Inférence fonctionnelle et prédiction de voies métaboliques.Application à la bactérie fixatrice d'azote Sinorhizobium meliloti.

Clotilde Claudel

THESE

Présentée devant

L'UNIVERSITE PAUL SABATIER TOULOUSE III

en vue de l'obtention du

DOCTORAT DE L'UNIVERSITE

Discipline: Bio-informatique

Ecole Doctorale Biologie Santé Biotechnologies

par

Clotilde Claudel-Renard

Inférence fonctionnelle et prédiction de voies métaboliques. Application à la bactérie fixatrice d'azote *Sinorhizobium meliloti*.

Date de soutenance : 19 décembre 2003

Composition du jury :

Dr. Claude Chevalet	Directeur de thèse
Dr. Daniel Kahn	Co-directeur de thèse
Dr. Claudine Médigue	Rapporteur
Pr. Jean-Loup Risler	Rapporteur
Pr. Gwenaele Fichant	Examinateur
Dr. Alain Viari	Examinateur

Laboratoire de Génétique Cellulaire, Institut National de la Recherche Agronomique, Centre de Toulouse, BP 27, Chemin de Borde-Rouge, 31326 Castanet-Tolosan cedex.

Résumé

Inférence fonctionnelle et prédiction de voies métaboliques

Application à la bactérie fixatrice d'azote *Sinorhizobium meliloti*.

Des génomes entiers de bactéries sont séquencés en nombre croissant. Parallèlement sont mis en place des programmes d'analyse systématique de l'expression des gènes et des protéines dans différentes conditions. La compréhension du fonctionnement d'un organisme nécessite une annotation des fonctions des gènes et l'intégration de ces données dans des schémas fonctionnels. Les voies métaboliques constituent une classe de fonctions permettant d'aborder ce problème d'intégration, elles sont bien répertoriées chez de nombreux organismes et sont accessibles à l'expérimentation.

Dans un premier temps, nous avons développé une méthode automatique de prédiction de fonction spécifique des enzymes. Cette méthode nommée PRIAM (PRofils pour l'Identification Automatique du Métabolisme) repose sur la nomenclature des enzymes et sur la construction automatique d'un jeu de profils spécifiques des fonctions enzymatiques. Puis, cette méthode permet d'identifier les enzymes dans un génome complet et de visualiser les résultats obtenus sur les graphes des voies métaboliques de la base de données KEGG.

Dans un second temps, cette méthode a été appliquée sur le génome de la bactérie fixatrice d'azote *Sinorhizobium meliloti* et nous a permis l'analyse des voies métaboliques spécifiques de cet organisme symbiote.

Remerciements

Ce travail a été réalisé au laboratoire de Génétique Cellulaire de l'INRA de Toulouse sous la direction de Messieurs Claude Chevalet et Daniel Kahn. Je tiens à leur exprimer mes remerciements pour la confiance qu'ils m'ont accordée, pour les conseils qu'ils m'ont prodigués et pour la méthode et la rigueur qu'ils m'ont enseignées.

Je remercie le Dr Claudine Médigue et le Professeur Jean-Loup Risler d'avoir accepté de rapporter le travail présenté dans ce manuscrit, ainsi que le Professeur Gwenaele Fichant et le Dr Alain Viari d'avoir accepté d'être membre du jury.

Je remercie le laboratoire de Génétique Cellulaire pour son amical accueil et tout particulièrement Thomas Faraut et Mireille Morisson pour leurs conseils tout au long de ma thèse.

Je remercie également l'équipe de Daniel Kahn du laboratoire des Interactions Plantes Microorganismes : Jérome Gouzy, Emmanuel Courcelle, Sébastien Carrere pour leur aide dans la construction de PRIAM, Hélène Berges pour son soutien et tout particulièrement Florence Servant pour ses conseils.

Je remercie le laboratoire de Biométrie et Intelligence Artificielle, plus particulièrement Monique Fallière pour son dynamisme et Patricia Thébault pour son soutien de tous les jours.

Enfin je dédie ce travail à mes parents, ma famille, Frédéric, Tristan et tous mes amis pour leur soutien tout au long de cette thèse.

TABLE DES MATIERES

LISTE DES FIGURES

LISTE DES TABLEAUX

ABREVIATIONS

ADN : Acide DésoxyRibonucléique

ARN : Acide RiboNucléique

ARNm: ARN messager

ARNr: ARN ribosomique

ATP: Adenosine TriPhosphate

BAC: Bacterial Artificial Chromosome

BLAST: Basic Local Alignement Search Tool

BLOSUM: BLOcks Sustitution Matrix

CGI: Common Gateway Interface

EST: Expressed Sequence Tag

FAD: Flavin Adenine Dinucleotide

GSS: Genome Survey Sequence

GTP: Guanosyl TriPhosphate

HMM: Hidden Markov Models

HSP: High-scoring Segment Pair

HTC: High Throughput cDNA

HTG: High Throughput Genomic

HTML: Hyper Text Markup Language

ORF: Open Reading Frame

PAM: Point Accepted Mutation

PCR: Polymerase Chain Reaction

PSI-BLAST: Profile Searches Iterated Basic Local Alignment Search Tool

PSSM: Position-Specific Scoring Matrices

STS: Sequence-Tagged Sites

WWW: World Wide Web

AVANT-PROPOS

Cette année 2003 correspond au cinquantième anniversaire de la découverte du modèle moléculaire de la double hélice de l'ADN par Watson et Crick. Cette découverte a permis de formuler le dogme central de la biologie moléculaire selon lequel l'information génétique est portée par les acides nucléiques et exprimée par les protéines. Cette connaissance de la structure du support de l'information génétique a été la première étape pour l'invention du séquençage de l'ADN par W. Gilbert et F. Sanger dans les années 1970. L'ère de la génomique commençait, c'est-à-dire l'identification de l'ensemble des gènes constituant le génome d'un organisme. En 1995, le premier génome d'un organisme vivant a été entièrement séquencé (*Haemophilus influenzae*). A ce jour, plus de 100 génomes de différentes espèces ont été séquencés dont de nombreuses bactéries, virus, organismes eucaryotes comme la levure en 1997, le nématode en 1998 et enfin le génome humain. La génomique a ainsi amené une masse énorme de données qu'il faut analyser pour en extraire la signification physiologique.

En cette période nommée de « post-génomique » ou « après-génome », la bioinformatique est une discipline qui a émergé pour gérer, extraire des connaissances, analyser et visualiser les données issues de ces programmes de séquençage. Une des grandes problématiques de la biologie et de la bioinformatique aujourd'hui est de prédire la ou les fonctions des gènes identifiés dans les génomes. Parmi les fonctions attribuées aux gènes, on trouve celles des enzymes. Elles participent à la plupart des processus cellulaires : réplication, transcription des gènes mais aussi tout le métabolisme. Une nouvelle approche de la bioinformatique est donc d'intégrer les fonctions des gènes dans des réseaux de relations : réseaux de régulation de gènes, transduction d'un signal externe vers le noyau, voies métaboliques, afin de mieux comprendre la contribution des fonctions des gènes au phénotype de la cellule puis de l'organisme.

L'objectif de cette thèse a été de créer un outil générique de prédiction des fonctions enzymatiques et d'intégrer les résultats dans les voies métaboliques d'un organisme, à partir de son génome complet. Le premier chapitre défini différents concepts, tel que la ou les fonctions d'un gène, l'homologie puis présente différents outils et méthodes pour prédire les fonctions des gènes et pour intégrer les fonctions comme les enzymes dans des voies

métaboliques. Dans cette thèse, nous avons développé une méthode de prédiction spécifique des fonctions enzymatiques et intégré

les résultats obtenus dans des graphes des voies métaboliques. Le deuxième chapitre présente les différentes étapes de construction et de validation de la méthode nommée PRIAM (PRofils pour l'Identification Automatique du Métabolisme). Nous avons ensuite appliqué l'outil PRIAM sur le génome complet, récemment séquencé, d'une bactérie d'intérêt agronomique : *Sinorhizobium meliloti*. Le troisième chapitre présente cet organisme et les résultats de l'analyse de ses voies métaboliques obtenus avec PRIAM.

I
INTRODUCTION

L'introduction de cette thèse comprend cinq parties. La première partie définit le concept de fonction d'un gène qui est une des problématiques actuelles et futures la plus importante en bioinformatique. La deuxième partie présente ensuite le concept d'homologie utilisé pour la prédiction de fonction. La troisième partie, décrit les différents outils de prédiction par similarité des fonctions à l'échelle des gènes. La quatrième partie présente les différentes méthodologies de prédiction de fonction à l'échelle des génomes. Enfin, la cinquième partie expose les méthodes d'intégration des fonctions pour le cas particulier des fonctions enzymatiques dans des voies métaboliques.

1 Fonctions des gènes

Une des grandes problématiques de la bioinformatique en cette phase post-génomique est de prédire la fonction des gènes détectés dans les génomes des nombreux organismes séquencés (plus de 100 génomes complets dans la base de données GOLD en mars 2003). La connaissance des fonctions des gènes permet de mieux comprendre le fonctionnement des organismes, si l'on intègre ces fonctions dans leur réseau d'action qui peut être une voie métabolique, une voie de transduction du signal, un réseau de régulation de l'expression des gènes,… La définition de la fonction d'un gène est complexe. Il existe en réalité plusieurs niveaux pour définir la fonction de gène.

1.1 Définitions de la fonction d'un gène

Le dictionnaire (Robert, 1982) définit la fonction comme : "une action, un rôle caractéristique d'un élément, d'un organe, dans un ensemble". La notion de fonction biologique d'un gène peut être décrite de différentes manières. En effet, on peut définir la fonction d'un gène à différentes échelles : celle de la molécule, de la cellule, de l'organisme ou de la population. Ainsi, le gène codant pour la chaîne bêta de l'hémoglobine participe :

- à l'échelle moléculaire, à la fixation de l'oxygène,
- à l'échelle de la cellule à la libération de l'oxygène pour la production d'énergie,
- à l'échelle de l'organisme à la circulation de l'oxygène dans toutes les cellules.
- Enfin, à l'échelle de la population, la mutation d'un allèle du gène de la chaîne bêta de l'hémoglobine chez l'homme peut avoir plusieurs conséquences. Cela peut être la cause d'une grave maladie, la thalassémie, mais aussi être une protection contre le parasite du paludisme.

Par ailleurs, un gène et la ou les protéines qu'il code, ont des caractéristiques biologiques qui leur permettent d'avoir plusieurs fonctions. Un gène a la caractéristique de pouvoir donner plusieurs transcrits. En effet, le mécanisme d'épissage alternatif dans certains organismes permet d'obtenir plusieurs transcrits à partir d'un gène. Le nombre de gènes dans le génome humain était estimé entre 50 à 100.000 gènes du fait du nombre important de protéines exprimées. Or, à ce jour, la base de données ENSEMBL estime ce nombre à seulement

32.284 gènes (1). Un gène exprimant une ou plusieurs protéines différentes peut donc avoir plusieurs fonctions biologiques.

D'autre part, certaines protéines, comme les enzymes multi-fonctionnelles, sont connues pour avoir plusieurs fonctions biologiques. Une protéine peut avoir des fonctions différentes selon le contexte cellulaire ou l'environnement protéique. Par exemple, la protéine Scute de la drosophile est un facteur de transcription impliqué dans la détermination du sexe, mais aussi dans la compétence des cellules à devenir des cellules neuronales. Ainsi, définir la ou les fonctions d'une protéine consiste à déterminer les interactions possibles avec d'autres molécules et à identifier leurs rôles à différents niveaux.

1.2 Le cas de la fonction enzymatique

De nombreux types de fonctions moléculaires peuvent être attribués à une protéine : facteur de transcription, récepteur, protéine de structure (microtubule, myosine,…). Les fonctions biologiques souvent attribuées sont des fonctions enzymatiques. En effet, à ce jour plusieurs milliers d'enzymes sont connues, plus de 3.500 enzymes différentes sont répertoriées dans la nomenclature de l'IUBMB (International Union of Biochemistry and Molecular Biology). Une enzyme est une protéine active qui permet la transformation d'une substance naturelle en une autre substance. Les enzymes (protéines) et les ribozymes (ARN) sont des molécules qui permettent la catalyse d'une réaction chimique spécifique et qui peuvent être régulées par différents facteurs (pH, température). L'action d'une enzyme fut mise en évidence pour la première fois en 1830. Elle fut isolée deux ans plus tard sous le nom de diastase, car elle séparait les dextrines et le maltose de l'amidon (du grec diastasis, séparation). Le terme d'enzyme fut créé par W. Kühne en 1878, à partir des termes grecs « zumé » (levain) et « en » (dedans), pour remplacer les termes de ferment et de diastase trop restrictifs. Les enzymes sont des acteurs essentiels de la plupart des processus biologiques survenant dans les cellules. De par ce rôle, elles sont des cibles privilégiées en industrie et en recherche biomédicale. En industrie, elles sont utilisées dans les productions de lessives, de jus de fruits, dans la production de sucre à partir d'amidon, dans le traitement des peaux, des fibres textiles, dans le traitement des déchets, etc. En recherche biomédicale, ce sont des cibles de maladies métaboliques héréditaires par exemple mais aussi des outils pour la recherche en biologie moléculaire (cf. l'ADN polymérase pour la technique de PCR, les enzymes de restriction, etc...).

1.2.1 La classification des enzymes

Dès les années 1960, de nombreuses enzymes sont déjà caractérisées biologiquement. Une nomenclature enzymatique devient nécessaire pour classer les enzymes selon la réaction catalysée ainsi que pour leur donner un nom systématique. En effet, il existe des enzymes identiques avec des noms différents et réciproquement des enzymes différentes avec le même nom. Les noms communs donnés n'indiquant pas toujours la réaction catalysée.

Une nomenclature enzymatique a été créée en 1961 par l'IUBMB (International Union of Biochemistry and Molecular Biology) (2), celle-ci comporte 6 catégories : les oxydoréductases, les transférases, les hydrolases, les lyases, les isomérases, et les ligases ou synthétases. Un code à 4 champs appelé numéro EC (Enzyme Commission) permet de décrire chaque enzyme. Le premier champ allant de 1 à 6 correspond à l'une des 6 catégories. Le deuxième champ correspond à une sous-classe et donne une information sur le type de composé ou de groupe impliqué dans la réaction. Le troisième champ donne une information sur le type de réaction. Enfin, le quatrième champ correspond à un numéro de série servant à identifier une enzyme individuelle. Par exemple, le numéro EC 1.1.1.1 de l'activité enzymatique alcool déshydrogénase correspond à la classe 1 des oxydoréductases. La sous-classe 1.1 signifie que l'enzyme agit sur le groupe CH-OH. Le niveau 1.1.1 signifie que l'accepteur est soit le NAD+ soit le NADP+. Enfin, le plus bas niveau de la hiérarchie, identifie la réaction particulière de l'alcool déshydrogénase avec comme accepteur le NAD+ alors que le numéro EC 1.1.1.2 correspond à l'activité alcool déshydrogénase avec comme accepteur le NADP+. L'autre apport de la création de cette nomenclature est la dénomination de chaque enzyme par un nom systématique. Ce nom systématique donne les informations sur la structure du substrat et sur la nature de la réaction. Par exemple la pyruvate déshydrogénase, transforme le pyruvate en acétyl-CoA par la perte d'un hydrogène récupéré par le cofacteur NADP.

