Nouvelles approches bioinformatiques pour l'étude à grande échelle de l'évolution des activités enzymatiques

Cécile Pereira

Université Paris-Sud

École Doctorale Gènes Génomes Cellules

Laboratoire : Institut de biologie intégrative de la cellule (I2BC)

Discipline : Sciences de la vie et de la santé

Thèse de doctorat

Soutenue le 11 mai 2015 par

Cécile Pereira

Nouvelles approches bioinformatiques pour l'étude à grande échelle de l'évolution des activités enzymatiques.

Directeur de thèse :	M. Olivier Lespinet	PR (Université Paris-Sud)
Co-encadrant de thèse :	M. Alain Denise	PR (Université Paris-Sud)

Composition du jury :

Rapporteurs :	M. Christophe Dessimoz	LECTURER (University College London)
	M. Marc Henri Lebrun	DR (INRA)
Examinateurs :	M. Jérôme Azé	PR (LIRMM)
	M. Armel Guyonvarch	PR (Université Paris-Sud)
	Mme Claudine Médigue	DR (CNRS)
	M. Philippe Silar	PR (Université Paris 7)

Remerciements

L'ensemble de mes travaux a été enrichi des discutions, commentaires, et retours que mes collaborateurs, relecteurs, ou collègues ont bien voulu me donner.

Je tenais donc tout d'abord à remercier tout particulièrement mes rapporteurs, Christophe Dessimoz et Marc-Henri Lebrun, ainsi que mes examinateurs, Jérôme Azé, Armel Guyonvarch, Claudine Médigue et Philippe Silar d'avoir accepter d'évaluer mon travail.

Ma thèse terminée, je regarde ces dernières années et je réalise que bien qu'un travail officiellement solitaire, cette thèse n'aurait pu se faire sans l'aide de toute une équipe que j'aimerais donc remercier ici. Je tiens à remercier Alain Denise et Olivier Lespinet pour la qualité de leur encadrement, leur disponibilité et leurs encouragements. Je remercie Anne Lopes qui m'aura appris que les coups durs sont de 'très bon exercices'. Merci à Christine Drevet pour son soutien les jours de gros temps et à Mathieu Barba et Samer Abboud pour leurs coups de main. Je n'oublie pas les 'nouveaux', Marie Hélène, Jean Pierre, Bruno, Annie et Jean Christophe sans qui les réunions d'équipe n'auraient pas la même saveur (merci Anne pour les chouquettes et Jean-Christophe pour les macarons). Merci également à mon équipe actuelle, l'équipe de bioinformatique du LRI, qui a su m'intégrer dès mon arrivée et me soutenir dans la dernière ligne droite.

Plus que mes équipes, je souhaite ici souligner la bonne ambiance du premier étage du 400 et du coup remercier l'ensemble des équipes de Daniel Gautheret,

d'Olivier Namy et de Jean-Luc Pernodet. Plus particulièrement, merci Drago pour les N à l'envers dans les grilles de mots croisés, merci Claire T pour la devise du labo 'boire son café debout ça rend fou', merci Rachel pour le soutien quand ma chaudière m'a abandonnée, et surtout merci à tous pour avoir supporté mon côté pipelette sans trop broncher.

D'un point de vue plus scientifique, j'aimerais remercier mon comité de thèse (Philippe Silar et Claudine Medigue). Ils n'avaient pas tords, j'avais à mes débuts des idées pour les dix années à venir, j'espère avoir fait ce qu'ils pouvaient attendre de moi dans ces trois ans et avoir la chance de m'attaquer à la suite dans les années futures.

Je remercie également l'équipe dirigeante, Jean Pierre Rousset et Monique Boulotin, qui se sont toujours montrés encourageants et enthousiastes pour mon travail, même si cela leur demandait d'écrire des lettres de recommandations ou leur impliquait des après midi de discussions.

Gérer un projet de 3 ans, avec ses hauts et ses bas, c'est un travail d'endurance que je n'aurais pu faire sans mes amis et ma famille qui m'ont aidé à décompresser et à tenir la distance. Je soulignerai donc l'importance des soirées du mardi chez Keffrey (merci John Hutinet et Kevin Versatrahers), celle du jeudi (chez moi), les pauses cafés inter-équipes ainsi que les soirées non statutaires. Ainsi dans le désordre, merci : Moano, Selima, Thibaut, Marine, Anne Clem, Anne So, Diane, Dimitri, Amanda, Rémi, Hélène ... (si je vous ai oublié et que vous devriez y être merci d'écrire votre nom ici : _____).

Je remercie également Marc Descrimes, mon compagnon de vie, qui m'a soutenue durant ces trois années et sans qui je n'aurais mangé que des plats préparés.

Merci tout particulièrement à ma famille, qui m'a fait confiance, et cela même si elle ne comprenait pas forcément ce en quoi consiste mon travail. Un clin d'œil à mon frère Sébastien, à qui j'ai du faire passer le virus puisqu'il est venu faire

son stage de troisième dans nos locaux.

Enfin, je remercie tout ceux que qui voudront bien lire ce manuscrit, ça fait chaud au cœur par avance de savoir que ma rédaction servira à d'autres.

Table des matières

I Introduction

Résumé

Cette thèse a pour objectif de proposer de nouvelles méthodes permettant l'étude de l'évolution du métabolisme. Pour cela, nous avons choisi de nous pencher sur le problème de comparaison du métabolisme de centaines de micro-organismes.

Afin de comparer le métabolisme de différentes espèces, il faut dans un premier temps connaître le métabolisme de chacune de ces espèces.

Les protéomes des micro-organismes avec lesquels nous souhaitons travailler proviennent de différentes bases de données et ont été séquencés et annotés par différentes équipes, *via* différentes méthodes. L'annotation fonctionnelle peut donc être de qualité hétérogène. C'est pourquoi il est nécessaire d'effectuer une ré-annotation fonctionnelle standardisée des protéomes des organismes que nous souhaitons comparer.

L'annotation de séquences protéiques peut être réalisée par le transfert d'annotations entre séquences orthologues. Il existe plus de 39 bases de données répertoriant des orthologues prédits par différentes méthodes. Il est connu que ces méthodes mènent à des prédictions en partie différentes. Afin de tenir compte des prédictions actuelles tout en ajoutant de l'information pertinente, nous avons développé la méta-approche MARIO. Celle-ci combine les intersections des résultats de plusieurs méthodes de détections de groupes d'orthologues et les enrichit grâce à l'utilisation de profils HMM. Nous montrons que notre méta–approche permet de prédire un plus grand nombre d'orthologues tout en améliorant la similarité de fonction des paires

d'orthologues prédites. Cela nous a permis de prédire le répertoire enzymatique de 178 protéomes de micro-organismes (dont 174 champignons).

Dans un second temps, nous analysons ces répertoires enzymatiques afin d'en apprendre plus sur l'évolution du métabolisme. Dans ce but, nous cherchons des combinaisons de présence/absence d'activités enzymatiques permettant de caractériser un groupe taxonomique donné. Ainsi, il devient possible de déduire si la création d'un groupe taxonomique particulier peut s'expliquer par (ou a induit) l'apparition de certaines spécificités au niveau de son métabolisme.

Pour cela, nous avons appliqué des méthodes d'apprentissage supervisé interprétables (règles et arbres de décision) sur les profils enzymatiques. Nous utilisons comme attributs les activités enzymatiques, comme classe les groupes taxonomiques et comme exemples les champignons. Les résultats obtenus, cohérents avec nos connaissances actuelles sur ces organismes, montrent que l'application de méthodes d'apprentissage supervisé est efficace pour extraire de l'information des profils phylogénétiques. Le métabolisme conserve donc des traces de l'évolution des espèces. De plus, cette approche, dans le cas de prédiction de classifieurs présentant un faible nombre d'erreurs, peut permettre de mettre en évidence l'existence de probables transferts horizontaux. C'est le cas par exemple du transfert du gène codant pour l'EC:3.1.6.6 d'un ancêtre des pezizomycotina vers un ancêtre d'*Ustilago maydis*.

Chapitre 1

Objectifs

Le métabolisme varie en fonction des espèces. Ainsi, deux espèces peuvent être issues du même milieu et présenter des voies métaboliques différentes. De plus, deux espèces proche taxonomiquement n'ont pas non plus un métabolisme identique.

La problématique de cette thèse est donc la suivante : Comment expliquer les variations métaboliques d'une espèce à l'autre ou d'un phénotype à l'autre?

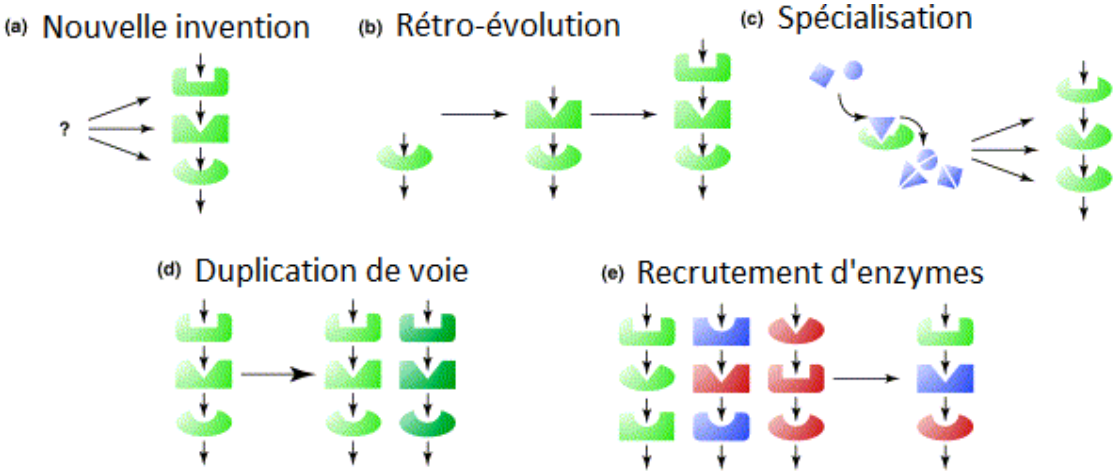


Figure 1.1 – *Mécanismes d'évolution du métabolisme : différentes hypothèses. Figure adaptée de (Schmidt et al., 2003)*

Il existe actuellement plusieurs hypothèses quand aux mécanismes d'évolu-

tion du métabolisme 1.1 (Schmidt et al., 2003, Alves et al., 2002). La première serait que les enzymes évoluent indépendamment avant d'être impliquées dans la même voie. Une seconde hypothèse, celle de rétro-évolution voudrait que les enzymes évoluent de manière à produire le plus possible d'un produit en bout de voie. Ainsi, de manière à augmenter ce rendement, il serait nécessaire d'utiliser plus d'un métabolite distant comme substrat initial de cette voie métabolique. Cette hypothèse est corroborée par des études de génomique comparative. Une autre hypothèse suppose que les enzymes présentaient à l'origine plusieurs fonctions. La duplication de cette enzyme aurait aboutie à la sélection d'enzymes plus spécifiques et efficaces pour chacune des étapes. Il est également possible que l'ensemble d'une voie se duplique aboutissant à deux voies métaboliques catalysées par des enzymes homologues. Enfin, il est possible que les voies métaboliques soient issues de l'utilisation de même enzymes dans différentes voies, aboutissant à un patchwork d'enzymes homologues catalysant des réactions dans des voies métaboliques différentes.

Ces différentes hypothèses ne sont pas incompatible entre elles, plusieurs études scientifiques tendant à valider plutôt l'une ou l'autre en fonction de la voie métabolique analysée (Rison and Thornton, 2002, Pfeiffer et al., 2005, Caetano-Anollés et al., 2009).

Afin de mieux comprendre les étapes d'évolution du métabolisme nous proposons une approche basée sur la comparaison des profils phylogénétiques d'activités enzymatiques. Nous recherchons des activités enzymatiques qui conjointement, par leurs profils de présence / absence, seraient capable de caractériser des groupes taxonomiques. Il pourrait s'agir d'activités enzymatiques maintenant la division du groupe taxonomique en deux sous groupes. Il pourrait aussi être question de métabolismes accessoires impliqués dans l'adaptation d'un groupe taxonomique à un environnement particulier. L'analyse des activités enzymatiques sélectionnées pourraient permettre de comprendre la dynamique d'évolution du métabolisme. Si plusieurs ac-

tivités enzymatiques sont capables de caractériser un même groupe taxonomique, sont elles impliquées dans la même voie métabolique? Il pourrait alors s'agir de modules fonctionnels.

Chapitre 2

Le métabolisme

Le métabolisme peut être défini ainsi : 'In an organism, the total physical and chemical processes that support energy molecule production from nutrients, and the converse use of energy molecules to support cellular and organismal homeostasis.' (nature education, 2015).

Les réactions métaboliques peuvent être spontanées, ou bien catalysées par des enzymes. Ainsi pour chaque organisme, la faculté à se développer dans un environnement particulier ou encore sa capacité de production de certains composés chimiques se reflètent dans son métabolisme.

Il implique l'utilisation et la production de **métabolites**. Un métabolite est un composé issu de la transformation biochimique d'une molécule initiale par le métabolisme. Cette réaction peut être catalysée par une **enzyme**. Une enzyme est une macromolécule, catalysant une ou plusieurs réactions chimiques, réduisant ainsi l'énergie d'activation de ces réactions et augmentant leurs vitesses.

Les micro-organismes se caractérisent par un métabolisme secondaire constitué d'une très grande variété d'enzymes aux nombreuses applications potentielles. Durant cette thèse nous nous sommes intéressés à la prédiction du métabolisme de centaines de champignons ainsi qu'à l'analyse de son évolution au travers d'applica-

tion de méthodes d'apprentissage supervisé.

Dans cette première partie, nous présentons les notions liées à la définition et à la représentation du métabolisme ainsi que celles associées à la taxonomie des champignons.

On appelle **voie métabolique** un réseau de réactions le long duquel les métabolites sont transformés. Historiquement, les voies métaboliques ont été définies en fonction de la production d'un composé donné (produit). Cette coupure du métabolisme global en sous–voies est cependant relativement arbitraire.

Le métabolisme peut être divisé en deux : métabolisme primaire, essentiel à la survie de l'organisme et métabolisme secondaire. Ces deux types de métabolismes, parce qu'ils sont plus ou moins nécessaires à l'organisme, présentent peut-être des mécanismes d'évolution différents.

2.1 Le métabolisme primaire

Le **métabolisme primaire** (aussi appelé essentiel) est composé de l'ensemble des réactions du métabolisme qui sont essentielles à la survie d'un organisme. Il regroupe les voies cataboliques (dégradation de molécules) et anaboliques (synthèse de molécules) indispensables à la croissance, au développement et à la reproduction normale de l'organisme. L'assimilation du glucose, par exemple, (voir figure 2.1) appartient au métabolisme primaire. Un métabolite primaire est un métabolite impliqué dans le métabolisme primaire. Parce qu'il est essentiel, le métabolisme primaire est généralement bien conservé entre les espèces.

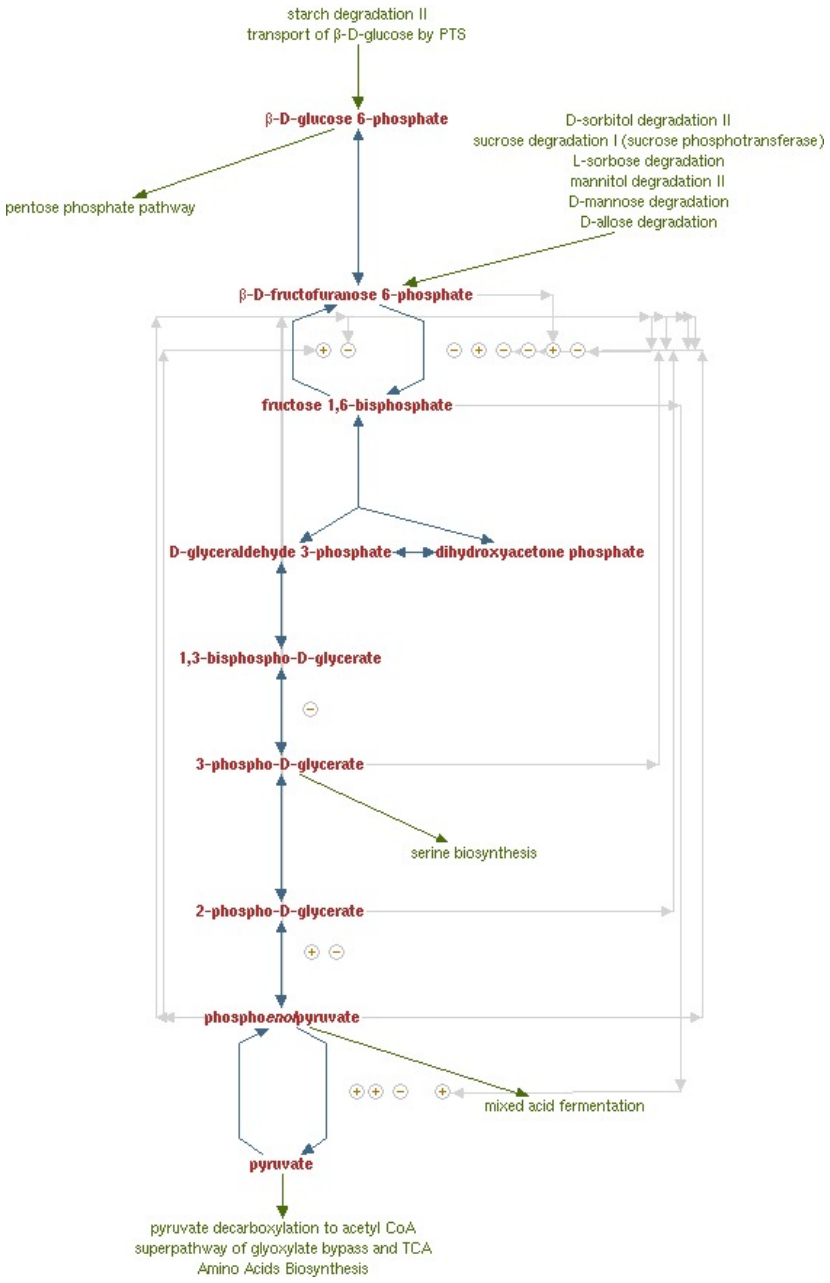


Figure 2.1 – *Représentation de la voie métabolique de la glycolyse (point d'entrée dans la voie : glucose–6–phosphate, point de sortie de la voie : pyruvate) dans la base de données MetaCyc (Caspi et al., 2008).*

2.2 Le métabolisme secondaire

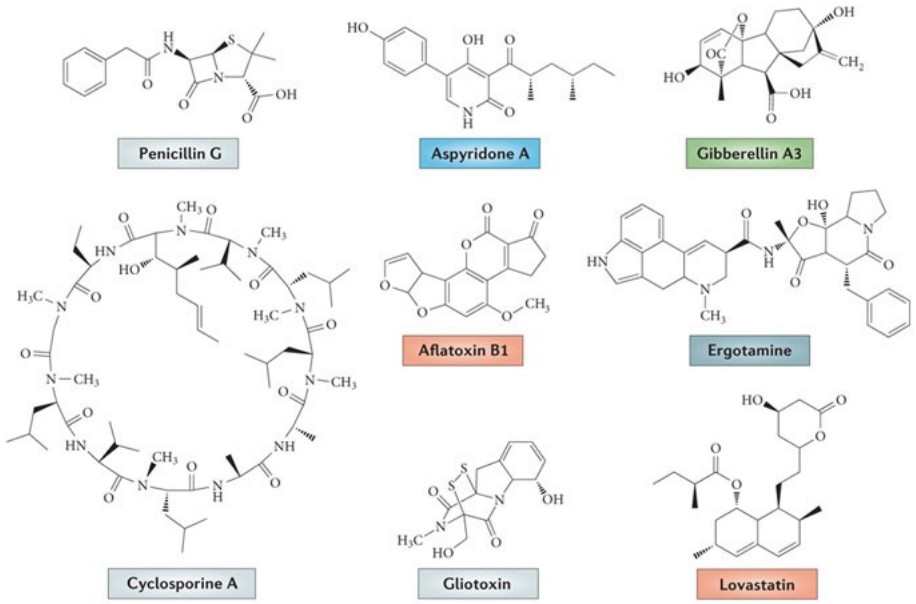


Figure 2.2 – *Exemples de métabolites secondaires produits par les champignons. Figure adaptée de (Brakhage, 2013). Le gris indique des dérivés de peptides non ribosomaux; le rouge représente des dérivés de polyketides; le bleu représente des composés résultants d'un mixte entre des polyketides non ribosomaux et des composés peptidiques; le vert représente la gibberelline dont la synthèse implique un terpène cyclase.*

Le **métabolisme secondaire** (aussi appelé métabolisme accessoire) est composé de l'ensemble des réactions du métabolisme qui ne sont pas essentielles à la survie de l'organisme dans les conditions habituelles. Il s'agit de voies permettant la synthèse de molécules non essentielles, mais contribuant à l'adaptation de l'organisme à son environnement.

Il existe une grande variété de métabolites secondaires (voir la figure 2.2). Ainsi, les champignons ont la capacité de produire une large gamme de composés ayant un impact pour l'homme. On retrouve entre autres des molécules d'intérêt

pharmaceutique telles que des antibiotiques (comme la pénicilline), des immunosuppresseurs (cyclosporines), des molécules diminuant le cholestérol (lovastatines) ou encore des molécules toxiques (aflatoxines, gliotoxines, aspyridones). On observe également des molécules sécrétées permettant la dégradation de la biomasse (enzymes dégradant la cellulose, la lignine) qui pourraient être utilisées par exemple pour la production de biocarburants.

Il existe plusieurs hypothèses quant au rôle du métabolisme secondaire (Firn and Jones, 2000, Jenke-Kodama et al., 2008).

Il a été proposé (Malik, 1980) que les réactions du métabolisme secondaire seraient plus importantes pour la cellule que les produits finaux. Par exemple, lors de conditions environnementales non optimales, le manque de certains composés peut induire un ralentissement d'une ou plusieurs étapes du métabolisme primaire, induisant l'accumulation d'un composé intermédiaire. Si ce composé est toxique à forte concentration, trouver une voie alternative pour sa dégradation (ou son stockage) devient avantageux pour l'organisme. La détoxification de l'acide phénylacétique (métabolite primaire) *via* la production du benzylpénicilline (métabolite secondaire) en est un exemple.

Dans une seconde hypothèse (Baba and Schneewind, 1998), les métabolites secondaires auraient été sélectionnés pour leur implication dans une activité biologique qui permet à son producteur d'acquérir des aptitudes (lutte contre les compétiteurs, signaux chimiques, défense de l'habitat). Rohlfs et al. (2007) ont par exemple testé l'attrait d'un prédateur (*Folsomia candida*) pour deux souches d'*Aspergillus fumigatus*, une souche sauvage, et une souche mutante pour le gène LaeA (un régulateur de production de métabolites secondaires lié à la virulence (Bok et al., 2005)). Ils ont observé que le prédateur (*Folsomia candida*) mangeait préférentiellement le champignon muté, appuyant l'hypothèse que les champignons sont devenus virulents afin de les protéger de leurs prédateurs.

Certains métabolismes secondaires, par exemple les antibiotiques, sont parfois produits en faible concentration. Il a été proposé qu'ils aient entre autres un rôle dans la signalisation et l'interaction entre cellules (Davies, 2006).

Les deux hypothèses sont complémentaires, la production de métabolites secondaires pouvant permettre la régulation de métabolites primaires tout en ayant une certaine fonction pour l'organisme (cas de la pénicilline par exemple).

Contrairement aux métabolites primaires, la production d'un métabolite secondaire particulier est généralement restreinte à un faible nombre d'espèces et liée à des conditions environnementales particulières (stress, contact avec un autre organisme, etc.). Les espèces de groupes taxonomiques différents partageant les mêmes métabolites secondaires ont pu les acquérir par convergence évolutive ou par transfert horizontal. En effet, les bactéries et les champignons présentent des **clusters** de gènes du métabolisme secondaire (Smith et al., 1990). On entend par là que les gènes du métabolisme secondaire ayant une fonction donnée sont généralement regroupés au même endroit sur le génome. Cette structure particulière pourrait expliquer une propagation des voies du métabolisme secondaire par transferts horizontaux (transferts de clusters).

2.3 Classification des enzymes : les *EC numbers*

Les réactions du métabolisme peuvent être catalysées par des enzymes. Ces enzymes sont classées en fonction des réactions qu'elles catalysent. Ce classement s'effectue *via* l'utilisation d'*enzyme commission numbers* (*EC numbers*) (Webb, 1992). Il s'agit d'un code commençant par EC et se poursuivant par une suite de quatre nombres séparés par des points (voir figure 2.3). Ces nombres représentent un raffinement progressif de la classification de la réaction.

Historiquement, cette classification a été mise en place par *the International Commission on Enzymes*. La définition de ces numéros de classification avait alors

été la suivante :

'To consider the classification and nomenclature of enzymes and coenzymes, their units of activity and standard methods of assay, together with the symbols used in the description of enzyme kinetics.'

3 Hydrolases
3.4 Peptidases
3.4.23 Aspartate-endopeptidases
3.4.23.41 peptidase de type yapsine 1

Figure 2.3 – *Exemple d'EC number : l'EC:3.4.23.41.*

Il existe six classes initiales d'enzymes (le premier nombre peut prendre six valeurs différentes).

1. les oxidoréductases catalysent des réactions d'oxydation ou de réduction,
2. les transférases permettent le transfert d'un groupe fonctionnel,
3. les hydrolases coupent un substrat par hydrolyse,
4. les lyases associent ou coupent le/les substrat(s) par des réactions n'impliquant pas l'eau,
5. les isomérases induisent des changements conformationnels au sein d'un substrat,
6. les ligases lient deux substrats en un produit par liaison covalente.

Chaque classe est divisée en sous–catégories formant une arborescence de profondeur quatre. Ainsi deux activités enzymatiques (ou réactions catalysées par un enzyme) ne différant que par le dernier nombre de l'*EC number* sont des activités enzymatiques fortement similaires.

Un *EC number* caractérise une réaction. On remarquera de ce fait que deux enzymes différentes (homologues ou non) peuvent catalyser la même réaction (même

EC number), on parle alors d'**isosymes**. De plus, une enzyme peut catalyser plusieurs réactions et ainsi être associée à plusieurs *EC numbers*

2.4 Représentations du métabolisme

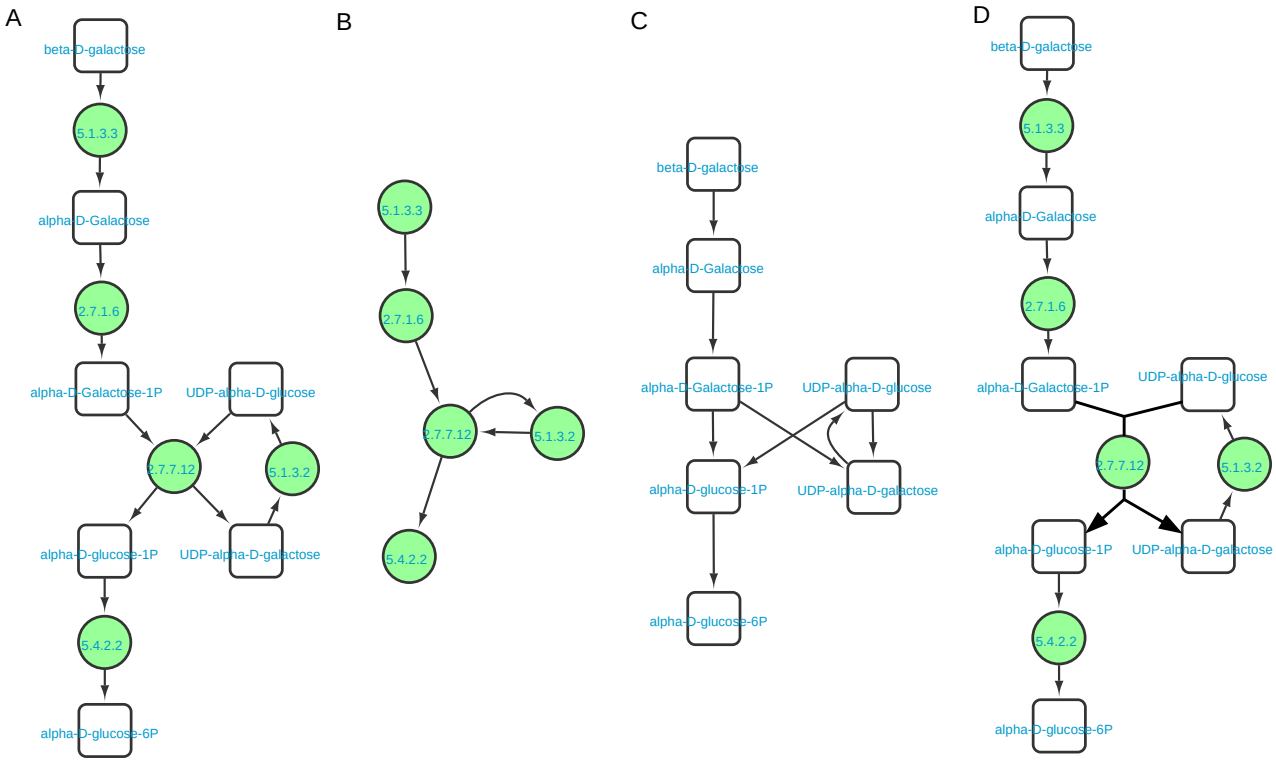


Figure 2.4 – *Représentation du métabolisme : types de graphes. Un exemple de quatre types de représentations graphiques de la voie de Leloir. A) Graphe biparti. B) Graphe centré sur les activités enzymatiques. C) Graphe centré sur les composés. D) Hypergraphe biparti. Figure réalisée à partir du logiciel cytoscape (Shannon et al., 2003).*

Le métabolisme peut se représenter par différents graphes appelés graphes métaboliques ou réseaux métaboliques (Takuji and Peer, 2009). Il existe plusieurs types de graphes métaboliques en fonction des entités représentées par les nœuds

et les arêtes (voir la figure 2.4). Les graphes bipartis présentent, comme leur nom l'indique, deux types de nœuds. Le premier type figure les réactions (symbolisée par les *EC numbers*) et le second, les composés. Dans ce type de graphe, les nœuds ne peuvent être reliés qu'à des nœuds de type différent. Un moyen de simplifier ce graphe est de ne conserver que l'un des deux types de nœuds. Il existe ainsi les graphes centrés sur les réactions (*reaction maps* ou *reaction-centric network*) dans lesquels les *EC numbers* forment les nœuds du graphe et les arêtes figurent l'utilisation du produit d'une des enzymes comme substrat de l'enzyme suivante (voir figure 2.4.C). De manière complémentaire, on retrouve des graphes centrés sur les composés (*compound-centric maps*) présentant les interactions entre les métabolites. Les métabolites sont alors symbolisés par les nœuds du graphe. Les enzymes sont figurées par les arêtes. Ce type de graphe est par exemple utilisé par les bases de données KEGG (Kanehisa and Goto, 2000), MetaCyc (Caspi et al., 2008) et Reactome (Joshi-Tope et al., 2005).

Enfin, les hypergraphes présentent des hyperarêtes pouvant connecter plus de deux nœuds. Il s'agit d'une généralisation de la notion de graphe, et, de ce fait, il peut y avoir des hypergraphes des trois types précédents. Les hypergraphes permettent de tenir compte du fait que plusieurs composés sont souvent nécessaires à une même réaction.

Sur ces graphes, le degré d'un nœud (*EC* ou composé) est le nombre de liens (ou arêtes) qui sont incidents à ce nœud. On dira d'un ensemble de nœuds qu'ils forment un sous graphe connexe si il existe un chemin entre chaque paire de sommets.

2.5 Bases de données de voies métaboliques

Il existe un grand nombre de bases de données dédiées aux voies métaboliques. Une liste non exhaustive est donnée dans le tableau 2.1. Ces bases ont pour

	MetaCyc	**keggdb**	**Model SEED**	**Reactome**	**BiGG**	**Uni-pathway**
Vérifications manuelles	+	–	–	+	+	+
# d'organismes	>1,000	>1,000	>200	21	6	>1,000
Génomes	+	+	+	–	–	–
Protéomes	+	+	+	+	–	–
Réactions	+	+	+	+	+	+
Métabolites	+	+	+	+	+	+
Voies	+	+	+	+	–	+
Connexion requise	–a	–	–a	–	+	-

Table 2.1 – *Caractéristique des principales bases de données portant sur le métabolisme (page web BioCyc, Kanehisa and Goto, 2000, page web model SEED, page web Reactome, page web BIGG, Morgat et al., 2012). 'Génomes' signifie que les séquences génomiques sont disponibles. 'Protéomes' signifie que les propriétés des enzymes, telles que la structure des sous-unités, les inhibiteurs, les cofacteurs sont accessibles. 'Voie' signifie que les informations sur les voies et leurs diagrammes sont disponibles. Le 'a' au niveau de la connexion signifie que cette dernière est nécessaire pour construire des modèles, mais pas pour voir des modèles existants.*

but de représenter de manière structurée les connaissances de la communauté. Cependant, elles ne font pas les mêmes choix d'espèces prises en compte et de gestion des données non vérifiées ou manquantes. Récemment, deux publications ont résumé les différences et similitudes entre plusieurs de ces bases (Jing et al., 2014, Stobbe et al., 2014). Pour nos travaux, nous avons principalement utilisé les bases de données KEGG (Kanehisa and Goto, 2000) et MetaCyc (Caspi et al., 2008).

2.5.1 KEGG

La *Kyoto Encyclopedia of Genes and Genomes* (KEGG) (Kanehisa et al., 2014) est une base de données fondée en 1995 et dédiée à la compréhension de systèmes biologiques. KEGG propose des voies métaboliques, des groupes d'orthologues ainsi que des modules de séquences génomiques (modules de voies, complexes structuraux, modules fonctionnels, marqueurs de phénotypes).

KEGG est un ensemble de bases de données constitué de 15 bases (voir tableau 2.2) parmi lesquelles KEGG PATHWAY, KEGG BRITE et KEGG MODULE ont pour vocation d'être des références dans leurs domaines. Les voies métaboliques décrites dans KEGG sont l'union des voies métaboliques connues pour différents organismes. Il s'agit donc de mosaïques n'étant pas présentes dans leur ensemble dans un unique organisme.

Les groupes d'orthologues de KEGG (KEGG Orthology) sont construits manuellement. Leur identifiant est appelé K.

KEGG définit des modules. Il s'agit de combinaisons de groupes d'orthologues. Il existe dans cette base quatre types de modules; les modules de voies, les complexes structuraux, les modules fonctionnels et les modules correspondant à des signatures de phénotypes particuliers.

Ces modules sont définis manuellement en analysant la conservation d'un schéma de présence/absence de gènes dans un ensemble de génomes ainsi que la

Catégorie	Nom de la base	Contenu
	KEGG PATHWAY	Graphes KEGG des voies métaboliques
	KEGG BRITE	Base de données d'ontologie représentant les hiérarchies fonctionnelles d'objets biologiques divers
	KEGG MODULE	Modules de voies, complexes structuraux, modules fonctionnels, marqueurs de phénotypes
Génomique	KEGG ORTHOLOGY	Groupes d'orthologues
	KEGG GENOME	Description des organismes aux génomes complets
	KEGG GENES	Catalogues des gènes
	KEGG SSDB	Gènes ayant des similarités de séquence
Chimie	KEGG COMPOUND	Métabolites et autres petites molécules
	KEGG GLYCAN	Structures expérimentalement déterminées de glycanes
	KEGG REACTION	Réactions biochimiques
	KEGG RPAIR	Paires de réactants
	KEGG RCLASS	Classification des réactions
	KEGG ENZYME	Nomenclature des enzymes
Santé	KEGG DISEASE	Maladies humaines
	KEGG DRUG	Médicaments
	KEGG DGROUP	Groupes de médicaments
	KEGG ENVIRON	Drogues brutes, huiles essentielles et substances impliquées dans la santé

Table 2.2 – *Description de la base de données KEGG*

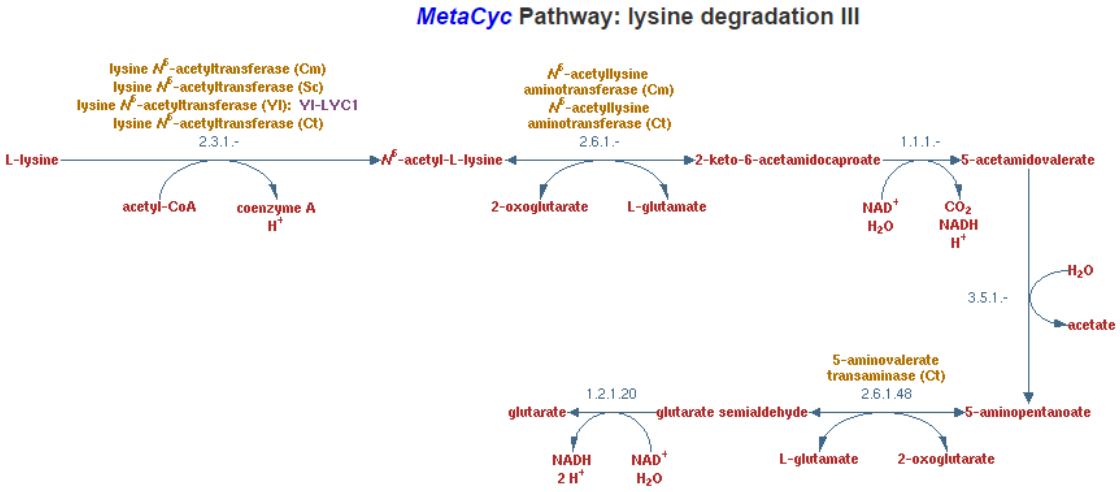


Figure 2.5 – *Voie MetaCyc de dégradation de la lysine (champignons).*

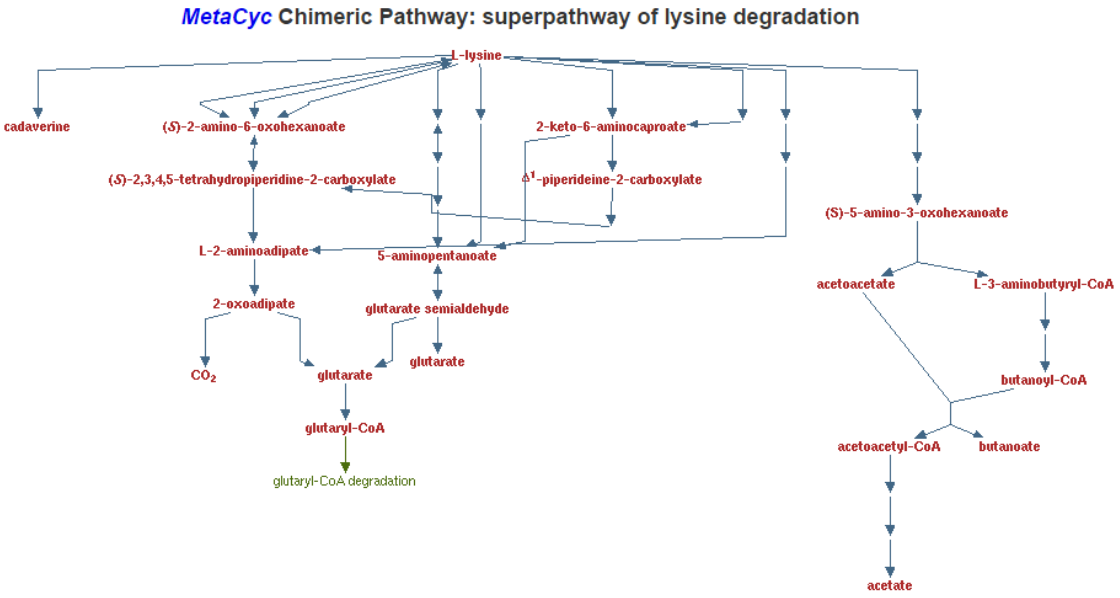


Figure 2.6 – *Super voie MetaCyc de dégradation de la lysine. Cette voie inclut la voie de dégradation de la lysine chez les champignons (voir figure 2.5).*

corrélation de leurs positions (structure de type opéron). Le module M00010 par exemple, correspond à la première oxydation du carbone de l'oxaloacétate produisant le 2–oxaloglutarate (cycle du cytrate). Il est défini dans KEGG par la formule suivante : K01647 et (K01681 ou K01682) et (K00031 ou K00030).

Un phénotype particulier peut être caractérisé par une expression booléenne faisant intervenir plusieurs modules. De ce fait, il est permis de définir des modules à partir d'une combinaison d'autres modules afin de caractériser un phénotype.

Une comparaison des gènes d'un organisme aux groupes d'orthologues de KEGG peut permettre d'extraire la ou les sous–voies correspondant à cet organisme ainsi que la liste des modules présents.

2.5.2 MetaCyc

MetaCyc (Caspi et al., 2008) est une base de données non redondante de voies métaboliques vérifiées expérimentalement. Elle présente une liste de réactions, de composés et de gènes. Plus de 2579 organismes sont représentés, majoritairement des micro-organismes et des plantes. Plus de 2255 voies métaboliques sont répertoriées, incluant plus de 12074 réactions métaboliques Cette base répertorie l'ensemble des réactions catalysées par une enzyme et présentant un *EC number*.

Les voies de MetaCyc figurent des voies réellement trouvées chez des organismes. On y retrouve les voies dites de 'base' qui sont des voies présentant des réactions individuelles (par exemple la voie de dégradation de la lysine chez les champignons, voir la figure 2.5) et les 'super' voies métaboliques contenant un ensemble de voies, de 'super' voies et de réactions (par exemple la dégradation de la lysine chez l'ensemble des organismes, voir la figure 2.6).

2.5.3 Comparaison des bases de données KEGG et MetaCyc d'un point de vue du métabolisme

Les voies de KEGG sont généralement trois à quatre fois plus grandes que les voies de MetaCyc, cela s'explique par le fait que MetaCyc cherche à modéliser les voies d'organismes particuliers alors que KEGG travaille sur l'union des voies de plusieurs organismes.

Les voies de MetaCyc sont l'équivalent des modules de KEGG (voir la figure 2.7). Les 'super' voies de MetaCyc sont comparables aux voies de KEGG.

Catégorie	MetaCyc (tout)	KEGG (tout)	MetaCyc (base)	KEGG (module)	MetaCyc (super voies)	KEGG (voies)
# voies			1846	179	296	237
# composés	11991	15161	5371	828	5523	4759
# réactions	10262	8692	6155	878	6348	6174
# réactions par voie			4.37	6.22	14.24	28.84
# composés par voie			11.49	15.27	25.63	37.45

Table 2.3 – *Comparaison de KEGG et MetaCyc : Nombre de voies, nombre total de composés décrits et de composés par types de voie, nombre total de réaction et nombre de réactions par type de voie, nombre moyen de réactions par voies et nombre moyen de composés par voies. Tableau adapté de (Altman et al., 2013) sur des versions de KEGG et MetaCyc datant de 2012.*

Une comparaison exhaustive de KEGG et MetaCyc a été réalisée en 2013 (Altman et al., 2013). Le tableau 2.3 résume une partie des résultats présentés dans cet article. Il a été observé que KEGG contient significativement plus de composés que MetaCyc. Cependant, MetaCyc présente plus de voies métaboliques.

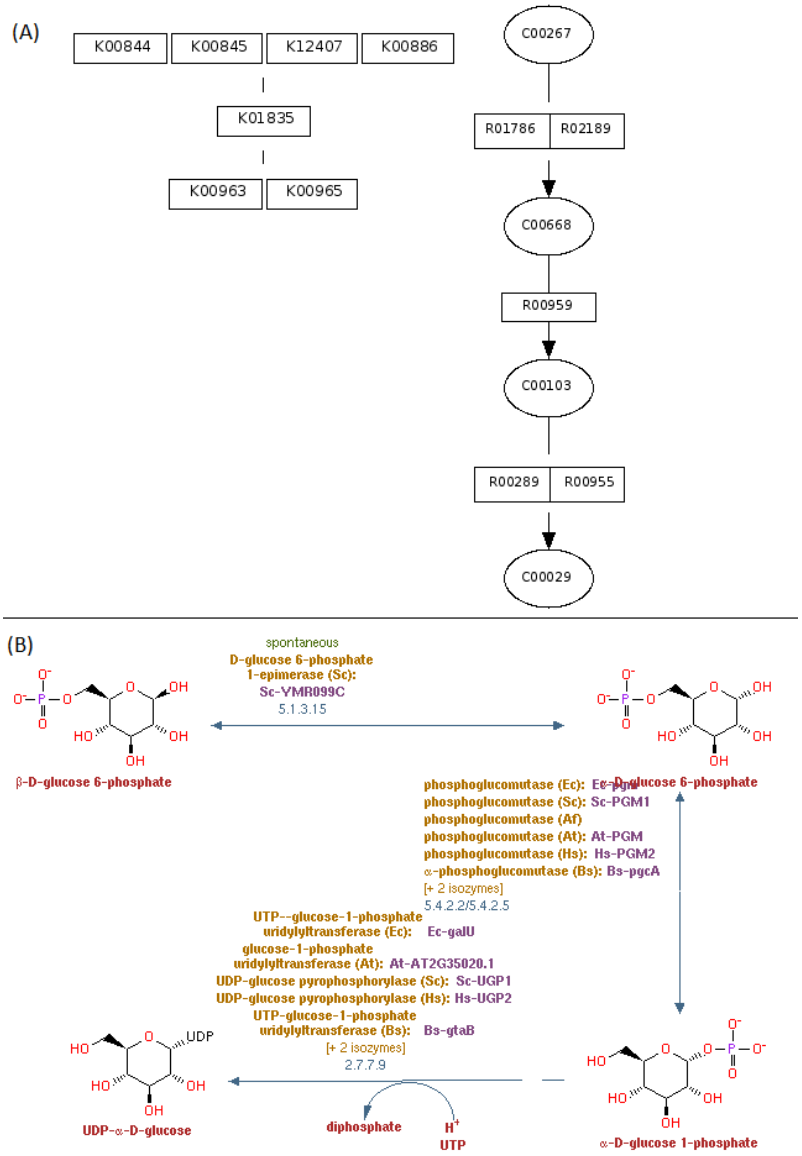


Figure 2.7 – *Biosynthèse de l'UDP-glucose dans les bases de données KEGG et MetaCyc. (A) Module KEGG. Ce module fait intervenir sept groupes d'orthologues (K), quatre composés (C) et cinq réactions (R). Pour qu'un organisme présente ce module, il lui faut avoir au moins l'un des groupes K00844 ou K00845 ou K12407 ou K00886, le groupe K01835 et l'un des groupes K00963 ou K00965. (B) Voie MetaCyc. Les flèches figurent les réactions, les noms des enzymes sont en jaune, les gènes sont en violet et les EC numbers en bleu. Le module KEGG et la voie MetaCyc sont ici identiques mis à part la réversion possible de certaines réactions dans MetaCyc qui sont indiquées irréversibles dans KEGG.*

En particulier, la partie module de KEGG est incomplète (718 modules KEGG contre 2255 voies MetaCyc en février 2015). Les deux bases de données présentent chacune environ 9000 réactions décrites et 6200 réactions présentes dans des voies.

Le nombre de réactions identiques dans les deux bases de données est relativement faible (1961 parmi les 6378 réactions de MetaCyc et les 6174 réactions de Kegg). Ces deux bases de données semblent donc antagonistes.

Taxon	**Voie métabolique**	**% Unique**
Organismes cellulaires	1840	47.7
Eukaryotes	1094	46.8
Champignons	219	35.6

Table 2.4 – *Pourcentages de voies de MetaCyc relatives à un taxon et absentes de KEGG. La colonne % unique représente la fraction des voies métaboliques de base de MetaCyc de ce taxon qui sont uniques par rapport aux voies de KEGG. Extrait d'un tableau de (Altman et al., 2013).*

La distribution de la taxonomie représentée dans les deux bases diffère. Lors de leur comparaison en 2013 (Altman et al., 2013), il a été observé que les voies présentes chez les plantes, les métazoaires, les actinobactéries et les champignons étaient plus représentées chez MetaCyc que chez KEGG (voir tableau 2.4). Il s'agit principalement de voies trouvées initialement chez les plantes, mais présentes également chez les vertébrés, les chordés, les métazoaires, les champignons, les archées et les protéobactéries. Les voies présentes dans KEGG et absentes dans MetaCyc sont des voies portant sur la dégradation xénobiotique, le métabolisme du glycane et le métabolisme des terpenoides et polyketides.

2.6 L'évolution du métabolisme

Deux grandes théories sont en concurrence pour expliquer l'évolution du métabolisme, le modèle rétrograde et le modèle patchwork (voir figure 2.8). Ces deux modèles, plutôt qu'opposés, sont certainement complémentaires.

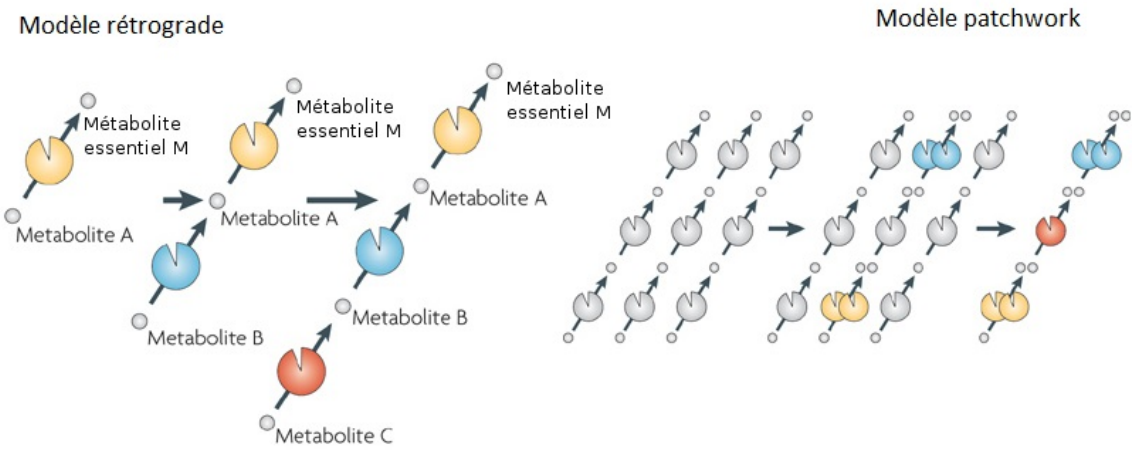


Figure 2.8 – *Les deux modèles d'évolution du métabolisme : le modèle rétrograde à gauche et le modèle patchwork à droite (figure adaptée de Takuji and Peer (2009))*

Le modèle d'évolution rétrograde (Horowitz, 1945) pose l'hypothèse que les voies métaboliques évoluent de manière rétrograde par rapport à la direction de la voie en réponse à une diminution de concentration d'un substrat dans l'environnement (voir la figure 2.8). Ce modèle présuppose que les métabolites intermédiaires sont présents dans l'environnement, épuisés par les réactions et régénérés par de nouvelles réactions.

Prenons l'exemple de la figure 2.8. Le métabolite M est essentiel à l'organisme. Il est produit à partir du substrat A, cette réaction est catalysée par l'enzyme jaune. Un organisme incapable de produire le métabolite A commencera par utiliser tout le métabolite A présent dans son environnement, induisant une forte diminution

de la concentration environnementale de A. Le recrutement d'une enzyme capable de produire le métabolite A à partir d'un substrat B présent dans l'environnement procurera alors un avantage sélectif à l'organisme. Cependant, comme pour A, la concentration en B dans l'environnement diminuera avec son utilisation, et de ce fait, une enzyme catalysant la production de B à partir d'un métabolite C sera sélectionnée.

Notons que puisque l'enzyme jaune est déjà capable de lier le métabolite A, il y a une plus grande chance que la seconde enzyme recrutée soit issue d'une duplication du gène codant pour l'enzyme jaune plutôt que d'un autre gène codant pour une enzyme sans affinité avec A. Ce modèle est étayé par des études qui ont montré que des enzymes homologues (issues de gènes dérivants d'un même gène ancestral) sont significativement plus souvent à une distance inférieure à 3 étapes l'une de l'autre dans le graphe métabolique que des enzymes non homologues (Alves et al., 2002).

Le modèle d'évolution patchwork (Jensen, 1976, Lazcano and Miller, 1996) suppose que les enzymes augmentent leur spécificité pour un substrat donné après duplication. Ainsi, il est considéré qu'initialement les enzymes présentent une large gamme de substrats et de produits, même si certains d'entre eux sont favorisés. Cette large gamme de substrats et de produits par enzyme permet la génération d'une large gamme de composés. La duplication de gènes dans ces voies métaboliques apporte un avantage sélectif par l'augmentation de la production du métabolite d'intérêt. Au final, aux événements de duplication succèdent des événements de spéciation permettant la spécialisation d'une de ces voies (voir la figure 2.8).

Ce modèle est étayé par plusieurs études. Il a par exemple été montré chez *Escherichia coli* que deux protéines homologues ont deux fois plus de chances d'intervenir dans deux voies différentes que d'intervenir dans la même voie (Teichmann et al., 2001).

Chapitre 3

Les champignons

3.1 Taxonomie

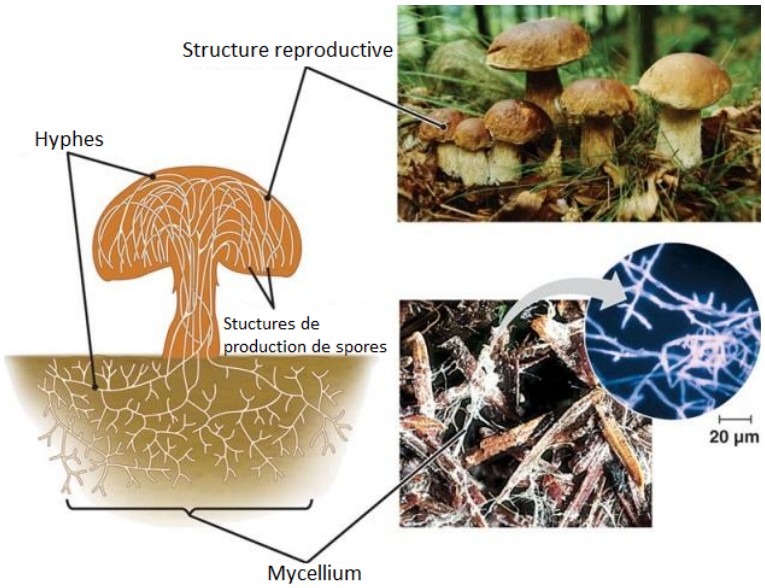


Figure 3.1 – *Exemple de morphologie d'un champignon.*

L'étude de la diversité des organismes est depuis longtemps un sujet de recherche en biologie. De morphologique, la classification des champignons est à

présent basée sur des méthodes phylogénomiques, les convergences évolutives étant très nombreuses chez ces organismes. Ainsi, les champignons qui appartiennent au même groupe taxonomique ne présentent pas forcément la même morphologie ni même la même stratégie nutritionnelle. Les champignons sont des organismes osmotrophes (ils transportent des nutriments par des transporteurs). Ils peuvent être saprophytes (se nourrir à partir de matière morte), parasites ou mutualistes. Ces trois stratégies peuvent être présentes dans le même groupe taxonomique.

On s'imagine communément les champignons sous leur forme comestible que l'on trouve sur les marchés (voir figure 3.1). Il ne s'agit cependant là que d'appareils de dispersion des spores appelés carpophores. Il existe également des champignons unicellulaires. Ils peuvent se présenter sous forme de levure (comme *Saccharomyces cerevisiae*), mais aussi sous forme de cellules avec ramifications (rhizoide) ou de cellules plurinucléées. Parmi les formes pluricellulaires, on retrouve des organismes unis ou plurinucléés.

Les organismes pluricellulaires pourvus de cellules plurinucléées sont les plus courants. Les cellules s'alignent et forment des hyphes (voir figure 3.1), structures polarisées permettant la colonisation du substrat. Ces hyphes sont capables de se brancher, générant un réseau. Chez certains champignons les hyphes sont de plus capables de fusionner, permettant la reproduction sexuée.

Les champignons se divisent en deux groupes principaux, les champignons dits 'supérieurs' et les champignons appelés communément 'inférieurs'.

Les champignons 'supérieurs' ou **dikarya** (composé des ascomycota et des basidiomycota voir la figure 3.2) dérivent d'un ancêtre commun ayant acquis la capacité d'avoir des cellules à deux noyaux. Ils ont acquis une aptitude croissante à la dégradation de la biomasse et la différenciation de sporophores pluricellulaires. Les dikarya regroupent les basidiomycetes et les ascomycètes. Les spores des deux groupes sont fondamentalement différentes (voir figure 3.3), ce qui a permis leur

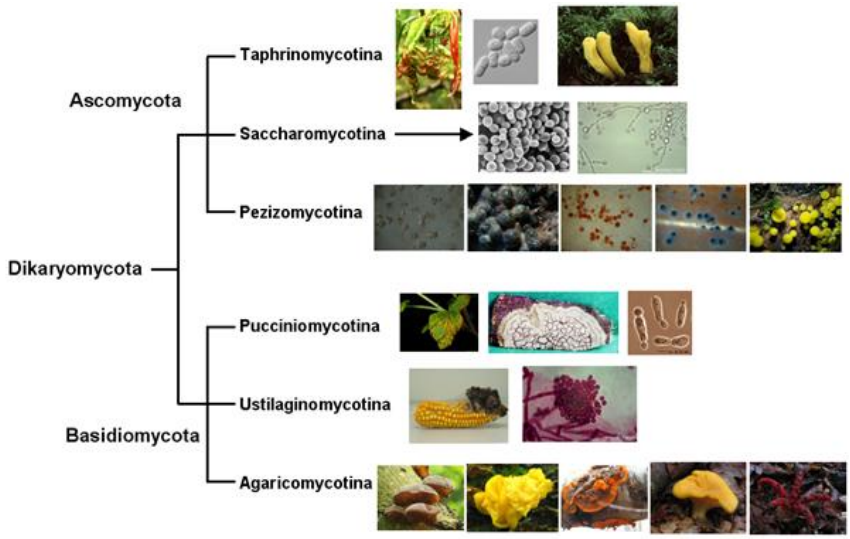


Figure 3.2 – *Sous-embranchement chez les champignons supérieurs (Dikarya). L'arbre se lit de gauche à droite. Il présente quelques photographies d'organismes appartenant à chaque sous-embranchement. Figure adaptée du cours de mycologie du Pr Philippe Silar (page web Philippe Silar).*

(a) Un des modèles d'asques typiques des ascomycètes (huit spores par asque) (photo D. Borganiro).

(b) Un type de baside typique des basidiomycètes avec des spores fixées à l'extérieur (photo S. Poumarat).

Figure 3.3 – *Illustration de la différence des spores d'ascomycètes et de basiodiomycètes. Les ascomycota ont des spores incluses dans des asques. Les basidiomycota présentent des spores produites à l'extérieur, sur des cellules spécialisées appelées basidia.*

distinction dès le XIXe siècle. Ces deux groupes présentent de nombreuses convergences évolutives. De nombreuses espèces des deux groupes ont par exemple évolué indépendamment des structures de reproduction (carpophores) pluricellulaires.

Basiodiomycota

Parmi les **basidiomycota** (liv, 2013), les pucciniomycota regroupent des saprophytes, des pathogènes de plantes et des parasites de champignons. Ils incluent les rouilles (parasites des cultures) dont certaines présentent des cycles biologiques extrêmement complexes (*Puccinia graminis* fait par exemple intervenir deux hôtes, le blé et l'épine-vinette).

Un second sous–embranchement des basidiomycota est celui des ustilagomycotina. Cet embranchement contient principalement des pathogènes de plantes dont *Ustilago maydis*, un pathogène du Maïs qui crée des tumeurs noires pouvant être consommées.

Les agaricomycotina représentent deux tiers des basidiomycota actuellement connus. Parmi eux, les tremellomycetes peuvent être saprophytes ou parasites. Les dacrymycetes se nourrissent des arbres en décomposition. Enfin, les agaricomycetes regroupent des saprophytes, des champignons mutualistes, des parasites ou encore des lichens. Parmi les champignons de ce groupe, on retrouve les champignons les plus efficaces pour dégrader les végétaux.

Ascomycota

Les **ascomycota** représentent 60% des champignons répertoriés. La moitié est des lichens (association avec des algues).

Le sous–embranchement des taphrinomycotina présente différentes classes dont certaines incluent des champignons produisant des cloques sur les feuilles des arbres, des levures (*Schizosaccharomyces pombe* permet de faire de la bière) ou

encore des champignons présents dans les poumons des mammifères (pneumocystidiomycetes).

Le sous–embranchement des saccharomycotina contient des champignons passant la majeure partie de leur vie sous forme de levure. On y trouve des champignons ayant colonisé une grande variété de niches écologiques (nectar de fleurs, fruits en décomposition, sol, vase etc.) et fortement utilisés dans l'industrie et l'agroalimentaire (levure du boulanger *Saccharomyces cerevisiae*, production de fromages *Geotrichum candidum* etc.).

Les pezizomycota, un troisième sous–embranchement des ascomycota, présentent une grande diversité de modes de dispersion et d'envahissement du substrat. Ce groupe présente de nombreuses convergences évolutives avec le groupe des *agaricomycetes (basidiomycota)* dont la capacité qu'ont certains organismes de dégrader la lignine. Ils produisent généralement de plus petits carpophores que les agaricomycetes. Les amateurs de champignons reconnaîtront dans la classe des *pezizomycetes* les morilles et les truffes. Les dothideomycetes sont principalement des pathogènes de plantes. Les eurotiomycetes regroupent les aspergillus et les penicillium. Ils ont la particularité de se disperser grâce à des spores asexuelles et présentent des capacités de production d'acides organiques et d'antibiotiques. On retrouve également dans ce groupe des champignons responsables de mycoses. Les leotiomycetes présentent des asques sans opercules et sont généralement des parasites de plantes. Les sordariomycetes présentent des organismes modèles ainsi que des pathogènes de plantes ou d'insectes. Enfin, d'autres classes présentent des lichens, des champignons pathogènes d'insectes ou encore capables de produire des pièges différenciés pour capturer les animaux.

Champignons inférieurs

Les champignons dits 'inférieurs' forment un groupe paraphylétique (n'incluant pas l'ensemble des organismes du taxon) regroupant l'ensemble des champignons n'appartenant pas au groupe taxonomique des dikarya. Ils sont incapables de différencier des structures de dispersion pluricellulaires hormis quelques truffes simples.

Les chytridiomycota, les neocallismastigomycota et les blastocladiomycota sont souvent aquatiques et se dispersent grâce à des zoospores (spores flagellés et mobiles). Il s'agit principalement de saprophytes.

Les entomophthromycotina, les zoopagomycotina, les kickxellomycotina et les microsporidies sont des parasites.

Parmi les mucoromycotina on retrouve des saprophytes et des mycorhiziens (association entre les myceliums des champignons et les racines des plantes terrestres). Ils sont très abondants dans les sols et présentent une différenciation de sporophores ressemblant à des petites truffes.

Les glomeromycota représentent 5 à 10% de la biomasse microbienne du sol. Ces organismes vivent en symbiose avec les racines de certaines plantes, formant ainsi des endomycorhizes. Il existe ainsi des échanges entre plantes et champignons au niveau d'hyphes ayant pénétré les parois des cellules de la périphérie des racines. Ces champignons ne présentent pas de sexualité, mais leurs hyphes sont capables de fusionner, ce qui leur permet de conserver une diversité génétique.

Il y a actuellement 100 000 espèces de champignons décrites, cependant, on estime leur nombre à environ 1,5 millions d'espèces (Hawksworth, 1991). La taxonomie que nous décrivons ici ne tient compte que des génomes de champignons entièrement séquencés (environ 300 actuellement) ainsi que des six gènes séquencés dans plusieurs organismes à des fins de phylogénies par la communauté *Assembling the fungal tree of life* (AFTOL). Le séquençage des champignons à des fins notam-

ment de recouvrement de l'ensemble de la taxonomie fongique est en cours.

3.2 Principales initiatives de séquençage des champignons

Saccharomyces cerevisiae a été le premier génome eucaryote séquencé (Goffeau et al., 1996). Son séquençage a offert la possibilité d'effectuer des études de génétique inverse à large échelle (Aparicio et al., 2001, Boone et al., 2007). S'en est suivi le séquençage d'autres champignons, permettant la mise en place de techniques de génomique comparative (Stajich et al., 2007). On a ainsi pu identifier de nouveaux gènes, déterminer des familles de protéines et détecter des éléments de régulation ainsi que de mécanismes moléculaires liés à l'évolution. Cela a montré la portée des méthodes de génomique comparative. Plusieurs centres de recherche travaillent actuellement activement au séquençage et à l'annotation structurale de ces organismes (Cuomo and Birren, 2010).

Le Broad Institute abrite la *Fungal Genome Initiative*. Ce groupe a été mis en place en 2000. Il séquence et analyse des génomes de champignons ayant un impact en médecine, en agriculture et en industrie. Ces génomes (plus de 50), incluent des pathogènes de l'homme et de végétaux ainsi que des organismes modèles. La liste des champignons étudiés par la 'Fungal Genome Initiative' est disponible sur leur site web (page web FGI).

Un second projet est à mettre en avant pour son initiative de séquençage de champignons : '1000 génomes de champignons'. Ce projet a pour but de permettre la compréhension de la diversité fongique. Le *Joint Genome Institute of the Department of Energy* (JGI) ainsi que onze équipes internationales ont lancé le projet de séquencer plus de 1 000 génomes de champignons en cinq ans (Grigoriev et al., 2011). L'échantillonnage a été fait de manière à séquencer au moins deux génomes de

référence appartenant à chacune des plus de 500 familles de champignons connues. La liste des génomes actuellement séquencés ou en cours de séquençage est disponible sur le site web (page web 1000 genomes). Le portail MycoCosm (Grigoriev et al., 2014) permet l'accès aux données, la visualisation et l'accès à des méthodes de génomique comparative. Il répertorie à la fois des génomes séquencés par le JGI, mais également des génomes issus d'autres ressources.

3.3 Liste et descriptions des espèces de travail

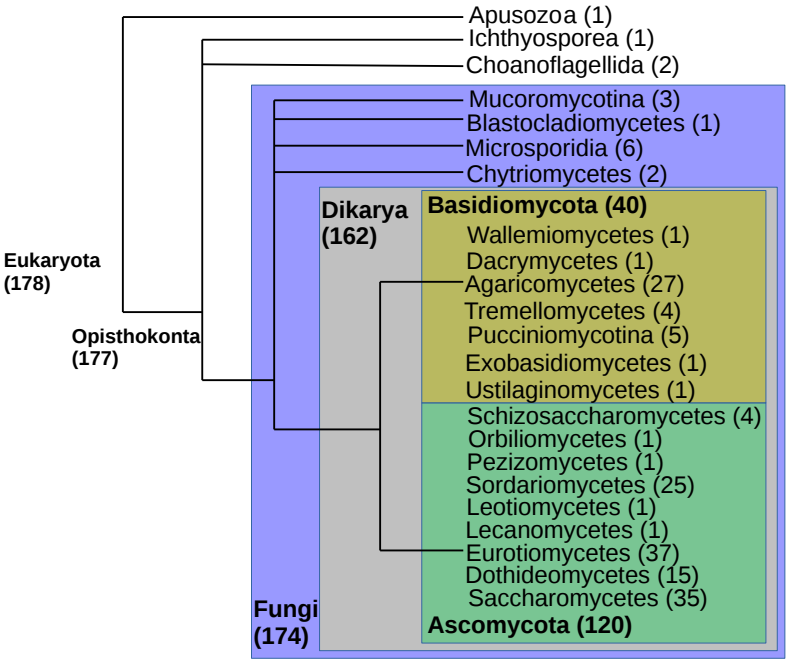


Figure 3.4 – *Résumé de la taxonomie des espèces utilisées. Pour chaque groupe taxonomique le nombre d'espèces appartenant au groupe est indiqué après son nom et entre parenthèses.*

Source	**# d'espèces téléchargées**
Antonospora locustae DB	1
Ashbya genome Database	1
Broad Institute	70
DSM Anti–Infectives	1
Genolevures	7
Genoscope	1
IGM (Orsay)	6
JGI	70
NCBI	8
NCBI	8
NITE	1
Sanger	2
Standford	1
TIGR	1
TuberDB	1
University of Kentuky	1
URGI	1
WashU	4
MIT	1

Table 3.1 – *Bases de données utilisées pour le téléchargement de protéomes et nombre de protéomes téléchargés.*

Dans le cadre de cette thèse, nous avons principalement travaillé sur 178 protéomes (majoritairement de champignons). Il s'agit de l'ensemble des champignons qui étaient complètement séquencés au 1er janvier 2011. Ces génomes ont été téléchargés sur dix-huit bases de données différentes, principalement sur le JGI et le Broad Institute (voir le tableau 3.1).

Nous travaillons actuellement à la mise à jour de nos travaux sur l'ensemble des champignons séquencés à ce jour et disponibles sur le site du JGI, à savoir plus

de 350 organismes.

La liste des protéomes avec lesquels nous avons travaillé ainsi que leur taxonomie détaillée est fournie dans le tableau 3 en annexe.

II Détection de groupes d'orthologues

Chapitre 4

Introduction

4.1 Motivations

Le but de ce travail a été de comparer les répertoires enzymatiques de différentes espèces de champignons. L'obtention de ces répertoires enzymatiques n'est pas triviale, les génomes des différentes espèces n'étant pas complètement annotés expérimentalement (du Plessis et al., 2011). Ainsi, la première étape de notre travail a logiquement consisté en la réannotation fonctionnelle automatique et homogène des protéomes de champignons.

Il existe un grand nombre de méthodes de prédiction de fonction protéiques. Ces méthodes sont basées sur différents types de données (une méthode peut utiliser une combinaison de ces données) :

- la similarité des séquences protéiques (Thomas et al., 2003).
- la synténie des gènes (Overbeek et al., 1999).
- les fusions de gènes (Enright et al., 1999, Reid et al., 2010).
- la similarité des profils phylogénétiques (que nous détaillerons dans la partie 3) (Pellegrini et al., 1999).

- la structure 3D des protéines (Laskowski et al., 2005).

- les données d'expression géniques (si des gènes sont co-exprimés dans différentes conditions on en déduit qu'ils sont impliqués dans la même fonction cellulaire).

- les données d'interaction protéine-protéine.

Nous nous sommes plus particulièrement intéressés aux approches basées sur la détection d'orthologues. En effet, il a été montré que les orthologues présentent une similarité de fonction (Altenhoff et al., 2012, Rogozin et al., 2014).

Différentes méthodes de prédiction de groupes d'orthologues existent. Elles sont basées sur deux principales approches (Kristensen et al., 2011) : (i) la similarité de séquences et (ii) la comparaison entre l'arbre des gènes et l'arbre des espèces.

Afin d'évaluer les différentes méthodes de prédiction de groupes d'orthologues, plusieurs jeux de données ont été proposés dont *OrthoBench* (Trachana et al., 2011) et *Protein reference database* (Gabaldon et al., 2009, Dessimoz et al., 2012, page web EBI). Afin de savoir si nous pouvions sans biais utiliser l'une des plus de 30 méthodes actuelles de prédiction de groupes d'orthologues, nous avons décidé d'en comparer un sous–ensemble composé des plus couramment utilisées. L'application de quatre méthodes différentes de prédiction de groupes d'orthologues sur ces ensembles de protéomes de référence a induit la prédiction de groupes d'orthologues dissemblables (voir Table 1 de l'article (Pereira et al., 2014)). Ainsi, se limiter à une méthode particulière induit un biais, et la sélection d'une d'entre elles n'est pas aisée, chacune ayant ses avantages (Hulsen et al., 2006, Dalquen et al., 2013).

De ce fait, nous proposons et évaluons ici une méta-approche de prédiction de groupes d'orthologues.

