1	1	1	4|7	66.0%
				SACE0	KLLA0	2	2	2	2	0	0	0	0	0|7	0.0%
				SACE1	KLLA1	1	1	1	1	1	1	1	1	0|7	0.0%
				SACE0	DEHA0	2	2	2	2	0	0	0	0	0|7	0.0%
				SACE1	DEHA1	1	1	1	1	1	1	1	1	0|7	0.0%
				SACE2	YALI1	1	1	1	1	1	1	1	1	4|7	66.0%
glucose fermentation	pathway6502	2.2	2.2	SACE0	CAGL0	8	7	8	8	5	5	3	3	1|22	6.2%
				SACE0	KLLA0	8	8	8	8	5	5	3	3	0|22	0.0%
				SACE0	DEHA0	8	7	8	8	5	5	3	3	1|22	6.2%
				SACE0	YALI0	8	8	8	8	5	5	3	3	0|22	0.0%
sucrose degradation	pathway6624	1	1.5	SACE0	NA	1	0	2	0	0	0	1	0	3|3	###
				SACE0	KLLA0	1	1	2	2	0	0	1	1	0|3	0.0%
				SACE0	DEHA0	1	1	2	2	0	0	1	1	0|3	0.0%
				SACE0	YALI0	1	1	2	2	0	0	1	1	0|3	0.0%
lactose degradation	pathway6664	1.7	1.8	SACE0	CAGL0	4	2	5	2	2	1	2	2	6|11	55.6%
				SACE0	KLLA0	4	4	5	5	2	2	2	2	0|11	0.0%
				SACE0	DEHA0	4	4	5	5	2	2	2	2	0|11	0.0%
				SACE0	YALI0	4	4	5	5	2	2	2	2	0|11	0.0%
galactose metabolism	pathway6768	1.5	1.6	SACE1	CAGL0	3	0	4	1	2	0	1	1	7|9	85.7%
				SACE0	KLLA0	3	3	4	4	2	2	1	1	0|9	0.0%
				SACE0	DEHA0	3	3	4	4	2	2	1	1	0|9	0.0%
				SACE0	YALI0	3	3	4	4	2	2	1	1	0|9	0.0%
UDP-glucose conversion	pathway6779	1.5	1.7	SACE0	CAGL0	2	2	1	0	1	1	0	0	1|5	33.3%
				SACE0	KLLA0	2	2	1	1	1	1	0	0	0|5	0.0%
				SACE0	DEHA0	2	2	1	1	1	1	0	0	0|5	0.0%
				SACE0	YALI0	2	2	1	1	1	1	0	0	0|5	0.0%
mannose catabolism	pathway6797	1	1	SACE0	CAGL0	1	1	1	1	0	0	1	1	0|2	0.0%
				SACE0	KLLA0	1	1	1	1	0	0	1	1		
				SACE0	DEHA0	1	1	1	1	0	0	1	1		
				SACE0	YALI0	1	1	1	1	0	0	1	1		
glycogen catabolism	pathway6822	1.2	2	SACE0	CAGL0	1	1	4	4	0	0	2	2	0|6	0.0%
				SACE0	KLLA0	1	1	4	4	0	0	2	2		
				SACE0	DEHA0	1	1	4	4	0	0	2	2		
				SACE0	YALI0	1	1	4	4	0	0	2	2		
glucose 1-phosphate metabolism	pathway6877	1.7	2	SACE0	CAGL0	2	2	3	3	1	1	3	3	0|5	0.0%
				SACE0	KLLA0	2	2	3	3	1	1	3	3		
				SACE0	DEHA0	2	2	3	3	1	1	3	3		
				SACE0	YALI0	2	2	3	3	1	1	3	3		

sorbitol degradation	pathway6898	1.6	1.8	SACE0	CAGL0	2	2	3	3	2	2	0	0	0|6	0.0%
				SACE0	KLLA0	2	2	3	3	2	2	0	0		
				SACE0	DEHA0	2	2	3	3	2	2	0	0		
				SACE0	YALI0	2	2	3	3	2	2	0	0		
mannitol degradation	pathway6913	1.6	1.8	SACE0	CAGL0	2	2	3	3	2	2	0	0	0|6	0.0%
				SACE0	KLLA0	2	2	3	3	2	2	0	0		
				SACE0	DEHA0	2	2	3	3	2	2	0	0		
				SACE0	YALI0	2	2	3	3	2	2	0	0		
hexitol degradation super-pathway	pathway6925	1.6	1.8	SACE0	CAGL0	2	2	3	3	2	2	0	0	0|6	0.0%
				SACE0	KLLA0	2	2	3	3	2	2	0	0		
				SACE0	DEHA0	2	2	3	3	2	2	0	0		
				SACE0	YALI0	2	2	3	3	2	2	0	0		
glucuronate degradation (to xylulose 5-phosphate)	pathway6931	1.3	1.3	SACE0	CAGL0	3	1	3	2	2	1	1	0	3|7	50.0%
				SACE0	KLLA0	3	3	3	3	2	2	1	1	0|7	0.0%
				SACE0	DEHA0	3	3	3	3	2	2	1	1	0|7	0.0%
				SACE0	YALI0	3	3	3	3	2	2	1	1	0|7	0.0%
methylglyoxal catabolism	pathway7003	1.5	2.2	SACE0	CAGL0	1	1	3	3	0	0	0	0	0|4	0.0%
				SACE0	KLLA0	1	1	3	3	0	0	0	0		
				SACE0	DEHA0	1	1	3	3	0	0	0	0		
				SACE0	YALI0	1	1	3	3	0	0	0	0		
polyamine degradation	pathway7052	1.7	1.7	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|5	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	2	2	1	1		
acrylonitrile degradation	pathway7105	1	1	SACE0	NA	1	0	1	0	0	0	0	0	2|2	###
				SACE0	KLLA0	1	1	1	1	0	0	0	0	0|2	0.0%
				SACE0	DEHA0	1	1	1	1	0	0	0	0	0|2	0.0%
				SACE0	YALI0	1	1	1	1	0	0	0	0	0|2	0.0%
octane oxidation	pathway7142	3.8	3.8	SACE0	CAGL0	3	3	3	3	1	1	1	1	0|6	0.0%
				SACE0	KLLA0	3	3	3	3	1	1	1	1		
				SACE0	DEHA0	3	3	3	3	1	1	1	1		
				SACE0	YALI0	3	3	3	3	1	1	1	1		
sulfur degradation	pathway7250	1.1	1.4	SACE0	CAGL0	2	2	3	3	0	0	1	1	0|12	0.0%
				SACE1	CAGL1	1	1	2	2	1	1	1	1		
				SACE2	CAGL2	1	1	2	2	0	0	2	2		
				SACE0	KLLA0	2	2	3	3	0	0	1	1		
				SACE1	KLLA1	1	1	2	2	1	1	1	1		
				SACE2	KLLA2	1	1	2	2	0	0	2	2		
				SACE0	DEHA0	2	2	3	3	0	0	1	1		
				SACE1	DEHA1	1	1	2	2	1	1	1	1		
				SACE2	DEHA2	1	1	2	2	0	0	2	2		
				SACE0	YALI0	2	2	3	3	0	0	1	1		
				SACE1	YALI1	1	1	2	2	1	1	1	1		
				SACE2	YALI2	1	1	2	2	0	0	2	2		
sulfate assimilation pathway II	pathway7378	1.2	1.6	SACE0	CAGL0	2	2	4	4	2	2	2	2	0|11	0.0%
				SACE1	CAGL1	1	1	2	2	0	0	2	2		
				SACE0	KLLA0	2	2	4	4	2	2	2	2		
				SACE1	KLLA1	1	1	2	2	0	0	2	2		
				SACE0	DEHA0	2	2	4	4	2	2	2	2		
				SACE1	DEHA1	1	1	2	2	0	0	2	2		
				SACE0	YALI0	2	2	4	4	2	2	2	2		
				SACE1	YALI1	1	1	2	2	0	0	2	2		
fatty acid oxidation pathway	pathway7444	2.6	2.4	SACE0	CAGL0	5	5	4	4	2	2	3	3	0|13	0.0%
				SACE0	KLLA0	5	5	4	4	2	2	3	3	0|13	0.0%
				SACE0	DEHA0	5	5	4	4	2	2	3	3	0|13	0.0%
				SACE0	YALI0	5	4	4	4	2	2	3	3	1|13	11.1%
lipases biosynthesis	pathway7525	1.3	2.3	SACE0	CAGL0	3	3	8	8	1	1	3	3	0|11	0.0%
				SACE0	KLLA0	3	3	8	8	1	1	3	3	0|11	0.0%
				SACE1	DEHA0	3	3	8	5	1	1	3	3	3|11	27.3%
				SACE1	YALI0	3	3	8	5	1	1	3	3	3|11	27.3%
glycerol biosynthesis	pathway7601	1.8	2.2	SACE0	CAGL0	1	1	2	2	1	1	1	1	0|3	0.0%
				SACE0	KLLA0	1	1	2	2	1	1	1	1		
				SACE0	DEHA0	1	1	2	2	1	1	1	1		
				SACE0	YALI0	1	1	2	2	1	1	1	1		
formaldehyde oxidation II (glutathione-dependent)	pathway7638	1.2	1.2	SACE2	CAGL0	3	1	3	2	1	1	1	1	3|8	50.0%
				SACE2	KLLA0	3	1	3	2	1	1	1	1		
				SACE2	DEHA0	3	1	3	2	1	1	1	1		
				SACE2	YALI0	3	1	3	2	1	1	1	1		