La liste d'enzymes de l'IUBMB comportait 712 entrées en 1961, elle en comporte aujourd'hui plus de 3.500. Plusieurs bases de données de connaissances sont structurées sur la nomenclature EC, comme les bases de données ENZYME (3), LIGAND (4) et BRENDA (5). Après quelques années, une enzyme peut être considérée comme mal classifiée, l'entrée EC est alors enlevée ou transférée. Les numéros EC ne sont alors pas réutilisés pour éviter des

confusions. La relation « à un polypeptide correspond un et un seul numéro EC» n'est pas toujours vérifiée. Pour un même polypeptide, il peut exister plusieurs numéros EC. C'est le cas des multi-enzymes qui possèdent plusieurs régions protéiques avec des activités enzymatiques différentes. De même, une enzyme avec un seul numéro EC peut correspondre à plusieurs polypeptides. C'est le cas des enzymes oligomériques qui sont des regroupements de plusieurs sous-unités protéiques.

1.2.2 Les multi-enzymes

Une multi-enzyme est une protéine avec au moins deux activités enzymatiques sur des régions de séquences protéiques différentes d'un même polypeptide. De nombreuses multi-enzymes sont connues chez les eucaryotes, un peu moins chez les procaryotes. Un exemple connu de multi-enzyme est le gène GART qui est une tri-enzyme chez l'homme, la souris et le poulet. Ce gène possède les trois activités suivantes :

- GARS (Phosphoribosylamine-glycine ligase, EC 6.3.4.14),
- AIRS (Phosphoribosylformylglycinamidine cyclo-ligase, EC 6.3.3.1) et
- GART (Phosphoribosylglycinamide formyltransferase, EC 2.1.2.2).

Ces trois enzymes sont impliquées dans la voie de synthèse des purines. Chez *E. coli* et *B. subtilis*, ces trois activités enzymatiques sont codées par trois gènes différents mais qui sont homologues aux trois régions du gène GART. Cette observation est à l'origine d'études sur ce type de fusion de gènes afin de comprendre leur origine. En effet, des études de la structure du gène GART ont mis en évidence un signal de polyadénylation non fonctionnel dans un intron (6). Ce signal pourrait être un fossile moléculaire d'un évènement de recombinaison de plusieurs gènes à l'origine de ce multi-enzyme.

1.2.3 Les enzymes oligomériques et les complexes enzymatiques

Les enzymes constituées de plusieurs polypeptides nommées sous-unités protéiques forment des enzymes oligomériques. Elles comportent en général au maximum une dizaine de sous-unités. Ainsi, par exemple l'ornithine carbamoyltransférase chez *Pseudomonas aeruginosa* est un dodécamère.

Par ailleurs, on appelle complexe enzymatique, un groupement de protéines réalisant une activité enzymatique. Il peut comporter des dizaines d'éléments, chaque élément pouvant avoir une activité enzymatique indépendante. Toutes les protéines d'un complexe

enzymatique ne possèdent pas une activité enzymatique mais elles contribuent à une certaine structure protéique favorable à la réaction enzymatique. Par exemple la NADH déshydrogénase humaine est composée de 40 sous-unités différentes.

2 L'homologie

Le concept d'homologie est utilisé en analyse de séquence comme principe pour réaliser de la prédiction de fonction. Deux gènes sont dits **homologues** s'ils divergent d'un ancêtre commun (7). En effet, si l'on peut prouver que deux séquences sont homologues c'est-à-dire issues d'une séquence ancestrale commune, on déduit en principe que leur fonction est similaire à la fonction de cet ancêtre. En réalité, d'une part deux séquences homologues n'ont pas toujours conservé la même fonction que leur ancêtre et d'autre part, on met en évidence une similarité de séquences qui ne correspond pas nécessairement à une relation d'homologie.

2.1 Homologues, orthologues et paralogues

Des gènes homologues sont issus d'un ancêtre commun par deux types de mécanismes observés au cours de l'évolution (figure 1) : soit les gènes homologues de deux espèces différentes sont issus d'un évènement de spéciation d'un ancêtre commun (fourche dans l'arbre), ils sont alors qualifiés de gènes **orthologues** et possèdent en principe une fonction biologique similaire ; soit ils sont issus d'un événement de duplication d'un ancêtre commun (ligne horizontale dans l'arbre), ils possèdent alors en principe des caractéristiques fonctionnelles similaires et sont nommés gènes **paralogues**.

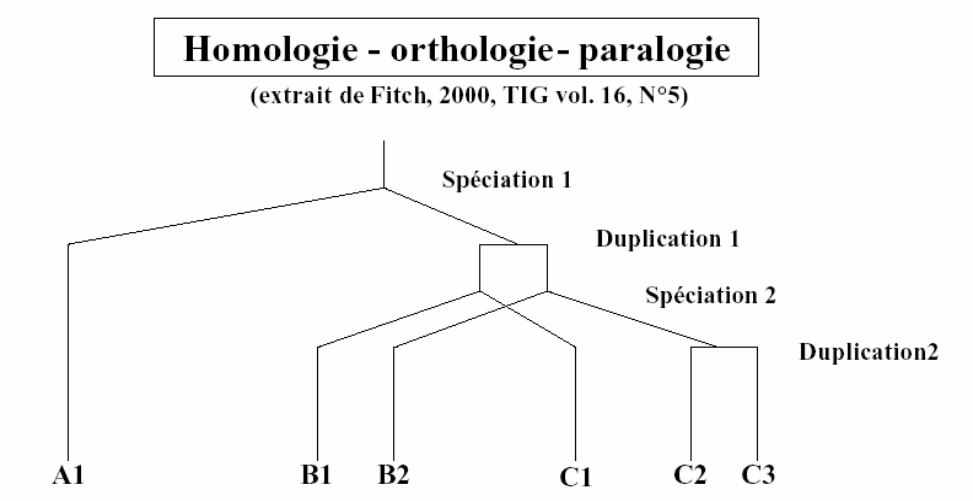


Figure 1: Représentation schématique de la distinction des orthologues et des paralogues.

L'homologie de gènes est une notion qualitative que l'on ne peut pas mesurer. L'homologie a la caractéristique d'être transitive. Si l'on considère une homologie sur des séquences entières, on peut déduire que deux gènes A et B sont homologues si les gènes A-C et les gènes B-C sont homologues. Mais un paramètre pouvant donner une indication sur l'homologie de deux séquences est de mesurer la ressemblance de deux séquences. Cette ressemblance entre deux séquences est nommée similitude, elle s'exprime en pourcentage d'identité entre deux séquences. Cela consiste à calculer le pourcentage de bases nucléiques ou d'acides aminés identiques entre deux séquences alignées de façon optimale.

Le problème qui se pose dans l'utilisation du concept d'homologie est qu'il faut être capable de discriminer les séquences orthologues des séquences paralogues afin d'être dans la meilleure situation pour inférer une fonction à un gène.

En effet, les gènes paralogues, séquences homologues issues d'un évènement de duplication d'un gène dans une même espèce, ne possèdent pas toujours la même fonction. De nombreuses familles de séquences similaires sont connues dans une même espèce. Certains groupes de gènes codant pour des enzymes réalisent le même type de réaction mais ne possèdent pas la même spécificité de substrats. Comme par exemple les kinases chez *E. coli* (8). La duplication des gènes et leur divergence au cours de l'évolution serait un des mécanismes à l'origine de la diversité des fonctions biologiques (9, 10).

2.2 Similarité de fonctions en l'absence d'homologie: les analogues

Dès 1943, des enzymes possédant la même activité fructose 1-6 biphosphate aldolase mais avec des structures différentes, ont été découvertes. Il existe en effet deux classes d'aldolases qui ne descendraient pas d'un ancêtre commun. Ces deux classes auraient évolué indépendamment et convergé vers une même fonction biochimique. Ces enzymes sont qualifiées d'analogues (11). Galperin *et al.* ont identifié 105 cas d'enzymes analogues sans similarité de séquence. Sur ces 105 cas, 34 présentent des architectures tridimensionnelles effectivement différentes. Des cas d'enzymes analogues observés correspondent à un type de séquences spécifique des organismes eucaryotes et un type de séquences spécifique des

bactéries. Les enzymes analogues sont des cibles prometteuses pour le développement de médicaments anti-bactériens.

2.3 Décomposition en domaines des protéines

Les séquences protéiques ont la caractéristique de pouvoir être décomposées en domaines. Le domaine structural est une unité de repliement autonome. Au niveau de la structure primaire, le domaine ne correspond pas toujours à une région contiguë de la séquence protéique. Une famille de domaines, correspondant par exemple à un domaine de liaison à un cofacteur, peut appartenir à différentes protéines de fonctions différentes. La modularité des protéines rend donc complexe la recherche d'homologies par similarité de séquences. On parle alors d'**homologie partielle**. Afin d'identifier des similitudes locales, il existe des algorithmes tel que l'algorithme de Smith et Waterman (12) et des heuristiques tel que BLAST (13). Par ailleurs il existe des bases de données manuelles ou automatiques tel que ProDom (14), Pfam (15), décomposant les protéines connues en différents domaines et permettant de mettre en évidence des homologies partielles entre les protéines.

3 Les outils de prédiction de fonction des gènes

L'inférence de la fonction d'une séquence inconnue est réalisée par la recherche d'une forte similarité de cette séquence avec d'autres séquences. Un fort degré de similarité permet de supposer des relations d'homologies et de proposer une similarité de fonction. Cette prédiction de la fonction d'un gène nécessite deux types d'outils : des algorithmes de comparaison de séquences, qui sont implémentés dans différents programmes et des catalogues de séquences qui sont présentés sous forme de bases de données biologiques. La comparaison de séquences est le plus souvent réalisée sur les séquences protéiques. En effet, la dégénérescence du code génétique combinée avec les possibilités de substitution des acides aminés, rend la détection d'homologie des séquences codantes plus délicate au niveau des séquences nucléiques. Il est donc préférable d'opérer cette détection au niveau des séquences protéiques. C'est pourquoi nous présentons plus spécifiquement dans cette partie les outils de comparaison des séquences protéiques et les bases de données relatives aux protéines.

3.1 La comparaison de séquences protéiques

Différents types d'alignements de séquences existent reposant tous sur un même principe. Il existe tout d'abord des alignements de paires de séquences réalisés soit de façon globale c'est-à-dire en alignant deux séquences sur toute leur longueur, soit de façon locale sur des segments de séquences. D'autre part, il est possible d'effectuer des alignements locaux d'une séquence avec une banque de séquences. Enfin, il est possible de réaliser des alignements multiples de séquences similaires afin de caractériser une famille. Ces alignements multiples peuvent être décrits par des profils et peuvent aussi être alignés avec une banque de séquences et servir à la décomposition en régions similaires des protéines.

3.1.1 Les alignements de séquences

3.1.1.1 Principe de l'alignement

La comparaison de deux séquences nécessite de les aligner de façon optimale c'est-à-dire de rechercher un segment commun avec un maximum de similarité entre les acides

aminés. La similarité entre les acides aminés peut-être mesurée de différentes manières. La similarité est représentée dans des matrices de scores d'équivalence, de taille 20 fois 20 acides aminés. La similarité de deux séquences est évaluée par le calcul d'un score de l'alignement. Ce score de comparaison est égal à la somme des scores élémentaires lus dans la matrice des scores. L'alignement optimal entre deux séquences peut-être amélioré en introduisant à certaines positions des insertions ou des délétions. La ressemblance entre les deux séquences alignées est considérée comme significative lorsque le score de comparaison est supérieur ou égal à un score seuil fixé.

- **Matrices de scores protéiques**

Différents types de matrices de scores existent. La matrice de scores la plus simple est la matrice d'identité. Un score de 1 est attribué lorsque les acides aminés sont identiques, un score de 0 lorsqu'ils sont différents.

Les matrices de scores, utilisées dans la construction d'arbres phylogénétiques sont basées sur le code génétique. Le score reflète le nombre minimum de changements de bases nécessaire pour passer d'un acide aminé à un autre.

D'autres matrices de scores sont basées sur les similarités de propriétés physico-chimiques des acides aminés car certaines de leurs propriétés ont un rôle important dans la fonction biologique de la protéine.

Deux types de matrices de scores les plus utilisées pour comparer les protéines et basées sur deux approches différentes sont : les matrices Dayhoff ou PAM (Point Accepted Mutation) (16) et les matrices BLOSUM (BLOcks Sustitution Matrix) (17). Elles reposent sur les substitutions observées dans un alignement de séquences protéiques.

Les matrices de Dayhoff sont des matrices de probabilité de mutation, ajustées à une certaine distance évolutive. Les fréquences de mutation pour une matrice 1 PAM c'est-à-dire 1% de mutations acceptées sont calculées à partir d'un alignement de séquences très similaires. A partir de la matrice 1 PAM, les valeurs des fréquences de mutation sont extrapolées pour une matrice 250 PAM. Une matrice 1 PAM sera utilisée pour comparer des séquences proches alors qu'une matrice 250 PAM permettra de comparer des séquences plus éloignées.

Les matrices BLOSUM reposent sur une approche différente. Elles sont construites à partir d'un alignement multiple de séquences plus éloignées sans insertions ni délétions et sans extrapolation. Les alignements de séquences représentent des groupes de séquences similaires

répertoriées dans la base de données BLOCKS (17). Les probabilités de substitution sont calculées à partir du comptage des paires d'acides aminés observés dans chaque block de la banque. Pour faire varier les fréquences observées, les séquences les plus similaires dans une famille sont regroupées et chaque groupe est compté comme une seule séquence. Ainsi, une matrice BLOSUM 62 est dérivée d'un alignement dans lequel les segments de séquences avec plus de 62% d'identité sont regroupés. De même que pour les matrices PAM, un ensemble de matrices BLOSUM a été calculé avec un seuil de regroupement allant de 30 à 90%. Ainsi, la matrice BLOSUM 30 sera utilisée pour la comparaison de séquences éloignées alors que la matrice BLOSUM 90 permettra de comparer des séquences proches.

Enfin, il existe des matrices dérivées d'alignement de structures tertiaires. Les alignements obtenus grâce à la connaissance de la structure peuvent permettre d'aligner des séquences plus distantes (18).

- **Insertions et délétions**

Afin d'améliorer l'alignement de séquences, certains programmes permettent l'introduction d'insertions ou de délétions à certaines positions de l'alignement. Les insertions sur une séquence correspondent à des délétions sur l'autre séquence alignée. Cette amélioration correspond à un évènement biologique qui survient au cours de l'évolution des génomes. Cet événement nommé **indel (INsertion-DElétion)** est intégré dans le calcul du score de comparaison par une pénalité. Cette pénalité est différente selon les programmes de comparaison. La pénalité P d'un indel est généralement calculée de la manière suivante :

$$P = x + yl$$

Avec l la longueur de l'indel,
x la pénalité fixe de l'indel et
y la pénalité d'extension de l'indel.
On peut faire varier les deux paramètres x et y. Par exemple, en augmentant la valeur de x, on défavorise les insertions et en diminuant la valeur de y, on favorise de longues insertions.

- **Evaluation statistique de l'alignement**

Il est important de pouvoir estimer si l'alignement de deux séquences obtenu a une signification biologique ou si c'est un évènement qui peut être obtenu par hasard. Pour cela, différents outils de statistiques peuvent être utilisés.

Le critère statistique le plus simple est le pourcentage d'identité entre deux segments de séquences.