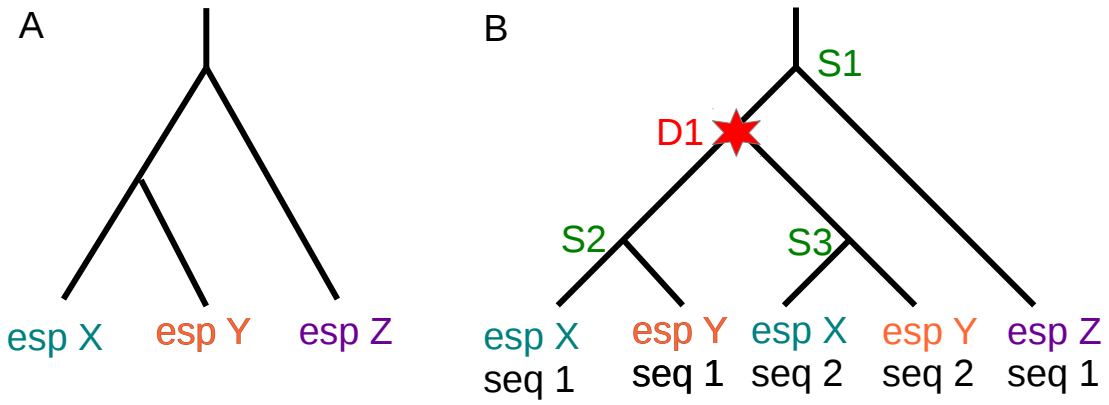


Figure 4.1 – *Comparaison de l'arbre des espèces (A) et d'un arbre obtenu avec un ensemble de protéines homologues (B) (figures adaptées de (Altenhoff and Dessimoz, 2012)). (A) Arbre des espèces avec 3 espèces : X, Y et Z. (B) Arbre représentant un scénario évolutif d'une séquence à travers les 3 espèces (X, Y et Z). Les 'S' verts indiquent des événements de spéciation. S2 et S3 représentent la même spéciation. Le 'D' ainsi que l'étoile rouge indiquent un événement de duplication.*

4.2 Définitions

Deux gènes partageant un ancêtre commun sont dits **homologues**. Une relation d'**orthologie** est définie pour une paire de gènes homologues ayant commencé à diverger par un événement de spéciation. Les orthologues peuvent être divisés en plusieurs groupes : les '*one-to-one*', les '*one-to-many*', les '*many-to-one*' et les '*many-to-many*'. Les qualificateurs '*one*' et '*many*' indiquent pour chacun des deux gènes s'ils ont subi une duplication ayant lieu après le premier événement de spéciation et ainsi s'il existe pour un gène ancestral un ou plusieurs gènes dans une espèce qui sont orthologues à un ou plusieurs gènes de l'autre espèce. La relation d'orthologie n'est pas transitive. Cela se comprend sur la figure 4.1 : les gènes espXseq1 et espZseq1 sont orthologues, les gènes espYseq2 et espZseq1 sont orthologues, mais les gènes espXseq1 et espYseq2 ne sont pas orthologues du fait de la duplication

D1.

La **paralogie** se définit pour une paire de gènes homologues ayant commencé leur divergence par un événement de duplication. Ainsi, sur la figure 4.1, les gènes espXseq1 et espYseq2 sont paralogues.

Plusieurs types d'orthologues et de paralogues permettent de préciser le propos :

- **In-paralogue** : relation impliquant une paire de gènes ainsi qu'un événement de spéciation de référence. Une paire de gènes sont in-paralogues s'ils impliquent deux gènes paralogues dont l'événement de duplication initial a eu lieu après l'événement de spéciation de référence (relation orthologues *one-to-many, many-to-many* ou *many-to-one*). Sur la figure 4.1 par exemple, les gènes X1 et X2 sont in-paralogues en référence à l'événement de spéciation S1.

- **Out-paralogue** : comme pour les in-paralogues, une relation d'out-paralogie se définit pour une paire de gènes et un événement de spéciation de référence. Une paire de gènes est dite out-paralogue si l'événement de duplication qui les relie est plus ancien qu'un événement de spéciation de référence. Sur la figure 4.1 par exemple, les gènes X1 et X2 sont out-paralogues en référence à l'événement de spéciation S2.

- **Co-orthologues** : relation définie entre trois gènes. Lorsque deux des gènes sont in-paralogues par rapport à un événement de spéciation associé au troisième gène, on dit alors que les deux gènes in-paralogues sont co-orthologues au troisième gène. Ainsi, sur la figure 4.1.B les gènes X1 et Y2 sont co-orthologues au gène Z1.

Les relations d'orthologie et de paralogie se définissent donc en fonction de paires ou de triplets de gènes. Cependant, à des fins d'annotation, de nombreuses équipes ont cherché à définir des groupes de gènes orthologues. Dans notre exemple

(figure 4.1) les gènes espXseq1 et espZseq1 ainsi que les gènes espXseq2 et espZseq1 sont orthologues. Cependant les séquences espXseq1 et espXseq2 ne sont pas orthologues. L'orthologie n'est donc pas une notion transitive. De ce fait, la notion de groupes d'orthologues peut avoir plusieurs définitions.

Inparanoid (Remm et al., 2001) définit des **groupes d'orthologues** entre deux espèces comme des groupes contenant des orthologues et des in-paralogues. Avec la publication de OrthoMCL (Li et al., 2003) la définition de 'groupes d'orthologues' devient celle de groupes de gènes incluant des orthologues ainsi que des paralogues 'récents'. La notion de 'récent' est censée limiter l'ajout de paralogues ayant divergé en fonction, mais n'est pas définie plus précisément. Plus tard, il a été proposé dans une étude évaluant différentes méthodes de prédiction d'orthologues (Altenhoff and Dessimoz, 2009) de n'autoriser dans ces groupes que les orthologues ainsi que les in-paralogues issus de duplications postérieures à tous les événements de spéciation du groupe.

Les groupes d'orthologues hiérarchiques sont définis en fonction d'un événement de spéciation donné (Boeckmann et al., 2011, Gabaldón and Koonin, 2013). Ainsi, tous les paralogues issus de duplications postérieures à l'événement de spéciation observé sont considérés comme appartenant au groupe d'orthologues (présence dans le groupe d'orthologues et d'in-paralogues par rapport à un événement de spéciation donné). Cette définition est utilisées de manière implicite par plusieurs bases de données. C'est le cas par exemple des bases COG, EggNog et OrthoDB.

Dans le cadre de notre étude, nous travaillons sur un groupe taxonomique particulier, celui des champignons. Notre but est d'obtenir des groupes de protéines ayant conservé la même fonction afin de pouvoir annoter les protéomes de ces organismes. Ces groupes de proteines peuvent être des groupes définis au niveau de la spéciation de ces organismes. Dans le cas de groupes incluant des changements de fonctions dus à l'apparition de fortes divergences au sein de paralogues, nous souhai-

tons séparer le groupe pour former deux groupes distincts homogènes en fonction. Nous obtiendrons donc possiblement un ensemble de groupes d'orthologues définis pour des niveaux hiérarchiques différents. Cependant nous limiterons l'appartenance d'une protéine à au plus un groupe.

Chapitre 5

État de l'art des méthodes de détection d'orthologues

Il existe des dizaines de méthodes de détection de paires ou de groupes d'orthologues. Ces méthodes sont souvent associées à leur propre base de données répertoriant les orthologues prédits pour l'ensemble ou des sous–partie du vivant. Il existe principalement deux grandes techniques de détection d'orthologues, les méthodes basées sur les graphes et celles basées sur l'analyse d'arbres phylogénétiques.

5.1 Méthodes de détection d'orthologues basées sur les graphes

Les méthodes basées sur les graphes partent de l'hypothèse que les séquences 'proches' en fonction d'un certain critère sont orthologues.

Elles fonctionnent généralement en trois étapes :

- la prédiction de paires d'homologues
- la création d'un graphe de relation d'homologie,

- et le *clustering* des séquences du graphe (induisant l'ajout ou le retrait de relations d'homologie).

5.1.1 Prédiction de paires d'orthologues

***Best reciprocal hit* (BRH)** (Overbeek et al., 1999) fut la première méthode proposée pour la détection de groupes d'orthologues. Elle consiste en l'extraction des paires de meilleurs hits réciproques sur les résultats blast de comparaison des protéomes deux à deux. Les couples de protéines sont ensuite filtrés (en fonction par exemple d'un ratio de scores blast ou d'un pourcentage d'alignement). Cette méthode présente de bons résultats sur différents benchmarks (Salichos and Rokas, 2011, Wolf and Koonin, 2012, Dalquen et al., 2013). Elle a cependant été critiquée, notamment par Koski et Goldin qui montrent que le hit blast le plus proche n'est souvent pas le plus proche voisin phylogénétique (Koski and Golding, 2001).

De ce fait, il a plus tard été proposé d'évaluer la distance évolutive entre deux séquences plutôt que leur similarité Blast. La méthode **'RSD'** (*Reciprocal Smallest Distance*) (Wall et al., 2003) évalue la distance entre deux protéines en fonction de l'estimation du nombre de substitutions nécessaires pour passer de la séquence de l'une à la séquence de l'autre.

Dans le but de limiter les protéines pour lesquelles le calcul du nombre de substitutions est fait, une première étape consiste à comparer par Blast une protéine à un protéome donné. Les hits répondant à un certain seuil d'e-value sont ensuite alignés deux à deux avec la protéine initiale afin d'estimer le nombre de substitutions. Les distances ainsi calculées sont alors utilisées pour déterminer la protéine la plus proche de la protéine initiale. Si cette relation est réciproque alors les deux protéines sont considérées comme orthologues.

Ces deux méthodes comparent les protéomes d'espèces différentes. La recherche du meilleur hit réciproque induit de ne trouver résultant de la comparaison

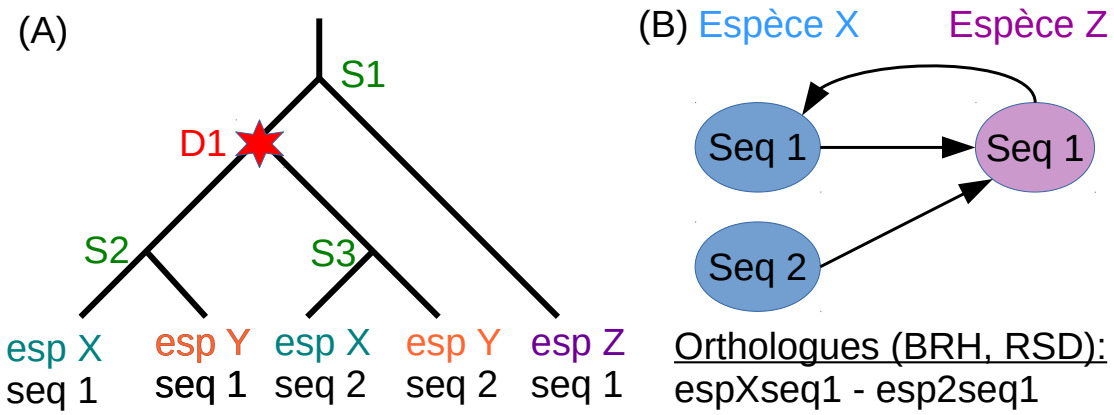


Figure 5.1 – *Exemple de la comparaison des espèces X et Z par les méthodes BRH ou RSD. (A) Arbre représentant un scénario évolutif d'une séquence à travers les 3 espèces (X, Y et Z) (même arbre qu'en figure 4.1. Les 'S' vert indiquent des événements de spéciation. Le 'D' ainsi que l'étoile rouge indiquent un événement de duplication. (B) Les meilleurs hits entre les différentes séquences sont figurés par des flèches. Les espXseq1 et espZseq1 sont trouvés orthologues, mais, la conservation de l'unique meilleur hit pour espZseq1 induit de ne pas trouver la relation d'orthologie espXseq2 – espZseq1.*

de deux espèces que deux gènes par groupe d'orthologues. Cela ne permet pas la prédiction d'in-paralogues (voir figure 5.1). De plus, cette méthode induit le manque de 60% des relations d'orthologie chez les plantes et les animaux (Dalquen and Dessimoz, 2013).

Afin de parer au manque de certaines relations des méthodes BRH et RSD (dans les cas d'orthologues de type *many-to-many* ou *one-to-many*) (voir figure 5.1) plusieurs méthodes ont été proposées.

La méthode **Inparanoid** (Remm et al., 2001) est basée sur une approche de type BRH. Comme BRH et RSD, elle travaille à partir de paires de génomes. Elle permet cependant la prédiction de relations d'orthologie de type *many-to-many*, *one-to-many* et *many-to-one*, non prises en compte par BRH *via* la prédiction d'in–paralogues. Pour chaque paire de protéomes, les Blastp réciproques sont effectués

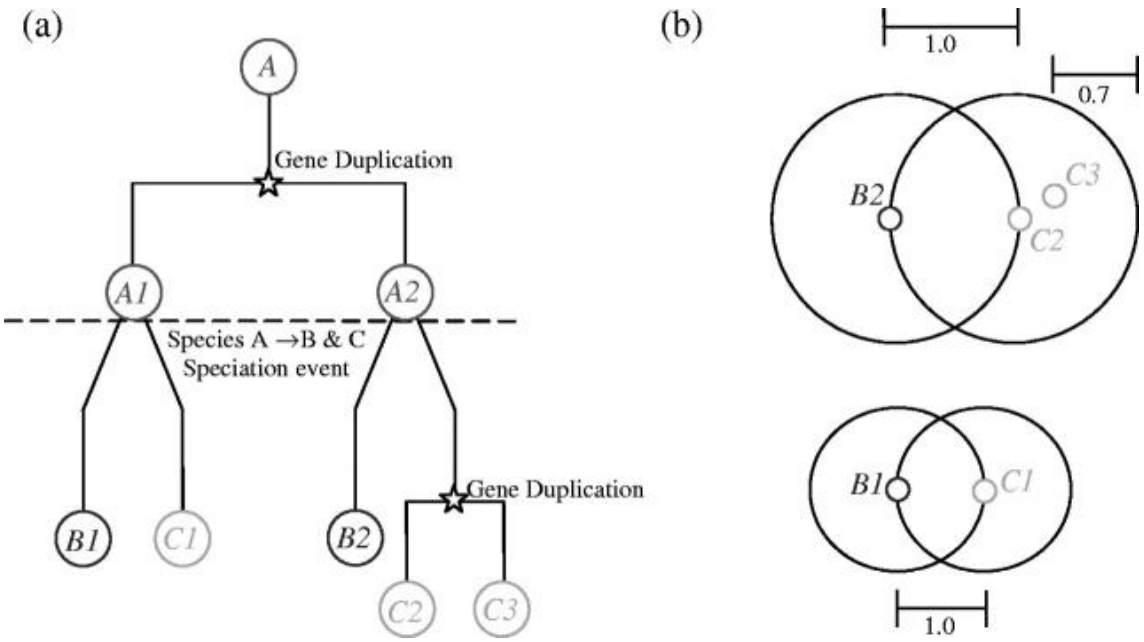


Figure 5.2 – *Principe de fonctionnement d'Inparanoid (figure tirée de l'article (Remm et al., 2001)). Exemple d'un arbre de gènes (a) et du cluster obtenu par Inparanoid (b). (a) Duplication du gène portant la protéine A dans l'espèce A. Des événements de spéciations induisent l'apparition des espèces B et C. Dans l'espèce C, les gènes C2 et C3 sont in–paralogues et co-orthologues à B2. B1 est out-paralogue à C2 et C3 par rapport au dernier événement de spéciation. (b) B2-C2 et B1-C1 sont les deux orthologues trouvés originalement. La comparaison d'une espèce contre elle-même permet de trouver les paires respectivement plus similaires que ces paires d'orthologues et d'ainsi prédire les inparalogues. On observe ici la prédiction de deux clusters, B2-C2-C3 et B1-C1. Si la distance entre C2 et C3 avait été supérieure à celle entre B2 et C2 alors C3 ne serait pas apparue dans le cluster B2-C2.*

ainsi que les Blastp de chaque protéome contre lui-même. Les meilleurs hits réciproques sont sélectionnés en tant qu'orthologues. Les scores Blast de ces hits réciproques sont alors utilisés comme seuil de définition des in-paralogues. L'analyse du résultat du Blastp de l'espèce contre elle-même pour le gène impliqué dans la relation d'orthologie permet de sélectionner comme in-paralogues l'ensemble des gènes ayant un score inférieur à celui des deux orthologues (voir figure 5.2). Cette méthode permet ainsi la prédiction de paires d'orthologues ou d'in-paralogues entre deux espèces. Inparanoid en est à présent à sa huitième version (Sonnhammer and Östlund, 2014)

Plutôt que d'ajouter des paralogues après avoir déterminé les orthologues, il a été proposé d'être moins stringent sur la notion de meilleur hit réciproque (utilisée par BRH) en sélectionnant les ensembles de séquences respectant certains seuils. La méthode **OMA** (*orthologs matrix project*) (Altenhoff et al., 2015), par exemple, est basée sur un algorithme qui compare les gènes en fonction de leur distance évolutive. Elle ne conserve pas les paires de protéines réciproquement les plus similaires, mais toutes celles respectant les critères de seuil (portant sur le score blast et la longueur d'alignement). Cela permet de ne pas perdre les liens issus de relations d'orthologie de type *many-to many*. Cette méthode produit une liste de paires d'orthologues et d'in-paralogues. La base de données OMA contient au 1er décembre 2014 des groupes d'orthologues prédits pour 1706 protéomes complets.

5.1.2 Méthodes de *clustering* de graphe : prédiction de groupes d'orthologues

Les méthodes de prédiction d'orthologues de type graphe prédisent généralement des paires d'orthologues (et possiblement d'in-paralogues) entre paires de génomes. Les paires de gènes orthologues et in-paralogues forment un graphe, les nœuds étant les gènes et les arêtes les relations d'orthologie. Il est intéressant lorsque

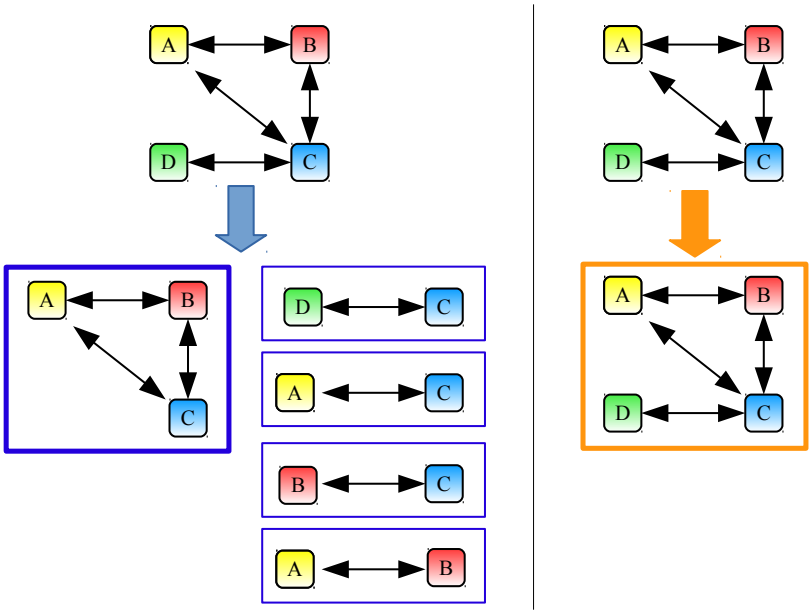


Figure 5.3 – *Comparaison méthodes de clustering "lien simple" (à gauche) / "lien complet" (à droite). Dans le cas des groupes obtenus par la méthode lien complet, il existe des groupes chevauchants. Dans notre approche, si deux groupes sont chevauchants nous ne conservons que les plus grands. Dans l'exemple, seul le groupe A-B-C serait conservé.*

l'on souhaite travailler avec plus de deux espèces d'obtenir des groupes d'orthologues. Afin d'obtenir des groupes à partir des relations prédites par paires d'espèces, des méthodes de *clustering* peuvent être appliquées.

La méthode la plus simple pour créer des groupes d'orthologues est d'agréger les paires d'orthologues présentant un gène en commun. Ainsi, la méthode **lien simple** (voir figure 5.3) consiste à rechercher l'ensemble des composantes connexes du graphe. Cette méthode permet d'obtenir de grands groupes, mais présente généralement l'inconvénient de regrouper des groupes d'orthologues différents, car une erreur de prédiction d'orthologie par la méthode de prédiction de paire induit la fusion de deux groupes. De plus, les cas de fusion de gènes induisent la fusion des deux groupes dont ils sont issus. Dans le but d'obtenir des groupes à partir des résultats

d'Inparanoid nous avons utilisé pour la publication (Pereira et al., 2014) la méthode "lien simple" au travers d'une heuristique permettant de sélectionner des groupes pour lesquels chaque protéine présente au moins 20% de liens avec les autres protéines du groupe. Notons que nous aurions également pu utiliser Multiparanoid, une méthode basée sur Inparanoid pour obtenir des groupes d'orthologues. Cependant, lors de sa thèse, Sandrine Grossetête–Lalami avait testé Multiparanoid sur quatre puis onze génomes (Lalami, 2010). Alors qu'elle n'avait eu aucun problème à obtenir les résultats avec quatre génomes, Multiparanoid ne retournait aucun résultat avec les onze génomes. De ce fait, nous n'avons pas travaillé avec ce programme. Depuis nos travaux, il a été publié la méthode Hieranoid (Schreiber and Sonnhammer, 2013), une nouvelle méthode basée sur Inparanoid afin de créer des groupes d'orthologues. Notons que nous pourrions à présent utiliser cette méthode.

Complémentaire à la méthode lien simple, la méthode **lien complet** (voir figure 5.3) consiste en la recherche de sous parties du graphe pour lesquels l'ensemble des nœuds sont connectés, appelées cliques. Ces dernières formeront les groupes d'orthologues prédits. Ainsi, si deux groupes sont chevauchants, le groupe le plus grand est sélectionné et le second groupe n'est pas conservé. La méthode "lien complet" permet d'obtenir des groupes plus spécifiques que ceux obtenus avec la méthode "lien simple". En effet, une prédiction d'orthologues entre A et B qui s'avérerait être un faux positif impliquerait certainement que le gène B ne serait pas trouvé orthologue à l'ensemble des orthologues de A dans les autres espèces. De ce fait, plus le nombre d'espèces est grand plus la méthode est stringente. La contrepartie est que cette méthode de *clustering* est sensible aux erreurs de type manque de liens dans le graphe, induisant des sur-divisions des groupes d'orthologues. La recherche de cliques a été utilisée par OMA pour extraire des groupes de leurs paires d'orthologues prédites. Dans notre travail, nous avons combiné cette méthode de recherche de clique avec la méthode BRH. Nous avons ainsi obtenu une méthode stringente à

la fois sur la prédiction de paires (pas de prédiction d'in-paralogues), mais également sur la prédiction de groupes.

Une alternative moins stringente que la méthode "lien complet" est la recherche au sein du graphe de triplets d'orthologues. Il s'agit du *clustering* de type **COG** (*Clusters of Orthologous Groups*) (Tatusov et al., 1997). Cela a été la première méthode proposée comme alternative à la méthode "lien simple" pour la construction de clusters de gènes. Elle se base sur un graphe de relation de meilleurs hits blasts. Ces meilleurs hits n'ont pas à être réciproques. Elle utilise en tant que graine les triangles présents dans le graphe des relations d'orthologies prédites. Ces triangles doivent donc faire intervenir trois protéines d'espèces différentes. Les triangles présentant une face commune sont joints. La complexité en temps est en $O(n^3)$ avec n le nombre de génomes. L'analyse de la base de données COGs a montré qu'environ un tiers des groupes d'orthologues contiendrait des paralogues (Dessimoz et al., 2006).

Le programme OrthoMCL (Li et al., 2003, page web OrthoMCL) utilise quant à lui un *clustering* utilisant une approche probabiliste, le **clustering Markovien**. Cette méthode est basée sur la recherche de clusters dans la matrice de similarité entre les séquences. Dans un premier temps, un graphe est construit, les nœuds sont les gènes, les arêtes symbolisent une similarité entre les séquences protéiques. Elles sont pondérées en fonction du score de cette similarité. Les groupes d'orthologues doivent donc former au sein de ce graphe des clusters de forte connexité. MCL calcul de manière déterministe les probabilités associés à une marche aléatoire au sein de ce graphe. Du fait de la forte connexité au sein d'un groupe d'orthologues, la probabilité de se déplacer et de rester au sein du même groupe d'orthologues est plus grande que celle de sortir du groupe. Ainsi le 'marcheur' aura tendance à visiter plusieurs fois des nœuds du même groupe d'orthologues avant de visiter un nouveau groupe. C'est ce phénomène qui est utilisé pour retirer les liens les moins souvent parcourus. Au final, les composantes connexes restantes formeront les groupes d'or-

thologues.

Nous avons listé dans cette partie un sous-ensemble des méthodes de type graphe. Pour ceux intéressés par la diversité des méthodes existantes, nous proposons de se référer au tableau 1 en annexe. Ce tableau présente et compare douze méthodes de détection de groupes d'orthologues de type graphe. Il est basé sur une combinaison de plusieurs articles de revues ainsi que sur des recherches personnelles.

5.2 Les méthodes de détection de groupes d'orthologues basées sur la phylogénie

Les méthodes basées sur les arbres utilisent un modèle explicite de l'évolution d'une famille de gènes (l'arbre de la famille de gènes) de manière à prédire des relations d'orthologies. Elles profitent avantageusement de l'information contenue dans l'alignement multiple des séquences protéiques de manière à déduire les événements de spéciation, duplication ou perte de gènes.

Ces méthodes utilisent généralement des groupes d'homologues pré–construits (ou des banques de données de domaines) et valident ou modifient ces groupes par l'analyse de l'arbre phylogénétique obtenu après l'alignement multiple des séquences du groupe. Elles cherchent ainsi à expliquer la nature des événements associés aux nœuds internes de l'arbre phylogénétique et travaillent généralement avec l'hypothèse que les événements de perte de gènes, de duplication ou de transferts horizontaux sont indépendants. Elles peuvent ou non faire intervenir l'arbre des espèces. Nous présenterons ici des approches des deux types. Parce qu'il existe un grand nombre de méthodes, nous ne les détaillerons pas toutes. Cependant, le tableau 2 à la fin de cette partie décrit et compare un sous–ensemble des méthodes publiées.

5.2.1 Méthodes basées sur la comparaison de l'arbre des espèces avec l'arbre du groupe d'homologues

La nature des événements associés aux nœuds internes de l'arbre des séquences peut être déduite par comparaison avec l'arbre des espèces. Cette procédure est appelée *tree reconciliation*.

Deux types d'approches ont été développées pour la réconciliation de l'arbre des gènes avec l'arbre des espèces : les méthodes utilisant la parcimonie et celles basées sur les probabilités.

Les méthodes basées sur la parcimonie utilisent un modèle discret de l'évolution des gènes. Chaque événement de duplication ou de perte de gène est associé à un coût. Leur but est de trouver le coût minimum.

L'algorithme LCA (Goodman et al., 1979), basé sur la parcimonie, permet l'association des nœuds de l'arbre des gènes à des événements de spéciation ou de duplication *via* sa comparaison avec un arbre des espèces complètement résolu. Soit G un arbre des gènes et S un arbre des espèces. LCA associe chaque gène u de G à l'espèce x de S la plus récente de manière à ce que chaque gène contemporain qui descend de u appartienne à une espèce contemporaine qui descend de x (voir figure 5.4). Ainsi, un nœud u de G est une duplication si et seulement si il est associé au même nœud de S que l'un de ses fils. L'arbre réconcilié peut alors être construit. Pour chaque nœud u de G associé à un nœud x de S, si u est une duplication alors elle est localisée sur la branche juste au-dessus de x, sinon u (événement de spéciation) est placé sur x.

La réconciliation de type LCA est implémentée pour tourner en temps linéaire en fonction du nombre d'espèces (Bender and Farach-Colton, 2000).

Un second modèle, le **modèle probabiliste** est utilisé pour associer une probabilité qu'un certain scénario d'évolution mène à un certain arbre des gènes G, connaissant l'arbre des espèces, dont les longueurs des branches sont connues. Il

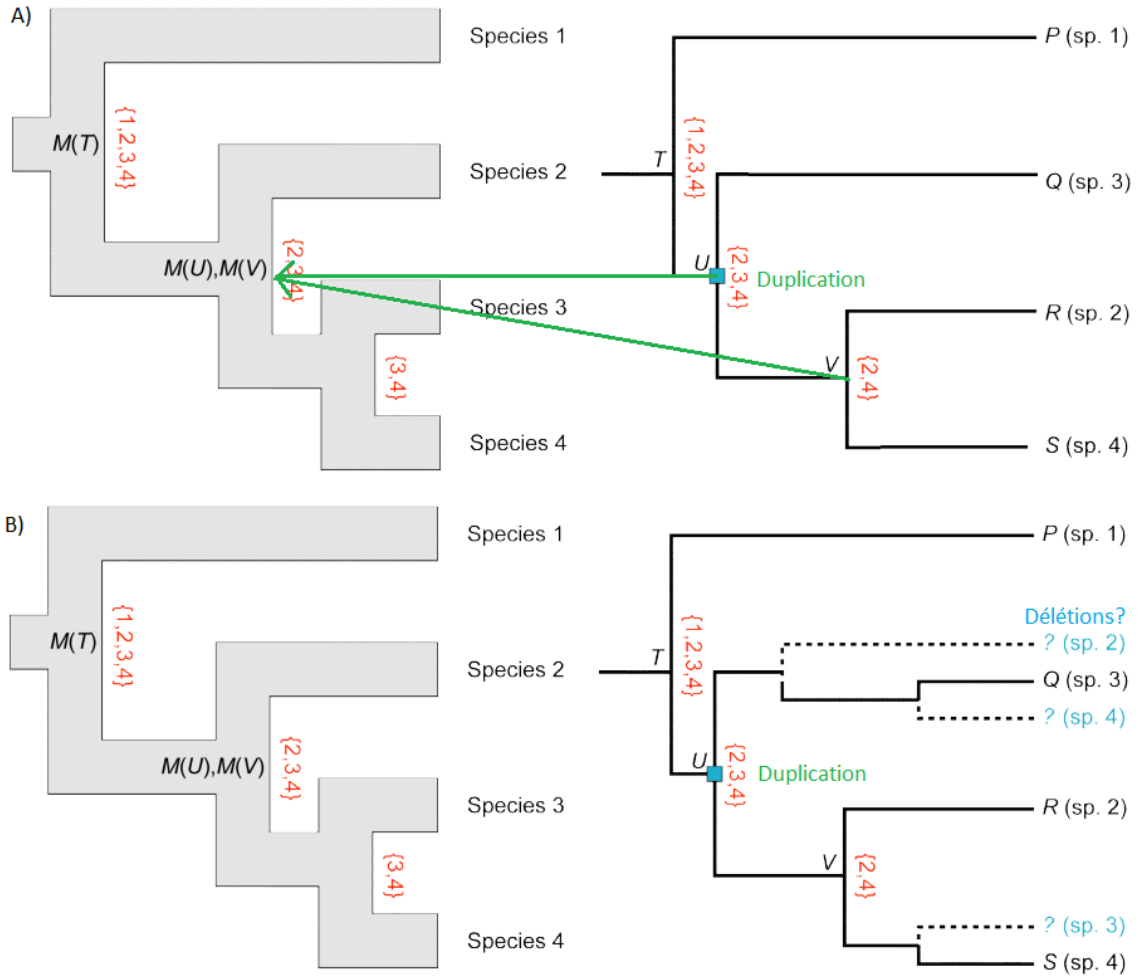


Figure 5.4 – *A) Recouvrement LCA entre arbre des espèces (à gauche) et arbre des gènes (à droite). Le nœud ancestral u de G est déduit comme étant une duplication, car il coïncide avec le même nœud M(U),M(V) de S que l'un de ses nœuds fils (V). B) Réconciliation de l'arbre des gènes avec l'arbre des espèces.*

s'agit d'un modèle continu de l'évolution des gènes.

Ce modèle simule l'évolution d'un gène u le long d'un sous–arbre S enraciné en x comme suit :

- un gène u passe d'un nœud x a son nœud fils x1 par un procédé *birth-and-death* le long de la branche (x, x1) reflétant sa survie, sa duplication ou sa perte,
- et pour chaque descendant de u qui survit après x1 l'étape 1 est répétée.

Ce procédé permet de prendre en considération des événements 'fantômes' comme la duplication suivie de la perte d'un gène.

Un algorithme efficace (Arvestad et al., 2009) permet de calculer les probabilités associées à tous les arbres de réconciliations en $O({n_g}^2 n_s)$ avec n_g et n_s le nombre de nœuds de l'arbre des gènes et de l'arbre des espèces. *PrimeGEM* utilise quant à lui un algorithme MCMC pour estimer les probabilités a posteriori des relations d'orthologie (Arvestad et al., 2009).

Les modèles probabilistes et parcimonieux ont été comparés sur 1278 arbres de gènes de 12 champignons (Doyon et al., 2012). Ils ont abouti en des résultats très similaires, montrant que la méthode basée sur la parcimonie (plus rapide) peut être utilisée dans le but d'obtenir des prédictions précises.

Réconcilier l'arbre des gènes et l'arbre des espèces est NP-complet lorsque l'on tient compte des transferts horizontaux (Conow et al., 2010), mais peut être réduit à un problème polynomial en imposant des contraintes réalistes (nombre de gènes par espèces, nombre de transferts, consistance de localité de temps entre les espèces (Merkle and Middendorf, 2005)). Une erreur dans l'arbre des espèces aboutit naturellement à des biais dans la réconciliation. Plusieurs méthodes ont été proposées pour parer à ce problème. Il a ainsi été proposé de fusionner les nœuds de branches peu fiables (Berglund-Sonnhammer et al., 2006, Durand et al., 2006). Cette technique de nœuds avec plus de deux fils est utilisée par *Ensemble compara project* (Vilella et al., 2009a).

Toutes ces méthodes tiennent compte des événements de duplication et de spéciation. Seules certaines d'entre elles gèrent également les événements de transferts horizontaux.

La gestion des transferts horizontaux est difficile, car elle implique un transfert entre deux espèces coexistantes à un moment donné. De plus, le gène transféré peut avoir été perdu ou avoir subi des duplications/spéciations dans les espèces descendantes de l'espèce donneuse.

5.2.2 Méthodes n'utilisant que l'arbre des gènes

L'arbre taxonomique peut contenir des erreurs. L'obtention d'un arbre 'parfait' est un sujet de recherche encore actuel. De ce fait, certaines méthodes font abstraction de l'arbre des espèces et n'utilisent que l'arbre obtenu avec les séquences du groupe d'orthologues.

Un premier ensemble d'algorithmes utilisent un étiquetage des événements évolutifs des nœuds internes de l'arbre des gènes sur la base du recouvrement entre les espèces des différents sous-arbres, on appelle cette méthode *species-overlap*. C'est le cas par exemple des méthodes **LOFT** (van der Heijden et al., 2007), **'Phylogeny'** (Lemoine et al., 2007), **Branchclust** (Poptsova and Gogarten, 2007) et **phylomeDB** (Huerta-Cepas et al., 2007).

Ces méthodes peuvent nécessiter un arbre des gènes enraciné ou effectuer l'enracinement elles-mêmes. Plusieurs méthodes d'enracinement de l'arbre ont été proposées. Basée sur un modèle d'évolution constante du génome, la méthode de détection d'orthologues **Orthostrapper** enracine l'arbre en son milieu. Nous noterons que cette heuristique est peu vraisemblable, car il est connu que toutes les familles de protéines ne suivent pas ce type d'évolution. D'autres méthodes proposent d'enraciner l'arbre de manière à minimiser un certain critère (parcimonie). Il peut s'agir par exemple de minimiser le nombre de duplications (Hallett and La-

gergren, 2000), avec dans le cas d'arbres équivalents le choix de l'arbre le moins profond (RIO (Zmasek and Eddy, 2002)). Enfin, l'enracinement peut être fait grâce aux gènes d'une espèce appartenant à un autre groupe taxonomique *outgroup* (phylomeDB (Huerta-Cepas et al., 2007)). Cependant, il est souvent difficile de choisir de manière automatique une séquence *outgroup* d'une espèce à la fois suffisamment éloignée pour avoir une position externe dans l'arbre et suffisamment proche pour s'aligner avec les séquences du groupe sans trop induire de gaps (régions non traitées pour l'arbre). De plus, dans le cas d'études sur l'ensemble du vivant le choix d'un *outgroup* devient impossible.

La méthode **COCO-CL** n'utilise pas l'arbre des gènes à proprement parler, mais analyse la matrice de corrélation entre les séquences. Elle effectue un *clustering* hiérarchique de la matrice de similarité des séquences, menant à des sous-groupes d'orthologues imposés par des événements de spéciation. Cette méthode ne tient pas compte de l'arbre des espèces, mais trouve des groupes hiérarchiques.

Récemment, la méthode **GETHOGs** (*Graph-based Efficient Technique for Hierarchical Orthologous Groups*) (Altenhoff et al., 2013) a été proposée. Il s'agit d'un algorithme permettant d'inférer des groupes d'orthologues hiérarchiques à partir d'un graphe d'orthologues. Elle utilise une correspondance formelle entre le graphe et la hiérarchie du groupe.

Enfin, certaines méthodes n'utilisent l'arbre des espèces que pour une analyse manuelle des familles de gènes. C'est le cas par exemple de la méthode **PANTHER** (*Protein ANalysis THrough Evolutionary Relationships*) (Thomas et al., 2003).

5.2.3 Méthodes n'utilisant que l'arbre des espèces

La méthode **Hieranoid** (Schreiber and Sonnhammer, 2013) est une variation de la méthode Inparanoid. La méthode multi-espèces Hiéranoid présente l'ori-

ginalité de comparer les génomes en fonction de la taxonomie. L'arbre est parcouru des feuilles vers la racine. Chaque nœud correspond à l'application d'Inparanoid. Si plusieurs séquences forment d'ores et déjà un groupe, ce groupe sera comparé *via* l'utilisation de son profil HMM. Les groupes grossissent donc de manière itérative en suivant le parcours de l'arbre. Les séquences ajoutées à l'étape n seront prises en compte pour la création du profil HMM du groupe à l'étape n+1.

5.3 Points forts et points faibles des deux types de méthodes

D'un point de vue pratique, les méthodes basées sur les graphes sont plus rapides en temps de calcul que les méthodes basées sur les arbres. Cependant, elles pèchent généralement dans la construction de groupes hiérarchiques, un groupe étant créé pour un ensemble d'espèces donné, il est difficilement adaptable à un sous ensemble d'espèces. Les méthodes EggNOG et OrthoDB parent à ce problème en intégrant l'arbre des espèces *via* la répétition de la recherche de groupes d'orthologues à différents niveaux taxonomiques.

Les méthodes basées sur les arbres tiennent compte, pour la création de l'arbre des espèces, de l'alignement multiple des séquences. Ainsi, elles se basent sur les régions les plus conservées. Cela permet une source d'information supplémentaire par rapport à la comparaison des séquences deux à deux. L'arbre de réconciliation obtenu permet d'avoir plus d'informations que les paires et les groupes d'orthologues *via* la spécification de la suite des événements de spéciation et de duplication. Cependant, ces méthodes sont sensibles au bruit ainsi qu'à des artefacts de branches longues et d'attractions des branches courtes. Tenir compte de l'alignement multiple des séquences est à la fois un point fort et un point faible. Il induit une sensibilité à la qualité de l'alignement multiple qui peut être un problème dans le cas des pro-

téines multi-domaines. De plus, l'absence de traitement des gaps de l'alignement multiple peut induire un traitement moins efficace des insertions et délétions dans les séquences.

Généralement, les méthodes basées sur les arbres travaillent à partir de groupes d'homologues pré-construits. Ainsi, elles utilisent une méthode basée sur les graphes en pré-traitement (paramètres peu stringents). Nous noterons de plus que les méthodes basées sur les arbres utilisent un modèle d'évolution considérant le nombre de duplications et de pertes de gènes faible par rapport au nombre de spéciations et ne prennent généralement pas en compte les transferts horizontaux. Or, on sait aujourd'hui que ces deux hypothèses de départ ne sont pas respectées dans le cas des procaryotes et des virus (Treangen and Rocha, 2011). Le cas de transferts horizontaux, même s'il a été moins décrit, a été décrit entre eucaryotes et bactéries ou virus, mais aussi entre deux eucaryotes (Nedelcu et al., 2008).

Il existe différentes méthodes de comparaison des résultats obtenus par les méthodes de prédictions de groupes d'orthologues. Il a par exemple été analysé au sein des groupes la conservation du niveau d'expression, la présence des mêmes domaines protéiques, les informations d'interactions protéines-protéines (Hulsen et al., 2006) et la conservation de la synténie. Plus récemment, OrthoBENCH (Trachana et al., 2011), a proposé d'évaluer les méthodes en fonction de groupes pré-établis. Lors de sa publication, OrthoBENCH avait fait intervenir 12 génomes de métazoaires et proposait 70 familles manuellement analysées dans le but d'obtenir des groupes surs. Il s'agit de groupes d'orthologies définis au niveau métazoaire. Ils incluent de ce fait des in-paralogues. Ils ont été construits *via* la sélection des protéines obtenues par l'intersection des résultats de cinq méthodes, l'ajout de séquences similaires par comparaison avec leurs profils HMMs et par l'analyse manuelle de l'arbre obtenu après l'alignement de ces séquences. Parmi ces groupes, 35 sont connus pour être difficiles à prédire (évolution rapide), 5 forment des familles bien alignées (évolution

plus lente du groupe) et 30 ont été choisies aléatoirement. Ces groupes ont été utilisés pour comparer cinq méthodologies (OMA (Altenhoff et al., 2015), EggNog (Powell et al., 2014), OrthoMCL (Li et al., 2003), OrthoDB (Waterhouse et al., 2013) et TreeFam (Li et al., 2006)). À présent, une nouvelle version d'orthobench (Trachana et al., 2014) propose également des groupes construits à partir de génomes de bactéries. L'*Orthology Benchmark Service* quant à lui, combine plusieurs méthodes de comparaisons tels que la comparaison de la similarité de fonction (Hulsen et al., 2006) ou la comparaison de l'arbre des gènes avec l'arbre des espèces (Altenhoff and Dessimoz, 2009). Enfin, il a été proposé d'utiliser un jeu de données simulées (Dalquen et al., 2013). Ces méthodes de benchmark peuvent être basées sur les mêmes critères que ceux utilisés pour la définition des orthologues. C'est le cas par exemple lors de l'évaluation de méthodes utilisant l'arbre des espèces sur le nombre de différences entre l'arbre des séquences et l'arbre des espèces. Ces benchmarks peuvent donc être biaisés en faveur d'un des types de méthodes.

Ils ont cependant permis d'établir que :

- La qualité des génomes ainsi que la couverture phylogénétique des espèces comparées influencent grandement la qualité des groupes obtenus par l'ensemble des méthodes (Trachana et al., 2011).

- Il existe un compromis entre la spécificité et la sensibilité. Ainsi, les méthodes basées sur les graphes sont généralement plus sensibles, mais moins spécifiques que les méthodes basées sur les arbres (Chen et al., 2007).

- Les événements de duplication/perte de gènes sont globalement bien gérés par l'ensemble des méthodes (Dalquen et al., 2013).

- La prédiction de transferts horizontaux reste un challenge pour l'ensemble des méthodologies (Dalquen et al., 2013).

- Analysées sur des données simulées, les méthodes basées sur les scores d'alignement (BRH, Inparanoid, OrthoMCL, OrthoInspector et QuartetS) sont plus impactées par les erreurs de séquençage et d'annotation structurale que les méthodes basées sur les distances (OMA et RSD). (Dalquen et al., 2013). Cela peut s'expliquer par le fait que les scores d'alignement sont plus perturbés par des caractères ambigus que les estimateurs de distances.

- Dans le cas de fort taux de paralogues (après par exemple la duplication complète du génome WGD chez les champignons) toutes les méthodes échouent dans la prédiction de groupes d'orthologues précis (Salichos and Rokas, 2011).

- Si le but est de retrouver un orthologue et pas les in-paralogues alors les méthodes les plus simples (type BRH) sont les plus appropriées (Salichos and Rokas, 2011).

La résolution d'arbre contenant des centaines de séquences est encore aujourd'hui un sujet problématique. Les méthodes de types arbre, car elles apportent une plus grande variété d'informations, sont à conseiller dans le cas de petits nombres de génomes d'animaux ou de plantes (présentant moins de 100 espèces), car les transferts horizontaux et les forts taux de duplication semblent moins présents dans ces espèces. Les méthodes de types graphes sont quant à elle à privilégier dans le cas d'études à large échelle ou dans le cas de la prise en compte de génomes contenant une proportion non négligeable de transferts horizontaux (études courantes chez les procaryotes) (Kristensen et al., 2011).

5.4 Combinaisons de méthodes

5.4.1 MetaPhOrs

MetaPhOrs (Pryszcz et al., 2011a) est une méthode de prédiction basée sur l'utilisation conjointe de sept bases de données contenant des prédictions d'orthologie. Les bases de données utilisées sont les suivantes : PhylomeDB, Ensembl, EggNOG, orthoMCL, COG, Fungal Orthogroups et TreeFam. La base de données MetaPhOrs recouvre 829 génomes complets. La prédiction d'orthologie et de paralogie est basée sur un score de cohérence. Celui-ci est compris entre 0 et 1, plus sa valeur est grande et plus la relation est prédite avec confiance.

Ce score de confiance (CS) est calculé de la manière suivante :

- $CS_{orthologues} = T_{orthologues}/(T_{orthologues} + T_{paralogues})$
- $CS_{paralogues} = T_{paralogues}/(T_{orthologues} + T_{paralogues})$

avec $T_{orthologues}$ le nombre d'arbres confirmant l'orthologie et $T_{paralogues}$ le nombre d'arbres confirmant la paralogie. Si le score de confiance dépasse 0,5 la relation d'orthologie est conservée.

Cette méta-approche ne permet pas la prédiction de couples d'orthologues absents des bases initialement sélectionnées.

5.4.2 FungiPATH, première version de l'algorithme

FungiPATH (Grossetete et al., 2010) fut la première méta-approche développée par l'équipe Bioinformatique Moléculaire de l'IGM (équipe dans laquelle j'ai effectué ma thèse). Son fonctionnement est basé sur l'utilisation des intersections des groupes d'orthologues obtenus avec quatre méthodes existantes (BRH, Inparanoid, OrthoMCL, Phylogeny) (voir la figure 5.5). Ces méthodes sont traitées comme des

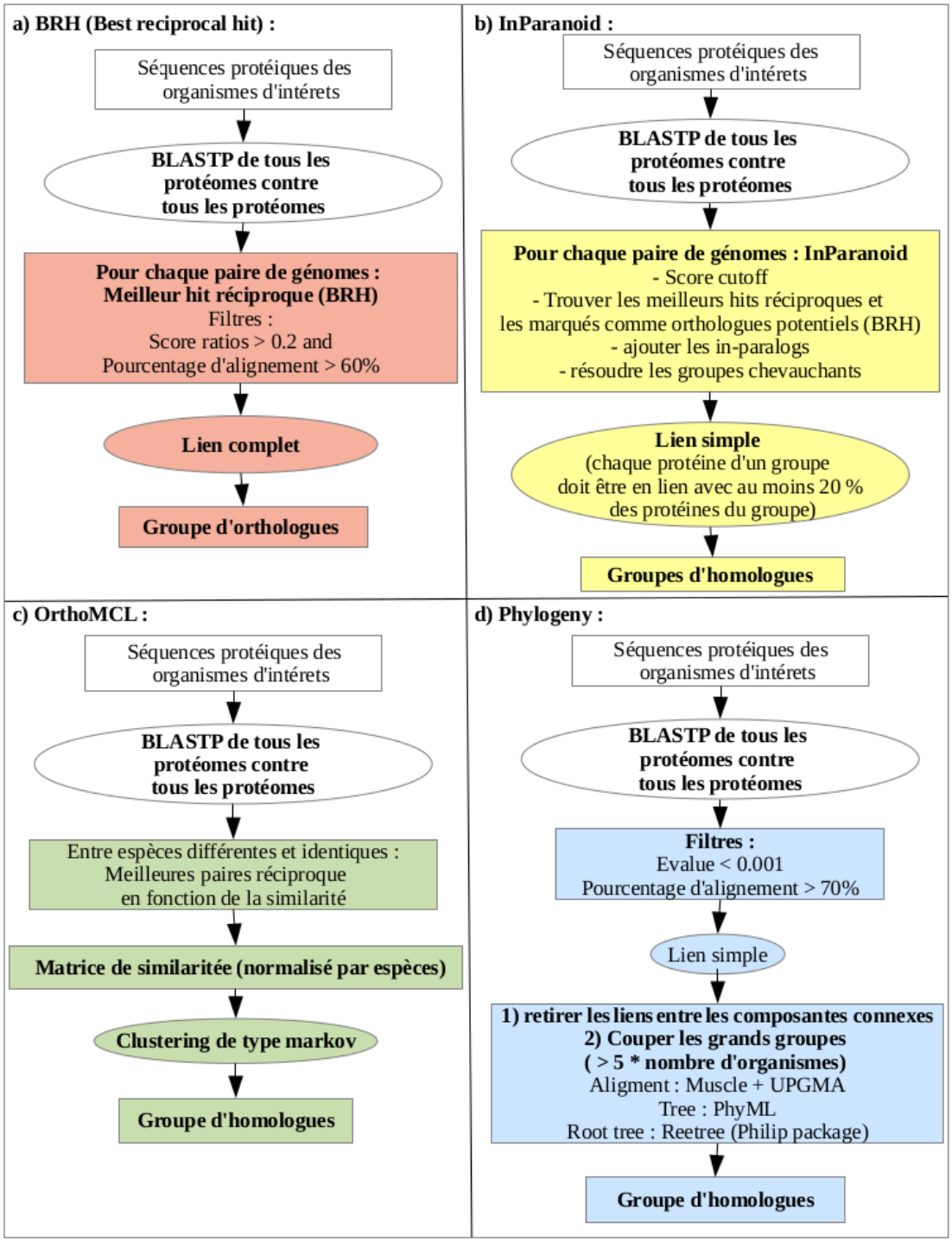


Figure 5.5 – *Présentation des quatre méthodes de détection de groupes d'orthologues utilisées en input de la méthode FungiPath*

méthodes indépendantes. Elle fut développée et optimisée pour l'étude du métabolisme des champignons.

5.4.2.1 Description de la méthode

Le but de la méthode FungiPath est de permettre l'obtention d'annotations pertinentes et homogènes d'un ensemble de protéomes de champignons. Afin de procéder à l'annotation des protéomes, la méthode commence par la détection des groupes d'orthologues chez ces espèces. Pour cela, elle prend en entrée les groupes d'orthologues obtenus avec quatre méthodes initiales de manière à les combiner. Les résultats des quatre méthodes sont filtrés de manière à retirer les groupes constitués uniquement de paralogues (séquences appartenant toutes à la même espèce). FungiPath utilise les intersections des différentes méthodes comme graines de groupes d'orthologues.

Les intersections sont construites de la manière suivante :

- Sélection des intersections des quatre méthodes (taille deux minimum)
- Sélection des intersections de trois méthodes (taille deux minimum) n'impliquant pas de protéines déjà présentes dans les intersections sélectionnées précédemment. Une protéine ne peut pas être présente dans deux intersections. Si le cas se présente, seule l'intersection présentant le plus de séquences est sélectionnée.
- Sélection des intersections de deux méthodes (taille deux minimum) de la même manière que pour les intersections de trois méthodes.

Les intersections de quatre, trois ou deux méthodes de taille deux minimum sont les graines des groupes d'orthologues. Elles sont considérées comme fiable, car consensus des résultats obtenus par au moins deux méthodes. Pour chaque intersection est construit son alignement multiple (Muscle) ainsi que son profil HMM

(Hmmer). L'idée sous-jacente étant que les profils HMM reflètent les propriétés de conservation de séquence des intersections. La compatibilité entre les séquences non assignées à une intersection et les profils HMM des intersections est ensuite évaluée (hmmsearch). Une séquence seule est ajoutée au groupe avec lequel elle match avec la meilleure evalue si cette evalue est inférieure ou égale à 10^{-10}.

Ainsi, un groupe d'orthologues est constitué des séquences présentes dans une intersection ainsi que des séquences ayant eu cette intersection en best-hit lors de la comparaison séquences/profils HMM avec une evalue inférieure à 10^{-10}. La méthode FungiPath permet la prédiction de relations d'orthologie absentes des groupes d'orthologues de départ.

Les groupes d'orthologues contiennent à la fois des orthologues et des inparalogues. Notons cependant que les arbres résultant des groupes ne sont pas analysés, il est possible que des paralogues ayant conservé la même fonction (profils HMM similaires) soient regroupés dans le même groupe.

5.4.2.2 Limites de la méthode :

D'un point de vue théorique, la méthode dite 'FungiPath' présente plusieurs limitations.

- En utilisant l'ensemble des intersections des résultats des méthodes initiales en une unique étape elle ne privilégie pas les intersections de N méthodes par rapport aux intersections de $N - M$ méthodes ($N > M > 0$) alors que les intersections faisant intervenir le plus de méthodes semblent naturellement plus fiables.

- Les profils HMM sont faits à partir des intersections ayant une taille 2 ou plus. Cette taille d'intersection minimum n'a pas été évaluée. Les profils effectués avec des intersections de taille 2 sont certainement peu représentatifs de l'ensemble du groupe et peuvent donc induire un biais dans l'ajout de séquences

à ces groupes.

- Lors de la comparaison séquence / profil seule la e-value est prise en compte, or ce critère à lui seul peut s'avérer insuffisant pour éviter l'ajout de protéines issues de fusion de gènes dans le groupe d'orthologues d'un des deux gènes.

D'un point de vue technique, l'implémentation de la méthode fonctionne uniquement avec les résultats des dites quatre méthodes. Les scripts à lancer sont nombreux et le readMe de lancement conséquent. Les quatre méthodes initiales avaient besoin de Blast pour fonctionner, mais bien que les paramètres soient similaires, les Blast devaient être lancés plusieurs fois afin d'obtenir pour chaque méthode les informations spécifiques requises.

Chapitre 6

Méta-approche de détection de groupes d'orthologues: article

Le méta-approche de détection de groupes d'orthologues a été publiée en 2014 dans le journal *BMC genomics* (Pereira et al., 2014). Ce chapitre contient une reproduction de cet article.

RESEARCH

A meta-approach for improving the prediction and the functional annotation of ortholog groups

Cécile Pereira[1,2,3], Alain Denise[1,2,3,4] and Olivier Lespinet[1,2,3*]

*Correspondence:
olivier.lespinet@igmors.u-psud.fr
[1]Univ Paris-Sud, Institut de Génétique et Microbiologie, UMR8621, 91405 Orsay, F
Full list of author information is available at the end of the article

Abstract

Background: In comparative genomics, orthologs are used to transfer annotation from genes already characterized to newly sequenced genomes. Many methods have been developed for finding orthologs in sets of genomes. However, the application of different methods on the same proteome set can lead to distinct orthology predictions.

Methods: We developed a method based on a meta-approach that is able to combine the results of several methods for orthologous group prediction. The purpose of this method is to produce better quality results by using the overlapping results obtained from several individual orthologous gene prediction procedures. Our method proceeds in two steps. The first aims to construct seeds for groups of orthologous genes; these seeds correspond to the exact overlaps between the results of all or several methods. In the second step, these seed groups are expanded by using HMM profiles.

Results: We evaluated our method on two standard reference benchmarks, OrthoBench and Orthology Benchmark Service. Our method presents a higher level of accurately predicted groups than the individual input methods of orthologous group prediction. Moreover, our method increases the number of annotated orthologous pairs without decreasing the annotation quality compared to twelve state-of-the-art methods.

Conclusions: The meta-approach based method appears to be a reliable procedure for predicting orthologous groups. Since a large number of methods for predicting groups of orthologous genes exist, it is quite conceivable to apply this meta-approach to several combinations of different methods.

Keywords: ortholog; homolog; meta-approach; sequences-profile comparison

Background

Performing an accurate gene/protein functional annotation is one of the crucial steps of any new genome project. It is partly achieved by performing the functional annotation of groups of orthologs.

Orthologs are genes in different species that arose from a common ancestral gene by speciation events [1]. Based on the 'orthology-function conjecture' [2, 3], the orthologs retain the same function and thus can be used for the transfer of functional annotation from experimentally characterized genes to uncharacterized genes [4].

In this article, an ortholog group contains all the genes that evolved by gene duplication since the most ancestral speciation event of a given set of genomes [4]. Thus, ortholog groups include orthologs, co-orthologs and paralogs that evolved by lineage specific duplication after the relevant speciation event (in-paralogs) [5] (see Additional file 1).

The prediction of orthologous genes is a difficult task because of non-uniform evolutionary rates, extensive gene duplication, gene loss and horizontal gene transfer [6]. Over the last decades, a large number of methods and tools have been developed to perform orthologous gene prediction, and nowadays not less than 37 databases offer groups of orthologs [7]. However, the results predicted by these various methods are often uncertain. In particular, users should be aware that the application of different methods on the same proteomes can lead to distinct orthology predictions [6, 8, 9]. Accordingly to these results, it is particularly difficult to know which method or database will be the most appropriate. In addition, we might reasonably question the relevance of biological findings drawn from the orthology prediction obtained by any single method.

Sequence similarity is a good predictor of homology but does not define homolog sequences. Like the similarity used to predict homolog sequences, the genome context could be used to predict toporthologs (orthologous genes that retain their ancestral genomic position). This precision is motivated by the biological significance of genomic context [10] (genes that are near each other are more likely to interact [11] and are possibly coordinately expressed [12]). Because the gene order changes rapidly [13] and can not be use for distant species, we focus on the prediction of ortholog groups, without subdividing this groups into toporthologs and atoporthologs.

Prediction of gene orthology is based on two main approaches, namely tree-based methods and graph-based methods [14].

Tree-based methods are based on a tree-like evolutionary scenario and the evolution of the entire group of homologous genes is performed at the same time. The pairs of orthologs are inferred from the analysis of gene family trees and these methods [15, 16, 17, 18, 19] use the definition of orthology in order to distinct orthologs and paralogs. Gene orthology selection is generally done by tree reconciliation [20] with a reference species tree [17, 18, 19]. However, this last step becomes an issue when horizontal gene transfer plays a major role in the evolution of the organisms [21]. Moreover, tree-based methods are sensitive to artifacts, such as long and short-branch attraction at large or small evolutionary distances [22]. The results also depend of the quality of the species tree, which can contain errors especially at large evolutionary distances.

Graph based methods rely on the assumption that orthologous genes or proteins are more similar than any other gene or protein coming from the same organisms. Thus in graph based methods the orthologs are clustered together according to a similarity measure between the sequences. Several similarity scoring methods are used to cluster the sequences, for example BLAST derived scores [23] or similarity scores computed from Smith-Waterman alignments [24]. These methods [25, 26, 27, 28] are generally much faster than tree-based methods and can deal with a larger number of species. However, they fail to detect differential gene losses [29, 30] and can create mixed groups in the case of complex mixtures of differently-related genes.

As stated above, tree-based and graph-based methods have their advantages and drawbacks. In this work we propose to combine results obtained by several different methods by developing a meta-approach. The purpose is to produce better quality results by using the overlapping results obtained from several individual methods. The rationale behind our approach is that when identical results are found by several methods then they are more likely accurate. This is especially true as the prediction methods use different approaches like tree-based or graph-based methods. However, the overlap between multiple orthology prediction methods may lead to the loss of many true positives orthologs, especially when the number of initial methods is high. To overcome this problem the meta-approach is performed in two steps. An initial step finds seeds for groups of orthologous genes that correspond to the exact

overlaps between all or at least several methods. In a second step we expand these seed groups by using HMM profiles.

Using acknowledged benchmark sets and procedures, we evaluated our meta-approach in relation to two aspects: the quality of our ortholog groups compared to known groups, and the relevance of functional sequence annotation based on our groups. The meta-approach allows to improve both.

Methods

The meta-approach

The entries of the meta-approach are ortholog groups obtained by several input methods. The general outline is as follows. First, we take into account only orthologs that are predicted by several methods, by selecting the intersections of their groups. From the sequences of the intersected groups, we build HMM profiles, possibly adding other sequences to the groups by comparing the sequences to the HMM profiles. Selection of the added sequences is based on the e-value and the percentage of alignment between the sequences and the HMM profiles. This whole process is performed several times, with the number of methods decreasing at each step, as detailed below.

At first, let us justify the meta-approach in a few words. It combines results from several methods, each of them having a given level of sensitivity and specificity. The first stage is stringent (specific), and tends to generate small orthologous groups, because each group is the intersection of the groups obtained by several methods. Recalling that our main objective at the end is annotation, what is important is not to have the largest possible groups, but to ensure that the genes that are in the same group will share the same function. From these small groups, which we call *intermediate groups*, HMM profiles are built. Proteins which are not in any intermediate group are called *unassigned proteins*. Each unassigned protein is compared to the profile HMM of each intermediate group and can be added to a group if the results of comparison satisfy conditions on the e-value and the minimum length of the alignment. Using the HMM profiles aims to improve the sensitivity of the results. Moreover, because the HMM profiles are made from several strongly selected protein sequences, we expect this step still to have a good specificity.

We present below the algorithm in more detail (see also Fig 1).

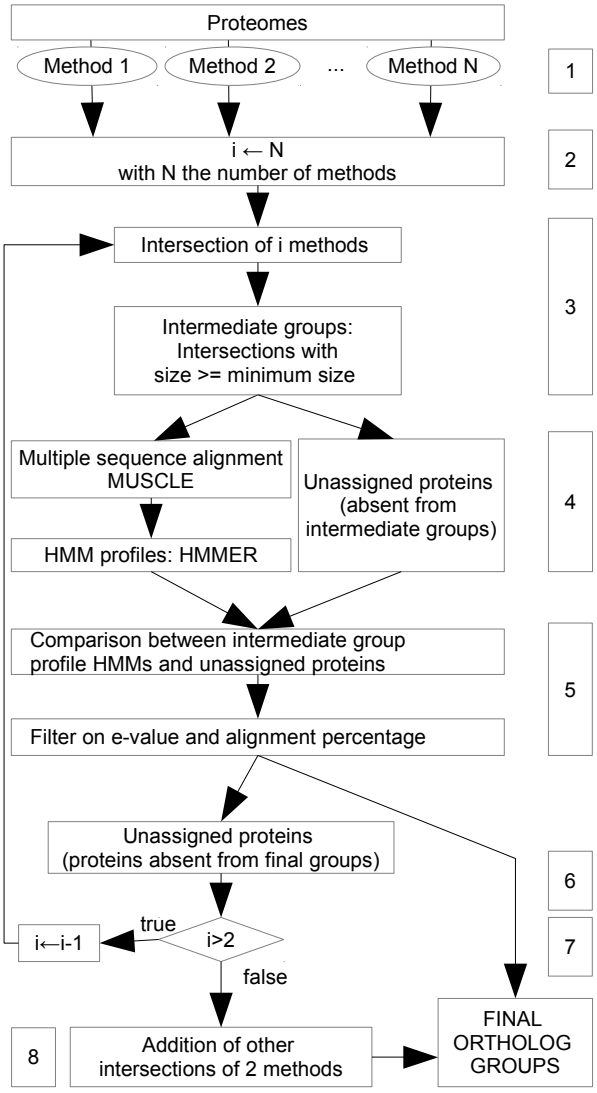


Figure 1 Overview of the meta-approach. The numbers in the boxes refer to the consecutive steps which are detailed in the text.

1 Collect ortholog groups from N input methods ($N \geq 2$)

2 set $i = N$.

3 Compute all sets of proteins that are intersections of groups of i methods: that is, two proteins are in the same set if and only if they are in the same group in i methods. Additionally, sets are selected according to their size: a set is selected only if its size is higher or equal to a given threshold (*minimum size* equals 4 as default). The selected sets form the intermediate groups. The proteins that do not belong to any intermediate group are called *unassigned proteins*. A protein cannot be in several intermediate groups. If this is the case, the largest intersection is kept (this occurs only when $i < N$). If there is some ambiguity for one protein (two distinct available groups of the same size) one of them is retained randomly.

4 For each intermediate group, a multiple alignment is generated with MUSCLE [31]. From each alignment, a profile HMM is computed using HMMER [32].

5 Each unassigned protein sequence is compared to each HMM profile. An unassigned protein is added to an intermediate group if: (i) the e-value of the comparison is lower than a given threshold (default $10E-10$) and (ii) the length of the alignment is above a given ratio compared to the lengths of the sequence and of the profile (default 40%). An unassigned protein can be added to one intermediate group at most. If several HMMs satisfy the thresholds for the same unassigned protein, the lower e-value is retained, then the higher length ratio if necessary.

6 The groups obtained after the previous step are kept aside. This means that the proteins contained in these groups are not used for the next steps. They will be final ortholog groups.

7 If $i > 2$ and if there still some unassigned proteins, then $i \leftarrow i-1$ and GOTO step 3.

8 Otherwise ($i = 2$), the loop stops. There can remain intersection groups that have not been selected as intermediate groups because their cardinality is smaller than the *minimum size*. These groups are added to the final ortholog groups (note that these are necessarily results of intersections of two methods only.)

The values of the three parameters (e-value, minimum length of the alignment and minimum intersection size) were determined by comparing results obtained with different parameter combinations from the same data set (see Additional file 2).

Sotware availability

The MARIO software which implements the meta-approach is freely available at http://bim.igmors.u-psud.fr/mario/.

The selected input methods

The meta-approach was performed by using the results of four methods (BRH [33], Inparanoid [26], OrthoMCL [25] and Phylogeny [34]). The three graph based methods that we selected (BRH, Inparanoid, OrthoMCL) present distinct approaches for predicting ortholog pairs and then for producing groups. They are among the most representative of graph-based methods. A method developed previously in our laboratory called 'Phylogeny' was used as a representative of tree-based methods. All these methods have been implemented in stand-alone programs.

The initial Best Reciprocal Hit (BRH) method [33] was modified by taking into account the sequence alignment length as well as the alignment score ratio between query and subject sequences. The score ratio is the ratio of the raw BLAST score of the alignment and the raw score of each sequence against itself. All pairs of best reciprocal hits *i.e.* were both filters are above the threshold values are considered as orthologs. Pairs of orthologs are clustered by identifying fully connected orthologous groups: each protein of any given ortholog group has an orthology relationship with every other protein in the group (in our case, searching for such cliques is computationally tractable because, in the BRH method, each group presents up to one protein per species). Inparanoid [26] was used with default parameters. This method predicts pairs of orthologs and inparalogs. The pairs are clustered into groups in such a way that each protein of any group is linked to at least 20% of the other proteins of the group. OrthoMCL [25] uses the percentage match length to obtain pairs of orthologous proteins. The method clusters the pairs into groups by using the MCL program [35]. OrthoMCL was used with default parameters. Phylogeny [34] is based on the phylogenetic analysis of homologous genes. No species tree is required. Homologs detected by BLAST are grouped transitively. Homologous sequences are aligned using the MUSCLE program. These multiple alignments are

processed with a maximum likelihood approach to reconstruct the phylogeny of the corresponding family, using the PhyML software. Group trees are rooted by using the program Retree from the Phylip package [36]. The analysis of the rooted tree allows to identify duplication and speciation events and to distinguish orthologs and paralogs.

Evaluation

In order to evaluate our meta-approach, we checked its consistency according to the ability to predict ortholog groups, and the quality of protein functional annotation within an ortholog group. We used two benchmarks: OrthoBench and the Orthology Benchmark Service. The values of the parameters used on both benchmark tests for the meta-approach were the same, as stated above: minimum e-value $10E-10$, minimum alignment length of 40%, minimum intersection size equal to four.

Evaluation on 70 reference ortholog groups

Taking orthoBENCH [37] as a reference benchmark, we compared the results of the four initial methods, and those obtained by the meta-approach, to the reference ortholog groups (RefOGs). The orthoBENCH dataset involves 1519 proteins from 12 metazoan species divided into 70 manually curated ortholog groups. For our analysis, we downloaded the proteome version of Ensembl 72 [38]. As orthoBENCH is based on Ensembl 62, the proteins removed or added between the versions 60 and 72 of Ensembl were not taken into account. Moreover, if a gene has splice variants, When comparing the groups produced by the meta-approach or the individual methods with those of orthoBench, two types of errors were defined: group fissions (proteins of a RefOG are in two or more ortholog groups), and group fusions (more than 3 proteins have been added to a RefOG) [37].

Functional annotation conservation test

The Orthology Benchmark Service is a recent web server (http://orthology.benchmarkservice.org/) allowing us to compare methods of orthologous gene prediction. This is based on a common set of 66 species (2011 quest for orthologs reference dataset) [39, 40]. The benchmark service proposes two types of procedures for evaluating orthologous groups: phylogeny-based definition tests and functional annotation conservation test. In the Phylogeny based tests, orthologous groups are defined in such a way

that every pair of genes in the group is either orthologous or in-paralogous with respect to the last speciation event in their clade. However, we refer to a different and more recent definition of ortholog groups [4]. Thus this test is not relevant for our purpose (see Additional file 1 for further details).

The web server proposes also evaluation procedures for measuring the homogeneity of the functional annotation of the pairs of orthologs [7]. For each pair, if both proteins are annotated, the similarity of the annotation is computed with the Schlicker similarity [41]. This measure allows partial matches, resulting in a robust similarity score for the comparison of gene products with incomplete annotation or for the comparison of multi-functional proteins. This score ranges between 0 and 1, from low to high functional similarity. We computed this measure for Enzyme Commission (EC) numbers [42] and for Gene Ontology (GO) terms [43]. For GO terms, only annotations with experimental support (EXP, IDA, IPI, IMP, IGI and IEP) were considered.

Results and discussion

At first we briefly present the results of the four initial methods and of the meta-approach on the orthoBENCH dataset. Then we compare the meta-approach to twelve other state-of-the-art methods from the functional similarity point of view.

Comparing the results of the four initial methods with those of the meta-approach

The results obtained for each of the four initial methods were compared with those obtained by the meta-approach (Table 1 and see Additional file 3) on the orthoBENCH dataset.

First we observe that all the methods produce different numbers of ortholog groups (ranging from 14 771 for the meta-approach to 25 384 for BRH), and, in addition predict orthology relationships for different numbers of proteins. In order to measure the similarity between the groups obtained by each method, we computed the Jaccard coefficient by dividing the number of ortholog pairs in common between two methods by the total number of pairs of orthologs ($|A \cap B|/|A \cup B|$). The Jaccard coefficient value is expected to be between 0 (no ortholog pairs in common) and 1 (all couples are identical). In our case, all the values range from 0.164 to 0.541. This means that all the methods individually produce rather different results. The Jaccard coefficient values between the meta-approach and any of the input meth-

ods are even lower (lower than 0.156). In other words, none of the selected methods alone can explain the result of the meta-approach.

Table 1 Comparison on OrthoBENCH [37], Jaccard similarity coefficient. Abbreviations : 'Meta' refers to Meta-approach, 'Phy.' to phylogeny, 'Inp.' to inparanoid and 'Ort.' to orthoMCL.

		BRH	Inp.	Ort.	Phy.	Meta
Jaccard	BRH		0.541	0.172	0.389	0.060
similarity	Inp.			0.248	0.340	0.093
coefficient	Ort.				0.164	0.156
	Phy.					0.079
#Proteins		140561	163850	155982	124206	187902

Quality of ortholog groups

Among the four input methods, BRH and Inparanoid give the highest level of accurately predicted groups (groups without fusion or fission events) (Figure 2.A). BRH presents the highest number of fissions and the smallest number of fusions (Figure 2.B). Inparanoid allows the detection of in-paralogs between each pair of proteomes and thus the number of fusions is higher than with BRH and Phylogeny. The Phylogeny approach presents the smallest number of groups impacted by fusions or fissions. The OrthoMCL method presents groups largely impacted by fusion events compared with the other three methods. The larger number of fusions is associated to a lower number of fissions. This result on orthoMCL is consistent with the results obtained by Dalquen *et al* [6] on a dataset of mammalian genomes. As for the meta-approach, it presents the lowest percentage of groups affected by fission or fusion events (Fig 2.C). It also allows an increase of 73.7% in the number of accurately predicted groups compared to the highest result obtained with the four initial methods (Fig 2.A). At the same time, the meta-approach presents the lowest number of fissions and a number of fusions lower than three of the initial methods alone (Fig 2.B). This demonstrate that the meta-approach improves the results obtained with any of the initial methods, either graph-based or tree-based.