serine-isocitrate lyase pathway	pathway7679	1.6	1.7	SACE0	CAGL0	1	1	1	1	0	0	0	0	0|18	0.0%
				SACE1	CAGL1	3	3	4	4	2	2	2	2		
				SACE2	CAGL2	2	2	2	2	2	2	2	2		
				SACE0	KLLA0	1	1	1	1	0	0	0	0		
				SACE1	KLLA1	3	3	4	4	2	2	2	2		
				SACE2	KLLA2	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	1	1	1	1	0	0	0	0		
				SACE1	DEHA1	3	3	4	4	2	2	2	2		
				SACE2	DEHA2	2	2	2	2	2	2	2	2		
				SACE0	YALI0	1	1	1	1	0	0	0	0		
				SACE1	YALI1	3	3	4	4	2	2	2	2		
				SACE2	YALI2	2	2	2	2	2	2	2	2		
carbon monoxide dehydrogenase pathway	pathway7744	2.8	2.2	SACE0	CAGL0	5	5	3	3	4	4	2	2	0|10	0.0%
				SACE0	KLLA0	5	5	3	3	4	4	2	2		
				SACE0	DEHA0	5	5	3	3	4	4	2	2		
				SACE0	YALI0	5	5	3	3	4	4	2	2		
toluene degradation, via catechol	pathway7801	3	3	SACE0	CAGL0	2	2	2	2	1	1	1	1	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	1	1		
				SACE0	YALI0	2	2	2	2	1	1	1	1		
p-cymene degradation	pathway7896	3	3	SACE0	CAGL0	2	2	2	2	1	1	1	1	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	1	1		
				SACE0	YALI0	2	2	2	2	1	1	1	1		
aldoxime metabolism	pathway8007	1	1	SACE0	NA	1	0	1	0	0	0	0	0	2|2	0.0%
				SACE0	KLLA0	1	1	1	1	0	0	0	0	0|2	0.0%
				SACE0	DEHA0	1	1	1	1	0	0	0	0	0|2	0.0%
				SACE0	YALI0	1	1	1	1	0	0	0	0	0|2	0.0%
m-cresol degradation	pathway8040	3	3	SACE0	CAGL0	2	2	2	2	1	1	1	1	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	1	1		
				SACE0	YALI0	2	2	2	2	1	1	1	1		
valine degradation	pathway8120	2	2	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	2	2	1	1		
tryptophan degradation via kynurenine	pathway8158	1.1	1.3	SACE1	CAGL0	2	0	4	1	2	0	3	1	11|14	90.0%
				SACE1	KLLA0	2	0	4	1	2	0	3	1	11|14	90.0%
				SACE0	DEHA0	2	2	4	4	2	2	3	3	4|14	40.0%
				SACE0	YALI0	2	2	4	4	2	2	3	3	0|14	0.0%
				SACE1	YALI1	2	2	2	2	1	1	1	1		
tryptophan degradation	pathway8272	2.5	2.5	SACE0	CAGL0	2	1	2	0	1	0	1	0	3|5	75.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1	0|5	0.0%
				SACE0	DEHA0	2	1	2	0	1	0	1	0	3|5	75.0%
				SACE0	YALI0	2	2	2	2	1	1	1	1	0|5	0.0%
threonine catabolism	pathway8330	1.2	1.2	SACE0	CAGL0	2	2	2	2	1	1	1	1	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	1	1		
				SACE0	YALI0	2	2	2	2	1	1	1	1		
L-serine degradation	pathway8450	1	1	SACE0	CAGL0	1	1	1	1	1	1	1	1	0|2	0.0%
				SACE0	KLLA0	1	1	1	1	1	1	1	1		
				SACE0	DEHA0	1	1	1	1	1	1	1	1		
				SACE0	YALI0	1	1	1	1	1	1	1	1		
proline utilization	pathway8457	1.3	1.3	SACE0	CAGL0	3	3	3	3	3	3	2	2	0|7	0.0%
				SACE0	KLLA0	3	3	3	3	3	3	2	2		
				SACE0	DEHA0	3	3	3	3	3	3	2	2		
				SACE0	YALI0	3	3	3	3	3	3	2	2		
phenylalanine degradation	pathway8506	1.6	1.7	SACE0	CAGL0	2	2	3	2	1	1	1	1	1|6	20.0%
				SACE0	KLLA0	2	2	3	3	1	1	1	1	0|6	0.0%
				SACE0	DEHA0	2	2	3	2	1	1	1	1	1|6	20.0%
				SACE0	YALI0	2	2	3	3	1	1	1	1	0|6	0.0%
leucine degradation	pathway8618	1.7	2	SACE0	CAGL0	2	2	3	3	1	1	1	1	0|5	0.0%
				SACE0	KLLA0	2	2	3	3	1	1	1	1		
				SACE0	DEHA0	2	2	3	3	1	1	1	1		
				SACE0	YALI0	2	2	3	3	1	1	1	1		
isoleucine degradation	pathway8660	1.8	2	SACE0	CAGL0	2	2	3	3	1	1	1	1	0|5	0.0%
				SACE0	KLLA0	2	2	3	3	1	1	1	1		
				SACE0	DEHA0	2	2	3	3	1	1	1	1		
				SACE0	YALI0	2	2	3	3	1	1	1	1		
homocysteine degradation	pathway8701	1.2	1.2	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|5	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	2	2	1	1		

glycine degradation	pathway8713	2.7	2.5	SACE0	CAGL0	1	1	0	0	1	1	0	0	0|10	0.0%
				SACE1	CAGL1	4	4	3	3	4	4	2	2		
				SACE0	KLLA0	1	1	0	0	1	1	0	0		
				SACE1	KLLA1	4	4	3	3	4	4	2	2		
				SACE0	DEHA0	1	1	0	0	1	1	0	0		
				SACE1	DEHA1	4	4	3	3	4	4	2	2		
				SACE0	YALI0	1	1	0	0	1	1	0	0		
				SACE1	YALI1	4	4	3	3	4	4	2	2		
glutamate degradation I	pathway8761	1.2	1.2	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|7	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	2	2	1	1		
asparagine degradation	pathway8804	1.8	1.8	SACE0	CAGL0	2	2	2	2	2	2	1	1	0|5	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	1	1	0|5	0.0%
				SACE0	DEHA0	2	2	2	2	2	2	1	1	0|5	0.0%
				SACE0	YALI0	2	1	2	2	2	1	1	1	1|5	25.0%
arginine degradation (aerobic)	pathway8854	1.4	1.5	SACE0	CAGL0	5	5	6	6	5	5	3	3	0|14	0.0%
				SACE0	KLLA0	5	5	6	6	5	5	3	3		
				SACE0	DEHA0	5	5	6	6	5	5	3	3		
				SACE0	YALI0	5	5	6	6	5	5	3	3		
alanine degradation	pathway8914	1.3	1.3	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
glutamate degradation IX	pathway8924	1.3	1.3	SACE0	CAGL0	2	2	2	2	2	2	2	2	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	2	2	2	2		
				SACE0	DEHA0	2	2	2	2	2	2	2	2		
				SACE0	YALI0	2	2	2	2	2	2	2	2		
4-aminobutyrate degradation	pathway8935	1.3	1.3	SACE0	CAGL0	3	3	3	3	2	2	2	2	0|7	0.0%
				SACE0	KLLA0	3	3	3	3	2	2	2	2		
				SACE0	DEHA0	3	3	3	3	2	2	2	2		
				SACE0	YALI0	3	3	3	3	2	2	2	2		
dissimilation of N-acetylglucosamine, N-acetylmannosamine and N-acetylneuraminic acid	pathway8939	1.3	1.3	SACE0	CAGL0	2	2	2	2	0	0	2	2	0|8	0.0%
				SACE1	CAGL1	2	2	2	2	2	2	1	1		
				SACE0	KLLA0	2	2	2	2	0	0	2	2		
				SACE1	KLLA1	2	2	2	2	2	2	1	1		
				SACE0	DEHA0	2	2	2	2	0	0	2	2		
				SACE1	DEHA1	2	2	2	2	2	2	1	1		
				SACE0	YALI0	2	2	2	2	0	0	2	2		
				SACE1	YALI1	2	2	2	2	2	2	1	1		
citrulline degradation	pathway9016	1.5	1	SACE0	CAGL0	2	2	1	1	0	0	0	0	0|3	0.0%
				SACE0	KLLA0	2	2	1	1	0	0	0	0		
				SACE0	DEHA0	2	2	1	1	0	0	0	0		
				SACE0	YALI0	2	2	1	1	0	0	0	0		
allantoin degradation	pathway9035	1.3	1.3	SACE0	CAGL0	2	1	2	1	1	1	1	0	4|8	50.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1	0|8	0.0%
				SACE0	DEHA0	2	2	2	2	1	1	1	1	0|8	0.0%
				SACE0	YALI0	2	2	2	2	1	1	1	1	0|8	0.0%
glycerol degradation	pathway9106	1.3	1.3	SACE0	CAGL0	3	3	3	3	3	3	0	0	0|7	0.0%
				SACE0	KLLA0	3	3	3	3	3	3	0	0		
				SACE0	DEHA0	3	3	3	3	3	3	0	0		
				SACE0	YALI0	3	3	3	3	3	3	0	0		
UDP-N-acetylglucosamine biosynthesis	pathway9125	1.3	1.3	SACE0	CAGL0	4	4	4	4	3	3	3	3	0|11	0.0%
				SACE0	KLLA0	4	4	4	4	3	3	3	3		
				SACE0	DEHA0	4	4	4	4	3	3	3	3		
				SACE0	YALI0	4	4	4	4	3	3	3	3		
UDP-N-acetylgalactosamine biosynthesis	pathway9147	1.3	1.3	SACE0	CAGL0	4	4	4	4	3	3	3	3	0|11	0.0%
				SACE0	KLLA0	4	4	4	4	3	3	3	3		
				SACE0	DEHA0	4	4	4	4	3	3	3	3		
				SACE0	YALI0	4	4	4	4	3	3	3	3		
polyamine biosynthesis	pathway9163	1.3	1.3	SACE0	CAGL0	2	2	2	2	1	1	1	1	0|7	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	1	1		
				SACE0	DEHA0	2	2	2	2	1	1	1	1		
				SACE0	YALI0	2	2	2	2	1	1	1	1		
ectoine synthesis	pathway9228	1.6	1.3	SACE0	CAGL0	4	4	3	3	2	2	1	1	0|7	0.0%
				SACE0	KLLA0	4	4	3	3	2	2	1	1		
				SACE0	DEHA0	4	4	3	3	2	2	1	1		
				SACE0	YALI0	4	4	3	3	2	2	1	1		
salvage pathways of pyrimidine deoxyribonucleotides	pathway9286	1.1	1.8	SACE0	CAGL0	2	2	5	5	1	1	0	0	0|7	0.0%
				SACE0	KLLA0	2	2	5	5	1	1	0	0		
				SACE0	DEHA0	2	2	5	5	1	1	0	0		
				SACE0	YALI0	2	2	5	5	1	1	0	0		