Un deuxième critère très souvent utilisé est le Z score qui rend compte de l'éloignement du score de l'alignement par rapport à une distribution aléatoire. Pour calculer ce Z score, on construit une base de séquences aléatoires à partir de l'une des deux séquences de l'alignement, afin de maintenir la composition en acides aminés.

Ainsi, le Z score est égal à :

$$Z = (s - m) / e$$

Avec s le score de l'alignement considéré,

m la moyenne des scores aléatoires et

e l'écart type des scores aléatoires.

Les scores obtenus suivent une loi de distribution des valeurs extrêmes avec une queue de distribution pour les scores élevés (19). Le Z score peut être considéré comme significatif s'il est élevé, c'est-à-dire supérieur à 2 e (écart types).

D'autres méthodes statistiques ont été développées pour la comparaison d'une séquence avec une banque de séquences. La méthode de Karlin et Altschul (20) est la plus répandue. Elle permet de calculer la probabilité de trouver le plus haut score parmi toutes les paires de segments possibles entre deux séquences. Un score S' normalisé est calculé à partir du score de similarité S, afin de tenir compte de la séquence et de la banque.

$$S' = (\lambda S - \ln K) / \ln 2$$

Les paramètres λ et K sont estimés d'après la loi de distribution des valeurs extrêmes et tiennent compte de la composition en acides aminés des séquences. A partir de ce score normalisé est calculé la valeur E qui détermine le nombre attendu d'HSP (High-scoring Segment Pair) avec un score d'au moins S :

$$E = mn\, 2^{-S'} = Kmn\, e^{-\lambda S}$$

Par ailleurs, la valeur P détermine la probabilité qu'un alignement obtenu par hasard ait un score supérieur ou égal à un seuil.

$$P = 1 - e^{-E} = 1 - \exp(-Kmne^{-\lambda S})$$

Ainsi, si la valeur de E est faible, on peut conclure que l'alignement obtenu met en évidence une similitude significative entre les deux séquences observées.

3.1.1.2 Alignement global et local

La programmation dynamique permet de réduire le temps nécéssaire à la comparaison de deux séquences de longueurs équivalentes N à un temps proportionnel à N^2, tout en permettant des insertions ou des délétions dans l'alignement. La programmation dynamique repose sur le principe de ne conserver que certains évènements de comparaison selon certains critères. Deux types d'algorithmes ont été développés sur ce principe de programmation dynamique adapté au problème biologique de la comparaison de séquences: l'algorithme de Needleman et Wunsch (21) pour aligner globalement deux séquences supposées similaires sur toute leur longueur et l'algorithme de Smith et Waterman (12) pour les alignements locaux de séquences similaires sur certaines régions.

- **Algorithme de Needleman et Wunsch**

L'implémentation de l'algorithme de Needleman et Wunsch se fait en trois étapes :
Tout d'abord, une matrice de comparaison à deux dimensions de la taille des deux séquences à aligner est construite. La matrice est remplie avec les valeurs des scores élémentaires (se) des matrices de substitution.
La deuxième étape est la transformation de la matrice par addition des scores. Le score somme S (i,j) correspondant à l'acide aminé i de la séquence représentée à l'horizontale et à l'acide aminé j de la séquence représentées à la verticale devient :

$$S(i,j) = se(i,j) + \max \begin{vmatrix} S(i+1, ,j+1) \\ S(x, j+1) - P \\ S(i+1, y) - P \end{vmatrix} \quad \text{avec } i+1 < x \leq m \text{ et } j+1 < y \leq n$$

Avec P la pénalité donnée pour une insertion.
La dernière étape consiste à identifier le chemin décrivant l'alignement optimal. Il correspond au passage par les scores sommes les plus élevés en autorisant les mouvements horizontaux, verticaux et diagonaux. Le mouvement diagonal est cependant privilégié car il n'entraîne pas la création d'un indel.

- **Algorithme de Smith et Waterman.**

Le même principe est appliqué à la recherche de segments locaux similaires dans l'algorithme de Smith et Waterman. Cependant, n'importe quelle case de la matrice de comparaison initiale peut être considérée comme point de départ pour le calcul des scores sommes. Le système de scores pour transformer la matrice devient :

$$S(i,j) = \max \begin{cases} se\ (i,j) + S(i+1, j+1) \\ se\ (i,j) + \max S(x,j+1) - P \\ se\ (i,j) + \max S(i+1,y) - P \\ 0 \end{cases} \quad \begin{array}{l} \\ \text{avec } i+1 < x \leq m \\ \text{et} \quad j+1 < y \leq n \\ \\ \end{array}$$

Cette équation permet d'initialiser à zéro lorsqu'un score devient négatif et devient un nouveau point potentiel de départ de similarité.

3.1.1.3 Les heuristiques

L'utilisation de l'algorithme de Smith et Waterman devient très coûteux en temps lorsqu'il faut comparer une séquence aux centaines de milliers de séquences se trouvant dans les banques. C'est pourquoi des stratégies heuristiques consistant à identifier les régions les plus similaires ont été développées. Ces heuristiques sont implémentées dans les programmes FASTA (A fast approximation to Smith-Waterman) (22) et BLAST (Basic Local Alignement Search Tool) (13). Ces deux programmes permettent d'obtenir une liste de séquences similaires parmi une base de données de séquences nucléiques ou protéiques. Les candidats sont classés grâce au calcul d'un score ou d'une probabilité d'obtenir cette séquence par hasard.

- **FASTA**

L'algorithme FASTA présente deux étapes principales. Tout d'abord, une recherche de mots similaires entre la séquence requête et les séquences d'une banque, est effectuée. Pour cela, deux listes de mots, de longueur 6 nucléotides pour les acides nucléiques et 2 acides aminés pour les protéines, sont construites à partir de la séquence requête et des séquences de la banque. Les mots sont comparés et un score initial est calculé. Puis dans un second temps, les mots de score supérieur à un certain seuil et localisés dans une fenêtre de taille déterminée par la longueur des insertions tolérées, sont reliés. Un nouveau score est calculé. Lorsqu'il dépasse un seuil, un alignement de type Smith et Waterman est réalisé sur la région et donne un score optimal.

Le principal avantage de FASTA est la possibilité d'insertions-délétions dans la recherche de similarité. De plus, la recherche de mots de 6 nucléotides est reconnue comme très efficace pour rechercher des similitudes entre séquences nucléiques.

- **BLAST**

L'algorithme BLAST comprend lui aussi deux étapes. Il consiste à repérer tous les HSP (High-scoring Segment Pair) entre la séquence requête et les séquences de la base. Un HSP est un segment commun, le plus long possible, entre deux séquences et correspond à une similitude sans insertion-délétion ayant au moins un score supérieur ou égal à un score seuil. Pour déterminer un HSP, des mots de longueur fixe sont identifiés dans un premier temps entre la séquence requête et les séquences de la banque. Dans le cas des acides nucléiques, cela revient à des recherches d'identité entre les deux séquences de segments de 11 nucléotides. Dans le cas des protéines, une liste de mots similaires de 3 acides aminés de la séquence requête est établie. Un mot similaire est un mot de la séquence requête qui obtient un score supérieur à un score seuil, en fonction d'une matrice de substitution. Puis les séquences qui possèdent au moins un de ces mots dans la banque sont repérées. Dans un deuxième temps, la similitude est étendue dans les deux directions le long de chaque séquence, à partir du mot commun, de manière à ce que le score cumulé puisse être amélioré. L'extension du segment de similarité s'arrête dans trois cas : si le score cumulé descend d'une quantité x donnée par rapport à la valeur maximale qu'il avait atteint, si le score cumulé devient inférieur ou égal à zéro, si la fin d'une des deux séquences est atteinte.

Le principal avantage de l'algorithme BLAST est d'être très sensible grâce à la recherche de mots similaires dans sa première étape. Cependant, des régions de faible complexité c'est-à-dire avec certains acides aminés répétés, peuvent être choisies comme HSP et mettre en évidence des similitudes sans signification biologique. Il convient alors d'utiliser des programmes comme DUST pour les séquences nucléiques ou SEG pour les séquences protéiques, afin de masquer ces régions.

3.1.1.4 Alignement multiple

Les programmes BLAST et FASTA permettent d'identifier un ensemble de séquences similaires à une séquence requête. Des caractéristiques communes à ces séquences peuvent être identifiées et mises en évidence en les alignant les unes par rapport aux autres. Cependant, il est difficile d'aligner plus de deux séquences dans un temps raisonnable par programmation dynamique. C'est pourquoi les programmes les plus utilisés d'alignement multiple tel que CLUSTAL W (23), MULTALIN (24) reposent sur une heuristique qui consiste en un alignement progressif plutôt que simultané. Ces méthodes d'alignement progressif consistent à réaliser tout d'abord une classification hiérarchique basée sur les distances 2 à 2 des séquences. Dans une deuxième étape, un alignement progressif des séquences est réalisé en suivant cette classification jusqu'à l'obtention d'un alignement multiple.

Récemment d'autres méthodes heuristiques progressives ont été développées, utilisant de nouvelles fonctions « objectives » comme dans DiAlign2 (25) ou dans T-COFFEE (26). Le logiciel T-COFFEE contient une fonction objective qui tient compte d'une bibliothèque d'alignements de paires de séquences pour évaluer l'alignement multiple à chaque étape de l'alignement progressif. D'autres logiciels utilisent des méthodes itératives comme le logiciel SAGA qui contient une méthode d'optimisation de la fonction objective par algorithme génétique traduisant l'intérêt biologique de l'alignement (27).

L'alignement multiple permet de caractériser des familles de protéines en mettant en évidence des positions conservées dans leurs séquences. Un bon alignement multiple peut être utilisé pour aider à la prédiction des structures secondaire et tertiaire d'une protéine. Il est l'étape préliminaire à la construction d'arbres phylogénétiques. Il permet aussi de calculer une séquence consensus ou un profil pour une famille de protéines.

3.1.2 Les profils

Les méthodes de comparaison par paires de séquences tel que BLAST et FASTA supposent que toutes les positions dans l'alignement sont d'égales importances. L'alignement multiple de séquences d'une même famille, grâce à leur similarité de séquence ou de structure, permet lui de mettre en évidence que certains acides aminés sont plus conservés. Ces familles de séquences alignées peuvent être caractérisées par des outils nommés profils. Deux types de profils, reposant sur des méthodologies différentes sont utilisés : les profils ou PSSM (Position-Specific Scoring Matrices) résumant l'information dans une matrice à 2 dimensions ou bien les profils HMM (Hidden Markov Models), structures composées d'états et de transitions, et d'un ensemble de distributions de probabilités de transitions. Les profils sont des outils permettant de détecter des séquences similaires aux séquences composant le profil et en même temps des séquences plus distantes que la séquence requête de départ, grâce au fait que le profil tient compte d'une composition en acides aminés de la famille de séquences.

3.1.2.1 Profil ou PSSM

Un profil ou matrice de score position spécifique (PSSM) peut être décrit comme une matrice composée de 20 ou 21 colonnes et N lignes (figure 2). Les 20 premières colonnes de chaque ligne contiennent des scores pour chaque position de l'alignement, pour les 20 acides aminés possibles et la 21ème colonne peut correspondre à la pénalité d'insertion ou de délétion. Les N lignes correspondent aux positions de l'alignement multiple. La valeur du score à une position p d'un acide aminé a dans un profil calculé par PROFMAKE (28) est de :

$$M(p,a) = \Sigma^{20}_{b=1} W(p,b) \, . \, Y(a,b)$$

Avec Y(a,b) la valeur du score de substitution de l'acide aminé a en b dans la matrice de Dayhoff et W(p,b) un poids pour l'apparition de l'acide aminé b à la position p.
W(p,b) peut être choisi comme la fréquence d'apparition de l'acide aminé b à la position p.

$$W(p,b) = n(b,p) / N_R$$

Avec N_R le nombre de séquences dans l'alignement multiple.

Un poids logarithmique peut aussi être utilisé tel que W(p,b) est proportionnel à

$$\log [n(b,p) / N_R].$$

Last position-specific scoring matrix computed

		A	R	N	D	C	Q	E	G	H	I	L	K	M	F	P	S	T	W	Y	V
1	M	-1	-1	-2	-3	-1	0	-2	-3	-2	0	1	-1	6	0	-2	-1	-1	-1	-1	0
2	K	-1	1	0	-1	-3	0	0	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
3	A	4	-1	-2	-2	0	-1	-1	0	-2	-1	-1	-1	-1	-2	-1	0	0	-3	-2	0
4	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	1	4	-2	1	0	-3	-2	-1	-2	-1	0
5	M	0	-1	0	-1	-1	0	-1	-1	-1	0	0	0	5	0	-1	2	0	-1	-1	0
6	K	-1	1	0	-1	-3	0	0	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
7	K	-1	0	-1	-2	-2	0	0	-2	-1	0	1	4	0	-1	-1	0	-1	-2	-1	0
8	D	-1	0	0	5	-3	0	1	-1	-1	-3	-3	3	-2	-3	-1	0	-1	-4	-2	-2
9	G	2	-1	0	-1	-1	-1	-1	5	-2	-2	-2	-1	-2	-2	-1	0	-1	-2	-2	-1
10	E	2	0	0	0	-2	1	4	-1	0	-2	-2	0	-1	-2	-1	0	0	-3	-2	-1
11	F	-1	-1	-1	0	-2	0	2	-2	0	-1	-1	-1	0	5	-2	-1	-1	0	1	-1
12	G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
13	A	3	-1	-2	-2	0	-1	-1	-1	-2	2	0	-1	0	-1	-1	0	0	-3	-1	1
14	W	-1	-3	-4	-4	-1	-2	-2	-2	-2	0	0	-2	0	0	-3	-2	-1	10	1	2
15	M	-1	-1	-2	-3	-1	0	-2	-3	-2	1	3	-1	5	0	-2	-1	-1	-1	-1	0
16	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	0	5	-2	-2	0
17	D	-1	-1	0	5	-3	0	3	-1	0	-3	-4	0	-2	-3	-1	0	-1	-4	-2	-2
18	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	0	-2	0	-1	-2	-2	0	-3	-1	4
19	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
20	I	0	-3	-3	-3	-1	-2	-2	-4	-3	4	1	-2	0	0	-2	-2	0	-3	-1	3
21	P	-1	-2	-2	-1	-3	-1	-1	-2	-2	-3	-3	-1	-2	-4	7	-1	-1	-4	-3	-2
22	E	-1	0	0	1	-4	1	5	-2	0	-3	-3	0	-2	-3	-1	0	-1	-3	-2	-2
23	I	-1	-2	-3	-3	-1	-2	-3	-4	-3	4	3	-2	1	0	-3	-2	-1	-2	-1	2
24	D	-1	-2	0	5	-3	0	0	3	-1	-3	-4	-1	-3	-3	-1	0	-1	-3	-3	-3
25	H	-1	0	0	-1	-3	0	0	-2	7	-3	-2	3	-1	-1	-1	0	-1	-2	0	-2
26	H	-2	0	4	0	-3	0	0	-1	7	-3	-3	0	-2	-1	-2	0	-1	-2	0	-3
27	D	-1	-1	0	5	-3	0	3	-1	0	-3	-4	0	-2	-3	-1	0	-1	-4	-2	-2
28	L	0	-2	-3	-4	-1	-2	-2	-4	-3	2	3	-2	1	0	-2	-2	0	-2	-1	2
29	L	-1	-2	-3	-4	-1	-2	-3	-4	-3	1	4	-2	1	0	-3	-2	-1	-2	-1	0
30	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	1	-3	0	0	-3	-2	-1	-3	-1	3
31	K	-1	1	0	-1	-3	0	0	-2	-1	-3	-2	5	-1	-3	-1	0	-1	-3	-2	-2
32	I	0	-3	-3	-3	-1	-2	-2	-4	-3	4	1	-2	0	0	-2	-2	0	-3	-1	3
33	R	-1	4	0	-1	-3	0	0	-2	0	-3	-2	3	-1	-3	-1	0	-1	-3	-2	-2
34	K	2	1	0	-1	-1	0	0	-1	-1	-2	-1	4	-1	-2	-1	0	0	-3	-2	-1
35	T	2	-1	0	-1	0	-1	-1	-2	-1	-1	-1	-1	-1	-2	-1	0	4	-2	-2	0
36	A	3	-1	0	-1	0	0	0	0	-1	-1	-1	0	-1	-2	-1	2	0	-3	-2	0
37	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	1	-3	0	0	-3	-2	-1	-3	-1	3
38	C	0	-3	-3	-3	10	-3	-4	-3	-3	-1	-1	-3	-1	-2	-3	-1	-1	-2	-2	-1
39	G	0	-2	0	-1	-3	-2	-2	6	-2	-4	-4	-2	-3	-3	-2	0	-2	-2	-3	-3
40	T	0	-1	0	-1	-1	-1	-1	-2	-2	-1	-1	-1	-1	-2	-1	0	5	-2	-2	0
41	D	-2	-2	0	6	-3	0	1	-1	-1	-3	-4	-1	-3	-3	-1	0	-1	-4	-3	-3
42	V	0	-3	-3	-3	-1	-2	-2	-3	-3	3	0	-2	0	-1	-2	-2	0	-3	-1	4
43	H	-2	0	0	-1	-3	0	0	-2	9	-3	-3	-1	-2	-1	-2	-1	-2	-2	1	-3
44	I	-1	-3	-3	-3	-1	-3	-3	-4	-3	4	1	-3	0	0	-3	-2	-1	-3	-1	3

Figure 2 : Profil ou matrice de score position spécifique obtenue avec PSI-BLAST.
A l'horizontal, sont représentés les 20 acides aminés et à la verticale la séquence consensus d'un alignement multiple.