Functional similarity performance comparison

We compared twelve methods including all those available in the orthology benchmark service and the four selected input methods for the analysis of the reference proteomes [40]. Additionally, in order to evaluate the impact of using HMM on a single method, we applied the profile HMM procedure (steps 4, 5 and 8) to the

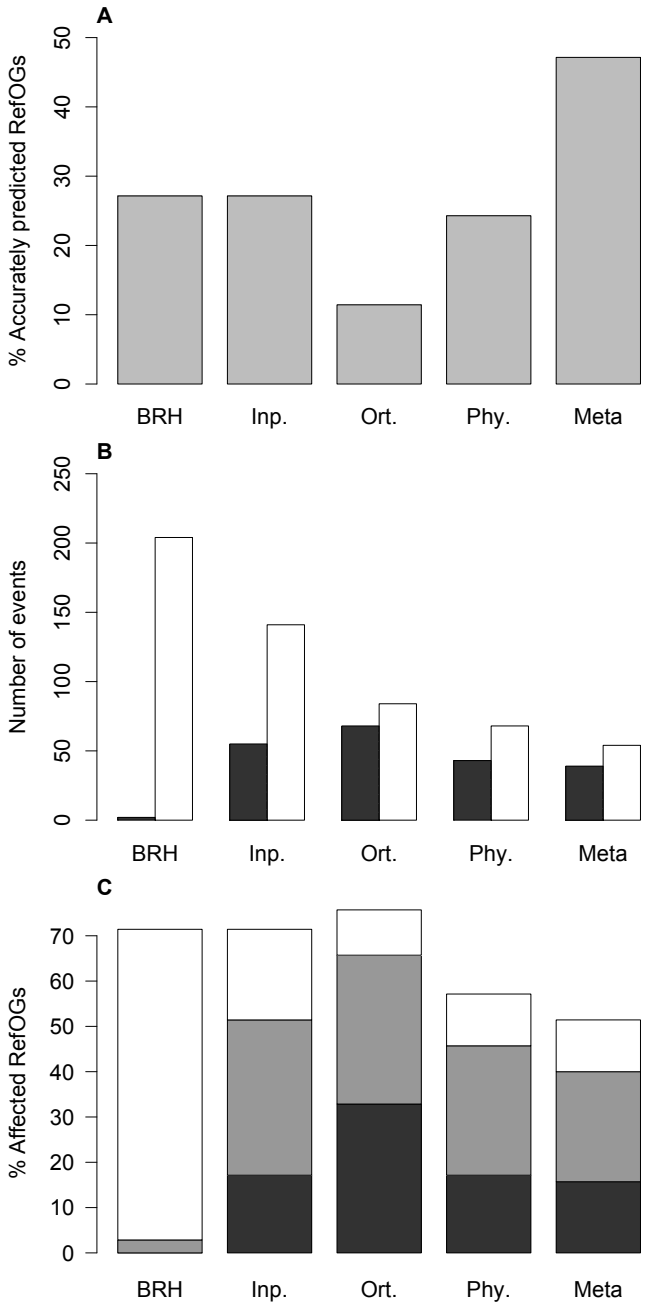


Figure 2 Comparison of the predicted ortholog groups quality (benchmark OrthoBENCH) (A) Percentage of accurately predicted RefOGs (groups predicted without fusion or fission events), (B) Number of fusions (in dark gray) or fissions (in white), (C) Percentage of RefOGs affected by a fusion event (in dark gray), by a fission event (in white) or by the booth (in light gray). A fusion of groups corresponds to the addition of more than 3 erroneously assigned genes to a RefOG. Fissions correspond to a RefOG split in several groups: n group gives $n-1$ fissions. Abbreviations: 'Meta' refers to Meta-approach, 'BRH' to BRH [33], 'Phy.' to Phylogeny [34], 'Inp.' to Inparanoid [26] and 'Ort.' to orthoMCL [25].

BRH groups. For the meta-approach, we used the same values of parameters as those used for the orthoBENCH analyses.

Enzyme classification conservation test

The Pearson correlation test was performed with and without the results of the meta-approach in order to determine the relationship between the number of annotated orthologs and the average Schlicker similarity obtained with the EC number annotations. The results of the metaPhors [18] method stored on the orthology benchmark service website was not available for all the species, therefore this approach was not used for the calculation of correlation. The Pearson correlation is significant whether we use results of the meta-approach or not. The Pearson correlation equals -0.971 (p-value $7.436E{-}8$) using the meta-approach and, -0.964 (p-value $8.573E{-}7$) without the meta-approach (negative correlation hypothesis). This means that increasing the number of ortholog relations is correlated with a decrease in the average Schlicker similarity (Figure 3.A). All methods present a percentage of Schlicker similarity higher than 90%, revealing that all methods succeed in predicting pairs of enzymes with a similar function. Finally, the meta-approach also finds the largest number of ortholog relationships.

Gene ontology conservation test

The Pearson correlation test was performed without taking into account the meta-approach in order to determine the relationship between the number of annotated orthologs and the average Schlicker similarity obtained on GO terms. The results of the metaPhors [18] method were not used for the same reason as indicated previously. The Pearson coefficient was -0.804 (p-value $1.419E{-}3$ with the negative correlation hypothesis). Thus, as for the EC number similarity, the larger number of ortholog relations is correlated to the decreasing of the average Schlicker similarity. The meta-approach detects an increased number of ortholog relations compared to orther methods (Fig 3.B). The Pearson correlation test was performed on the results of all methods (the meta-approach plus the twelve others) in order to determine if the meta-appraoch presents results that are compatible with the same linear regression curve as obtained with the other methods. The Pearson coefficient is not significant (-0.471 and p-value 0.06121 with the negative correlation hypothesis), showing that the result obtained with the meta-approach is not compatible with the

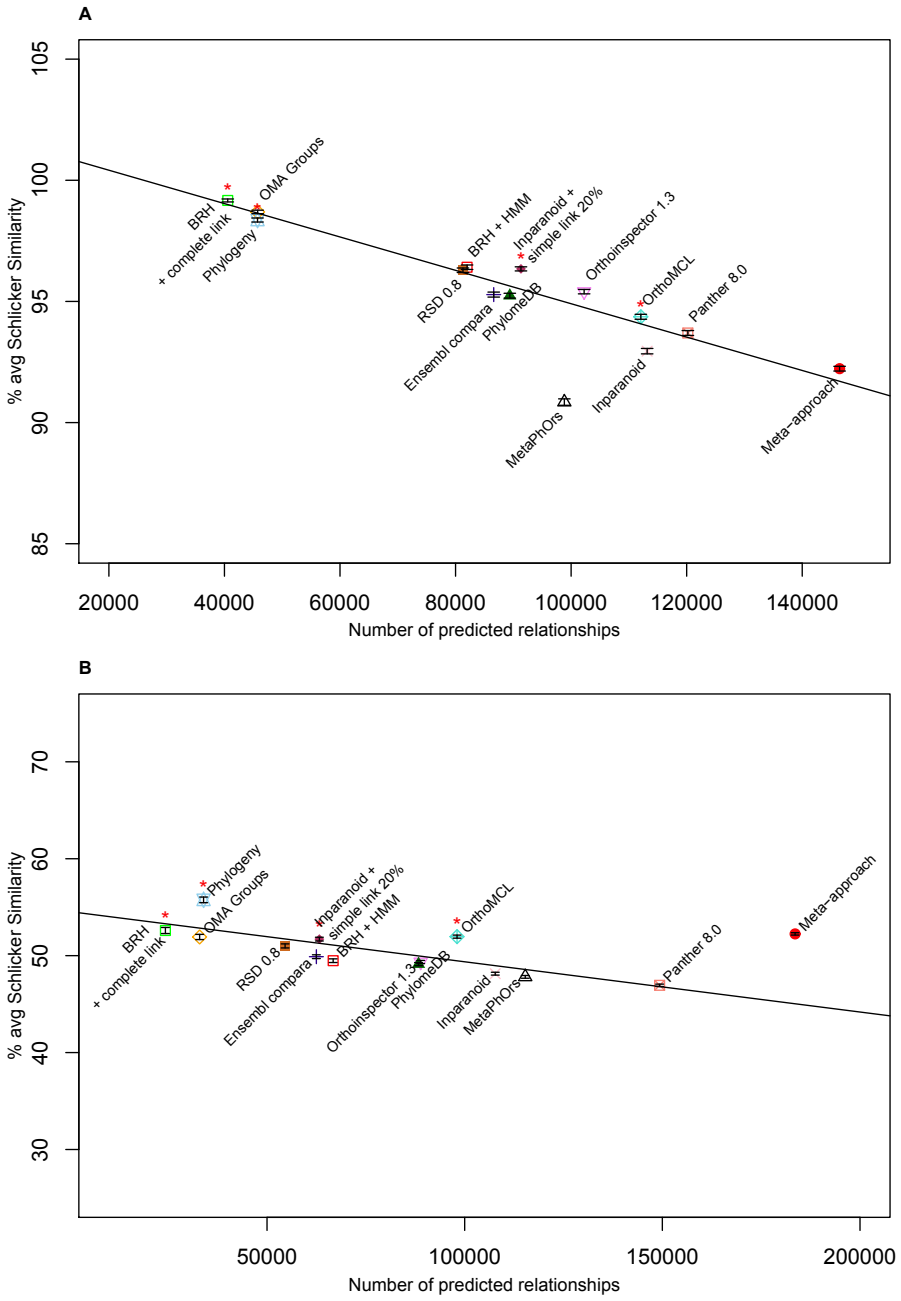


Figure 3 Function-Based Tests (A) Enzyme classification conservation test. The linear regression curve has an intercept value of 101.8 and a regression coefficient of $-6.887E{-}05$. (B) Gene ontology conservation test. The linear regression curve obtained on the GO term annotation has an intercept value of 54,55 and a regression coefficient of $-5.184E{-}05$. The black lines are the linear regression obtained on all methods except the meta-approach, the metaPhors and the BRH plus HMM profiles approach. Error bars for each method are in black. Fourteen methods were compared: ●: Meta-approach. △: metaPhOrs [18]. □: BRH [33] with complete link. □: BRH with complete link plus HMM steps. +: Ensembl compara v2 [38]. ×: Inparanoid pairs [26]. ◇: OMA Groups [27]. ▽: orthoinspector 1.30 [44]. ⊠: PANTHER 8.0 [45]. ▲: phylomeDB [19]. ■: Roundup (RSD 0.8) [28]. ◆: Inparanoid with 20% simple link. ◇: OrthoMCL. ★: Phylogeny. ✳: The four methods used for the meta-approach.

linear regression. Furthermore, the point representing the meta-approach is above the linear regression curve (Fig 3.B), showing that the meta-approach outperforms the other methods on this dataset. Thus, the meta-approach increases both the average Schlicker similarity and the number of ortholog relationships. In addition, the application of the HMM steps on the BRH groups increases the number of annotated ortholog pairs. However and contrary to the meta-approach, the Schlicker similarity decreases when the number of ortholog relation increase, as predicted by the linear regression. Therefore, the combination of the results of several methods is necessary to improve the quality of the final prediction.

Conclusions

The meta-approach appears to be a reliable method of prediction of ortholog groups. Based on the combination of existing methods, the meta-approach finds a consensus of higher quality. Both ortholog group quality and consistence of group annotation have been positively tested. We showed with the orthoBench dataset [37] that, compared to the initial methods, the meta-approach reduce the number of incorrect groups as well as the number of fission and fusion events. Furthermore, the meta-approach presents the largest GO term similarity compared to twelve of the thirteen state-of-the-art methods on the protein reference dataset [40]. Phylogeny's Schlicker similarity is larger than the meta-approach, but Phylogeny predicts many less pairs of annotated orthologs. All other methods present both a smaller Schlicker similarity and an smaller number of pairs of annotated orthologs.

The meta-approach combines the results of several methods in order to obtain specific intersections and adds to these intersections similar sequences (by using profile HMMs). The user has to be well aware that results depend of the selected input methods and on the selected parameters for the HMM profiles.

The meta-approach presented in this article takes the benefits from the particular four methods used here, but as a large number of methods for predicting groups of orthologous genes exist, it would be interesting to apply this meta-approach to different methods or to more methods.

Competing interests

The authors declare that they have no competing interests.

Author's contributions

CP developed and tested the meta-approach. OL and AD supervised the work. All authors (CP, AD, and OL) drafted, read and approved the final manuscript.

Acknowledgements

We thank the eBio platform of University Paris-Sud for resources support. We thank Sandrine Grossetête for her preliminary work on the meta-approach. We thank Anne Lopes for reading this manuscript. We are very indebted to Barry Holland for his invaluable help in improving the English.

Declarations

The publication charges for this article were funded by CNRS-INSERM-INRIA grant PEPS Bio-Math-Info (BMI) 2012-2013.

Author details

[1]Univ Paris-Sud, Institut de Génétique et Microbiologie, UMR8621, 91405 Orsay, F. [2]Univ Paris-Sud, Laboratoire de Recherche en Informatique, UMR8623, 91405 Orsay, F. [3]CNRS, 91405 Orsay, F. [4]INRIA team AMIB, 91120 Palaiseau, F.

References

1. Fitch, W.M.: Distinguishing homologous from analogous proteins. Systematic zoology **19**(2), 99–113 (1970). doi:10.2307/2412448
2. Altenhoff, A.M., Studer, R.a., Robinson-Rechavi, M., Dessimoz, C.: Resolving the ortholog conjecture: orthologs tend to be weakly, but significantly, more similar in function than paralogs. PLoS computational biology **8**(5), 1002514 (2012). doi:10.1371/journal.pcbi.1002514
3. Rogozin, I.B., Managadze, D., Shabalina, S.A., Koonin, E.V.: Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. Genome Biology and Evolution (2014). doi:10.1093/gbe/evu051
4. Gabaldón, T., Koonin, E.V.: Functional and evolutionary implications of gene orthology. Nature reviews. Genetics **14**(5), 360–6 (2013). doi:10.1038/nrg3456
5. Sonnhammer, E.L.L., Koonin, E.V.: Orthology, paralogy and proposed classification for paralog subtypes. Trends in Genetics **18**(12), 619–620 (2002). doi:10.1016/S0168-9525(02)02793-2
6. Dalquen, D.a., Altenhoff, A.M., Gonnet, G.H., Dessimoz, C.: The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. PloS one **8**(2), 56925 (2013). doi:10.1371/journal.pone.0056925
7. Altenhoff, A.M., Dessimoz, C.: Phylogenetic and functional assessment of orthologs inference projects and methods. PLoS computational biology **5**(1), 1000262 (2009). doi:10.1371/journal.pcbi.1000262
8. Chen, F., Mackey, A.J., Vermunt, J.K., Roos, D.S.: Assessing performance of orthology detection strategies applied to eukaryotic genomes. PLoS ONE **2**(4), 12 (2007)
9. Salichos, L., Rokas, A.: Evaluating ortholog prediction algorithms in a yeast model clade. PloS one **6**(4), 18755 (2011). doi:10.1371/journal.pone.0018755
10. Dewey, C.N.: Positional orthology: putting genomic evolutionary relationships into context. Briefings in bioinformatics **12**(5), 401–12 (2011). doi:10.1093/bib/bbr040
11. Huynen, M., Snel, B., Lathe, W., Bork, P.: Exploitation of gene context. Current opinion in structural biology **10**(3), 366–70 (2000)
12. Hurst, L.D., Pál, C., Lercher, M.J.: The evolutionary dynamics of eukaryotic gene order. Nature reviews. Genetics **5**(4), 299–310 (2004). doi:10.1038/nrg1319
13. Wolf, Y.I., Rogozin, I.B., Kondrashov, a.S., Koonin, E.V.: Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. Genome research **11**(3), 356–72 (2001). doi:10.1101/gr.161901
14. Kristensen, D.M., Wolf, Y.I., Mushegian, A.R., Koonin, E.V.: Computational methods for gene orthology inference. Briefings in bioinformatics **12**(5), 379–91 (2011). doi:10.1093/bib/bbr030
15. Storm, C.E.V., Sonnhammer, E.L.L.: Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics (Oxford, England) **18**(1), 92–9 (2002)

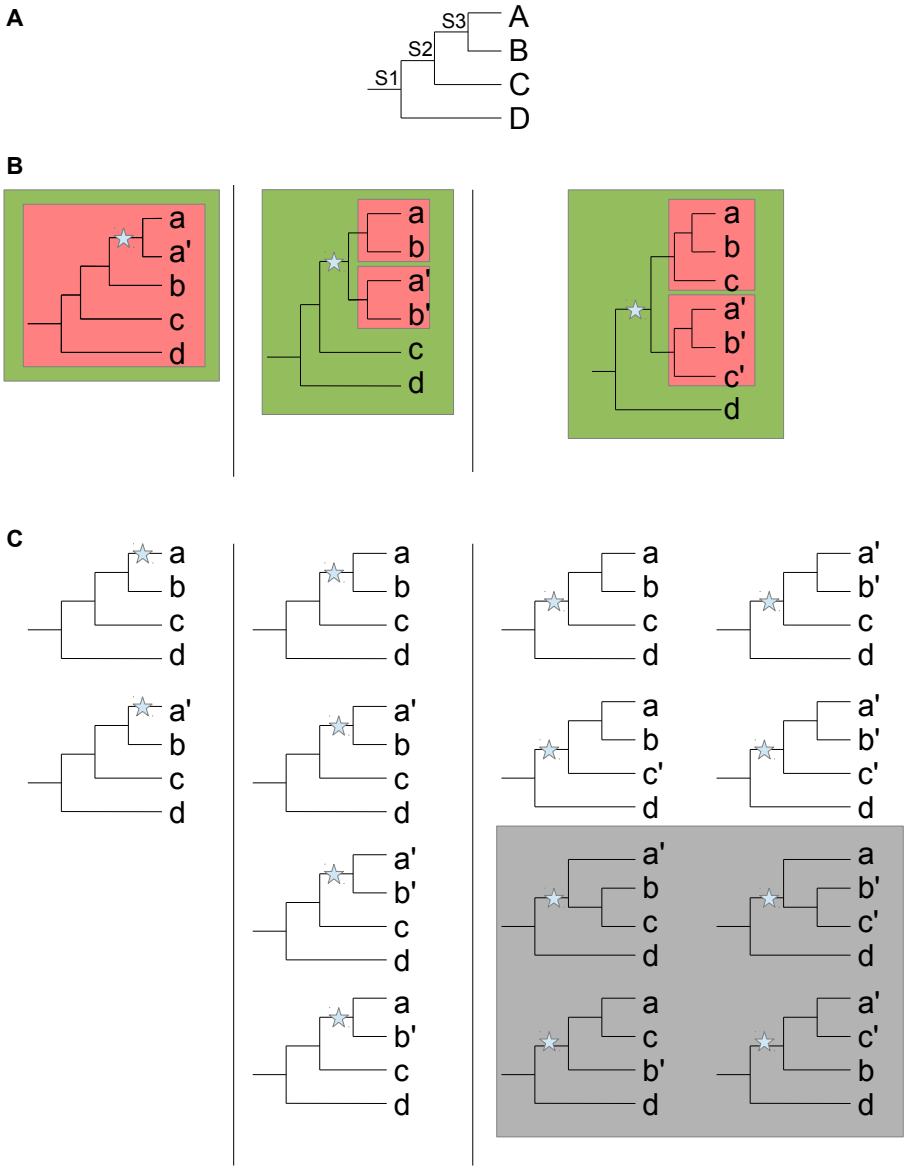
16. Zmasek, C.M., Eddy, S.R.: Rio: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. BMC bioinformatics **19**, 1–19 (2002). doi:10.1186/1471-2105-3-14
17. Ruan, J., Li, H., Chen, Z., Coghlan, A., Coin, L.J.M., Guo, Y., Hériché, J.-K., Hu, Y., Kristiansen, K., Li, R., et al.: Treefam: 2008 update. Nucleic Acids Research **36**(Database issue), 735–740 (2008)
18. Pryszcz, L.P., Huerta-Cepas, J., Gabaldón, T.: Metaphors: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. Nucleic Acids Research **39**(5), 32 (2011)
19. Huerta-Cepas, J., Capella-Gutierrez, S., Pryszcz, L.P., Denisov, I., Kormes, D., Marcet-Houben, M., Gabaldón, T.: Phylomedb v3.0: an expanding repository of genome-wide collections of trees, alignments and phylogeny-based orthology and paralogy predictions. Nucleic Acids Research **39**(Database issue), 556–60 (2011). doi:10.1093/nar/gkq1109
20. Page, R.D., Charleston, M.A.: From gene to organismal phylogeny: reconciled trees and the gene tree/species tree problem. Molecular phylogenetics and evolution **7**(2), 231–240 (1997)
21. Treangen, T.J., Rocha, E.P.C.: Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. PLoS Genetics **7**(1) (2011)
22. O'Connor, T., Sundberg, K., Carroll, H., Clement, M., Snell, Q.: Analysis of long branch extraction and long branch shortening. BMC genomics **11 Suppl 2**, 14 (2010)
23. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of molecular biology **215**(3), 403–410 (1990). doi:10.1016/S0022-2836(05)80360-2
24. Smith, T.F., Waterman, M.S.: Identification of common molecular subsequences. Journal of molecular biology **147**(1), 195–197 (1981). doi:10.1016/0022-2836(81)90087-5
25. Li, L., Stoeckert, C.J., Roos, D.S.: Orthomcl: identification of ortholog groups for eukaryotic genomes. Genome Research **13**(9), 2178–89 (2003). doi:10.1101/gr.1224503
26. O'Brien, K.P., Remm, M., Sonnhammer, E.L.L.: Inparanoid: a comprehensive database of eukaryotic orthologs. Nucleic Acids Research **33**(Database Issue), 476–480 (2005)
27. Altenhoff, A.M., Schneider, A., Gonnet, G.H., Dessimoz, C.: Oma 2011: orthology inference among 1000 complete genomes. Nucleic Acids Research **39**(Database issue), 289–94 (2011). doi:10.1093/nar/gkq1238
28. Deluca, T.F., Wu, I.-H., Pu, J., Monaghan, T., Peshkin, L., Singh, S., Wall, D.P.: Roundup: a multi-genome repository of orthologs and evolutionary distances. Bioinformatics (Oxford, England) **22**(16), 2044–2046 (2006). doi:10.1093/bioinformatics/btl286
29. Koonin, E.V., Wolf, Y.I.: Genomics of bacteria and archaea: the emerging dynamic view of the prokaryotic world. Nucleic Acids Research **36**(21), 6688–6719 (2008)
30. Wolf, Y.I., Novichkov, P.S., Karev, G.P., Koonin, E.V., Lipman, D.J.: The universal distribution of evolutionary rates of genes and distinct characteristics of eukaryotic genes of different apparent ages. Proceedings of the National Academy of Sciences of the United States of America **106**(18), 7273–7280 (2009)
31. Edgar, R.C.: Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Research **32**(5), 1792–1797 (2004)
32. Finn, R.D., Clements, J., Eddy, S.R.: Hmmer web server: interactive sequence similarity searching. Nucleic Acids Research **39**(Web Server issue), 29–37 (2011)
33. Overbeek, R., Fonstein, M., D'Souza, M., Pusch, G.D., Maltsev, N.: The use of gene clusters to infer functional coupling. Proceedings of the National Academy of Sciences of the United States of America **96**(6), 2896–901 (1999)
34. Lemoine, F., Lespinet, O., Labedan, B.: Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. BMC Evolutionary Biology **7**(1), 237 (2007)
35. van Dongen, S.: Mcl - a cluster algorithm for graphs. National Research Institute for Mathematics and Computer Science, in the Netherlands, Amsterdam **Technical**(10), 1–40 (2000)
36. Felsenstein, J.: Phylip (phylogeny inference package) version 3.6. Technical report, Department of Genome Sciences, University of Washington, Seattle (2005)
37. Trachana, K., Larsson, T.a., Powell, S., Chen, W.-H., Doerks, T., Muller, J., Bork, P.: Orthology prediction methods: a quality assessment using curated protein families. BioEssays: news and reviews in molecular, cellular and developmental biology **33**(10), 769–80 (2011). doi:10.1002/bies.201100062
38. Flicek, P., Ahmed, I., Amode, M.R., Barrell, D., Beal, K., Brent, S., Carvalho-Silva, D., Clapham, P., Coates,

G., Fairley, S., et al.: Ensembl 2013. Nucleic Acids Research **41**(Database issue), 48–55 (2013). doi:10.1093/nar/gks1236

39. Gabaldón, T., Dessimoz, C., Huxley-Jones, J., Vilella, A.J., Sonnhammer, E.L., Lewis, S.: Joining forces in the quest for orthologs. Genome biology **10**(9), 403 (2009)
40. Dessimoz, C., Gabaldòn, T., Roos, D.S., Sonnhammer, E.L.L., Herrero, J.: Toward community standards in the quest for orthologs. Bioinformatics (Oxford, England) **28**(6), 900–4 (2012). doi:10.1093/bioinformatics/bts050
41. Schlicker, A., Domingues, F.S., Rahnenführer, J., Lengauer, T.: A new measure for functional similarity of gene products based on gene ontology. BMC bioinformatics **7**, 302 (2006). doi:10.1186/1471-2105-7-302
42. Tipton, K., Boyce, S.: History of the enzyme nomenclature system. Bioinformatics (Oxford, England) **16**(1), 34–40 (2000). doi:10.1093/bioinformatics/16.1.34
43. Ashburner, M., Ball, C.A., Blake, J.A., Botstein, D., Butler, H., Cherry, J.M., Davis, A.P., Dolinski, K., Dwight, S.S., Eppig, J.T., et al.: Gene ontology: tool for the unification of biology. the gene ontology consortium. Nature genetics **25**(1), 25–29 (2000)
44. Linard, B., Thompson, J.D., Poch, O., Lecompte, O.: Orthoinspector: comprehensive orthology analysis and visual exploration. BMC bioinformatics **12**, 11 (2011)
45. Mi, H., Muruganujan, A., Thomas, P.D.: Panther in 2013: modeling the evolution of gene function, and other gene attributes, in the context of phylogenetic trees. Nucleic Acids Research **41**(Database issue), 377–86 (2013). doi:10.1093/nar/gks1118

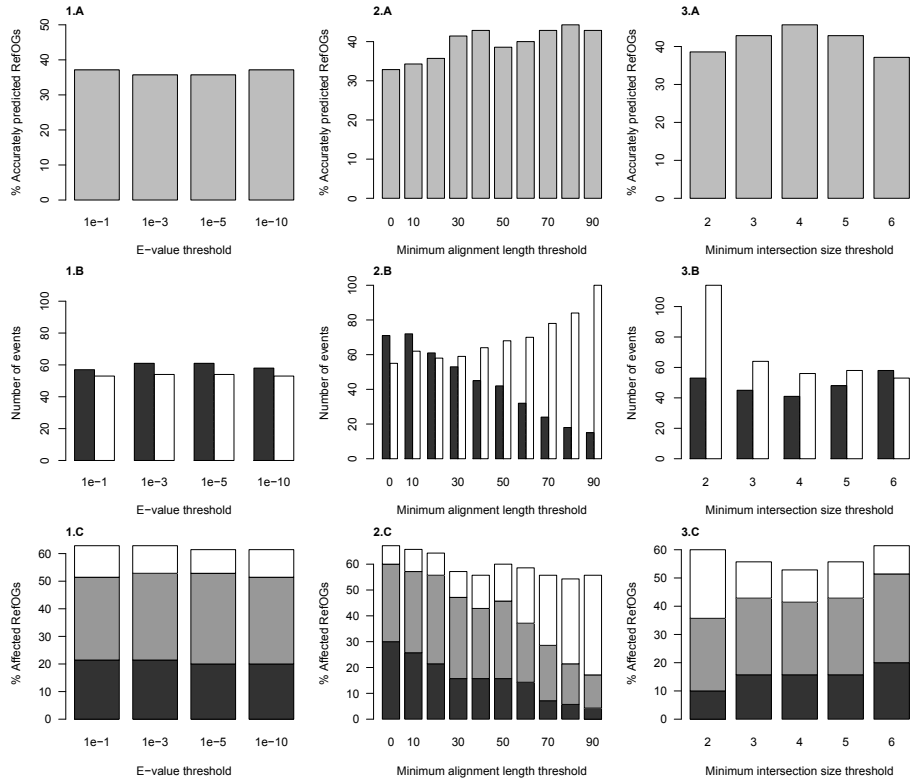
Additional Files

Additional file 1



Comparison of group trees obtained with two definitions of ortholog groups. The phylogeny-based definition tests select ortholog groups in which at least one protein of each of the n species is present. If several proteins are available, one of them is selected randomly, which can lead to differences between the species tree topology and the gene tree topology depending on the ortholog group definition. (A) Example of a specie tree with four species. Each speciation event is presented by a 'S' and a number associated. (B) Possibles associated gene trees and ortholog groups. Green : ortholog group at the S1 level, pink: ortholog group with in-paralogs allowed only if the duplication occurred after the last speciation event (phylogeny tree test definition). Stars: duplication events. (C) Gene trees possibly evaluates with the phylogenetic tree test. This gene trees results from the random selection of one sequence of each species from the ortholog group at the S1 level (green) presented in sub-figure B. In grey, gene tree inducing high Robinson-Foulds distance while the ortholog group is coherent at the S1 level. The larger the number of species used and the more this type of error will occur.

Additional file 2



The impact of each parameter of the meta-approach evaluated on orthoBench. Each column (1,2,3) corresponds to the evaluation of a parameter. Like in the figure 2, the graphs (A) corresponds to the percentage of accurately predicted RefOGs, graphs (B) corresponds to the number of fusions (in dark gray) or fissions (in white) and graphs (C) corresponds to the percentage of RefOGs affected by a fusion event (in dark gray) or by a fission event (in white) or by both types of events (in light gray). The impact of the e-value threshold is observed in the column 1. The two other parameters are fixed (minimum alignment length of 40% and minimum intersection size of six). The variation of the e-value does not involve a large variation in the quality of the predicted groups. The selected value is $1E{-}10$ (highest accurately predicted RefOGs and smallest percentage of groups affected by fission or fusion events). The impact of the minimum alignment length parameter (used in step 5 of the meta-approach) is observed in the column 2. The e-value is fixed to $1E{-}5$ and the minimum intersection size equal three. According to this chart, the increase of the required alignment induces the decrease of the number of fusion and the increase of the number of fission. The highest accurately predicted RefOGs is obtained with the values 40% and 80%. The impact of the minimum size parameter (used in step 3 of the meta-approach) is observed on the column 3. The two other parameters are $1E{-}5$ for the e-value and 40% for the minimum alignment length. Results obtained with intersections of size four or more presents the highest number of accurately predicted groups. However, this evaluation was performed on only 12 species. Thus, the number of ortholog groups containing more than 4 sequences could have induce an under-evaluation of the value of this parameter.

Additional file 3

		BRH	Inparanoid	OrthoMCL	Phylogeny	Meta-approach
# Identical groups	BRH	25384				
	Inparanoid	7543	21342			
	OrthoMCL	5272	7268	17524		
	Phylogeny	4082	3260	2696	17944	
	Meta-approach	3322	4705	4463	2654	14771
# Proteins		140561	163850	155982	124206	187902

Identical groups on OrthoBENCH. Number of identical groups finds on OrthoBENCH for every pair of methods.

Chapitre 7

Comparaison de MARIO avec la première méthode développée pour FungiPath

7.1 Comparaison d'un point de vue méthodologique

Les approches MARIO (Meta-Approche Recherche Intersection Orthologues) et FungiPath v1 ont toutes les deux été développées ces dernières années par notre équipe. L'approche MARIO, la plus récente, a été enrichie des observations et évaluations faites sur les résultats de la méthode FungiPath.

Toutes les deux basées sur l'utilisation de l'intersection de plusieurs méthodes, elles diffèrent cependant sur plusieurs points. L'approche MARIO travaille de manière incrémentale, utilisant en premier lieu l'intersection de l'ensemble des méthodes (N) avant de tester de manière successive l'intersection d'un nombre de méthodes décroissant. Ainsi, en premier lieu sont comparées aux intersections de N méthodes toutes les autres séquences, y compris les séquences appartenant à l'intersection de N-i méthodes (avec i compris entre N-1 et 2 méthodes). Elle favorise

ainsi l'ajout de séquences aux intersections du plus grand nombre de méthodes. Cela induit une diminution du nombre de groupes d'orthologues finaux trouvés, les séquences de certaines intersections de N-i méthodes pouvant être ajoutées précédemment à d'autres groupes. De plus, la méthode MARIO permet la prise en compte des paramètres négligés par la méthode FungiPath, à savoir la taille minimum des intersections et la longueur minimum de l'alignement séquence / profil HMM. À des fins de comparaison, l'approche MARIO a été testée en utilisant les résultats des mêmes méthodes initiales que FungiPath. Cependant, elle est implémentée de manière à pouvoir prendre en compte les résultats d'autant de méthodes que désirées, qu'elles fournissent des groupes ou des paires d'orthologues. Dans le cas de méthodes fournissant des paires d'orthologues, un pré-traitement est appliqué de manière à créer des groupes d'orthologues. Pour le moment la création de groupes se fait *via* l'application de la méthode de *clustering* de type "lien complet". Comme pour BRH, seules les cliques non chevauchantes sont conservées, dans le cas de cliques chevauchantes, la plus grande est conservée.

7.2 Comparaison sur les protéomes de référence

Il existe une communauté de scientifiques travaillant sur les orthologues et se rejoignant en conférence une fois tous les deux ans afin de regrouper leurs efforts pour traiter le problème de recherche d'orthologues. Cette communauté s'est appelée la communauté *Quest for Orthologs*. Lors d'une de ces réunions, il a été mis en exergue qu'il était nécessaire de mettre au point un jeu de données recouvrant l'ensemble de la taxonomie et commun aux différentes méthodes.

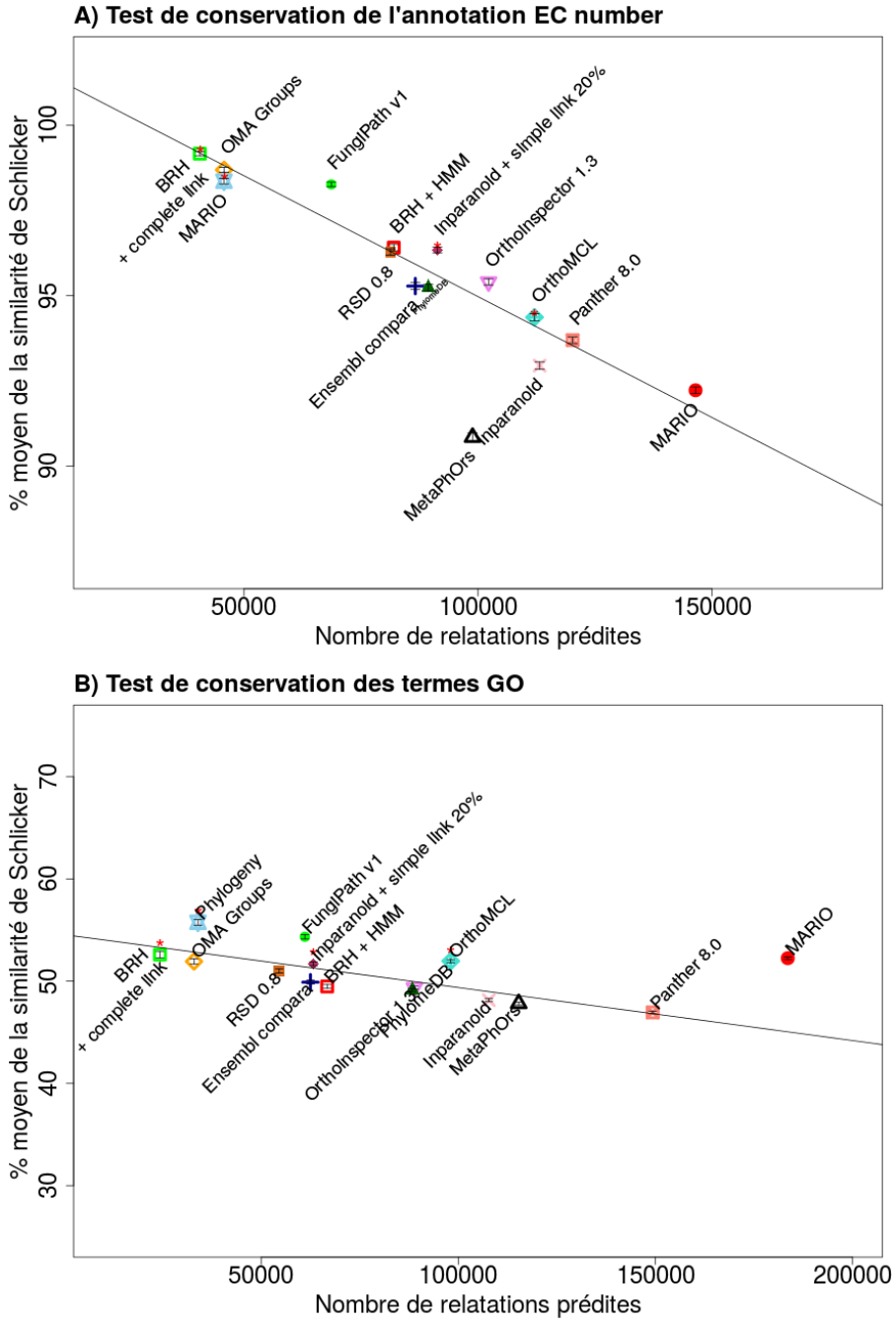


Figure 7.1 – *Comparaison de la similarité de fonction des différentes méthodes. A) Similarité de l'annotation en EC numbers en fonction du nombre de paires dont les deux éléments sont annotés avec au moins un EC number. Attention, l'axe des ordonnées débute à 85% de similarité. B) Similarité de l'annotation en termes GO en fonction du nombre de paires dont les deux éléments sont annotés avec au moins un terme GO. Attention, l'axe des ordonnées est compris entre 30 et 70%.*

	PRD, 66 espèces		178 eucaryotes		12 métazoaires	
Méthode	nb. de groupes	nb. de prot.	nb. de groupes	nb. de prot.	nb. de groupes	nb. de prot.
MARIO	30603	467404	68182	1336810	14771	187902
FungiPath	63231	519757	126752	1489500	-	-
BRH	55749	294145	122393	937067	25383	140566
Inparanoid	43087	387024	67986	1008313	21342	163853
OrthoMCL	38308	433044	71118	1289189	17524	155983
Phylogeny	38004	282140	75240	926124	17944	124208

Table 7.1 – *Pour chaque méthode, nombre de groupes d'orthologues et de protéines impliquées dans ces groupes.*

Protein reference dataset (PRD) (Gabaldon et al., 2009, Dessimoz et al., 2012, page web EBI) est cet ensemble de protéomes. Mis à jour régulièrement, il contient actuellement les protéomes de 147 espèces (avril 2013) dont 120 eucaryotes. Il s'agit d'une proposition d'un ensemble idéal de protéomes, couvrant un large spectre de gammes et de taux d'évolution. Le but de cet ensemble est de refléter les diverses applications courantes de l'orthologie (par exemple la reconstruction phylogénétique ou la prédiction de fonctions). Ces protéomes sont non redondants, un gène correspondant à une protéine. Ils ont été obtenus à partir des bases de données UniProt, Ensembl et Ensembl Genome.

Cet ensemble de protéines permet entre autres la comparaison des résultats obtenus avec différentes méthodes sur les mêmes protéomes initiaux. Le serveur web *orthology benchmark service* y est dédié (version *protein reference dataset* à 66 génomes). Nous l'avons utilisé dans le cadre de notre étude afin de comparer nos résultats à ceux de la communauté et l'utilisons ici pour comparer FungiPath et MARIO (voir figure 7.1 et le tableau 7.1).

La comparaison de la similarité d'annotation des méthodes FungiPath v1

et MARIO sur les données PRD (voir figure 7.1) montre que la méthode MARIO permet d'augmenter le nombre de relations d'orthologie prédites par rapport à la méthode FungiPath. Cette augmentation n'est pas due à la prise en compte de plus de protéines dans les groupes, mais à la formation de plus grands groupes. Elle s'accompagne d'une faible diminution de la similarité d'annotation des paires prédites. Cependant on observe sur la figure 7.1.B que la méthode MARIO présente un point statistiquement au-dessus de la régression linéaire obtenue avec l'ensemble des méthodes de la communauté. Cette régression linéaire reflète le compromis entre la spécificité des relations prédites et la sensibilité des méthodes. On observe ainsi que la méthode MARIO, en se distinguant de manière positive de l'ensemble des méthodes, permet une amélioration des résultats comparés à ceux obtenus avec la méthode FungiPath.

La méthode MARIO permet d'obtenir un plus faible nombre de groupes d'orthologues tout en augmentant leurs tailles comparées aux groupes obtenus avec la méthode FungiPath (voir tableau 7.1). Cette diminution du nombre de groupes est cruciale pour l'étape d'analyse de ces groupes par apprentissage supervisé. Cependant, la qualité de ces groupes peut être mise en doute. Il est possible que l'on observe la fusion de groupes de paralogues au sein d'un même groupe appelé 'orthologue' si les deux groupes ont conservé des fonctions similaires, ce qui se reflète dans des profils HMM similaires.

Chapitre 8

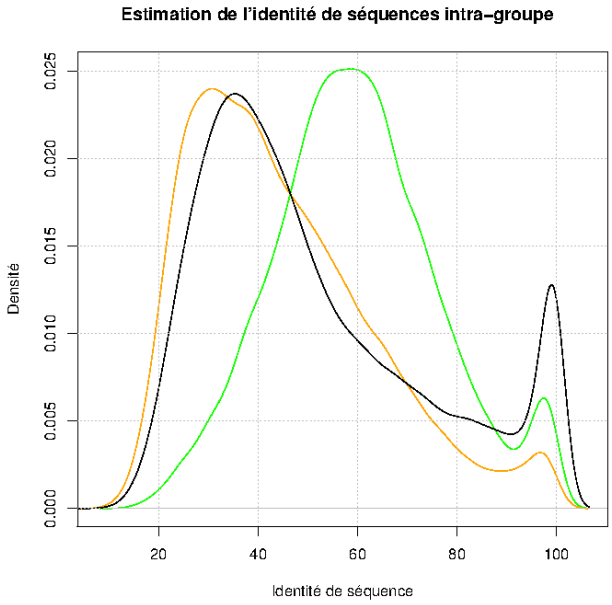
Analyse des résultats de MARIO

8.1 Analyse de l'identité de séquences intra-groupes

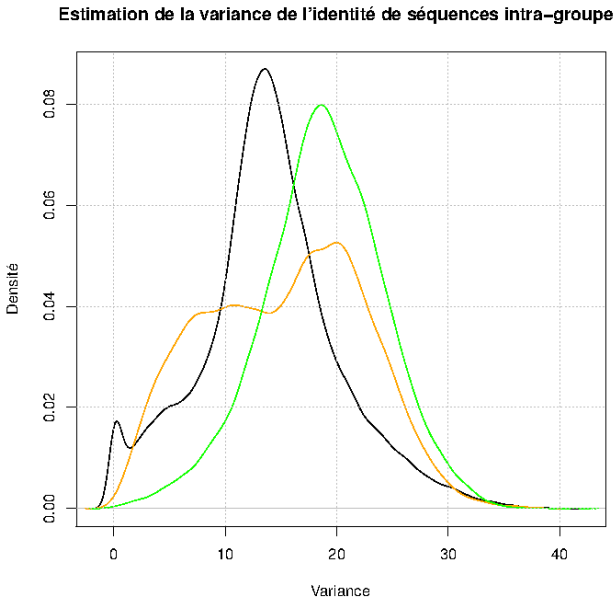
L'identité de séquences moyenne intra-groupe a été calculée pour chaque groupe en faisant la moyenne des identités de paires des séquences du groupe. L'identité de séquence entre deux protéines correspond au nombre de positions identiques divisé par le nombre de positions alignées et multiplié par cent. Elle se calcule à partir d'un alignement multiple. Les séquences étant dans des groupes d'orthologues, nous avons supposé qu'elles étaient relativement proches et utilisons de ce fait MAFFT, une méthode d'alignement linéaire en temps de calcul dans le cas de séquences très similaires (Katoh et al., 2002). Cette supposition a été vérifiée à postériori.

La figure 8.1 présente la densité de l'identité de séquences intra-groupes en fonction des protéomes étudiés. Chaque courbe correspond à l'identité moyenne obtenue sur un jeu de données différent, orange pour PRD (66 espèces), noir pour les 178 eucaryotes (dont 174 champignons) et vert pour les génomes d'orthobench (12 métazoaires).

Appliquée sur l'ensemble du règne ou sur un sous ensemble composé de champignons, la méta-approche permet l'obtention d'orthologues ayant une identité



(a) Densité de l'identité de séquences moyenne intra-groupe



(b) Variance des similarités de séquences par paires au sein de chaque groupe.

Figure 8.1 – *Identité de séquences intra-groupe obtenue sur 178 protéomes d'eucaryotes (dont 174 champignons) en noir (max=35.25%), 66 protéomes de référence (version 2011) en orange (max=30.73%), et sur 12 métazoaires (orthobench) en vert (max=56.8%). La variance n'est calculée que sur les groupes contenant au moins trois protéines. L'identité est calculée en divisant le nombre de positions identiques par le nombre de positions alignées.*

de séquences proche. En effet, les courbes orange et noires présentent un profil proche et un pic d'identité de séquence intra-groupe similaire (entre 30 et 35%). Dans le cas de l'application de la méta-approche aux données de orthoBENCH, l'identité de séquence est plus grande (56.85%). Cette augmentation est cohérente avec l'étude d'un clade couvrant douze organismes relativement peu divergents comparés au groupe des champignons (les deux éléments les plus éloignés sont *Caenorhabditis elegans* et *Rattus norvegicus*).

De manière générale, la similarité de séquence intra-groupe dépend donc de la distance évolutive des espèces étudiées. Sur les courbes représentant le pourcentage d'identité des groupes construits à partir de protéomes de champignons ou de l'ensemble du vivant, on observe une distribution similaire suggérant la présence de groupes présentant majoritairement des séquences appartenant à une des grandes branches du vivant.

On observe également sur la figure 8.1 un pic correspondant à des groupes présentant des pourcentages d'identité moyenne très élevée (aux alentours de 95%). Il s'agit majoritairement de groupes constitués uniquement d'une paire de séquences.

8.2 Analyse de différences obtenues sur les groupes d'orthoBENCH

Dans l'article "A meta-approach for improving the prediction and the functional annotation of ortholog groups" (Pereira et al., 2014), nous avons comparé les groupes d'orthologues que nous avons obtenus avec les groupes d'orthologues vérifiés (70 groupes d'orthoBENCH). Nous avons présenté dans l'article que l'application de la méta-approche permet une amélioration de la prédiction des groupes (moins de fissions et de fusions que sur les groupes fournis en entrée). Cependant, l'analyse de la qualité des groupes accepte un faible nombre d'erreurs. En effet, les auteurs de la

Figure 8.2 – *Arbre phylogénétique obtenu sur le groupe RefOG019. Les séquences considérées sont celles d'orthoBENCH ainsi que celles ajoutées par la méta-approche. Une séquence de Nematostella vectensis (en bleu) a été utilisée en outgroup. La séquence prédite, mais non présente dans le groupe RefOG019 est indiquée en vert. Méthode alignement : muscle, Arbre : PhyML, Bootstrap : 500.*

publication d'orthoBENCH avaient défini la notion de fusion comme l'ajout de plus de trois protéines au groupe d'orthologues et la fission comme la présence de protéine du groupe d'orthologues analysé dans un autre groupe. Il peut ainsi manquer des séquences n'ayant été assimilées à aucun autre groupe. Il peut également y avoir des gènes ajoutés au groupe s'ils sont en nombre inférieur à quatre. Afin d'expliquer ce choix d'analyse au niveau groupe et non pas au niveau gène nous avons choisi de présenter ici quelques cas de groupes présentant ce type d'erreurs.

Le groupe *PHd Finger family* (RefOG019) est un groupe classé par orthoBENCH comme un groupe à évolution lente. Le groupe correspondant prédit par notre méta-approche est issu de l'intersection de 3 méthodes. Aucune séquence n'a été ajoutée à cette intersection à l'étape de comparaison profil HMM / séquences non assignées. Le groupe prédit est donc la pure intersection des résultats obtenus avec 3 des 4 méthodes initiales. Il contient l'ensemble des séquences du groupe orthoBENCH RefOG019 ainsi qu'une séquence supplémentaire FBpp0078894. Le groupe présente une identité de séquence moyenne de 96.48% et une variance de cette identité de 4,06%. La séquence FBpp0078894 présente une identité de séquence avec les autres séquences du groupe allant de 88 à 95%. Elle est donc similaire aux séquences du groupe orthoBENCH RefOG019. L'arbre obtenu avec les séquences du groupe prédit (voir la figure 8.2) présente la séquence ajoutée (entourée de ***) aussi éloignée des séquences du groupe que de la séquence *outgroup* de *Nematostella vectensis*. L'*outgroup* a été choisi comme dans orthoBENCH chez *Nematostella vectensis*. La séquence Y54F10BM.14 est placée par orthoBENCH comme une séquence appartenant à ce groupe, cependant, elle est plus éloignée du reste du groupe que la séquence que nous ajoutons. Il y a donc ici deux hypothèses; la séquence que nous ajoutons a été manquée par orthoBENCH, ou la séquence Y54F10BM.14 ainsi que la séquence que nous ajoutons ne devraient pas faire partie de ce groupe.

Le groupe correspondant aux séquences de la famille Sec13 (RefOG001) est constitué de 11 séquences. Il s'agit de protéines impliquées dans la formation de vésicules du réticulum endoplasmique (Alex J et al., 2013). La prédiction de ce groupe par la méta-approche a abouti à un groupe de 13 séquences. Il est prédit à partir d'une intersection des 4 méthodes composée de 12 séquences ainsi que d'une séquence ajoutée par comparaison du profil HMM de l'intersection avec les séquences non assignées. Les deux séquences présentes dans le groupe prédit mais pas dans le groupe de référence sont des séquences qui ont été prédites comme

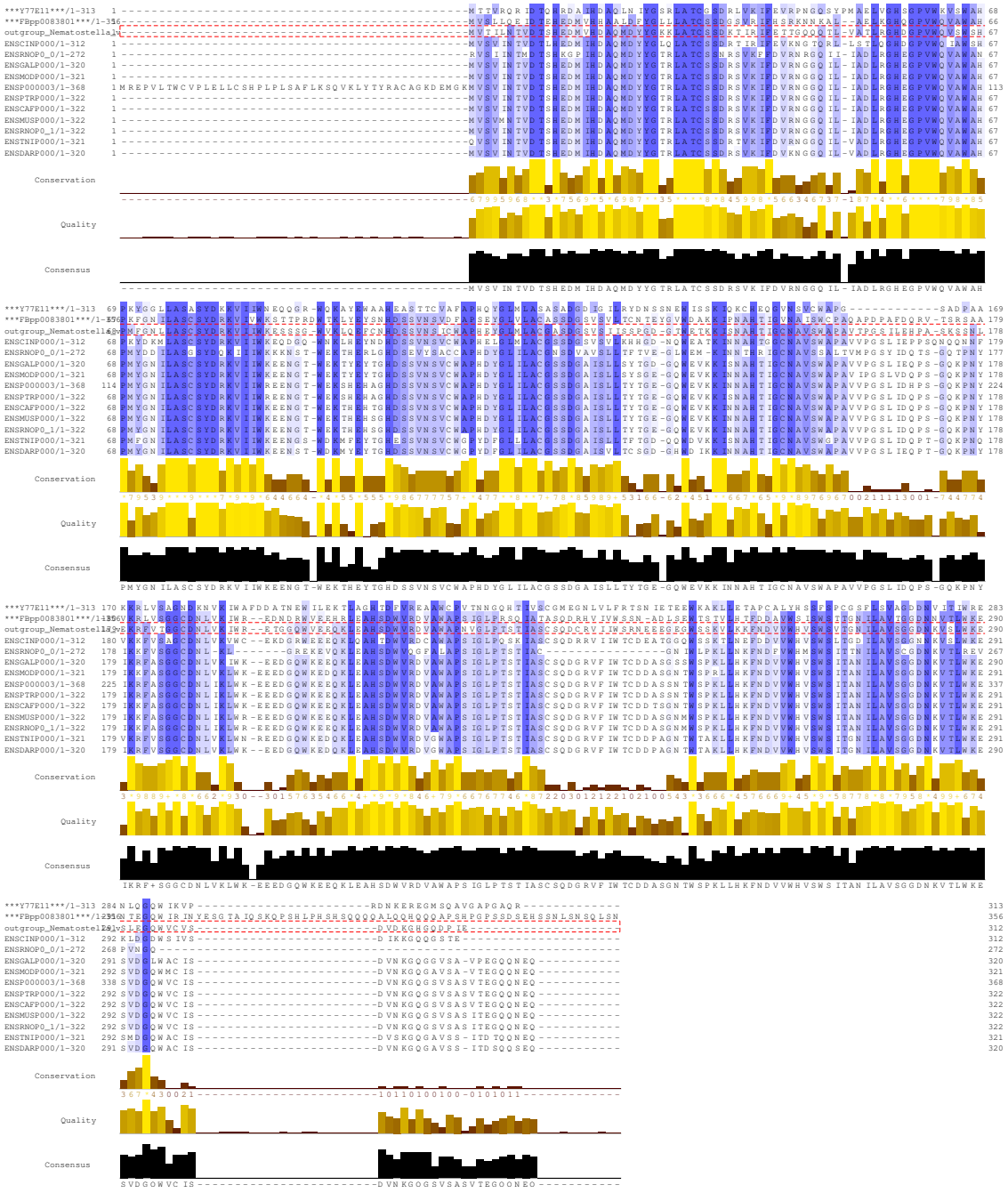


Figure 8.3 – *Alignement multiple du groupe d'orthologues Sec13 (RefOG001) (MUSCLE). Les séquences considérées sont celles d'orthoBENCH ainsi que celles ajoutées par la méta-approche. La séquence outgroup est encadrée en rouge.*

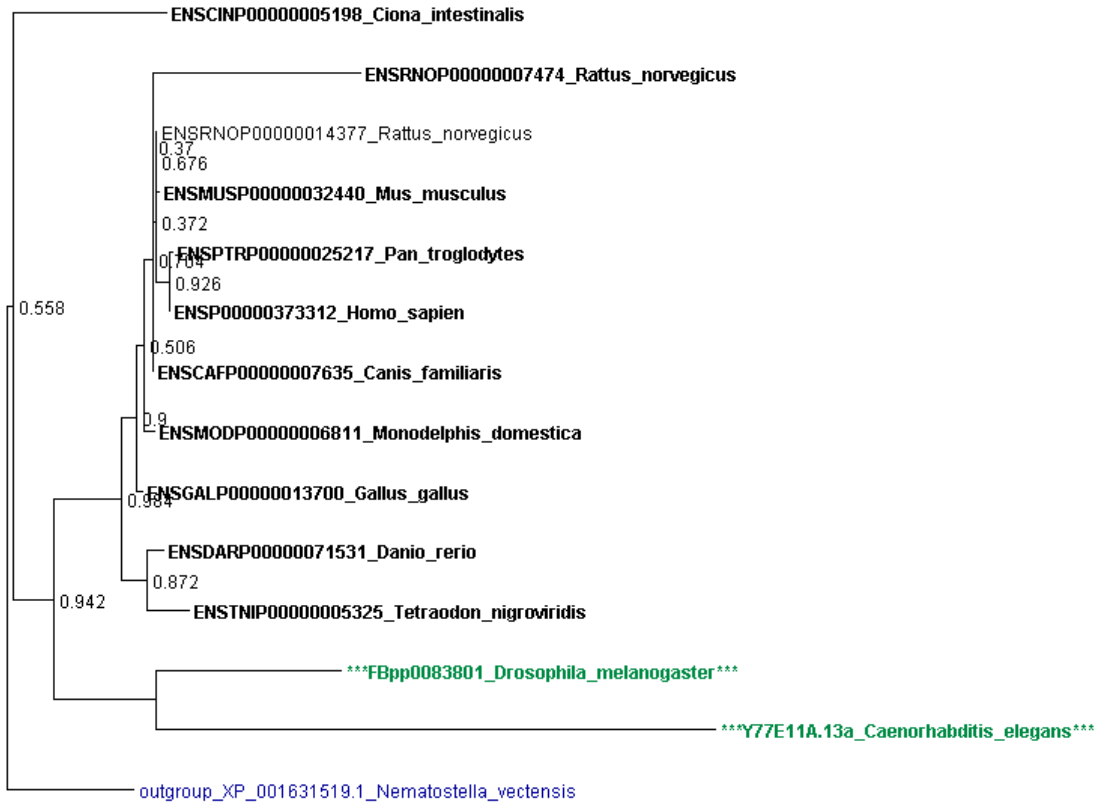


Figure 8.4 – *Arbre phylogénétique obtenu sur le groupe RefOG001. Les séquences considérées sont celles d'orthoBENCH ainsi que celles ajoutées par la méta-approche. Les séquences prédites, mais non présentes dans le groupe RefOG001 sont indiquées en vert. Les séquences appartenant à l'intersection de 4 méthodes sont figurées en gras. Comme dans l'article d'orthoBENCH, l'enracinement a été fait à l'aide d'une séquence appartenant à Nematostella vectensis (outgroup en rouge dans l'arbre). Méthode alignement : muscle, Arbre : PhyML, bootstrap : 500.*

appartenant à ce groupe d'orthologues par les 4 méthodes. Le groupe prédit a une identité moyenne de 72,3% et une variance de 21%. Les deux séquences ajoutées, FBpp0083801 et Y77E11 présentent des identités moyennes de respectivement 49,8% et 39,16%. L'alignement multiple (voir figure 8.3) montre que les sites conservés chez les séquences de RefOG001 sont également conservés dans les deux séquences que nous ajoutons. L'arbre phylogénétique obtenu avec les 13 séquences (voir figure 8.4) ainsi qu'avec une séquence homologue de *Nematostella vecensis* (*outgroup*) montre que les 2 séquences non associées à RefOG001, sont relativement éloignées des 11 autres. Ces séquences forment une deuxième branche, inclue entre séquences du groupe. Il est possible qu'il s'agisse de paralogues. Cependant, il s'agit des seules copies de gènes de ce groupe chez *D. melanogaster* et *C. elegans*. Il pourrait donc également s'agir de séquences orthologues ayant plus divergées que chez les autres espèces.

Le groupe d'orthologues correspondant au **récepteur TNFRSF1A du facteur de nécrose tumorale** (RefOG057) est un groupe contenant 11 séquences pour lequel la méta-approche trouve 11 séquences, dont deux différentes de celle du groupe de référence (alignement présenté en figure 8.5 et arbre fourni en figure 8.6). Il n'y a ainsi au niveau global du groupe ni fusion ni fission, cependant il manque deux séquences et deux autres séquences ont été ajoutées. Le groupe prédit par la méta-approche a un pourcentage d'identité moyen de 74,08% (variance = 14,66%). Si les 2 séquences non prédites par la méta-approche, mais présentes dans orthoBENCH y sont ajoutées, le pourcentage d'identité moyen diminue et tombe à 53,95% avec une augmentation de la variance (variance = 27,29). Cela montre que les deux séquences non prédites sont plus distantes des autres séquences que les séquences prédites par la méta-approche. Afin d'analyser ce groupe, les séquences du groupe RefOG057 ainsi que celle du groupe prédit par la méta-approche ont été alignées (MUSCLE, voir figure 8.5) et un arbre phylogénétique a été produit à partir de cet alignement

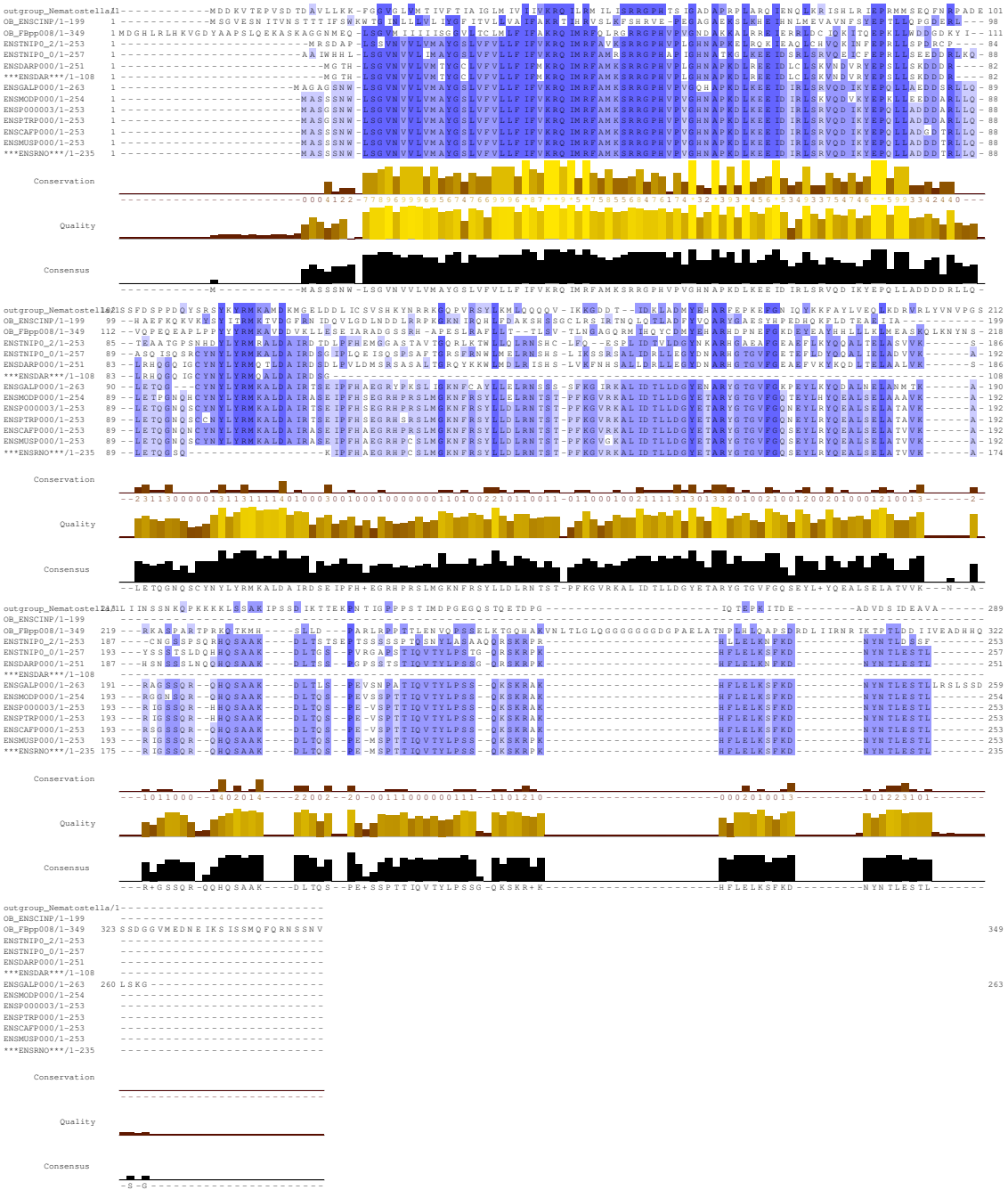


Figure 8.5 – *Alignement multiple du groupe RefOG057. Les séquences considérées sont celles d'orthoBENCH ainsi que celles ajoutées par la méta-approche. Les séquences de RefOG057 non prédites par la méta-approche sont précédées de OB, les séquences prédites, mais non présentes dans le groupe RefOG057 sont encadrées par des '***'. Nous avons utilisé la même espèce qu'orthoBENCH pour définir l'outgroup.*

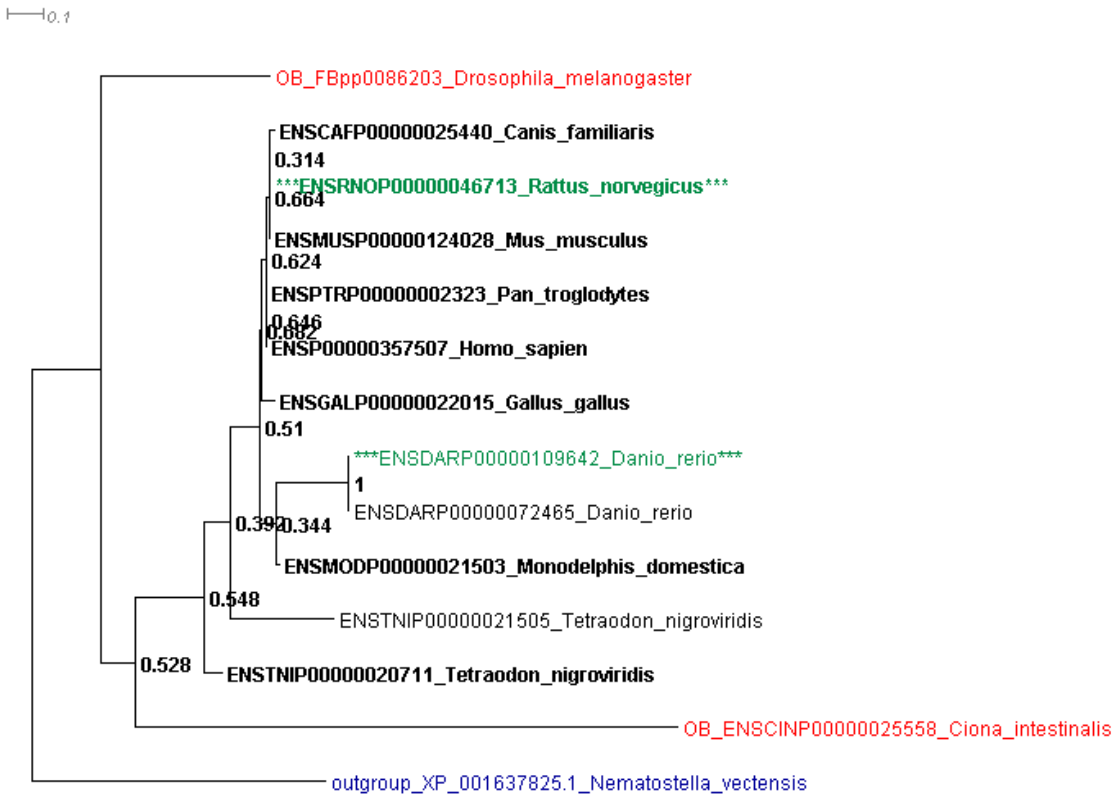


Figure 8.6 – *Arbre phylogénétique obtenu sur le groupe RefOG057. Les séquences considérées sont celles d'orthoBENCH ainsi que celles ajoutées par la méta-approche. Les séquences de RefOG057 non prédites par la méta-approche sont signifiées en rouge et précédées de OB, les séquences prédites, mais non présentes dans le groupe RefOG057 sont indiquées en vert et suivie d'étoiles. Les séquences appartenant à l'intersection de 3 méthodes sont figurées en gras. L'enracinement a été fait à l'aide d'une séquence appartenant à Nematostella vectensis (outgroup, en bleu). Méthode alignement : muscle, Arbre : PhyML. Bootstrap : 500.*

(PhyML) (voir figure 8.6). L'enracinement a été fait à l'aide d'une séquence appartenant à *Nematostella vectensis*, métazoaire éloigné des 12 présents dans le jeu de données initial. La prédiction du groupe d'orthologues correspondant au groupe RefOG057 est issue de l'intersection de 3 méthodes. Cette intersection est composée de 8 séquences sur les 12. Parmi ces 8 séquences (en gras sur la figure 8.6) , la séquence ENSRNOP00000046713 n'appartient pas au groupe RefOG057. Cependant, le fait que 3 méthodes aient retrouvé cette séquence et qu'elle se situe au cœur de l'arbre semble indiquer le contraire. Dans sa dernière mise à jour (avril 2014) orthoBENCH a ajouté à ce groupe un splice-variant de ENSRNOP00000046713. La seconde séquence ajoutée par la méta-approche, ENSDARP00000109642 (en vert et suivie de '***'), est plus proche phylogénétiquement du reste du groupe RefOG057 que les deux séquences non trouvées. Les deux séquences non trouvées (ENSCINP00000025558 et FBpp0086203, en rouge et précédées par OG) appartiennent vraisemblablement au groupe d'orthologues car elles se situent sur la même branche que les séquences du groupe. Cependant la longueur des branches fait penser qu'elles ont vraisemblablement plus divergé que les autres séquences. De ce fait, la méta-approche a dans ce cas été trop stringente mais a permis de trouver deux séquences qui n'étaient pas présentes dans le groupe de référence alors qu'elles semblent pertinentes.

Ces exemples illustrent notre choix de travailler au niveau des fissions ou fusions de groupes. Ainsi, une faible marge de manœuvre permet de tenir compte de possibles erreurs dans orthoBENCH. On observe cependant que l'absence de prédictions de certaines séquences (exemple de RefOG057) ne sera pas prise en compte si les séquences manquantes ne sont pas associées à un autre groupe. L'analyse de fusion ou de fission de groupes ne permet donc pas d'observer l'ensemble des erreurs de type faux négatif (manque de relations). Cette évaluation favorise de ce fait les petits groupes (spécifiques mais moins sensibles). La méta-approche produit de plus grand groupes que les méthodes initiales et n'est donc pas favorisée par ce type

d'analyse. Les bon résultats de la méta-approche s'interprètent donc comme le reflet de l'ajout de vrai positifs.

Chapitre 9

MARIO : le programme implémentant la méthode

Le programme MARIO (Pereira et al., 2014) est disponible gratuitement (page web MARIO) (voir figure 9.1). Cette page met à disposition à la fois les scripts, mais également le fichier d'aide (ReadMe) et les groupes d'orthologues générés par les méthodes sur les benchmarks *orthobench* et *orthology benchmark service*.

9.1 Les paramètres du programme

Le programme MARIO utilise les logiciels MUSCLE (programme d'alignement multiple) et HMMER (programme de création de profils HMM et comparaison séquence/profils) (Eddy, 2011). Pour plus de facilité, ces programmes sont fournis dans le dossier contenant les scripts de MARIO. De plus, les librairies Perl Math::Combinatorics, Cwd, Bio::SeqIO, File::Basename, File::Spec et XML::Simple doivent être disponible pour Perl lors de l'exécution du programme. Perl recherche ses librairies dans les dossiers listés dans son 'classpath' (variable d'environnement). Pour ajouter des librairies à Perl deux solutions : les installer *via* cpan ou ajouter au

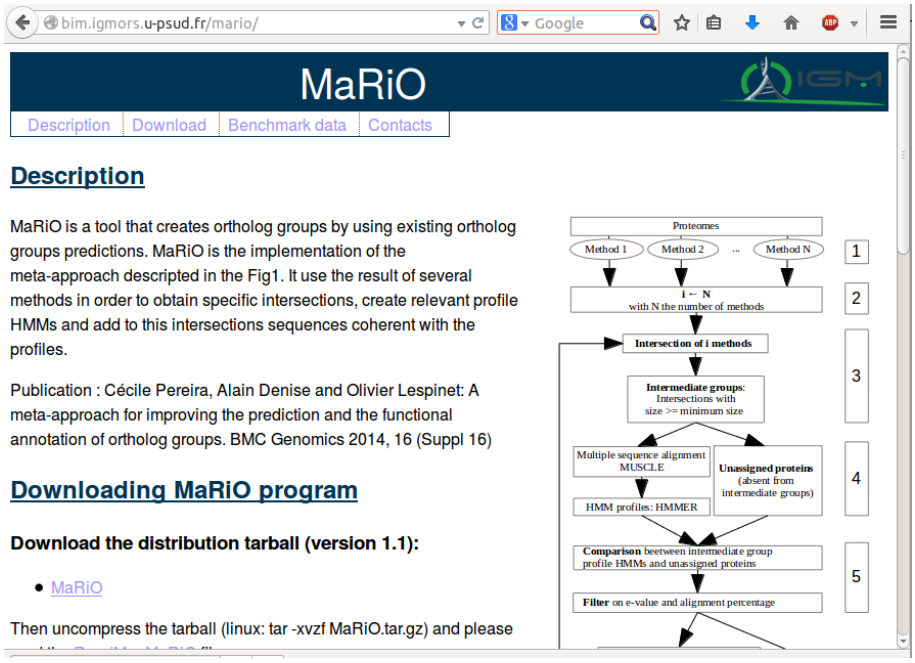


Figure 9.1 – *Impression-écran du site web MARIO (page web MARIO).*

classpath de Perl le dossier contenant ces librairies (dans notre cas le dossier contenant nos scripts). Dans le but de permettre l'utilisation des nouveaux standards de la communauté, les protéomes peuvent être fournis dans deux formats différents, le format fasta et le format seqXML. Pour chaque espèce, il est nécessaire de fournir un unique fichier contenant l'ensemble des séquences avec lesquelles nous devons travailler.