salvage pathways of pyrimidine ribonucleotides	pathway9402	1.6	1.8	SACE0	CAGL0	5	5	7	7	2	2	2	2	0|15	0.0%
				SACE0	KLLA0	5	5	7	7	2	2	2	2		
				SACE0	DEHA0	5	5	7	7	2	2	2	2		
				SACE0	YALI0	5	5	7	7	2	2	2	2		
de novo biosynthesis of pyrimidine deoxyribonucleotides	pathway9484	1.5	1.8	SACE0	CAGL0	3	3	5	5	2	2	1	1	0|13	0.0%
				SACE0	KLLA0	3	3	5	5	2	2	1	1		
				SACE0	DEHA0	3	3	5	5	2	2	1	1		
				SACE0	YALI0	3	3	5	5	2	2	1	1		
salvage pathways of guanine, xanthine, and their nucleosides	pathway9594	1.6	1.6	SACE0	CAGL0	3	3	3	3	1	1	1	1	0|6	0.0%
				SACE0	KLLA0	3	3	3	3	1	1	1	1		
				SACE0	DEHA0	3	3	3	3	1	1	1	1		
				SACE0	YALI0	3	3	3	3	1	1	1	1		
salvage pathways of adenine, hypoxanthine, and their nucleosides	pathway9650	1.7	1.7	SACE0	CAGL0	4	4	4	4	3	3	1	1	0|10	0.0%
				SACE0	KLLA0	4	4	4	4	3	3	1	1		
				SACE0	DEHA0	4	4	4	4	3	3	1	1		
				SACE0	YALI0	4	4	4	4	3	3	1	1		
de novo biosynthesis of pyrimidine ribonucleotides	pathway9745	2	2	SACE0	CAGL0	8	8	8	8	4	4	3	3	0|18	0.0%
				SACE0	KLLA0	8	8	8	8	4	4	3	3		
				SACE0	DEHA0	8	8	8	8	4	4	3	3		
				SACE0	YALI0	8	8	8	8	4	4	3	3		
de novo biosynthesis of purine nucleotides	pathway9757	2.1	2.1	SACE0	CAGL0	10	10	9	9	6	6	5	5	0|33	0.0%
				SACE0	KLLA0	10	10	9	9	6	6	5	5		
				SACE0	DEHA0	10	10	9	9	6	6	5	5		
				SACE0	YALI0	10	10	9	9	6	6	5	5		
ppGpp metabolism	pathway9784	2.4	2.4	SACE0	CAGL0	2	2	2	2	1	1	0	0	0|4	0.0%
				SACE0	KLLA0	2	2	2	2	1	1	0	0		
				SACE0	DEHA0	2	2	2	2	1	1	0	0		
				SACE0	YALI0	2	2	2	2	1	1	0	0		
fatty acid elongation, unsaturated	pathway9825	1.3	1.3	SACE0	CAGL0	1	1	1	1	1	1	1	1	0|2	0.0%
				SACE0	KLLA0	1	1	1	1	1	1	1	1		
				SACE0	DEHA0	1	1	1	1	1	1	1	1		
				SACE0	YALI0	1	1	1	1	1	1	1	1		
triglyceride biosynthesis	pathway9880	1.2	1.4	SACE0	CAGL0	2	2	4	4	0	0	1	1	0|12	0.0%
				SACE1	CAGL1	1	1	1	1	0	0	1	1		
				SACE0	KLLA0	2	2	4	4	0	0	1	1		
				SACE1	KLLA1	1	1	1	1	0	0	1	1		
				SACE0	DEHA0	2	2	4	4	0	0	1	1		
				SACE1	DEHA1	1	1	1	1	0	0	1	1		
				SACE0	YALI0	2	2	4	4	0	0	1	1		
				SACE1	YALI1	1	1	1	1	0	0	1	1		
sphingolipid metabolism	pathway9897	2.1	2.1	SACE0	CAGL0	6	6	6	6	2	2	0	0	0|20	0.0%
				SACE0	KLLA0	6	6	6	6	2	2	0	0	0|20	0.0%
				SACE0	DEHA0	6	6	6	6	2	2	0	0	0|20	0.0%
				SACE0	YALI0	6	5	6	5	2	2	0	0	2|20	16.7%

Pathway Name	Pathway number	Average in degree	Average out degree	CC SACE	CC species	Input SACE	Input species	Output SACE	Output species	I/O SACE	I/O Species	O/I SACE	O/I Species	Isolated compounds| Total compounds	% of ports lost
Glycolysis/Gluconeogenesis	sce00010	1.3	1.2	0	CAGL0	3	2	2	2	0	0	0	0	1|23	12.5%
				1	CAGL1	2	2	1	1	0	0	0	0		
				0	KLLA0	3	3	2	2	0	0	0	0	0|23	0.0%
				1	KLLA1	2	2	1	1	0	0	0	0		
				0	DEHA0	3	2	2	2	0	0	0	0	1|23	12.5%
				1	DEHA1	2	2	1	1	0	0	0	0		
				0	YALI0	3	3	2	2	0	0	0	0	0|23	0.0%
				1	YALI1	2	2	1	1	0	0	0	0		
Citrate cycle (TCA cycle)	sce00020	1.4	1.4	0	CAGL0	4	4	5	5	4	4	2	2	0|20	0.0%
				0	KLLA0	4	4	5	5	4	4	2	2		
				0	DEHA0	4	4	5	5	4	4	2	2		
				0	YALI0	4	4	5	5	4	4	2	2		
Pentose phosphate pathway	sce00030	1.3	1.3	0	CAGL0	6	6	5	5	1	1	2	2	0|19	0.0%
				0	KLLA0	6	6	5	5	1	1	2	2		
				0	DEHA0	6	6	5	5	1	1	2	2		
				0	YALI0	6	6	5	5	1	1	2	2		
Pentose and glucuronate interconversions	sce00040	1.1	1.1	0	CAGL1	2	1	2	1	1	0	0	0	2|9	40.0%
				0	CAGL3	2	1	2	1	1	1	0	0		
				2	CAGL4	1	1	1	1	0	0	0	0		
				0	KLLA1	2	2	2	2	1	1	0	0	2|9	0.0%
				2	KLLA3	1	1	1	1	0	0	0	0		
				0	DEHA1	2	2	2	2	1	1	0	0	2|9	0.0%
				2	DEHA3	1	1	1	1	0	0	0	0		
				0	YALI1	2	2	2	2	1	1	0	0	2|9	0.0%
				2	YALI3	1	1	1	1	0	0	0	0		
Fructose and mannose metabolism	sce00051	1.2	1.2	0	CAGL0	5	2	5	1	0	0	1	0	0|18	0.0%
				0	CAGL1	5	3	5	4	0	0	1	1		
				0	KLLA0	5	5	5	5	0	0	1	1	0|18	0.0%
				0	DEHA0	5	5	5	5	0	0	1	1		
				0	YALI0	5	5	5	5	0	0	1	1		
Galactose metabolism	sce00052	1.6	1.7	0	CAGL2	6	1	8	1	0	0	1	1	18|24	75.0%
				1	CAGL0	1	1	1	1	0	0	0	0		
				0	KLLA0	6	1	8	2	0	0	1	0	8|24	73.3%
				0	KLLA2	6	1	8	2	0	0	1	1		
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	6	1	8	2	0	0	1	0	8|24	73.3%
				0	DEHA2	6	1	8	2	0	0	1	1		
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI1	6	1	8	1	0	0	1	0	10|24	76.7%
				0	YALI2	6	1	8	2	0	0	1	1		
				1	YALI0	1	1	1	1	0	0	0	0		
Ascorbate and aldarate metabolism	sce00053	1.0	1.0	0	CAGL0	1	1	1	1	0	0	0	0	0|4	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
Fatty acid biosynthesis	sce00061	2.2	2.2	0	CAGL0	2	2	2	2	1	1	1	1	0|36	0.0%
				0	KLLA0	2	2	2	2	1	1	1	1		
				0	DEHA0	2	2	2	2	1	1	1	1		
				0	YALI0	2	2	2	2	1	1	1	1		
Fatty acid metabolism	sce00071	1.5	2.1	0	CAGL0	8	8	15	15	2	2	1	1	1|28	3.7%
				1	CAGL1	2	1	2	2	1	1	0	0		
				0	KLLA0	8	8	15	15	2	2	1	1	0|28	0.0%
				1	KLLA1	2	2	2	2	1	1	0	0		
				0	DEHA0	8	8	15	15	2	2	1	1	1|28	3.7%
				1	DEHA1	2	1	2	2	1	1	0	0		
				0	YALI0	8	8	15	15	2	2	1	1	0|28	0.0%
				1	YALI1	2	2	2	2	1	1	0	0		
Synthesis and degradation of ketone bodies	sce00072	1.2	1.2	0	CAGL0	1	1	1	1	1	1	1	1	0|3	0.0%
				0	KLLA0	1	1	1	1	1	1	1	1		
				0	DEHA0	1	1	1	1	1	1	1	1		
				0	YALI0	1	1	1	1	1	1	1	1		
Biosynthesis of steroids	sce00100	1.2	1.2	0	CAGL0	1	1	2	2	1	1	0	0	0|15	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	2	2	1	1	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	2	2	1	1	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	2	2	1	1	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
Bile acid biosynthesis	sce00120	1.2	1.2	0	CAGL1	3	3	3	3	0	0	0	0	1|15	10.0%
				1	CAGL0	1	0	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA1	3	3	3	3	0	0	0	0	0|15	0.0%
				1	KLLA0	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA1	3	3	3	3	0	0	0	0	1|15	10.0%
				1	DEHA0	1	0	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI1	3	3	3	3	0	0	0	0	0|15	0.0%
				1	YALI0	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		