On peut calculer une séquence consensus à partir du profil tel que l'acide aminé c choisi comme consensus à la position p est celui qui a le plus grand score M(p,c). C'est-à-dire, l'acide aminé le plus similaire aux autres résidus de la colonne de l'alignement multiple et non pas le plus observé dans la colonne. Les profils sont utilisés par des programmes de recherche de similarité dans des banques de séquences tel que PROFILESEARCH (29) ou PSI-BLAST (30). Ils ont montré dans plusieurs analyses leur grande sensibilité, grâce aux scores dépendant de la position dans un alignement multiple de séquences. Cependant, comme pour la recherche de similarité avec une séquence contre une banque, il est difficile de discriminer les candidats vrais positifs des faux positifs

3.1.2.2 Profil HMM

Les modèles de Markov cachés (HMM) (31) sont une classe de modèles probabilistes généralement utilisés en reconnaissance vocale. Ils ont été introduits en bioinformatique dans les années 80 et utilisés pour la construction de profils dans les années 1990 (32). Un profil HMM comprend trois états pour chaque colonne d'un alignement multiple (figure 3). Un état 'match' modélisant la distribution des résidus observés dans la colonne. Un état 'insertion' permettant l'insertion de un ou plusieurs acides aminés entre la colonne considérée et la suivante et un état 'délétion' pour enlever la colonne correspondant à l'acide aminé consensus. Les états 'match' et 'insertion' ont chacun 20 probabilités d'émission correspondant aux 20 acides aminés alors que les états 'délétion' n'ont pas de probabilité d'émission. Les probabilités du profil HMM sont transformées en scores avant de pouvoir aligner une nouvelle séquence avec le profil. Le score est un score comparable à celui de BLAST ou FASTA. Ainsi le score pour un résidu x dans l'état 'match' est égal à :

$$S = \log (p_x / f_x)$$

Avec p_x la probabilité d'émission pour le résidu x dans l'état match et f_x la fréquence attendue pour le résidu x dans la banque de séquence.

En revanche, les scores d'insertion ne sont pas les mêmes que les pénalités de gap des programmes d'alignement standard. Ainsi, le coût d'ouverture d'un gap pour un HMM est égal à :

$$a = \log t_{MI} + \log t_{IM}$$

Avec t_{MI} la probabilité de l'état de transition pour aller de l'état 'match' à l'état 'insertion' et t_{IM} la probabilité de l'état de transition pour quitter l'état 'insertion'.
Le coût d'extension d'un gap est égal à b :

$$b = \log t_{II}$$

Avec t_{II} la probabilité de la transition.
Ainsi, les coûts de gaps ne sont pas arbitraires, ils sont optimisés sur le jeu d'entraînement du profil HMM.

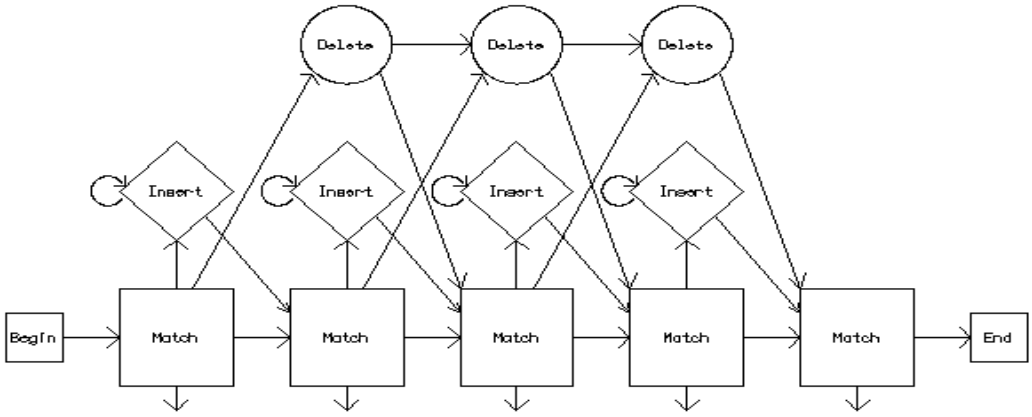


Figure 3: Profil HMM à trois états: matche, insertion et délétion.

Les profils PSSM et HMM sont assez équivalents pour décrire des familles de domaines. Les profils HMM permettent de faire varier selon la position les pénalités d'insertion délétion (33). Cependant, des profils généralisés ont été développés, se rapprochant de la structure d'un HMM (34) et il existe d'ailleurs des programmes de conversion dans les deux sens.

3.1.2.3 Programmes de recherche de similarité avec des profils

- **PSI-BLAST**

Plus récemment, la méthode du programme BLAST a été développée pour améliorer la sensibilité de la recherche de similarité. Ceci afin de trouver des séquences moins similaires à la séquence de départ mais pouvant être le résultat d'une divergence évolutive. L'équipe d'Altschul a créé le programme PSI-BLAST (Profile Searches Iterated Basic Local Alignement Search Tool) (30). Ce programme est utilisé pour détecter des similarités de séquences dans une base de données de séquences protéiques. Sa stratégie consiste à construire un alignement multiple à partir du résultat obtenu avec BLAST, puis de calculer une matrice de scores position spécifiques ou profil à partir de cet alignement multiple et de

l'utiliser comme requête à l'itération suivante. Le processus peut être itéré jusqu'à ce qu'il n'y ait plus de similarité significative trouvée. D'autre part, ce programme peut-être utilisé pour calculer une matrice de scores position spécifiques directement à partir d'un alignement multiple prédéfini de séquences.

i) Construction de l'alignement multiple

Les séquences récupérées au premier tour de BLAST pour construire un alignement multiple ont une E-value inférieure à la valeur par défaut 0.01. Un alignement multiple est construit puis à partir de celui-ci est extrait un alignement réduit M_c dans lequel toutes les séquences recrutées sont représentées soit par un résidu soit par un gap. Ce qui permet de calculer une matrice de longueur égale à la séquence requête d'origine.

ii) Poids des séquences

Pour éviter de donner un poids naturel trop important aux séquences similaires et permettre aux séquences divergentes d'être prises en compte dans le calcul du profil, un poids est attribué aux différentes séquences. Dans PSI-BLAST, la méthode de Henikoff et Henikoff modifiée est implémentée (35). Elle consiste à tenir compte du nombre moyen de types de résidus observés dans les différentes colonnes de l'alignement (21 types possibles, pour 20 résidus et l'indel).

iii) Estimation des fréquences

Le score calculé à une position de l'alignement multiple est le logarithme des ratios entre la fréquence observée d'un acide aminé et la fréquence attendue de cet acide aminé, il est de la forme:

$$\log (Q_i / P_i)$$

Avec Q_i la probabilité observée de trouver un acide aminé i dans la colonne de l'alignement multiple. Q_i est estimé en utilisant la méthode des pseudocounts (36). P_i est la fréquence attendue pour l'acide aminé i dans la banque de séquence.

Le programme PSI-BLAST (version 2.2.1) a été récemment modifié afin d'améliorer son efficacité (37). Les modifications principales sont la possibilité d'appliquer l'algorithme de Smith-Waterman sur chaque alignement obtenu ainsi que le filtrage automatique des segments de séquences de faible complexité dans la banque de séquences. D'autre part, l'utilisation de statistiques basées sur la composition en acides aminés de la séquence requête et de la banque permettent de mieux évaluer si les alignements locaux sont significatifs.

L'utilisation de PSI-BLAST a été évaluée pour l'annotation structurale de génomes bactériens (38). Pour cela, un ensemble de domaines ont été sélectionnés dans la banque SCOP (39). PSI-BLAST a été testé sur ce jeu de séquences pour vérifier qu'il permet bien de trouver tous les domaines d'une superfamille sans détecter trop de faux positifs. Ce jeu de références a été utilisé avec PSI-BLAST pour analyser la composition en repliements structuraux des génomes de *Mycoplasma genitalium* et *Mycobacterium tuberculosis* et a permis de mettre en évidence que seulement 20 à 30% des résidus de ces génomes ne correspondent pas à des protéines de structures connues.

- **IMPALA et RPS-BLAST**

Pour faciliter les analyses de génomes complets, des logiciels complémentaires à PSI-BLAST ont été développés tels que IMPALA (Integrating Matrix Profiles And Local Alignments) (40) et RPSBLAST (Reversed Position Specific BLAST) (41). Ils consistent à comparer une séquence protéique requête contre une banque de profils ou PSSM construite préalablement avec PSI-BLAST.

IMPALA est un logiciel qui utilise une grande partie de la théorie et du code de PSI-BLAST, excepté pour le calcul du score de l'alignement local qui est effectué en utilisant l'algorithme de Smith-Waterman (aujourd'hui intégré dans la nouvelle version 2.2.1 de PSI-BLAST). En effet, le fait de construire une banque de profils a pour objectif d'éviter la redondance d'information de séquence et donc d'effectuer une recherche sur un ensemble de données plus petit qu'une banque de séquences. Ce qui permet de tolérer un temps plus long de recherche. IMPALA comprend trois programmes : makemat et copymat pour le formatage des profils et impala pour la recherche d'une séquence requête contre la banque de PSSM. La recherche de similarité est une recherche d'alignement local optimal avec insertion délétion, de type BLAST.

RPS-BLAST est un logiciel reposant sur l'algorithme de BLAST en recherchant des mots sans gap puis en réalisant une extension des matchs. Lorsqu'un alignement sans gap est trouvé avec un score supérieur à une valeur seuil alors une extension avec gap est réalisée et les alignements avec une E-value inférieure à un seuil sont apportés. Ainsi RPS-BLAST est plus rapide qu'IMPALA puisqu'il utilise à la fois une banque de mots de type BLAST pour la séquence requête mais aussi une table précalculée pour les profils. RPS-BLAST est utilisé pour faire des recherches d'homologies contre la banque CDD (41) de domaines conservés.

3.1.3 Les programmes de décomposition des protéines en domaines

Etant donné la modularité des protéines, la comparaison de séquences nécessite l'utilisation d'algorithmes de classification des protéines en familles de domaines. De nombreux algorithmes ont été développés tels que DOMAINER (42), DIVCLUS (43). Nous présentons deux algorithmes : MKDOM 2 (44) développé par Jérome Gouzy au laboratoire des Interactions Plantes Microorganismes à Toulouse et PICASSO (45) développé à l'Institut Européen de Bioinformatique (EBI, UK).

3.1.3.1 MKDOM

MKDOM version 2 (44) repose sur le programme PSI-BLAST en utilisant comme séquence requête la plus petite séquence de la banque. Cette séquence est supposée être composée d'un seul domaine, dans la banque de séquences choisie pour définir des familles de domaines. L'algorithme repose sur trois étapes : (figure 4) sélection de la plus petite séquence comme requête, recherche de similarité avec PSI-BLAST contre la banque pour générer la famille de domaines puis retrait des segments de séquences de ces domaines de la banque. Le processus est itéré jusqu'à épuisement de la banque. Le programme produit une séquence consensus et un alignement des séquences d'une famille de domaines avec le programme d'alignement multiple MULTALIN (24). Le résultat de la décomposition en domaines des protéines de la banque peut être visualisé en utilisant le programme XDOM (46). MKDOM2 a permis la création des bases de données ProDom et ProDom-CG (Protein Domains of Complete Genomes). La dernière version de ProDom de 2002 comprend 481.952 séquences décomposées en 365.172 familles de domaines par MKDOM2 dont 138.322 familles avec plus de deux séquences.