La méta-approche utilise plusieurs paramètres dont les valeurs ont été optimisées sur le jeu de données orthoBENCH. Ces paramètres sont modifiables. Ainsi peuvent être modifiés : la taille minimum des intersections (-inter), le pourcentage d'alignement séquence / profil (-pa) et l'evalue associée au hit séquence / profil HMM (-e). Le programme a la possibilité de tourner sur plusieurs cœurs (paramètre -cpu). Le nom du dossier contenant les fichiers intermédiaires ainsi que le nom du dossier contenant les résultats finaux sont défini par défaut (TMP et RESULT), mais ils restent cependant modifiables par l'utilisateur *via* les options -tmp et -res.

Enfin, les résultats sont fournis au format fasta (un dossier par groupe d'orthologues) ainsi qu'au format orthoXML (un fichier contenant l'ensemble des groupes).

Exemple de commande de lancement du programme MARIO :

>perl -I SCRIPTS/perl_library/ MARIO.pl -i /INPUT_GROUPS/ -f /FASTA_PROTEOMES/ -cpu 6 1>MARIO.out 2> MARIO.err

L'option -I permet d'ajouter le dossier SCRIPTS/perl_library temporairement au classpath de l'environnement perl. Les groupes obtenus par différentes méthodes (**i**nputs) sont disponibles dans le dossier INPUT_GROUPS (paramètre -i) et les protéomes (**f**asta) dans le dossier FASTA_PROTEOMES (paramètre -f). Les paramètres de filtres sont ceux par défaut. Le programme tourne ici sur 6 cpu, le fichier output est MARIO.out et le fichier de sorties erreurs MARIO.err.

9.2 Temps de calcul

Notre méta-approche utilise les groupes d'orthologues obtenus par plusieurs autres méthodes (méthodes inputs). Il est naturel de se demander la nature de la corrélation entre le nombre de méthodes utilisées et le temps de calcul. La méta-approche boucle sur les intersections de N à 2 méthodes. L'intuition est donc que plus le nombre de méthodes est grand et plus le temps de calcul sera long. Les résultats ne sont cependant pas ceux-là. La figure 9.2 présente le temps de calcul de MARIO en fonction du nombre de méthodes dont les résultats ont été utilisés en input. On observe ainsi que le temps de calcul ne semble pas lié de manière linéaire au nombre

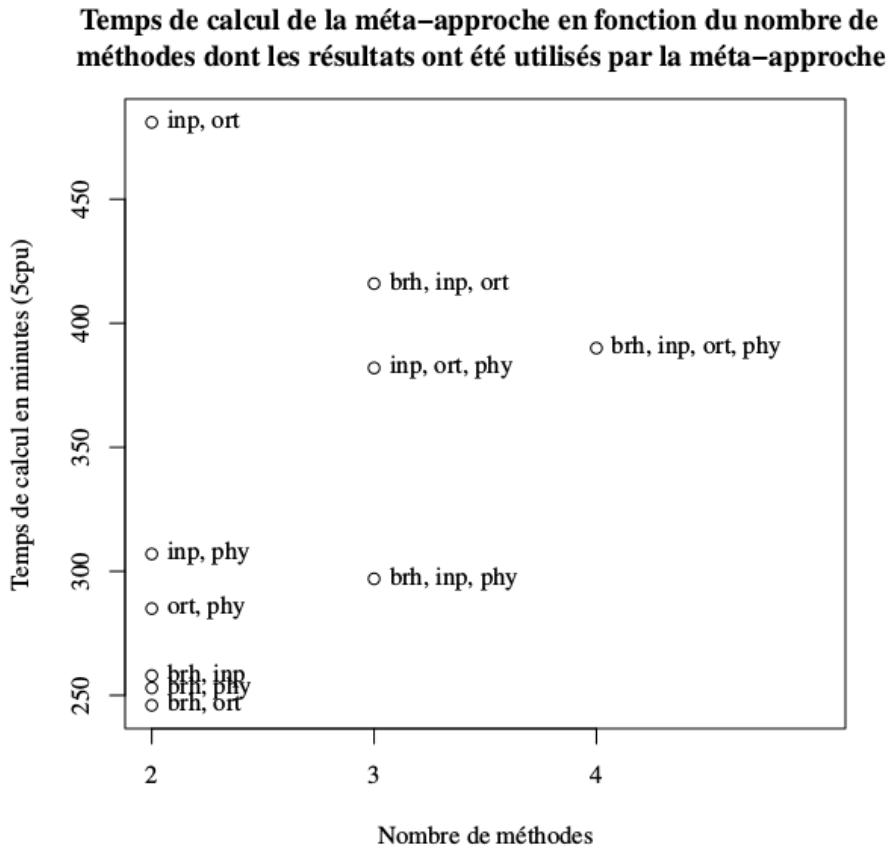
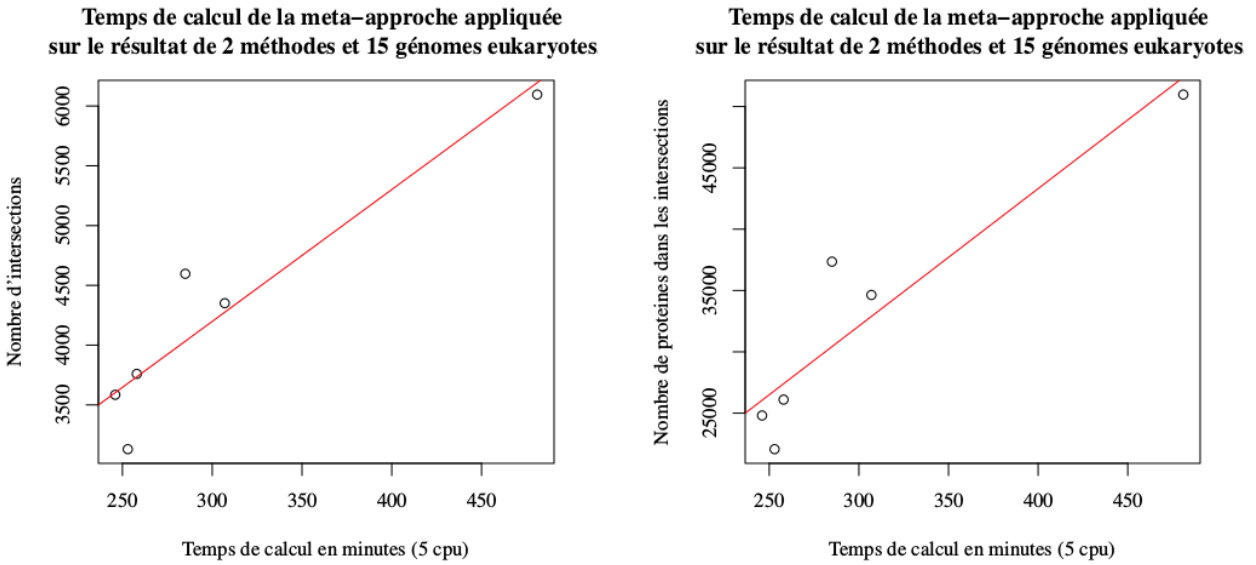


Figure 9.2 – *Temps de calcul de MARIO en minutes en fonction du nombre de méthodes dont les groupes sont utilisés par la méta-approche. Quinze protéomes d'eucaryotes ont été utilisés pour l'analyse. Abréviations : inp : inparanoid + lien simple (20%), ort : orthoMCL, phy : phylogeny, brh : best reciprocal hits.*

de méthodes. En effet, on peut observer un temps de calcul de la méta-approche appliquée aux résultats de deux méthodes (Inparanoid et OrthoMCL) plus long que le temps de calcul associé au travail sur les résultats des 4 méthodes et cela avec les mêmes protéomes de départ.

Au vu de ce résultat, nous avons analysé le temps de calcul en fonction du nombre et de la taille des intersections (voir la figure 9.3). On observe une corrélation entre le nombre d'intersections (et le nombre de protéines impliquées dans

des intersections) et le temps de calcul. Ainsi, la partie du programme effectuant les alignements multiples ainsi que les profils HMM des intersections semble significativement plus longue que la comparaison des séquences seules à ces profils. L'utilisation de la dernière version de HMMER (Eddy, 2011) pourrait modifier cette observation, les développeurs de ce programme défendant dans l'article une division du temps de calcul par 10.



(a) Temps de calcul de MARIO en minutes en fonction du nombre d'intersections. En rouge : régression linéaire (constante de régression : 889.36, pente : 11.03, Pearson's product-moment corrélation test, greater, correlation : 0.94, p-value : 0.002644).

(b) Temps de calcul de MARIO en minutes en fonction du nombre protéines impliquées dans des intersections. En rouge : régression linéaire (constante de régression: -1511, pente : 112, Pearson's product-moment correlation test, greater, correlation: 0.929, p-value: 0.003679)

Figure 9.3 – *Temps de calcul en fonction du nombre d'intersections et du nombre de protéines impliquées dans les intersections. MARIO a été appliqué sur un serveur en parallèle sur 5 cpu. Quinze protéomes d'eucaryotes ont été utilisés pour l'analyse.*

Indépendamment de la première étude, nous avons analysé le temps de

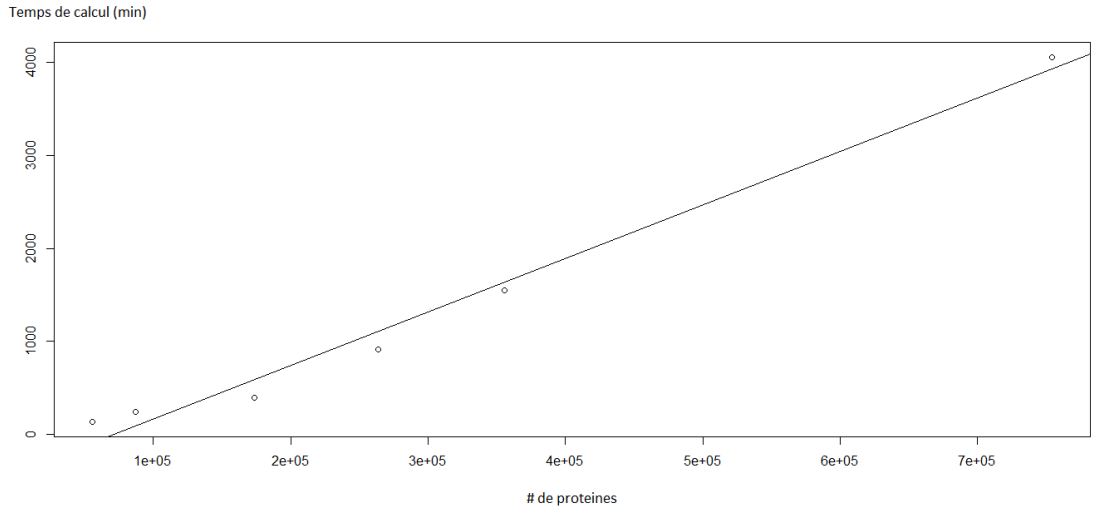


Figure 9.4 – *Temps de calcul de MARIO en fonction du nombre de protéines. Temps de calcul sur un serveur de calcul, 1 cpu utilisé. Test effectué pour 5, 10, 15, 20 et 66 espèces choisies parmi les espèces de référence. Le temps de calcul suit une régression linéaire :* $temps\ de\ calcul = 1.119 * \#species - 6.866$*,*
Le coefficient de corrélation est de 0.996 (Pearson's product-moment correlation p-value de 3e−4).

calcul en fonction du nombre de protéomes. Cette analyse a été réalisée sur les protéomes de référence version 2011. Pour cette analyse, le script MARIO a été appliqué plusieurs fois sur un nombre de protéomes croissant. Les protéomes inclus dans le jeu de données n-1 étant inclus dans le jeu de données n. On observe figure 9.4 que le temps de calcul semble linéaire en fonction du nombre d'espèces. Cependant, nous avons montré plus haut que le temps de calcul est fortement impacté par le nombre d'intersections. Il serait donc intéressant, afin de vérifier cette relation de linéarité, de ré-effectuer l'analyse en plusieurs réplications sur des groupes d'organismes choisis par *bootstrap* afin de limiter le biais lié aux variations de la taille des intersections. Cette analyse est en cours.

Nous avons montré dans ce paragraphe que ce n'est pas le nombre de mé-

thodes, mais le nombre d'intersections ainsi que le nombre de protéines par intersection qui semble induire une augmentation du temps de calcul. De ce fait, l'utilisation des résultats de méthodes très corrélées, outre le fait qu'ils n'induisent pas la formation d'intersections très différentes des groupes initiaux, induit un temps de calcul plus long.

Il pourrait également être intéressant de tester l'utilisation d'autres programmes d'alignement (tels que MAFFT) ainsi que d'autres programmes de créations de profils (tels que PSSM ou HH-suite) afin d'observer s'il est possible de diminuer le temps de calcul tout en préservant (ou améliorant) la qualité des groupes prédits.

Chapitre 10

FungiPATH : mise à jour du site web et de la base de données

La base de données FungiPath a été mise à jour *via* l'application de la méta-approche (MARIO) sur 178 protéomes d'eucaryotes, dont 174 protéomes de champignons (voir la partie liste et description des espèces de travail dans l'introduction). La mise à jour de la base de données s'est associée à la mise à jour du site web Fungipath.

Comme pour la version précédente de FungiPath, l'annotation des groupes d'orthologues a été faite par comparaison des groupes d'orthologues finaux avec les séquences annotées dans l'union de SwissProt et Metacyc (voir la figure 10.1).

Une nouveauté a consisté en l'annotation avec des termes GO. Deux bases indépendantes de séquences annotées ont été créées, une pour les séquences annotées avec au moins un *EC number* et une pour les séquences annotées avec au moins un terme GO. Ces bases sont ensuite comparées indépendamment aux profils des groupes d'orthologues prédits par la méta-approche. La séquence annotée présentant la plus faible *e-value* lors de la comparaison avec le profil HMM permettra le transfert de l'annotation si l'*e-value* de comparaison est inférieure à 1E-80 (seuil

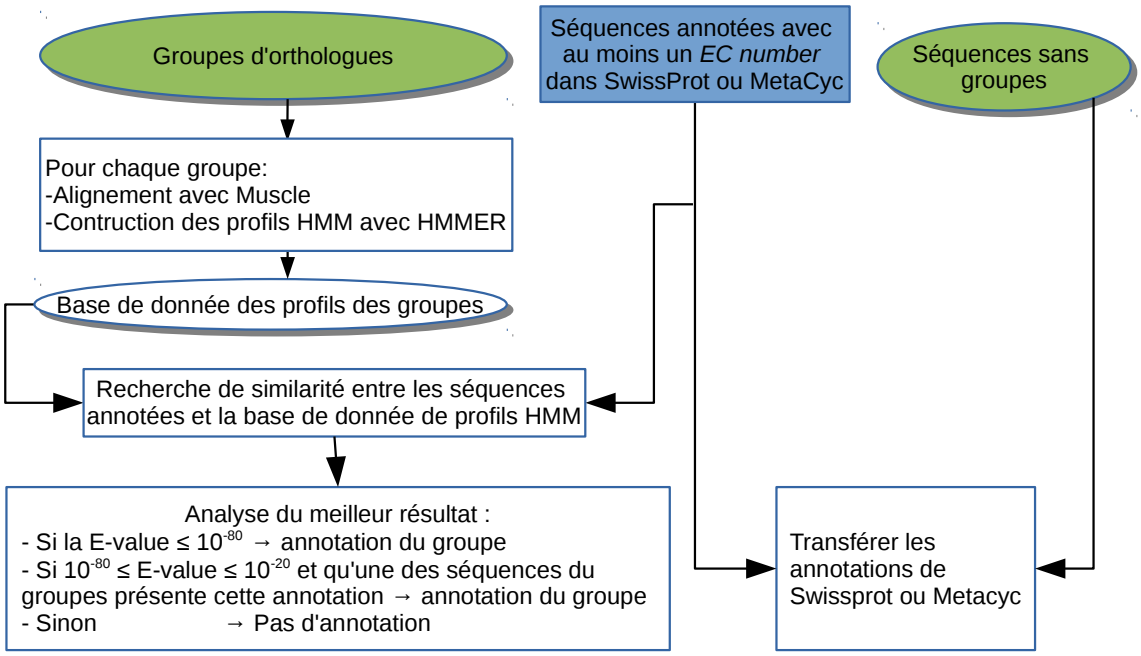


Figure 10.1 – *Pipeline d'annotation des groupes d'orthologues. Les étapes sur fond vert représentent les résultats obtenus avec la méta-approche. Le carré sur fond bleu les données à télécharger sur les sites SwissProt et MetaCyc.*

stringent). Dans le cas où celle-ci serait comprise entre 1E-20 et 1E-80 l'annotation n'est transférée que si l'une des séquences du groupe porte cette annotation dans Swissprot ou Metacyc. Il s'agit d'annotations déjà vérifiées pour au moins une des séquences du groupe. Vu que l'ensemble du groupe est censé avoir la même annotation (similarité de fonction) et que l'on connaît la fonction d'une des séquences du groupe, on devrait pouvoir transférer cette annotation à l'ensemble du groupe.

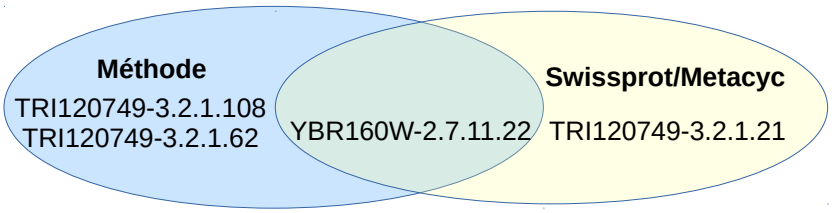
Nous réfléchissons actuellement à améliorer cette annotation. Pour cela, nous proposons de ne pas conserver deux seuils d'e-value, mais un seul et cela afin d'éviter un biais d'annotation induisant une surévaluation du nombre de groupes correctement prédits. De plus, nous réfléchissons à ne plus conserver uniquement le meilleur hit, mais un ensemble de meilleurs hits afin d'effectuer l'annotation. Cette proposition repose sur trois observations : un enzyme peut porter plusieurs

annotations, les annotations SwissProt et Métacyc ne sont pas exhaustives et le meilleur hit de comparaison profil HMM/ séquence annotée peut être légèrement biaisé entre autres par la qualité de l'alignement multiple. Le choix de prendre les N meilleurs hits présentant une e-value inférieure à un certain seuil peut paraître arbitraire. Ainsi, nous proposons d'utiliser prochainement en tant que seuil, en plus d'un seuil minimum d'e-value, la présence, quand il existe, d'un saut dans les e-values résultats.

Le site web FungiPath est un site associé à une base de données dont le but est d'aider à l'analyse de l'évolution du métabolisme. Il permet à la fois l'accès aux groupes d'orthologues ainsi qu'aux annotations *EC numbers* et termes GO. Chaque voie métabolique de KEGG et MetaCyc est présente sur notre site web. Sur ces voies métaboliques, les EC numbers sont colorés en fonction de la proportion d'organismes sélectionnés présentant au moins une séquence annotée avec cet *EC number*. L'utilisateur a la possibilité de sélectionner les espèces l'intéressant plus particulièrement.

Les caractéristiques de la base de données obtenue sur ces 178 protéomes sont disponibles sur le tableau 10.3. On observe ainsi que 4504 groupes d'orthologues sont annotés avec au moins un *EC number*. La qualité de ces annotations est évaluée *via* la comparaison des couples ID-EC vérifiés (dans SwissProt et Metacyc) et ceux prédits par la méta-approche. L'annotation des groupes est faite à partir de la comparaison des profils des groupes d'orthologues prédits *via* la comparaison avec les bases de données SwissProt et Metacyc. Bien que l'ensemble des deux bases de données (et pas uniquement les séquences de champignons) est utilisé dans l'étape d'annotation, on notera que, vérifier les annotations prédites en utilisant SwissProt sur leur correspondance avec SwissProt elle-même, induit certainement une surestimation de la qualité de ces annotations.

Séquences	Annotations		Annotations dans Swissprot / Metacyc	
Identifiant	EC	ID-EC	EC	ID-EC
TRI120749	3.2.1.108 3.2.1.62	TRI120749-3.2.1.108 TRI120749-3.2.1.62	3.2.1.21	TRI120749-3.2.1.21
YBR160W	2.7.11.22	YBR160W-2.7.11.22	2.7.11.22	YBR160W-2.7.11.22
AAL003Wp	2.7.8.5	AAL003Wp-2.7.8.5	∅	∅



	Exemple
Annotations spécifiques à Swissprot ou Metacyc	1
Annotations identiques	1
Nouvelles annotations proposées	2

Figure 10.2 – *Définition et exemples de couples ID-EC. Le premier tableau donne l'exemple de trois séquences présentant une annotation dans SwissProt ou MetaCyc (en jaune) et une annotation que nous prédisons (en bleu). Pour chaque séquence et chacune des annotations associées, nous formons un couple 'identifiant de la séquence - une de ces annotations'. Il est ainsi possible d'analyser le nombre d'annotations de Swissprot et MetaCyc non retrouvées par notre annotation, le nombre d'annotations en commun ainsi que le nombre d'annotations nouvelles.*

Statistique des groupes d'orthologues prédits	
# de groupes	68 182
# de groupes annotés avec au moins un *EC number*	4 504
# d'*EC numbers* différents	1 294
# de groupes annotés avec au moins un terme GO	14 057
# de termes GO différents	7 544
Qualité de l'annotation	
# de couples ID-EC annotés spécifiques à SwissProt	359 (dont 251 avec id dans un groupe)
# de couples ID-EC prédits et vérifiés dans SwissProt	10 275 (dont 6489 avec id dans un groupe)
# de couples ID-EC prédits et absents de SwissProt	351 088

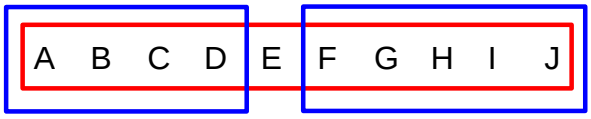
Figure 10.3 – *Caractéristiques de la base de données FungiPath 178 (méta-approche MARIO)*

On observe que parmi les couples ID-EC présents dans SwissProt moins de 3% des annotations SwissProt sont manquées. Ce faible nombre d'erreurs, même s'il est certainement sous-estimé, est cohérent avec la forte similarité de Schliker moyenne obtenue sur les 66 protéomes de référence. Les groupes d'orthologues prédits par la méta-approche sont annotés en fonction de leurs profils. Cette annotation peut se faire avec des séquences de l'ensemble du règne. Ainsi, il est possible de transférer au groupe l'annotation de séquences déjà connues chez les champignons (séquence intra-groupe annotée), mais aussi de proposer de nouvelles fonctions encore inconnues chez ces organismes. L'application de la méta-approche sur 178 génomes d'eucaryotes a permis la prédiction de 351 088 nouveaux couples ID-EC. Ce nombre est lié à l'hétérogénéité du nombre d'annotations vérifiées chez ces organismes. Ainsi, nous retrouvons l'annotation connue présente dans SwissProt et ajoutons des annotations potentielles pour les organismes faiblement annotés.

Chapitre 11

Perspectives

11.1 Comparaison des intersections entre elles



Intersection des deux méthodes : [ABCD] et [FGHIJ]

Figure 11.1 – *Exemple d'intersections de deux méthodes. Dix séquences sont figurées (A-J). La méthode rouge forme un unique groupe, la méthode bleue propose deux groupes. L'intersection des deux méthodes aboutie à la formation de deux ensembles [ABCD] et [FGHIJ].*

Notre méta-approche est basée sur l'utilisation des intersections. Or si parmi l'ensemble des méthodes l'une d'elles a tendance à plus diviser les groupes alors des séquences ayant été placées dans un groupe par l'une des méthodes et deux groupes par l'autre, seront placées dans deux intersections différentes et donc deux groupes différents. Ainsi sur l'exemple de la figure 11.1 la première méthode (méthode rouge) ne forme qu'un groupe alors que la seconde méthode (méthode bleue) en forme

E-value	% alignement	min. seq	Groupes modifiés
1e-10	40	3	43,07%
1e-10	60	3	23,72%
1e-10	80	3	7,85%
1e-10	90	3	2,66%
1e-30	60	3	18,05%
1e-5	60	3	24,79%
1e-5	75	3	11,65%
1e-5	80	3	8,11%
1e-5	90	3	2,70%

Table 11.1 – *Pourcentage de groupes touchés par un regroupement en fonction des paramètres de seuil. Le seuil de la taille minimum de l'intersection fixé à 3.*

deux. La méta-approche en formera alors deux sans tester si les deux intersections n'appartiennent en fait pas au même groupe.

Afin d'observer si ce cas de figure arrive souvent, nous avons souhaité comparer les intersections de quatre méthodes entre elles.

Les comparaisons profils HMM/ profils HMM permettent de regrouper des séquences très éloignées phylogénétiquement, et ce n'est pas notre but puisque nous cherchons des groupes d'orthologues à des fins de comparaison fonctionnelle.

Nous avons donc comparé les séquences de chacune des intersections de quatre méthodes contre la base de données de profils HMMs de ces mêmes intersections. Si toutes les séquences de l'intersection A obtiennent une *e-value* de comparaison au groupe B inférieur à un certain seuil et un pourcentage d'alignement supérieur à un second seuil et inversement alors on considérera que les groupes A et B auraient pu être fusionnés (voir la figure 11.2).

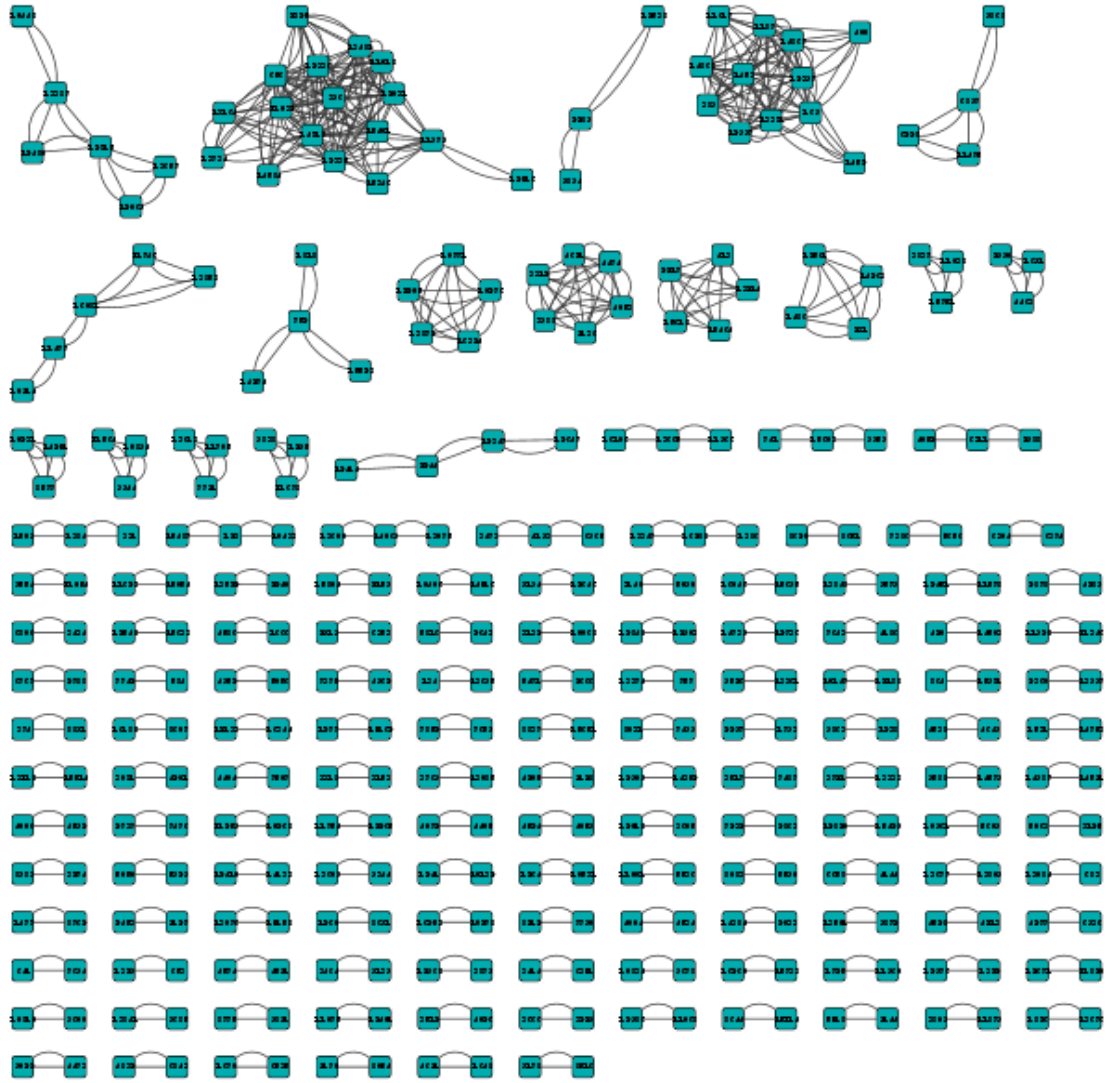


Figure 11.2 – *Graphe des liens possibles entre les intersections de quatre méthodes (355/13 348 groupes). Deux intersections (nœuds du graphe représentés par des carrés bleus) sont reliées par un lien si l'ensemble des protéines de l'une des deux intersections à un résultat de comparaison séquence/profil HMM du second groupe présentant une e-value inférieure à 10^{-10}, un pourcentage d'alignement supérieur à 90% (taille minimum d'intersection égale à 3).*

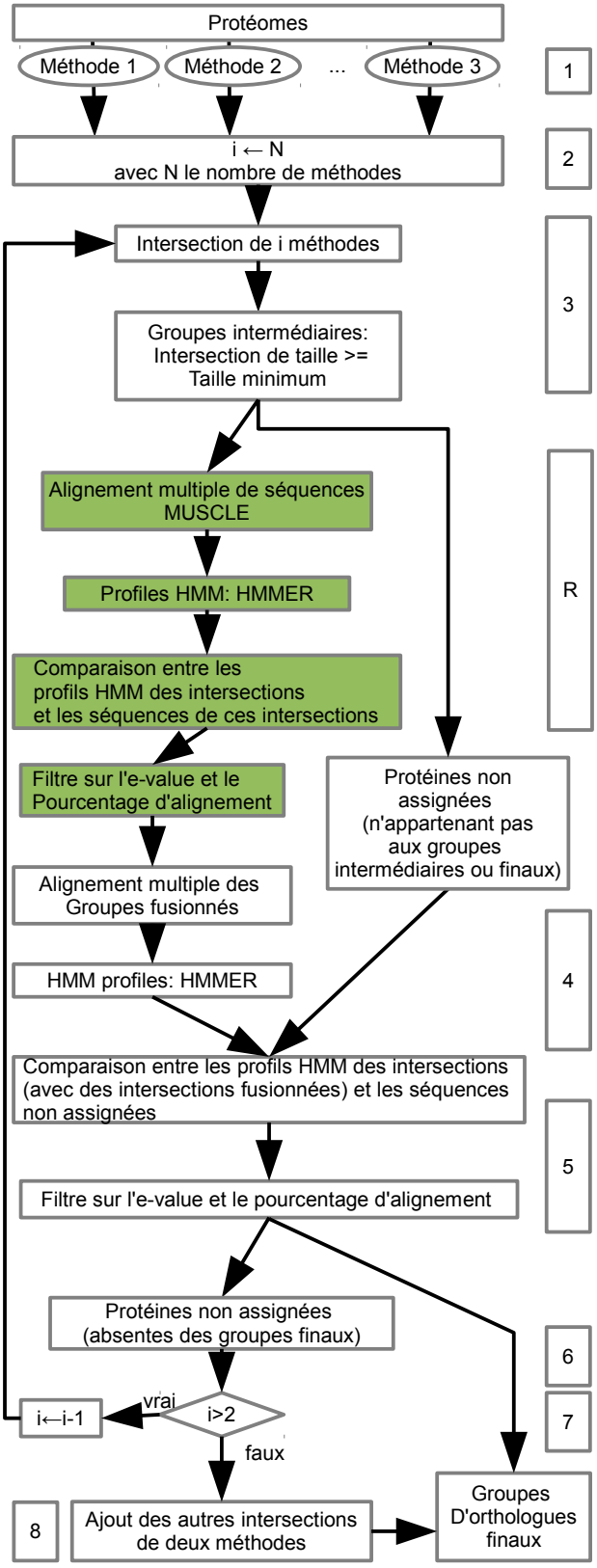


Figure 11.3 – *Algorithme d'une version alternative de la méta-approche incluant une comparaison des intersections entre elles. L'étape ajoutée est l'étape R pour rajout (cadres sur fond vert).*

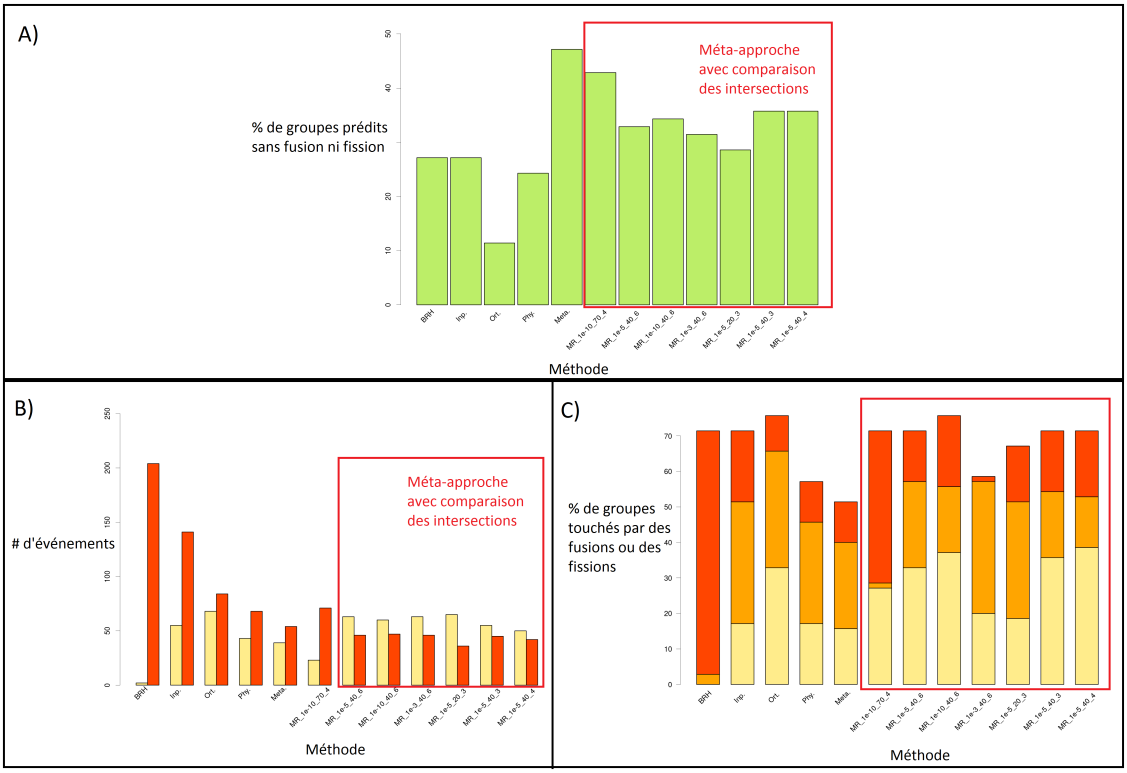


Figure 11.4 – *Analyse des résultats de la méta-approche comparant les intersections entre elles. A) Pourcentage de groupes d'orthoBENCH (RefOG) bien prédits en fonction de la méthode. B) Nombre de fusions ou de fissions en fonction de la méthode. C) Pourcentage de groupes touchés par des fusions ou des fissions. Abréviations : Inp. pour inparanoid, Ort. pour orthoMCL, Phy. pour phylogeny, Meta pour la méta-approche (MARIO) et MR pour méta-approche avec regroupement. Pour chaque test de méta-approche avec regroupement, nous indiquons après MR la liste des seuils. Par exemple, $MR\_1e-10\_70\_4$ signifie que l'évalue seuil testé était $(1e-10)$, le pourcentage d'alignement seuil 70% et la taille minimum des intersections était de quatre séquences.*

Cette étude a été réalisée sur les données des protéomes de référence (version 2011) en parallèle du paramétrage de la méta-approche. De ce fait, les seuils testés ont été fixés de manière arbitraire (voir le tableau 11.1).

Le résultat de la comparaison des intersections de quatre méthodes entre elles au seuil d'e-value 1E-10 et de pourcentage d'alignement supérieur ou égal à 90% est visible figure 11.2. On observe que sur les 13 348 intersections de quatre méthodes de taille supérieure ou égale à 3, 355 (soit 2,66%) auraient été fusionnés. Le nombre de groupes touchés passe à 5 749 (43,07%) si l'on diminue le seuil du pourcentage d'alignement à 40% (voir le tableau 11.1).

Ainsi, la question de regrouper des intersections semble pertinente.

Á cette fin, nous avons testé une méta-approche alternative de détection de groupes d'orthologues. Celle-ci présente une étape supplémentaire par rapport à la méta-approche publiée (Pereira et al., 2014). Cette étape correspond à la comparaison des intersections entre elles (voir figure 11.3).

Comme pour l'évaluation de la méta-approche initiale, cette méta-approche modifiée (que nous appellerons MR pour méthode avec regroupements) a été évaluée sur les résultats obtenus sur orthoBENCH (voir la figure 11.4) ainsi que sur la similarité d'annotation des orthologues prédits sur 66 protéomes de référence (voir la figure 11.5).

La comparaison sur le jeu de données orthoBENCH a été faite en utilisant différents paramètres. On observe cependant que les combinaisons de paramètres testés ne permettent pas d'augmenter le nombre de groupes d'orthoBENCH prédits sans erreurs par rapport à la méta-approche sans regroupements (voir figure 11.4).

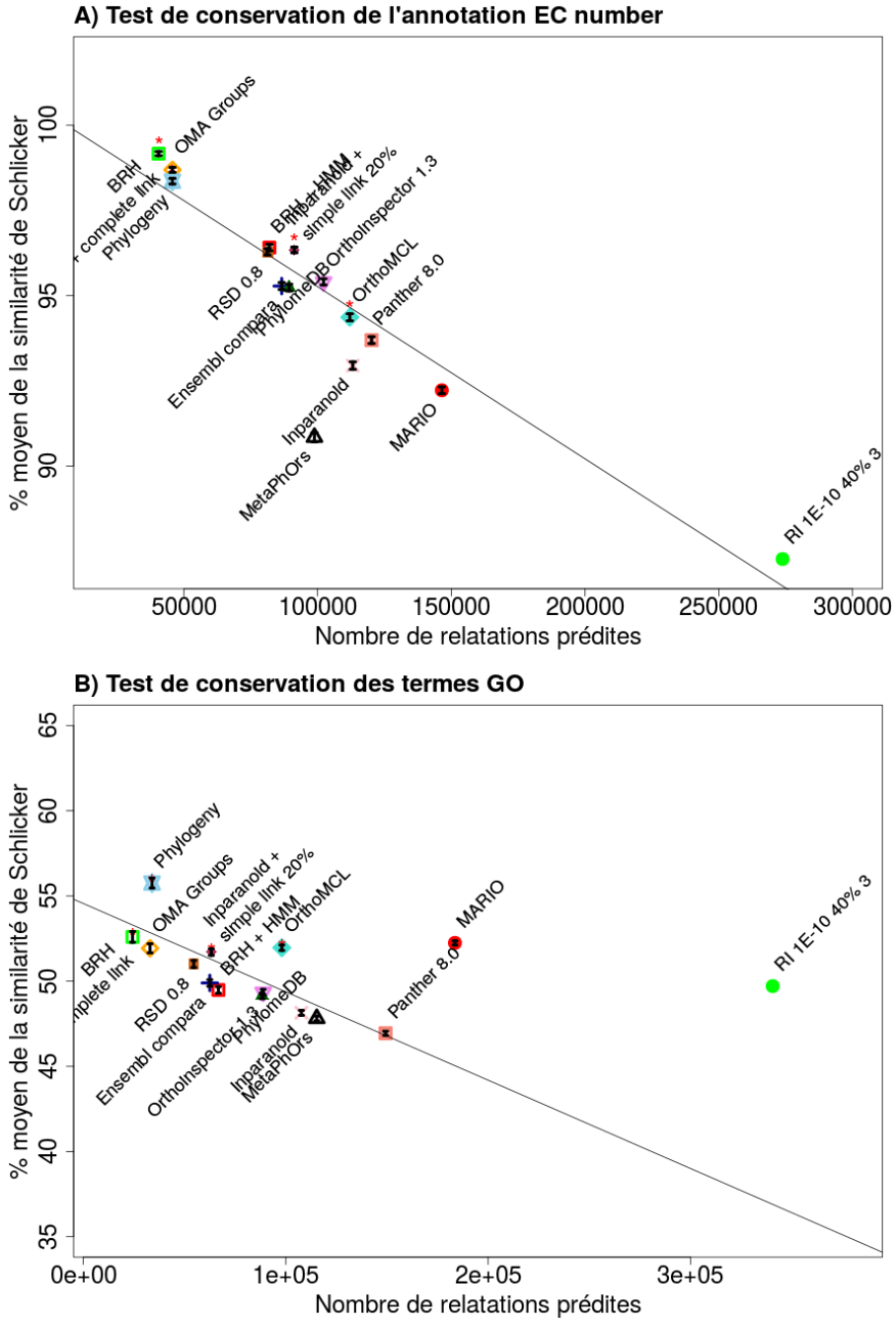


Figure 11.5 – *Analyse de la similarité de fonction des paires d'orthologues de la méta-approche regroupent des intersections. A) Test de conservation des EC numbers. B) Test de conservation des termes GO.*

La méthode de regroupement testée est figurée par RI (Regroupement des Intersections) suivit des paramètres appliqués pour cette méthode (mêmes paramètres que pour la méta-approche MARIO).

L'analyse des types d'erreurs de la méthode avec regroupement montre que cette approche à tendance à augmenter le nombre de fusions ainsi qu'à diminuer le nombre de fissions. Cela n'est pas étonnant puisque l'on permet une étape de regroupement entre les intersections. On observe que l'augmentation des regroupements est liée au seuil de MR. De ce fait, l'augmentation du pourcentage d'alignement séquence/ profile HMM induit une diminution du nombre de groupes fusionnés. Ainsi, une évaluation plus exhaustive des jeux de paramètre de la méthode MR induirait peut-être l'obtention de meilleurs résultats que ceux de la méta-approche actuelle.

L'analyse de MR sur la similarité d'annotation *EC numbers* et termes GO a été faite à titre d'exemple avec un jeu de paramètres non optimum (voir la figure 11.5). L'augmentation du nombre de fusions induites par cette méthode induit l'augmentation du nombre de relations prédites. Cependant, on observe que cette augmentation du nombre de relations est accompagnée d'une similarité d'annotations plus faible que celle des autres méthodes.

Cependant, le résultat obtenu sur la similarité d'annotation termes GO pour l'approche MR est significativement éloigné de la droite de régression. Ce résultat est encourageant pour cette approche. Il pourrait cependant s'agir du regroupement de paralogues ayant conservé une relative similarité de fonction.

L'optimisation de l'approche MR pourrait se faire de différentes manières. Un algorithme génétique pourrait par exemple être utilisé. Une seconde solution serait de travailler sur un niveau donné (par exemple l'intersection de quatre méthodes) et d'observer sur le graphe de test de similarité termes GO l'évolution de la similarité en fonction des paramètres de regroupement des intersections de ces quatre méthodes. L'évolution de la similarité termes GO des intersections de quatre méthodes plus ou moins regroupées est observable sur la figure 11.6. Le paramètre d'evalue impacte peu le regroupement. Le paramètre le plus discriminant semble être la longueur d'alignement. Sur ce graphe, une longueur d'alignement à 90% permet

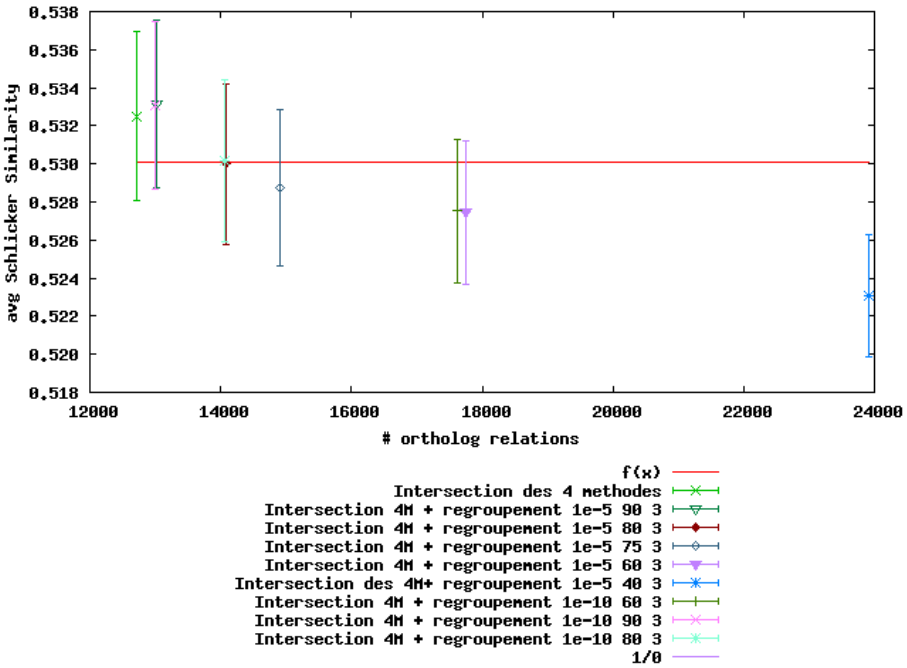


Figure 11.6 – *Variation du nombre de relations annotées et de la similarité de Schliker (termes GO) associée en fonction des regroupements d'intersections de quatre méthodes. Ces regroupements sont faits sur la base de comparaison des séquences des intersections avec les profils HMM de l'ensemble des intersections. Les intersections sont regroupées par la méthode "lien complet". Il faut que l'ensemble des séquences d'un groupe match en respectant les critères d'e-value et de pourcentage d'alignement contre l'autre groupe et inversement pour que le regroupement effectif de deux intersections soit possible. La ligne rouge représente la similarité finale obtenue avec la version initiale de la méta-approche (MARIO).*

l'augmentation de la similarité termes GO et un pourcentage minimum d'alignement à 80% permet d'obtenir une similarité comparable à celle de la méta-approche.

Au vu de son temps de calcul plus long, ainsi que de l'absence d'impact flagrant sur la qualité des résultats évalués sur orthoBENCH nous n'avons pas sélectionné cette version de la méta-approche pour la publication. Cependant, nous n'avions pas optimisé les paramètres de cette méthode. Cette optimisation pourrait

permettre d'obtenir une variante de la méta-approche présentant moins de fissions. L'obtention d'un plus petit nombre de groupes d'aussi bonne similarité de fonctions intra-groupe permettra de faciliter leurs analyses par des méthodes d'apprentissages (diminution du nombre d'attributs).

11.2 MARIO au niveau des domaines

La méta-approche a été développée pour travailler avec des groupes d'orthologues. Or, il est connu que certains gènes peuvent subir des événements de fusion, rendant l'évaluation du groupe d'orthologues associé discutable. Il a de ce fait été proposé de non plus travailler sur des groupes d'orthologues, mais sur des **domaines orthologues**. La méta-approche peut être appliquée sur les domaines si ceux-ci ont été préalablement définis. Pour cela, il faudrait donner en entrée du programme les protéines de chaque espèce divisées en domaines (format fasta, une séquence par domaine d'une protéine donnée) ainsi que les groupes d'orthologues de chaque domaine fournis par des méthodes inputs. La méthodologie de la méta-approche serait alors inchangée. Cependant, le paramètre de longueur de l'alignement demandé entre un profil HMM et une séquence non regroupée devra certainement être réévalué.

11.3 Prise en compte de la synténie

Si les gènes d'un groupe sont prédits orthologues et qu'ils présentent le même environnement dans plusieurs espèces alors cela renforce cette prédiction. La synténie est un paramètre peu conservé à l'échelle de l'ensemble des règnes. Cependant, elle peut être utile dans le cas de l'application de l'approche à un sous ensemble d'organismes appartenant à un clade restreint.

Dans notre étude, la prise en compte de la synténie pourrait être intégrée à différentes étapes.

- Dans le cas de l'utilisation de résultats obtenus avec des méthodes peu stringentes, on pourrait imaginer un post traitement sur les intersections (étape 3 de la méthode) afin de ne conserver que les orthologues prédits présentant le même environnement génique.

- Dans le cas où l'utilisateur souhaiterait obtenir un groupe de top-orthologues (c'est-à-dire des orthologues ayant conservé le même environnement génique) on pourrait imaginer un post-traitement des groupes d'orthologues finaux afin de les extraire.

L'évaluation de la conservation de la synténie se fait sur l'observation des orthologues entourant un gène donné. Il faut donc pour cela avoir prédit les orthologues de ces organismes. Ainsi ces pré- et post–traitements auraient une qualité dépendante de la qualité de la méthode utilisée pour prédire l'orthologie.

11.4 Meta-approche : recycler les connaissances

L'utilisation d'une méta-approche peut paraître fastidieuse. Elle implique d'avoir obtenu au préalable la prédiction de groupes d'orthologues par d'autres méthodes. Ainsi, elle pourrait paraître longue à mettre en œuvre.

Ces dernières années, la communauté *Quest for Orthologs* , a travaillé sur la mise en place de protéomes ainsi que de formats standards (Dessimoz et al., 2012). Ainsi, la base de données "Proteine reference database" référence des protéomes avec lesquels la communauté est encouragée à travailler. Cette base est mise à jour tous les ans. De ce fait, plusieurs méthodes utilisent ces protéomes et fournissent en ligne leurs résultats (méthodes et résultats répertoriés sur les sites de QFO).

L'existence de protéomes de référence ainsi que l'homogénéité des formats de sauvegarde des groupes d'orthologues induisent une comparaison aisée des résultats de plusieurs méthodes sur ce jeu de données de référence.

De ce fait, notre méta-approche est une manière de recycler le temps de calcul dépensé pour obtenir les groupes initiaux afin de prédire des groupes de plus grande qualité.

III Caractérisation d'un groupe de champignons en fonction de profils phylogénétiques

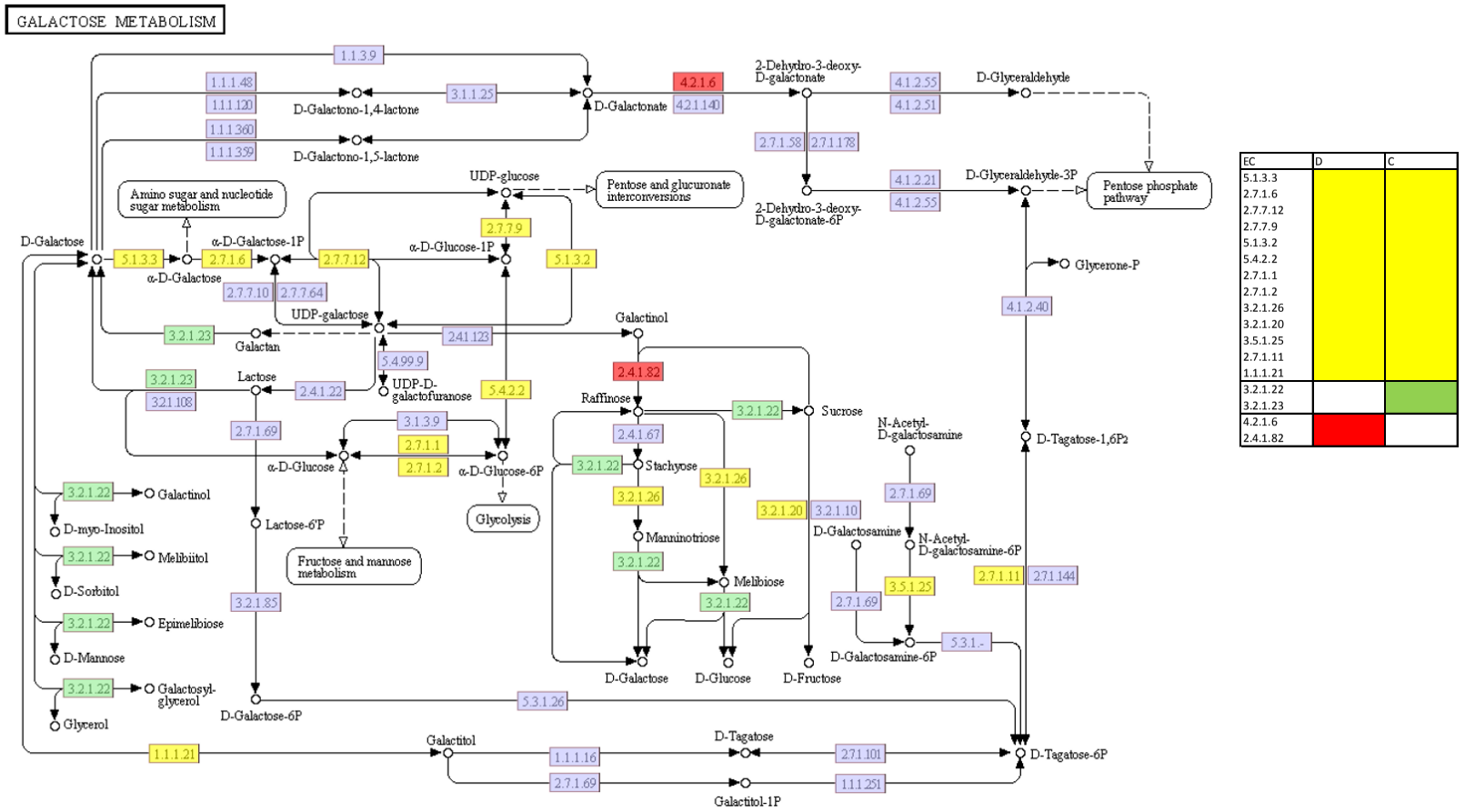


FIGURE 11.7 – *Comparaison de la voie du métabolisme du galactose chez Debaryomyces hansenii (D) et Cryptococcus neoformans (C). Les activités enzymatiques présentes chez les deux organismes sont figurées en jaune, celles spécifiques à Cryptococcus neoformans sont en rouge et celles spécifiques à Debaryomyces hansenii sont en vert. Données issues de FungiPath v1 178 génomes.*

Dans le chapitre précédent, nous avons présenté les étapes d'annotation fonctionnelle des génomes de champignons. La base de données FungiPath contient à la fois les prédictions de groupes d'orthologues, mais aussi les profils enzymatiques de centaines d'organismes. La comparaison des profils enzymatiques de plusieurs espèces peut nous permettre d'améliorer nos connaissances sur l'évolution du métabolisme.

À titre d'exemple, comparons la voie du galactose chez *Debaryomycetes hansenii* (ascomycota) et *Cryptococcus neoformans* (basidiomycota)(figure 11.7). La comparaison de leurs profils permet d'observer que *Debaryomyces hansenii* contrairement à *Cryptococcus neoformans* présente la possibilité d'utiliser le lactose comme source d'énergie. En effet, nous avons prédit chez *Debaryomyces hansenii* la présence de l'EC:3.2.1.23 permettant la transformation du lactose en D-galactose. Cette activité enzymatique est absente de *Cryptococcus neoformans*. De plus le D-galactose est métabolisable par les deux organismes *via* la voie de Leloir ayant pour substrat le D-galactose et pour produit l'alpha-D-glucose 1P(en jaune sur la figure 11.7).

Notre but est de caractériser l'évolution. Pour cela, nous ne pouvons nous contenter de comparer deux espèces. L'obtention de larges quantités de données sur des centaines de génomes de champignons induit la difficulté d'en extraire de l'information pertinente. En effet, il ne s'agit plus ici de la comparaison des profils de deux espèces comme en figure 11.7 mais de la comparaison systématique des profils enzymatiques et phylogénétiques de centaines d'organismes.

L'application de méthodes d'apprentissage est une solution possible pour résoudre ce problème. Nous cherchons à comprendre les mécanismes d'évolution du métabolisme. Quelles sont les voies les plus impactées ? Comment évoluent les activités enzymatiques ? Comment apparaissent de nouvelles voies ? Pour cela, nous proposons d'appliquer des méthodes d'apprentissage supervisé sur les profils phylogénétiques afin de caractériser la taxonomie.

Chapitre 12

Analyse des profils phylogénétiques : état de l'art des méthodes existantes

Les profils phylogénétiques s'intègrent dans un ensemble de méthodes dites de **contexte génomique**. Ces méthodes ont été développées afin de permettre l'inférence de fonctions de gènes pouvant ne pas présenter de similarité de séquences. Elles permettent la prédiction de protéines fonctionnellement liées uniquement à partir de données pouvant être extraites de manière automatique à partir d'un génome annoté structurellement. Pour cela, ces méthodes se basent sur la reproduction de mêmes schémas entre différentes espèces.

Qu'appelle-t-on des protéines 'fonctionnellement liées' ? Deux protéines peuvent être dites **fonctionnellement liées** pour plusieurs raisons. Elles peuvent interagir *via* la formation de complexes protéine-protéine (de nature stable ou non), être impliquées dans la même voie métabolique ou la même voie de signalisation ou encore être exprimées dans la cellule sous les mêmes conditions environnementales. Les méthodes de contexte génomique permettent généralement d'identifier des associations fonctionnelles sans pour autant être capable d'identifier le type d'interaction.

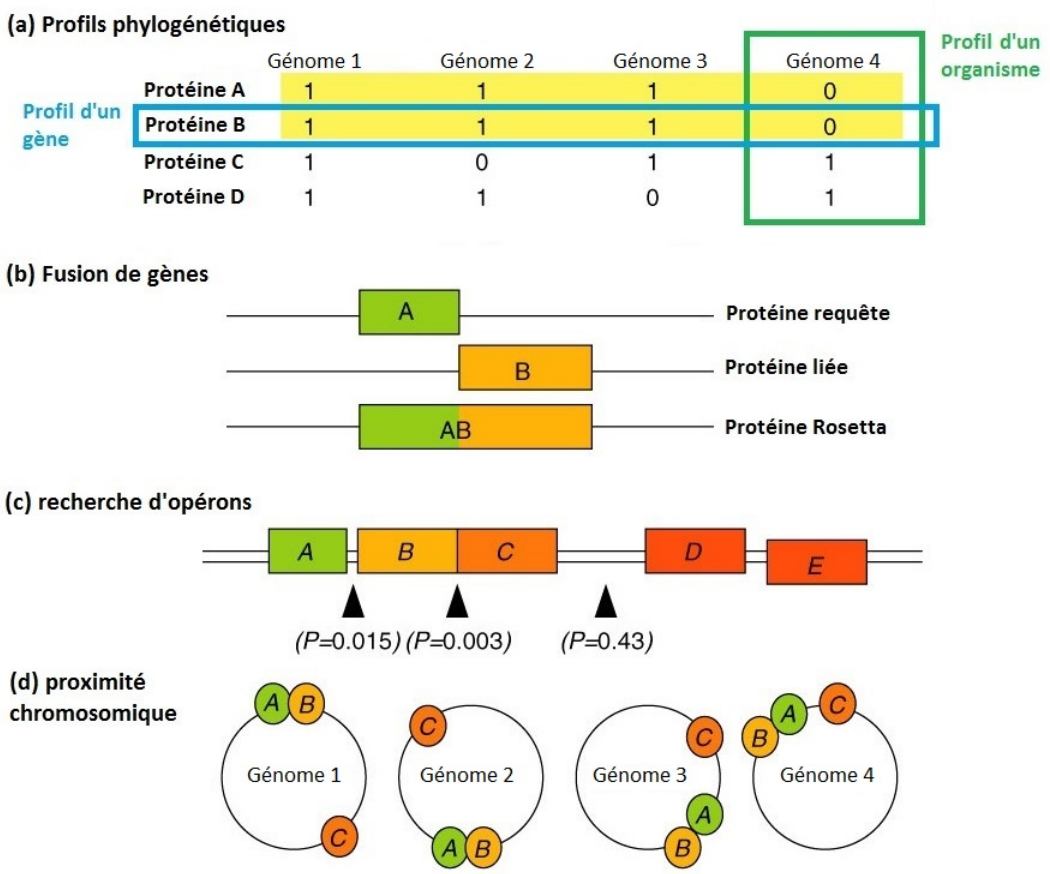


Figure 12.1 – *Exemple des quatre méthodes dites de contexte génomiques. Un profil phylogénétique encadré en bleu. Un profil enzymatique est encadré en vert. Figure adaptée de (Bowers et al., 2004)*

Il existe quatre types principaux de méthodes de contexte génomique (Ferrer et al., 2010):

(a) Les méthodes basées sur les **profils phylogénétiques** (Pellegrini et al., 1999) utilisent les motifs d'occurrence d'un gène dans différentes espèces (voir figure 12.1.a). Le profil phylogénétique d'un gène G donné est un vecteur binaire encodant la présence (1) ou l'absence (0) d'une séquence homologue à G dans une liste de génomes de référence. Les profils phylogénétiques de l'ensemble des gènes d'un génome donné peuvent être vus comme une matrice. Cela permet de définir le profil d'un organisme (une colonne sur la figure 12.1.a) ainsi que le profil d'un gène (une ligne sur la figure 12.1.a). L'étude des profils phylogénétiques est basée sur l'hypothèse que deux gènes impliqués dans la même fonction ont tendance à être conservés ou délétés conjointement. Différentes mesures de similarité des profils phylogénétiques ont été proposées tels que par exemple l'utilisation de l'information mutuelle, du coefficient de corrélation de Pearson, du coefficient de Jaccard ou encore de la *weigthed hypergeometric p-value with runs* (Cokus et al., 2007). Cette dernière mesure utilise de l'information basée sur la taxonomie des espèces. C'est également celle ayant permis les meilleures prédictions de fonctions protéiques (Ferrer et al., 2010).

(b) Les méthodes basées sur **la fusion de gènes** (Enright et al., 1999) partent de l'hypothèse que deux gènes fusionnés dans certains génomes ont une forte probabilité d'être impliqués dans la même fonction. On parle de gène "Pierre de Rosette" (voir figure 12.1.b).

(c) Les méthodes basées sur la recherche **d'opérons** (voir figure 12.1.c) utilisent le fait que les gènes co-localisés sur le même opéron ont de plus grandes chances d'être fonctionnellement liés que des gènes ne l'étant pas.

(d) les méthodes basées sur la conservation d'une **proximité chromosomique**

(synténie) (voir figure 12.1.d) (Dandekar et al., 1998, Overbeek et al., 1999, Pellegrini et al., 2001, Yanai and DeLisi, 2002, Bowers et al., 2004), prennent en compte la distance "physique" entre des gènes sur chaque génome. Cette distance est calculée comme le nombre de gènes séparant deux homologues plus 1. On calcule ensuite la probabilité jointe que deux gènes soient à une distance inférieure à celles observées sur l'ensemble des génomes. Si des gènes ont une localité proche sur les génomes alors il est probable qu'ils soient fonctionnellement liés.

Les méthodes de recherche de fusion de gènes ainsi que de clusters de gènes ont une couverture limitée, car elles ne permettent pas de prédire la fonction des gènes ne présentant pas ces spécificités. Les quatre types de méthodes ont parfois été combinés (Ferrer et al., 2010) sans apporter cependant une amélioration significative des résultats.

Notre travail a principalement porté sur l'étude des profils phylogénétiques.

12.1 Les profils phylogénétiques et leurs applications

Les profils phylogénétiques ont été utilisés dans de nombreux buts différents (Kotaru et al., 2013). Ils ont historiquement permis d'annoter des protéines (Pellegrini et al., 1999, Vert, 2002, Mikkelsen et al., 2005, Chen and Vitkup, 2006). Ils ont également été utilisés afin de prédire la localisation de protéines au sein de la cellule. Pour cela, les profils phylogénétiques de protéines de localisation inconnue ont été comparés avec ceux de protéines de localisation connue (Marcotte et al., 2000).

La comparaison des profils de différentes espèces permet d'analyser l'évolution de génomes sur la base du contenu en gènes (Vitulo et al., 2007) et ainsi de prédire des génomes ancestraux (Csűös, 2010).

Récemment, ils ont été utilisés afin d'identifier des modules fonctionnels ou évolutifs (Wu et al., 2006, Li et al., 2014) en tenant compte de l'arbre des espèces. La recherche de corrélation entre profils et phénotypes a permis d'associer génotype et phénotype (Antonov and Mewes, 2008, MacDonald and Beiko, 2010). On a par exemple étudié la corrélation entre un profil phénotypique d'organismes séquencés avec les profils phylogénétiques de chaque gène (Cookson, 2003, Jim et al., 2004, Slonim et al., 2006). Ainsi, un gène est prédit associé à un certain phénotype si la similarité entre le profil du gène et celui du phénotype est observée supérieure à un certain seuil. On a également étudié la corrélation de groupes de gènes présents simultanément (*complex phylogenetic profiling*) avec un phénotype donné (Antonov and Mewes, 2008). Dans une autre étude, des règles d'association ont été inférées à partir des profils afin de prédire le lieu de vie de différentes bactéries (MacDonald and Beiko, 2010).

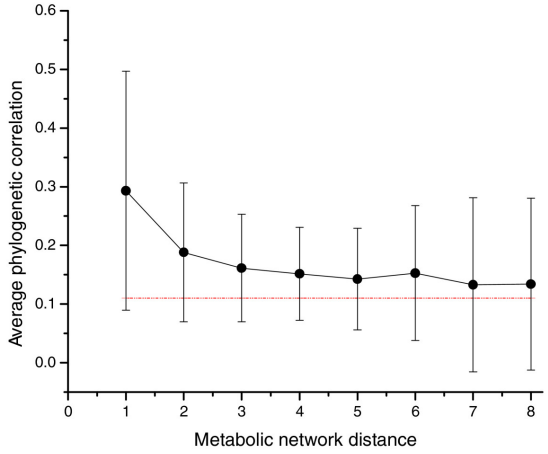


Figure 12.2 – *Corrélation moyenne des profils phylogénétiques entre un gène cible et tous les autres gènes codant pour des enzymes situées à une certaine distance sur le graphe métabolique. Les barres d'erreurs représentent l'écart-type. La ligne rouge présente la corrélation du bruit de fond. Figure tirée de (Chen and Vitkup, 2006)*

Des protéines ont été associées à des enzymes orphelines (Chen and Vitkup, 2006) grâce à la recherche de gènes dont le profil est corrélé avec celui de gènes connus pour exprimer des protéines dont les activités enzymatiques sont à faible distance de celle de l'enzyme orpheline dans le graphe métabolique (voir figure 12.2).

De petites voies d'ARN interférant ont été prédites *via* la recherche de *clusters* enrichis en micro-ARN et en petits ARN interférant au sein des profils phylogénétique des gènes (Tabach et al., 2013)

Les profils ont également été utilisés afin de prédire des interactions physiques entre des protéines (Shoemaker and Panchenko, 2007).

Dans nos travaux, nous comparons les profils phylogénétiques dans le but d'extraire de l'information sur l'évolution du métabolisme. Nous cherchons à caractériser son évolution. Pour cela, nous appliquons des méthodes d'apprentissage supervisé dans le but de caractériser la taxonomie. Nos travaux sont donc à mettre en parallèle avec d'autres portant sur l'évolution (Wu et al., 2006, Li et al., 2014).

Chapitre 13

Apprentissage

Nous nous posons plusieurs questions relatives à l'évolution du métabolisme.

Le métabolisme garde-t-il des traces de son évolution ? Plus précisément, y a-t-il des activités enzymatiques qui conjointement, par leur profil de présence / absence, seraient capables de caractériser des groupes taxonomiques ? Il pourrait s'agir d'activités enzymatiques maintenant la division du groupe en deux sous-groupes.

Si nous trouvons des activités enzymatiques caractéristiques de la taxonomie, leur position au sein des graphes métaboliques peut être intéressante. S'agit-il d'EC situés au niveau de nœuds fortement connectés ? En effet, dans ce cas la présence ou l'absence de ces activités enzymatiques induit un fort changement dans le graphe métabolique (perte ou gain d'un grand nombre de liens). Cette forte différence avec le métabolisme précédent, si elle est conservée, pourrait induire ou renforcer la divergence entre des groupes taxonomiques.

Si plusieurs activités enzymatiques sont capables de caractériser un même groupe taxonomique, quels sont leurs liens les unes avec les autres ? Sont-elles majoritairement regroupées au sein de certaines voies ? Ces voies seraient-elles alors des voies métaboliques plus souvent touchées par des évolutions ? Il est possible que ces voies majoritairement touchées par l'évolution, si elles existent, diffèrent en fonction

du groupe taxonomique étudié et qu'il n'existe pas de règle générale. Nous devrons donc garder cela à l'esprit durant nos analyses.

Nous pouvons également nous demander si ces EC caractéristiques forment un sous-graphe connexe au sein du graphe métabolique. Leur connexité pourrait appuyer une notion d'évolution du métabolisme par l'ajout ou la perte de modules fonctionnels (von Mering et al., 2003).

Afin de répondre à l'ensemble de ces questions, la première étape consiste en la sélection d'activités enzymatiques caractérisant des groupes taxonomiques. Pour cela, nous avons choisi d'appliquer des méthodes de fouille de données.

13.1 Définitions

Nous travaillons sur un ensemble d'**exemples** qui sont l'ensemble de nos champignons.

Nous connaissons les groupes taxonomiques auxquels appartiennent ces champignons. Cette taxonomie reflète l'évolution. C'est cette taxonomie que nous allons chercher à caractériser. On parle de **concept cible**. Un concept cible peut avoir plusieurs valeurs que l'on appellera **classes**. Par exemple, 'ascomycota' est une classe possible du concept cible 'groupe taxonomique de champignons'. Nous traitons les groupes taxonomiques comme des concepts binaires. Un champignon peut appartenir ou non à un groupe taxonomique. Nous avons de ce fait codé chaque concept cible correspondant à un groupe taxonomique par '1' si le champignon appartient au groupe taxonomique analysé ou '0' si le champignon appartient à un autre groupe.

Nous connaissons pour ces champignons un ensemble de caractéristiques (ou **attributs**) qui sont la présence ou l'absence d'activités enzymatiques chez ces organismes. Nous cherchons à caractériser les groupes taxonomiques en fonction de la valeur des attributs. Plusieurs activités enzymatiques peuvent avoir la même

distribution au sein des champignons. Lorsque c'est le cas, nous n'avons conservé qu'un de ces profils (annoté avec l'ensemble des EC concernés).

Nous recherchons ici des synapomorphies (caractère dérivé partagé par plusieurs taxons) enzymatiques caractérisant un groupe taxonomique.