TAB. 11 – Tableau des résultats complets pour la prédiction de voies métaboliques par analyse structurale de graphes (données KEGG). La colonne "CC SACE" indique l'identifiant de la composante connexe chez *S. cerevisiae* et "CC species" indique l'identifiant de son équivalent dans l'espèce étudiée. I/O et O/I représentent les points d'interconnexion des voies. Les lignes en bleu clair représentent les voies métaboliques universellement conservées, en vert les voies conservées dans une ou plusieurs espèces, en bleu les possibles conservations, en violet les possibles pertes et en rouge, les pertes probables. Les lignes non colorées indiquent les cas ambigus. Ce tableau nous montre notamment que 43 voies sont universellement conservées chez les quatre espèces étudiées.

Ubiquinone biosynthesis	sce00130	1.1	1.1	0	CAGL0	2	2	2	1	0	0	1	1	2|13	14.3%
				1	CAGL1	2	2	1	1	0	0	0	0		
				0	KLLA0	2	2	2	1	0	0	1	1	2|13	14.3%
				1	KLLA1	2	2	1	1	0	0	0	0		
				0	DEHA0	2	2	2	2	0	0	1	1	0|13	0.0%
				1	DEHA1	2	2	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	2	2	2	1	0	0	1	1	2|13	14.3%
				1	YALI1	2	2	1	1	0	0	0	0		
Urea cycle and metabolism of amino groups	sce00220	1.2	1.2	0	CAGL0	4	4	5	5	3	3	2	2	0|20	0.0%
				0	KLLA0	4	4	5	5	3	3	2	2		
				0	DEHA0	4	4	5	5	3	3	2	2		
				0	YALI0	4	3	5	5	3	2	2	2	1|20	11.1%
Purine metabolism	sce00230	1.6	1.7	0	CAGL0	12	11	14	11	0	0	1	0	6|61	14.3%
				1	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	12	12	14	14	0	0	1	1	0|61	0.0%
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	12	12	14	14	0	0	1	1		
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI0	12	12	14	14	0	0	1	1	2|61	14.3%
Pyrimidine metabolism	sce00240	1.6	1.6	0	CAGL0	8	8	10	10	0	0	0	0	0|35	0.0%
				0	KLLA0	8	8	10	10	0	0	0	0		
				0	DEHA0	8	8	10	10	0	0	0	0		
				0	YALI0	8	5	10	7	0	0	0	0	0|35	0.0%
				0	YALI1	8	3	10	3	0	0	0	0		
Glutamate metabolism	sce00251	1.1	1.6	0	CAGL0	1	1	10	10	0	0	1	1	0|17	0.0%
				0	KLLA0	1	1	10	10	0	0	1	1		
				0	DEHA0	1	1	10	9	0	0	1	1	0|17	0.0%
				0	DEHA1	1	0	10	1	0	0	1	0		
				0	YALI0	1	1	10	10	0	0	1	1	0|17	0.0%
Alanine and aspartate metabolism	sce00252	1.1	1.4	0	CAGL0	1	1	8	8	0	0	1	1	0|17	0.0%
				1	CAGL1	1	1	1	1	1	1	0	0		
				0	KLLA0	1	1	8	8	0	0	1	1		
				1	KLLA1	1	1	1	1	1	1	0	0		
				0	DEHA0	1	1	8	7	0	0	1	1	0|17	0.0%
				0	DEHA2	1	0	8	1	0	0	1	0		
				1	DEHA1	1	1	1	1	1	1	0	0		
				0	YALI0	1	1	8	8	0	0	1	1	0|17	0.0%
				1	YALI1	1	1	1	1	1	1	0	0		
Glycine, serine and threonine metabolism	sce00260	1.2	1.4	0	CAGL0	4	4	13	13	1	1	5	5	0|36	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				3	CAGL3	1	1	1	1	1	1	0	0		
				0	KLLA0	4	4	13	13	1	1	5	5		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				3	KLLA3	1	1	1	1	1	1	0	0		
				0	DEHA0	4	4	13	13	1	1	5	5		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				3	DEHA3	1	1	1	1	1	1	0	0		
				0	YALI0	4	4	13	13	1	1	5	5		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
				3	YALI3	1	1	1	1	1	1	0	0		
Methionine metabolism	sce00271	1.3	1.3	0	CAGL0	5	5	3	3	2	2	1	1	0|13	0.0%
				0	KLLA0	5	5	3	3	2	2	1	1		
				0	DEHA0	5	5	3	3	2	2	1	1		
				0	YALI0	5	5	3	3	2	2	1	1		
Cysteine metabolism	sce00272	1.5	1.2	0	CAGL0	7	7	4	4	1	1	0	0	0|16	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	1	1	0	0		
				0	KLLA0	7	7	4	4	1	1	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	1	1	0	0		
				0	DEHA0	7	7	4	4	1	1	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	1	1	0	0		
				0	YALI0	7	7	4	4	1	1	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	1	1	0	0		
Valine, leucine and isoleucine degradation	sce00280	1.4	1.4	0	CAGL0	5	5	5	5	0	0	0	0	0|19	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	2	2	2	2	1	1	2	2		
				0	KLLA0	5	5	5	5	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	2	2	2	2	1	1	2	2		
				0	DEHA0	5	5	5	5	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	2	2	2	2	1	1	2	2		
				0	YALI0	5	5	5	5	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	2	2	2	2	1	1	2	2		
Valine, leucine and isoleucine biosynthesis	sce00290	1.3	1.5	0	CAGL0	6	6	8	8	2	2	0	0	0|22	0.0%
				0	KLLA0	6	6	8	8	2	2	0	0		
				0	DEHA0	6	6	8	8	2	2	0	0		
				0	YALI0	6	6	8	8	2	2	0	0		