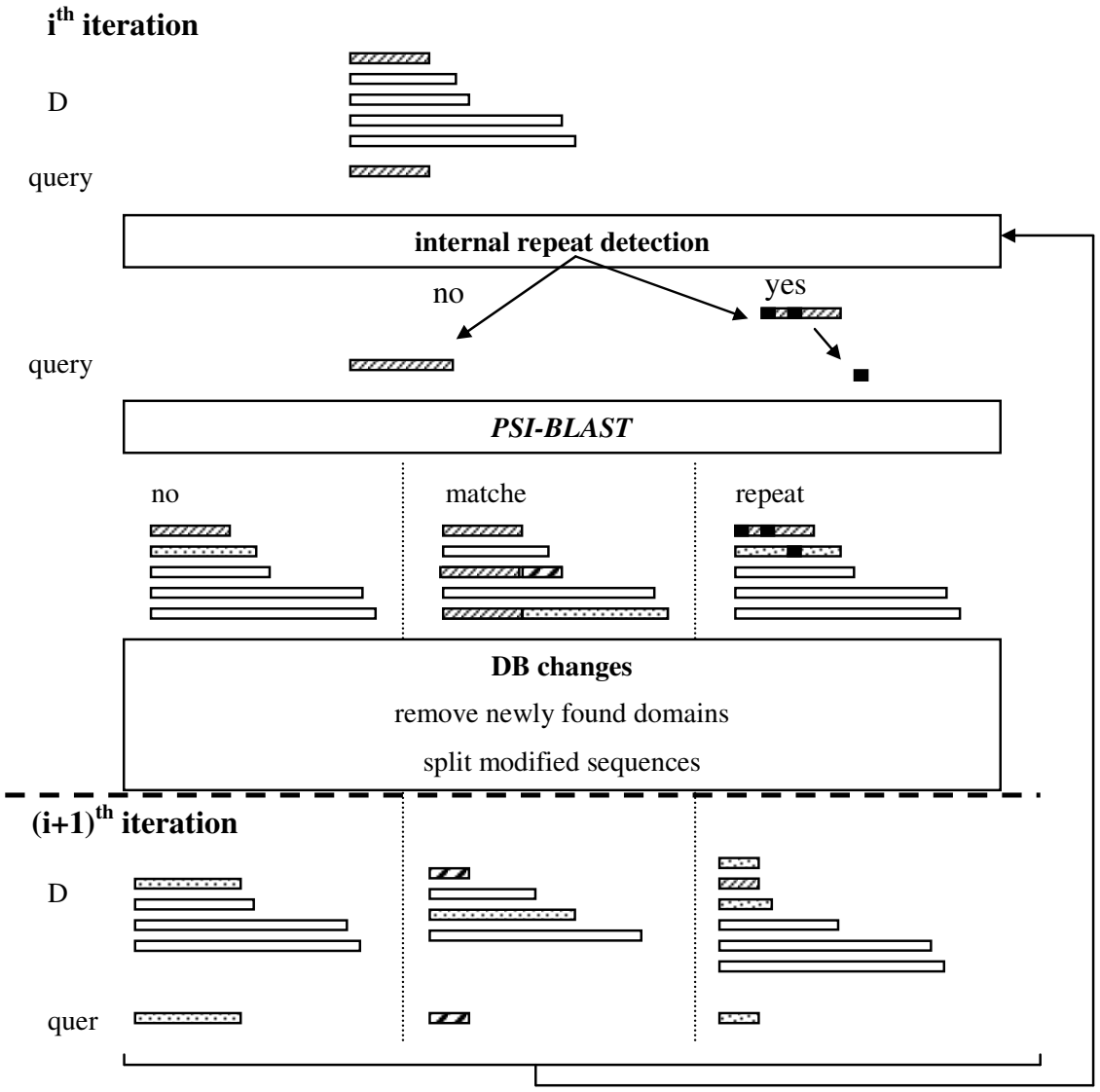


Figure 4: Représentation schématique de la décomposition en domaines par MKDOM2.

3.1.3.2 PICASSO

L'algorithme PICASSO (Protein Incremental Classification And Sequence Space Organization) a pour objectif de trouver un nombre minimum de profils pour couvrir un ensemble de séquences (45). Il comprend trois étapes : une recherche BLAST de tout contre tout permettant de faire des alignements multiples, une unification hiérarchique de profils obtenus à partir des alignements et un découpage des modules mobiles c'est-à-dire observés

dans différentes familles avec des combinaisons de voisins différents. PICASSO a permis d'identifier environ 33.000 familles à partir d'une banque de 150.000 séquences.

3.2 Les bases de données biologiques

Depuis les années 70, avec la création d'une méthode de séquençage par W. Gilbert, il devient nécessaire de répertorier les séquences biologiques dans des bases de données (base de données PIR en 1968) pour gérer les données devenues importantes et complexes.
On distingue deux types de bases de données biologiques :

- les bases de données généralistes qui répertorient tous les types de séquences nucléiques ou protéiques et les données de structures tridimensionnelles,
- les bases de données spécialisées ou thématiques qui classent ces données de séquences ou de structures selon une caractéristique biologique particulière, tel que des motifs, des domaines, des classifications structurales, des voies biochimiques,...

Ces bases de données peuvent être interrogées en utilisant des systèmes d'interrogations croisés comme SRS (Sequence Retrieval System) (47), ENTREZ (48) ou DBGET (49). Ils permettent de sélectionner des séquences dans une ou plusieurs banques selon un ou plusieurs champs grâce aux opérateurs logiques ET, OU, NON.

Dans cette introduction aux bases de données biologiques, un sous ensemble des bases de données existantes est présenté. Ce sont des bases de données très utilisées ou qui concernent la thématique de recherche de cette thèse. En effet, dans des catalogues de bases de données comme DBcat (50), 511 bases de données biologiques sont inventoriées en mars 2003.

3.2.1 Les bases généralistes de séquences

3.2.1.1 Bases de séquences nucléiques

Il existe trois bases de données biologiques généralistes principales, de séquences nucléiques : GenBank aux USA (51), EMBL (European Molecular Biology Laboratory) (52) et DDBJ (DNA Data Bank of Japan) (53). Grâce à la collaboration internationale INSD (International Nucleotide Sequence Databases) de ces trois bases de données depuis 1982, les séquences nucléiques sont échangées chaque jour pour assurer une synchronisation des informations. Les laboratoires ou les centres de séquençage envoient les séquences à l'aide

des outils de soumission Bankit pour GenBank, Webin pour EMBL et Sakura pour DDBJ ou avec le programme de soumission multi-plateforme Sequin pour GenBank et EMBL.

GenBank est une base de données biologique généraliste de séquences nucléiques provenant de 119.000 organismes différents dans la version 131 d'août 2002 (51). Elle a été créée et est maintenue par le National Center for Biotechnology Information (NCBI, USA). La version d'août 2002 contient 18.2 millions de séquences provenant entre autre de 29 génomes complets microbiens, 134 génomes de microorganismes et 33 génomes eucaryotes en cours de séquençage. Une entrée GenBank consiste en un fichier texte comprenant des champs qui contiennent une description de la séquence, le nom scientifique, la taxonomie de l'organisme d'origine de la séquence, des références bibliographiques et une liste de caractéristiques biologiques telles que les traductions en protéines, les unités de transcription, les régions répétées et les sites de mutations. GenBank est distribué sous forme de 17 divisions, dont 4 correspondent aux groupes taxonomiques (bactéries, virus, primates, rongeurs). Les autres divisions correspondent à différentes stratégies de séquençage. La division des EST (Expressed Sequence Tag), par exemple, est la source principale de nouvelles entrées. Pour organiser les données d'EST sans avoir de redondance, le NCBI a développé une base spécialisée de clusters de séquences, UniGene, comprenant 9 animaux et l'homme. Les autres divisions contiennent entre autres les STS (Sequence-Tagged Sites), les GSS (Genome Survey Sequence), les séquences HTG (High Throughput Genomic), les séquences HTC (High Throughput cDNA).

EMBL (European Molecular Biology Laboratory) (52) est une base de données biologique généraliste de séquences nucléiques, développée par l'European Bioinformatics Institute (EBI). La structure de cette banque est similaire à celle de GenBank pour le format des fichiers et pour sa construction en divisions.

En collaboration avec les autres banques nucléiques généralistes, un nouveau type de soumission de données a été créé : les séquences TPA (Third Party Annotation). Ces soumissions permettent l'annotation de séquences nucléiques existantes par un autre auteur que l'auteur de la séquence. Elles peuvent être de différentes origines, des séquences d'ARNm assemblées à partir d'EST chevauchants, un ARNm à partir d'une séquence génomique non annotée, des annotations d'exons, d'introns, de régions codantes à partir d'une séquence non annotée.

DDBJ (DNA Data Bank of Japan) est la base de données biologique généraliste de séquences nucléiques, développée au Japon. En 2002, 17,3% des entrées de l'INSD sont

apportées par la DDBJ (53). Ces nouvelles entrées proviennent essentiellement des génomes du riz, de la souris, du chimpanzé et du nématode. Les données de génomes complets qui seront incorporées dans l'INSD sont tout d'abord classées dans la banque spécialisée GIB (Genome Information Broker). La nouveauté 2002 de cette banque est l'utilisation du métalangage XML (eXtensible Markup Langage) à la place des formats fichiers plats. En effet, les fichiers plats ne sont pas très efficaces pour extraire des informations des entrées de séquences de génome car une entrée contient souvent plusieurs gènes. Le format DDBJ-XML permet de faciliter l'écriture de requêtes complexes avec la base de données car ce langage permet d'inventer des nouvelles balises pour isoler toutes les informations élémentaires. Ce format est tout d'abord appliqué à la base de données GIB qui contient les génomes complets de 90 espèces.

3.2.1.2 Bases de séquence protéiques

Il existe deux bases de données de séquences protéiques très utilisées : Swiss-Prot qui est certainement la base de donnée la plus utilisée par la qualité de son annotation, et PIR également utilisée car elle couvre un plus grand nombre d'espèces.

- **Swiss-Prot et TrEMBL**

Swiss-Prot est une base de données non redondante de séquences protéiques annotées, créée par Amos Bairoch en 1986 et maintenue aujourd'hui par le SIB (Swiss Institute of Bioinformatics) et l'EBI (European Bioinformatics Institute). Dans la version 40.39 du 10 Janvier 2003, on compte 120 961 entrées (54). La spécificité de cette base repose sur trois critères : une annotation de grande qualité, un niveau minimal de redondance, des liens avec de nombreuses bases de données. Une entrée Swiss-Prot comprend 2 parties : tout d'abord, les données principales que sont la séquence, son origine taxonomique et des citations, puis des informations d'annotation qui peuvent être la ou les fonctions de la protéine, des modifications post-transcriptionnelles, des sites et des domaines, des structures secondaire, tertiaire et quaternaire, des conflits de séquences. Certaines banques contiennent plusieurs entrées pour une même séquence protéique, ceci à partir d'informations bibliographiques. Ces différences de séquences protéiques peuvent être dues à différentes causes biologiques : des variants d'épissage, des polymorphismes, des mutations causant des maladies, mais aussi des erreurs de séquençage. Ces informations sont indiquées dans le champ FT (Feature Table) d'une seule entrée de Swiss-Prot. Une entrée Swiss-Prot contient aussi des références vers

d'autres bases de données (43 bases de données). Des bases de données généralistes avec le lien vers la séquence nucléique et des bases de données thématiques telles que des bases de signatures, de familles de domaines, de structures pour améliorer l'annotation.

Pour compléter Swiss-Prot qui contient peu de séquences par rapport aux bases de séquences nucléiques, EMBL a développé en 1996 la base de données TrEMBL (Translation of EMBL nucleotide sequence database) de traduction automatique de ses régions codantes (CDS), qui ne sont pas encore intégrées dans SWISS-PROT. Il existe deux sections dans TrEMBL : SP-TrEMBL et REM-TrEMBL. SP-TrEMBL contient les séquences qui seront éventuellement incorporées dans Swiss-Prot. REM-TrEMBL contient les séquences qui ne le seront pas c'est-à-dire les immunoglobulines, des fragments de séquences, … (55). Chaque séquence TrEMBL est annotée de façon automatique à partir de trois types de données : la base de données de référence Swiss-Prot, la base de données de familles de protéines et de signature INTERPRO pour classer la protéine non annotée dans un groupe, et la base de données de règles d'annotation RuleBase. Les données de Swiss-Prot et TrEMBL sont sous forme de fichiers plats pour le moment, mais une base de données relationnelle a été développée et les fichiers sont transformés en documents avec un nouveau format de type XML : SP-ML (SWISS-PROT Markup Language).

- **PIR**

La base de données PIR (Protein Information Ressource), créée en 1968, maintient deux bases de données généralistes de séquences protéiques avec la collaboration du MIPS (Munich Information Center for Protein Sequence) et du JIPID (Japan International Protein Sequence Database). La banque PSD (Protein Sequence Database) contient dans la dernière version de 2002 (56) 283 000 entrées annotées et classées couvrant toute la taxonomie, tandis que la banque NREF (Non-Redondant Reference Database) regroupe de façon non redondante toutes les données contenues dans les banques de séquences protéiques : PIR-PSD, SWISS-PROT, TrEMBL, RefSeq, GenPept et PDB. NREF contient plus d'un million d'entrées.

La banque PIR-PSD est annotée en utilisant une classification automatique en familles et superfamilles d'après le travail de Margaret Dayhoff. 99% des séquences de la PSD appartiennent à une famille et les séquences appartenant à une famille ont au moins 45% d'identité entre elles. Les superfamilles sont constituées de familles de séquences présentant

une même architecture en domaines. Comme SWISS-PROT, PIR existe sous forme de fichier plat et est depuis peu distribué en format XML.

3.2.1.3 Base de structures tridimensionnelles

Une protéine n'est pas seulement un polymère linéaire de 20 types d'acides aminés possibles, ce polymère se replie en une conformation tridimensionnelle lui donnant sa fonction. Ainsi, cette structure tridimensionnelle (3D) peut mettre en évidence un site actif de catalyse d'une enzyme ou un site de fixation d'une protéine à l'ADN. La connaissance de la structure 3D d'une protéine est donc un moyen, en recherche fondamentale, pour comprendre son mode d'action, c'est-à-dire sa fonction, et en recherche appliquée, pour découvrir de nouveaux médicaments. La première structure 3D obtenue est celle de la myoglobine en 1960, par la technique de diffraction au rayon X du cristal de la protéine. On utilise aussi la technologie de Résonance Magnétique Nucléaire (RMN) pour déterminer les coordonnées des atomes d'une macromolécule, qui peut être une protéine, mais aussi de l'ADN ou de l'ARN.
La Protein Data Bank (PDB) est la plus ancienne archive de données de structures tridimensionnelle, elle contenait 7 entrées en 1971, en octobre 2002 (57) elle contient 18 000 structures. Le processus d'acquisition des données est effectué par l'organisme Research Collaboratory for Structural Bioinformatics (RCSB) depuis 1999. Les données sont uniformisées pour faciliter les recherches dans la base de données. Les champs d'une entrée donnent des informations sur le nom de la macromolécule, l'organisme source des données, la méthode utilisée et sa résolution, les coordonnées des atomes, …
De nombreuses bases de données secondaires sont issues de la PDB. Elles exploitent ces informations pour en extraire d'autres comme la décomposition en domaines structuraux des protéines dans SCOP et CATH.

3.2.2 Les bases de domaines et de signatures

3.2.2.1 Bases de domaines à partir de structures

A ce jour, les technologies employées permettent d'élucider beaucoup plus de structures de protéines globulaires que de protéines membranaires. Les protéines globulaires peuvent être classées selon leur contenu et l'arrangement des éléments de structures secondaires dans un domaine. Un domaine peut ne contenir que des hélices α, que des feuillets β, des hélices α et feuillets β mélangés, un groupe d'hélices α puis un groupe de

feuillets β ou des structures irrégulières. Des bases de données comme SCOP et CATH utilisent cette classification structurale des domaines pour hiérarchiser les protéines. Les classifications SCOP et CATH sont très utiles pour déduire des fonctions à partir de relations d'homologies structurales c'est-à-dire souvent avec des similarités plus faibles.

- **SCOP**

La base de données SCOP (Structural Classification of Protein) est une classification de toutes les protéines dont la structure est connue, selon des relations de structure et d'évolution. Les protéines sont classées d'après l'unité de base qui est le domaine structural défini par des experts. Le domaine structural est une région capable de se replier indépendamment dans le cas d'une protéine multi-domaines.

Le plus haut niveau de la hiérarchie, les classes, correspond à la classification des structures secondaires observées dans un domaine. Puis SCOP divise les classes en repliements, superfamilles et familles (58). Le repliement représente le niveau de similarité structurale. Ainsi, des séquences de faible similarité peuvent être regroupées dans le même repliement à cause de deux types de processus survenus au cours de l'évolution. Soit les séquences sont issues d'un événement de divergence à partir d'un ancêtre commun, en ayant conservé la même structure. Soit les séquences sont issues d'un événement de convergence vers une même structure. Le niveau le plus bas de la hiérarchie, la famille, regroupe les séquences à partir de résultats de similarité de séquence. Les séquences d'une même famille ont au moins 30% d'identité. Pour quelques exceptions, la similarité des séquences est plus faible mais la fonction est très proche, c'est le cas des globines avec 15% d'identité. Les familles sont ensuite regroupées en superfamilles car les structures et les fonctions des protéines sont très proches.