13.2 Fouille de données : présentation des méthodes

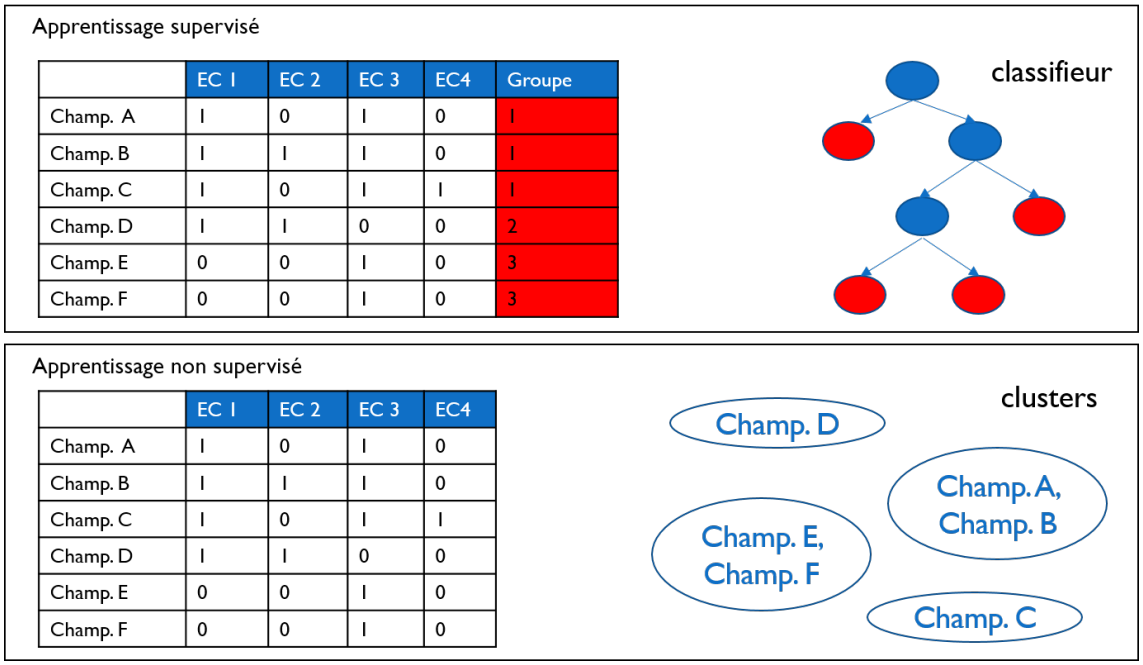


Figure 13.1 – *Illustration d'apprentissage supervisé et non supervisé. A) Arbre de décision (apprentissage supervisé). Les données sont composées d'exemples (en blanc) d'attributs (les colonnes) et d'un concept cible (en rouge). L'arbre de décision (à droite). B) Clustering : à gauche les données (pas de classe), à droite une représentation des clusters.*

De manière générale, les méthodes d'apprentissage peuvent être de deux types : supervisé ou non supervisé.

L'objectif de la classification est d'identifier les classes auxquelles appartiennent des objets à partir de traits descriptifs (attributs). Les méthodes d'**appren-**

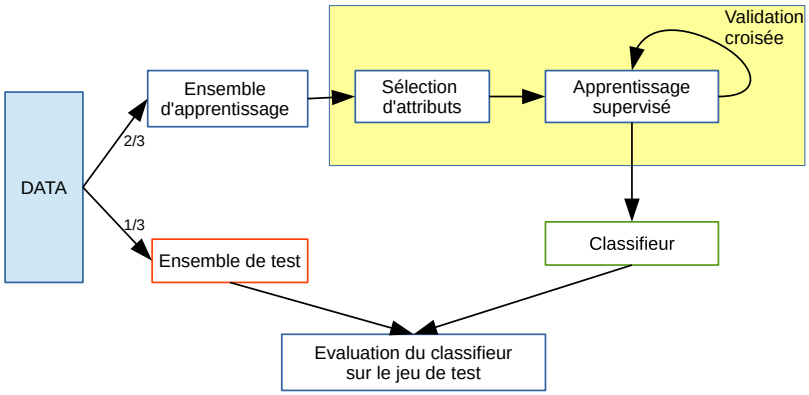


Figure 13.2 – *Protocole d'apprentissage supervisé.*

tissage supervisé cherchent à obtenir un moyen d'inférer des classes prédéfinies (dans notre cas un groupe taxonomique) en fonction de combinaisons d'attributs (nos activités enzymatiques) (voir la figure 13.1). Elles fournissent en sortie un classifieur qui permet de prédire la classe d'un nouvel exemple. Les méthodes d'apprentissage supervisé nécessitent d'avoir un concept cible pour lequel on dispose d'exemples de chaque classe.

Dans le cadre de cette thèse, notre problème pourrait se définir ainsi : "Étant donné un ensemble de champignons pour lesquels on connaît la taxonomie, quels sont les modèles capables de prédire la taxonomie d'autres champignons ?". Le but de cette approche est de trouver des règles capables de définir des groupes taxonomiques. Nous souhaitons ensuite pouvoir analyser ces règles afin d'en extraire de l'information sur l'évolution du métabolisme.

La recherche de classifieurs peut se faire de différentes manières. Elle peut être basée sur des hypothèses probabilistes (classification de Bayes (John and Langley, 1995)), sur des notions de proximité de l'exemple à classer par rapport aux exemples d'apprentissage (k plus proches voisins (Aha and Kibler, 1991)) ou encore rechercher une structure (arbre de décision (Quinlan, 1993), réseaux de neurones (Rosenblatt, 1957)).

Sur la figure 13.2 sont représentées les différentes étapes classiques de l'apprentissage supervisé. Il s'agit basiquement de diviser le jeu de données en deux sous–ensembles, un ensemble destiné à apprendre le classifieur (ensemble d'apprentissage) et un ensemble destiné à son évaluation (ensemble de test). L'évaluation se fait en comparant les classes prédites sur le jeu de données test avec les classes connues de ces exemples.

En bioinformatique, l'apprentissage supervisé est utilisé pour répondre au besoin d'analyse de grands jeux de données obtenus par exemple *via micro-array*, spectrométrie de masse, ou séquençage de nouvelle génération (Larrañaga et al., 2006).

Les méthodes d'apprentissage supervisé se basent généralement sur une hypothèse d'indépendance des attributs (pour nous les *EC numbers*). Dans le cas présent, les activités enzymatiques interagissent au sein du métabolisme. Cette hypothèse d'indépendance n'est donc pas vérifiée. Elles sont plus efficaces si le nombre d'exemples (pour nous de champignons) est plus grand que le nombre d'attributs. Le problème d'un nombre d'exemples faible par rapport au nombre d'attributs tient au fait que l'intérêt d'un attribut par rapport à un autre peut parfois être plus difficilement discriminé. Avec au plus 174 exemples pour plus de 1000 profils différents, nous ne sommes donc pas dans des conditions optimales. Cependant, il s'agit de profils reflétant un rôle biologique. On espère que l'information portant sur l'évolution au sein de ces profils est suffisamment forte pour ne pas être impactée par le bruit possiblement induit par le grand nombre de profils. De plus, on testera l'application de méthodes de sélection d'attributs (type filtre) afin de retirer les attributs non pertinents. Cependant, du fait de ces limitations, il faudra être prudent dans l'analyse et la validation des classifieurs obtenus.

Il nous faut trouver un bon compromis entre la minimisation des erreurs de classification et la complexité du modèle de classification. Plus que la recherche de

classifieurs, nous cherchons à comprendre les mécanismes biologiques liés à l'évolution ayant induit l'obtention de ces classifieurs. Nous recherchons donc des méthodes de classification permettant d'obtenir des classifieurs interprétables. C'est le cas des arbres de décision et des règles de classification.

Le second type de méthode d'apprentissage est appelé apprentissage **non supervisé**. Il désigne un ensemble de méthodes ayant pour but de trouver la typologie existante caractérisant un ensemble d'observations (nos champignons) en fonction d'un ensemble de caractéristiques (présence d'activités enzymatiques). Elles n'utilisent pas de classe préétablie (voir la figure 13.1).

Parmi elles, les méthodes de type *clustering* recherchent au sein des données des sous-ensembles partageant des similitudes. Le but est de regrouper les individus dans des classes, chacune la plus homogène possible et, entre elles, les plus distinctes possibles. Ces méthodes redéfinissent les classes (groupes d'exemples homogènes) en même temps qu'elles les caractérisent. Un second type de méthode non supervisé, les règles d'associations, recherche des co-occurence entre plusieurs attributs.

Ces méthodes ont été utilisées entre autres pour prédire la fonction de protéines sur la base des profils phylogénétiques et cela parce que des groupes de protéines avec des profils similaires peuvent correspondre à des protéines partageant la même fonction (Pellegrini et al., 1999). Le *clustering* a également servi à suggérer les cibles de médicaments dont le mécanisme d'action était incertain (Perlman et al., 2004). Dernièrement, l'application d'une méthode de *clustering* sur des profils a permis de construire des chaînes de Markov cachées caractéristiques de voies métaboliques données et, grâce à elles, de rechercher de nouvelles protéines potentiellement impliquées dans ces voies (Li et al., 2014).

Notre but est d'apprendre de l'information compréhensible sur le métabolisme à partir des profils enzymatiques des champignons. Nous souhaitons caractériser la taxonomie, nous avons donc un concept cible. De ce fait, nous avons focalisé

notre travail sur l'apprentissage supervisé. Afin de pouvoir interpréter les classifieurs nous avons choisi de nous concentrer sur les arbres de décision et les règles de classification. Ces deux types de classifieurs fournissent des règles de types 'SI combinaison d'activités enzymatiques présentes/absentes ALORS groupe taxonomique prédit'.

13.2.1 Méthodes de classifications interprétables

EC	Fungi A	Fungi B	Fungi C	Fungi D	Fungi E	Fungi F
EC1						
EC2						
EC3						
EC4						
Taxonomic groups	1	1	1	2	3	3

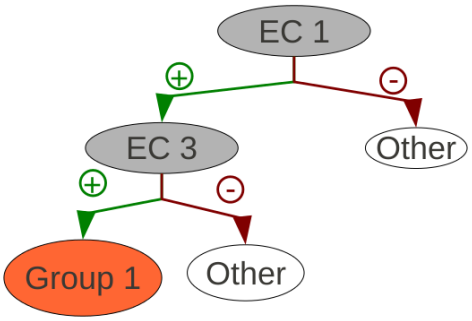


Figure 13.3 – *Jeu de données et arbre de décision associé. Classe : groupe taxonomique 1, attributs : EC. Sur le tableau, la présence d'une activité enzymatique est signifiée par un fond jaune, son absence par un fond blanc. Sur l'arbre, les nœuds sont représentés par des disques gris, et les feuilles par des disques blancs ou orange. L'arbre ci-dessus présente trois règles : SI l'EC1 et l'EC3 sont présents alors le champignon appartient au groupe 1. SI l'EC1 est absente alors on n'appartient pas au groupe 1. Enfin, si l'EC1 est présent, mais que l'EC3 est absent alors on n'appartient pas au groupe 1.*

Les méthodes de type classification permettent de prédire la classe d'un nouvel exemple.

Les arbres de décision impliquent des tests sur des attributs particuliers

(voir figure 13.3). Leur but est de diviser successivement et le plus efficacement possible les exemples de l'ensemble d'apprentissage, par des tests définis à l'aide d'attributs. Les nœuds de l'arbre représentent les tests. Les feuilles sont les concepts appris. Seule une partie de l'ensemble des attributs est sélectionnée pour faire partie de l'arbre. Les arbres de décision sont des procédures de classification interprétables par l'utilisateur. Ce type d'approche produit des arbres dont les branches sont des règles du type 'Si le champignon possède les activités enzymatiques A et B alors il appartient au groupe taxonomique testé'.

D'autres algorithmes prédisent des règles de classification. Une règle est composée d'une série de tests aboutissant, s'ils sont tous vrais, à une certaine classe du concept cible. Des règles de classification peuvent être déduites d'un arbre de décision.

Comme souvent, plusieurs méthodes ont été proposées afin d'extraire les règles ou arbres de classification les plus pertinents. Nous n'avons pas d'*a priori* sur la méthode la plus efficace sur nos données. Nous avons donc choisi de tester trois méthodes que nous décrivons dans les paragraphes suivants.

Arbre de décision

Le but des méthodes produisant un arbre de décision est d'apprendre une fonction discrète à plusieurs valeurs qui approxime le concept cible. Les arbres appris permettent d'obtenir des règles de décision. Une branche (suite de nœuds qui va de la racine vers une feuille) donnera une règle.

Principe : Les instances sont triées par des tests effectués aux nœuds de l'arbre, ces tests portent sur les attributs décrivant les instances. Sur nos données, il s'agit donc de trier les champignons en fonction de leur groupe taxonomique uniquement grâce à des combinaisons d'activités enzymatiques devant être présentes ou absentes chez ces organismes. À chaque branche sortant d'un nœud étiqueté

avec une activité enzymatique correspond une valeur possible pour ces activités enzymatiques : 'présente' ou 'absente'. Un arbre de décision va représenter une disjonction de conjonctions de paires (activité enzymatique, présence) (voir figure 13.3).

Un exemple d'algorithme permettant d'obtenir un arbre de décision est l'**algorithme C4.5** (Quinlan, 1993). La première étape de cet algorithme consiste à sélectionner l'attribut (le profil phylogénétique) permettant le meilleur gain d'information [1].

Cet attribut est placé au niveau d'un nœud et une branche est créée pour chaque valeur possible de l'attribut (dans notre cas activité enzymatique présente ou absente). Le jeu de données est alors divisé au niveau de chaque nouvelle branche afin de respecter la valeur de l'attribut sélectionné comme nœud. Il s'agit d'une recherche d'optimum local.

Le classifieur peut-être appris sur l'ensemble des données. Cependant, dans ce cas, il est possible que le classifieur soit tellement spécifique du jeu de données d'apprentissage qu'il présentera de mauvaises prédictions sur un nouveau jeu de données. Il s'agit du phénomène de sur–apprentissage.

Afin d'éviter le sur-apprentissage deux choix sont possibles. Le **pré-élagage** consiste à arrêter l'expansion de l'arbre lorsqu'une classe est suffisamment majoritaire. Le **post-élagage** consiste à construire l'arbre jusqu'au bout et à retirer ensuite certains sous-arbres caractérisant une trop petite proportion des exemples.

L'algorithme C4.5 utilise le post-élagage, l'ensemble d'apprentissage est uti-

1. Soit E un ensemble d'exemples, A l'attribut et x la classe. Le calcul du gain d'information relatif à l'attribut A (entre A et x) se fait de la manière suivante : $GI(E,A) = H(E) - \sum_{v \in Valeurs(A)} \frac{|E_v|}{|E|} H(E_v)$ avec $H(E)$ l'entropie de l'ensemble des exemples E et $H(E_v)$ est l'entropie des exemples présentant la valeur v pour l'attribut A ($E_v \in E$). $H(E) = -\sum_c p(A_c) \ln(p(E_c))$, avec $p(E_c)$ la proportion d'exemples appartenant à la classe c. Pour plus de détail voir la partie méthodes de type filtre.

lisé pour élaguer l'arbre. Le critère d'élagage est basé sur une heuristique permettant d'estimer l'erreur réelle sur un sous-arbre donné. Bien qu'il paraisse peu pertinent d'estimer l'erreur réelle sur l'ensemble d'apprentissage, il semble que la méthode donne des résultats corrects.

L'utilisation des arbres de décision en biologie est en croissance constante (voir figure 13.4). Les arbres de décision ont l'avantage de permettre l'obtention d'un classifieur facilement interprétable.

Règle de décision

Tout comme les arbres de décision, les règles de décision sont interprétables. Parce qu'il existe plusieurs méthodes différentes de construction de règles de décisions pouvant mener à des classifieurs différents, nous avons choisi d'en tester deux parmi les plus courantes.

L'algorithme **PART** (Frank and Witten, 1998) construit un ensemble de règles *via* une méthode *separate–and–conquer*. Pour chaque règle un arbre de décision complet est construit avec l'algorithme C4.5. La meilleure règle de l'arbre est alors extraite et sélectionnée comme règle par PART. On apprend ensuite un nouvel arbre à partir des exemples non couverts par la première règle. Les résultats obtenus par PART peuvent différer de ceux obtenus par C4.5 si plus d'une règle est nécessaire pour classer les exemples.

L'algorithme **Repeated incremental pruning to produce error reduction** (RIPPER) (Cohen, 1995) permet d'obtenir une collection de règles de type 'SI ... ALORS ....'. Il présente deux phases pour la construction d'une règle, une phase d'expansion et une phase d'élagage. Durant la phase d'expansion, des attributs sont ajoutés à la règle de manière gloutonne jusqu'à ce que la règle soit parfaite. Durant la phase d'élagage, les attributs les moins pertinents sont retirés de la règle. RIPPER utilise le gain d'information pour évaluer la pertinence d'un attribut. Les règles

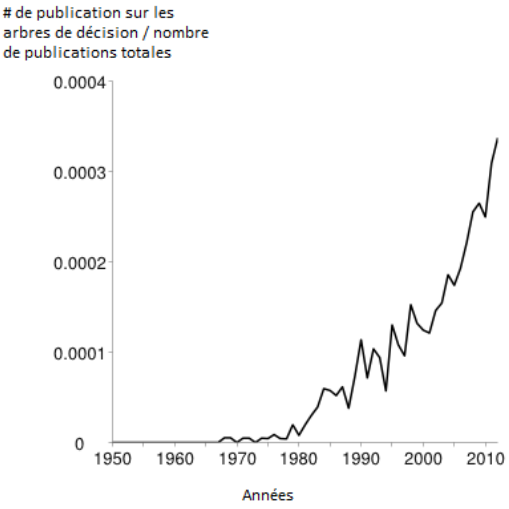


Figure 13.4 – *Proportion dans pubmed de publications portant sur l'apprentissage supervisé de type arbre (nombre de publications portant sur l'apprentissage de type arbre divisé par le nombre total de publications annuel). Abscisse : années. Ordonnées : nombre de publications normalisé par le nombre de publications par an . Figure réalisée grâce à l'outil ML-Trends (Palidwor and Andrade-Navarro, 2010) et la recherche des mots clés '"classification tree" or "decision tree" or "random forest" dans le titre ou le résumé des publications'.*

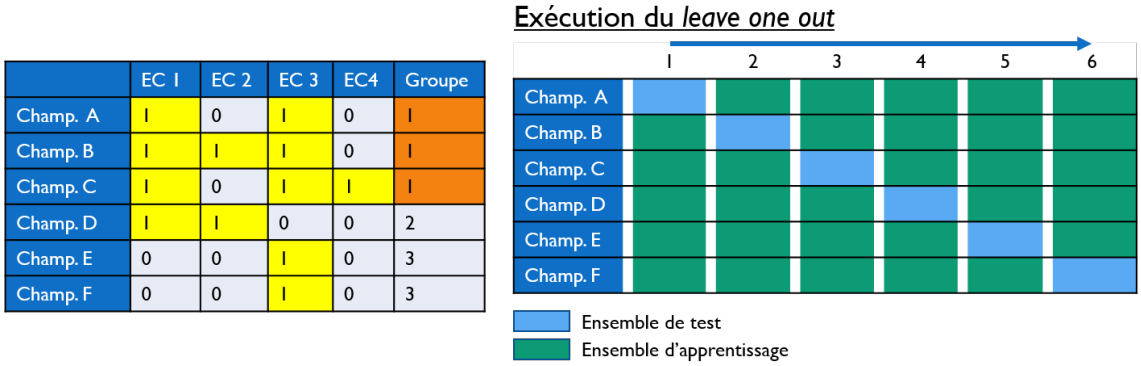
suivantes sont apprises sur les données non couvertes par les règles précédentes.

13.2.2 Méthodes d'évaluation des classifieurs obtenus

Prédire un classifieur et l'évaluer sur le même jeu de données induit une surévaluation de sa qualité. C'est une conséquence du 'sur-apprentissage' ou *overfitting*. De manière à éviter ce phénomène, on divise classiquement le jeu de données en deux parties, un ensemble d'apprentissage sur lequel est appris le classifieur et un ensemble de test sur lequel il est évalué (voir figure 13.2).

Validation croisée et *leave one out*

La validation croisée consiste en la division du jeu de données en deux ensembles (ensemble d'apprentissage et ensemble de test) de manière répétée. Étant donné un échantillon d'exem-ples étiquetés, les exemples seront à tour de rôle dans l'ensemble de test ou dans l'ensemble d'apprentissage. Pour ce faire, les champignons sont divisés en k sous-ensembles. Les champignons du groupe taxonomique que l'on cherche à caractériser sont distribués aléatoirement et de manière proportionnelle dans chacun des k paquets à peu près égaux en taille. Chaque paquet sert alors tour à tour d'ensemble de test et les k-1 autres paquets servent d'ensemble d'apprentissage. Il faut alors apprendre le classifieur sur toutes les données d'apprentissage. Ainsi, après k itérations, chaque champignon aura été placé une fois dans l'ensemble de test. Le taux d'erreur est calculé successivement sur chaque paquet test. L'estimation de l'erreur totale se fait en moyennant les taux d'erreurs obtenus sur les k tests. Cette validation n'indique pas quel classifieur choisir parmi les k classifieurs créés. Elle donne une approximation du taux d'erreurs du classifieur obtenu sur l'ensemble des données.



	EC 1	EC 2	EC 3	EC4	Groupe
Champ. A	1	0	1	0	1
Champ. B	1	1	1	0	1
Champ. C	1	0	1	1	1
Champ. D	1	1	0	0	2
Champ. E	0	0	1	0	3
Champ. F	0	0	1	0	3

FIGURE 13.5 – *Exemple leave one out avec n=6 exemples.*

Le **leave one out** est un type de validation croisée où l'ensemble de test se limite à 1 exemple. Le classifieur est appris et évalué successivement sur les $(n-1)$ exemples (voir figure 13.5). Ce type particulier de validation croisé est à privilégier dans le cas où il y a un faible nombre d'exemples, ce qui est le cas de nos données de travail.

Les différentes mesures d'évaluation

	Classe prédite		
		positif	negatif
classe réelle	positif	VP	FN
	negatif	FP	VN

TABLE 13.1 – *Matrice de confusion*

Afin de sélectionner les meilleurs classifieurs possible il est nécessaire de les évaluer. Un moyen de représenter la qualité des résultats est d'utiliser une matrice de confusion (voir le tableau 13.1). Cette matrice présente en ligne le nombre d'occurrences de la classe réelle et en colonne le nombre d'occurrences d'une classe prédite.

Il existe alors quatre cases correspondantes à quatre cas possibles : les exemples positifs correctement prédits (vrais positifs ou VP) les exemples négatifs correctement prédits (vrais négatifs ou VN), les exemples positifs manqués par le classifieur (faux négatifs ou FN) et les exemples négatifs prédits comme positifs (faux positifs ou FP). Ces quatre cas permettent la définition de différentes mesures d'évaluation. Nous présentons ici les mesures les plus courantes ainsi que leur intérêt.

La **précision** représente le pourcentage de prédictions correctes associées à la classe positive. Précision $= \frac{VP}{VP+FP}$.

Le **rappel** représente le pourcentage d'exemples d'un certain type (dans notre cas appartenant ou pas au groupe taxonomique testé) correctement prédits. Dans le cas de l'observation des exemples positifs on parlera de sensibilité (sensibilité $= \frac{VP}{VP+FN}$), dans le cas de l'observation des exemples négatifs on parlera de spécificité

La courbe **ROC** (Receiver Operating Characteristic) est une courbe représentant le pourcentage de faux positifs en fonction du pourcentage de vrais positifs (voir figure 13.6).

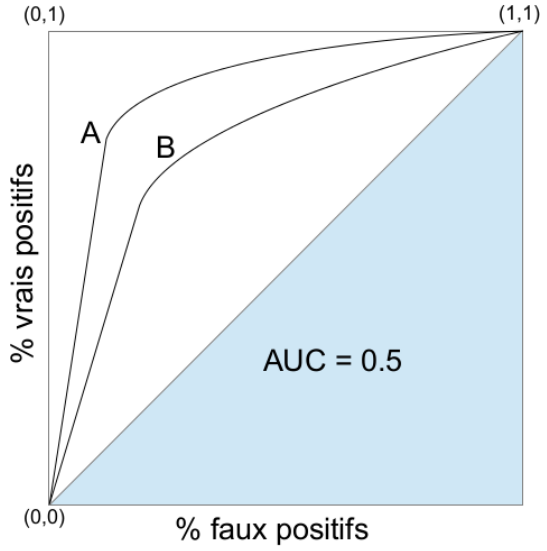


Figure 13.6 – *Aire sous la courbe ROC, figure tirée de (Azé, 2012). La droite allant du point (0,0) au point (1,1) représente la droite obtenue avec un classifieur aléatoire.*

13.3 Arbres de décision et règles de classification appliqués à la caractérisation de la taxonomie

13.3.1 Résultats obtenus sur FungiPath 50 génomes

Nos premiers résultats d'applications de méthodes d'apprentissage supervisé sur des données fongiques ont été obtenus sur les données de FungiPath dans la version 50 génomes (Grossetete et al., 2010). La méthodologie appliquée consistait en l'application de trois algorithmes (ID3, C4.5 et PRISM) sur chaque groupe taxonomique présentant au moins trois individus dans nos données initiales. Les attributs étaient évalués par la mesure du gain d'information. Nous avons ainsi pu obtenir plusieurs classifieurs.

L'algorithme C4.5 propose pour le groupe des pezizomycetes un classifieur faisant intervenir un unique EC : si un champignon possède l'EC:3.1.6.6 alors c'est un pezizomycete, sinon il appartient à un autre groupe. L'application de ce classifieur sur les 48 champignons de cette version de FungiPath aboutit à une erreur : *Ustilago maydis* est prédit comme étant un pezizomycotina. Après alignement des séquences annotées EC:3.1.6.6 et analyse de l'arbre phylogénétique obtenu il semblerait qu'il s'agisse d'une observation d'un transfert horizontal entre un ancêtre des pezizomycotina et un ancêtre d'*Ustilago maydis*.

Les ascomycota et les basidiomycota sont les deux sous-groupes de dikaryomycetes. Les recherches de classifieurs pour ces deux groupes ont abouti à l'obtention de deux classifieurs (trouvés par les trois méthodes) :

- Si le champignon possède l'activité enzymatique 3.4.23.41 alors il s'agit d'un ascomycota, sinon il appartient à un autre groupe.
- Si le champignon possède l'activité enzymatique 3.4.23.5 alors il s'agit d'un basidiomycota, sinon il appartient à un autre groupe.

Les deux activités enzymatiques trouvées ici comme caractérisantes portent sur la transformation et le maintien de la paroi cellulaire. Or, ces deux groupes de champignons sont incapables de fusionner leurs membranes. Cette incompatibilité pourrait être due à une différence de composition/structure de ces membranes. La spécificité de présence de l'EC 3.4.23.41 ou de l'EC:3.4.23.5 pourrait ainsi être en lien avec la différenciation en deux groupes taxonomiques.

Nous avons choisi de tester la perte de l'activité enzymatique 3.4.23.41 chez *Podospora Anserina*. Ce travail s'est fait en collaboration avec l'équipe Génétique et Épigénétique des Champignons (université Paris Diderot).

Dans la version de FungiPath à 50 génomes, *Podospora anserina* présentait deux groupes d'orthologues annotés avec l'activité enzymatique 3.4.23.41. Le premier présentait une unique séquence chez *P. anserina* (HpYPS1), ce groupe ayant au moins un représentant chez l'ensemble des ascomycota excepté chez *tricoderma reesei*, le second présentait deux séquences chez cet organisme (in-paralogue) et ce groupe n'était représenté que parmi 11 ascomycota sur 22. Nous avons dans un premier temps délété le gène appartenant au premier groupe, car un orthologue était présent chez un plus grand nombre d'ascomycota.

Validation expérimentale : délétion de HpYPS1 chez *Podospora anserina*. J'ai réalisé la délétion de HpYPS1 comme décrite dans (Grognet et al., 2012) (voir la figure 13.7). Je l'ai vérifiée par PCR et Southern blot en utilisant le kit "DIG system" de Roche (voir figure 13.8). J'ai suivi le protocole décrit dans Brygoo and Debuchy (1985). La souche transformée est la souche S mus51Δ. Dans cette souche, les recombinaisons ne se font que de façon homologue. Les amorces utilisées pour les délétions suivent la nomenclature suivante : les amorces A + B servent à amplifier la séquence flanquante en 5', les amorces C + D servent à amplifier la séquence flanquante en 3' et les amorces E + F servent à amplifier la cassette de résistance (voir le tableau des amorces 13.2 et la figure 13.7).

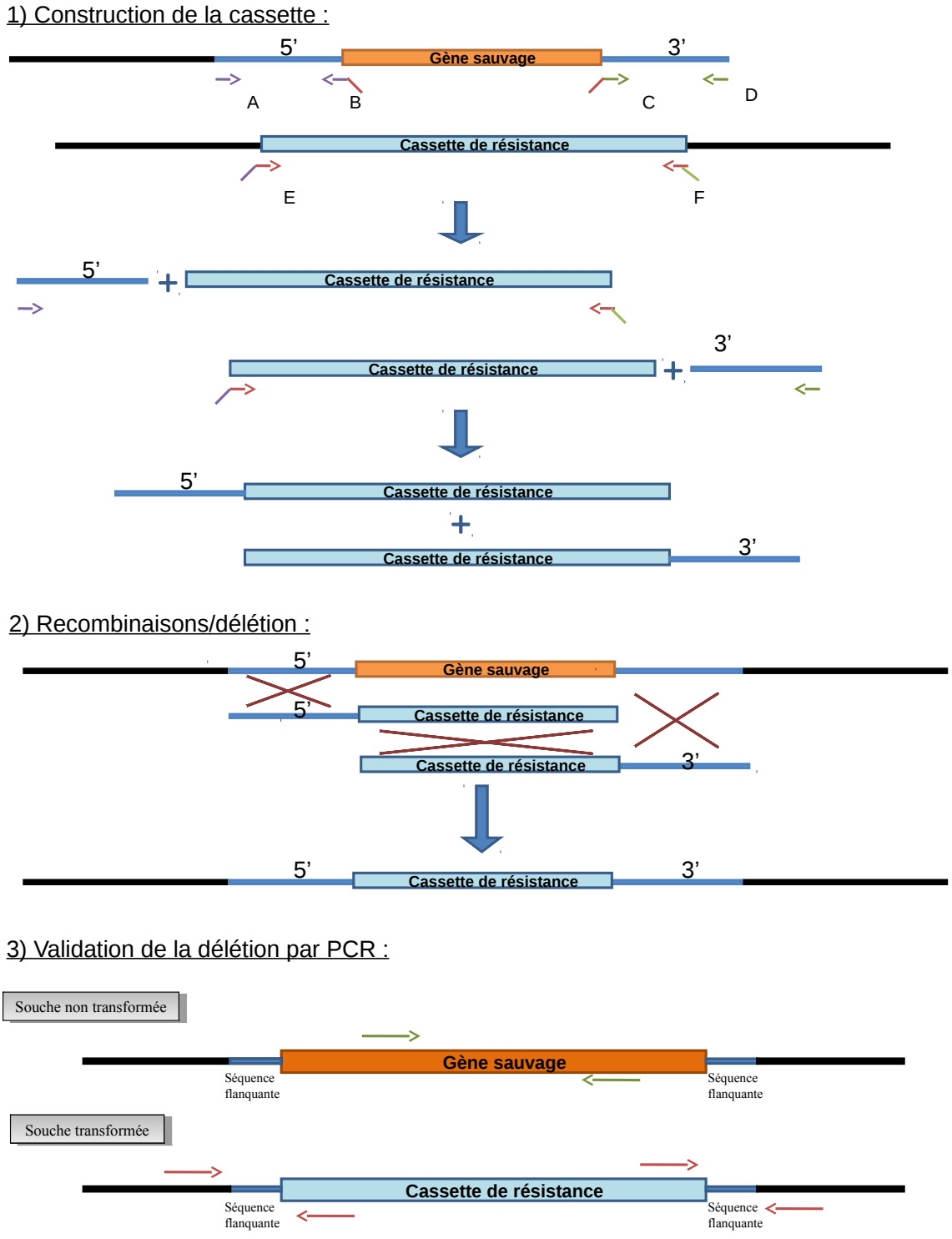


Figure 13.7 – *Schéma de la construction du mutant de délétion et de sa validation par PCR*

Nom de l'amorce	séquence
8310_A	cgcgtacacttgataataccctttgtttcc
8310_B	CTATTTAACGACCCTGCCCTGAACCGgaacagacgttcagaaaaagagcaaaaagc
8310_mk_E	GCTTTTTGCTCTTTTTCTGAACGTCTGTTCcggttcagggcagggtcgttaaatag
8310_mk_F	CTTTCCTATAACCAGCAATGATGCACATCTcatcgaactggatctcaacagcggtaag
8310_C	CTTACCGCTGTTGAGATCCAGTTCGATGagatgtgcatcattgctggttataggaaag
8310_D	tgactcttcaaaagcctaaaagctggctac

Table 13.2 – *Tableau des amorces*

La construction de la cassette de résistance s'est faite en plusieurs étapes. La première consiste en l'amplification par PCR des régions 5' et 3' du gène ainsi que de la cassette de résistance (étape 1 sur la figure 13.7). L'amplification est faite avec des amorces PCR présentant une complémentarité avec la région 5', la région 3' ou la cassette. L'amorce B et l'amorce C présentent une complémentarité respectivement avec la partie 5' et 3' de la cassette de résistance ainsi qu'avec respectivement la région 5' ou 3' du gène sauvage. Les amorces E et F, qui permettent l'amplification de l'amorce, présentent, en plus de leur complémentarité avec l'amorce, une complémentarité avec les régions 5' et 3' du gène. Les amplifications ont été vérifiées par Southern Blot. La seconde étape fut la ligation de la cassette avec la séquence de la région 5' du gène d'une part et celle de la région 3' du gène d'autre part. Cette ligation est faite *via* une PCR et a été vérifiée par Southern Blot.

Nous avons préparé 10 transformants. La délétion de HpYPS1 a été vérifiée *via* digestion de l'ADN, PCR avec sondes marquées et analyse du résultat par Southern Blot. Parmi les 10 transformants, 2 ne présentaient aucune bande correspondant à la présence de HpYPS1 dans le génome (transformants 1 et 9), tous présentaient une bande correspondant à la présence de la région 5' du gène HpYPS1 suivie du début de la cassette HpYPS1 et sept (1,2,5,7,8,9,10) présentait une bande

correspondant à la fin de la cassette de résistance suivie du début de la partie 3' du gène HpYPS1.

Les transformants 1, 5, 8, 9 et 10 ont été croisés (mat + / mat −). De nouveau, une vérification par digestion de l'ADN (par les enzymes AgeI et HindIII) nous a permis d'évaluer la qualité de nos transformants. On montre ainsi (figure 13.8) que les souches 1 et 10 présentent des bandes correspondant à la présence de la cassette de résistance et pas de bande correspondant au gène HpYPS1. Pour ces deux souches, l'insertion a donc eu lieu comme voulu. Les tests phénotypiques ont été réalisés sur ces deux souches.

Nous avons ainsi testé l'incompatibilité végétative, l'interférence hyphale, la présence et la forme du périthèce (fructification) ainsi que la pousse sur différents milieux de culture (dextrine, glucose, cellulose, MYG, papier et bois). Tous ces tests ont donné le même résultat que pour la souche sauvage. Cependant, comme dit dans l'introduction, seul un des deux groupes d'orthologues a été délété. Ces résultats négatifs ne peuvent donc induire de conclusion sur l'effet de la suppression de l'activité enzymatique 3.4.23.41. Ayant mis au point parallèlement une nouvelle version de FungiPath avec plus de génomes, nous avons choisi de ne tester la délétion complète des gènes annotés comme permettant de produire des protéines pourvues de cette activité enzymatique que si ces résultats étaient maintenus après analyse de cette nouvelle version (voir la suite des résultats).

13.3.2 Résultats obtenus sur 174 champignons dont les groupes d'orthologues avaient été prédits avec la méthode FungiPath développée par S. Grossetête

La caractérisation de groupes taxonomiques sur les données de FungiPath version à 178 génomes a été publiée dans les actes de JOBIM 2013 (Pereira et al., 2013). Nous montrons dans cet article que l'application de méthode d'apprentis-

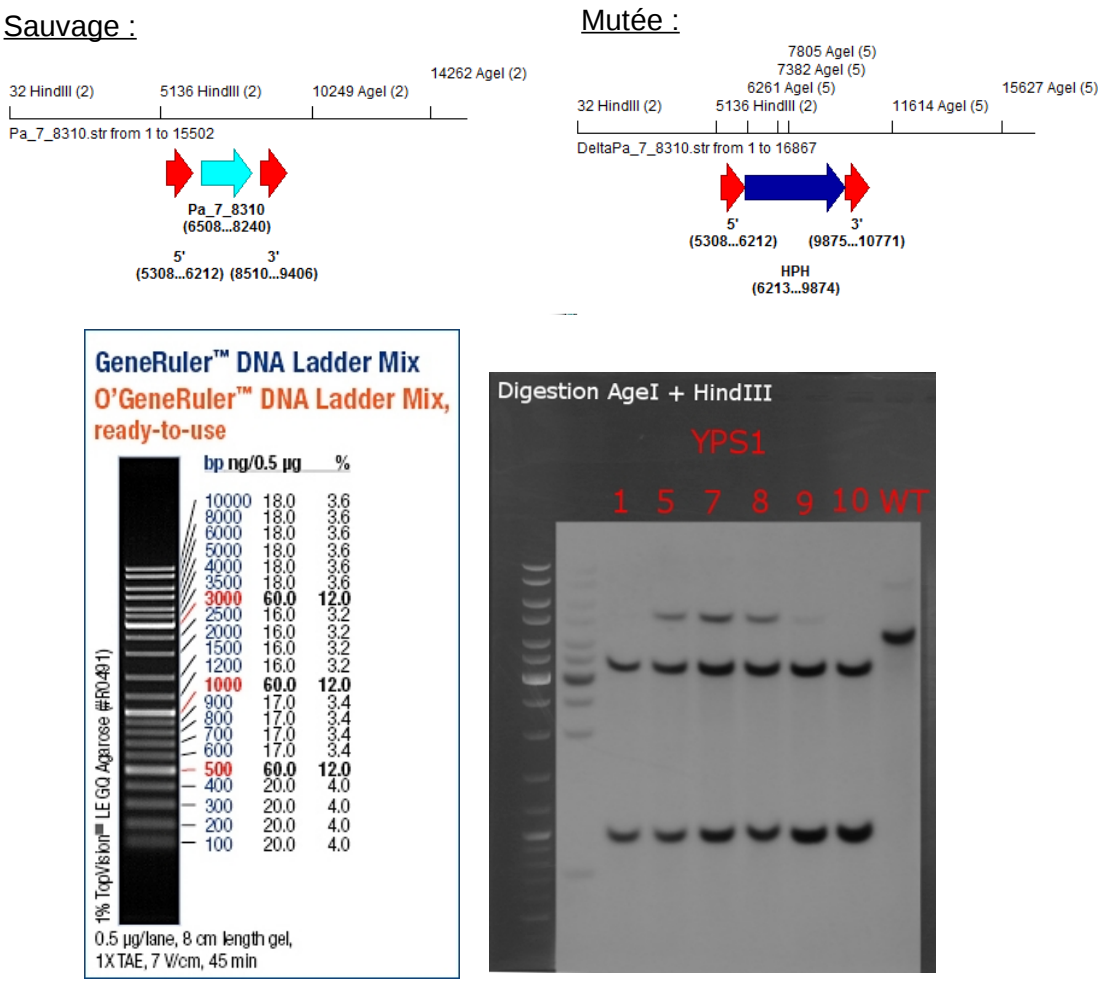


Figure 13.8 – *Southern blot obtenu sur les 10 souches transformées afin de déléter HpYPS1. Les tailles des fragments attendus chez la souche délétée sont de 3709 et 1125 paires de bases et chez une souche sauvage une bande est attendue à 5113 paires de bases. Les transformants 1 et 10, qui ne présentent que les deux bandes (correspondant aux bandes attendues après délétion) sont validés.*

sage sur des profils enzymatiques permet de caractériser des groupes taxonomiques avec un faible nombre d'activités enzymatiques. De plus, nous observons que les cas d'apparente mauvaise prédiction du groupe taxonomique d'un champignon peuvent correspondre à la présence de transferts horizontaux d'un champignon ayant appartenu au groupe taxonomique prédit vers un ancêtre du champignon pour lequel le groupe taxonomique est prédit.

Comparative analysis of phylogenetic profiles for the enzymatic characterization of fungal groups

Cécile PEREIRA[1,2], Jérôme AZÉ[2,3], Alain DENISE[1,2,3], Christine DREVET[1], Christine FROIDEVAUX[2,3], Philippe SILAR[1,4], Olivier LESPINET[1,2]

[1] IGM, UMR 8621 CNRS, Université Paris-Sud, Bât 400, 91405 Orsay Cedex, France
cecile.pereira@igmors.u-psud.fr
[2] LRI, UMR 8623 CNRS, Université Paris-Sud, Bât 650, 91405 Orsay Cedex, France
[3]INRIA AMIB, Saclay, France
4 LIED, FRE, Université Paris Diderot, Sorbonne Paris Cité, Paris, France

Abstract *We try to characterize the evolutionary origin of the enzymatic repertoire of different fungal groups. The characteristics for each of the groups studied are determined through the application of* data mining *method on enzyme profiles previously determined by comparative genomics. Through the presentation of results for taxonomic groups Agaricomycetes and Pezizomycota, we show that the application of supervised learning methods is effective in extracting information from phylogenetic* profiles. *We extract specific enzyme activities combinations for each taxonomic groups covered by our analysis. Our approach also enables us to highlight the existence of probable horizontal gene transfers.*

Keywords Data Mining, fungi, phylogenetic profiles, evolution, enzymes.

Étude comparative des profils phylogénétiques dans le but de définir les spécificités enzymatiques de différents groupes de champignons.

Résumé *Nous essayons de caractériser l'origine évolutive du répertoire enzymatique de différents groupes de champignons. Les caractéristiques de chacun des groupes étudiés sont déterminées grâce à l'application de méthodes de* fouille de données *sur les* profils enzymatiques *préalablement déterminés par génomique comparée. À travers la présentation des résultats obtenus pour les groupes taxonomiques des Agaricomycetes et des Pezizomycota, nous montrons que l'application de méthodes d'apprentissage supervisé est efficace pour extraire de l'information des* profils phylogénétiques. *Notre approche permet également de mettre en évidence l'existence de probables transferts horizontaux.*

Mots-clés Fouille de données, champignons, profils phylogénétiques, évolution, enzymes.

1 Introduction

Les champignons possèdent un vaste répertoire enzymatique leur permettant de dégrader et de synthétiser de nombreux composés organiques. Nous nous intéressons à l'origine évolutive de ce répertoire. Plus précisément, notre but est de caractériser les spécificités enzymatiques de différents groupes de champignons. Ceux-ci sont constitués soit à partir de critères taxonomiques, soit à partir de critères ayant trait aux modes de vie.

La comparaison des voies métaboliques de différents organismes a déjà fait l'objet de plusieurs travaux, ainsi ce type d'analyse a déjà permis de reconstruire des arbres phylogénétiques cohérents, preuve de la persistance d'information évolutive dans ces voies [1], la topologie des voies métaboliques a également été étudiée dans le but de proposer de nouvelles phylogénies [2] enfin, de nouvelles cibles thérapeutiques ont pu être définies grâce à la comparaison des voies métaboliques d'organismes pathogènes et non pathogènes [3]. Cependant aucune étude portant sur la totalité du répertoire enzymatique de plus d'une centaine d'espèces n'a encore été menée à ce jour.

Dans ce travail, nous essayons de déterminer la spécificité du répertoire enzymatique des différents groupes de champignons par apprentissage supervisé à partir de profils enzymatiques préalablement établis par détection des gènes homologues entre 165 espèces de champignons [4] (Tableau supplémentaire 1).

Cette approche nous a permis de définir quelles étaient les activités enzymatiques caractéristiques par exemple des *Agaricomycetes* et des *Pezizomycotina*. Elle a également permis de mettre en évidence de probables transferts horizontaux entre plusieurs des espèces ou groupes étudiés.

2 Méthodologie

2.1 Les données

Nous travaillons à partir de 165 espèces eucaryotes complètement séquencées dont 161 espèces de champignons (Table supplémentaire 1). Les espèces ont été choisies en fonction soit de leur position taxonomique (de manière à échantillonner l'ensemble de la diversité des *Eumycota*), soit de leurs caractéristiques biologiques (de manière à couvrir des modes de vies et d'habitats différents).

La comparaison exhaustive des 1 748 866 protéines constitutives des 165 espèces nous a permis de constituer 139 004 groupes de protéines homologues. Tous les groupes ont été annotés avec le même protocole d'annotation fonctionnelle [4] afin d'éliminer d'éventuels biais liés aux protocoles d'annotation initialement utilisés pour chacun des génomes. Nous obtenons ainsi 12 505 groupes possédant une annotation fonctionnelle de type enzymatique caractérisée par un ou plusieurs *Enzyme Commission* (EC) *numbers* [5]. 1 412 *EC numbers* différents sont utilisés pour définir la fonction de ces 12 505 groupes.

À partir de la distribution des EC numbers nous construisons des profils enzymatiques et des profils phylogénétiques. Le profil enzymatique d'un génome donné est défini par la liste des activités enzymatiques présentes dans ce génome. De la même façon, le profil phylogénétique d'une activité enzymatique donnée est définie par la liste des génomes qui possèdent un gène portant cette activité enzymatique. Ces deux types de profils peuvent être représentés par une matrice (Table 1) à deux dimensions dans laquelle chaque case indique la présence ou l'absence d'une activité enzymatique donnée dans une espèce donnée. Les 1 412 EC *numbers* différents retrouvés parmi les 12 505 groupes d'homologues présentant une activité enzymatique peuvent se répartir en 1 155 profils distincts puisque plusieurs EC *numbers* peuvent présenter un même profil.

Génome \ EC number	1.1.1.1	1.1.1.108	1.1.1.116	1.1.1.138	1.1.1.14	1.1.1.157	1.1.1.158,4.3.1.2
Tuber melanosporum			#				
Arthrobotrys oligospora			#				
Aspergillus nidulans			#				
Talaromyces stipitatus			#				
Penicillium marneffei			#				
Penicillium chrysogenum	*	*	# *	*	*	*	*
Neosartorya fischeri			#				
Aspergillus aculeatus			#				
Aspergillus flavus			#				

Table 1 : Exemple de profils enzymatiques et de profils phylogénétiques. Les génomes sont disposés en ligne et les EC *numbers* en colonne (listes non exhaustives). Lorsque une activité enzymatique donnée est présente chez une espèce donnée la case du tableau est colorée en gris. Le profil enzymatique de *Penicillium chrysogenum* est symbolisé par des '*'. Le profil phylogénétique de l'EC 1.1.1.116 est symbolisé par des '#'.

2.2 Apprentissage supervisé

Nous utilisons des méthodes d'apprentissage supervisé afin d'extraire des informations sur la spécificité du répertoire enzymatique de différents groupes de champignons à partir des profils phylogénétiques.

Il existe un large panel de méthodes d'apprentissage de natures différentes (modèle bayésien, arbres de décision, règles de classification, k-plus proches voisins, *etc*.). Nous avons choisi d'utiliser des approches fournissant des modèles de classification interprétables et à fort pouvoir explicatif tels que des arbres de décision ou des règles de classification. Les arbres de décision sont des approches *top-down*, c'est-à-dire qu'à chaque étape, le triplet (attribut, test, valeur) qui optimise le critère est retenu et deux partitions disjointes sont créées. Les règles utilisent quant à elles une approche *bottom-up*, c'est-à-dire effectuent une généralisation à partir d'un exemple de manière à couvrir le plus possible d'exemples positifs, sans couvrir d'exemples négatifs. Afin d'exploiter les profils phylogénétiques nous combinons l'utilisation des trois algorithmes d'apprentissage supervisés C4.5 [4] (arbre de décision), PART [5] (règle de décision) et RIPPER [6] (règles construites directement) avec un système de vote majoritaire. Le critère d'évaluation choisi pour ces 3 algorithmes est le gain d'information. L'implémentation de ces algorithmes est fournie par la boîte à outils WEKA [7].

Les méthodes d'apprentissage supervisé appliquées sur les champignons décrits par les profils enzymatiques renvoient comme résultat des classifieurs permettant de mettre en évidence les combinaisons d'enzymes spécifiques d'un groupe taxonomique donné. En d'autres termes, elles permettent de mettre en évidence les « synapomorphies enzymatiques » d'un groupe d'espèces. Dans ce but, nous définissons les exemples positifs comme étant l'ensemble des génomes appartenant à un groupe donné, et les exemples négatifs comme étant l'ensemble des génomes n'appartenant pas à ce groupe (apprentissage de deux classes).

Afin d'estimer la qualité des classifieurs obtenus par chacune des trois méthodes de classification choisies, nous avons utilisé la méthode « *Leave One Out* ». Il est à noter que nous avons prédit les groupes d'orthologues à partir de la totalité des génomes. Ainsi, *stricto sensu,* il existe un lien entre les données d'apprentissage et de test, ce qui peut être une source de biais dans l'évaluation. Cependant, ce n'est pas sur les groupes d'orthologues que ce fait l'évaluation, mais sur les profils phylogénétiques. Ces profils ont été obtenus par l'annotation des groupes d'orthologues. Ces groupes sont annotés indépendamment et plusieurs groupes peuvent porter la même annotation. Ainsi l'utilisation des profils permet une diminution de ce possible biais.

A)

Réel (R)\ Prédit (P)	**Positif**	**Négatif**
Positif	Vrais positifs (VP)	Faux négatifs (FN)
Négatif	Faux positifs (FP)	Vrais négatifs (VN)

B)

$$\text{Précision} = \frac{VP}{VP + FP}$$

$$\text{Sensibilité} = \frac{VP}{VP + FN}$$

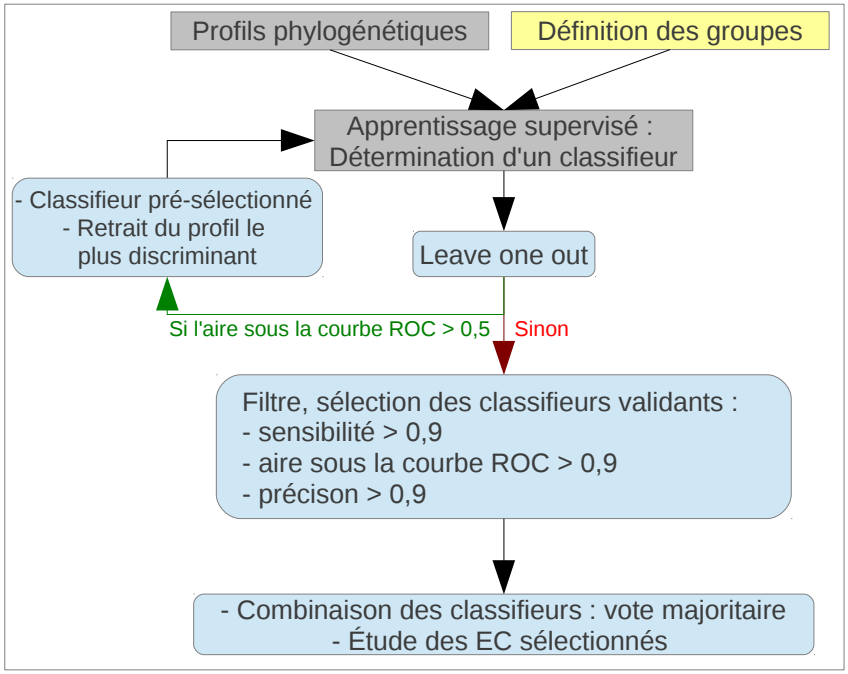
Figure 1 : (A)Matrice de confusion et (B) Critères d'évaluation associés.

La matrice de confusion (Figure 1A) est obtenue en comparant le groupe taxonomique associé à chaque champignon de l'ensemble de test (méthode *Leave One Out*) avec le groupe taxonomique de référence de cette espèce. Elle se construit en mettant respectivement sur les lignes et sur les colonnes les groupes taxonomiques de référence et la classification faite par le classifieur.

La majorité des groupes étudiés ont un nombre d'exemples positifs très inférieur au nombre d'exemples négatifs, c'est pourquoi nous avons fait le choix de critères d'évaluation décrivant majoritairement la qualité de prédiction des exemples positifs. Nous évaluons donc la qualité des classifieurs en fonction de trois critères différents (Figure 1B) : la sensibilité, la précision et l'aire sous la courbe ROC (*Receiver Operating Characteristic*) . Ces critères ont des valeurs comprises entre 0 et 1. Les classifieurs d'un groupe taxonomique sont conservés s'ils ont, pour ces trois critères, une valeur supérieur à 0,9.

Nous cherchons à caractériser un groupe d'organismes en fonction des activités enzymatiques. Or il est possible que plusieurs combinaisons d'activités enzymatiques différentes soient pertinentes pour caractériser ce groupe. De ce fait, afin d'obtenir un large panel de classifieurs pertinents nous appliquons de façon itérée la recherche de classifieurs sur les données. La recherche des classifieurs s'effectue en deux étapes. La première consiste à collecter l'ensemble des classifieurs meilleurs que des classifieurs aléatoires (aire sous la courbe ROC supérieure à 0,5). Le seconde étape à pour objectif de filtrer les classifieurs en fonction dess

seuils fixés pour la sensibilité, la précision et l'aire sous la courbe ROC. Les profils phylogénétiques sélectionnés en premier par l'algorithme d'apprentissage sont retirés du jeu de données et un nouveau classifieur est appris (Figure 2).



Pour un groupe taxonomique donné, les classifieurs obtenus avec les trois approches prédisent chacun l'appartenance du champignon à un groupe (le groupe taxonomique en cours de caractérisation ou « autre », c'est-à-dire tout sauf ce groupe). L'ensemble des activités enzymatiques sélectionnées par les classifieurs forment les activités enzymatiques caractéristiques de ce groupe. La prédiction du groupe taxonomique d'un nouveau champignon sera faite en conservant la prédiction majoritaire de l'ensemble des classifieurs finaux.

Figure 2 : Pipeline d'apprentissage supervisé pour un algorithme d'apprentissage donné. Les différentes étapes sont décrites sur les rectangles aux angles arrondis à fond bleu, les données que nous générons sont représentées sur fond gris et les données de la littérature sur fond jaune.

3 Résultats

D'une manière générale les résultats de l'application de cette méthodologie ont donné des classifieurs faisant intervenir un faible nombre d'activités enzymatiques avec de forts pourcentages de bonne classification. L'obtention de ces classifieurs s'effectue en un temps raisonnable, l'apprentissage d'un classifieur pour un groupe donné (une itération de la boucle de la figure 2) prenant moins d'une minute.

Nous traitons ici en tant qu'exemple la caractérisation des *Agaricomycetes* (phylum des *Basidiomycota*) à partir de l'ensemble des espèces présentes dans nos données ainsi qu'une partie de la caractérisation des *Pezizomycotina (*phylum *Ascomycota)* parmi les champignons.

3.1 Caractérisation des Agaricomycetes

L'application de notre méthode (Figure 2) à la détermination des *Agaricomycetes* a permis de construire les 13 classifieurs de la table supplémentaire 2. Les profils phylogénétiques sélectionnés par les classifieurs sont présentés dans la table 2. Les activités enzymatiques sélectionnées par les classifieurs comme caractéristiques des *Agaricomycetes* sont impliquées dans différentes processus biologiques que nous décrivons ci-dessous.

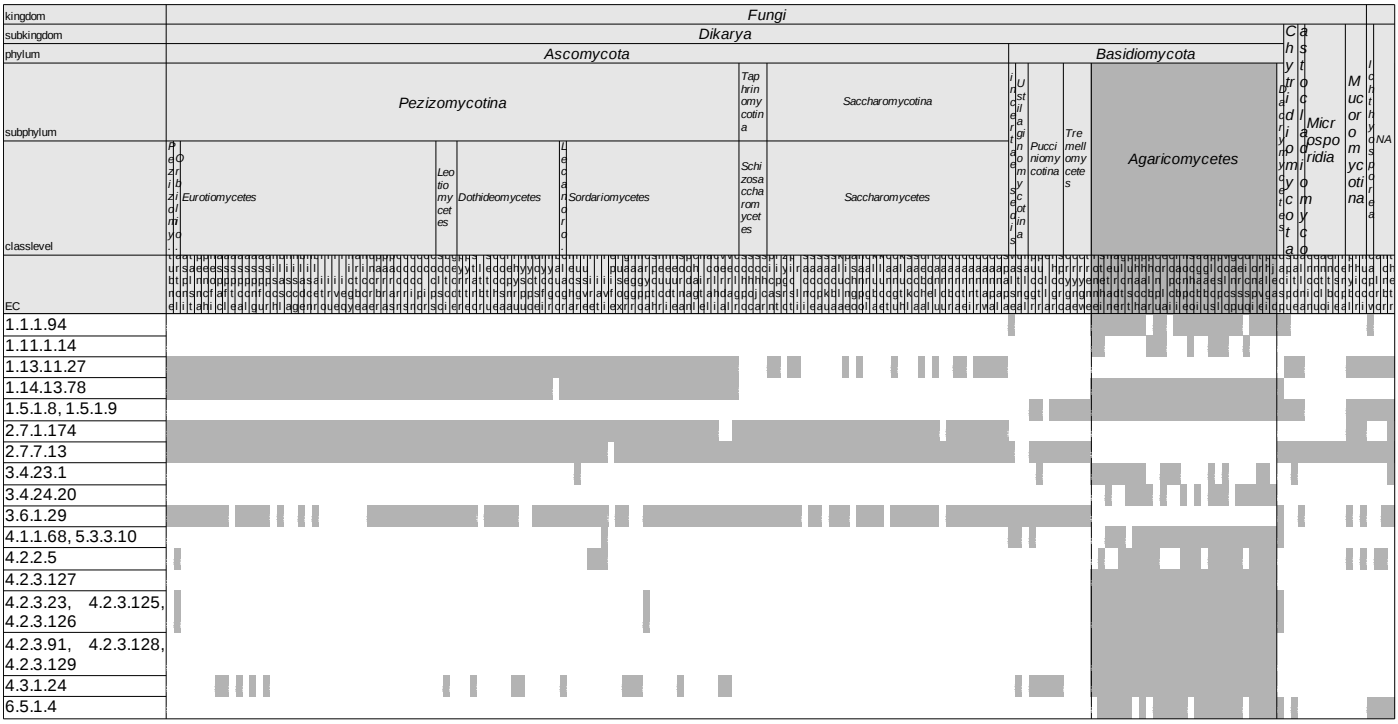


Table 2 : Profils phylogénétiques des activités enzymatiques présentes dans les classifieurs caractérisant la classe *Agaricomycetes*. Les activités enzymatiques sont disposées en ligne et les génomes sont disposés en colonne. Les cases sur fond gris signifient que l'activité enzymatique est présente dans le génome, sur fond blanc qu'elle est absente. Le groupe taxonomique des *Agaricomycetes* est sur fond gris foncé.

3.1.1 Création de composés biologiquement actifs

Les EC :4.2.3.91 (cuberol synthase), EC :4.2.3.127 (beta-copaenz synthase), EC :4.2.3.128 (beta-cubebene synthase) et 4.2.3.129 ((+)-sativene synthase) ont été retrouvés caractéristiques des *Agaricomycetes* par l'ensemble des méthodes. Elles correspondent aux activités enzymatiques de la Sesquiterpene synthase COP4 connue pour être impliquée dans la catalyse de la cyclisation du farnesyl diphosphate en plusieurs produits, incluant la germacrene D, la beta-copaene, la bete-cubebene, la (+)-sativene et le cuberol. Elles catalysent la formation de terpenoids intermédiaires à la création de composés biologiquements actifs tels que les antibiotiques et les toxines [13]. De plus, l'activité enzymatique EC :1.14.13.78 (ent-kaurene oxidase) retrouvée caractéristique des *Agaricomycetes*, a elle aussi un rôle dans la synthèse des terpenoides (gibberellins). Les *Agaricomycetes* se caractérisent donc en partie par leur capacité à produire de tels types de composés.

Trois autres activités enzymatiques correspondant aux activités enzymatiques de la Sesquiterpene synthases COP3 sont retrouvées, il s'agit des EC : 4.2.3.126 (Alpha-muurolene synthase), EC : 4.2.3.23 (Gamma-muurolene synthase) et EC : 4.2.3.125 (Germacrene-A synthase). Les produits des réactions catalysées par ces enzymes sont des composés (ou des intermédiaires de composés) biologiquement actifs. Le germacrene est par exemple un agent antimicrobien.

3.1.2 Dégradation de la biomasse

Les activités enzymatiques EC :1.11.1.14 (lignine peroxidase) ,EC :3.4.24.20 (Peptidyl-Lys metalloendopeptidase), EC :4.2.2.5 (chondroitine AC lyase) et EC:3.4.23.1 (pepsin A) sont retrouvées comme caractéristiques des *Agaricomycetes*. La peptidyl-Lys metalloendopeptidase est une protéase sécrétée. La pepsin A est une endopeptidase. La lignin peroxidase est impliquée dans la dégradation de la lignine [13]. La chondroitine AC lyase permet la dépolymérisation de la chondroitine sulfate (constituant du

cartilage) et du dermatan sulfate (constituant de la peau), ainsi cette activité enzymatique pourrait permettre la dégradation de la biomasse d'origine animale.

Le groupe des *Agaricomycetes* contient à la fois les pourritures brunes et les pourritures blanches. Elles sont toutes deux connues pour leur capacité de dégradation de la biomasse. Ainsi trouver ces activités enzymatiques caractéristique de ce groupe est cohérent avec les connaissances actuelles sur ces organismes.

3.1.3 Paroi et membrane cellulaire

Une activité enzymatique spécifiquement présente au niveau de la paroi et de la membrane cellulaire a été sélectionnée. Il s'agit de activité enzymatique EC :2.7.7.13 (mannose-1-phosphate guanylyltransferase).

L'absence de la mannose-1-phosphate guanylyltransferase est caractéristique des *Agaricomycetes*. Cette activité enzymatique permet la formation de GDP-mannose, lui même impliqué dans la formation de la paroi. Ainsi la paroi des Agaricomycetes semble être particulière ou sa synthèse implique d'autres activités enzymatiques (sous-voies différentes) induisant la production de GDP-mannose.

3.1.4 Metabolisme des nucléotides

Les *Agaricomycetes* se caractérisent également comme ne possédant pas l'activité enzymatique EC : 3.6.1.29 (bis(5'-adenosyl)-triphosphatase). Cette activité enzymatique est en lien avec le métabolisme des nucléotides. Ainsi ce groupe d'organisme semble posséder un métabolisme des nucléotides particulier différent de celui des autres *Basidiomycetes*.

3.1.5 Stockage de glycerolypides

L'activité enzymatique EC :1.1.1.94 (glycerol-3-phosphate dehydrogenase [NAD(P)+]) est impliquée dans le stockage des triglycérides. Les *Agaricomycetes* ne sont pas spécialement connus pour avoir cette capacité. Ainsi il s'agit donc soit d'une erreur d'annotation soit de la découverte d'une nouvelle capacité caractéristique de se groupe de champignons qu'il serait intéressant de tester.

3.1.6 Métabolisme des vitamines

L'absence de la mannose-1-phosphate guanylyltransferase (EC: 2.7.7.13) chez les *Agaricomycetes* pourrait également avoir des conséquences dans les possibilités de biosynthèse de la vitamine C.

3.1.7 Métabolisme de acides aminés

Plusieurs activités enzymatiques sélectionnées sont impliquées dans le métabolisme de la tyrosine et de la phenylalanine. Les activités enzymatiques EC :4.1.1.68 et EC :5.3.3.10 agissent de manière consécutive dans la métabolisme de la tyrosine. Elles appartiennent à une sous-voie pouvant avoir comme précurseur la phenylalanine ou la tyrosine. L'EC :4.3.1.24 (phenylalanine ammonia-lyase) permet la formation du trans-Cinnamate, substrat initial d'une partie des voies de dégradation de la phenylalanine. Enfin, l'activité enzymatique EC :1.13.11.27 permet la transformation du phenylpyruvate en 2-hydroxy-phenylacetate, précurseur de la voie de dégradation des styrenes.

Les activités enzymatiques EC:1.5.1.8 (saccharopine dehydrogenase (NADP+, L-lysine-forming)) et EC:1.5.1.9 (saccharopine dehydrogenase (NAD+, L-glutamate-forming)) ont un rôle consécutif dans le métabolisme de la lysine. Elles permettent l'entrée possible de la lysine vers les le cycle du citrate (dégradation) et vers la voie de production de penicilline et cepholosporine.

Les mécanismes de modification et de dégradation de certains acides aminés semble donc caractéristique des *Agaricomycetes.*

3.1.8 Regulation de la structure de la membrane nucléaire

La diacylglycerol kinase (CTP dependant) est connue chez *Saccharomyces cerevisiae* pour réguler la

synthèse de phospholipides et la croissance de la membrane nucléaire [15]. Son absence est caractéristique du groupe des *Agaricomycetes*. Ainsi, ces champignons utilisent donc certainement d'autres mécanismes pour effectuer cette régulation.

3.2 Caractérisation des *Pezizomycotina*

La caractérisation des *Pezizomycotina* sur les données constituées de l'ensemble des champignons permet la sélection de 381 classifieurs. Parmi ceux-ci nous ne présenterons ici que le premier classifieur obtenu par les 3 algorithmes. Ce classifieur est le suivant : si l'activité enzymatique 3.1.6.6 est présente alors le champignon appartient au groupe des *Pezizomycotina* sinon il appartient à un autre groupe. L'EC 3.1.6.6 (choline-sulfatase) intervient dans la voie de dégradation de la choline-o-sulfate, permettant une source de soufre endogène mobilisable pendant la croissance [16]. Les *Pezizomycotina* se caractérisent donc en partie par leur capacité à croître sur un milieu pauvre en soufre.

La choline-sulfatase est présente chez l'ensemble des *Pezizomycotina* ainsi que chez *Ustilago maydis* (un *Basidiomycota*). Nous cherchons à comprendre d'où provient cette activité enzymatique chez U. *maydis.* Au total, 85 protéines annotées EC:3.1.6.6 appartiennent au même groupe d'orthologue que la séquence d'*U. Maydis*. Sur l'arbre phylogénétique obtenu à partir de ces 85 séquences (Figure 3) la séquence correspondant à la protéine de *U. maydis* n'est pas sur une branche externe mais est nichée proche d'un groupe de séquences de *Pezizomycotina*. Cela conduit à penser que la présence de l'activité enzymatique EC:3.1.6.6 est probablement due à un transfert horizontal entre une souche de *Pezizomycotina* et *U. maydis.*

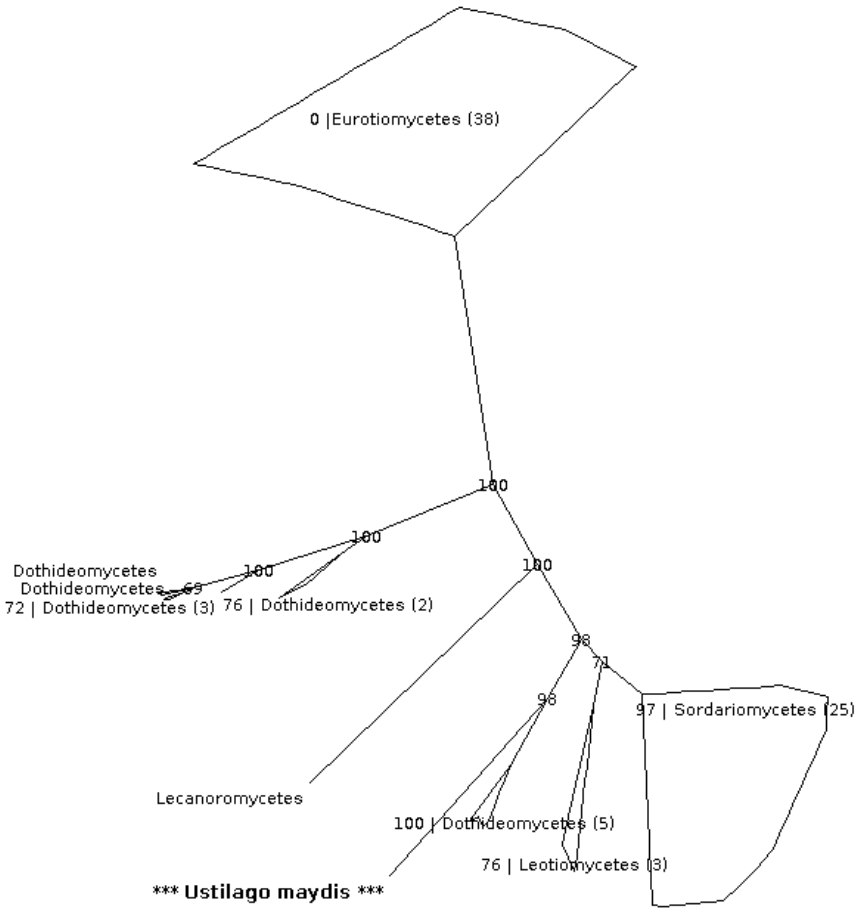


Figure 3 : Arbre phylogénétique construit avec le groupe d'orthologue de la sequence d'*U. maydis* annoté 3.1.6.6. L'identifiant de la sequence de *U. Maydis* est représentée en gras et encadré par des étoiles ('***'). Pour les autres séquences seul la classe taxonomique est indiquée. Les séquences ont été alignées avec

MUSCLE [8] et les parties les moins bien alignées ont été retirées à l'aide du logiciel Gblocks [9]. L'arbre phylogénétique a été construit à l'aide du programme PhyML [10]. Les nombres indiqués sur certain nœuds correspondent à la valeur de bootstrap.

4 Conclusion

L'application de méthodes d'apprentissage supervisé sur les groupes taxonomiques à partir de profils enzymatiques nous permet de caractériser un groupe d'organismes en fonction de ses activités enzymatiques. À travers la caractérisation des *Agaricomycetes* et des *Pezizomycotina*, nous mettons en évidence la validité de l'application de méthodes d'apprentissage supervisé dans le but d'extraire de l'information des profils phylogénétiques. Cette méthode permet également de poser de nouvelles hypothèses sur l'évolution du répertoire enzymatique de ces organismes, notamment en détectant des transferts horizontaux.

Par la suite, cette méthodologie sera appliquée à la caractérisation d'autres groupes partageant une caractéristique commune dans le but de comprendre les mécanismes induisant cette caractéristique. Nous l'appliquerons par exemple à la caractérisation des champignons ayant de fortes capacités de dégradation de la biomasse. L'étude de la biomasse se fera sur les profils phylogénétiques des activités enzymatiques ainsi que sur l'étude des groupes d'orthologues afin de trouver de nouvelles protéines impliquées dans ce processus. Nous pourrons ainsi prédire les capacités de dégradation de la biomasse de nouvelles espèces de champignons.

Remerciments

Ce travail a bénéficié d'un financement par le PEPS Bio-Maths-Info (BMI) CNRS-INSERM-INRIA .

References

[1] Y. Zhang, S. Li, G. Skogerbo, Z. Zhang, X. Zhu, Z. Zhang, S. Sun, H. Lu, B. shi and R. Cher, Phylogenetic properties of metabolic pathway topologies as revealed by global analysis. *BMC Bioinformatics*, doi:10.1186/1471-2105-7-252, 2006.

[2] A. Mano, T. Tuller, O. Béjà and R. Y. Pinter, Comparative classification of species and the study of pathway evolution based on the alignment of métabolic pathways. *BMC Bioinformatics*, doi:10.1186/1471-2105-11-S1-S38, 2010

[3] D. Perumal, C. S. Lim and M. K. Sakharkar, A comparative study of metabolic network topology between a pathogenic and a non-pathogenic bacterium for potential drug target identificaion. *Translat Bioinforma*, 2009: 100-104, 2009

[4] S. Grossetête, B. Labedan and O. Lespinet, FUNGIpath, a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics*, doi:10,1186/1471-2164-11-81, 2010.

[5] IUPAC-IUBMB. IUBMB Joint Commission on Biochemical Nomenclature (JCBN) and Nomenclature Committee of IUBMB, Newsletter 1999. *Eur. J Biochem.* 1999;264:607-609

[6] J.R. Quinlan, C4.5: Programs for Machine Learning. *Morgan Kaufmann Publishers*, 1993.

[7] E. Frank, I. H. Witten, Generating Accurate Rule Sets Without Global Optimization. *Fifteenth International Conference on Machine Learning*, 144-151, 1998.

[8] W. W. Cohen, Fast Effective Rule Induction. *Twelfth International Conference on Machine Learning*, 115-123, 1995.