Lysine biosynthesis	sce00300	1.3	1.2	0	CAGL0	5	5	4	4	1	1	1	1	0|19	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	1	1	0	0		
				0	KLLA0	5	5	4	4	1	1	1	1		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	1	1	0	0		
				0	DEHA0	5	5	4	4	1	1	1	1		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	1	1	0	0		
				0	YALI0	5	5	4	4	1	1	1	1		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	1	1	0	0		
Lysine degradation	sce00310	1.0	1.1	0	CAGL0	1	1	1	1	0	0	0	0	0|16	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	2	2	0	0	1	1		
				3	CAGL3	1	1	1	1	0	0	0	0		
				4	CAGL4	1	1	1	1	1	1	1	1		
				5	CAGL5	1	1	1	1	1	1	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	2	2	0	0	1	1		
				3	KLLA5	1	1	1	1	0	0	0	0		
				4	KLLA3	1	1	1	1	1	1	1	1		
				5	KLLA4	1	1	1	1	1	1	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	2	2	0	0	1	1		
				3	DEHA5	1	1	1	1	0	0	0	0		
				4	DEHA3	1	1	1	1	1	1	1	1		
				5	DEHA4	1	1	1	1	1	1	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	2	2	0	0	1	1		
				3	YALI5	1	1	1	1	0	0	0	0		
				4	YALI3	1	1	1	1	1	1	1	1		
				5	YALI4	1	1	1	1	1	1	0	0		
Arginine and proline metabolism	sce00330	1.2	1.4	0	CAGL0	7	7	11	11	3	3	2	2	2|33	0.0%
				1	CAGL1	2	2	2	2	0	0	0	0		
				0	KLLA0	7	7	11	11	3	3	2	2	0|33	0.0%
				1	KLLA1	2	2	2	2	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	7	7	11	11	3	3	2	2		
				1	DEHA1	2	2	2	2	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	7	6	11	11	3	2	2	2	1|33	7.7%
				1	YALI1	2	2	2	2	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
Histidine metabolism	sce00340	1.1	1.2	0	CAGL0	2	2	3	3	0	0	0	0	0|16	0.0%
				1	CAGL1	1	1	2	2	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				3	CAGL3	1	1	1	1	0	0	0	0		
				4	CAGL4	1	1	1	1	0	0	0	0		
				0	KLLA0	2	2	3	3	0	0	0	0		
				1	KLLA1	1	1	2	2	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				3	KLLA3	1	1	1	1	0	0	0	0		
				4	KLLA4	1	1	1	1	0	0	0	0		
				0	DEHA2	2	2	3	2	0	0	0	0	1|16	7.1%
				1	DEHA1	1	1	2	2	0	0	0	0		
				2	DEHA3	1	1	1	1	0	0	0	0		
				3	DEHA4	1	1	1	1	0	0	0	0		
				4	DEHA5	1	1	1	1	0	0	0	0		
				0	YALI0	2	2	3	3	0	0	0	0	0|16	0.0%
				1	YALI1	1	1	2	2	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
				3	YALI3	1	1	1	1	0	0	0	0		
				4	YALI4	1	1	1	1	0	0	0	0		
Tyrosine metabolism	sce00350	1.1	1.0	0	CAGL1	1	1	1	1	0	0	0	0	2|11	0.0%
				1	CAGL0	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				4	CAGL3	2	2	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0	0|11	0.0%
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				3	KLLA3	1	1	1	1	0	0	0	0		
				4	KLLA4	2	2	1	1	0	0	0	0		
				0	DEHA1	1	1	1	1	0	0	0	0	2|11	0.0%
				1	DEHA0	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				4	DEHA3	2	2	1	1	0	0	0	0		
				0	YALI1	1	1	1	1	0	0	0	0	0|11	0.0%
				1	YALI0	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
				3	YALI3	1	1	1	1	0	0	0	0		
				4	YALI4	2	2	1	1	0	0	0	0		

Phenylalanine metabolism	sce00360	1.0	1.0	0	CAGL0	1	1	1	1	0	0	0	0	2|6	0.0%
				1	CAGL1	1	1	1	1	0	0	1	1		
				0	KLLA0	1	1	1	1	0	0	0	0	0|6	0.0%
				1	KLLA1	1	1	1	1	0	0	1	1		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	1	1		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	1	1		
				2	YALI2	1	1	1	1	0	0	0	0		
gamma-Hexachlorocyclohexane degradation	sce00361	1.1	1.1	0	CAGL0	2	2	2	2	0	0	0	0	0|16	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				3	CAGL3	1	1	1	1	0	0	0	0		
				4	CAGL4	1	1	1	1	0	0	0	0		
				0	KLLA0	2	2	2	2	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				3	KLLA3	1	1	1	1	0	0	0	0		
				4	KLLA4	1	1	1	1	0	0	0	0		
				0	DEHA0	2	2	2	2	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				3	DEHA3	1	1	1	1	0	0	0	0		
				4	DEHA4	1	1	1	1	0	0	0	0		
				0	YALI0	2	2	2	2	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
				3	YALI3	1	1	1	1	0	0	0	0		
				4	YALI4	1	1	1	1	0	0	0	0		
Benzoate degradation via hydroxylation	sce00362	1.2	1.4	0	CAGL0	1	1	2	2	0	0	1	1	0|7	0.0%
				1	CAGL1	2	2	2	2	0	0	0	0		
				0	KLLA0	1	1	2	2	0	0	1	1		
				1	KLLA1	2	2	2	2	0	0	0	0		
				0	DEHA0	1	1	2	2	0	0	1	1		
				1	DEHA1	2	2	2	2	0	0	0	0		
				0	YALI0	1	1	2	2	0	0	1	1		
				1	YALI1	2	2	2	2	0	0	0	0		
Tryptophan metabolism	sce00380	1.3	1.5	0	CAGL0	7	1	12	1	1	0	0	0	11|31	71.6%
				0	CAGL2	7	2	12	4	1	0	0	0		
				0	CAGL9	7	1	12	1	1	1	0	0		
				1	CAGL6	4	3	2	2	0	0	0	0		
				2	CAGL12	1	1	1	1	1	1	1	1		
				3	CAGL13	1	1	1	1	1	1	0	0		
				0	KLLA0	7	1	12	1	1	0	0	0	10|31	70.1%
				0	KLLA2	7	2	12	4	1	0	0	0		
				0	KLLA9	7	1	12	1	1	1	0	0		
				1	KLLA6	4	4	2	2	0	0	0	0		
				2	KLLA12	1	1	1	1	1	1	1	1		
				3	KLLA13	1	1	1	1	1	1	0	0		
				0	DEHA0	7	7	12	12	1	1	0	0	0|31	0.0%
				1	DEHA1	4	4	2	2	0	0	0	0		
				2	DEHA2	1	1	1	1	1	1	1	1		
				3	DEHA3	1	1	1	1	1	1	0	0		
				0	YALI0	7	7	12	12	1	1	0	0		
				1	YALI1	4	4	2	2	0	0	0	0		
				2	YALI2	1	1	1	1	1	1	1	1		
				3	YALI3	1	1	1	1	1	1	0	0		
Phenylalanine, tyrosine and tryptophan biosynthesis	sce00400	1.3	1.3	0	CAGL0	5	5	4	4	2	2	0	0	0|24	0.0%
				1	CAGL1	1	1	2	2	0	0	1	1		
				0	KLLA0	5	5	4	4	2	2	0	0		
				1	KLLA1	1	1	2	2	0	0	1	1		
				0	DEHA0	5	5	4	4	2	2	0	0		
				1	DEHA1	1	1	2	2	0	0	1	1		
				0	YALI0	5	5	4	4	2	2	0	0		
				1	YALI1	1	1	2	2	0	0	1	1		
Novobiocin biosynthesis	sce00401	1.0	1.0	0	CAGL0	6	6	4	4	0	0	0	0	0|33	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				0	KLLA0	6	6	4	4	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	6	6	4	4	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI0	6	6	4	4	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
beta-Alanine metabolism	sce00410	1.1	1.1	0	CAGL0	2	2	2	2	1	1	0	0	0|7	0.0%
				1	CAGL1	1	1	1	1	1	1	0	0		
				0	KLLA0	2	2	2	2	1	1	0	0		
				1	KLLA1	1	1	1	1	1	1	0	0		
				0	DEHA2	2	1	2	2	1	0	0	0	1|7	16.7%
				1	DEHA1	1	1	1	1	1	1	0	0		
				0	YALI0	2	2	2	2	1	1	0	0	0|7	0.0%
				1	YALI1	1	1	1	1	1	1	0	0		
Taurine and hypotaurine metabolism	sce00430	1.2	1.2	0	CAGL0	2	2	2	2	0	0	0	0	0|5	0.0%
				0	KLLA0	2	2	2	2	0	0	0	0		
				0	DEHA1	2	0	2	1	0	0	0	0	3|5	75.0%
				0	YALI0	2	2	2	2	0	0	0	0	0|5	0.0%