La version 1.55 de SCOP (39) contient 30 403 domaines répartis en 7 classes, 279 repliements, 947 superfamilles et 1557 familles. Ces domaines correspondent à 12 794 entrées de la PDB.

- **CATH**

La base de données CATH (Class, Architecture, Topology and Homologous superfamily) est, comme SCOP, une hiérarchisation des protéines avec des données de

structure, à partir de la classification des domaines structuraux. Alors que SCOP délimite les domaines uniquement par des experts, CATH utilise à la fois l'avis des experts et le résultat d'algorithmes. La classification des domaines comprend aussi 4 niveaux : la classe, l'architecture, la topologie ou repliement et la famille homologue.

Les classes de CATH comportent les grandes classes définies par les structures secondaires des domaines : tout α, tout β (59). Les classes α / β et α + β sont regroupées en une seule classe car l'équipe de CATH considère qu'il y a trop de chevauchements entre ces classes. D'autres classes contiennent les structures irrégulières et les protéines multi-domaines.

L'architecture regroupe manuellement les protéines avec des arrangements tridimensionnelles de repliements similaires. La topologie (ou repliement) regroupe automatiquement, par la méthode SSAP (60), des protéines qui possèdent une homologie structurale significative (score SSAP > 70). La famille homologue regroupe des protéines avec une forte similarité de séquence (35% d'identité) ou avec une faible similarité de séquence (20% d'identité) mais avec une forte homologie de structure.

La dernière version de CATH contient 34 287 domaines, répartis en 1 383 familles homologues (61).

Précédemment, nous avons présenté des bases de données regroupant des protéines selon leur décomposition en domaines, ceci seulement pour des protéines ayant une structure 3D résolue. Or, un grand nombre de bases de données protéiques classent les protéines selon leur décomposition en domaines, à partir de résultats de recherche de similarité de séquences entre protéines. Les protéines peuvent être caractérisées par différents types de descripteurs : des motifs, des empreintes, des alignements de régions de séquences représentés par des profils ou des HMM. Tous ces descripteurs sont utilisés pour classer en familles les protéines dans des bases de données spécialisées. Quelque unes de ces bases sont décrites.

3.2.2.2 Base de signatures

- **PROSITE**

PROSITE (62) est la première base de données de familles de protéines et de domaines, créée en 1988 par Amos Bairoch et Philipp Bucher du Swiss Institute of

Bioinformatics (SIB). PROSITE est une **collection de motifs** qui possède une signification biologique décrite par des expressions régulières ou des profils, liés à une documentation sur la famille de protéines ou le domaine qu'ils permettent de détecter.

Les motifs permettent de détecter des similarités entre des séquences éloignées, car ce sont des résidus particuliers qui ont été conservés au cours de l'évolution pour leur rôle biologique. Ce sont des segments courts de séquences puisqu'ils ont une longueur de 10 à 20 acides aminés. Ils sont impliqués dans différentes fonctions biologiques telles que le site catalytique d'un enzyme, le site de fixation d'un groupe prosthétique (hème, biotine,…), un site de liaison à un ion, des cystéines impliquées dans des ponts disulfures, un site de liaison à une molécule ou à une autre protéine. Les motifs sont obtenus à partir d'un alignement de séquences ayant une fonction similaire. La région la plus conservée de l'alignement est réduite à une expression régulière décrivant le motif. L'expression est de type M-x-G-x(3)-[IV]2-{FWY}. Cette expression signifie que le motif commence par une méthionine. Le x correspond ensuite à un résidu indifférent, puis il y a une glycine, puis trois résidus indifférents, on trouve ensuite entre crochets l'isoleucine ou la valine deux fois, et enfin entre accolades n'importe quel résidu sauf la phénylalanine, le tryptophane ou la tyrosine. Il n'y a pas de seuil permettant d'évaluer la significativité statistique d'un motif, mais une évaluation du nombre de motifs trouvés contre Swiss-Prot est donnée.

Un certain nombre de familles de domaines ne peuvent pas être retrouvées en utilisant des motifs car les séquences ont beaucoup divergé. Pour ce type de problème, PROSITE a construit des **profils ou matrices de scores**. En plus de la possibilité de détecter des séquences moins conservées, les profils permettent de caractériser des régions plus grandes que les motifs. Cet outil sera développé dans la partie « Profil ou PSSM ». En résumé, un profil est une matrice reflétant la distribution des acides aminés à chaque position de l'alignement multiple des séquences d'un domaine. Le profil est utilisé pour effectuer une recherche de similarité contre une banque et permet d'identifier des candidats grâce au calcul d'un score.

La version 17.35 de janvier 2003 contient 1167 documentations pour 1600 motifs et profils différents.

- **PRINTS**

PRINTS (63) est une **collection d'empreintes** caractérisant des familles de protéines. Les empreintes sont des groupes de motifs conservés sans indels de 10 à 20 acides aminés dans un alignement multiple d'une famille de protéines (figure 5). Ils permettent de tolérer des non-appariements au niveau du motif mais aussi au niveau de la signature complète. Ce type de descripteur peut aussi être utilisé pour discriminer les protéines au niveau des super-familles, des familles ou des sous-familles en identifiant des régions différentes entre ces niveaux. Cette base de données est construite manuellement comme PROSITE, c'est pourquoi dans la version 36.0 de septembre 2002, elle ne contient que 1800 empreintes caractérisant environ 11 000 motifs.

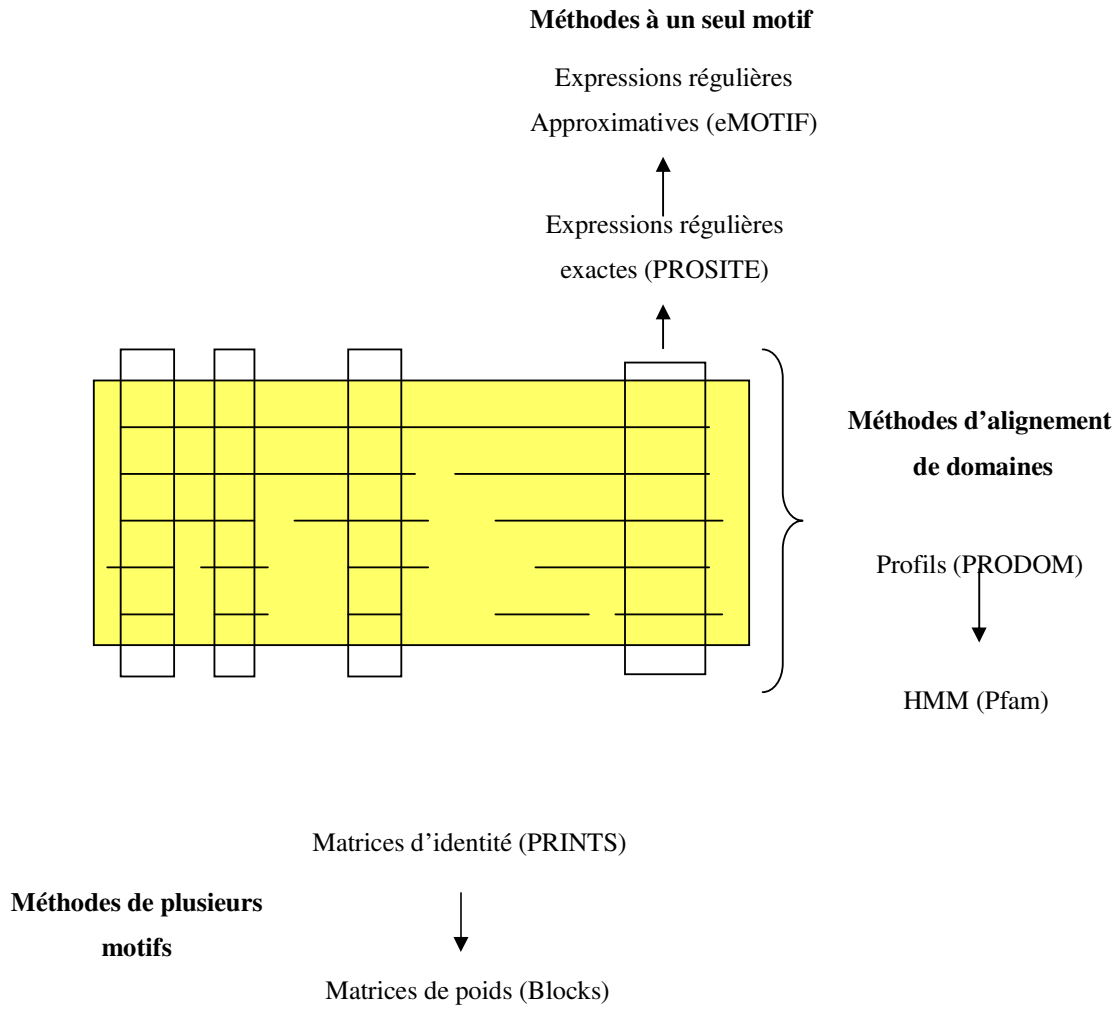


Figure 5 : Vue d'ensemble de trois approches d'analyse de séquence et les bases de données correspondantes.

3.2.2.3 Bases de domaines

- **Pfam**

Pfam (15) est une base de données **d'alignements multiples** de séquences protéiques et de **profils de type HMM** (Hidden Markov Models). La méthodologie des HMM est développée dans la partie « Profil HMM ». La version 7.8 de novembre 2002 contient 5 049 familles. Une famille Pfam contient des annotations sur la fonction, des références bibliographiques, des liens vers d'autres bases de données et deux types d'alignements multiples : un alignement résumé contenant seulement les membres représentatifs de la famille, utilisé pour calculer le profil HMM, et un alignement complet avec tous les membres de la famille trouvés avec le profil dans les banques Swiss-Prot et TrEMBL. Pfam est composée de deux parties, la partie Pfam-A est construite manuellement, la partie Pfam-B est un supplément de Pfam-A construit automatiquement à partir de la base ProDom. Afin d'améliorer les limites des domaines, Pfam utilise depuis peu les informations de structures données dans la base de données SCOP.

- **ProDom**

ProDom (14) est une base de données de **familles de domaines** protéiques générée automatiquement à partir des séquences de Swiss-Prot et TrEMBL. La construction de ProDom repose sur une méthode de regroupement automatique des régions homologues. Le programme MKDOM2 (44) effectue cette décomposition en domaines des protéines. L'algorithme repose sur l'hypothèse que la plus petite séquence protéique correspond à un domaine. Cette séquence est utilisée comme requête dans la recherche d'homologie contre Swiss-Prot et TrEMBL avec le programme PSI-BLAST (30). Les régions de séquences obtenues avec une homologie significatives sont regroupées dans une famille. Les fragments de séquences restants sont triés selon leur taille et la plus petite séquence est utilisée pour réitérer le processus. Certaines contraintes sont appliquées pour améliorer cette méthode. Tout d'abord, les fragments de séquences présents dans les banques Swiss-Prot et TrEMBL sont enlevés pour ne pas fausser l'hypothèse de départ. Le programme SEG est utilisé pour filtrer

les régions de faible complexité dans les séquences car elles sont la cause de similarités non significatives. Les zones de répétitions internes sont détectées pour être utilisées comme requêtes au lieu de la séquence entière. Enfin, 21 familles de domaines expertisées par le groupe de ProDom et les familles de Pfam-A sont utilisées pour construire des profils. Ces profils sont utilisés en première étape de la construction de ProDom pour détecter des domaines homologues par PSI-BLAST. La dernière version 2002.1 de ProDom utilise aussi les domaines de la base de données SCOP pour initier la construction, cette version contient 365 172 familles de domaines.

- **InterPro**

Il existe de nombreuses méthodes d'identification de motifs ou de domaines. Chaque descripteur a ses avantages sur les autres. Les motifs décrivent bien précisément un site fonctionnel dans une protéine mais ne tiennent pas compte du contexte des séquences, alors que les profils, ou HMM, peuvent tenir compte de la variabilité des séquences, ceci sur une longueur plus importante. Mais il est difficile de déterminer quel est le score ou la probabilité permettant de conclure à l'appartenance d'un candidat à une famille. D'où l'utilité de combiner tous les résultats de ces méthodes dans une seule source de données. Le consortium InterPro (64) a été créé en 1998 pour regrouper les données des 4 premières banques PROSITE, Pfam, PRINTS, ProDom. Puis en 2000, SMART et TIGRFAMs ont rejoint InterPro. Les données sont intégrées manuellement par une équipe de biologistes. Une entrée InterPro correspondant à un numéro d'accession qui comprend une ou plusieurs signatures pour différents types de données : une famille, un domaine, une répétition ou des sites de modification post-transcriptionnelle. Deux types de relations peuvent exister dans une entrée INTERPRO pour pouvoir regrouper les différentes données : la relation « parent / enfant », décrivant des ancêtres communs entre plusieurs entrées, les signatures enfants caractérisant un sous-ensemble des séquences caractérisées par la signature parent, la relation « contient / trouvé dans » qui indique la composition en domaines des séquences. Ce type de relation permet de caractériser des domaines mobiles c'est-à-dire trouvés dans différentes familles sans relation parent / enfant.

3.2.3 Les bases spécialisées dans le métabolisme

De nombreux types de fonctions biologiques sont inventoriés dans des bases de données spécialisées telles que les signaux de transcription (TRANSFAC), des récepteurs couplés aux protéines G (GCRdb) et bien d'autres. Dans cette section sont présentées les principales bases de données spécialisées dans les informations du métabolisme. Il existe d'une part des bases répertoriant un ou plusieurs types d'acteurs du métabolisme tels que ENZYME, BRENDA et LIGAND, et d'autre part, des bases de connaissances sur le métabolisme de nombreux organismes telles que KEGG, WIT, EcoCyc, MetaCyc et UMBBD.

3.2.3.1 Les bases enzymatiques

- **ENZYME**

ENZYME (3) est une base de données répertoriant les enzymes selon la nomenclature EC de l'IUBMB (International Union of Biochemistry and Molecular Biology). ENZYME est une base sous forme de fichiers texte dans un format correspondant à celui de Swiss-Prot et TrEMBL. Pour chaque numéro EC correspondant au champ ID, une liste d'informations sur l'enzyme telles que le ou les différents noms connus pour cette enzyme, l'activité catalytique décrite par la réaction enzymatique, les cofacteurs, des liens vers les entrées Swiss-Prot ayant cette activité indiquée dans le champs DE, est donnée. La version 30 de la base de données ENZYME de mars 2003 contient 4.136 numéros EC différents.

- **BRENDA**

BRENDA (BRaunschweig Enzyme DAtabase) (5) était à l'origine, en 1987, une collection sous format texte d'informations moléculaires et fonctionnelles sur les enzymes. Depuis quelques années ces informations sont accessibles via un système de base de données relationnel par le web (http://www.brenda.uni-koeln.de/). La version de BRENDA en mars 2003 contient 3.973 numéros EC différents. BRENDA contient de nombreuses informations pour chaque EC, les mêmes informations que la base ENZYME : et des informations complémentaires sur les différents substrats et cofacteurs possibles pour la réaction effectuée par l'enzyme. Par ailleurs, elle contient de nombreuses informations issues de références

bibliographiques sur les différents paramètres de la réaction (KM, activité, pH, température), sur la structure de l'enzyme (liens vers la PDB, poids moléculaire, les différentes sous unités selon les espèces, …), les propriétés moléculaires de la réaction (conditions de stabilité de la réaction, purification, …). Toutes ces informations sont extraites manuellement de références bibliographiques.