[9] M. Hall, E. Frank, G. Holmes, B. Pfahringer, P. Reutemann, I. H. Witten, The WEKA Data Mining Software: An Update; *SIGKDD Exploration*, Volume 11, Issue 1, 2009

[10] R.C. Edgar, MUSCLE: multiple sequence alignment with high accuracy and high throughput *Nucleic Acids Res.* **32**(5):1792-1797, 2004

[11] J. Castresana, Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. *Molecular Biology and Evolution* 17, 540-552. 2000

[12] S. Guindon, J.F. Dufayard, V. Lefort, M. Anisimova, W. Anisimova, O. Gascuel, New algorithms and Methods to Estimate Maximum-Likelihood Phylogenies: Assessing the Performance of PhyML 3.0. *Systematic Biology*, 59(3):307-21, 2010

[13] F. Lopes-Gallego, S. A. Agger, D. A. Pella, M.D. Distefano, C. Schmidt-Dannert, Sesquiterpene synthase Cop4 and Cop6 from *Coprinus cinereus*: Catalytic promiscuity and cyclization of farnesyl pyrophosphate geometrical isomers. *Chembiochem*, 11(8):1093-1106, 2010

[14] G. Daniel, J. Volc, L. Filonova, O. Plíhal, E. Kubátová, P. Halada. Characteristics of Gloeophyllum trabeum alcohol oxidase, an extracellular source of H2O2 in brown rot decay. *Appl Environ Microbiol 73(19):6241-53*, 2007.

[15] G. Han, L. O'Hara, G. M. Carman and S. Siniossoglou. An unconventional diacylglycerol kinase that regulates phospholipid synthesis and nulcear membrane growth. *J Biol Chem. 283(29):20433-20442*, 2008.

[16] R. A. Gravel. Choline-O-sulphate utilization in *Aspergillus nidulans. Genetical Research, 28(3):261-76, 1976*

Annexes

Groupe taxonomique						nombre de génomes
Eumycota	Dikarya	Ascomycota	Pezizomycotina	Dothideomycetes	Capnodiales	5
					Hysteriales	2
					Pleosporales	8
				Eurotiomycetes	Eurotiales	13
					Onygenales	24
				Lecanoromycetes	Lecanorales	1
				Leotiomycetes	Helotiales	3
				Orbiliomycetes	Orbiliales	1
				Pezizomycetes	Pezizales	1
				Sordariomycetes	Diaporthales	1
					Glomerellales	5
					Hypocreales	8
					Magnaporthales	3
					Sordariales	8
			Saccharomycotina	Saccharomycetes	Saccharomycetales	35
			Taphrinomycotina	Schizosaccharomycetes		4
		Basidiomycota	Basidiomycota incertae sedis	Wallemiomycetes	Wallemiales	1
			Agaricomycetes		Agaricales	7
					Auriculariales	1
					Boletales	2
					Corticiales	4
					Gloeophyllales	1
					Hymenochaetales	1
					Polyporales	9
					Russulales	2
			Dacrymycetes		Dacrymycetales	1
			Tremellomycetes		Tremellales	4
			Pucciniomycotina		Pucciniales	3
			Pucciniomycotina		Sporidiobolales	2
			Ustilaginomycotina	Exobasidiomycetes	Malasseziales	1
				Ustilaginomycetes	Ustilaginales	1
	Blastocladiomycota		Blastocladiomycetes		Blastocladiales	1
	Chytridiomycota		Chytridiomycetes		Spizellomycetales	1
					Rhizophydiales	1
	Microsporidia					6
	Mucoromycotina				Mucorales	3
Apusozoa						1
Filasterea						1
Choanoflagellida						2

Table supplémentaire 1 : Distribution taxonomique des génomes. Les groupes taxonomiques de champignons sont sur fond gris, les autres sur fond blanc.

Algorithmes	Données	Classifieurs
C4.5	Tous les profils	Si les EC : 4.2.3.127 et 4.2.3.91 et 4.2.3.128 et 4.2.3.129 sont présents alors *Agaricomycetes* (151/0) (un seul profil) Sinon autre (27/0)
	Tous les profils sauf celui de EC : 4.2.3.127	Si les EC : 4.2.3.126 et 4.2.3.23 et 4.2.3.125 sont absents alors autre (148.0/0) (un seul profil) Sinon si l'EC : 4.3.1.24 est absent autre (3.0) Sinon *Agaricomycetes* (27.0)
	Tous les profils sauf ceux de EC : 4.2.3.127 et EC :4.2.3.126	Si les EC : 2.7.7.13 et EC : 1.5.1.9 et EC : 1.5.1.8 sont absents alors autre (3.0/0) Sinon si l'EC : 2.7.7.13 est absent et que les EC :1.5.1.9 et 1.5.1.8 sont présents alors (27.0) Sinon autre (148.0)
	7 profils phylogénétiques retirés	Si les EC : 4.1.1.68, EC : 5.3.3.10 et EC : 1.11.1.14 sont absents alors autre (147.0/1.0) Sinon, si les EC : 4.1.1.68, EC : 5.3.3.10 sont absents et que l'EC : 1.11.1.14 est présent alors *Agaricomycetes* (2.0/0) Sinon si les EC : 4.1.1.68 et EC : 5.3.3.10 sont présents et que l'EC : 3.6.1.29 est absent alors *Agaricomycetes* (24.0/0) Sinon autre (5.0/0)
PART	Tous les profils	Si EC : les EC 4.2.3.127 et 4.2.3.91 et 4.2.3.128 et 4.2.3.129 sont présents alors *Agaricomycetes* Sinon autre
	Tous les profils sauf celui de EC : 4.2.3.127	Si les EC : 4.2.3.126 et 4.2.3.23 et 4.2.3.125 sont absents alors autre (148.0/0) (un seul profil) Sinon si l'EC :4.3.1.24 est absent autre (3.0) Sinon *Agaricomycetes* (27.0)
	Tous les profils sauf ceux de EC : 4.2.3.127 et EC :4.2.3.126	Si l'EC :2.7.7.13 est présent alors autre (148.0/0) Sinon, si l'EC :1.5.1.9 et l'EC : 1.5.1.8 sont présents alors *Agaricomycetes* (27.0/0) (même profil) Sinon autre (3.0/0)
	5 profils phylogénétiques retirés	Si l'EC :1.1.1.94 et l'EC : 1.11.1.14 sont absents alors autre (148.0/1.0) Sinon, si l'EC :4.3.1.24 est présent alors *Agaricomycetes* (26.0/0) Sinon autre (4.0/0)
	7 profils phylogénétiques retirés	Si l'EC :4.1.1.68, l'EC : 5.3.3.10 et l'EC : 1.11.1.14 sont absent alors autre (147.0/1.0) Sinon, si l'EC : 3.6.1.29 est absent alors *Agaricomycetes* (26.0/0) Sinon autre (5.0/0)
RIPPER	Tous les profils	Si EC : les EC 4.2.3.127 et 4.2.3.91 et 4.2.3.128 et 4.2.3.129 sont présents alors *Agaricomycetes* Sinon autre
	15 profils phylogénétiques retirés	Si les EC :6.5.1.4 et 4.3.1.24 sont présents alors *Agaricomycetes* (22.0/0.0) Sinon si l'EC : 3.4.24.20 est présent alors *Agaricomycetes* (3.0/0.0) Sinon si l'EC : 1.11.1.14 est présent alors Agaricomycetes (2.0/0.0) Sinon autre (151.0/0.0)
	41 profils phylogénétiques retirés	Si l'EC : 2.7.1.174 est absent et que l'EC : 1.14.13.78 est présent alors *Agaricomycetes* (30.0/3.0) Sinon autre (148.0/0.0)
	45 profils phylogénétiques retirés	Si l'EC :4.2.2.5 est présent et que l'EC : 1.13.11.27 est absent alors *Agaricomycetes* (19.0/0.0) Sinon, si l'EC : 3.4.23.1 est présent alors *Agaricomycetes* (10.0/4.0) Sinon autre (149.0/2.0)

Table supplémentaire 2 : Classifieurs caractérisant les Agaricomycetes. Les nombres entre parenthèses indiquent dans le cas où une règle génère des erreurs sur le jeu de données complet combien de génomes sont couverts par la règle puis le nombre d'erreurs.

Comparaison avec les résultats obtenus sur FungiPath version 50 génomes

La méthodologie appliquée dans les deux analyses diffère. En effet, pour l'analyse faite sur FungiPath 50, seul le premier classifieur obtenu avec chaque méthode était conservé. Dans l'approche proposée à JOBIM nous avons ajouté une étape de suppression dans les données du premier attribut sélectionné par la méthode de classification. Nous espérions ainsi obtenir plus de règles pouvant nous aider à comprendre les mécanismes d'évolution du métabolisme.

Toujours dans ce but, nous avons été relativement peu stringents. Nous avons conservés les classifieurs tant qu'ils présentaient un meilleur résultat que celui attendu aléatoirement (*leave one out*, aire sous la courbe ROC>0,5). Un grand nombre des classifieurs présélectionnés sont donc certainement peu pertinents. De manière à ne conserver que les classifieurs les plus informatifs, nous avons effectué un second filtre sélectionnant cette fois-ci les classifieurs présentant des seuils sensibilité, aire sous la courbe ROC et précision supérieurs à 0,9 en *leave one out*. L'ensemble des classifieurs obtenu après la boucle ne sont donc pas au final sélectionnés. Cette approche fut proposée afin de ne pas être limitée par la présence de profils induisant du bruit dans les données. C'est un des moyen que nous avons proposé afin d'obtenir différents classifieurs *via* l'application de méthode de classification sur différents ensembles de profils.

La méthode présentée dans cet article permet d'obtenir 300 classifieurs portant sur le concept 'ascomycota'. Parmi eux, quatre font intervenir l'EC:3.4.23.41. Cependant, dans cette nouvelle version de FungiPath, cette activité enzymatique n'est plus suffisante à elle seule pour caractériser le groupe des ascomycota. Elle n'est effectivement présente que chez des ascomycota mais pas chez l'ensemble des ascomycota. Certains des nouveaux ascomycota ajoutés dans la base ne possèdent pas cette activité enzymatique, c'est pourquoi elle n'est plus sélectionnée. Nous sommes donc dans un cas où le nombre d'exemples que nous avions dans nos don-

nées n'était pas suffisant pour caractériser le groupe taxonomique. De plus, certaines séquences anciennement annotées avec cette activité enzymatique ne le sont plus.

Pour rappel la prédiction des groupes d'orthologues de FUNGIpath 50 et FUNGIpath 178 v1 (FP178 v1) a été réalisée grâce à l'application de la première méta-approche développée dans notre l'équipe (Grossetete et al., 2010). L'annotation de ces groupe avait ensuite été réalisée à partir de la comparaison des profils HMM des groupes avec les séquences annotée manuellement dans Swissprot et MetaCyc.

Les groupes d'orthologues de la version FungiPath 50 ne sont pas strictement inclus dans la version FungiPath 178 v1. C'est de ce type d'observation que nous est apparu un des problèmes de la première méthode FungiPath. L'utilisation de toutes les intersections de toutes les méthodes (intersection de 4, 3 et 2 méthodes) en une seule étape induit un trop grand nombre de graines (plusieurs graines peuvent être représentatives d'un même groupe d'orthologues). Cela aboutit à l'obtention de groupes d'orthologues divisés. C'est l'une des raisons qui nous ont poussés à développer la méta-approche MARIO.

Nous obtenons également dans cette analyse 72 classifieurs permettant de caractériser les basidiomycota. Parmi eux, aucun ne fait intervenir l'activité enzymatique 3.4.23.5. Nous observons ici le même type de résultats que pour les ascomycota. L'ensemble des 12 basidiomycota de FungiPath 50 sont prédits comme présentant l'activité enzymatique 3.4.23.5 dans la nouvelle base. Le groupe d'orthologues annoté avec l'activité 3.4.23.5 dans la base de données à 178 génomes ne contient que des séquences de basidiomycota. Malgré cela, cette activité enzymatique n'est pas sélectionnée car elle n'est retrouvée que parmi 33 des 40 basidiomycota de la base. Les méthodes de classification n'ont pas trouvé de combinaisons satisfaisantes d'activités enzymatiques faisant intervenir l'activité enzymatique 3.4.23.5. Nous sommes ici dans un cas où le faible nombre de champignons dans nos données initiales à induit l'apprentissage de classifieurs trop peu efficaces sur de nouveaux exemples pour

pouvoir être de nouveau prédits en les utilisant dans le jeu de données d'apprentissage.

L'un des classifieurs des pezizomycota sur FP178 v1 (83 champignons parmi les 174) fait intervenir la même activité enzymatique que celle obtenue sur FungiPath 50, l'EC:3.1.1.6. Comme nous l'avions observé pour FungiPath 50, nous retrouvons un unique cas de mauvaise classification avec cette activité enzymatique, le champignon *Ustilago Maydis* classé en tant que pezizomycota alors qu'il s'agit d'un basidiomycota. Comme avec les données de FungiPath 50, l'analyse de l'arbre phylogénétique obtenu avec ces séquences (voir article JOBIM) aboutit à conclure à un transfert horizontal entre un ancêtre des pezizomycota et un ancêtre d'*Ustilago maydis*. L'analyse faite sur 50 génomes est ici confirmée. Les pezizomycota semblent bien se caractériser par la présence de l'activité enzymatique 3.1.6.6.

13.3.3 Résultats obtenus avec MARIO sur 173 protéomes de champignons

Comme nous l'avons expliqué dans la partie précédente, nous avons observé que la première version de FungiPath avait tendance à diviser des groupes d'orthologues. Pour pallier ce problème, nous avons développé une nouvelle approche, l'approche MARIO. Nous présentons par la suite les résultats obtenus sur ces nouvelles données.

Les données d'entrée : FungiPath obtenu avec la méthode MARIO

Le développement de la version actuelle de FungiPath est basé sur les groupes d'orthologues prédits en appliquant la méthode MARIO présentée dans la partie 2. Cette base a été créée à partir de 178 protéomes, dont 174 de champignons. Cependant, après vérification du contenu de la base il s'est avéré que le protéome de *Candida dublieniensis* avait été téléchargé dans une version incomplète. Ainsi nous avons utilisé pour l'analyse les 173 autres protéomes de champignons, ce qui correspond à 1 706 330 protéines et 67 045 groupes d'orthologues.

Parmi eux, 4 373 groupes (321 113 protéines) ont été annotés avec une activité enzymatique (*EC number*). Nous avons prédit 1 240 activités enzymatiques différentes chez ces organismes.

Les profils obtenus pour les différentes activités enzymatiques ne sont pas indépendants. Deux activités enzymatiques peuvent présenter le même profil. Dans ce cas, nous ne conservons pour l'analyse qu'un seul profil que nous étiquetons avec l'ensemble des *EC numbers* concernés. Les activités enzymatiques partageant le même profil sont, dans un tiers des cas, impliquées dans la même voie métabolique (voie de KEGG) (voir le tableau 4 en annexe).

On observe dans la table 13.3 la distribution des données dans la matrice

Nombre de profils d'enzymes différents	1075
Nombre de champignons (tous les champignons ont des profils différents)	173
Nombre de 0 dans la matrice	67068 (36.16%)
Nombre de 1 dans la matrice	118388 (63.84%)

Table 13.3 – *Les profils en chiffres*

champignons/activités enzymatiques. Lorsqu'une activité enzymatique est prédite pour un organisme, elle a tendance à être également présente chez plusieurs autres champignons (nombre de 1 dans la matrice > nombre de 0). Cela est dû en partie à notre méthode d'annotation. En effet, les séquences seules ne sont annotées que si leur fonction est connue dans SwissProt ou MetaCyc. De nouvelles annotations ne sont proposées que pour des groupes d'orthologues ce qui induit l'annotation simultanée de plusieurs séquences.

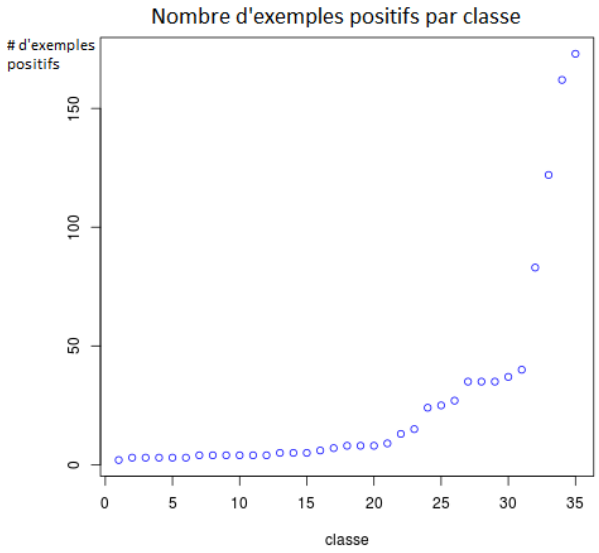


Figure 13.9 – *Distribution du nombre d'espèces par groupe taxonomique parmi les 173 champignons de la base de données FungiPath. Le point le plus haut correspond au groupe taxonomique des dikarya.*

Les **exemples** utilisés pour l'apprentissage sont les espèces présentes dans FungiPath. Nous avons 29 classes ou groupes taxonomiques pour lesquels il y a au moins trois représentants, ce concept-cible 'groupe taxonomique' est qualifié de multi–classe.

Les différentes classes ne sont pas présentes dans les mêmes proportions (voir figure 13.9). La majorité des classes que nous cherchons à caractériser présentent un faible nombre d'exemples positifs, c'est-à-dire que, généralement, peu de champignons appartiennent à un groupe taxonomique. Ce faible nombre d'exemples positifs peut aboutir à l'apprentissage de classifieurs représentatifs des espèces de ce groupe dans FungiPath mais pas représentatifs du groupe en lui même. Nous devrons donc être prudents dans nos conclusions faites pour les groupes présentant peu ou trop de représentants. Cette notion de trop peu d'exemples positifs ou négatifs est difficile à évaluer *a priori*. Nous avons choisi de travailler avec l'ensemble des groupes présentant au moins trois champignons. Nous garderons à l'esprit que nous pourrons avoir certains groupes pour lesquels nous n'observons pas de classifieurs, pas parce qu'aucun classifieur n'est pertinent pour caractériser ce groupe dans l'absolu, mais parce que trop peu d'exemples sont présents pour obtenir un classifieur. Nous observons ce phénomène dans nos résultats mais n'avons pas évalué strictement ce problème.

Résultats obtenus avec une recherche de classifieurs itérée sur un ensemble d'attributs à la taille décroissante (algorithme proposé à JOBIM)

Il est possible que plusieurs combinaisons d'activités enzymatiques soient pertinentes pour la caractérisation d'un groupe taxonomique. Ainsi, comme expliqué dans la présentation de l'approche dans l'article Pereira et al. (2013), nous avons fait le choix de retirer de manière itérative la première activité enzymatique sélectionnée afin d'obtenir un ensemble d'arbre et de règles de décision. Pour cela, nous

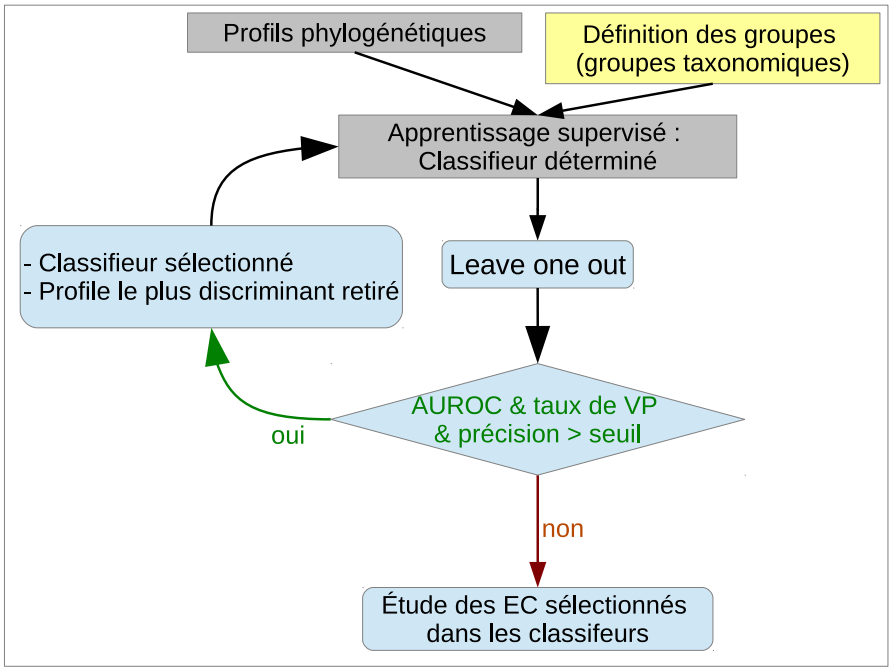


Figure 13.10 – *Pipeline d'apprentissage supervisé pour un algorithme donné. Les différentes étapes sont décrites sur des rectangles aux angles arrondis et sur fond bleu, le test est sur figuré par un losange, les données que nous générons sont sur fond gris et les données de la littérature sur fond jaune.*

Table 13.4 – *Nombre d'EC numbers différents présents dans au moins un classifieur. Seuls les groupes taxonomiques présentant au moins un classifieur (seuils 0,9) sont listés.*

Groupe taxonomique	nombre d'*EC numbers* différents dans les classifieurs	nombre de classifieurs C4.5	nombre de classifieurs RIPPER	nombre de classifieurs PART
Agaricomycotina	3	0	1	0
Ascomycota	114	18	62	18
Basidiomycota	2	1	1	1
Dikarya	8	5	2	5
Onygenales	5	1	2	1
Pezizomycotina	97	0	45	0
Taphrinomycotina	1	1	1	1

appliquons le pipeline présenté en figure 13.10. Nous avons choisi d'appliquer des seuils plus stringents que ceux proposés dans Pereira et al. (2013) avec un seuil de 0,9 pour l'aire sous la courbe ROC, le taux de vrais positifs et la précision.

La méthode que nous avions proposé précédemment induisait la sélection d'un grand nombre de classifieurs qui, du fait même de leurs nombres, étaient difficilement interprétables. Pour rappel, elle présentait deux étapes de filtre, une lors de la boucle permettant l'apprentissage itératif de classifieurs (seuil AUC>0,5) et une seconde afin de ne conserver au final que les classifieurs aux seuils sensibilité, précision et AUC supérieurs à 0,9.

L'étape de pré-sélection se faisait avec des seuils tout juste supérieurs aux résultats attendus d'un classifieur aléatoire (AUC>0,5), ce qui pourrait être discutable (trop peu spécifiques). En ne conservant dans cette nouvelle analyse que les classifieurs passants successivement des seuils fixés à 0,9 (seuil élevé) nous nous assurons de n'obtenir et d'analyser que les résultats les plus spécifiques. Cependant, il est possible que nous n'observions pas l'ensemble des règles possiblement pertinentes.

En additionnant le nombre de classifieurs obtenus par chaque méthode nous arrivons à un total de 158 classifieurs (plusieurs méthodes pouvant aboutir à un même classifieur) et 230 activités enzymatiques (voir tableau 13.4).

Analyse du groupe taxonomique des agaricomycotina (seuil 0,9). Ce groupe a été caractérisé par un unique classifieur obtenu par l'algorithme RIPPER. La règle est la suivante :

Si les activités enzymatiques 1.5.1.9 et 1.5.1.8 sont présentes et si l'activité enzymatique 2.1.1.205 est absente chez le champignon, alors il appartient aux agaricomycotina (33 champignons couverts, 2 faux positifs), sinon il appartient à un autre groupe (140 champignons couverts, 1 faux négatif) (notons que les activités enzymatiques 1.5.1.8 et 1.5.1.9 ont le même profil).

Les activités enzymatiques 1.5.1.9 et 1.5.1.8 sont présentes dans la voie de

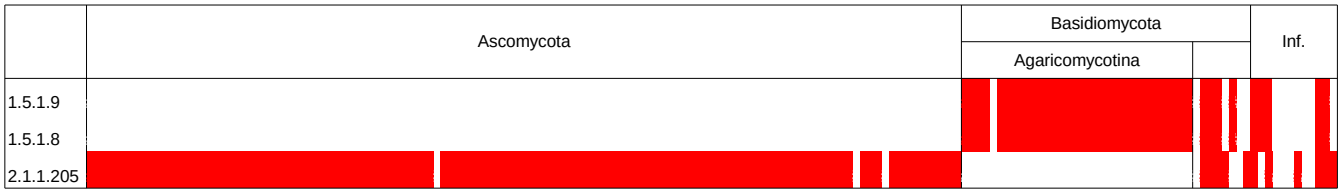


Figure 13.11 – *Profils des EC caractéristiques des agaricomycotina (jeu de données : 173 champignons traités avec MARIO).*

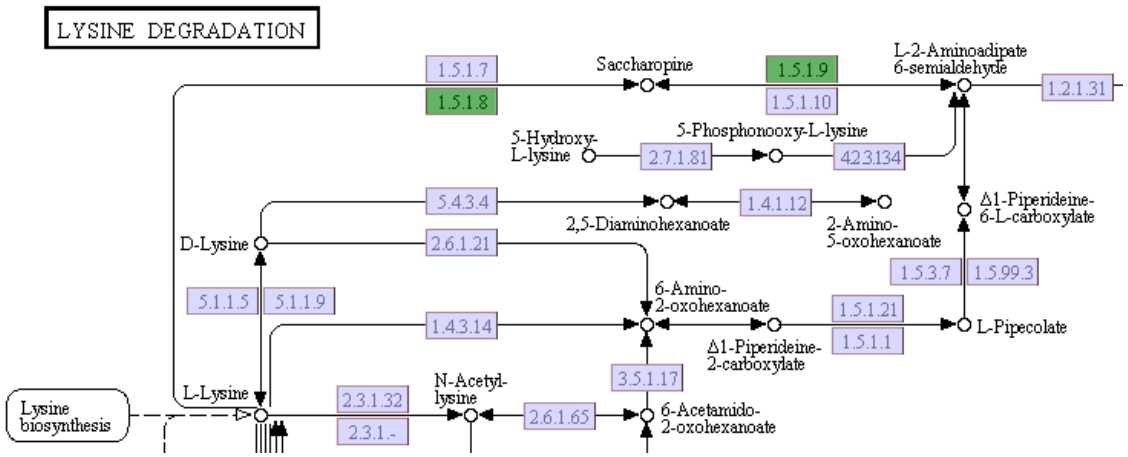


Figure 13.12 – *Caractérisation des agaricomycotina, EC numbers sélectionnés appartenant à la voie de dégradation de la lysine (KEGG).*

dégradation de la lysine (voir figure 13.12). Ces deux activités enzymatiques avaient déjà été sélectionnées lors de l'analyse présentée à JOBIM. Les agaricomycotina regroupent les pourritures blanches et brunes. Cette dégradation de la lysine est une entrée possible vers le cycle du cytrate ainsi que vers la voie de production de la penicilline et de la cepholosporine (antibiotiques).

L'activité enzymatique 2.1.1.205 correspond à une tRNA (cytidine32 / guanosine34–2'–O)-methyltransferase. Dans la base de données SwissProt, il n'existe pour le moment que deux champignons annotés avec cette activité enzymatique, *S. cerevisiae* et *S. pombe*. Ces deux champignons appartiennent au groupe taxonomique des ascomycota (alors que les agaricomycotina appartiennent aux basidiomycota).

L'activité enzymatique permet de discriminer d'autres basidiomycota et des champignons inférieurs des agaricomycotina. Cette activité enzymatique n'avait pas été sélectionnée dans la version présentée à JOBIM.

Analyse du classifieur obtenu pour les basidiomycota (seuil 0,9). L'application des trois algorithmes de classification à la caractérisation des basidiomycota aboutissent aux même règles (dans le cas de C4.5 il est possible d'extraire de l'arbre les règles associées).

Si l'activité enzymatique 2.7.1.36 est absente et que l'activité enzymatique 4.2.1.17 est présente, alors le champignon appartient au groupe taxonomique des basidiomycota, sinon il appartient à un autre groupe. Ce classifieur ne présente pas d'erreur sur le jeu de données. Les profils associés à ce classifieur sont visibles sur la figure 13.13.

Figure 13.13 – *Profils des EC caractéristiques des basidiomycota (jeu de données : 173 champignons traités avec MARIO)*

L'activité enzymatique 2.7.1.36 correspond à la mevalonate kinase, présente dans la voie de biosynthèse des terpénoïdes (voie du métabolisme secondaire). Elle avait déjà été sélectionnée lors de l'analyse présentée à JOBIM sur la version précédente de FungiPath. Cette enzyme est une transférase, elle catalyse l'ajout d'un groupement phosphate sur le mevalonate. La voie du mevalonate est présente à la fois chez les eucaryotes et chez les procaryotes (Lombard and Moreira, 2011) il est donc étonnant de ne pas trouver cette activité enzymatique chez les basidiomycota. Cependant, elle n'est décrite dans SwissProt que chez *S. cerevisiae*. Si cette voie est effectivement présente chez les basidiomycota, il est possible qu'un groupe

d'orthologues non annoté dans FungiPath permette la production de cette activité enzymatique (groupe possiblement homologue au groupe annoté 2.7.1.36) ou bien qu'une voie alternative soit utilisée chez ces organismes.

L'activité enzymatique 4.2.1.17 correspond à l'Enoyl-CoA hydratase. Elle est présente dans vingt voies métaboliques de KEGG, dont la voie décrivant la biosynthèse de métabolites secondaires. Elle intervient dans la béta–oxydation, voie permettant la dégradation des molécules d'acide gras. Contrairement à l'EC:2.7.1.36, celle-ci n'avait pas été sélectionnée dans les analyses précédentes. Il s'agit d'un attribut sélectionné sur l'analyse des champignons ne présentant pas l'EC:2.7.1.36. Il pourrait s'agir d'un cas de sur-apprentissage.

Les basidiomycota se caractériseraient donc en partie par une absence d'homologue connu de la mevalonate kinase.

Analyse du classifieur obtenu pour les taphrinomycotina (seuil 0,9). Dans la base de données FungiPath, quatre champignons appartiennent au groupe des taphrinomycotina. Le classifieur les caractérisant est composé d'une unique activité enzymatique. Si le champignon possède l'activité enzymatique 2.4.1.134 (3-beta-galactosyltransferase), alors il appartient au groupe des taphrinomycotina, sinon il appartient à un autre groupe (100% de bonne classification sur l'ensemble de nos données). Cette activité enzymatique n'est décrite chez les champignons que chez *S. pombe*, un taphrinomycotina. Elle est impliquée dans la création de la paroi cellulaire ainsi que dans la méiose.

Figure 13.14 – *Profils des EC caractéristiques des dikarya (jeu de données : 173 champignons traités avec MARIO)*

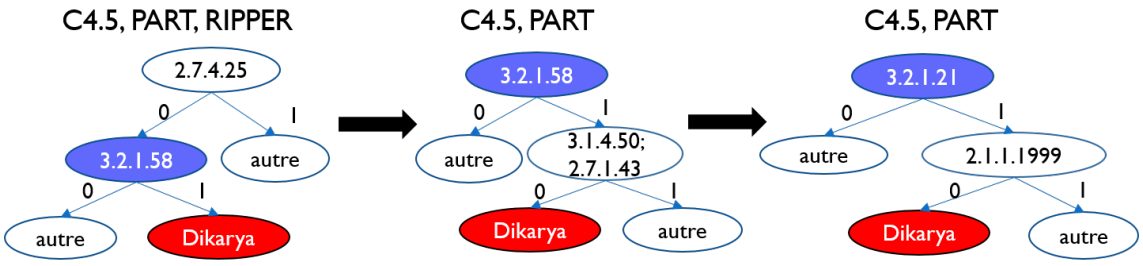


Figure 13.15 – *Classifieurs obtenus pour le groupe taxonomique des dikarya faisant intervenir au moins une activité enzymatique présente dans le métabolisme du saccharose et de l'amidon. Le 0 figure une activité enzymatique absente, le 1 une activité enzymatique présente.*

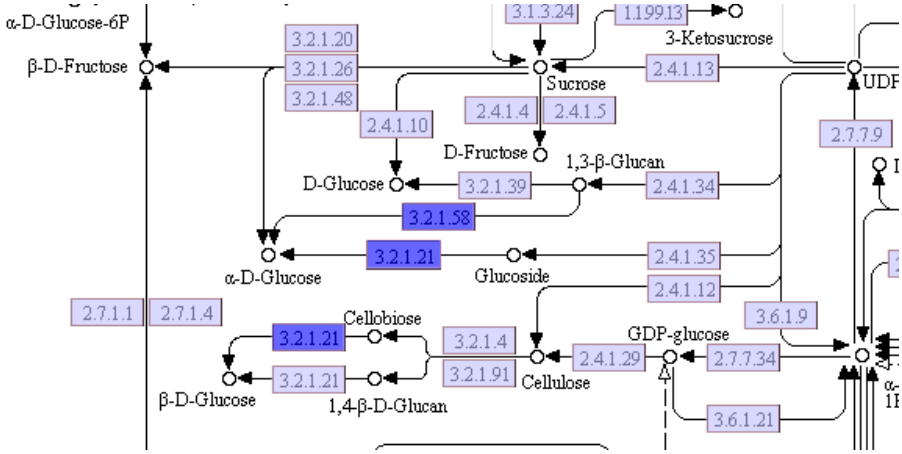


Figure 13.16 – *Caractérisation des dikarya, EC numbers sélectionnés appartenant au métabolisme du saccharose et de l'amidon.*

Analyse des classifieurs obtenus pour le groupe des dikarya (seuil 0,9). La caractérisation des dikarya a permis de trouver plusieurs classifieurs pertinents (voir les figures 13.14 et 13.15). Nous n'avions pas caractérisé ce groupe sur les données à 50 génomes. Si l'on compare les résultats obtenus sur cette nouvelle version de FUNGIpath (MARIO) et ceux obtenus sur FungiPath 178 v1 on observe que 6 des 7 profils sélectionnés sur la version MARIO l'étaient déjà dans les analyses en partie présentées à JOBIM.

Parmi les activités enzymatiques sélectionnées par les classifieurs obtenus sur FungiPath 178 version MARIO, les EC:3.2.1.21 et l'EC:3.2.1.58 appartiennent à une même voie, la voie métabolique du saccharose et de l'amidon. Ces activités enzymatiques sont impliquées dans la dégradation de la biomasse, que ce soit de la cellobiose (produit de dégradation de la cellulose) ou de beta-glucanes (ex : cellulose). On remarque qu'elles ont été sélectionnées dans deux classifieurs différents. Ces deux activités enzymatiques pourraient faire partie d'un même module fonctionnel, aux profils semblables, et pour lequel l'ensemble des activités enzymatiques est pertinent. Ainsi, la capacité à dégrader la cellulose est caractéristique des dikarya.

Plusieurs activités enzymatiques sélectionnées permettent de modifier la paroi des champignons (EC: 3.2.1.58, EC: 2.1.1.1, EC: 3.2.1.21). Il avait déjà été montré que la paroi présente des différences majeur en fonction des groupes taxonomiques (**?**). Que l'ensemble des dikarya possèdent ces activités enzymatiques pourrait signifier qu'elles sont nécessaire aux dikarya ou encore qu'il s'agit d'une famille de protéines ayant différentes mutations en fonction du groupe taxonomique.

Analyse des classifieurs obtenus pour le groupe des ascomycota. Les 94 classifieurs obtenus pour les ascomycota sont trop nombreux pour que nous puissions tous les analyser séparément. Cependant, nous observons que l'activité enzymatique 3.4.23.41, trouvée caractéristique sur l'analyse de la version à 50 génomes, ne l'est plus ici. La séquence que nous avions délétée chez *Podospora anserina* (HpYPS1) appartient dans cette nouvelle version à un groupe de 359 protéines contenant majoritairement des ascomycota mais présentant également des séquences pour trois zygomycota et deux basidiomycota. Ce groupe, bien qu'il soit préférentiellement présent chez les ascomycota, n'a pas été sélectionné par les méthodes de classification. Le groupe d'HpYPS1 a, comme précédemment, été annoté avec une protéine de la famille des pepsines. Cette famille de protéines a des substrats différents en fonction de l'espèce dans laquelle elle est décrite. Cela a induit l'annotation de ce groupe avec un *EC number* différent (3.4.23.24) dans la dernière version de FungiPath.

Analyse générale des classifieurs. Lorsqu'il existe plusieurs dizaines de classifieurs, il devient difficile de les analyser un à un.

Nous avons procédé de la manière suivante: pour chaque groupe taxonomique, nous avons listé l'ensemble des activités enzymatiques se trouvant dans au moins un classifieur caractéristique de ce groupe. Nous obtenons une liste d'activités enzymatiques par groupe taxonomique. Pour chacune de ces listes, nous avons ensuite recherché les activités enzymatiques présentes dans la même voie KEGG.

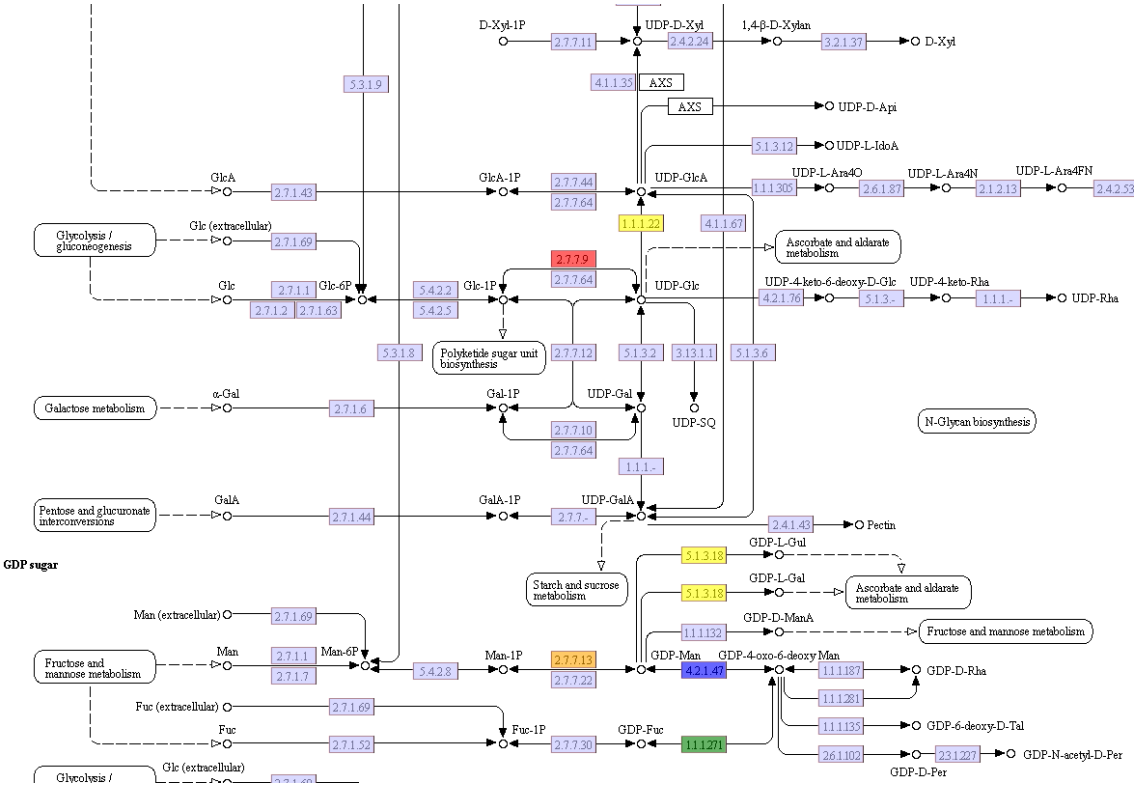


Figure 13.17 – *Caractérisation des ascomycota, les EC numbers sélectionnés appartiennent à la voie 'Amino sugar and nucleotide sugar metabolism' (KEGG). Deux EC sont présentés de la même couleur si ils ont été séléctionnés conjointement dans un même classifieur. La couleur des EC number dépend donc du classifieur dans lequel ils ont été sélectionnés.*

Nous obtenons ainsi 57 paires 'groupe taxonomique – voie métabolique' pour lesquelles au moins deux *EC numbers* sont présents. Parmi elles, 25 présentent au moins un lien entre deux des *EC numbers* sélectionnés et 11 forment une unique composante connexe.

Nous observons que lorsque plusieurs *EC numbers* caractéristiques d'un même groupe taxonomique sont connexes dans les graphes métaboliques de KEGG ils sont généralement issus de classifieurs différents (voir la figure 13.17).

Ces composantes connexes pourraient être des modules fonctionnels. La sélection d'une seule activité enzymatique suffirait dans ce cas à caractériser l'ensemble du module. Ainsi, la recherche itérative de différents classifieurs permet de retrouver ces possibles modules.

Nous observons la sélection d'EC impliqués dans des voies métaboliques. Nous pouvons de ce fait analyser si des voies sont sur-représentées dans ces ensembles de profils.

Nous avons cherché à déterminer si une des voies trouvées était sur-représentée au sein de la liste des activités enzymatiques décrivant chacun des groupes. Pour cela, nous avons utilisé un test exact de Fisher de comparaison des proportions (comparaison de la proportion de la voie présente dans l'échantillonnage avec la proportion d'EC de la voie parmi les listes des EC des champignons). Nous effectuons ce test de comparaison de proportions pour chacune des voies présentes dans la liste. Pour chaque test, nous acceptons une marge d'erreurs de 5%. De manière à ce que la multitude de tests n'induise pas la sélection de résultats appartenant aux 5% d'erreurs, nous avons choisi de corriger les résultats avec la méthode de Bonferroni. Cette méthode, simple et conservative, consiste en la multiplication de la p-value par le nombre de tests effectués. On obtient une p-value corrigée.

Les résultats de cette analyse sont observables sur le tableau 13.5.

TABLE 13.5 – *Voies sur-représentées parmi les profils sélectionnés par les classifieurs.*

Groupe taxonomique	# EC sélectionnés	nom de la voie	# d'EC sélectionnés appartenant à la voie	% d'EC sélectionnés appartenant à la voie	p-value Fisher	p-value corrigée	liste des EC sélectionnés
Agaricomycotina	3	Lysine degradation	2	66,67%	0,0024	0,0072	1.5.1.8; 1.5.1.9
Taphrinomycotina	1	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	1	100,00%	0,0016	0,0047	2.4.1.134
Taphrinomycotina	1	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	1	100,00%	0,0016	0,0047	2.4.1.134

Parmi les voies sélectionnées, on retrouve les voies de dégradation de la lysine et de synthèse de glycosaminoglycan que nous avons décrites précédemment lors de l'analyse des taphrinomycotina.

13.3.4 Application de classifieurs : conclusions et discussions

Nous observons que lorsque plusieurs *EC numbers* de la même voie sont sélectionnés pour caractériser un même groupe taxonomique, ces *EC numbers* ont tendance à former un sous-graphe connexe (dans 25 cas sur 57) et à avoir été sélectionnés par des classifieurs différents. Une hypothèse pourrait être que nous observons ici la sélection des mêmes modules fonctionnels par différents classifieurs.

Les groupes taxonomiques n'ont pas tous pu être caractérisés par des classifieurs. La proportion de champignons appartenant au groupe taxonomique testé est peut-être souvent trop faible pour arriver à un classifieur pertinent. En effet, on observe que parmi les sept groupes taxonomiques caractérisés quatre sont ceux contenant le plus grand nombre d'exemples positifs. Une solution pourrait être d'augmenter le nombre d'exemples dans la base afin d'augmenter dans le même temps les informations disponibles pour apprendre les classifieurs. De plus, augmenter le nombre de champignons pourrait nous permettre de vérifier les classifieurs obtenus dans cette analyse.

Nous observons que les activités enzymatiques sélectionnées en premier par les classifieurs sont généralement celles qui sont les mieux conservées dans les résultats obtenus sur les différents jeux de données (profil des EC:1.5.1.8 et EC:1.5.1.9 pour les agaricomycetes ou profils de l'EC:2.3.1.36 des basidiomycota). Les attributs suivants sont sélectionnés sur la base d'observation sur une partie des données. Il y a alors deux possibilités. On peut supposer que cette partie des données est trop petite pour induire la sélection d'attributs pertinents ou bien que plusieurs profils sont pertinents et que la méthode n'en retourne qu'un parmi ceux-ci.

Il existe un phénomène d'échantillonnage (*Peaking phenomenon*). Lorsqu'il y a un grand nombre d'attributs par rapport au nombre d'exemples, certains attributs qui ne représentent que du bruit peuvent être pris en compte dans les classifieurs. Afin de limiter ce phénomène, l'une des solutions peut être de présélectionner les attributs (Patil and Bichkar, 2012). De ce fait, les attributs trop peu discriminants des données initiales ne seront pas sélectionnés sur les sous-jeux de données induits par les premières règles formées par un classifieur. Dans les sections suivantes, nous testons l'impact de méthodes de sélection d'attributs sur les classifieurs.

13.4 Méthodes de sélection d'attributs

13.4.1 Description des méthodes

Bien que les méthodes de type arbre de décision puissent être utilisées comme des méthodes de filtres des attributs, il est connu qu'elles présentent de moins bons résultats s'il existe des attributs non pertinents. Ce phénomène s'explique par le fait que la sélection d'attributs se fait un attribut à la fois et seulement sur les exemples couverts par la condition du nœud précédent. De ce fait, les attributs sont sélectionnés en fonction de leur pouvoir prédictif local, celui-ci pouvant avoir été évalué sur un faible nombre d'exemples (Perner, 2001). Ainsi, effectuer un pré-filtre des données avant d'utiliser un algorithme de type arbre de décision peut permettre d'obtenir de meilleurs résultats (Perner, 2001).

Le but de la sélection d'attributs est de diminuer le nombre d'attributs tout en conservant autant d'information que possible pour résoudre dans un second temps le problème de classification (Jain and Zongker, 1997) (voir figure 13.2). Les méthodes de sélection d'attributs permettent de retirer les attributs redondants ou non-signifiants (Saeys et al., 2007, Bolón-Canedo et al., 2013). Ces méthodes ont été utilisées en bioinformatique (Saeys et al., 2007, Ma and Huang, 2008) dans différents

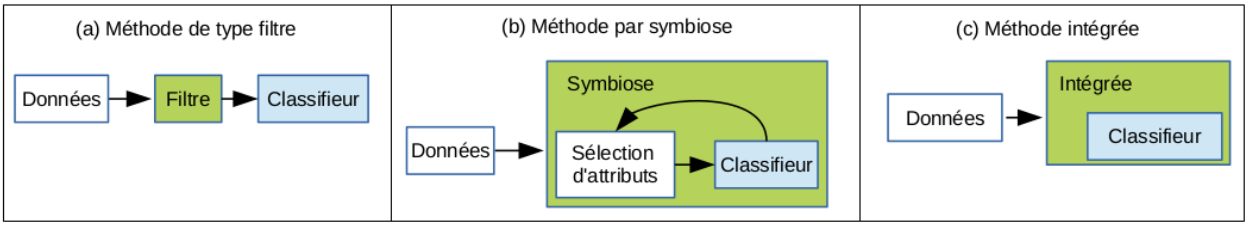


Figure 13.18 – *Schéma des techniques de sélection d'attributs. Figure adaptée de (Bolón-Canedo et al., 2013)*

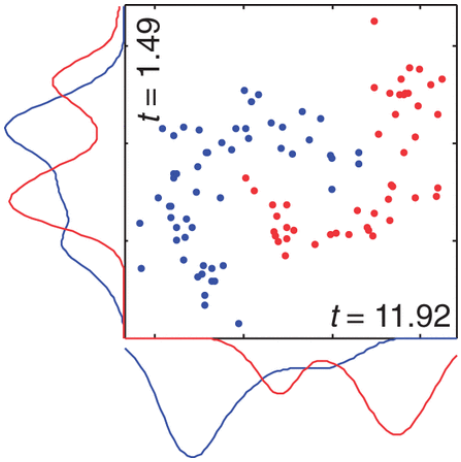


Figure 13.19 – *Sélection d'attributs basés sur le test de Student. L'attribut x (abscisses) discrimine mieux les données que l'attribut y (ordonnées). Figure tirée de (de Ridder et al., 2013)*

buts tels que la prédiction de séquences codantes (Salzberg et al., 1998) ou l'analyse de données de microarray (Jafari and Azuaje, 2006).

Les données de FungiPath présentent la particularité d'avoir un plus grand nombre d'attributs (les activités enzymatiques) que d'exemples (les espèces de champignons), appliquer de telles méthodes peut donc permettre de rééquilibrer les données.

Il existe trois types de méthodes de sélections d'attributs que nous décrivons dans les trois paragraphes suivants (voir la figure 13.18).

Les méthodes de type filtre (Guyon and Elisseeff, 2003, Bolón-Canedo et al., 2013) utilisent les caractéristiques générales des données. Elles ordonnent les attributs en fonction de leur pertinence. L'évaluation de la pertinence d'un attribut est faite grâce au calcul d'un score dépendant de la méthode choisie. Différentes mesures de score sont possibles, tel que par exemple *Student t test* (voir figure 13.19). Ce type de filtre est simple à implémenter et rapide en temps de calcul.

Parmi les méthodes de type filtre nous pouvons décrire les méthodes ReliefF et gain d'information que nous réutiliserons par la suite.

L'algorithme **ReliefF** recherche les attributs les plus pertinents et les classes indépendamment de tout algorithme d'apprentissage. Cette sélection d'attributs nécessite que les exemples soient étiquetés. Il approxime le poids de chaque attribut en utilisant en modèle sous-jacent un modèle à plus proche voisin. Les attributs les plus pertinents sont alors ceux qui varient plus lorsque l'exemple change de classe que lorsqu'il n'en change pas. Le poids des attributs varie entre -1 et 1. Les grandes valeurs positives étant assignées aux attributs les plus importants.

Le **gain d'information** (ou réduction d'entropie) est utilisé pour mesurer la dépendance entre un attribut et une classe. L'entropie est une mesure de l'incertitude moyenne d'un résultat. Il s'agit de l'incertitude *a priori* que l'output d'une expérience aléatoire décrite par P soit observé. Le gain d'information

correspond à la réduction d'entropie due à un tri suivant les valeurs de l'attribut. Soit E un ensemble d'exemples, A l'attribut et x la classe. Le calcul du gain d'information relatif à l'attribut A (entre A et x) se fait de la manière suivante : $GI(E,A) = H(E) - \sum_{v \in Valeurs(A)} \frac{|E_v|}{|E|} H(E_v)$ avec $H(E)$ l'entropie de l'ensemble des exemples E et $H(E_v)$ est l'entropie des exemples présentant la valeur v pour l'attribut A ($E_v \in E$). $H(E) = -\sum_c p(A_c) \ln(p(E_c))$, avec $p(E_c)$ la proportion d'exemples appartenant à la classe c. L'une des critiques communes faites aux méthodes de type filtre est qu'elles induisent la sélection d'attributs redondants. Ainsi une performance équivalente pourrait être obtenue avec un plus petit ensemble d'attributs complémentaires. Cependant, en biologie, trouver un ensemble d'attributs redondant n'est pas problématique, cela pouvant refléter une réalité biologique de synergie entre ces différents attributs.

Les approches par symbiose (ou méthode de type ***wrapper***) (voir figure 13.18.b) optimisent un prédicteur dans la phase de sélection des attributs. Le classifieur fonctionne alors comme une boite noire. Ces approches ont pour but de sélectionner **conjointement** un ensemble de variables présentant un bon pouvoir prédictif. Dans le cas d'un travail avec m attributs, l'ensemble des combinaisons possibles d'attributs est $O(2^m)$. Tester l'ensemble des combinaisons n'est donc pas envisageable à moins que m soit petit. Les méthodes wrapper utilisent généralement une recherche gloutonne [2] dans l'espace des attributs. Ces méthodes présentent de meilleurs résultats que les méthodes de type filtre, mais leur temps de calcul est en contrepartie beaucoup plus long. Elles nécessitent en effet d'effectuer un grand nombre de fois la boucle sélection d'attributs / évaluation par le classifieur. C'est pourquoi nous n'avons pas travaillé avec ce type d'approche.

Les méthodes intégrées (ou ***embedded***)(voir figure 13.18.c) effectuent des sélections d'attributs lors de l'apprentissage (un seul passage). Elles permettent

2. principe de faire, étape par étape, un choix optimum local, dans l'espoir d'obtenir un résultat optimum global

ainsi d'obtenir les avantages des deux types de méthodes précédemment présentées, elles interagissent avec le classifieur (comme les méthodes de symbiose) et elles sont plus rapides que les méthodes de type symbiose (comme les méthodes de type filtre). Les méthodes intégrées sont généralement spécifiques à une méthode d'apprentissage supervisée donnée. Contrairement à la plupart des méthodes de type filtre, elles sont capables de détecter des dépendances entre attributs.

Parmi les méthodes intégrées, les méthodes de type élagage éliminent de manière récursive l'attribut le moins pertinent. Cette méthode a été utilisée avec l'algorithme SVM (*support vector machines*) (Guyon et al., 2002). Les algorithmes de type arbres de décision (ID3, C4.5) apprennent le classifieur et sélectionnent les attributs dans le même temps. Les attributs sont sélectionnés au moment de la création des nœuds de l'arbre. Enfin, il existe des méthodes basées sur la régularisation. C'est le cas par exemple de la sélection d'attributs de type lasso. L'addition de contraintes au moment de l'optimisation du problème permet de réduire sa complexité.

Nous souhaitons appliquer les méthodes de type filtre afin de limiter la sélection d'attributs peu pertinents par nos classifieurs. La première étape de l'analyse est d'observer quels sont les profils sélectionnés par ces méthodes.

13.4.2 Résultats de l'application de filtre

Nous avons testé deux méthodes de filtre, le ratio de gain d'information et reliefF, afin de sélectionner les profils phylogénétiques pertinents pour la caractérisation de la taxonomie.

13.4.2.1 Ratio de gain d'information

On appelle 'ratio de gain d'information' (abrégé RIG) (Yang and Pedersen, 1997) le gain d'information normalisé par l'information générée en divisant le jeu de

donnée S en v partitions correspondantes aux v partitions induites par l'application de l'attribut A.

$$SplitInfo_A(S) = (-\sum_{i=1}^{v} |S_i|/|S|)log_2(|S_i|/|S|)$$

$$GainRatio(A) = Gain(A)/SplitInfo_A(S)$$

Le ratio de gain d'information permet de classer l'ensemble des attributs en fonction de leur pertinence pour caractériser une classe donnée sans tenir compte des interactions possibles entre attributs (combinaison d'attributs).

Le gain d'information est le score utilisé par nos méthodes de classification. Nous avons choisi de tester ce filtre afin d'observer si une sélection portant sur le même score que la méthode de classification aboutit à l'apprentissage de classifieurs différents.

Pour chaque profil, la valeur du ratio de gain d'information pour une classe donnée est comprise entre 0 (pas d'information) et 1 (profil discriminant parfaitement les classes). Les courbes de densité des valeurs de ratio de gain d'information sont observables sur la figure 13.20. Ces courbes présentent un pic autour de 0,025 (ce qui est très inférieur à 1) et cela pour chacun des groupes taxonomiques testés. La majorité des profils pris séparément ne sont donc pas à eux seuls pertinents pour la caractérisation de ces groupes. Filtrer ces profils afin de retirer les profils non informatifs semble donc une approche cohérente. Il est à noter que ces courbes ne donnent d'information que sur la pertinence des profils pris séparément, elles ne fournissent pas d'information sur la pertinence d'une combinaison de profils. Il est donc possible de sélectionner trop ou trop peu de profils. Dans le but de comparer les résultats induits par la sélection d'ensembles plus ou moins limités de profils nous avons choisi de tester la présélection de 10 ou 100 profils. C'est un moyen de calibrer notre méthode et d'analyser la différence des résultats obtenus. Nous pourrons ainsi déduire si la sélection de 100 profils ou même 10 profils est suffisante pour obtenir des classifieurs.

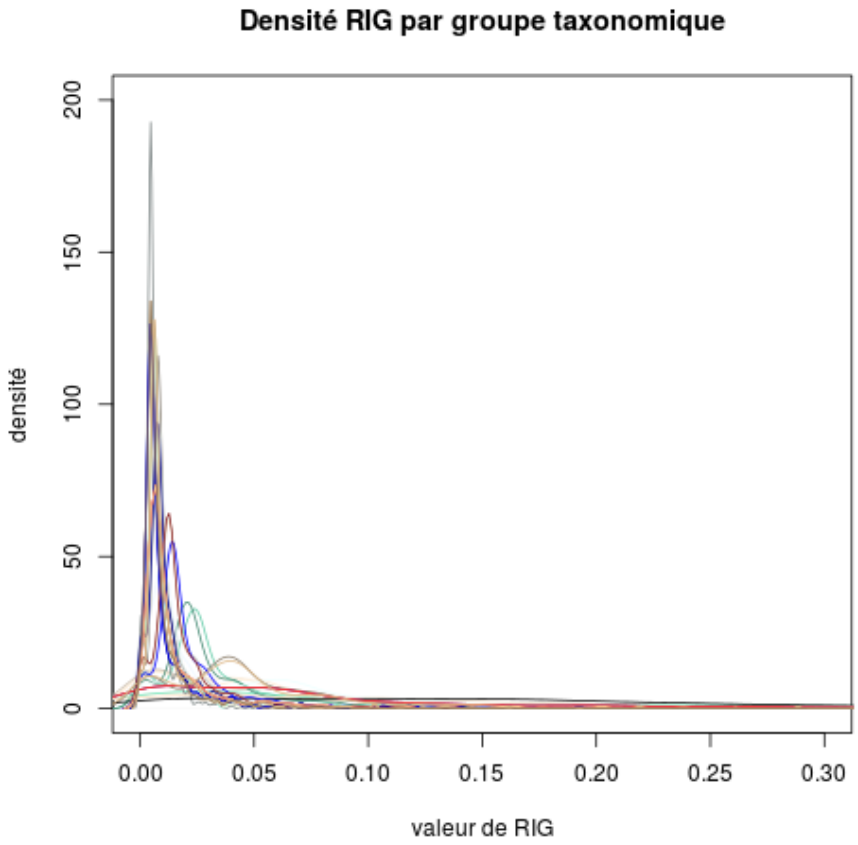


Figure 13.20 – *Densité de la valeur de ratio de gain d'information obtenu avec les différents groupes taxonomiques. Chaque courbe correspond aux valeurs obtenues pour un groupe taxonomique.*

Nous avons donc ici sélectionné pour chacun des 29 groupes taxonomiques les 10 ou les 100 profils les plus pertinents (soit près de 1/10 des profils initiaux). Le RIG moyen des profils sélectionnés est donné en annexe dans le tableau IV.

Analyse pour chaque groupe taxonomique des caractéristiques des 10 profils au plus grand RIG. Plusieurs activités enzymatiques sélectionnées des profils sélectionnés apparaissent dans au moins une voie métabolique de KEGG. Nous avons analysé indépendamment les EC sélectionnés pour chacun des groupes

taxonomiques. Pour chaque voie métabolique de KEGG, nous avons ensuite observé la connexité entre ces EC caractéristiques d'un groupe donné.

Parmi nos résultats, nous trouvons 78 paires 'groupe taxonomique – voie métabolique contenant au moins deux *EC numbers*'. Dans 28,2% des cas (22 sur 78), nous observons que les activités enzymatiques ne sont pas isolées (nombre d'EC $>$ nombre de composantes connexes). Il s'agit de paires d'ECs.

À des fins de comparaison, nous avons tiré aléatoirement mille fois deux activités enzymatiques dans chacun des graphes KEGG et calculé le pourcentage de fois où ces activités enzymatiques sont connexes. Nous obtenons une moyenne de 21,32%. La proportion de composantes connexes retrouvées dans nos paires semble donc supérieure à ce qui pourrait être attendu.

Nous avons également traité les EC présentant les profils sélectionnés indépendamment les uns des autres et observé leurs degrés (nombre de connexions) dans chacune des voies métaboliques de KEGG. Pour cela, nous utilisons la librairie KEGGgraph (Zhang and Wiemann, 2009) disponible pour R. Nous avons comparé ce nombre avec le degré moyen des *EC numbers* de chacune de ces voies avec un test de Welch. Nous n'observons pas de différence significative entre le degré moyen des *EC numbers* des voies et le degré des *EC numbers* sélectionnés.

Voies sur-représentées parmi les 10 premiers profils caractéristiques de chaque groupe taxonomique Nous avons échantillonné pour chaque groupe taxonomique les 10 profils présentant la plus forte valeur de RIG.

Afin d'observer si certaines voies sont plus souvent impliquées dans l'évolution du métabolisme, nous comparons par groupe taxonomique et pour chaque voie de KEGG la proportion d'*EC numbers* des profils échantillonnés appartenant à cette voie avec la proportion d'*EC number* appartenant à cette même voie dans FungiPath.

Nous avons pour cela utilisé des tests exacts de Fisher (avec l'alternative

'plus grande que') et nous les avons corrigés avec la méthode de Bonferroni (H0 : la voie x n'est pas sur-représentée au sein des EC associés aux profils sélectionnés, H1 : la voie x est sur-représentée au sein des EC associés aux profils sélectionnés). Pour chaque test de Fisher, nous acceptons une marge d'erreur de 5%. Nous avons appliqué la correction afin d'éviter la sélection de voies due au grand nombre de tests.

Table 13.6 – *Pour chaque groupe taxonomique, liste des voies KEGG sur-représentées dans la liste d'EC appartenant aux profils classés parmi les 10 meilleurs par la méthode de filtre 'Ratio gain d'information' (RGI).*

Groupe taxonomique	# EC select.	nom voie	# de la voie	# d'EC select. ∈ voie	% d'EC select. ∈ voie	liste des EC select.	p-value Fisher	p-value corrigée
Agaricales	20	Aminoacyl-tRNA biosynthesis	970	5	25,00%	6.1.1.5 ; 6.1.1.1 ; 6.1.1.17 ; 6.1.1.19 ; 6.1.1.12	4,48E-05	6,71E-04
Agaricomycetes	15	Sesquiterpe-noid and triterpenoid biosynthesis	909	2	13,33%	4.2.3.23 ; 3.1.7.6	4,35E-03	3,48E-02
Basidiomycota	14	Lysine degradation	310	3	21,43%	1.5.1.8 ; 2.6.1.39 ; 1.5.1.9	2,11E-03	2,74E-02
Dothideomycetes	13	Arachidonic acid metabolism	590	2	15,38%	1.1.1.189 ; 1.1.1.184	3,38E-03	4,73E-02
Eurotiales	13	Lipopoly-saccharide biosynthesis	540	2	15,38%	2.7.1.167 ; 2.7.7.70	1,27E-04	1,40E-03
Eurotiales	13	Peptidoglycan biosynthesis	550	2	15,38%	6.3.2.8 ; 1.3.1.98	3,77E-04	4,15E-03
Eurotiomycetes	11	Lipopoly-saccharide biosynthesis	540	2	18,18%	2.7.1.167 ; 2.7.7.70	9,43E-05	1,13E-03
Glomerellales	14	Thiamine metabolism	730	2	14,29%	2.7.1.50 ; 2.5.1.3	3,85E-03	4,62E-02
Pleosporales	12	Arachidonic acid metabolism	590	2	16,67%	1.1.1.189 ; 1.1.1.184	2,94E-03	2,94E-02
Taphrinomycotina	16	Ubiquinone and other terpenoid-quinone biosynthesis	130	3	18,75%	2.1.1.114 ; 2.1.1.64 ; 2.1.1.222	1,44E-04	2,88E-03
Tremellomycetes	10	N-Glycan biosynthesis	510	3	30,00%	2.4.1.267 ; 2.4.1.256 ; 3.2.1.106	9,25E-04	2,78E-02

On observe onze couples 'groupe taxonomique / voie sur-représentée' (voir tableau 13.6).

Vu que les profils peuvent parfois être annotés avec plusieurs *EC numbers*, il est possible que cette évaluation seule ne soit pas suffisante. En effet, l'échantillonnage par profil induit peut-être la sélection préférentielle de certaines voies. Dans le but d'observer ce phénomène, nous testons si la sélection aléatoire de profils a statistiquement tendance à sélectionner des profils d'*EC numbers* impliqués dans des voies particulières.

Nous avons effectué 1000 tirages aléatoires sans remise de 10 profils et nous avons analysé la liste des *EC numbers* ainsi sélectionnés de la même manière que pour les profils échantillonnés avec le filtre RIG. Un profil peut figurer la présence/ absence de plusieurs *EC number*. Le nombre d' *EC number* sélectionné est donc plus grand que le nombre de profils. Celui-ci varie entre 10 et 33 (moyenne de 16.64). Parmi les 1000 tirages, on trouve 195 couples 'un tirage aléatoire / une voie sur-représentée' (166 tirages différents). Ainsi, 16,6% des tirages aléatoires ont abouti à la sélection d'au moins une voie.

Ces 195 couples 'tirage aléatoire / voie sur-représentée' font intervenir 60 voies différentes (voir la figure 13.21). Parmi elles, certaines sont très fortement représentées, il pourrait s'agir de voies présentant des *EC* ayant le même profil. En effet, dans le cas d' *EC numbers* partageant le même profil, nous avions observé que un tiers était présent dans la même voie.

Nous avons précédemment sélectionné les 10 profils au plus grand RIG pour 21 groupes taxonomiques. Parmi les voies sur-représentées au sein de ces 10 profils sélectionnés et peu ou pas présentes parmi les voies sur-représentées dans les tirages aléatoires (voir le tableau 13.6), on trouve : la biosynthèse de peptidoglycanes, la voie de dégradation de la lysine, la biosynthèse d'aminoacyl-tRNA, la biosynthèse de sesquiterpenoid, et de triterpenoid et la biosynthèse de peptidoglycanes.

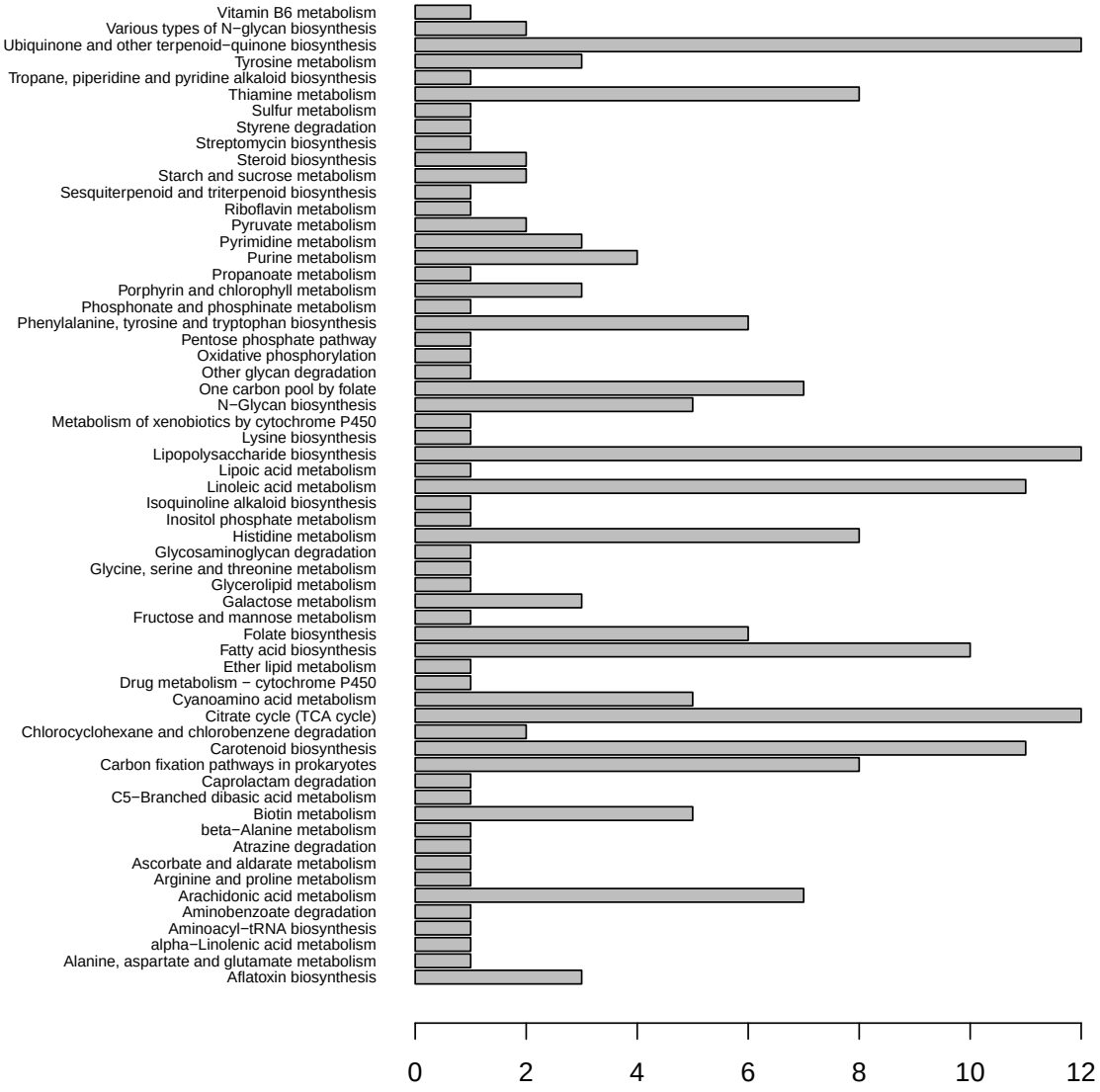


Figure 13.21 – *Nombre de fois où chacune des voies a été retrouvée de manière significative parmi les EC sélectionnés par un tirage aléatoire de 10 profils.*

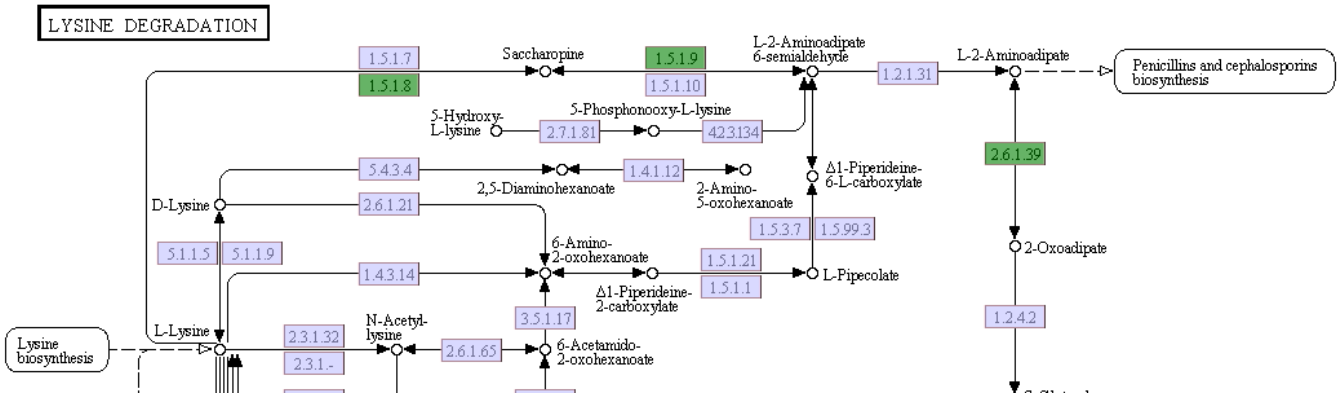


Figure 13.22 – *Exemple de voie sur-représentée parmi les 10 profils au plus fort RIG pour le groupe taxonomique des Basidiomycota. La voie de dégradation de la Lysine est statistiquement sur-représentée chez les Basidiomycota. Les EC numbers sur fond vert représentent les EC numbers de profils ayant parmi les 10 plus fortes valeurs de RIG.*

Les voies caractérisant la taxonomie des champignons (d'après cette méthode) sont donc majoritairement des voies de synthèse de métabolites et appartiennent au métabolisme secondaire.

La voie de dégradation de la lysine (voir figure 13.22) permet par exemple l'entrée dans la voie de synthèse de la pénicilline (antibiotique découvert pour la première fois chez les champignons). La possibilité de dégrader la lysine afin d'entrer dans la voie de production de cet antibiotique semble plus présente chez les basidiomycota que chez les autres champignons.