Aminophosphonate metabolism	sce00440	1.0	1.0	0	CAGL0	1	1	1	1	0	0	0	0	0|8	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
Selenoamino acid metabolism	sce00450	1.1	1.1	0	CAGL0	3	3	3	3	0	0	1	1	0|16	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	3	3	3	3	0	0	1	1		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	3	3	3	3	0	0	1	1		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	3	3	3	3	0	0	1	1		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
Cyanoamino acid metabolism	sce00460	1.3	1.3	0	CAGL0	1	1	1	1	0	0	1	1	1|18	5.6%
				1	CAGL1	1	1	2	2	0	0	0	0		
				2	CAGL3	4	3	3	3	0	0	0	0		
				3	CAGL4	1	1	1	1	1	1	1	1		
				4	CAGL5	2	2	2	2	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	1	1	0|18	0.0%
				1	KLLA1	1	1	2	2	0	0	0	0		
				2	KLLA2	4	4	3	3	0	0	0	0		
				3	KLLA3	1	1	1	1	1	1	1	1		
				4	KLLA4	2	2	2	2	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	1	1		
				1	DEHA1	1	1	2	2	0	0	0	0		
				2	DEHA2	4	4	3	3	0	0	0	0		
				3	DEHA3	1	1	1	1	1	1	1	1		
				4	DEHA4	2	2	2	2	0	0	0	0		
				1	YALI1	1	1	2	2	0	0	0	0	2|18	0.0%
				2	YALI2	4	4	3	3	0	0	0	0		
				3	YALI4	1	1	1	1	1	1	1	1		
				4	YALI5	2	2	2	2	0	0	0	0		
Glutathione metabolism	sce00480	1.4	1.4	0	CAGL0	4	4	3	3	2	2	0	0	0|13	0.0%
				0	KLLA0	4	4	3	3	2	2	0	0		
				0	DEHA0	4	4	3	3	2	2	0	0		
				0	YALI0	4	4	3	3	2	2	0	0		
Starch and sucrose metabolism	sce00500	1.4	1.3	0	CAGL0	4	2	3	3	0	0	0	0	4|20	33.3%
				0	KLLA0	4	4	3	3	0	0	0	0	2|20	0.0%
				0	DEHA0	4	4	3	3	0	0	0	0	2|20	0.0%
				0	YALI0	4	4	3	3	0	0	0	0	2|20	0.0%
N-Glycan biosynthesis	sce00510	1.4	1.3	0	CAGL0	1	1	1	1	0	0	0	0	0|28	0.0%
				1	CAGL1	6	6	2	2	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	6	6	2	2	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	6	6	2	2	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	6	6	2	2	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
Nucleotide sugars metabolism	sce00520	1.4	1.6	0	CAGL0	2	1	3	2	0	0	0	0	2|8	40.0%
				0	KLLA0	2	2	3	3	0	0	0	0	0|8	0.0%
				0	DEHA0	2	2	3	3	0	0	0	0		
				0	YALI0	2	2	3	3	0	0	0	0		
Streptomycin biosynthesis	sce00521	1.2	1.0	0	CAGL1	1	1	1	1	1	1	1	1	0|6	0.0%
				1	CAGL0	2	2	2	2	0	0	1	1		
				0	KLLA1	1	1	1	1	1	1	1	1		
				1	KLLA0	2	2	2	2	0	0	1	1		
				0	DEHA1	1	1	1	1	1	1	1	1		
				1	DEHA0	2	2	2	2	0	0	1	1		
				0	YALI1	1	1	1	1	1	1	1	1		
				1	YALI0	2	2	2	2	0	0	1	1		
Aminosugars metabolism	sce00530	1.1	1.2	0	CAGL0	3	3	4	4	0	0	0	0	0|12	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				0	KLLA0	3	3	4	4	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	3	3	4	4	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI0	3	3	4	4	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
Peptidoglycan biosynthesis	sce00550	1.5	1.5	0	CAGL0	3	3	3	3	0	0	0	0	0|6	0.0%
				0	KLLA0	3	3	3	3	0	0	0	0		
				0	DEHA0	3	3	3	3	0	0	0	0		
				0	YALI0	3	3	3	3	0	0	0	0		

Glycerolipid metabolism	sce00561	1.0	1.1	1	CAGL1	1	1	1	1	0	0	0	0	2|18	7.7%
				2	CAGL2	1	1	1	1	1	1	0	0		
				3	CAGL3	1	1	1	1	0	0	0	0		
				4	CAGL4	1	0	2	2	0	0	1	1		
				5	CAGL5	1	1	1	1	1	1	0	0		
				6	CAGL6	1	1	1	1	0	0	1	1		
				1	KLLA1	1	1	1	1	0	0	0	0	2|18	7.7%
				2	KLLA2	1	1	1	1	1	1	0	0		
				3	KLLA3	1	1	1	1	0	0	0	0		
				4	KLLA4	1	0	2	2	0	0	1	1		
				5	KLLA5	1	1	1	1	1	1	0	0		
				6	KLLA6	1	1	1	1	0	0	1	1		
				1	DEHA1	1	1	1	1	0	0	0	0	2|18	7.7%
				2	DEHA2	1	1	1	1	1	1	0	0		
				3	DEHA3	1	1	1	1	0	0	0	0		
				4	DEHA4	1	0	2	2	0	0	1	1		
				5	DEHA5	1	1	1	1	1	1	0	0		
				6	DEHA6	1	1	1	1	0	0	1	1		
				1	YALI0	1	1	1	1	0	0	0	0	2|18	7.7%
				2	YALI1	1	1	1	1	1	1	0	0		
				3	YALI2	1	1	1	1	0	0	0	0		
				4	YALI4	1	0	2	2	0	0	1	1		
				5	YALI5	1	1	1	1	1	1	0	0		
				6	YALI6	1	1	1	1	0	0	1	1		
Inositol phosphate metabolism	sce00562	1.1	1.3	0	CAGL0	1	1	1	1	1	1	1	1	0|14	0.0%
				1	CAGL1	2	2	5	5	1	1	2	2		
				0	KLLA0	1	1	1	1	1	1	1	1		
				1	KLLA1	2	2	5	5	1	1	2	2		
				0	DEHA0	1	1	1	1	1	1	1	1		
				1	DEHA1	2	2	5	5	1	1	2	2		
				0	YALI0	1	1	1	1	1	1	1	1		
				1	YALI1	2	2	5	5	1	1	2	2		
Glycerophospholipid metabolism	sce00564	1.2	1.2	0	CAGL0	5	5	5	5	1	1	1	1	0|23	0.0%
				0	KLLA0	5	5	5	5	1	1	1	1		
				0	DEHA0	5	5	5	5	1	1	1	1		
				0	YALI0	5	5	5	5	1	1	1	1		
Ether lipid metabolism	sce00565	1.2	1.4	0	CAGL0	3	3	5	5	0	0	0	0	0|11	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				0	KLLA0	3	3	5	5	0	0	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	3	3	5	5	0	0	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI0	3	3	5	5	0	0	0	0		
				1	YALI1	1	1	1	1	0	0	0	0		
Arachidonic acid metabolism	sce00590	1.1	1.1	0	CAGL0	1	1	1	1	0	0	0	0	0|8	0.0%
				1	CAGL1	2	2	2	2	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	2	2	2	2	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	2	2	2	2	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	2	2	2	2	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
Sphingolipid metabolism	sce00600	1.4	1.3	0	CAGL0	5	5	4	4	1	1	1	1	2|17	0.0%
				0	KLLA0	5	5	4	4	1	1	1	1		0.0%
				0	DEHA0	5	5	4	4	1	1	1	1		0.0%
				0	YALI0	5	5	4	4	1	1	1	1		0.0%
Pyruvate metabolism	sce00620	1.1	1.3	0	CAGL0	2	2	6	6	0	0	3	3	0|17	0.0%
				0	KLLA0	2	2	6	6	0	0	3	3		
				0	DEHA0	2	2	6	6	0	0	3	3		
				0	YALI0	2	2	6	6	0	0	3	3		
1- and 2-Methylnaphthalene degradation	sce00624	1.3	1.5	0	CAGL0	1	1	2	2	0	0	0	0	4|16	0.0%
				1	CAGL1	4	4	5	5	0	0	1	1		
				0	KLLA0	1	1	2	2	0	0	0	0	0|16	0.0%
				1	KLLA1	4	4	5	5	0	0	1	1		
				2	KLLA2	2	2	2	2	0	0	0	0		
				0	DEHA0	1	1	2	2	0	0	0	0	4|16	0.0%
				1	DEHA1	4	4	5	5	0	0	1	1		
				0	YALI0	1	1	2	2	0	0	0	0	0|16	0.0%
				1	YALI1	4	4	5	5	0	0	1	1		
				2	YALI2	2	2	2	2	0	0	0	0		
Tetrachloroethene degradation	sce00625	1.0	1.0	0	CAGL0	2	2	2	2	0	0	1	1	0|8	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	2	2	2	2	0	0	1	1		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	2	2	2	2	0	0	1	1		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	2	2	2	2	0	0	1	1		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		