- **LIGAND**

LIGAND est une base de données de composés chimiques et de réactions enzymatiques (4). Elle comprend trois sections : COMPOUND, REACTION et ENZYME. La section COMPOUND répertorie les métabolites et les composés chimiques qui peuvent être aussi des toxines ou des polluants environnementaux. La section REACTION répertorie toutes les relations substrats - produits entre les métabolites ou composés chimiques. La section ENZYME contient toutes les enzymes présentées selon la nomenclature EC de l'IUBMB. Cette base est la source de données pour la construction des voies métaboliques de KEGG. Elle est disponible en fichier texte dans un format similaire à celui de GenBank. Un grand nombre d'informations est donné pour chaque entrée d'ENZYME, COMPOUND et REACTION, entre autres de nombreux liens entre ces 3 sections, mais aussi avec les autres sections de la base de données KEGG telles que les informations sur les voies métaboliques. La version de mars 2003 de la section ENZYME de la base de données LIGAND contient 4 171 numéros EC différents.

3.2.3.2 Les bases de connaissances métaboliques

- **UM-BBD**

UM-BBD (University of Minnesota Biocatalysis/Biodegradation Database) (65) contient des informations annotées sur les enzymes microbiennes et leur intégration dans des voies métaboliques. En mars 2003, la base de données comporte 530 enzymes intégrées dans 131 voies métaboliques. Ces voies sont des voies de synthèse et de dégradation de composés organiques et récemment ont été ajoutées des voies de transformations de métaux et métalloïdes. Par ailleurs, il est possible de prédire la voie métabolique de transformation d'un composé en un autre composé en choisissant à chaque étape une réaction parmi plusieurs proposées à partir des données de la base.

- **EcoCyc et MetaCyc**

EcoCyc (66, 67) est une base de connaissances créée en 1996 conjointement par le groupe de Monica Riley et le groupe de Peter Karp. Cette base est consacrée uniquement à la souche K-12 d' *E. coli* considérée comme un organisme modèle car un grand nombre de ses séquences ont une fonction attribuée expérimentalement. Cependant, dans EcoCyc en juin 2003, 28 % des séquences n'ont pas de fonction. A ce jour, cette base contient des données de gènes, de protéines, de voies métaboliques, de fonctions de transport et de la régulation de l'expression des gènes. La base est construite avec un schéma de données orienté objet et est représentée grâce au logiciel Pathway Tools. Ce logiciel permet d'avoir une vue globale du métabolisme de l'organisme, de faire des requêtes par nom de gène, de protéine, de numéro EC, de composé, de voie métabolique, de chromosome. Cette base est très utilisée pour inférer des fonctions à de nouvelles séquences, du fait de la bonne validation des fonctions des gènes d' *E. coli*, mais aussi pour tester différents types d'algorithmes permettant de détecter des interactions protéines-protéines, d'inférer des réseaux génétiques ou des données d'expression. Récemment une nouvelle base de données a été créée : MetaCyc (67, 68). Cette dernière contient les données du métabolisme pour un grand nombre d'organismes de différents phylum : microorganismes, plantes et homme. Comme pour EcoCyc, les données métaboliques sont issues de recherches bibliographiques. Les voies métaboliques pour les différents organismes sont obtenues grâce à un algorithme de prédiction de voies à partir d'un génome annoté appelé PathoLogic (69). Cet algorithme sera décrit dans la partie « méthodes de prédictions des voies métaboliques ».

- **KEGG**

La base de données KEGG (Kyoto Encyclopedia of Genes and Genomes) (70) créée en 1995 est composée de trois bases de données principales : PATHWAY décrivant les voies métaboliques et des complexes protéiques, GENES contenant les gènes et protéines des génomes complets et en cours de séquençage, LIGAND contenant les composés et les réactions chimiques. Par ailleurs, la base KEGG comprend des résultats de profils d'expression issus de microarray dans la base EXPRESSION et de systèmes double hybride chez la levure dans la base BRITE.

La base SSDB répertorie des liens fonctionnels entre protéines. Elle est construite en réalisant des comparaisons de toutes les paires de séquences protéiques de la banque GENES

avec le programme SSEARCH. Toutes les paires avec un score Smith-Waterman supérieur à 100 sont répertoriées. D'autre part les relations 'best-best hit' sont recherchées, c'est-à-dire lorsqu'un gène x d'un génome A a la meilleure homologie avec le gène y du génome B et que la relation est réciproque. Les graphes de relations obtenus sont à la base de la construction d'un tableau représentant des groupes de gènes orthologues (KEGG Orthology). Ces groupes d'orthologues sont aussi déduits d'une comparaison au niveau des voies métaboliques et au niveau des localisations chromosomiques. Ainsi, il est recherché si tous les membres d'un groupe appartiennent à une même voie ou complexe moléculaire ou bien s'ils sont localisés physiquement sur une même portion de chromosome. Ces informations sont utilisées pour compléter l'annotation des séquences de la base GENES. En avril 2003, GENES contient 7 génomes complets eucaryotes, 100 bactéries, 16 Archées et 352 génomes en cours de séquençage.

La base PATHWAY représente les réseaux d'interactions de protéines par des graphes dessinés manuellement. Dans ces graphes sont représentées d'une part les relations enzymes-enzymes par des voies métaboliques et d'autre part d'autres relations protéines–protéines telles que des liaisons, des phosphorylations, des interactions pour la régulation de l'expression des gènes. En avril 2003, 155 graphes se trouvent dans PATHWAY. Dans le cas des graphes des voies métaboliques, l'ensemble des informations du métabolisme de plusieurs espèces est représenté sur un même graphe. Les résultats de l'annotation de la base GENES pour un organisme peuvent être représentés sur un graphe en coloriant les nœuds (les enzymes) du graphe.

Toutes les données de la base KEGG peuvent être recherchées grâce au système DBGET/LinkDB (49). DBGET/LinkDB est un système de gestion de base de données réalisé par l'équipe de KEGG qui permet d'intégrer toutes les bases de données qui composent KEGG ainsi que d'autres bases sous forme de fichiers texte. Elle repose sur la conception de relations binaires entre les entrées d'une ou de différentes bases de données. LinkDB est la base de données contenant toutes les relations binaires obtenues par les liens annotés dans chaque base mais aussi par des liens de similarité calculés par des recherches BLAST et FASTA. DBGET est un système d'extraction des données avec deux commandes bfind et bget pour chercher et extraire les entrées dans plusieurs bases de données.

3.2.3.3 Discussion sur les bases spécialisées dans le métabolisme

La comparaison du contenu des trois bases de données KEGG, ENZYME et LIGAND n'est pas facile du fait de l'existence de différents synonymes pour décrire les substrats des réactions. Des travaux d'analyse des voies métaboliques en utilisant des graphes de PETRI (71) en unifiant le contenu de ces trois bases de données, permettent de mettre en évidence des différences. Ainsi la base BRENDA contient plus de réactions et de substrats car elle contient tous les substrats possibles pour des réactions avec différentes spécificités de substrats alors que les bases ENZYME et LIGAND ne contiennent que le nom générique de la famille de substrats.

Parmi les trois bases de données sur les enzymes présentées, la base de données BRENDA est la plus riche en informations mais elle possède un peu moins de numéros EC répertoriés par rapport aux deux autres bases de données. Les bases de données ENZYME et la section ENZYME de LIGAND sont assez similaires par le type d'informations données sur chaque enzyme. Cependant les annotations EC des séquences n'ont pas la même origine. En effet, dans ENZYME les séquences données pour un numéro EC proviennent des résultats de l'annotation par les banques Swiss-Prot et TrEMBL. Alors que les séquences proposées par la section ENZYME de LIGAND proviennent de différents résultats d'analyse de séquences obtenus par le groupe de KEGG. Il manque des informations sur l'origine de l'annotation (expérimentale ou prédiction) et des indications qualitatives sur l'homologie pour pouvoir interpréter les résultats. Or on observe des informations de séquences pouvant être le résultat de surprédictions par la méthodologie de KEGG.

Par ailleurs, la couverture des séquences génomiques par ces bases de données n'est pas parfaite. En effet, la version d'ENZYME d'octobre 2001 présente 3870 numéros EC parmi lesquels seulement 1662 numéros EC ont une information de séquences provenant de Swiss-Prot (43%). Alors qu'avec la récupération des informations de numéros EC supplémentaires et des annotations correspondant à des enzymes sans numéros EC, on trouve par exemple 250 gènes en plus pour la levure. Il est certain que l'information de séquence associée à la nomenclature EC n'est pas encore à jour mais c'est un long travail d'annotation et de vérification si l'on veut obtenir une nomenclature et des bases de données de qualité.

Le principal apport des bases de données de voies métaboliques est la représentation des informations biologiques sous forme de graphes. Deux types de représentation graphique sont observés, soit les voies sont dessinées manuellement comme dans KEGG, soit automatiquement et selon l'utilisateur comme dans EcoCyc. La difficulté de la représentation

des voies est de pouvoir représenter un maximum d'informations tout en restant lisible par l'œil humain.

Des erreurs sont observées dans ces différentes bases de données dues à l'utilisation de noms variés du fait du manque de nomenclature pour les noms de gènes, de protéines, de description de la fonction. Mais aussi un manque de classification des voies métaboliques, chaque base de données a choisi sa propre classification. Ainsi, il est difficile de comparer ces bases de données de voies métaboliques autrement que par la nomenclature EC de l'IUBMB (72).

Pour interpréter la complexité du métabolisme, des méthodologies (présentées dans la partie « Reconstruction des réseaux métaboliques à partir des bases de données métaboliques » de l'introduction) ont été développées afin de représenter toutes les informations du métabolisme de façon dynamique et permettre l'analyse des voies métaboliques.

4 Les méthodes de prédiction de fonction des génomes

4.1 Les méthodes d'annotation globale d'un génome

4.1.1 Définition de l'annotation d'un génome

L'annotation d'un génome est un processus de détermination de la structure et des fonctions des gènes, protéines ou tout autre élément structuré dans un génome (73). Ce processus est effectué sur la base de la similarité de séquences avec des séquences déjà caractérisées dans des banques. Il existe trois manières d'annoter un génome : soit manuellement par des annotateurs utilisant des logiciels d'analyse de séquences, soit de façon semi-automatique par des annotateurs avec des environnements présentant des synthèses de résultats obtenus avec les logiciels, soit de façon automatique par des programmes. Quelques exemples de ces trois manières d'annoter un génome sont présentés avec leurs avantages et inconvénients.

4.1.2 Annotation manuelle

La plupart des premiers génomes complètement séquencés ont été annotés manuellement par une équipe d'annotateurs en effectuant une recherche d'homologie pour chaque gène avec une banque de séquences généraliste. Par exemple le génome complet de la bactérie pathogène *Chlamydia trachomatis* a été annoté manuellement en collaboration avec 9 laboratoires. Le génome a été réparti entre les 9 groupes pour l'annotation mais celle-ci a été vérifiée par l'ensemble des équipes (74). Ainsi, l'annotation a été réalisée au moins 3 fois par des personnes différentes. Chaque séquence du génome a été utilisée comme requête du programme PSI-BLAST contre une banque de séquences non redondante. L'analyse du résultat et l'assignation d'une fonction à la séquence ont été réalisés par l'annotateur.

Le principal avantage de l'annotation manuelle est la qualité des annotations obtenues même si des séquences peuvent être mal annotées (faux positifs), des séquences peuvent être manquées (faux négatifs) et des erreurs de dénomination de la fonction sont observées. Mais

le principal inconvénient de l'annotation manuelle est la lenteur du processus étant donné le nombre toujours croissant de génomes séquencés.

4.1.3 Annotation semi-automatique

Différents types de systèmes d'annotation semi-automatique existent : certains favorisent les résultats de méthodes d'analyses de séquences comme dans le projet HAMAP (55), d'autres un environnement graphique d'annotation proposant la synthèse de différentes méthodes comme l'environnement iANT (75). Il existe aussi des environnements plus complexes intégrants les données et les méthodes dans un modèle orienté objet tel que Imagene (76).

Le groupe Swiss-Prot est en train de développer un projet d'annotation semi-automatique des génomes microbiens : HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes) (55). Une annotation automatique des séquences sans similarité de séquences est réalisée par la recherche de caractéristiques telles que des séquences signal, des domaines transmembranaires, des sites de liaison, … De même une annotation automatique des séquences appartenant à une famille connue et décrite par une règle est réalisée (77). Par ailleurs, des protéines exceptionnelles par rapport à une famille, c'est-à-dire avec plusieurs caractéristiques communes à la famille mais avec une caractéristique singulière, sont examinées par un expert pour une annotation manuelle.

L'environnement d'annotation iANT intègre un ensemble d'outils d'analyse de séquences nucléiques et de séquences protéiques (75). L'interface sous forme de documents hypertexte permet l'exécution des programmes, l'inspection et l'édition des résultats. La séquence ADN annotée est représentée par une carte cliquable pour chaque élément de la séquence (protéines, tRNA, rRNA, etc…) et permet ainsi à l'annotateur de voir l'environnement de chaque objet et influencer son annotation qui est toujours modifiable.

L'environnement Imagene est basé sur une intégration des données, qui sont les séquences d'un génome et des méthodes d'analyse de séquences dans un modèle orienté objet, et comprend aussi une interface graphique (76). L'interface graphique permettant à l'annotateur de lier les différents objets afin de produire l'annotation des séquences et permettant aussi de décider de sa stratégie d'annotation. Celle-ci consiste à choisir un ensemble de tâches à effectuer c'est-à-dire un ensemble de méthodes.

Chacun de ces environnements présente des avantages. Le point fort d'HAMAP est méthodologique, il permet de caractériser des séquences difficiles à annoter. L'avantage d'iANT est graphique, l'interface très synthétique permettant aux biologistes de faire une annotation experte et rapide. Le point fort d'Imagene est sa flexibilité, chaque annotateur pouvant décider de sa stratégie selon les différents objets à annoter. Tous ces environnements d'annotation semi-automatique requièrent encore beaucoup de temps pour annoter un génome complet mais permettent en général d'éviter les erreurs humaines.

4.1.4 Annotation automatique

De nombreux systèmes d'annotations automatiques se sont développés comme Ensembl (1), GeneQuiz.

Le système GeneQuiz permet une annotation fonctionnelle automatique de séquences protéiques (78). Les bases de données utilisées sont mises à jour régulièrement de façon automatique. Un grand nombre de méthodes sont utilisées automatiquement pour chaque séquence : filtrage des séquences, recherche de similarité BLAST et FASTA, recherche de répétitions, de motifs, prédiction d'hélices transmenbranaires, de structures, etc.… La principale différence par rapport à une annotation semi-automatique est que les fonctions des séquences sont inférées automatiquement grâce à un algorithme de transfert de la fonction, à partir d'un lexique de mots clés. Ces mots clés sont ceux observés le plus souvent dans le champ de description des séquences homologues obtenues avec les méthodes d'analyse de séquences. Un système automatique comme GeneQuiz permet une annotation rapide des génomes complets et de répéter celle-ci au fur et à mesure des mises à jour des données de séquences. Cependant, ce système est très peu flexible et ne permet donc pas de détecter les cas exceptionnels.