Pour chaque groupe taxonomique, analyse des caractéristiques des 100 profils au plus grand RIG. Comme nous en avons discuté dans l'introduction de cette section, l'analyse des 10 profils au plus grand RIG aboutit peut-être à une sélection trop stringente des profils. C'est pourquoi nous avons testé dans le même temps la sélection de 100 profils. Nous avons ainsi la possibilité d'analyser les résultats obtenus avec ces deux ensembles de profils.

Nous avons trouvé 782 paires associant un groupe taxonomique et une voie

contenant aux moins deux *EC* associés aux 100 profils ayant le plus fort RIG pour ce groupe.

Dans chacune des voies KEGG, le degré par voie des *EC numbers* sélectionnés est significativement plus grand que les degrés moyens des EC des voies (test de Welch alternative greater, p-value=4,00e-11).

Ainsi, ce serait plutôt au niveau de HUBS que les événements d'évolution conservés ou perdus pour l'ensemble d'un groupe taxonomique auraient lieu. Les nœuds à fort degré induisent certainement, au moment de leur gain ou de leur perte, une modification du phénotype de l'organisme. Cette modification induit peut-être la création d'une nouvelle branche dans la taxonomie.

Les 782 paires 'groupe taxonomique–voie sur-représentée parmi les *EC* des 100 profils à la plus forte valeur de RIG' correspondent à 92 voies différentes. Parmi ces paires 'groupe taxonomique – voie KEGG', 372 (48%) présentent au moins une paire d'*EC number* sélectionnés qui s'avèrent former un sous-graphe connexe et 143 (18%) ne sélectionnent que des *EC number* reliés. La caractérisation des microsporidia par exemple montre un exemple de sélection d'*EC numbers* connexes dans la voie du cycle du citrate (voir figure 13.23). Dans ce cas, c'est l'absence de cette voie qui caractérise ces champignons pathogènes obligatoires.

Lors de la sélection des 100 profils au plus grand RIG, la proportion d'*EC numbers* sélectionnés car ils présentent exactement le même profil diminue. Le tirage aléatoire de 1000 fois 100 profils n'a induit que l'obtention de deux paires 'tirage aléatoire / voie sur-représentée', ces deux voies étant 'Valine, leucine and isoleucine biosynthesis' et 'Butanoate metabolism'.

Parmi les listes de 100 profils au plus fort RIG obtenus pour chaque groupe taxonomique, trois présentent une sur-représentation d'au moins une voie métabolique (voir tableau 13.7). On observe (comme pour les 10 profils au plus fort RIG) la sélection de voies liées à la biosynthèse de composés. On retrouve par exemple

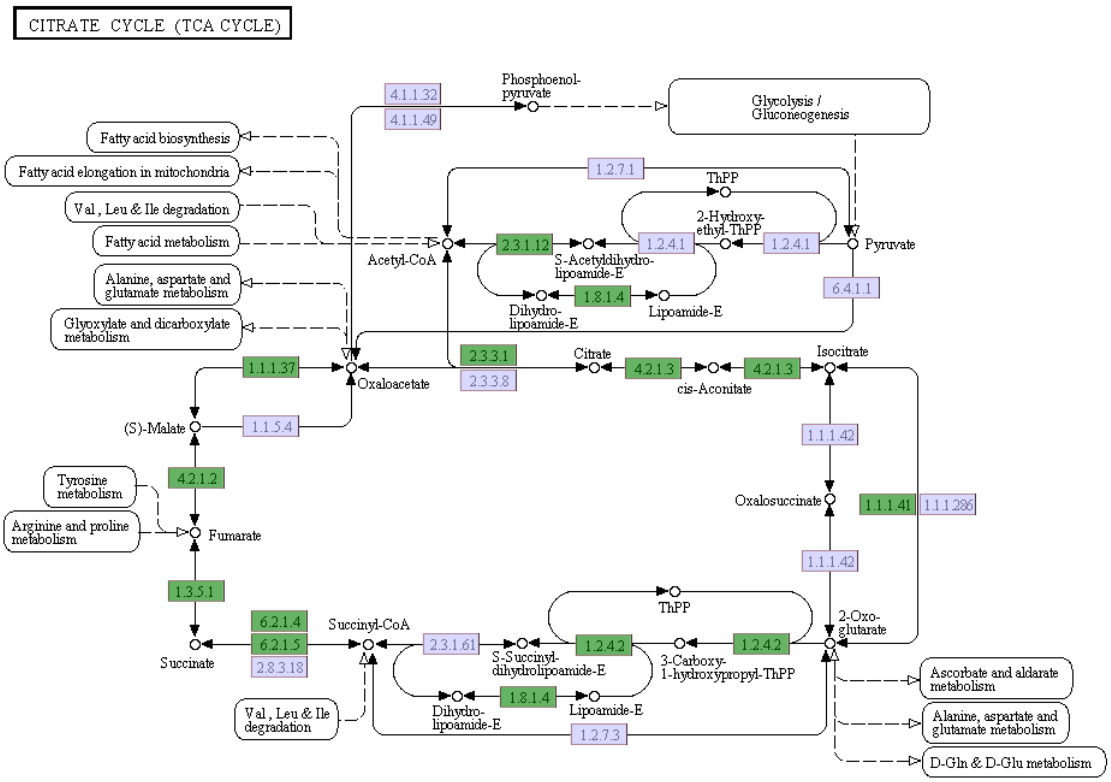


Figure 13.23 – *Représentation du cycle du citrate dans KEGG. Les EC numbers associés aux 100 profils ayant la plus grande valeur de RIG associé à la caractérisation du groupe taxonomique des Microsporidia sont représentés sur fond vert. On observe que l'ensemble de ces EC numbers sélectionnés forment un sous-graphe connexe.*

la sélection de la voie de biosynthèse de métabolites secondaires, à la fois parmi les *EC numbers* sélectionnés comme caractéristique des dikarya, mais aussi parmi ceux caractéristiques des microsporidia.

Les voies caractéristiques de nos groupes sont donc plutôt des voies du métabolisme secondaire, exception faite dans le cas des microsporidia. Les microsporidia sont des pathogènes obligatoires. Des voies du métabolisme primaire s'avèrent caractéristiques de ce groupe taxonomique, car elles ont en partie été perdues (utilisation de la machinerie cellulaire de l'hôte).

Table 13.7 – *Liste des voies sur-représentées dans la liste d'EC appartenant aux 100 premiers profils de chaque groupe taxonomique par la méthode de filtre 'Ratio gain d'information' (RIG).*

Groupe taxonomique	# EC select.	nom voie	# de la voie	# d'*EC* select. ∈ voie	% d'*EC* select. ∈ voie	p-value Fisher	p-value corrigée
Dikarya	184	Biosynthesis of secondary metabolites	1110	73	39,67%	1,11E-07	9,55E-06
Dikarya	184	Citrate cycle (TCA cycle)	20	11	5,98%	3,56E-06	3,06E-04
Dikarya	184	Phenylalanine, tyrosine and tryptophan biosynthesis	400	13	7,07%	7,59E-06	6,53E-04
Dikarya	184	Carbon fixation pathways in prokaryotes	720	10	5,43%	3,32E-04	2,85E-02
Dikarya	184	Valine, leucine and isoleucine biosynthesis	290	6	3,26%	5,32E-04	4,57E-02
Eurotiales	105	Peptidoglycan biosynthesis	550	3	2,86%	5,75E-04	3,68E-02
Microsporidia	164	Biosynthesis of secondary metabolites	1110	73	44,51%	1,98E-09	1,60E-07
Microsporidia	164	Citrate cycle (TCA cycle)	20	11	6,71%	1,15E-06	9,34E-05
Microsporidia	164	Phenylalanine, tyrosine and tryptophan biosynthesis	400	12	7,32%	1,53E-05	1,24E-03
Microsporidia	164	Metabolic pathways	1100	115	70,12%	5,57E-04	4,52E-02

13.4.2.2 ReliefF

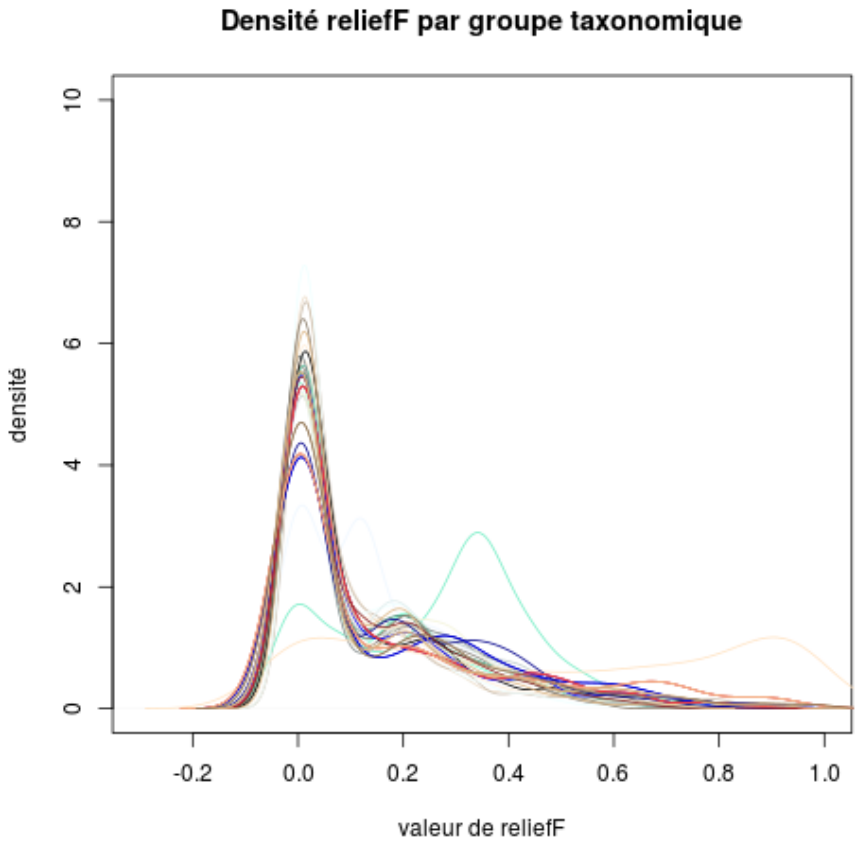


Figure 13.24 – *Densité de la valeur de reliefF en fonction des groupes taxonomiques.*

Afin de comparer l'impact de différents scores sur nos observations, nous avons également testé la présélection d'attributs en fonction de leur score reliefF. Les densités de score reliefF en fonction du groupe taxonomique sont visibles sur la figure 13.24. Comme pour le RIG, on observe une majorité des pics de densité autour d'une valeur faible du score (ici 0,1). Nous avons analysé indépendamment les caractéristiques des *EC numbers* associés aux profils ayant les meilleurs scores

reliefF (10 premiers et 100 premiers scores) pour chaque groupe taxonomique. Les scores moyens de ces profils sont observables sur le tableau 5 en annexe.

Les deux méthodes de sélection d'attributs aboutissent à des choix de profils différents. Cependant, plus le nombre de profils sélectionnés augmente plus ces différences tendent à se lisser (la médiane du pourcentage de profils sélectionnés par les deux méthodes lors de la sélection de 100 profils est supérieure à 70%).

Analyse pour chaque groupe taxonomique des caractéristiques des 10 profils au plus grand score reliefF. On trouve 62 paires 'groupe taxonomique – voie de kegg présentant au moins 2 *EC numbers* annotant les 10 meilleurs profils'. Ces 62 paires font intervenir 31 voies de KEGG. Parmi ces 62 paires, 21 présentent au moins 2 *EC numbers* connexes et 14 ne font intervenir qu'une composante connexe.

Comme pour le RIG, pour chaque groupe taxonomique la majorité des voies sur-représentées parmi les 10 premiers profils sélectionnés sont en lien avec de la biosynthèse de composés (voir le tableau 13.8). Cependant, pour un même groupe taxonomique ce ne sont pas les mêmes voies qui sont sur-représentées, les deux méthodes n'effectuant pas le même type de sélection. Cette observation est problématique. La limite de 10 profils est peut-être trop basse pour sélectionner l'ensemble de l'information pertinente pour la caractérisation de l'évolution du métabolisme. De plus, les deux méthodes ne sont pas basées sur le même type d'approche. Il est donc possible qu'elles trouvent chacune des sous-ensembles chevauchants de profils pertinents.

On observe que les voies sur-représentées parmi les EC des 10 meilleurs profils sont dans la moitié des cas non trouvées sur-représentées dans les données obtenues par tirage aléatoire (voir la figure 13.21 et le tableau 13.8). Il s'agit des voies de biosynthèse de terpenoid backbone, du métabolisme du glyoxylate et du dicarboxylate, du cycle du cytrate et de la voie de biosynthèse de métabolites secondaires.

Table 13.8 – *Liste des voies sur représentées parmi les EC appartenant aux profils classés dans le top 10 (filtre RELIEF)*

Groupe taxonomique	# *EC* select.	nom voie	# de la voie	# d'*EC* select. ∈ voie	% d'*EC* select. ∈ voie	liste des *EC* select.	p-value Fisher	p-value corrigé
Agaricales	15	Sesquiterpenoid and triterpenoid biosynthesis	909	2	13,33%	3.1.7.6;4.2.3.23	4,35E-03	4,35E-02
Agaricomycotina	14	Terpenoid backbone biosynthesis	900	4	28,57%	1.8.3.6;3.1.7.6;2.7.1.36;2.5.1.31	1,07E-04	6,44E-04
Ascomycota	11	Terpenoid backbone biosynthesis	900	3	27,27%	2.5.1.31; 2.7.1.36; 1.8.3.6	1,01E-03	8,06E-03
Dothideomycetes	13	Glyoxylate and dicarboxylate metabolism	630	3	23,08%	1.1.1.29; 3.5.1.10; 1.1.1.81	3,27E-03	4,91E-02
Glomerellales	13	Glyoxylate and dicarboxylate metabolism	630	3	23,08%	1.1.1.81; 1.1.1.29; 3.5.1.10	3,27E-03	4,25E-02
Microsporidia	37	Citrate cycle (TCA cycle)	20	6	16,22%	1.3.5.1; 6.2.1.5; 1.8.1.4; 2.3.1.12; 1.1.1.37; 4.2.1.3	8,59E-06	4,38E-04
Microsporidia	37	Biosynthesis of secondary metabolites	1110	21	56,76%	2.1.2.3; 2.7.6.1; 6.2.1.5; 3.5.4.10; 1.8.1.4; 6.3.3.1; 1.2.1.3; 4.2.1.3; 2.3.1.12; 2.5.1.54; 1.14.13.132; 1.3.5.1; 4.1.3.27; 2.6.1.1; 6.3.4.13; 2.6.1.42; 1.1.1.205; 5.3.1.24; 1.1.1.37; 4.2.1.10; 2.1.1.201	1,87E-04	9,53E-03
Microsporidia	37	Phenylalanine, tyrosine and tryptophan biosynthesis	400	5	13,51%	4.1.3.27; 5.3.1.24; 2.6.1.1; 2.5.1.54; 4.2.1.10	7,22E-04	3,68E-02
Taphrinomycotina	12	Ubiquinone and other terpenoid-quinone biosynthesis	130	3	25,00%	2.1.1.114; 2.1.1.222; 2.1.1.64	6,84E-05	1,44E-03
Tremellomycetes	10	N-Glycan biosynthesis	510	3	30,00%	2.4.1.256; 3.2.1.106; 2.4.1.267	9,25E-04	2,41E-02

La voie de biosynthèse de terpenoides est sélectionnée comme sur-représentées parmi les *EC* correspondant aux 10 profils sélectionnés pour caractériser les agaricomycotina et les ascomycota. Cette voie correspond respectivement à la sélection de quatre et trois profils (voir figure 13.25). On remarque que deux des quatre *EC* sélectionnés utilisent le même substrat.

Analyse pour chaque groupe taxonomique des caractéristiques des 100 profils au plus grand score ReliefF. Analyser un à un les 29 listes de 100 profils au plus grand score reliefF obtenues pour les 29 groupes taxonomiques n'est pas envisageable. Nous cherchons donc à analyser les propriétés générales de ces profils.

Nous pouvons dans un premier temps observer le degré des activités enzymatiques associé à ces profils. Comme pour les résultats obtenus avec le RIG, le degré moyen des *EC numbers* liés aux 100 profils au plus haut score reliefF pour chacun des groupes taxonomiques est significativement plus grand que le degré moyen des *EC numbers* apparaissant dans les mêmes voies KEGG (p-value : $7,67e-10$).

La sélection pour chaque groupe taxonomique des 100 profils présentant le meilleur score reliefF aboutit à l'obtention de 724 paires 'groupe taxonomique – voie métabolique présentant au moins deux *EC numbers* sélectionnés'.

Parmi les 724 paires, 351 présentent la sélection d'au moins 2 *EC numbers* connectés au sein du graphe et 173 contiennent des *EC numbers* ne formant qu'une unique composante connexe.

Les méthodes de type filtre ayant la possibilité de sélectionner des attributs redondants, ces composantes connexes pourraient être l'observation de modules fonctionnels.

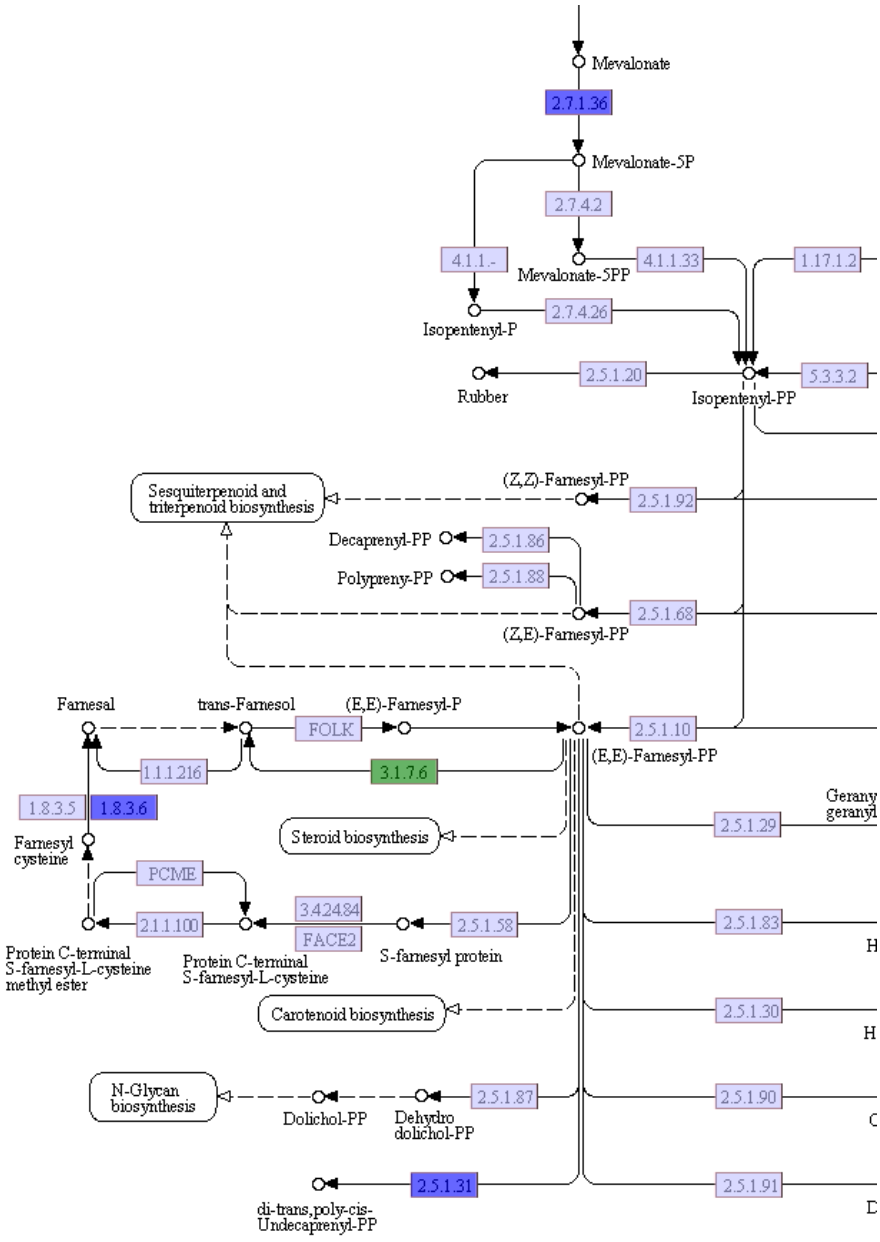


Figure 13.25 – *Sous-partie de la voie 'terpenoid backbone biosynthesis' de KEGG. En bleu foncé les EC numbers sélectionnés comme caractéristiques des agaricomycotina et des ascomycota, en vert : l' EC number qui n'est caractéristique que des agaricomycotina.*

Table 13.9 – *Pour chaque groupe taxonomique, liste des voies sur-représentées dans la liste d'EC appartenant aux profils classés parmi les 100 premiers par la méthode de filtre reliefF.*

Groupe taxonomique	# *EC* select.	nom voie	# de la voie	# d'*EC* select. ∈ voie	% d'*EC* select. ∈ voie	p-value Fisher	p-value corrigée
Eurotiomycetes	105	Betalain biosynthesis	965	3	2,86%	5,75E-04	3,68E-02
Microsporidia	148	Biosynthesis of secondary metabolites	1110	66	44,59%	1,51E-08	1,15E-06
Microsporidia	148	Citrate cycle (TCA cycle)	20	12	8,11%	2,89E-08	2,20E-06
Microsporidia	148	Carbon fixation pathways in prokaryotes	720	10	6,76%	5,50E-05	4,18E-03
Microsporidia	148	Metabolic pathways	1100	112	75,68%	7,52E-05	5,71E-03
Microsporidia	148	Alanine, aspartate and glutamate metabolism	250	12	8,11%	1,24E-04	9,43E-03
Microsporidia	148	Phenylalanine, tyrosine and tryptophan biosynthesis	400	10	6,76%	2,16E-04	1,64E-02
Microsporidia	148	Microbial metabolism in diverse environments	1120	35	23,65%	4,75E-04	3,61E-02
Microsporidia	148	Steroid biosynthesis	100	7	4,73%	5,01E-04	3,81E-02
Taphrinomycotina	105	Valine, leucine and isoleucine degradation	280	8	7,62%	1,93E-04	1,29E-02

Pour chaque groupe taxonomique, on observe sur le tableau 13.9 les voies sur-représentées parmi les *EC numbers* correspondant aux 100 meilleurs profils.

Les voies sur-représentées au sein des 100 meilleurs profils sélectionnés par les deux méthodes diffèrent. Microsporidia est le seul groupe taxonomique pour lequel les mêmes voies sont parfois sélectionnées *via* l'application des deux scores. Les quatre voies ainsi retrouvées parmi les résultats des deux méthodes sont les voies de biosynthèse de métabolites secondaires, le métabolisme de manière générale, le cycle du citrate ainsi que la biosynthèse de phenylalanine, tyrosine et tryptophane.

13.4.3 Discussions et conclusions sur l'analyse des résultats obtenus avec les méthodes de type filtre.

Les degrés dans le graphe des *EC numbers* associés aux 100 profils les plus caractéristiques de chaque groupe taxonomique (méthode RIG ou méthode reliefF) sont plus grands que les degrés moyens des *EC numbers* des voies KEGG. La modification de protéines portant un *EC number* fortement connexe dans le graphe est certainement souvent délétère. Dans le cas où cette modification n'est pas délétère, le changement phénotypique est peut-être suffisamment important pour induire une nouvelle branche dans la taxonomie. Cela pourrait donc expliquer leur conservation au sein d'un groupe taxonomique donné.

Pour chaque groupe taxonomique, les voies sur-représentées parmi les 10 ou les 100 premiers profils obtenus avec les deux méthodes sont principalement des voies impliquées dans la biosynthèse de composés du métabolisme secondaire. Les différents groupes taxonomiques de champignons semblent donc avoir des capacités métaboliques caractéristiques maintenues au cours de l'évolution.

Malgré la sélection d'un grand nombre de profils en commun , les voies sélectionnées ne sont pas constantes en fonction de la méthode de filtre. Plusieurs hypothèses peuvent expliquer ce résultat :

- hypothèse n°1 : il n'y a pas de voie métabolique évoluant plus rapidement que les autres ou ayant une importance plus grande que les autres dans la caractérisation des groupes taxonomiques.

- hypothèse n°2 : le découpage en voies est arbitraire. Les *EC* sélectionnés ne font peut-être pas partie de la même voie, mais ils sont peut-être connexes ou impliqués dans un même phénomène décrit par un ensemble de voies.

- hypothèse n°3 : les voies métaboliques de KEGG sont trop grandes. Un travail avec la définition de voies de MetaCyc ou avec la définition de modules aurait

peut-être permis de trouver des parties de voies sur-représentées parmi les profils sélectionnés.

- hypothèse n°4 : l'application de méthodes de filtre n'est peut-être pas adaptée à nos données, les résultats observés ne sont peut-être pas significatifs.

Nous avons analysé les types de profils sélectionnés par deux méthodes de type filtre. Nous avons abouti à leur utilisation afin de limiter l'impact de l'existence d'un possible bruit dans les données, bruit lié à la présence d'un grand nombre d'attributs par rapport au nombre d'exemples. Maintenant que nous avons observé le type de profils sélectionnés par les méthodes de type filtre nous pouvons les associer aux méthodes de type arbres et règles de décisions afin d'observer si cela permet d'obtenir de meilleurs classifieurs.

13.5 Application conjointe des filtres puis des règles et arbres de décisions

Nos données présentent plus d'attributs que d'exemples. Il est probable que certains attributs ne soient pas pertinents. Il peut donc être intéressant d'effectuer une étape de filtre des attributs avant de rechercher un classifieur. Nous avons de ce fait testé l'ajout d'une étape de filtre à notre *pipeline* (voir figure 13.26).

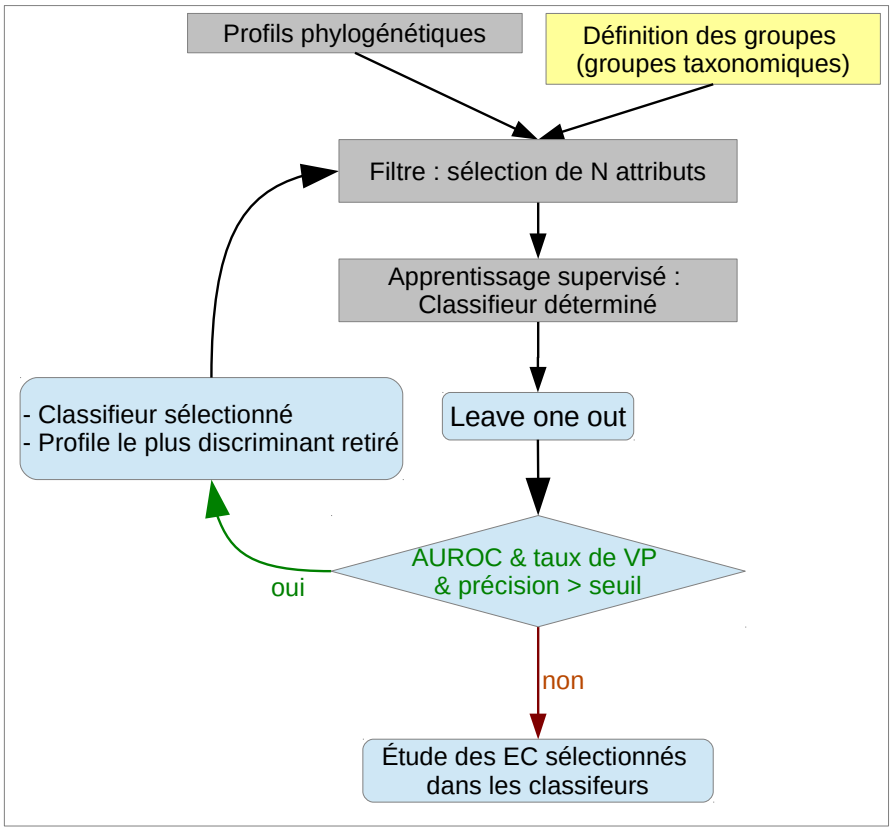


Figure 13.26 – *Pipeline d'apprentissage supervisé contenant une étape de filtre. Les différentes étapes sont décrites sur des rectangles aux angles arrondis et sur fond bleu, le test est figuré par un losange, les données que nous générons sont sur fond gris et les données de la littérature sur fond jaune.*

Pour cette première analyse, nous avons appliqué ce *pipeline* avec les méthodes RIG et reliefF (sélection de 10 ou 100 attributs) combiné à la recherche des classifieurs (seuils fixés à 0,9). Il s'agit d'une analyse préliminaire. Nous testerons dans les mois à venir le sélection d'attributs de 10 à 100 par pas de 10. L'ajout de l'étape de sélection d'attributs a permis l'obtention d'un plus grand nombre de classifieurs (voir les tableaux 13.27 et 13.10). Ainsi, retirer des profils non pertinents a permis de sélectionner un plus grand nombre de règles caractérisant la taxonomie.

	Apprentissage				
Méthode	classique	itérative	itérative seuil 0.9	itérative filtre 10	itérative filtre 100
Données	FUNGIpath 50	FUNGIpath 178 v1	FUNGIpath 178 v2 (MARIO)		
Itération de la recherche de classifieurs	-	x	x	x	x
Seuil de sélection des classifieurs (LOO)	-	1er filtre (itération) seuil 0.5, 2ème filtre (sélection) seuil 0.9	Seuil 0.9	Seuil 0.9	Seuil 0.9
Filtre appliqué aux attributs	-	-	-	Sélection de 10 attributs	Sélection de 100 attributs
méthodes de filtre	-	-	-	RGI, Relieff	RGI, Relieff
Nombre de groupes taxonomiques caractérisés	9	20	9	8 (RIG) | 17 (Relieff)	11 (RIG) | 20 (Relieff)
Nombre moyen de classiifeurs par groupe taxonomique caractérisé	1	113	30	20.5 (RIG) | 14.7 (Relieff)	16.9 (RIG) | 11.45 (Relieff)

Figure 13.27 – *Comparaison des résultats obtenus avec les différents pipelines.*

Table 13.10 – *Nombre de classifieurs obtenus par groupe taxonomique et nombre d'EC numbers présents dans au moins un de ces classifieurs.*

Groupe taxonomique	nombre d'*EC numbers* différents dans les classifieurs	nombre de classifieurs C4.5	nombre de classifieurs RIPPER	nombre de classifieurs PART
méthode: filtre ratio gain d'information (sélection des 10 meilleurs attributs) et classifieurs seuil 0,9				
Agaricomycetes	9	3	3	3
Agaricomycotina	13	1	0	6
Ascomycota	33	2	14	26
Basidiomycota	5	1	2	1
Dikarya	11	6	6	6
Onygenales	3	1	1	1
Pezizomycotina	45	20	39	21
méthode: filtre ratio gain d'information (sélection des 100 meilleurs attributs) et classifieurs seuil 0,9				
Agaricomycetes	10	4	1	2
Agaricomycotina	14	3	2	3
Ascomycota	83	7	56	7
Basidiomycota	21	1	10	1
Dikarya	19	6	10	6
Onygenales	4	2	2	2
Pezizomycotina	61	10	37	10
méthode: filtre reliefF (sélection des 10 meilleurs profils) et classifieurs seuil 0,9				
Agaricomycetes	7	3	1	3
Agaricomycotina	8	1	2	3
Ascomycota	28	22	16	22
Dikarya	5	2	3	2
Eurotiomycetes	10	0	0	5
Mucorales	1	1	1	1
Mucoromycotina	1	1	1	1
Onygenales	6	2	3	4
Pezizomycotina	34	19	27	26
Saccharomycetales	15	0	7	9
Saccharomycetes	15	0	7	9
Saccharomycotina	15	0	7	9
Schizosaccharomycetales	1	1	1	1
Schizosaccharomycetes	1	1	1	1
Taphrinomycotina	1	1	1	1
Tremellales	4	3	3	3
Tremellomycetes	4	3	3	3
Zygomycota	1	1	1	1

ce tableau continue sur la page suivante

Groupe taxonomique	nombre d'*EC numbers* différents dans les classifieurs	nombre de classifieurs C4.5	nombre de classifieurs RIPPER	nombre de classifieurs PART
méthode: filtre reliefF (sélection des 100 meilleurs profils) et classifieurs seuil 0,9				
Agaricomycetes	5	1	1	1
Agaricomycotina	12	3	1	2
Ascomycota	92	9	67	9
Basidiomycota	10	1	2	4
Dikarya	20	3	9	3
Eurotiomycetes	15	0	1	6
Mucorales	1	1	1	1
Mucoromycotina	1	1	1	1
Onygenales	10	2	2	4
Pezizomycotina	55	10	37	10
Saccharomycetales	4	0	2	0
Saccharomycetes	4	0	2	0
Saccharomycotina	4	0	2	0
Schizosaccharomycetales	1	1	1	1
Schizosaccharomycetes	1	1	1	1
Sordariales	2	1	1	1
Taphrinomycotina	1	1	1	1
Tremellales	4	3	1	3
Tremellomycetes	4	3	1	3
Zygomycota	1	1	1	1

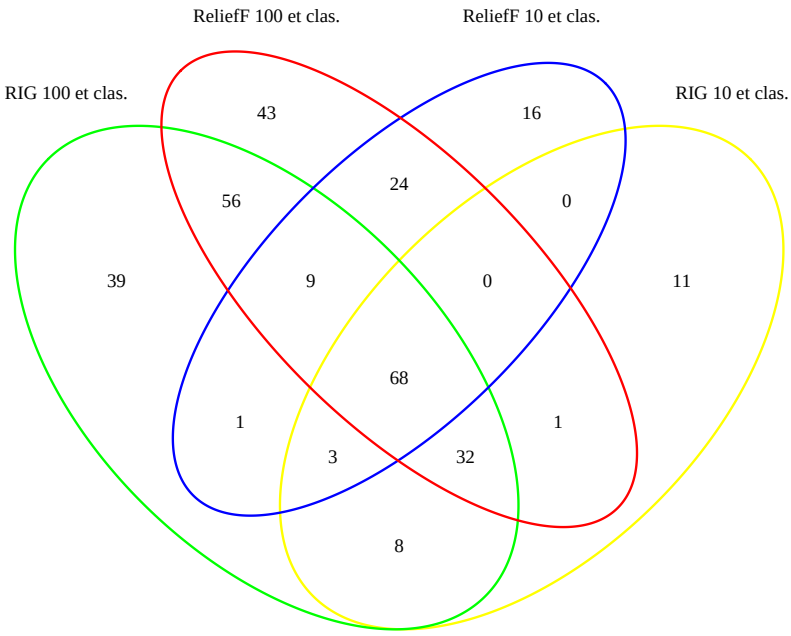


Figure 13.28 – *Diagramme de Venn des activités enzymatiques trouvées par les quatre expériences.*

Les classifieurs obtenus par ces quatre expériences ne sont pas identiques. Le diagramme de Venn indiquant le nombre d'activités enzymatiques trouvées *via* plusieurs expériences est présenté figure 13.28.

Ces quatre expériences permettent de trouver un trop grand nombre de classifieurs pour qu'il soit possible de les analyser ici individuellement.

Nous pouvons cependant observer si des classifieurs précédemment sélectionnés par la méthode sans filtre sont inclus dans ces résultats. Parmi les EC sélectionnés par les quatre alternatives du *pipeline* proposé dans cette dernière partie (voir figure 13.28 et 13.29) on retrouve comme pour chacune des méthodes précédentes la sélection de l'EC:3.1.6.6 comme étant caractéristique des pezizomycotina.

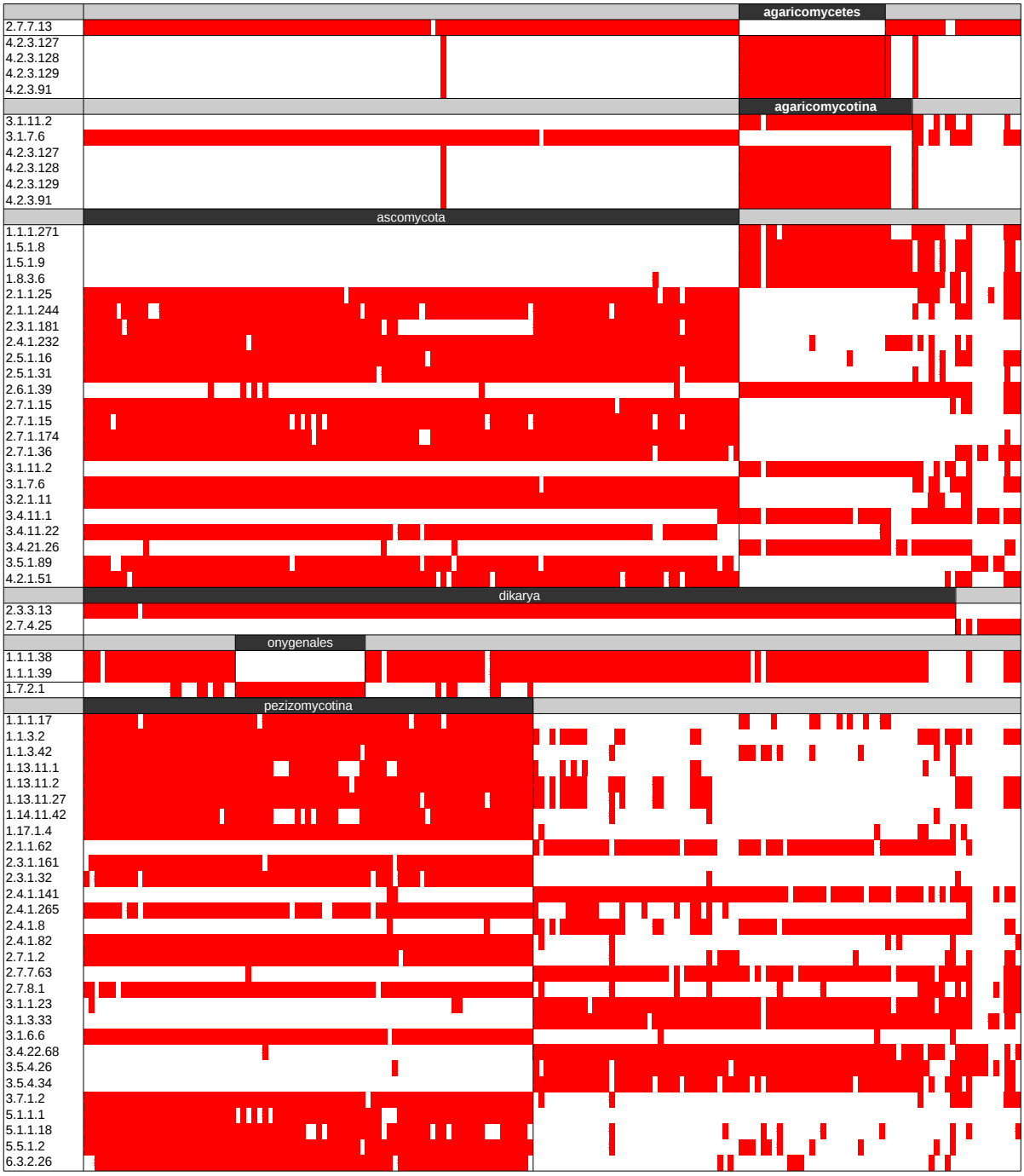


Figure 13.29 – *Profils des 68 EC sélectionnés dans l'intersection des quatre approches.*

Ce résultat, retrouvé par l'ensemble de nos approches, est donc l'un des plus sûrs que nous ayons obtenu.

De plus, les approches utilisant en filtre les 10 meilleurs profils obtenus avec RIG/ReliefF ou les 100 meilleurs profils obtenus avec RIG (soit 3/4 des approches) trouvent les EC:1.5.1.8 et 1.5.1.9 caractéristiques des agaricomycetes. Pour rappel, nous avions sélectionné les mêmes *EC* avec l'application de classifieurs sur les deux versions de FungiPath à 178 génomes.

Parmi les EC sélectionnés pour caractériser les dikarya dans l'approche présentée à JOBIM et appliquée aux données MARIO, un des sept EC est retrouvé par les quatre approches testées ici et trois sont retrouvées avec le 3/4 des approches (toutes sauf RIG10). Les autres correspondent à la sélection de deux profils reflétant chacun la présence / absence de deux *EC* et ayant été sélectionné sur une sous partie des données (deuxième règle d'un classifieur). Cette observation tend à nous encourager dans l'utilisation de filtres. Il semblerait en effet qu'ils permettent à la fois d'obtenir un grand nombre de classifieurs et d'éviter la sélection de profils dont la pertinence était précédemment mise en doute lors de leurs analyses.

Dans l'article présenté à JOBIM, nous avions détaillé les classifieurs obtenus pour les agaricomycetes. L'ensemble des profils sélectionnés par les quatre méthodes filtres et classifieurs avaient été sélectionné lors de l'analyse précédente. Il s'agit de profils impliqués dans la création de composés biologiquement actifs (Sesquiterpene synthase COP4) et de la manose-1-phosphate, impliquée dans la formation de la paroi cellulaire.

Propriété des EC sélectionnés Les *EC numbers* sélectionnés par chacune des expériences ne présentent statistiquement pas un degré différent du degré moyen.

Pour chacun des quatre types d'expériences caractérisant chacun des groupes taxonomiques, nous avons étudié la connexion entre les *EC* sélectionnés pour caractériser un même groupe taxonomique dans une même voie. Lorsqu'au moins deux

EC de la même voie sont sélectionnés, il y a en fonction du test entre 33% et 64% de chance qu'il y ait au moins un lien entre elle. Á nouveau, il pourrait s'agir de modules fonctionnels.

Voies sur-représentées Le nombre de classifieurs obtenus est grand et il serait fastidieux d'en analyser l'ensemble manuellement. De ce fait, par groupe taxonomique, nous avons recherché si des voies étaient sur-représentées. Pour cela, nous avons utilisé le test exact de Fisher (test de comparaison de proportions) et nous avons corrigé les p-values avec la méthode de Bonferroni. Nous obtenons les résultats présentés sur les tableaux 13.11 et 13.12. Parmi les voies sur-représentées pour un groupe taxonomique, quatre couples impliquant trois voies avaient déjà été observés. Il s'agit de la voie de biosynthèse des terpenoides (caractérisation des agaricomycotina et ascomycotina), de la voie de dégradation de la lysine (basidiomycotina), et de la voie de biosynthèse des glycosaminoglycanes (taphrinomycotina). Le fait de retrouver ces voies grâce à plusieurs méthodes confirme qu'elles sont peut-être des cibles dans l'évolution du métabolisme.

Table 13.11 – *Liste des voies sur représentées obtenu sur les données correspondant aux classifieurs appris sur des données préalablement filtrées avec un filtre de type ratio gain d'information. Abréviations : sélect. : sélectionnés, ∈ : appartenant, # : nombre.*

Groupe taxonomique	# *EC* select.	nom voie	# de la voie	# d'*EC* select∈ voie	% d'*EC* ∈ voie	liste des *EC* sélect.	p-value Fisher	p-value corrigée
sélection de 10 profils								
Agaricomycotina	13	Terpenoid backbone biosynthesis	900	3	23,08%	1.8.3.6; 3.1.7.6; 2.5.1.31	1,52E-03	1,06E-02
Ascomycota	33	Terpenoid backbone biosynthesis	900	4	12,12%	2.5.1.31; 1.8.3.6; 2.7.1.36; 3.1.7.6	1,92E-03	4,41E-02
Onygenales	3	Pyruvate metabolism	620	2	66,67%	1.1.1.38; 1.1.1.39	4,52E-03	2,26E-02
sélection de 100 profils								
Basidiomycota	21	Lysine degradation	310	4	19,05%	1.5.1.9; 1.5.1.8; 2.6.1.48; 2.6.1.39	5,03E-04	1,06E-02
Onygenales	4	Pyruvate metabolism	620	2	50,00%	1.1.1.39; 1.1.1.38	6,68E-03	4,01E-02
Sordariales	2	Styrene degradation	643	1	50,00%	1.14.13.63	1,62E-02	4,85E-02

Table 13.12 – *Liste des voies sur-représentées obtenues sur les données correspondant aux classifieurs appris sur des données préalablement filtrées avec un filtre de type reliefF.*

Groupe taxonomique	# *EC* select.	nom voie	# de la voie	# d'*EC* select∈ voie	% d'*EC* ∈ voie	liste des *EC* sélect.	p-value Fisher	p-value corrigée
sélection de 10 profils								
Ascomycota	28	Terpenoid backbone biosynthesis	900	4	14,29%	2.5.1.31; 2.7.1.36; 1.8.3.6; 3.1.7.6	1,10E-03	2,52E-02
Ascomycota	28	Lysine degradation	310	4	14,29%	2.6.1.48; 1.5.1.9; 1.5.1.8; 2.6.1.39	1,33E-03	3,06E-02
Taphrinomycotina	1	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	534	1	100,00%	2.4.1.134	1,56E-03	4,67E-03
Taphrinomycotina	1	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	532	1	100,00%	2.4.1.134	1,56E-03	4,67E-03
sélection de 100 profils								
Basidiomycota	10	Fructose and mannose metabolism	51	3	30,00%	2.7.7.13; 2.7.1.105; 1.1.1.271	1,74E-03	2,09E-02
Sordariales	2	Styrene degradation	643	1	50,00%	1.14.13.63	1,62E-02	4,85E-02
Taphrinomycotina	1	Glycosaminoglycan biosynthesis - chondroitin sulfate / dermatan sulfate	532	1	100,00%	2.4.1.134	1,56E-03	4,67E-03
Taphrinomycotina	1	Glycosaminoglycan biosynthesis - heparan sulfate / heparin	534	1	100,00%	2.4.1.134	1,56E-03	4,67E-03

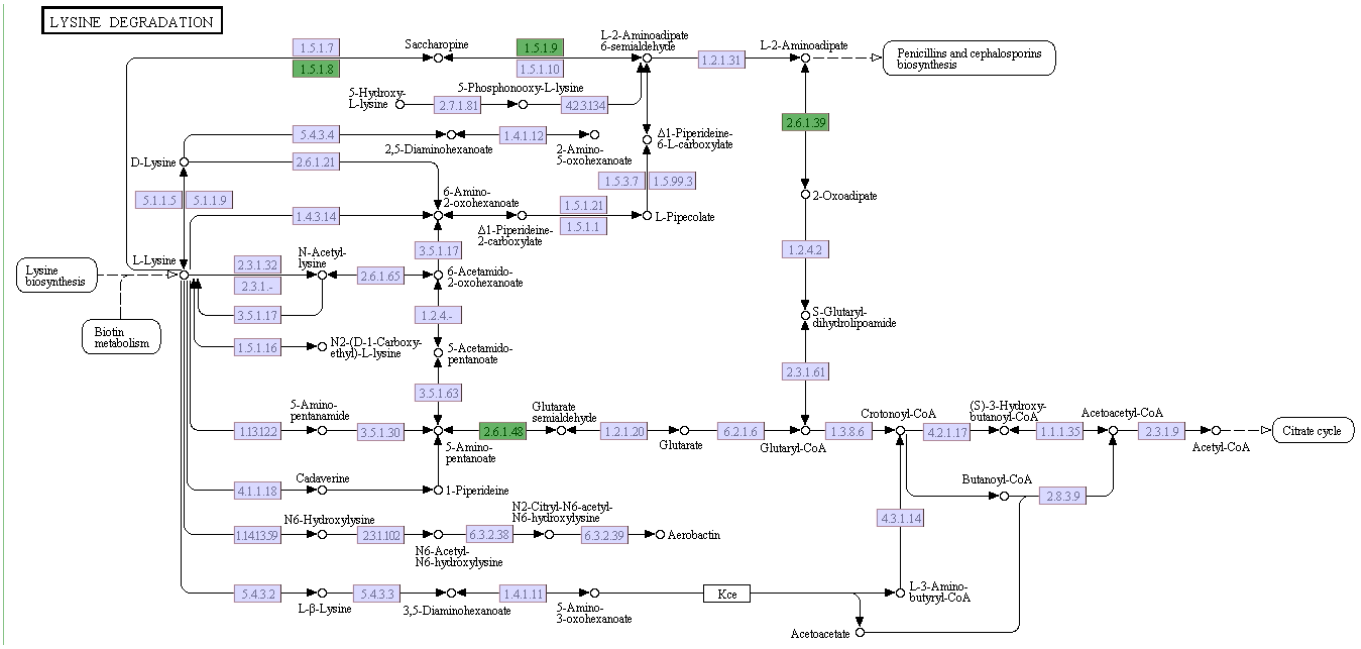


Figure 13.30 – *Les activités enzymatiques présentent dans au moins un classifieur caractéristique des ascomycota / basidiomycota sont sur fond vert. Ces activités enzymatiques sont généralement absentes chez les ascomycota et présentes chez les basidiomycota.*

Les classifieurs peuvent sélectionner des profils aussi bien pour la présence de l'activité enzymatique dans le groupe taxonomique que pour son absence dans ce groupe. Ainsi, les ascomycota et les basidiomycota sont en partie caractérisés par l'absence chez les ascomycota et la présence chez les basidiomycota de quatre activités enzymatiques impliquées dans la dégradation de la lysine (voir figure 13.30). Il s'agit des activités enzymatiques EC:2.6.1.48, EC:2.6.1.39, EC:1.5.1.8 et EC:1.5.1.9. Les EC:1.5.1.8 et EC:1.5.1.9 avaient été sélectionnés dans l'étude présentée à JOBIM ainsi que dans les résultats des classifieurs obtenus sans filtres. Ils étaient trouvés caractéristiques des agaricomycotina, un sous-groupe des basidiomycota pour lequel nous avons le plus de représentants.

On en déduit que les ascomycota ont tendance à ne pas pouvoir dégrader la lysine, du moins par les voies de dégradation connues, contrairement aux basi-

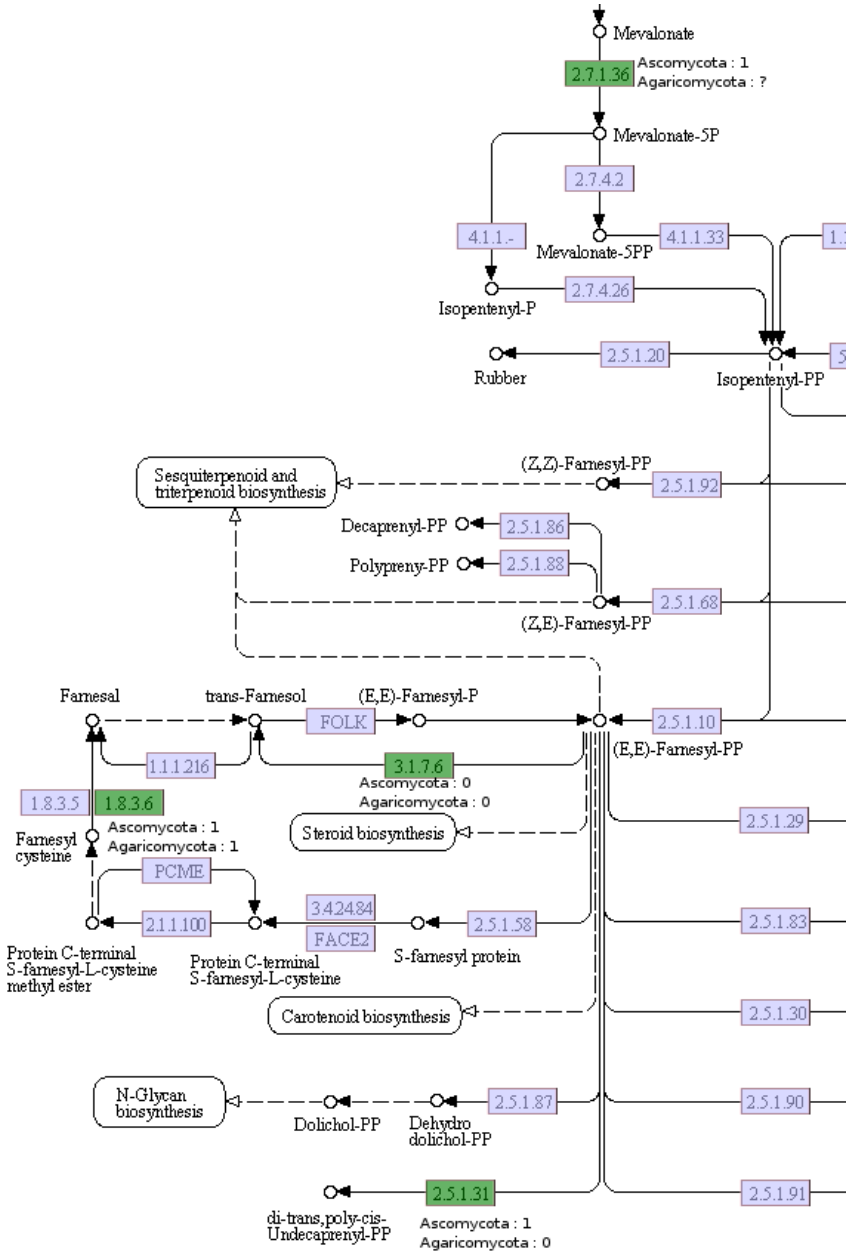


Figure 13.31 – *Sous-partie de la voie 'Terpenoid backbone biosynthesis' de KEGG. Les activités enzymatiques présentes dans des classifieurs d'ascomycota ou d'agaricomycota sont sur fond vert. Près de chaque activité enzymatique sur fond vert, du texte indique si l'activité enzymatique doit plutôt être présente (1) ou absente (0).*

diomycota. Cette observation est cohérente avec le fait que les ascomycota et les basidiomycota sont les deux seuls embranchements du sous-règne dikarya. Ils présentent de ce fait des synapomorphies opposées. De plus, parmi les basidiomycota on retrouve entre autres les pourritures blanches et brunes qui sont connues pour leurs capacités de dégradation de la biomasse.

L'application de filtres permet d'obtenir plus de classifieurs et donc plus de règles permettant de caractériser la taxonomie. On retrouve les profils sélectionnés par les autres méthodes (EC:3.1.6.6 pour les pezizomycota par exemple) mais aussi de nouveaux profils. Il nous faudra analyser si ces nouveaux profils sont plus pertinents que les anciens, pour cela nous pourrions tester les classifieurs obtenus sur ces 173 champignons sur la version de FungiPath en cours de calcul (augmentation de la taille de la base de 50%).

Chapitre 14

Apprentissage supervisé: conclusions et perspectives

14.1 Le point sur nos questions initiales

Nous avions introduit cette troisième partie avec plusieurs questions portant sur l'évolution du métabolisme. Nous pouvons à présent tenter d'y répondre.

Le métabolisme garde-t-il des traces de son évolution? L'application de classifieurs a permis la sélection de profils caractéristiques des groupes taxonomiques (profil de EC:3.1.6.6 caractéristique des pezizomycotina). Comme attendu, des traces de l'évolution sont donc persistantes dans le métabolisme.

Notre approche ne se limite pas à la sélection du meilleur classifieur. Nous recherchons un ensemble de classifieurs pertinents. Pour cela, de manière à obtenir un nouveau classifieur, nous retirons des données le premier attribut sélectionné par le classifieur précédent. Nous observons la sélection successive par ces différents classifieurs d'activités enzymatiques connexes dans le graphe du métabolisme. Nous posons de ce fait l'hypothèse de présence de modules fonctionnels. Si une activité

enzymatique du module est sélectionnée, elle suffit à caractériser l'ensemble du module (pas de sélection d'une seconde activité enzymatique du même module dans le même classifieur). Notre approche itérative nous permet de sélectionner successivement les activités enzymatiques des modules caractéristiques de la taxonomie. Il pourrait être intéressant d'analyser les arbres obtenus avec les groupes d'orthologues des protéines codant ces activités enzymatiques. L'hypothèse de module fonctionnel pourrait alors être corroborée par des histoires évolutives communes de ces gènes.

Les activités enzymatiques caractérisant la taxonomie sont-elles dans le graphe métabolique au niveau de nœuds à fort degré ? L'observation des 100 profils ayant les plus grandes valeurs de RIG ou reliefF a montré la sélection préférentielle d'activités enzymatiques présentant un degré plus élevé que la moyenne. Cependant, cette observation n'est pas faite sur les activités enzymatiques des classifieurs. Il semble que les activités enzymatiques présentes au niveau de nœuds à forte connexité tendent à être conservées au sein d'un groupe taxonomique. Cependant, cette conservation ne permet pas de caractériser efficacement ce groupe. Une hypothèse serait que l'apparition ou la disparition d'une activité enzymatique fortement connexe est un des mécanismes et pas l'unique mécanisme d'évolution du métabolisme. Cela explique que ces *EC* sont préférentiellement sélectionnés dans les 100 meilleurs profils. Les classifieurs ne les sélectionnent cependant pas car il existe d'autres profils aux meilleurs pouvoir prédictif dans les données. Ces autres activités enzymatiques semblent préférentiellement être impliquées dans le métabolisme secondaire.

Voies sur-représentées parmi les activités enzymatiques des profils caractérisant un groupe taxonomique. Chacune des méthodes a montré la présence de quelques voies sur-représentées (plus d'*EC numbers* appartenant à la voie qu'attendu dans les filtres et classifieurs). Ces voies peuvent être sélectionnées, car elles

sont absentes (cycle du citrate absent chez les microsporidia), présentes (voie de dégradation de la lysine chez les basidiomycota) ou les deux à la fois (voie de biosynthèse des terpenoides). Dans ce dernier cas, cela ne semble pas refléter la présence de voies alternatives. La distribution des enzymes dans la voie de biosynthèse des terpenoides est différente de celle observée dans les autres voies. Les enzymes ne sont pas connexes dans le graphe, certaines sont sélectionnées pour leur présence, d'autres pour leur absence et leurs profils attendus sont parfois les mêmes pour deux groupes taxonomiques différents (ascomycota et agaricomycotina). Cette voie, parce qu'elle est particulière par rapport aux autres, mériterait une étude plus approfondie.

14.2 Limites des méthodes actuelles et perspectives

14.2.1 Analyse des classifieurs

14.2.1.1 Consistance des classifieurs

Dans cette troisième partie, nous avons regroupé les résultats obtenus avec les différents classifieurs dans le but d'analyser les activités enzymatiques sélectionnées. Nous avons discuté le nombre d'activités enzymatiques trouvées *via* les différentes méthodes et nous avons comparé les profils sélectionnés. Cependant, nous n'avons pas comparé les différentes règles obtenues. Il n'existe pas à notre connaissance de logiciel proposant cette fonctionnalité.

La comparaison de ces règles n'est pas triviale. Par exemple, dans le cas des approches de type RIPPER, les règles portent sur les données non couvertes par les règles précédentes. Cela induit d'exprimer la règle courante par l'information donnée par le classifieur ainsi que par la négation des règles précédentes. De plus, la dernière règle est souvent un simple 'sinon' qu'il nous faut alors traduire comme la somme des cas non traités par les règles précédentes. Il sera également nécessaire de simplifier au maximun les règles obtenues afin de retrouver les règles équivalentes.

Une perspective logique de notre travail est donc d'aboutir à un programme de comparaison de classifieurs nous permettant d'analyser plus en détail nos résultats.

Afin d'évaluer nos règles, il sera alors intéressant d'observer leur consistance trouvée sur les différents jeux d'apprentissage *leave one out*.

14.2.1.2 Taille des classifieurs

Lors de l'analyse sur les données de FungiPath 50 génomes, nous avions montré que les classifieurs caractérisant la taxonomie étaient plus petits que des

classifieurs caractérisant des groupes aléatoires de même taille (voir la figure 14.1). Il serait intéressant d'étudier si l'on observe le même phénomène sur l'ensemble des classifieurs appris avec et sans filtre sur 173 espèces. En effet, leur taille particulière par rapport à celle des classifieurs obtenus sur groupes aléatoires renforce la confiance que l'on peut leur apporter.

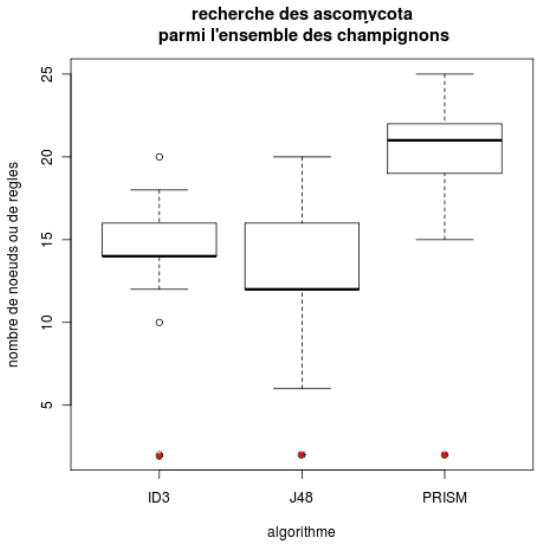


Figure 14.1 – *Boxplot du nombre de nœuds (arbre de décision C45, ID3) ou de règles (Prism) obtenues sur les premiers classifieurs caractérisant des classes aléatoires de même taille que les groupes taxonomiques. Les points rouges figurent la taille du premier classifieur caractérisant le groupe taxonomique des ascomycota.*

14.2.1.3 Impact de la taille de la classe

Nous avons choisi de ne caractériser que les groupes taxonomiques présentant au moins trois champignons. Il serait intéressant de faire une étude pour analyser la corrélation entre la qualité des classifieurs et le nombre d'instance de classe disponibles. En effet, dans nos résultats, nous observons de classifieurs que pour les classes ayant au moins 4 exemples positifs. De plus, le plus grand nombre de classifieurs est obtenu pour les ascomycota, un groupe taxonomique pour lequel

les 2/3 des instances appartiennent à cette classe.

Pour faire l'analyse de la corrélation entre le nombre d'exemples positifs et le nombre de classifieurs obtenu nous pourrions travailler à partir de classes présentant un relatif grand nombre d'exemples positifs (comme la classe des ascomycota) afin de comparer les classifieurs obtenus *via* des échantillonnages répétés de taille variée (approche de type cross validation). Cela nous permettrait d'évaluer de manière expérimentale le nombre d'exemples positifs ou négatifs minimums nécessaires à l'obtention de classifieurs de qualité.

14.2.1.4 Échantillonnage des espèces

Nous avons choisi de travailler avec l'ensemble des espèces de champignons séquencées en 2012 (date de la mise à jour de FungiPath). Or, on observe dans ces données la présence de plusieurs souches caractérisant la même espèce. Se pose donc la question d'un possible biais d'échantillonnage.

Une solution serait de retirer de l'analyse les souches les plus proches. La notion de distance entre les souches pouvant être étudiée grâce à la comparaison de leurs ARN18S.

Une deuxième solution pourrait être d'appliquer des méthodes de type *Bagging* (exemple *Random Forest* (Breiman, 2001)). Ces méthodes consistent à rechercher des classifieurs sur des jeux de données obtenus par tirage avec remise dans le jeu de données d'apprentissage. Cela permet de modifier le poids associé à chaque exemple et à chaque attribut. La recherche des règles les plus consistantes nous permettrait ensuite de pondérer les règles à analyser pour comprendre l'évolution du métabolisme.

14.2.2 Prise en compte de la structure des données

14.2.2.1 Attributs structurés

Le métabolisme structure les *EC numbers* en graphe. Pour le moment, nous l'avons utilisé *a posteriori* pour l'analyse des caractéristiques des *EC numbers* sélectionnés. Cependant, l'hypothèse d'indépendance des attributs n'est pas vérifiée. Il pourrait de ce fait être intéressant de prendre en compte cette structure lors de l'apprentissage.

14.2.2.2 Exemples structurés

Les groupes taxonomiques présentent une structure de type multi classe hiérarchique (arbre). En caractérisant les groupes taxonomiques un par un, nous ne l'avons pour le moment pas prise en compte. Pour ce faire, nous pourrions par exemple appliquer l'algorithme *hierarchical multi-label classification* (Clus-HMC) (Vens et al., 2008). Ce dernier prend en entrée une hiérarchie ainsi qu'un ensemble d'exemples étiquetés avec cette hiérarchie. Il permet d'obtenir un arbre de décision prédisant plusieurs classes de la hiérarchie pour un exemple. Chaque classe est alors associée à une probabilité (voir figure 14.2).

Cela nous permettrait de nous assurer qu'une classe ne peut être prédite que si les classes parentes le sont également.

14.2.3 Passage au quantitatif

Nous travaillons actuellement sur la notion de présence/absence d'une activité enzymatique. Cette notion ne tient pas compte du nombre de copies prédites. Nous avions obtenu l'information sur la présence d'activités enzymatiques en annotant les groupes d'orthologues que nous avions prédits. Nous avons l'information du nombre de groupes et du nombre de copies par groupe, nous pourrions donc

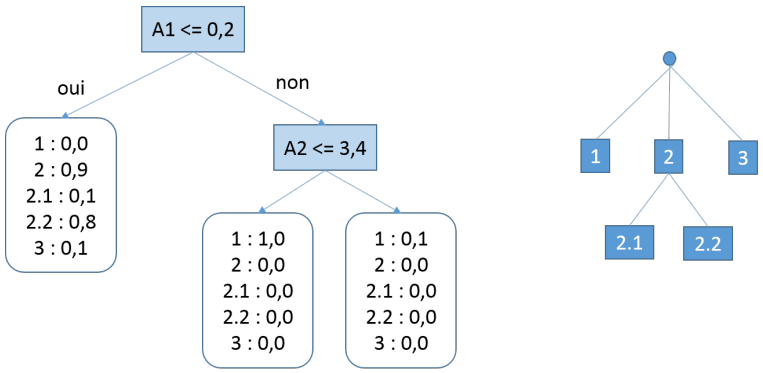


Figure 14.2 – *Exemple d'un résultat de clus-HMC. Á gauche, le classifieur, à droite, la hiérarchie. Il y a deux niveaux dans cette hiérarchie et six classes au total. La présence de ces six classes induit la prédiction par clus-HMC d'un vecteur de taille 6 indiquant pour chaque classe la probabilité qu'elle soit vraie au niveau de cette feuille.*

traiter la présence d'activités enzymatiques par une notion quantitative de nombre de copies. Cette observation pourrait pallier un possible regroupement au sein d'un même groupe de paralogues récents ayant une nouvelle fonction (on obtiendrait alors par exemple une règle imposant que l'activité enzymatique soit présente au moins N fois). Les méthodes de classification que nous avons utilisées jusqu'à présent sont capables de gérer des attributs quantitatifs, nous pourrions donc imaginer les appliquer à cette version des profils. Cependant, il faudrait être prudent, car le nombre de profils différents risque d'augmenter (moins de profils identiques) et cela pourrait biaiser les classifieurs.

IV Conclusions et perspectives

Conclusions

Cette thèse a porté sur deux aspects, l'annotation fonctionnelle de protéomes de champignons et l'analyse et la comparaison du métabolisme de ces champignons.

Nous souhaitions comparer le métabolisme de centaines de champignons. Afin de parer aux différences de qualité d'annotation en fonction des organismes, nous avons décidé de les ré–annoter fonctionnellement. Pour cela, nous avons développé une méta–approche de détection des groupes d'orthologues. Celle-ci est basée sur l'enrichissement des intersections de plusieurs méthodes *via* l'application de profils HMM. Nous avons évalué positivement cette approche sur les *benchmarks* proposés par la communauté. Nous montrons que cette nouvelle méthode permet de prédire plus de relations d'orthologie tout en conservant une forte similarité de fonction au sein des groupes.

Une fois les groupes d'orthologues obtenus et annotés, nous les avons analysés en appliquant des méthodes d'apprentissage supervisé. Nous avons caractérisé la taxonomie grâce aux profils enzymatiques des champignons. Pour cela, nous avons appliqué une méthode originale de recherche de classifieurs de type arbre de décision et règles de décisions sur un ensemble d'attributs décroissant. Comme attendu, nous avons ainsi observé que le métabolisme garde des traces de l'évolution des organismes. Nous observons par exemple que les agaricomycetes se caractérisent par leur capacité de dégradation de la lysine et les pezizomycotina par la présence de

l'EC:3.1.6.6. Cette approche nous a permis de sélectionner des activités enzymatiques caractéristiques de la taxonomie. Elles sont préférentiellement trouvées dans des voies permettant la synthèse de métabolites secondaires.

Dans l'analyse des 100 profils phylogénétiques les plus pertinents, nous avons mis en évidence que les profils les plus discriminants ont tendance à être les profils d'enzymes présentant un fort degré dans le graphe métabolique. Une hypothèse serait que ces activités enzymatiques sont une cible des changement majeurs du métabolisme pouvant induire l'apparition de nouveaux groupes taxonomique.

La combinaison des approches filtres et classifieurs nous permet d'augmenter le nombre de classifieurs obtenus pour chaque groupe taxonomique ainsi que le nombre de groupes taxonomiques pour lesquels on observe au moins un classifieur. Cette approche retrouve la partie des profils sélectionnés en commun par les approches précédentes et propose également de nouveaux profils. Le test prochain de ces classifieurs sur un plus grand jeu de données (calculs de l'approche MARIO sur 350 génomes en cours) nous permettra de valider la qualité de ces classifieurs.

Perspectives

Application de la méta–approche de détection de groupes d'orthologues à de nouveaux organismes

Nous avons montré dans le chapitre 2 que notre méta–approche permet d'obtenir des groupes d'orthologues de fonctions similaires. Nous l'avons pour le moment appliquée à deux jeux de données de *benchmark* ainsi qu'à 178 génomes (principalement des champignons). Nous travaillons actuellement à son application à plus de 350 génomes de champignons.

Nous développons également une nouvelle version du site web FungiPath. Celui-ci permet entre autre de visualiser la présence / absence des activités enzymatiques sur les voies métaboliques en fonction des espèces.

Nous proposerons rapidement une version des groupes d'orthologues pour la dernière version des protéomes de référence de la communauté *Quest for orthologs*.

Apprentissage à partir de nouveaux attributs

Regroupement d'attributs

Les profils que nous analysons ne sont pas indépendants. Ils représentent des activités enzymatiques possiblement impliquées conjointement dans les mêmes fonctions.

Une conservation ou une perte conjointe de plusieurs activités enzymatiques

peut être liée à une fonctionnalité commune. De ce fait, l'étape de filtre peut induire, si elle est trop stringente, une perte d'informations.

Afin de diminuer le nombre d'attributs, nous proposons donc de les regrouper. Un moyen de regrouper les attributs pourrait être d'utiliser les modules de KEGG ou les sous-voies de MetaCyc. Il nous faudrait alors définir quand un module est présent ou absent ou bien accepter de travailler avec des valeurs quantitatives associées à la proportion d'*EC numbers* du module présent.

Un autre moyen, celui-ci sans *a priori* sur le métabolisme, pourrait être de procéder par recherche dans les données de motifs fréquents.

Application de la méthode à d'autres attributs

Il est très probable que la communauté scientifique n'ait pas encore découvert l'ensemble du métabolisme des champignons. Notre annotation, bien que prédisant un grand nombre de nouvelles annotations, est tout de même limitée aux activités enzymatiques décrites au moins une fois dans SwissProt ou Metacyc.

Afin de pallier ce problème, on pourrait appliquer les méthodes d'apprentissage au niveau des groupes d'orthologues de tous les gènes. Les groupes d'orthologues sont bien plus nombreux que les activités enzymatiques prédites. Cependant, le développement de la méta–approche MARIO a permis de diviser par trois le nombre de groupes prédits par rapport à la méthode précédente. Il devient possible grâce à elle, en combinant classifieurs et filtres sur les attributs, d'appliquer notre approche sur les groupes d'orthologues.

Apprentissage de nouveaux concepts cibles

Nous présentons les résultats obtenus pour la caractérisation de la taxonomie. Il pourrait être intéressant de l'appliquer à d'autres concepts tels que par exemple les capacités de dégradation de la biomasse. On pourrait ainsi apprendre

de nouveaux mécanismes de dégradation de la biomasse aux travers des *EC* et des groupes d'orthologues sélectionnés. Cela permettrait également de proposer de nouveaux organismes ayant ces capacités.

Le concept "dégradation de la biomasse" est de type binomial (un champignon dégrade ou non un certain composé). Du fait du manque de connaissance sur les champignons, il s'agit d'une classe pour laquelle des informations sont manquantes. Ce manque d'informations devra être pris en compte pour la suite de l'analyse. On pourra ainsi choisir de ne pas utiliser les exemples non étiquetés, de les utiliser comme des exemples négatifs, de les utiliser comme des exemples positifs ou comme des exemples non labellés. Ces choix modifieront les classifieurs obtenus.