1,4-Dichlorobenzene degradation	sce00627	1.5	1.5	0	CAGL0	1	1	1	1	0	0	0	0	0|24	0.0%
				1	CAGL1	4	4	3	3	0	0	0	0		
				2	CAGL2	3	3	4	4	0	0	0	0		
				3	CAGL3	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	4	4	3	3	0	0	0	0		
				2	KLLA2	3	3	4	4	0	0	0	0		
				3	KLLA3	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	4	4	3	3	0	0	0	0		
				2	DEHA2	3	3	4	4	0	0	0	0		
				3	DEHA3	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	4	4	3	3	0	0	0	0		
				2	YALI2	3	3	4	4	0	0	0	0		
				3	YALI3	1	1	1	1	0	0	0	0		
Glyoxylate and dicarboxylate metabolism	sce00630	1.2	1.5	0	CAGL0	2	2	4	4	1	1	3	3	0|13	0.0%
				1	CAGL1	1	1	2	2	0	0	1	1		
				2	CAGL2	1	1	1	1	0	0	1	1		
				0	KLLA0	2	2	4	4	1	1	3	3		
				1	KLLA1	1	1	2	2	0	0	1	1		
				2	KLLA2	1	1	1	1	0	0	1	1		
				0	DEHA0	2	2	4	4	1	1	3	3		
				1	DEHA1	1	1	2	2	0	0	1	1		
				2	DEHA2	1	1	1	1	0	0	1	1		
				0	YALI0	2	2	4	4	1	1	3	3		
				1	YALI1	1	1	2	2	0	0	1	1		
				2	YALI2	1	1	1	1	0	0	1	1		
Benzoate degradation via CoA ligation	sce00632	1.2	1.1	0	CAGL0	2	1	1	1	0	0	0	0	1|13	7.7%
				1	CAGL2	1	1	1	1	0	0	0	0		
				2	CAGL1	1	1	1	1	0	0	0	0		
				3	CAGL3	2	2	2	2	0	0	0	0		
				4	CAGL4	1	1	1	1	1	1	1	1		
				0	KLLA0	2	2	1	1	0	0	0	0	0|13	0.0%
				1	KLLA2	1	1	1	1	0	0	0	0		
				2	KLLA1	1	1	1	1	0	0	0	0		
				3	KLLA3	2	2	2	2	0	0	0	0		
				4	KLLA4	1	1	1	1	1	1	1	1		
				0	DEHA0	2	2	1	1	0	0	0	0	0|13	0.0%
				1	DEHA2	1	1	1	1	0	0	0	0		
				2	DEHA1	1	1	1	1	0	0	0	0		
				3	DEHA3	2	2	2	2	0	0	0	0		
				4	DEHA4	1	1	1	1	1	1	1	1		
				0	YALI0	2	2	1	1	0	0	0	0	0|13	0.0%
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
				3	YALI3	2	2	2	2	0	0	0	0		
				4	YALI4	1	1	1	1	1	1	1	1		
Propanoate metabolism	sce00640	1.1	1.1	0	CAGL0	1	1	1	1	1	1	0	0	0|14	0.0%
				1	CAGL1	1	1	1	1	1	1	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				3	CAGL3	1	1	1	1	0	0	0	0		
				4	CAGL4	1	1	2	2	1	1	1	1		
				5	CAGL5	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	1	1	0	0	0|14	0.0%
				1	KLLA1	1	1	1	1	1	1	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				3	KLLA3	1	1	1	1	0	0	0	0		
				4	KLLA4	1	1	2	2	1	1	1	1		
				5	KLLA5	1	1	1	1	0	0	0	0		
				0	DEHA1	1	1	1	1	1	1	0	0	2|14	0.0%
				1	DEHA0	1	1	1	1	1	1	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				3	DEHA3	1	1	1	1	0	0	0	0		
				4	DEHA5	1	1	2	2	1	1	1	1		
				0	YALI0	1	1	1	1	1	1	0	0	0|14	0.0%
				1	YALI1	1	1	1	1	1	1	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
				3	YALI3	1	1	1	1	0	0	0	0		
				4	YALI4	1	1	2	2	1	1	1	1		
				5	YALI5	1	1	1	1	0	0	0	0		

Butanoate metabolism	sce00650	1.1	1.1	0	CAGL0	1	1	1	1	1	1	0	0	0|16	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	1	1	0	0		
				3	CAGL3	1	1	1	1	1	1	1	1		
				4	CAGL4	1	1	1	1	0	0	0	0		
				5	CAGL5	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	1	1	0	0	0|16	0.0%
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	1	1	0	0		
				3	KLLA3	1	1	1	1	1	1	1	1		
				4	KLLA4	1	1	1	1	0	0	0	0		
				5	KLLA5	1	1	1	1	0	0	0	0		
				0	DEHA0	1	0	1	1	1	0	0	0	1|16	8.3%
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	1	1	0	0		
				3	DEHA3	1	1	1	1	1	1	1	1		
				4	DEHA4	1	1	1	1	0	0	0	0		
				5	DEHA6	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	1	1	0	0	0|16	0.0%
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	1	1	0	0		
				3	YALI3	1	1	1	1	1	1	1	1		
				4	YALI4	1	1	1	1	0	0	0	0		
				5	YALI5	1	1	1	1	0	0	0	0		
One carbon pool by folate	sce00670	1.5	1.5	0	CAGL0	1	1	1	1	0	0	0	0	0|7	0.0%
				0	KLLA0	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
Methane metabolism	sce00680	1.1	1.2	0	CAGL0	1	1	2	2	0	0	1	1	2|11	0.0%
				2	CAGL2	2	2	2	2	2	2	1	1		
				3	CAGL4	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	2	2	0	0	1	1	2|11	0.0%
				2	KLLA2	2	2	2	2	2	2	1	1		
				3	KLLA4	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	2	2	0	0	1	1	2|11	0.0%
				2	DEHA2	2	2	2	2	2	2	1	1		
				3	DEHA4	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	2	2	0	0	1	1	2|11	0.0%
				2	YALI2	2	2	2	2	2	2	1	1		
				3	YALI4	1	1	1	1	0	0	0	0		
Carbon fixation	sce00710	1.2	1.3	0	CAGL0	1	1	1	1	0	0	1	1	0|20	0.0%
				1	CAGL1	3	3	3	3	0	0	1	1		
				2	CAGL2	1	1	3	3	1	1	3	3		
				3	CAGL3	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	1	1		
				1	KLLA1	3	3	3	3	0	0	1	1		
				2	KLLA2	1	1	3	3	1	1	3	3		
				3	KLLA3	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	1	1		
				1	DEHA1	3	3	3	3	0	0	1	1		
				2	DEHA2	1	1	3	3	1	1	3	3		
				3	DEHA3	1	1	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	1	1		
				1	YALI1	3	3	3	3	0	0	1	1		
				2	YALI2	1	1	3	3	1	1	3	3		
				3	YALI3	1	1	1	1	0	0	0	0		
Reductive carboxylate cycle (CO2 fixation)	sce00720	1.2	1.1	0	CAGL0	2	2	1	1	1	1	0	0	0|10	0.0%
				1	CAGL1	1	1	1	1	1	1	1	1		
				2	CAGL2	1	1	1	1	1	1	0	0		
				0	KLLA0	2	2	1	1	1	1	0	0		
				1	KLLA1	1	1	1	1	1	1	1	1		
				2	KLLA2	1	1	1	1	1	1	0	0		
				0	DEHA0	2	2	1	1	1	1	0	0		
				1	DEHA1	1	1	1	1	1	1	1	1		
				2	DEHA2	1	1	1	1	1	1	0	0		
				0	YALI0	2	2	1	1	1	1	0	0		
				1	YALI1	1	1	1	1	1	1	1	1		
				2	YALI2	1	1	1	1	1	1	0	0		
Thiamine metabolism	sce00730	1.5	1.3	0	CAGL0	1	1	1	1	0	0	0	0	0|7	0.0%
				1	CAGL1	2	2	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	2	2	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	2	2	1	1	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	2	2	1	1	0	0	0	0		
Riboflavin metabolism	sce00740	1.1	1.1	0	CAGL1	1	1	1	1	1	1	0	0	0|11	0.0%
				1	CAGL0	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	1	1	1	1	1	1	0	0		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	1	1	1	1	1	1	0	0		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI1	1	1	1	1	1	1	0	0		
				1	YALI0	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		

Vitamin B6 metabolism	sce00750	1.5	1.5	0	CAGL0	1	1	2	2	0	0	0	0	0|9	0.0%
				1	CAGL1	2	2	1	1	0	0	0	0		
				0	KLLA0	1	1	2	2	0	0	0	0		
				1	KLLA1	2	2	1	1	0	0	0	0		
				0	DEHA0	1	1	2	2	0	0	0	0		
				1	DEHA1	2	2	1	1	0	0	0	0		
				0	YALI0	1	1	2	2	0	0	0	0		
				1	YALI1	2	2	1	1	0	0	0	0		
Nicotinate and nicotinamide metabolism	sce00760	1.4	1.2	0	CAGL0	4	3	2	2	0	0	0	0	1|11	12.5%
				1	CAGL1	1	1	1	1	0	0	0	0		
				0	KLLA0	4	3	2	2	0	0	0	0	1|11	12.5%
				1	KLLA1	1	1	1	1	0	0	0	0		
				0	DEHA0	4	3	2	2	0	0	0	0	1|11	12.5%
				1	DEHA1	1	1	1	1	0	0	0	0		
				0	YALI0	4	4	2	2	0	0	0	0	0|11	0.0%
				1	YALI1	1	1	1	1	0	0	0	0		
Pantothenate and CoA biosynthesis	sce00770	1.2	1.1	0	CAGL0	5	5	2	2	1	1	0	0	0|10	0.0%
				0	KLLA0	5	5	2	2	1	1	0	0		
				0	DEHA0	5	5	2	2	1	1	0	0		
				0	YALI0	5	5	2	2	1	1	0	0		
Biotin metabolism	sce00780	1.1	1.1	0	CAGL0	1	0	1	1	0	0	0	0	1|7	50.0%
				0	CAGL2	1	1	1	0	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0	0|7	0.0%
				0	DEHA0	1	1	1	1	0	0	0	0	0|7	
				0	YALI0	1	0	1	1	0	0	0	0	0|7	50.0%
				0	YALI1	1	1	1	0	0	0	0	0		
Folate biosynthesis	sce00790	1.5	1.5	0	CAGL0	1	1	1	1	0	0	0	0	0|18	0.0%
				1	CAGL1	3	3	3	3	1	1	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0	0|18	0.0%
				1	KLLA1	3	3	3	3	1	1	0	0		
				1	DEHA1	3	3	3	3	1	1	0	0	2|18	33.0%
				0	YALI0	1	1	1	1	0	0	0	0	0|18	0.0%
				1	YALI1	3	3	3	3	1	1	0	0		
Porphyrin and chlorophyll metabolism	sce00860	1.1	1.1	0	CAGL0	3	3	3	3	1	1	0	0	0|17	0.0%
				1	CAGL1	1	1	1	1	1	1	0	0		
				0	KLLA0	3	3	3	3	1	1	0	0		
				1	KLLA1	1	1	1	1	1	1	0	0		
				0	DEHA0	3	3	3	3	1	1	0	0		
				1	DEHA1	1	1	1	1	1	1	0	0		
				0	YALI0	3	3	3	3	1	1	0	0		
				1	YALI1	1	1	1	1	1	1	0	0		
Terpenoid biosynthesis	sce00900	1.3	1.4	0	CAGL0	1	1	2	2	0	0	0	0	0|7	0.0%
				0	KLLA0	1	1	2	2	0	0	0	0		
				0	DEHA0	1	1	2	2	0	0	0	0		
				0	YALI0	1	1	2	2	0	0	0	0		
Limonene and pinene degradation	sce00903	1.1	1.1	0	CAGL0	1	1	1	1	0	0	0	0	0|18	0.0%
				1	CAGL1	2	2	2	2	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				3	CAGL3	1	1	1	1	0	0	0	0		
				4	CAGL4	4	4	4	4	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	2	2	2	2	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				3	KLLA3	1	1	1	1	0	0	0	0		
				4	KLLA4	4	4	4	4	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	2	2	2	2	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				3	DEHA3	1	1	1	1	0	0	0	0		
				4	DEHA4	4	4	4	4	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	2	2	2	2	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		
				3	YALI3	1	1	1	1	0	0	0	0		
				4	YALI4	4	4	4	4	0	0	0	0		
Nitrogen metabolism	sce00910	1.4	1.2	0	CAGL0	4	4	2	2	0	0	2	2	0|9	0.0%
				0	KLLA0	4	4	2	2	0	0	2	2	0|9	
				0	DEHA0	4	4	2	2	0	0	2	2	0|9	
				0	YALI1	4	3	2	1	0	0	2	1	2|9	50.0%
Sulfur metabolism	sce00920	1.2	1.2	0	CAGL0	2	2	2	2	0	0	2	2	0|13	0.0%
				1	CAGL1	1	1	1	1	0	0	0	0		
				2	CAGL2	1	1	1	1	0	0	0	0		
				0	KLLA0	2	2	2	2	0	0	2	2		
				1	KLLA1	1	1	1	1	0	0	0	0		
				2	KLLA2	1	1	1	1	0	0	0	0		
				0	DEHA0	2	2	2	2	0	0	2	2		
				1	DEHA1	1	1	1	1	0	0	0	0		
				2	DEHA2	1	1	1	1	0	0	0	0		
				0	YALI0	2	2	2	2	0	0	2	2		
				1	YALI1	1	1	1	1	0	0	0	0		
				2	YALI2	1	1	1	1	0	0	0	0		

Alkaloid biosynthesis II	sce00960	1.2	1.2	0	CAGL0	1	1	1	1	0	0	0	0	0|6	0.0%
				1	CAGL1	2	2	2	2	0	0	0	0		
				0	KLLA0	1	1	1	1	0	0	0	0		
				1	KLLA1	2	2	2	2	0	0	0	0		
				0	DEHA0	1	1	1	1	0	0	0	0		
				1	DEHA1	2	2	2	2	0	0	0	0		
				0	YALI0	1	1	1	1	0	0	0	0		
				1	YALI1	2	2	2	2	0	0	0	0		
Aminoacyl-tRNA biosynthesis	sce00970	1.6	1.0	0	CAGL0	2	2	1	1	1	1	0	0	0|63	0.0%
				1	CAGL1	2	2	1	1	1	1	0	0		
				10	CAGL10	3	3	2	2	0	0	1	1		
				11	CAGL11	2	2	1	1	0	0	0	0		
				12	CAGL13	2	2	1	1	1	1	0	0		
				13	CAGL12	2	2	1	1	0	0	0	0		
				14	CAGL14	2	2	1	1	0	0	0	0		
				15	CAGL15	2	2	1	1	0	0	0	0		
				16	CAGL16	2	2	1	1	1	1	0	0		
				17	CAGL18	2	2	1	1	0	0	0	0		
				18	CAGL17	2	2	1	1	0	0	0	0		
				19	CAGL19	2	2	1	1	0	0	0	0		
				2	CAGL2	2	2	1	1	0	0	0	0		
				3	CAGL4	2	2	1	1	1	1	0	0		
				4	CAGL3	2	2	1	1	0	0	0	0		
				5	CAGL5	2	2	1	1	0	0	0	0		
				6	CAGL6	2	2	1	1	0	0	0	0		
				7	CAGL8	2	2	1	1	1	1	0	0		
				8	CAGL7	2	2	1	1	0	0	0	0		
				9	CAGL9	2	2	1	1	1	1	0	0		
				0	KLLA0	2	2	1	1	1	1	0	0		
				1	KLLA1	2	2	1	1	1	1	0	0		
				10	KLLA10	3	3	2	2	0	0	1	1		
				11	KLLA11	2	2	1	1	0	0	0	0		
				12	KLLA13	2	2	1	1	1	1	0	0		
				13	KLLA12	2	2	1	1	0	0	0	0		
				14	KLLA14	2	2	1	1	0	0	0	0		
				15	KLLA15	2	2	1	1	0	0	0	0		
				16	KLLA16	2	2	1	1	1	1	0	0		
				17	KLLA18	2	2	1	1	0	0	0	0		
				18	KLLA17	2	2	1	1	0	0	0	0		
				19	KLLA19	2	2	1	1	0	0	0	0		
				2	KLLA2	2	2	1	1	0	0	0	0		
				3	KLLA4	2	2	1	1	1	1	0	0		
				4	KLLA3	2	2	1	1	0	0	0	0		
				5	KLLA5	2	2	1	1	0	0	0	0		
				6	KLLA6	2	2	1	1	0	0	0	0		
				7	KLLA8	2	2	1	1	1	1	0	0		
				8	KLLA7	2	2	1	1	0	0	0	0		
				9	KLLA9	2	2	1	1	1	1	0	0		
				0	DEHA0	2	2	1	1	1	1	0	0		
				1	DEHA1	2	2	1	1	1	1	0	0		
				10	DEHA10	3	3	2	2	0	0	1	1		
				11	DEHA11	2	2	1	1	0	0	0	0		
				12	DEHA13	2	2	1	1	1	1	0	0		
				13	DEHA12	2	2	1	1	0	0	0	0		
				14	DEHA14	2	2	1	1	0	0	0	0		
				15	DEHA15	2	2	1	1	0	0	0	0		
				16	DEHA16	2	2	1	1	1	1	0	0		
				17	DEHA18	2	2	1	1	0	0	0	0		
				18	DEHA17	2	2	1	1	0	0	0	0		
				19	DEHA19	2	2	1	1	0	0	0	0		
				2	DEHA2	2	2	1	1	0	0	0	0		
				3	DEHA4	2	2	1	1	1	1	0	0		
				4	DEHA3	2	2	1	1	0	0	0	0		
				5	DEHA5	2	2	1	1	0	0	0	0		
				6	DEHA6	2	2	1	1	0	0	0	0		
				7	DEHA8	2	2	1	1	1	1	0	0		
				8	DEHA7	2	2	1	1	0	0	0	0		
				9	DEHA9	2	2	1	1	1	1	0	0		
				0	YALI0	2	2	1	1	1	1	0	0		
				1	YALI1	2	2	1	1	1	1	0	0		
				10	YALI10	3	3	2	2	0	0	1	1		
				11	YALI11	2	2	1	1	0	0	0	0		
				12	YALI13	2	2	1	1	1	1	0	0		
				13	YALI12	2	2	1	1	0	0	0	0		
				14	YALI14	2	2	1	1	0	0	0	0		
				15	YALI15	2	2	1	1	0	0	0	0		
				16	YALI16	2	2	1	1	1	1	0	0		
				17	YALI18	2	2	1	1	0	0	0	0		
				18	YALI17	2	2	1	1	0	0	0	0		
				19	YALI19	2	2	1	1	0	0	0	0		
				2	YALI2	2	2	1	1	0	0	0	0		
				3	YALI4	2	2	1	1	1	1	0	0		
				4	YALI3	2	2	1	1	0	0	0	0		
				5	YALI5	2	2	1	1	0	0	0	0		
				6	YALI6	2	2	1	1	0	0	0	0		
				7	YALI8	2	2	1	1	1	1	0	0		
				8	YALI7	2	2	1	1	0	0	0	0		
				9	YALI9	2	2	1	1	1	1	0	0		

Metabolism of xenobiotics by cytochrome P450	sce00980	1.0	1.3	0	CAGL0	1	1	2	1	0	0	0	0	1|3	33.0%
				0	KLLA0	1	1	2	2	0	0	0	0	0|3	0.0%
				0	DEHA0	1	1	2	1	0	0	0	0	1|3	33.0%
				0	YALI0	1	1	2	2	0	0	0	0	0|3	0.0%
Phosphatidylinositol signaling system	sce04070	1.1	1.2	0	CAGL0	2	2	4	4	1	1	1	1	0|14	0.0%
				1	CAGL1	1	1	1	1	1	1	0	0		
				0	KLLA0	2	2	4	4	1	1	1	1		
				1	KLLA1	1	1	1	1	1	1	0	0		
				0	DEHA0	2	2	4	4	1	1	1	1		
				1	DEHA1	1	1	1	1	1	1	0	0		
				0	YALI0	2	2	4	4	1	1	1	1		
				1	YALI1	1	1	1	1	1	1	0	0		