Test d'autres méthodes d'apprentissage

Nous utilisons actuellement nos classifieurs à des fins de compréhension de l'évolution du métabolisme. Il nous est donc important de travailler avec des méthodes interprétables. Cependant, à des fins de prédiction, il serait intéressant de combiner plusieurs méthodes d'apprentissage présentant des biais différents. Obtenir de meilleurs classifieurs mais non interprétables permettant de prédire le concept 'dégradation de la biomasse' pourrait par exemple nous permettre de nous aiguiller sur les organismes à tester pour cette caractéristique et ainsi ajouter de l'information dans nos données (apprentissage semi–supervisé).

Ces nouvelles expériences et améliorations proposées devraient nous permettre d'améliorer les connaissances portant sur les mécanismes d'évolution du métabolisme.

Annexes

Table 1 – *Tableau des méthodes et bases de données basées sur une approche de type graphe (Altenhoff and Dessimoz, 2012, Kuzniar et al., 2008, Trachana et al., 2011, Kristensen et al., 2011, Boeckmann et al., 2011)*

Méthode basée sur les graphes	Description	Appliqué à	In-paralogues	Basée sur	Stratégie de regroupement	Base de donnée	Algorithme/Base de donnée disponible	Buts	site web
COG (Tatusov et al., 1997)	Le graphe des relations potentielles d'homologie est créé à partir des meilleurs hits BLAST (BeTs). La méthode COG identifie des triangles de gènes dans ce graphe impliquant 3 espèces différentes. Les triangles sont regroupés si ils partagent une arrête commune.	711 génomes qui représentent la diversité des bactéries et des archées	oui	Scores blast	regroupement de triangles adjacents de meilleurs hits réciproques	COG/ KOG	X/X	Classification phylogénétique	http://www.ncbi.nlm.nih.gov/COG/
BRH (Overbeek et al., 1999)	Extraction des paires de meilleurs hits réciproques sur les résultats blast de comparaison des protéomes deux à deux.		non	Scores blast	/		-/-	-	-
Inparanoid (Remm et al., 2001)	Prédictions sur des paires de génomes. BRH entre les paires de génomes puis application de règles statistiques afin d'ajouter les in-paralogues	Version 8: 273 génomes	oui	Scores blast	Regroupements par paires de génomes	Inparanoid	X/X	comparaison de paires d'espèces	http://inparanoid.sbc.su.se/cgi-bin/index.cgi
Multiparanoid (Alexeyenko et al., 2006)	Combine les prédictions de Inparanoid afin de faire des prédiction de groupes avec plus de deux espèces via une approche de *clustering* lien simple.	Article: 4 génomes d'eukaryotes	oui	Scores blast	"lien simple"	Multiparanoid	X/X	Comparaison de groupes d'espèces relativement proches.	http://multiparanoid.sbc.su.se/
RSD (Wall et al., 2003)	La méthode RSD combine alignement local et global de séquences et l'estimation de la distance évolutive par maximum de vraisemblance pour prédire des paires d'orthologues.	Version avril 2013: 2044 génomes.	non	Distance évolutive évaluée par maximum de vraisemblance	/	RoundUp	X/X	Comparaison de paires de génomes	http://roundup.hms.harvard.edu/
OMA (Altenhoff et al., 2015)	OMA permet de prédire des groupes d'orthologues. Blast 'all against all' puis recherche de paires d'orthologues avec la méthode RSD et l'utilisation d'un intervalle de confiance pour éviter d'exclure les orthologues n:m. Formation des groupes par recherche de cliques. Peut détecter des pertes de gènes.	Version 2015 : 1706 génomes	oui	Distance évolutive évaluée par maximum de vraisemblance / SIMD et distance évolutive	toutes les paires sont des orthologues	OMA Browser	-/X	Comparative genomics, phyletic profiles, species phylogeny	http://omabrowser.org/
OrthoMCL (Li et al., 2003)	Graphe des intermédiaires des relations d'orthologies crée de manière similaire à Inparanoid. Le *clustering* se fait via un procédé de Markov impliquant la simulation d'une marche aléatoire dans le graphe.	Version 5 : 150 génomes	oui	Scores blast	Clusters MCL	OrthoMCL-DB	X/X		http://www.orthomcl.org/orthomcl/

ce tableau continue sur la page suivante

Méthode basée sur les graphes	Description	Appliqué à	In-paralogues	Basée sur	Stratégie de regroupement	Base de donnée	Algorithme/Base de donnée disponible	Buts	site web
EggNOG (Powell et al., 2014)	Etend les groupes COGs de manière incrémentale à différents niveaux de l'arbre taxonomique.	Version 4 : 3686 génomes	oui	Scores blast	regroupement de triangles adjacents de meilleurs hits réciproques	EggNog	-/X	Comparative genomics, species phylogeny	http://eggnog.embl.de
OrthoDB (Waterhouse et al., 2013)	*clustering* les BRH des gènes de chaque paire d'espèces déterminé grâce à un alignement des séquences protéiques via un *all-against-all* Smith–Waterman utilisant SWIPE. Les clusters sont construits de manière progressive avec des seuils différents s'il s'agit de BRH triangulaires ou non. Les in-paralogues sont ensuite ajouté sur le même principe qu'inparanoid. Cette méthode est effectué à tous les niveaux de l'arbre taxonomique	Version 8: 87 arthropodes, 61 vertébrés, 227 champignons et 2627 bactéries	oui	Smith-Waterman, RBH,	regroupement de triangles adjacents de meilleurs hits réciproques	OrthoDB	X/X	Comparative genomics, species phylogeny	http://www.orthodb.org/ et http://www.orthodb.org/software
Ortho-Inspector (Linard et al., 2011)	Crée des groupes d'in-paralogues et examine ensuite les matches réciproques 1-to-1, 1-to-many, ou many-to-many entre chaque paires de groupes, avec une détection supplémentaire des informations contradictoires entre les groupes (comme un protéome incomplet).	La base "Prokaryotes," contient les orthologues entre les protéomes de 120 Archées et 1568 Bacteries. La base "Eukaryotes" dataset contient 259 protéomes complets et couvre l'ensemble des phylomes eucaryotes	oui	Scores blast	uniquement entre paires d'espèces	OrthoInspector	X/-	-	http://www.lbgi.fr/orthoinspector/
Domain-based detection of orthologs (DODO) (Chen et al., 2010)	Approche de type BRH mais basée sur les domaines protéiques (identifiés par comparaison à InterPro) et non plus sur les séquences protéiques.	Benchmarked against InParanoid's 100 génomes.	oui	RPS-BLAST et Blast	Regroupe des séquences partageant les mêmes domaines dans le même ordre	-	X/-	-	http://140.109.42.19:16080/dodo_web/home.htm

ce tableau continue sur la page suivante

Méthode basée sur les graphes	Description	Appliqué à	Inparalogues	Basée sur	Stratégie de regroupement	Base de donnée	Algorithme/Base de donnée disponible	Buts	site web
Ortholuge (Fulton et al., 2006)	Le projet Ortholuge a pour but d'améliorer prédictions de BRH en tenant compte de la perte de gènes. La méthode est similaire à Inparanoid mais utilise une distance phylogénétique au lieu du score de similarité blast. Limité à des comparaisons par paires mais utilise un outgroup.	Triplet de génomes étudiés : mouse, rat, human / cattle, human, mouse / Escherichia coli K12, Pseudomonas syringae pv. tomato str. DC3000, and Pseudomonas putida KT2440/ 13 gamma-proteobacteria genomes	oui	distance phylogénétique	-	-	X/-	Prédiction d'orthologues sur des génomes incomplets	http://www.pathogenomics.ca/ortholuge

Table 2 – *Tableau des méthodes et bases de données basées sur une approche de type arbre (Altenhoff and Dessimoz, 2012, Kuzniar et al., 2008, Trachana et al., 2011, Kristensen et al., 2011, Boeckmann et al., 2011)*

	Programme	Méthode	Arbres des gènes	Arbre des espèces	enracinement de l'arbre des gènes	incertitude de l'arbre des gènes	Cadre	Algorithme / Base de donnée disponible	site web
Méthodes faisant intervenir l'arbre des espèces	TreeMap (Jackson and Charleston, 2004)	Réconcilie l'arbre des gènes avec l'arbre des espèces.	binaire et enraciné	binaire	n.a.	n.a.	Parcimonie	X/-	http://sydney.edu.au/engineering/it/~mcharles/software/treemap/treemap3.html
	Notung (Durand et al., 2006)	Enracinement de l'arbre des gènes en minimisant le nombre de duplications	binaire	binaire et non binaire	nombre minimum de duplication / perte de gènes	Bootstrap	Parcimonie	X/-	http://ppod.princeton.edu/help/help_notung_ortho_para.html
	Korak	Calcule le nombre de réconciliations, la vraisemblance des réconciliations de type LCA et la probabilité postérieure de chaque réconciliation. Nécessite du taux de duplication et de délétion.	enraciné	daté	n.a.	n.a.	Probabilités	X/X	http://paleogenomics.irmacs.sfu.ca/KORAK/
	CoRe-Pa (Merkle et al., 2010)	Approche de type réconciliation, ne nécessite pas de fournir les coûts associés aux différents évènements. Evènement gérés : cospeciations, sortings, duplications, and (host) switches.	enraciné	enraciné	n.a.	n.a.	Probabilités	-/-	-
	Jane(Conow et al., 2010)	Les dates sur l'arbre des espèces peuvent être fournies en entrée ou être calculées par un algorithme génétique. La distance évolutive entre deux espèces pouvant échanger des gènes est contrôlable.	binaire	binaire et daté	n.a.	n.a.	Probabilités	X/-	http://www.cs.hmc.edu/~hadas/jane/
	Mowgli (Doyon et al., 2010)	MPR et nombre de réconciliations équivalentes.	binaire	binaire et daté	n.a.	Nearest Neighbor Interchange	Parcimonie	X/-	http://www.atgc-montpellier.fr/Mowgli/

ce tableau continue sur la page suivante

	Programme	Méthode	Arbres des gènes	Arbre des espèces	enracinement de l'arbre des gènes	incertitude de l'arbre des gènes	Cadre	Algorithme / Base de donnée disponible	site web
	AnGST (analyser of gene and speciestree) (David and Alm, 2011)	Gère les parties de l'arbre phylogénétiques incertaines en inférant l'arbre des gènes via une combinaison de sous arbres obtenus par Bootstrap de manière à aboutir à la réconciliation de coût minimum. Capable de gérer: horizontal gene transfer (HGT), gene duplication (DUP), gene loss (LOS), speciation (SPC) and exactly one gene birth or genesis event (GEN)	binaire	binaire et daté	nombre minimum de duplication / perte de gènes	bootstrap	Parcimonie	X/-	http://almlab.mit.edu/angst/
	Tarzan	Merkle et al (Merkle and Middendorf, 2005) et Merkle et Middendorf (Merkle et al., 2010)	binaire ou non, intervalle de temps	binaire, daté	n.a.	n.a.	Parcimonie	X/-	http://pacosy.informatik.uni-leipzig.de/51-0-Tarzan.html
	OrthoStrapper (Storm and Sonnhammer, 2002)	Le programme Orthostrapper utilise une heuristique de recherche de similarité de séquence pour prédire des orthologues avec des scores de confiance à partir d'un arbre des gènes sur lequel a été appliqué un bootstrap. Orthostrapper n'utilise pas d'arbre des espèces à proprement parlé, à la place, les séquences sont assignées à un groupe taxonomique. Orthologues basés sur les domaines.	binaire	complètement résolu	nombre minimum de duplication	Bootstrap	Parcimonie	X/-	ftp://ftp.cgb.ki.se/pub/prog/orthostrapper/
	GSR (Åkerborg et al., 2009)	modèle probabiliste intégrant la duplication de gènes, l'évolution des séquences et un modèle d'horloge relaxée pour les taux de substitution.		complètement résolu	n.a.	n.a.	Probabilités	X/-	http://prime.sbc.su.se/primeGSR/
	HOGENOM (Dufayard et al., 2005, Penel et al., 2009)	Tient compte de la longueur des branches de l'arbre et des nœuds ayant plus de deux enfants		partiellement résolu	nombre minimum de duplication	Multifurcate	Parcimonie	X/X	
	Ensembl (Vilella et al., 2009b)/TreeBeST (Li et al., 2006)	Gestion des partie non résolue de l'arbre des espèces en permettant la présence de nœuds ayant plus de deux fils.		partiellement résolu	nombre minimum de duplication et nombre minimum de délétions	non	Parcimonie	-/X	

ce tableau continue sur la page suivante

	Programme	Méthode	Arbres des gènes	Arbre des espèces	enracinement de l'arbre des gènes	incertitude de l'arbre des gènes	Cadre	Algorithme / Base de donnée disponible	site web
	TreeFam (Li et al., 2006)	Les clusters TreeFam sont créés par *clustering* hiérarchique de résultats de blats all-versus-all. Les arbres des gènes sont construits en utilisant différentes approches incluant le maximum de vraisemblance et neighbor-joining. Les orthologues et paralogues sont inférés en utilisant la méthode DLI avec l'arbre taxonomique du NCBI. Des experts vérifient les arbres manuellement.	maximum de vraisemblance et neighbor-joining				Parcimonie	X/X	http://www.treefam.org/
Méthodes ne faisant intervenir que l'arbre des gènes	BranchClust (Poptsova and Gogarten, 2007)	Super-familles construites à partir des BH de blast. L'arbre est construit et analysé par BranchClust pour détecter les familles contenant plus de X% des taxons de l'étude.		Species overlap	Nombre minimum de clusters	non	n.a.	-/X	
	LOFT (van der Heijden et al., 2007)	Étiquetage des événements évolutifs des nœuds internes de l'arbre des gène sur la base du recouvrement entre les espèces des différents sous arbres (*species-overlap*)		Species overlap	nombre minimum de duplication	non	Parcimonie	X/-	
	Phylogeny (Lemoine et al., 2007)	creation de groupes par lien simple et évaluation/division des groupes (Muscle + UPGMA, PhyML, Reetree)	resolut		milieu	non	Parcimonie	X/-	
que l'arbre des espèces	Hieranoid (Schreiber and Sonnhammer, 2013)	Compare les génomes en fonction de la taxonomie. Applique Inparanoid récursivement. Groupes intermédiaires comparés par HMM	-	resolut	-	-		X/-	http://hieranoid.sbc.su.se/
Méta-approche (Pryszcz et al., 2011a)	MetaPhOrt (Pryszcz et al., 2011b)	Méthode utilisant conjointement plusieurs bases de données afin de pondérer leurs résultats	-	-	-	-	-	X/X	http://orthology.phylomedb.org/

Table 3 – *Classification des eumycota. Les espèces listées sur la dernière colonnes sont les espèces disponibles dans la base*

	Embranchements	Sous embranchements	Classes	Ordres	
Out-groups	Apusozoa				Thecamonas trahens ATCC 50062
	Filasterea				Capsaspora owczarzaki
	Choanoflagellida				Salpingoeca sp. ATCC 50818
					Monosiga brevicollis
Champignons inférieurs	Chytridiomycota		Chytridiomycetes	Chytridiales	Batrachochytrium dendrobatidis
				Cladochytriales	
				Lobulomycetales	
				Rhizophydiales	
				Spizellomycetales	Spizellomyces punctatus
			Monoblepharidomycetes	Monoblepharidales	
	Neocallimastigomycota		Neocallimastigomycetes	Neocallimastigales	
	Blastocladiomycota		Blastocladiomycetes	Blastocladiales	Allomyces macrogynus
	Zygomycota	Microsporidia			Nematocida parisii
					Antonospora locustae
					Nosema ceranae
					Enterocytozoon bieneusi
					Encephalitozoon intestinalis
					Encephalitozoon cuniculi
		Entomophthoromycotina	Entomophthoromycetes	Entomophthorales	
		Zoopagomycotina	Zoopagomycetes	Zoopagales	
		Kickxellomycotina	Kickxellomycetes	Asellariales	
				Kickxellales	

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
				Dimargaritales	
				Harpellales	
		Mucoromycotina	Mucoromycetes	Mucorales	Rhizopus oryzae
					Mucor circinelloides
					Phycomyces blakesleeanus
				Mortierellales	
				Endogonales	
	Glomeromycota		Glomeromycetes	Glomerales	
				Diversisporales	
				Paraglomerales	
				Archaeosporales	
				Geosiphonales	
	Ascomycota	Taphrinomycotina	Neolectomycetes	Neolectales	
			Pneumocystidomycetes	Pneumocystidales	
			Schizosaccharomycetes	Schizosaccharomycetales	Schizosaccharomyces octosporus
					Schizosaccharomyces cryophilus
					Schizosaccharomyces japonicus
					Schizosaccharomyces pombe
			Taphrinomycetes	Protomycetales	
				Taphrinales	
				Coryneliales	
				Medeolariales	
				Mycocaliciales	

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
				Pyrenulales	
			Archaeorhizomycetes	Archaeorhizomycetales	
		Saccharomycotina	Saccharomycetes	Saccharomycetales	Ogataea angusta
					Clavispora lusitaniae
					Lipomyces starkeyi
					Candida nivariensis
					Candida bracarensis
					Candida albicans WO-1
					Candida albicans SC5314
					Candida caseinolytica
					Candida tenuis
					Candida dubliniensis
					Candida tropicalis
					Candida parapsilosis
					Yarrowia lipolytica
					Eremothecium gossypii
					Kluyveromyces lactis
					Naumovozyma castellii
					Nakaseomyces bacillisporus
					Nakaseomyces delphensis
					Candida castellii
					Candida glabrata
					Zygosaccharomyces rouxii
					Lachancea thermotolerans
					Lachancea kluyveri

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
					Saccharomyces mikatae
					Saccharomyces kudriavzevii
					Saccharomyces paradoxus
					Saccharomyces cerevisiae
					Saccharomyces bayanus
					Wickerhamomyces anomalus
					Pichia membranifaciens
					Spathaspora passalidarum
					Lodderomyces elongisporus
					Debaryomyces hansenii
					Meyerozyma guilliermondii
					Scheffersomyces stipitis
		Pezizomycotina	Orbiliomycetes	Orbiliales	Arthrobotrys oligospora
			Pezizomycetes	Pezizales	Tuber melanosporum
			Geoglossomycetes	Geoglossales	
			Arthoniomycetes	Arthoniales	
			Dothideomycetes	Botryosphaeriales	
				Capnodiales	Baudoinia compniacensis
					Zymoseptoria tritici
					Pseudocercospora fijiensis
					Mycosphaerella populorum SO2202
					Mycosphaerella pini
				Dothideales	
				Hysteriales	Hysterium pulicare
					Rhytidhysteron rufulum

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
				Jahnulales	
				Microthyriales	
				Myriangiales	
				Mytilinidiales	
				Patellariales	
				Pleosporales	Phaeosphaeria nodorum
					Leptosphaeria maculans
					Setosphaeria turcica
					Pyrenophora teres f. teres
					Pyrenophora tritici-repentis
					Alternaria brassicicola
					Cochliobolus sativus
					Cochliobolus heterostrophus
				Trypetheliales	
			Eurotiomycetes	Chaetothyriales	
				Verrucariales	
				Arachnomycetales	
				Eurotiales	Emericella nidulans
					Neosartorya fischeri
					Talaromyces stipitatus
					Penicillium marneffei
					Penicillium chrysogenum
					Aspergillus fumigatus
					Aspergillus carbonarius
					Aspergillus terreus

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
					Aspergillus oryzae
					Aspergillus niger
					Aspergillus flavus
					Aspergillus clavatus
					Aspergillus aculeatus
				Onygenales	Ajellomyces dermatitidis ATCC 18188
					Ajellomyces dermatitidis SLH14081
					Ajellomyces dermatitidis ER-3
					Ajellomyces capsulatus H143
					Ajellomyces capsulatus H88
					Ajellomyces capsulatus G186AR
					Ajellomyces capsulatus NAm1
					Paracoccidioides sp. 'lutzii' Pb01
					Paracoccidioides brasiliensis Pb18
					Paracoccidioides brasiliensis Pb03
					Coccidioides posadasii RMSCC 3488
					Coccidioides posadasii str. Silveira
					Coccidioides immitis RMSCC 3703

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
					Coccidioides immitis RMSCC 2394
					Coccidioides immitis H538.4
					Coccidioides immitis RS
					Uncinocarpus reesii
					Microsporum gypseum
					Arthroderma otae
					Arthroderma benhamiae
					Trichophyton equinum
					Trichophyton verrucosum
					Trichophyton tonsurans
					Trichophyton rubrum
			Lecanoromycetes	Acarosporales	
				Lecanorales	Cladonia grayi
				Peltigerales	
				Agyriales	
				Gyalectales	
				Ostropales	
				Pertusariales	
				Teloschistales	
				Trichotheliales	
			Xylonomycetes	Xylonomycetales	
			Lichinomycetes	Lichinales	
			Leotiomycetes	Cyttariales	
				Erysiphales	
				Helotiales	Botryotinia fuckeliana

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
					Sclerotinia sclerotiorum
					Geomyces destructans
				Rhytismatales	
				Thelebolales	
			Laboulbeniomycetes	Laboulbeniales	
				Pyxidiophorales	
			Sordariomycetes	Calosphaeriales	
				Lulworthiales	
				Meliolales	
				Phyllachorales	
				Trichosphaeriales	
				Coronophorales	
				Glomerellales	Colletotrichum higginsianum
					Glomerella graminicola
					Acremonium alcalophilum
					Verticillium dahliae
					Verticillium albo-atrum
				Halosphaeriales	
				Hypocreales	Epichloe festucae
					Trichoderma atroviride
					Trichoderma reesei
					Hypocrea virens
					Nectria haematococca
					Gibberella moniliformis
					Gibberella zeae

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
					Fusarium oxysporum
				Microascales	
				Boliniales	
				Chaetosphaeriales	
				Coniochaetales	
				Diaporthales	Cryphonectria parasitica
				Ophiostomatales	
				Sordariales	Podospora anserina
					Chaetomium globosum
					Thielavia terrestris
					Myceliophthora thermophila
					Sordaria macrospora
					Neurospora tetrasperma
					Neurospora discreta
					Neurospora crassa
				Magnaporthales	Gaeumannomyces graminis
					Magnaporthe grisea
					Magnaporthe poae
				Spathulosporales	
				Xylariales	
	Basidiomycota	Puccinomycotina	Agaricostilbomycetes	Agaricostilbales	
				Spiculogloeales	
			Attractiellomycetes	Attractiellales	
			Classeiculomycetes	Classeiculales	
			Cryptomycocolacomycetes	Cryptomycocolacales	

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
			Cystobasidiomycetes	Cystobasidiales	
				Erythrobasidiales	
				Naohideales	
			Microbotryomycetes	Heterogastridiales	
				Leucosporidiales	
				Microbotryales	
				Sporidiobolales	Rhodotorula graminis
					Sporobolomyces roseus
			Mixiomycetes	Mixiales	
			Pucciniomycetes	Helicobasidiales	
				Pachnocybales	
				Platygloeales	
				Septobasidiales	
				Pucciniales	Puccinia triticina
					Puccinia graminis
					Melampsora larici-populina
			Tritirachiomycetes	Tritirachiales	
		Ustilaginomycotina	Entorrhizomycetes	Entorrhizales	
			Ustilaginomycetes	Urocystales	
				Ustilaginales	Ustilago maydis
			Exobasidiomycetes	Doassansiales	
				Entylomatales	
				Exobasidiales	
				Georgefischeriales	
				Malasseziales	Malassezia globosa

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
				Microstromatales	
				Tilletiales	
		Agaricomycotina	Wallemiomycetes	Wallemiales	Wallemia sebi
			Bartheletiomycetes	Bartheletiales	
			Tremellomycetes	Cystofilobasidiales	
				Filobasidiales	Cryptococcus neoformans
					Cryptococcus gattii WM276
					Cryptococcus gattii R265
				Tremellales	Tremella mesenterica
			Dacrymycetes	Dacrymycetales	Dacryopinax sp. DJM-731 SS1
			Agaricomycetes	Agaricales	Agaricus bisporus var. bisporus H97
					Agaricus bisporus var. burnettii JB137-S8
					Moniliophthora perniciosa
					Coprinopsis cinerea
					Schizophyllum commune
					Pleurotus ostreatus
				Amanitales	
				Atheliales	
				Auriculariales	Auricularia delicata
				Boletales	Serpula lacrymans
					Coniophora puteana
				Cantharellales	
				Clavariales	
				Corticiales	Punctularia strigosozonata

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
					Phlebia brevispora HHB-7030 SS6
					Phanerochaete carnosa
					Phanerochaete chrysosporium
				Cortinariales	
				Entolomatales	
				Geastrales	
				Gloeophyllales	Gloeophyllum trabeum
				Gomphales	
				Hymenochaetales	Fomitiporia mediterranea
				Hydnangiales	
				Hysterangiales	
				Jaapiales	
				Lepidostromatales	
				Lycoperdales	
				Nidulariales	
				Phallales	
				Pluteales	
				Polyporales	Ganoderma sp. 10597 SS1
					Dichomitus squalens
					Phlebiopsis gigantea
					Ceriporiopsis subvermispora
					Bjerkandera adusta
					Postia placenta
					Wolfiporia cocos
					Fomitopsis pinicola

ce tableau continue sur la page suivante

	Embranchements	Sous embranchements	Classes	Ordres	
					Trametes versicolor
				Russulales	Stereum hirsutum FP-91666 SS1
					Heterobasidion annosum
				Sebacinales	
				Thelephorales	
				Trechisporales	
				Tricholomatales	Laccaria bicolor
				Tulostomatales	

Table 4 – *Analyse des profils identiques pour plusieurs activités enzymatiques.*

# ec dans le profil	moy # voies par ec	# voies dans le profil	max ec dans une voie	proportion max ec dans la même voie	nom de la voie majoritaire
4	1,0	1	4	100,00%	Biotin metabolism
4	1,0	1	4	100,00%	Phenylalanine, tyrosine and tryptophan biosynthesis
3	1,0	1	3	100,00%	Folate biosynthesis
3	1,0	1	3	100,00%	Ubiquinone and other terpenoid-quinone biosynthesis
3	2,3	3	3	100,00%	Purine metabolism
3	2,0	2	3	100,00%	One carbon pool by folate
3	1,0	1	3	100,00%	Histidine metabolism
2	1,0	1	2	100,00%	Amino sugar and nucleotide sugar metabolism
2	1,0	1	2	100,00%	Inositol phosphate metabolism
2	1,0	1	2	100,00%	Pyrimidine metabolism
2	1,0	1	2	100,00%	Carotenoid biosynthesis
2	1,0	1	2	100,00%	Tryptophan metabolism
2	1,0	1	2	100,00%	Starch and sucrose metabolism
2	1,0	1	2	100,00%	Biotin metabolism
2	1,0	1	2	100,00%	Linoleic acid metabolism
2	2,0	2	2	100,00%	beta-Alanine metabolism
2	1,0	1	2	100,00%	Lipopolysaccharide biosynthesis
2	1,5	2	2	100,00%	Aminoacyl-tRNA biosynthesis
2	1,0	1	2	100,00%	Thiamine metabolism
2	1,0	1	2	100,00%	Lysine degradation
2	1,0	1	2	100,00%	Thiamine metabolism
2	4,0	7	2	100,00%	Glycolysis / Gluconeogenesis
2	1,0	1	2	100,00%	Nicotinate and nicotinamide metabolism
2	1,0	1	2	100,00%	Aflatoxin biosynthesis
2	2,5	3	2	100,00%	Glycine, serine and threonine metabolism
2	3,0	5	2	100,00%	Cyanoamino acid metabolism
2	3,5	6	2	100,00%	Fatty acid biosynthesis

ce tableau continue sur la page suivante

# ec dans le profil	moy # voies par ec	# voies dans le profil	max ec dans une voie	proportion max ec dans la même voie	nom de la voie majoritaire
2	4,0	7	2	100,00%	Phenylalanine metabolism
2	1,0	1	2	100,00%	Arginine and proline metabolism
2	1,0	1	2	100,00%	Porphyrin and chlorophyll metabolism
2	1,5	2	2	100,00%	Purine metabolism
2	1,5	2	2	100,00%	N-Glycan biosynthesis
2	1,5	2	2	100,00%	Pyruvate metabolism
2	2,0	2	2	100,00%	N-Glycan biosynthesis
2	1,0	1	2	100,00%	Tyrosine metabolism
2	1,0	1	2	100,00%	Pentose phosphate pathway
4	1,0	2	3	75,00%	Fatty acid biosynthesis
3	0,7	1	2	66,67%	Terpenoid backbone biosynthesis
3	1,0	2	2	66,67%	Arachidonic acid metabolism
3	1,0	2	2	66,67%	Sulfur metabolism
4	1,0	3	2	50,00%	Histidine metabolism
4	1,5	5	2	50,00%	Purine metabolism
5	1,2	4	2	40,00%	Cysteine and methionine metabolism
6	1,3	7	2	33,33%	Aminoacyl-tRNA biosynthesis
7	1,6	10	2	28,57%	Aminoacyl-tRNA biosynthesis
20	2,9	37	4	20,00%	Citrate cycle (TCA cycle)
21	1,0	19	2	9,52%	Tryptophan metabolism
54	1,1	39	5	9,26%	Pyrimidine metabolism
28	0,4	9	2	7,14%	Aminoacyl-tRNA biosynthesis
2	2,0	4	1	50,00%	Pentose and glucuronate interconversions
2	3,5	7	1	50,00%	Butirosin and neomycin biosynthesis
2	0,5	1	1	50,00%	Penicillin and cephalosporin biosynthesis
2	1,0	2	1	50,00%	Styrene degradation
2	1,0	2	1	50,00%	Metabolism of xenobiotics by cytochrome P450
2	0,5	1	1	50,00%	Galactose metabolism

ce tableau continue sur la page suivante

# ec dans le profil	moy # voies par ec	# voies dans le profil	max ec dans une voie	proportion max ec dans la même voie	nom de la voie majoritaire
2	0,5	1	1	50,00%	Glyoxylate and dicarboxylate metabolism
2	1,0	2	1	50,00%	Purine metabolism
2	1,0	2	1	50,00%	Cysteine and methionine metabolism
2	1,5	3	1	50,00%	Methane metabolism
2	2,5	5	1	50,00%	Glycine, serine and threonine metabolism
2	2,0	4	1	50,00%	Butanoate metabolism
2	0,5	1	1	50,00%	Other types of O-glycan biosynthesis
2	1,0	2	1	50,00%	Alanine, aspartate and glutamate metabolism
2	1,0	2	1	50,00%	Glycerolipid metabolism
2	0,5	1	1	50,00%	Methane metabolism
2	1,0	2	1	50,00%	Arginine and proline metabolism
2	2,0	4	1	50,00%	Pantothenate and CoA biosynthesis
2	1,5	3	1	50,00%	Glycerophospholipid metabolism
3	1,3	4	1	33,33%	Glycerophospholipid metabolism
3	0,3	1	1	33,33%	Purine metabolism
3	0,3	1	1	33,33%	Lysine degradation
3	0,7	2	1	33,33%	Glycosylphosphatidylinositol(GPI)-anchor biosynthesis
3	1,7	5	1	33,33%	Phenylalanine, tyrosine and tryptophan biosynthesis
3	0,3	1	1	33,33%	Inositol phosphate metabolism
3	1,0	3	1	33,33%	Glycolysis / Gluconeogenesis
3	1,0	3	1	33,33%	Porphyrin and chlorophyll metabolism
3	0,3	1	1	33,33%	Sesquiterpenoid and triterpenoid biosynthesis
3	0,3	1	1	33,33%	Steroid biosynthesis
4	1,0	4	1	25,00%	Arginine and proline metabolism

ce tableau continue sur la page suivante

# ec dans le profil	moy # voies par ec	# voies dans le profil	max ec dans une voie	proportion max ec dans la même voie	nom de la voie majoritaire
4	3,5	14	1	25,00%	Pyruvate metabolism
4	2,3	9	1	25,00%	Steroid biosynthesis
5	1,0	5	1	20,00%	Linoleic acid metabolism
8	0,8	6	1	12,50%	Valine, leucine and isoleucine degradation
4	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	
2	0,0	0	0	0,00%	

Table 5 – *Score RELIEF moyen en fonction du nombre de profils sélectionnés.*

Groupe taxonomique	RELIEF moyen	RELIEF 1er	RELIEF moyen top 5	RELIEF moyen top 10	RELIEF moyen top 100
Agaricales	0.157	0.792	0.739	0.710	0.534
Agaricomycetes	0.133	0.959	0.883	0.830	0.585
Agaricomycotina	0.129	0.877	0.841	0.812	0.580
Ascomycota	0.123	0.918	0.885	0.855	0.596
Basidiomycota	0.125	0.934	0.873	0.842	0.597
Capnodiales	0.121	0.784	0.715	0.668	0.485
Corticiales	0.132	0.817	0.772	0.752	0.560
Dikarya	0.267	0.984	0.858	0.788	0.581
Dothideomycetes	0.121	0.855	0.703	0.660	0.470
Eurotiales	0.123	0.791	0.702	0.649	0.472
Eurotiomycetes	0.093	0.675	0.610	0.566	0.375
Glomerellales	0.128	0.763	0.698	0.662	0.469
Hypocreales	0.131	0.890	0.746	0.696	0.485
Leotiomycetes	0.116	0.743	0.694	0.657	0.452
Magnaporthales	0.129	0.869	0.786	0.738	0.497

ce tableau continue sur la page suivante

Groupe taxonomique	RELIEF moyen	RELIEF 1er	RELIEF moyen top 5	RELIEF moyen top 10	RELIEF moyen top 100
Microsporidia	0.460	1.000	0.990	0.987	0.960
Mucorales	0.152	1.000	0.984	0.957	0.660
Onygenales	0.105	0.834	0.669	0.609	0.433
Pezizomycotina	0.131	0.923	0.892	0.860	0.586
Pleosporales	0.128	0.786	0.747	0.709	0.516
Polyporales	0.129	0.854	0.789	0.760	0.566
Pucciniomycetes	0.166	0.924	0.909	0.877	0.649
Pucciniomycotina	0.154	0.874	0.815	0.764	0.544
Saccharomycetes	0.139	0.887	0.857	0.824	0.594
Sordariales	0.120	0.855	0.736	0.682	0.482
Sordariomycetes	0.110	0.739	0.710	0.684	0.447
Taphrinomycotina	0.181	1.000	0.965	0.941	0.766
Tremellomycetes	0.138	0.932	0.911	0.871	0.631
Zygomycota	0.152	1.000	0.984	0.957	0.660

Table 6 – *Ratio Gain d'information (RGI) moyen en fonction du nombre d'EC conservés après les avoir réordonnés.*

Groupe taxonomique	RGI moyen	RGI 1er	RGI moyen top 5	RGI moyen top 10	RGI moyen top 100
Agaricales	0.021	0.533	0.289	0.230	0.095
Agaricomycetes	0.065	0.817	0.722	0.620	0.311
Agaricomycotina	0.068	0.717	0.654	0.612	0.329
Ascomycota	0.119	0.837	0.780	0.734	0.437
Basidiomycota	0.072	0.785	0.697	0.658	0.367
Capnodiales	0.014	0.210	0.166	0.138	0.055
Corticiales	0.013	0.159	0.122	0.105	0.051
Dikarya	0.136	0.868	0.813	0.743	0.507
Dothideomycetes	0.037	0.460	0.371	0.311	0.133
Eurotiales	0.040	0.579	0.486	0.416	0.164
Eurotiomycetes	0.066	0.519	0.418	0.362	0.207
Glomerellales	0.014	0.257	0.152	0.120	0.050
Hypocreales	0.025	0.510	0.422	0.336	0.112
Leotiomycetes	0.008	0.263	0.154	0.109	0.032
Magnaporthales	0.011	0.690	0.313	0.209	0.049
Microsporidia	0.175	1.000	0.905	0.849	0.649
Mucorales	0.017	1.000	0.713	0.560	0.108
Onygenales	0.050	0.528	0.425	0.335	0.157
Pezizomycotina	0.159	0.892	0.833	0.779	0.468
Pleosporales	0.024	0.300	0.241	0.203	0.093
Polyporales	0.027	0.396	0.275	0.235	0.113
Pucciniomycetes	0.014	0.266	0.223	0.191	0.072
Pucciniomycotina	0.016	0.222	0.203	0.173	0.069
Saccharomycetes	0.099	0.846	0.797	0.743	0.440
Sordariales	0.021	0.588	0.371	0.259	0.087
Sordariomycetes	0.054	0.494	0.425	0.384	0.201
Taphrinomycotina	0.025	1.000	0.653	0.464	0.152
Tremellomycetes	0.015	0.288	0.255	0.201	0.072
Zygomycota	0.017	1.000	0.713	0.560	0.108

Bibliographie

2013. Belin.

Aha, D. and D. Kibler

1991. Instance-based learning algorithms. *Machine Learning*, 6:37–66.

Alex J, N., Z. Qian, O. Jason, H. Hanaa, B. Nilakshee, M. Alan G, and S. Scott M

2013. The COPII cage sharpens its image. *Nat Struct Mol Biol*, 20(2):139–140.

Alexeyenko, A., I. Tamas, G. Liu, and E. L. Sonnhammer

2006. Automatic clustering of orthologs and inparalogs shared by multiple proteomes. *Bioinformatics*, 22(14):e9–e15.

Altenhoff, A. and C. Dessimoz

2012. Inferring orthology and paralogy. In *Evolutionary Genomics*, M. Anisimova, ed., volume 855 of *Methods in Molecular Biology*, Pp. 259–279. Humana Press.

Altenhoff, A. M. and C. Dessimoz

2009. Phylogenetic and functional assessment of orthologs inference projects and methods. *PLoS Comput Biol*, 5(1):e1000262.

Altenhoff, A. M., M. Gil, G. H. Gonnet, and C. Dessimoz

2013. Inferring hierarchical orthologous groups from orthologous gene pairs. *PLoS ONE*, 8(1):e53786.

Altenhoff, A. M., N. Škunca, N. Glover, C.-M. Train, A. Sueki, I. Piližota, K. Gori, B. Tomiczek, S. Müller, H. Redestig, G. H. Gonnet, and C. Dessimoz

2015. The oma orthology database in 2015: function predictions, better plant support, synteny view and other improvements. *Nucleic Acids Research*, 43(D1):D240–D249.

Altenhoff, A. M., R. A. Studer, M. Robinson-Rechavi, and C. Dessimoz

2012. Resolving the ortholog conjecture: Orthologs tend to be weakly, but significantly, more similar in function than paralogs. *PLoS Comput Biol*, 8(5):e1002514.

Altman, T., M. Travers, A. Kothari, R. Caspi, and P. Karp

2013. A systematic comparison of the metacyc and kegg pathway databases. *BMC Bioinformatics*, 14(1):112.

Alves, R., R. A. Chaleil, and M. J. Sternberg

2002. Evolution of enzymes in metabolism: A network perspective. *Journal of Molecular Biology*, 320(4):751 – 770.

Antonov, A. V. and H. W. Mewes

2008. Complex phylogenetic profiling reveals fundamental genotype-phenotype associations. *Computational biology and chemistry*, 32(6):412–6.

Aparicio, O., J. V. Geisberg, and K. Struhl

2001. *Chromatin Immunoprecipitation for Determining the Association of Proteins with Specific Genomic Sequences In Vivo*, chapter 17. John Wiley & Sons, Inc.

Arvestad, L., J. Lagergren, and B. Sennblad

2009. The gene evolution model and computing its associated probabilities. *J. ACM*, 56(2):7:1–7:44.

Azé, J.

2012. *Prédiction d'Interactions et Amarrage Protéine-Protéine par combinaison de classifieurs.*

Baba, T. and O. Schneewind

1998. Instruments of microbial warfare: Bacteriocin synthesis, toxicity and immunity. *Trends in Microbiology*, 6(2):66 – 71.

Bender, M. and M. Farach-Colton

2000. The lca problem revisited. In *LATIN 2000: Theoretical Informatics*, G. Gonnet and A. Viola, eds., volume 1776 of *Lecture Notes in Computer Science*, Pp. 88–94. Springer Berlin Heidelberg.

Berglund-Sonnhammer, A.-C., P. Steffansson, M. J. Betts, and D. A. Liberles

2006. Optimal gene trees from sequences and species trees using a soft interpretation of parsimony. *J Mol Evol*, 63(2):240–50.

Boeckmann, B., M. Robinson-Rechavi, I. Xenarios, and C. Dessimoz

2011. Conceptual framework and pilot study to benchmark phylogenomic databases based on reference gene trees. *Briefings in bioinformatics*, 12(5):423–35.

Bok, J. W., S. A. Balajee, K. A. Marr, D. Andes, K. F. Nielsen, J. C. Frisvad, and N. P. Keller

2005. Laea, a regulator of morphogenetic fungal virulence factors. *Eukaryotic Cell*, 4(9):1574–1582.

Bolón-Canedo, V., N. Sánchez-Maroño, and A. Alonso-Betanzos

2013. A review of feature selection methods on synthetic data. *Knowledge and Information Systems*, 34(3):483–519.

Boone, C., H. Bussey, and B. J. Andrews

2007. Exploring genetic interactions and networks with yeast. *Nature Reviws Genetics*, 8:437–449.

Bowers, P. M., M. Pellegrini, M. J. Thompson, J. Fierro, T. O. Yeates, and D. Eisenberg

2004. Prolinks: a database of protein functional linkages derived from coevolution. *Genome biology*, 5(5):R35.

Brakhage, A. A.

2013. Regulation of fungal secondary metabolism. *Nat Rev Micro*, 11.

Breiman, L.

2001. Random forests. *Machine Learning*, 45(1):5–32.

Brygoo, Y. and R. Debuchy

1985. Transformation by integration in podospora anserina. *Molecular and General Genetics MGG*, 200(1):128–131.

Caetano-Anollés, G., L. S. Yafremava, H. Gee, D. Caetano-Anollés, H. S. Kim, and J. E. Mittenthal

2009. The origin and evolution of modern metabolism. *The International Journal of Biochemistry and Cell Biology*, 41(2):285 – 297. Molecular and Cellular Evolution: A Celebration of the 200th Anniversary of the Birth of Charles Darwin.

Caspi, R., H. Foerster, C. A. Fulcher, P. Kaipa, M. Krummenacker, M. Latendresse, S. Paley, S. Y. Rhee, A. G. Shearer, C. Tissier, T. C. Walk, P. Zhang, and P. D. Karp

2008. The metacyc database of metabolic pathways and enzymes and the biocyc collection of pathway/genome databases. *Nucleic Acids Research*, 36(suppl 1):D623–D631.

Chen, F., A. J. Mackey, J. K. Vermunt, and D. S. Roos

2007. Assessing Performance of Orthology Detection Strategies Applied to Eukaryotic Genomes. *PLoS ONE*, 2(4):12.

Chen, L. and D. Vitkup

2006. Predicting genes for orphan metabolic activities using phylogenetic profiles. *Genome Biology*, 7(2):R17.

Chen, T.-w., T. Wu, W. Ng, and W.-c. Lin

2010. Dodo: an efficient orthologous genes assignment tool based on domain architectures. domain based ortholog detection. *BMC Bioinformatics*, 11(Suppl 7):S6.

Cohen, W.

1995. Fast effective rule induction. In *Twelfth International Conference on Machine Learning*, Pp. 115–123.

Cokus, S., S. Mizutani, and M. Pellegrini

2007. An improved method for identifying functionally linked proteins using phylogenetic profiles. *BMC bioinformatics*, 8 Suppl 4:S7.

Conow, C., D. Fielder, Y. Ovadia, and R. Libeskind-Hadas

2010. Jane: a new tool for the cophylogeny reconstruction problem. *Algorithms for Molecular Biology*, 5(1):16.

Cookson, W.

2003. A new gene for asthma: would you ADAM and Eve it? *Trends in genetics : TIG*, 19(4):169–72.

Csűös, M.

2010. Count: evolutionary analysis of phylogenetic profiles with parsimony and likelihood. *Bioinformatics*, 26(15):1910–1912.

Cuomo, C. A. and B. W. Birren

2010. Chapter 34 - the fungal genome initiative and lessons learned from genome sequencing. In *Guide to Yeast Genetics: Functional Genomics, Proteomics, and Other Systems Analysis*, J. W. C. Guthrie and G. R. Fink, eds., volume 470 of *Methods in Enzymology*, Pp. 833 – 855. Academic Press.

Dalquen, D. a., A. M. Altenhoff, G. H. Gonnet, and C. Dessimoz

2013. The impact of gene duplication, insertion, deletion, lateral gene transfer and sequencing error on orthology inference: a simulation study. *PloS one*, 8(2):e56925.

Dalquen, D. A. and C. Dessimoz

2013. Bidirectional best hits miss many orthologs in duplication-rich clades such as plants and animals. *Genome Biology and Evolution*, 5(10):1800–1806.

Dandekar, T., B. Snel, M. Huynen, and P. Bork

1998. Conservation of gene order: A fingerprint of proteins that physically interact.

David, L. A. and E. J. Alm

2011. Rapid evolutionary innovation during an Archaean genetic expansion. *Nature*, 469(7328):93–96.

Davies, J.

2006. Are antibiotics naturally antibiotics? *Journal of Industrial Microbiology and Biotechnology*, 33(7):496–499.

de Ridder, D., J. de Ridder, and M. J. T. Reinders

2013. Pattern recognition in bioinformatics. *Briefings in Bioinformatics*, 14(5):633–647.

Dessimoz, C., B. Boeckmann, A. C. J. Roth, and G. H. Gonnet

2006. Detecting non-orthology in the cogs database and other approaches grouping orthologs using genome-specific best hits. *Nucleic Acids Research*, 34(11):3309–3316.

Dessimoz, C., T. Gabaldón, D. S. Roos, E. Sonnhammer, J. Herrero, and the Quest for Orthologs Consortium

2012. Toward community standards in the quest for orthologs. *Bioinformatics*.

Doyon, J.-P., S. Hamel, and C. Chauve

2012. An efficient method for exploring the space of gene tree/species tree reconciliations in a probabilistic framework. *Computational Biology and Bioinformatics, IEEE/ACM Transactions on*, 9(1):26–39.

Doyon, J.-P., C. Scornavacca, K. Y. Gorbunov, G. Szollosi, V. Ranwez, and V. Berry

2010. An efficient algorithm for gene/species trees parsimonious reconciliation with losses, duplications and transfers. *RECOMB-CG*.

du Plessis, L., N. Škunca, and C. Dessimoz

2011. The what, where, how and why of gene ontology—a primer for bioinformaticians. *Briefings in Bioinformatics*, 12(6):723–735.

Dufayard, J.-F., L. Duret, S. Penel, M. Gouy, F. Rechenmann, and G. Perrière

2005. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, 21(11):2596–2603.

Durand, D., B. V. Halldórsson, and B. Vernot.

2006. A hybrid micro–macroevolutionary approach to gene tree reconstruction. *Journal of Computational Biology.*, 13(2):320–335.

Eddy, S. R.

2011. Accelerated profile hmm searches. *PLoS Comput Biol*, 7(10):e1002195.

Enright, a. J., I. Iliopoulos, N. C. Kyrpides, and C. a. Ouzounis

1999. Protein interaction maps for complete genomes based on gene fusion events. *Nature*, 402(6757):86–90.

Ferrer, L., J. M. Dale, and P. D. Karp

2010. A systematic study of genome context methods: calibration, normalization and combination. *BMC bioinformatics*, 11(1):493.

Firn, R. D. and C. G. Jones

2000. The evolution of secondary metabolism - a unifying model. *Molecular microbiology 37*, Pp. 989–994.

Frank, E. and I. H. Witten

1998. Generating accurate rule sets without global optimization. In *Fifteenth International Conference on Machine Learning*, J. Shavlik, ed., Pp. 144–151. Morgan Kaufmann.

Fulton, D., Y. Li, M. Laird, B. Horsman, F. Roche, and F. Brinkman

2006. Improving the specificity of high-throughput ortholog prediction. *BMC Bioinformatics*, 7(1):270.

Gabaldón, T. and E. V. Koonin

2013. Functional and evolutionary implications of gene orthology. *Nature reviews. Genetics*, 14(5):360–6.

Gabaldon, T., C. Dessimoz, J. Huxley-Jones, A. Vilella, E. Sonnhammer, and S. Lewis

2009. Joining forces in the quest for orthologs. *Genome Biology*, 10(9):403.

Goffeau, A., B. G. Barrell, H. Bussey, R. W. Davis, B. Dujon, H. Feldmann, F. Galibert, J. D. Hoheisel, C. Jacq, M. Johnston, E. J. Louis, H. W. Mewes, Y. Murakami, P. Philippsen, H. Tettelin, and S. G. Oliver

1996. Life with 6000 genes. *Science*, 274(5287):546–567.

Goodman, M., J. Czelusniak, G. W. Moore, A. E. Romero-Herrera, and G. Matsuda

1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. *Systematic Biology*, 28(2):132–163.

Grigoriev, I., G. S. B. Cullen, Danie and, D. Hibbett, T. W. Jeffries, C. P. Kubicek, C. Kuske, J. K. Magnuson, F. Martin, J. W. Spatafora, A. Tsang, and S. E. Baker

2011. Fueling the future with fungal genomics. *Mycology*, 2(3):192–209.

Grigoriev, I. V., R. Nikitin, S. Haridas, A. Kuo, R. Ohm, R. Otillar, R. Riley, A. Salamov, X. Zhao, F. Korzeniewski, T. Smirnova, H. Nordberg, I. Dubchak, and I. Shabalov

2014. Mycocosm portal: gearing up for 1000 fungal genomes. *Nucleic Acids Research*, 42(D1):D699–D704.

Grognet, P., H. Lalucque, and P. Silar

2012. The paalr1 magnesium transporter is required for ascospore development in podospora anserina. *Fungal Biology*, 116(10):1111 – 1118.

Grossetete, S., B. Labedan, and O. Lespinet

2010. Fungipath: a tool to assess fungal metabolic pathways predicted by orthology. *BMC Genomics*, 11(1):81.

Guyon, I. and A. Elisseeff

2003. An introduction to variable and feature selection. *J. Mach. Learn. Res.*, 3:1157–1182.

Guyon, I., J. Weston, S. Barnhill, and V. Vapnik

2002. Gene selection for cancer classification using support vector machines. *Machine Learning*, 46(1-3):389–422.

Hallett, M. T. and J. Lagergren

2000. New algorithms for the duplication-loss model. In *Proceedings of the Fourth Annual International Conference on Computational Molecular Biology*, RECOMB '00, Pp. 138–146, New York, NY, USA. ACM.

Hawksworth, D.

1991. The fungal dimension of biodiversity: magnitude, significance, and conservation. *Mycological Research*, 95(6):641 – 655.

Horowitz, N.

1945. On the evolution of biochemical syntheses. *Proc Natl Acad Sci USA*.

Huerta-Cepas, J., H. Dopazo, J. Dopazo, and T. Gabaldon

2007. The human phylome. *Genome Biology*, 8(6):R109.

Hulsen, T., M. A. Huynen, J. De Vlieg, and P. M. Groenen

2006. Benchmarking ortholog identification methods using functional genomics data. *Genome Biology*, 7(4):R31.

Jackson, A. P. and M. A. Charleston

2004. A cophylogenetic perspective of rna–virus evolution. *Molecular Biology and Evolution*, 21(1):45–57.

Jafari, P. and F. Azuaje

2006. An assessment of recently published gene expression data analyses: reporting experimental design and statistical factors. *BMC Medical Informatics and Decision Making*, 6(1):27.

Jain, A. and D. Zongker

1997. Feature selection: Evaluation, application, and small sample performance. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 19:153–158.

Jenke-Kodama, H., R. Müller, and E. Dittmann

2008. Evolutionary mechanisms underlying secondary metabolite diversity. In *Natural Compounds as Drugs Volume I*, F. Petersen and R. Amstutz, eds., volume 65 of *Progress in Drug Research*, Pp. 119–140. Birkhäuser Basel.

Jensen, R. A.

1976. Enzyme Recruitment in Evolution of New Function. *Annual Review of Microbiology*, 30(1):409–425.

Jim, K., K. Parmar, M. Singh, and S. Tavazoie

2004. A cross-genomic approach for systematic mapping of phenotypic traits to genes. *Genome Res.*, Pp. 109–115.

Jing, L., F. Shah, M. Mohamad, N. Hamran, A. Salleh, S. Deris, and H. Alashwal

2014. Database and tools for metabolic network analysis. *Biotechnology and Bioprocess Engineering*, 19(4):568–585.

John, G. H. and P. Langley

1995. Estimating continuous distributions in bayesian classifiers. In *Eleventh Conference on Uncertainty in Artificial Intelligence*, Pp. 338–345, San Mateo. Morgan Kaufmann.

Joshi-Tope, G., G. M., V. I, D. E. P., S. E., d. B. B., and J. e. a. B.

2005. Reactome: a knowledgebase of biological pathways. *Nucleic acids research 33*, Database issue:D428–D432.

Kanehisa, M. and S. Goto

2000. Kegg: Kyoto encyclopedia of genes and genomes. *Nucleic Acids Research*, 28(1):27–30.

Kanehisa, M., S. Goto, Y. Sato, M. Kawashima, M. Furumichi, and M. Tanabe

2014. Data, information, knowledge and principle: back to metabolism in kegg. *Nucleic Acids Research*, 42(D1):D199–D205.

Katoh, K., K. Misawa, K. Kuma, and T. Miyata

2002. Mafft: a novel method for rapid multiple sequence alignment based on fast fourier transform. *Nucleic Acids Research*, 30(14):3059–3066.

Åkerborg, r., B. Sennblad, L. Arvestad, and J. Lagergren

2009. Simultaneous bayesian gene tree reconstruction and reconciliation analysis. *Proceedings of the National Academy of Sciences*, 106(14):5714–5719.

Koski, L. and G. Golding

2001. The closest blast hit is often not the nearest neighbor. *Journal of Molecular Evolution*, 52(6):540–542.

Kotaru, A. R., K. Shameer, P. Sundaramurthy, and R. C. Joshi

2013. An improved hypergeometric probability method for identification of functionally linked proteins using phylogenetic profiles. *Bioinformation*, 9(7):368–74.

Kristensen, D. M., Y. I. Wolf, A. R. Mushegian, and E. V. Koonin

2011. Computational methods for gene orthology inference. *Briefings in Bioinformatics*, 12(5):379–391.

Kuzniar, A., R. C. H. J. van Ham, S. Pongor, and J. A. M. Leunissen

2008. The quest for orthologs: finding the corresponding gene across genomes. *Trends in Genetics*, 24(11):539–551.

Lalami, S. G.

2010. *Génomique comparée et développement de nouveaux outils bioinformatiques permettant l'analyse de la diversité métabolique des Eumycota*. PhD thesis, Université Paris-Sud.

Larrañaga, P., B. Calvo, R. Santana, C. Bielza, J. Galdiano, I. Inza, J. A. Lozano, R. Armañanzas, G. Santafé, A. Pérez, and V. Robles

2006. Machine learning in bioinformatics. *Briefings in Bioinformatics*, 7(1):86–112.

Laskowski, R. a., J. D. Watson, and J. M. Thornton

2005. ProFunc: a server for predicting protein function from 3D structure. *Nucleic acids research*, 33(Web Server issue):W89–93.

Lazcano, A. and S. L. Miller

1996. The Origin and Early Evolution of Life: Prebiotic Chemistry, the Pre-RNA World, and Time. *Cell*, 85(6):793–798.

Lemoine, F., O. Lespinet, and B. Labedan

2007. Assessing the evolutionary rate of positional orthologous genes in prokaryotes using synteny data. *BMC Evolutionary Biology*, 7(1):237.

Li, H., A. Coghlan, J. Ruan, L. J. Coin, J.-K. Hériché, L. Osmotherly, R. Li, T. Liu, Z. Zhang, L. Bolund, G. K.-S. Wong, W. Zheng, P. Dehal, J. Wang, and R. Durbin

2006. Treefam: a curated database of phylogenetic trees of animal gene families. *Nucleic Acids Research*, 34(suppl 1):D572–D580.

Li, L., C. J. Stoeckert, and D. S. Roos

2003. Orthomcl: Identification of ortholog groups for eukaryotic genomes. *Genome Research*, 13(9):2178–2189.

Li, Y., S. E. Calvo, R. Gutman, J. S. Liu, and V. K. Mootha

2014. Expansion of biological pathways based on evolutionary inference. *Cell*, 158(1):213 – 225.

Linard, B., J. Thompson, O. Poch, and O. Lecompte

2011. Orthoinspector: comprehensive orthology analysis and visual exploration. *BMC Bioinformatics*, 12(1):11.

Lombard, J. and D. Moreira

2011. Origins and early evolution of the mevalonate pathway of isoprenoid biosynthesis in the three domains of life. *Molecular Biology and Evolution*, 28(1):87–99.

Ma, S. and J. Huang

2008. Penalized feature selection and classification in bioinformatics. *Briefings in Bioinformatics*, 9(5):392–403.

MacDonald, N. J. and R. G. Beiko

2010. Efficient learning of microbial genotype–phenotype association rules. *Bioinformatics*, 26(15):1834–1840.

Malik, V.

1980. Microbial secondary metabolism. *Trends in Biochemical Sciences*, 5(3):68 – 72.

Marcotte, E. M., I. Xenarios, A. M. van der Bliek, and D. Eisenberg

2000. Localizing proteins in the cell from their phylogenetic profiles. *Proceedings of the National Academy of Sciences*, 97(22):12115–12120.

Merkle, D. and M. Middendorf

2005. Reconstruction of the cophylogenetic history of related phylogenetic trees with divergence timing information. *Theory in Biosciences*, 123(4):277–299.

Merkle, D., M. Middendorf, and N. Wieseke

2010. A parameter-adaptive dynamic programming approach for inferring cophylogenies. *BMC Bioinformatics*, 11(Suppl 1):S60.

Mikkelsen, T. S., J. E. Galagan, and J. P. Mesirov

2005. Improving genome annotations using phylogenetic profile anomaly detection. *Bioinformatics (Oxford, England)*, 21(4):464–70.

Morgat, A., E. Coissac, E. Coudert, K. B. Axelsen, G. Keller, A. Bairoch, A. Bridge, L. Bougueleret, I. Xenarios, and A. Viari

2012. Unipathway: a resource for the exploration and annotation of metabolic pathways. *Nucleic Acids Research*, 40(D1):D761–D769.

nature education

2015. *Glossary*. nature education.

Nedelcu, A. M., I. H. Miles, A. M. Fagir, and K. Karol

2008. Adaptive eukaryote-to-eukaryote lateral gene transfer: stress-related genes of algal origin in the closest unicellular relatives of animals. *Journal of Evolutionary Biology*, 21(6):1852–1860.

Overbeek, R., M. Fonstein, M. D'Souza, G. D. Pusch, and N. Maltsev

1999. Use of contiguity on the chromosome to predict functional coupling. *In silico biology*, 1(2):93–108.

page web 1000 genomes

. Liste des génomes du projet 1000 génomes de champignons. http://genome.jgi.doe.gov/pages/fungi-1000-projects.jsf. Accessed: 2014-09-05.

page web BIGG

. Site web de la base de donnee bigg. bigg.ucsd.edu. Accessed: 2014-12-09.

page web BioCyc

. Site web de la base de donnee biocyc. Biocyc.org. Accessed: 2014-12-09.

page web EBI

. Site web répertoriant les protéomes de référence *reference proteomes*. http://www.ebi.ac.uk/reference_proteomes. Accessed: 2014-09-05.

page web FGI

. Liste des génomes du *Fungal Genome Initiative*. http://www.broadinstitute.org/scientific-community/science/projects/fungal-genome-initiative/status-fgi-projects. Accessed: 2014-09-05.

page web MARIO

. Site web de distribution des scripts de la méta-approche mario ainsi que des données prédites sur les benchmarks. http://bim.igmors.u-psud.fr/mario/. Accessed: 2014-09-05.

page web model SEED

. Site web de model seed. http://seed-viewer.theseed.org/. Accessed: 2014-12-09.

page web OrthoMCL

. Site web répertoriant les scripts et prédictions de orthomcl. http://www.orthomcl.org/orthomcl/. Accessed: 2014-09-05.

page web Philippe Silar

. Les eucaryotes: origine, évolution, diversité et biologie. http://cgdc3.igmors.u-psud.fr/microbiologie/partie1/chap3_02_eumycota.htm. Accessed: 2014-09-14.

page web Reactome

. Site web de la base de donnée réactome. http://www.reactome.org/. Accessed: 2014-12-09.

Palidwor, G. and M. Andrade-Navarro

2010. Mltrends: Graphing medline term usage over time. *Journal of Biomedical Discovery and Collaboration*, 5(0).

Patil, D. V. and R. S. Bichkar

2012. Article: Issues in optimization of decision tree learning: A survey. *International Journal of Applied Information Systems*, 3(5):13–29. Published by Foundation of Computer Science, New York, USA.

Pellegrini, M., E. M. Marcotte, M. J. Thompson, D. Eisenberg, and T. O. Yeates

1999. Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proceedings of the National Academy of Sciences of the United States of America*, 96(8):4285–8.

Pellegrini, M., M. Thompson, J. Fierro, and P. Bowers

2001. Computational method to assign microbial genes to pathways. In *Journal of Cellular Biochemistry*, volume 84 SUPPL. 37, Pp. 106–109.

Penel, S., A.-M. Arigon, J.-F. Dufayard, A.-S. Sertier, V. Daubin, L. Duret, M. Gouy, and G. Perriere

2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*, 10(Suppl 6):S3.

Pereira, C., J. Azé, A. Denise, C. Drevet, C. Froidevaux, P. Silar, and O. Lespinet

2013. Comparative analysis of phylogenetic profiles for the enzymatic characterization of fungal group. In *JOBIM 2013*, Toulouse, France.

Pereira, C., A. Denise, and O. Lespinet

2014. A meta-approach for improving the prediction and the functional annotation of ortholog groups. *BMC Genomics*, 15(Suppl 6):S16.

Perlman, Z. E., M. D. Slack, Y. Feng, T. J. Mitchison, L. F. Wu, and S. J. Altschuler

2004. Multidimensional drug profiling by automated microscopy. *Science*, 306(5699):1194–1198.

Perner, P.

2001. Improving the accuracy of decision tree induction by feature preselection. *Applied Artificial Intelligence*, (15 (8)):747–760.

Pfeiffer, T., O. S. Soyer, and S. Bonhoeffer

2005. The evolution of connectivity in metabolic networks. *PLoS Biol*, 3(7):e228.

Poptsova, M. and J. P. Gogarten

2007. BranchClust: a phylogenetic algorithm for selecting gene families. *BMC Bioinformatics*, 8(1):120.

Powell, S., K. Forslund, D. Szklarczyk, K. Trachana, A. Roth, J. Huerta-Cepas, T. Gabaldón, T. Rattei, C. Creevey, M. Kuhn, L. J. Jensen, C. von Mering, and P. Bork

2014. eggnog v4.0: nested orthology inference across 3686 organisms. *Nucleic Acids Research*, 42(D1):D231–D239.

Pryszcz, L., J. Huerta-Cepas, and T. Gabaldón

2011a. Metaphors: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Res.*, 39(5):e32.

Pryszcz, L. P., J. Huerta-Cepas, and T. Gabaldón

2011b. Metaphors: orthology and paralogy predictions from multiple phylogenetic evidence using a consistency-based confidence score. *Nucleic Acids Research*, 39(5):e32.

Quinlan, J. R.

1993. C4.5: Programs for machine learning. *San Mateo: Morgan Kaufmann.*

Reid, A. J., J. a. G. Ranea, A. B. Clegg, and C. a. Orengo

2010. CODA: accurate detection of functional associations between proteins in eukaryotic genomes using domain fusion. *PloS one*, 5(6):e10908.

Remm, M., C. E. Storm, and E. L. Sonnhammer

2001. Automatic clustering of orthologs and in-paralogs from pairwise species comparisons. *Journal of Molecular Biology*, 314(5):1041 – 1052.

Rison, S. C. and J. M. Thornton

2002. Pathway evolution, structurally speaking. *Current Opinion in Structural Biology*, 12(3):374 – 382.

Rogozin, I. B., D. Managadze, S. A. Shabalina, and E. V. Koonin

2014. Gene family level comparative analysis of gene expression in mammals validates the ortholog conjecture. *Genome Biology and Evolution*, 6(4):754–762.

Rohlfs, M., M. Albert, N. P. Keller, and F. Kempken

2007. Secondary chemicals protect mould from fungivory. *Biology Letters*, 3(5):523–525.

Rosenblatt, F.

1957. The perceptron–a perceiving and recognizing automaton. *Report 85-460-1.*

Saeys, Y., I. Inza, and P. Larrañaga

2007. A review of feature selection techniques in bioinformatics. *Bioinformatics*, 23(19):2507–2517.

Salichos, L. and A. Rokas

2011. Evaluating ortholog prediction algorithms in a yeast model clade. *PloS one*, 6(4):e18755.

Salzberg, S. L., A. L. Delcher, S. Kasif, and O. White

1998. Microbial gene identification using interpolated markov models. *Nucleic Acids Research*, 26(2):544–548.

Schmidt, S., S. Sunyaev, P. Bork, and T. Dandekar

2003. Metabolites: a helping hand for pathway evolution? *Trends in Biochemical Sciences*, 28(6):336 – 341.

Schreiber, F. and E. L. Sonnhammer

2013. Hieranoid: Hierarchical orthology inference. *Journal of Molecular Biology*, 425(11):2072 – 2081.

Shannon, P., A. Markiel, O. Ozier, N. S. Baliga, J. T. Wang, D. Ramage, N. Amin, B. Schwikowski, and T. Ideker

2003. Cytoscape: A software environment for integrated models of biomolecular interaction networks. *Genome Research*, 13(11):2498–2504.

Shoemaker, B. A. and A. R. Panchenko

2007. Deciphering protein–protein interactions. part ii. computational methods to predict protein and domain interaction partners. *PLoS Comput Biol*, 3(4):e43.

Slonim, N., O. Elemento, and S. Tavazoie

2006. Ab initio genotype-phenotype association reveals intrinsic modularity in genetic networks. *Molecular systems biology*, 2:2006.0005.

Smith, D. J., M. K. Burnham, J. H. Bull, J. E. Hodgson, P. Ward, J. M.and Browne, J. Brown, B. Barton, A. J. Earl, and G. Turner

1990. Beta-lactam antibiotic biosynthetic genes have been conserved in clusters in prokaryotes and eukaryotes. *The EMBO journal*, 9:741–747.

Sonnhammer, E. L. and G. Östlund

2014. Inparanoid 8: orthology analysis between 273 proteomes, mostly eukaryotic. *Nucleic Acids Research*.

Stajich, J., F. Dietrich, and S. Roy

2007. Comparative genomic analysis of fungal genomes reveals intron-rich ancestors. *Genome Biology*, 8(10):R223.

Stobbe, M. D., G. A. Jansen, P. D. Moerland, and A. H. van Kampen

2014. Knowledge representation in metabolic pathway databases. *Briefings in Bioinformatics*, 15(3):455–470.

Storm, C. E. V. and E. L. L. Sonnhammer

2002. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. *Bioinformatics (Oxford, England)*, 18(1):92–9.

Tabach, Y., A. C. Billi, G. D. Hayes, M. A. Newman, O. Zuk, H. Gabel, R. Kamath, K. Yacoby, B. Chapman, S. M. Garcia, M. Borowsky, J. K. Kim, and G. Ruvkun

2013. Identification of small RNA pathway genes using patterns of phylogenetic conservation and divergence. *Nature*, 493(7434):694–698.

Takuji, Y. and B. Peer

2009. Evolution of biomolecular networks: lessons from metabolic and protein interactions. *Nature reviews. Molecular cell biology 10*, 11:791–803.

Tatusov, R. L., E. V. Koonin, and D. J. Lipman

1997. A genomic perspective on protein families. *Science*, 278(5338):631–637.

Teichmann, S. A., S. C. Rison, J. M. Thornton, M. Riley, J. Gough, and C. Chothia

2001. The evolution and structural anatomy of the small molecule metabolic pathways in escherichia coli. *Journal of Molecular Biology*, 311(4):693 – 708.

Thomas, P. D., M. J. Campbell, A. Kejariwal, H. Mi, B. Karlak, R. Daverman, K. Diemer, A. Muruganujan, and A. Narechania

2003. Panther: A library of protein families and subfamilies indexed by function. *Genome Research*, 13(9):2129–2141.

Trachana, K., K. Forslund, T. Larsson, S. Powell, T. Doerks, C. von Mering, and P. Bork

2014. A phylogeny-based benchmarking test for orthology inference reveals the limitations of function-based validation. *PLoS ONE*, 9(11):e111122.

Trachana, K., T. A. Larsson, S. Powell, W.-H. Chen, T. Doerks, J. Muller, and P. Bork

2011. Orthology prediction methods: A quality assessment using curated protein families. *BioEssays*, 33(10):769–780.

Treangen, T. J. and E. P. C. Rocha

2011. Horizontal transfer, not duplication, drives the expansion of protein families in prokaryotes. *PLoS Genet*, 7(1):e1001284.

van der Heijden, R., B. Snel, V. van Noort, and M. Huynen

2007. Orthology prediction at scalable resolution by phylogenetic tree analysis. *BMC Bioinformatics*, 8(1):83.

Vens, C., J. Struyf, L. Schietgat, S. Džeroski, and H. Blockeel

2008. Decision trees for hierarchical multi-label classification. *Machine Learning*, 73(2):185–214.

Vert, J.-P.

2002. A tree kernel to analyse phylogenetic profiles. *Bioinformatics (Oxford, England)*, 18 Suppl 1:S276–84.

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney

2009a. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335.

Vilella, A. J., J. Severin, A. Ureta-Vidal, L. Heng, R. Durbin, and E. Birney

2009b. Ensemblcompara genetrees: Complete, duplication-aware phylogenetic trees in vertebrates. *Genome Research*, 19(2):327–335.

Vitulo, N., A. Vezzi, C. Romualdi, S. Campanaro, and G. Valle

2007. A global gene evolution analysis on Vibrionaceae family using phylogenetic profile. *BMC bioinformatics*, 8 Suppl 1:S23.

von Mering, C., E. M. Zdobnov, S. Tsoka, F. D. Ciccarelli, J. B. Pereira-Leal, C. A. Ouzounis, and P. Bork

2003. Genome evolution reveals biochemical networks and functional modules. *Proceedings of the National Academy of Sciences*, 100(26):15428–15433.

Wall, D. P., H. B. Fraser, and A. E. Hirsh

2003. Detecting putative orthologs. *Bioinformatics*, 19(13):1710–1711.

Waterhouse, R. M., F. Tegenfeldt, J. Li, E. M. Zdobnov, and E. V. Kriventseva

2013. Orthodb: a hierarchical catalog of animal, fungal and bacterial orthologs. *Nucleic Acids Research*, 41(D1):D358–D365.

Webb, E.

1992. Enzyme nomenclature 1992: Recommendations of the nomenclature committee of the international union of biochemistry and molecular biology. *Academic Press, San Diego, CA*.

Wolf, Y. I. and E. V. Koonin

2012. A tight link between orthologs and bidirectional best hits in bacterial and archaeal genomes. *Genome Biology and Evolution*, 4(12):1286–1294.

Wu, J., Z. Hu, and C. DeLisi

2006. Gene annotation and network inference by phylogenetic profiling. *BMC Bioinformatics*, 7(1):80.

Yanai, I. and C. DeLisi

2002. The society of genes: networks of functional links between genes from comparative genomics. *Genome biology*, 3(11):research0064.

Yang, Y. and J. O. Pedersen

1997. A comparative study on feature selection in text categorization. In *Proceedings of the Fourteenth International Conference on Machine Learning*, ICML '97, Pp. 412–420, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Zhang, J. D. and S. Wiemann

2009. Kegggraph: a graph approach to kegg pathway in r and bioconductor. *Bioinformatics*, 25(11):1470–1471.

Zmasek, C. and S. Eddy

2002. RIO: Analyzing proteomes by automated phylogenomics using resampled inference of orthologs. *BMC Bioinformatics*, 3(1):14.