4.1.5 Discussion

L'analyse de ces différentes stratégies d'annotation, permet de mettre en évidence qu'elles contiennent des erreurs d'annotations d'origines variées. La principale cause de ces erreurs est certainement le manque d'ontologie de la fonction biologique. Or celle-ci est nécessaire pour effectuer une re-annotation des génomes à partir des nouvelles données qui sont disponibles de jour en jour.

4.1.5.1 Les erreurs d'annotation

Il est primordial d'éviter les erreurs d'annotation car elles sont intégrées dans les banques de données et deviennent alors des sources de propagation d'erreurs pour l'annotation des génomes suivants.

Un certain nombre d'origines de ces erreurs ont été référencées (79). D'une part, on trouve des erreurs au niveau de l'ADN comme les erreurs de séquençage, de l'ordre de 0,1% et les erreurs de prédiction des régions codantes. En effet, les programmes de prédiction de gènes ne sont efficaces qu'à 60-70% pour les génomes eucaryotes et à 90% pour les génomes procaryotes. D'autre part, on observe des erreurs d'annotation de description de la fonction des protéines, obtenues à partir d'un mauvais choix des méthodes automatiques ou obtenues suite à des erreurs d'interprétation humaine. Ainsi, la comparaison des annotations du génome de *Mycoplasma genitalium* par trois groupes différents met en évidence 8% de désaccords importants (80). Ces erreurs d'annotation sont dues le plus souvent, à des attributions de mauvaises fonctions, d'une part à partir d'un résultat de faible similarité de séquences et d'autre part à des alignements partiels sur des régions non significatives pour la détermination de l'annotation (81). Une méthodologie permettant de détecter des relations d'homologie éloignées est l'utilisation de profils (82). Peu de méthodologies d'annotation utilisent ce principe pour annoter les génomes alors que la construction de plusieurs bases de données comme ProDom, Pfam, etc… reposent sur leur utilisation. Ainsi, l'environnement d'annotation semi-automatique iANT utilise les résultats d'homologies avec la base de données ProDom, PROSITE.

Enfin les erreurs d'annotation des fonctions sont aussi dues à un manque de nomenclature. On observe des gènes de familles non homologues avec les mêmes noms, du fait de l'utilisation d'un grand nombre de synonymes pour décrire un seul gène. Ce manque de nomenclature provient de la difficulté à créer une bonne classification des fonctions.

4.1.5.2 Nécessité d'ontologie

Afin de faciliter la classification des fonctions, il est nécessaire de définir une ontologie de la fonction biologique (83). En effet l'ontologie est une description formelle de concepts et de leurs relations dans un domaine d'intérêt. Or, il y a différents niveaux de définition d'une fonction biologique : la fonction biologique locale, au niveau des interactions d'une molécule avec d'autres molécules ; la fonction biologique intégrée, c'est-à-dire la contribution de la molécule dans le fonctionnement de l'organisme. Un modèle de base de

données comme EcoCyc propose une ontologie de la fonction biologique locale. Une ontologie définitive de la fonction biologique permettra de l'utiliser pour de nombreuses applications comme la ré-annotation des génomes. D'ailleurs l'EBI développe le projet 'Gene Ontology Annotation' (GOA) (84) afin d'intégrer les termes de vocabulaire récemment déterminé par le consortium d'ontologie de gènes (GO) (85), dans plusieurs de ses bases de données telles que Swiss-Prot, TrEMBL et InterPro.

4.1.5.3 La ré-annotation des génomes

De nombreux génomes annotés dans les années 1990 sont maintenant ré-annotés (86, 87). La ré-annotation des génomes permet de tester la performance de nouvelles méthodes mais permet aussi d'utiliser les dernières données pour trouver de nouvelles fonctions manquées dans les analyses précédentes (73). La comparaison de méthodes d'annotation de génomes est difficile de part l'évolution rapide des données et la difficulté de comparaison des annotations.

4.2 Les méthodes de génomique comparée

Différentes méthodologies utilisant la comparaison de nombreux génomes ont été développées afin de prédire les fonctions peu caractérisées expérimentalement. Les premières méthodologies reposent sur l'identification de groupes de gènes orthologues possédant souvent une même fonction. D'autres méthodologies reposent sur l'observation que certains gènes localisés dans une même région dans plusieurs génomes possèdent une fonction impliquée dans un même processus biologique. Enfin, des méthodologies se basent sur l'existence de gènes fusionnés dans certains génomes et codant pour des protéines qui interagissent entre elles.

4.2.1 Groupes de gènes orthologues et profils phylogénétiques

L'identification de familles de protéines hautement conservées chez les bactéries et manquantes chez les eucaryotes permet de mettre en évidence des candidats pour des cibles anti-bactérienne potentielles.

Les familles de gènes ou protéines homologues comprennent à la fois des gènes orthologues et des gènes paralogues. Les gènes paralogues sont issus de la duplication d'un gène dans un génome et permettent la création de nouvelles fonctions. Les gènes orthologues dans

différentes espèces sont issus de la spéciation d'un gène ancestral et peuvent posséder la même fonction. C'est pourquoi il est nécessaire d'identifier les groupes de gènes orthologues pour prédire une fonction biologique.

Des méthodologies comme COG (Clusters of Orthologous Group) déterminent des groupes de gènes orthologues dans trois ou plusieurs lignées phylogénétiques (88). Elles nécessitent la comparaison de toutes les paires de séquences de génomes complets. Pour chaque protéine, le meilleur match (best hit) dans chaque autre génome est identifié et permet de construire un graphe des relations « best hit » c'est-à-dire des meilleurs matchs de chaque comparaison. Un COG minimal est un triangle permettant de relier trois phylums différents. Tous les triangles avec une arrête commune sont joints. Ainsi les protéines de différents phylums dans un groupe sont orthologues. Cependant cette méthode échoue dans le cas de protéines multidomaines. C'est pourquoi, les domaines sont isolés et utilisés pour effectuer une deuxième itération du processus. Certains gènes paralogues avec de fortes similarités sont quand même trouvés dans quelques COG. La distribution phylogénétique des COG peut être représentée par un profil phylogénétique. Un profil phylogénétique est un motif ou un vecteur présentant l'absence ou la présence d'un orthologue dans chaque espèce considérée. Des équipes comme celle de Pellegrini et Marcotte utilisent la comparaison de ces profils phylogénétiques pour la prédiction de fonction (89). Ainsi, ils ont observé que certaines protéines non homologues présentant un profil phylogénétique similaire, sont fonctionnellement liées. Cependant, l'utilisation de cette méthodologie produit environ 30% de faux positifs (90).

4.2.2 Contexte des gènes

Une autre méthodologie de génomique comparée permettant de prédire des fonctions est la recherche de contexte des gènes (91, 92). Elle repose sur le concept que l'ordre conservé de certains gènes au cours de l'évolution est corrélé à des interactions physiques entre les protéines que codent ces gènes. C'est le cas pour les opérons chez les procaryotes ou des paires de gènes conservées dans plusieurs génomes. L'équipe de Dandekar met en évidence que 75% des paires de gènes trouvées par cette méthode correspondent expérimentalement à des interactions physiques (91).

4.2.3 Fusion de gènes ou pierre de rosette

La méthode nommée « pierre de rosette » ou fusion de gènes est aussi basée sur la recherche du contexte des gènes. L'hypothèse est que des gènes distincts dans certains génomes et trouvés fusionnées dans d'autres, interagissent physiquement. Ainsi, en recherchant les fusions de gènes orthologues il est possible de prédire des interactions. Cette méthode semble être très efficace mais présente une faible occurrence dans les génomes (93, 94). Cette méthode a été étendue en utilisant les paralogues mais bien qu'elle permette de prédire 26% de fonctions connues, elle produit aussi environ 36% de faux positifs (90, 95).

4.2.4 Conclusion

Les méthodes d'annotation globale d'un génome peuvent être complétées par l'utilisation de méthodes de génomique comparée comme les profils phylogénétiques, la localisation de gènes, les fusions de domaines qui permettent de trouver des liens entre des protéines non homologues. Il semble que chacune des méthodes de génomique comparée utilisée seule ne soit pas très efficace (peu sensibles et peu spécifiques). Par contre la combinaison de plusieurs méthodes semble améliorer l'efficacité. Les résultats de ces méthodes doivent être confirmées par des expérimentations, telles que des co-expressions de gènes sur des micro-array ou des expériences de protéomique.

5 Les méthodes de prédiction des voies métaboliques

Une voie métabolique est une suite de réactions enzymatiques transformant un ou plusieurs substrats en produits. Le résultat d'une voie métabolique peut être l'assimilation de composés, la production d'autres composés ou d'énergie, tout ceci dans le but de permettre la croissance de l'organisme. On distingue deux grands types de voies métaboliques : les voies cataboliques qui dégradent les molécules complexes en générant souvent de l'énergie, et les voies anaboliques qui utilisent de l'énergie pour la synthèse de molécules complexes. Ainsi, la glycolyse est la voie métabolique de dégradation du glucose en 2 molécules de pyruvate avec formation d'énergie sous forme de 2 molécules d'ATP.

L'analyse des voies métaboliques est un champ de recherche intéressant pour plusieurs raisons. Tout d'abord, les méthodes de prédiction de fonction des gènes permettent d'identifier l'ensemble des acteurs possibles du métabolisme d'un organisme à partir de son génome complet. Ceci est réalisable avec l'utilisation des bases de données de connaissances sur le métabolisme. Des méthodologies identifient les relations entre les fonctions enzymatiques prédites pour reconstruire dynamiquement le réseau global du métabolisme (Pathologic, PathFinder, etc…). Ainsi, avec le grand nombre de génomes complets disponibles, il est possible de comparer les voies métaboliques de différentes espèces et trouver des voies spécifiques de certains organismes.

Deuxièmement, la modélisation des voies est très utile pour guider l'ingénierie métabolique de cellules en maximisant la production d'un substrat spécifique. Ainsi, des méthodologies analysent la structure du réseau global. Elles permettent, par exemple, d'extraire toutes les voies fonctionnelles possibles ou de trouver la voie la plus courte pour aller d'un substrat à un produit. Ces méthodologies sont basées sur la connectivité des enzymes ou sur des principes mathématiques tels que l'algèbre linéaire et l'analyse convexe.

Cependant, le métabolisme est un réseau complexe d'interactions soumises à des stimuli. C'est pourquoi il est intéressant de pouvoir combiner les résultats d'analyses d'expressions de gènes ou de protéomiques et des méthodologies bio-informatiques pour déterminer les sous-réseaux exprimés dans différentes conditions physiologiques.

5.1 Reconstruction des réseaux métaboliques

L'annotation des génomes peut être améliorée en intégrant les fonctions prédites dans des réseaux métaboliques. Des systèmes comme KEGG, UMBBD proposent la représentation des annotations par des visualisations statiques des données du métabolisme. Des méthodes comme Pathologic et PathFinder proposent des représentations plus ou moins dynamiques du métabolisme à partir de données d'annotations.

5.1.1 PathoLogic

PathoLogic est un logiciel transformant les annotations d'un génome de microorganisme en une base de données de ses voies métaboliques, en utilisant comme référence la base de données EcoCyc ou MetaCyc (96). Cette méthode comprend à la fois des étapes automatiques et des étapes interactives avec l'utilisateur. L'algorithme implémenté dans PathoLogic extrait les informations de fonctions enzymatiques de fichiers au format Genbank, tels les numéros EC, les noms de produit de gène. Une vérification de la cohérence des ces annotations est réalisée et permet de créer une liste d'enzymes pour un génome. Ensuite, deux étapes interactives permettent tout d'abord à l'utilisateur d'identifier les complexes protéiques dans cette liste d'enzymes puis d'ajouter des voies métaboliques spécifiques de l'organisme, grâce à des éditeurs graphiques. Enfin, l'existence des voies est vérifiée par le calcul d'un score prenant en compte le nombre d'enzymes trouvées dans l'organisme sur le nombre total d'enzymes dans la voie métabolique. Des graphes représentent les voies métaboliques spécifiques de l'organisme. Cet algorithme a comme avantage la possibilité de détecter des enzymes mal annotées : numéro EC oublié, nom d'enzyme atypique. Il a été utilisé sur un grand nombre de microorganismes qui sont répertoriés dans la base de données MetaCyc. La comparaison de cette méthodologie sur le genome d'*Helicobacter Pylori* (69) avec l'annotation et la reconstruction manuelle de ses voies métaboliques met en évidence son efficacité.

5.1.2 PathFinder

PathFinder (97) est une méthode automatique pour la représentation dynamique des voies métaboliques à partir de données d'annotation qui peuvent être : soit une liste d'EC, soit des annotations au format EMBL ou Genbank. Les données métaboliques issues de la base de

données KEGG sont représentées par des graphes acycliques orientés, grâce à un logiciel d'agencement de graphe, VCG-tool, adapté à la visualisation des voies métaboliques. Par ailleurs, un algorithme permet d'identifier dans un graphe les plus petites suites de réactions sans branchements, appelées « chunks » et de connaître le pourcentage de réactions réalisées dans un chunk.

5.2 Analyse de la structure des réseaux métaboliques

Une analyse quantitative des réseaux métaboliques n'est pas réalisable pour le moment à cause du nombre limité d'informations sur les cinétiques des enzymes. Cependant, il est possible de faire une analyse qualitative des réseaux métaboliques avec les données de génomes complets, en réalisant une approche structurale par l'analyse de la topologie des voies métaboliques. L'analyse stœchiométrique des voies métaboliques pose un certain nombre de problèmes tel que : l'identification des flux possibles dans un système en fonction d'un ensemble d'enzymes caractérisées, de substrats disponibles, ou l'évaluation du sous-ensemble d'enzymes nécessaires pour réaliser un flux particulier dans le réseau.

Etant donné un réseau métabolique, il est intéressant de connaître à partir de la connectivité des métabolites, tous les chemins possibles entre un métabolite A et un métabolite B et le plus court chemin entre eux. Ces problèmes sont résolus avec des méthodes de théorie des graphes ou d'algèbre linéaire. En effet, un réseau métabolique peut être décrit par un graphe dirigé, dont les nœuds sont les métabolites et les arrêtes les réactions enzymatiques.

5.2.1 Analyse de la structure des réseaux métaboliques avec les graphes

Il est possible d'exploiter les connaissances disponibles dans les bases de données en les intégrant dans des graphes. Cette représentation permet de prédir la fonction et d'interpréter des données d'expression pour des génomes complets.

Küffner et al (71) ont choisi de rassembler les informations de différentes bases de données métaboliques (BRENDA, ENZYME et KEGG) dans un graphe appelé "PETRI net". Ce graphe représente toutes les relations et interconnections entre protéines. Puis, un algorithme génère et recherche les voies et les réseaux satisfaisant certaines contraintes. Un "Metabolic Display" (MD) est un graphe présentant toutes les voies possibles entre une source et un produit final selon un jeu de contraintes définies. D'autre part, un "Differential

Metabolic Display" (DMD) permet de visualiser et de comparer les réseaux de deux systèmes différents tels que deux organismes, deux tissus ou deux états différents.

Ainsi, l'utilisation des graphes de PETRI appliqués aux bases de données métaboliques permet:

- de représenter les différents contenus des bases de données,
- de comparer les informations génomiques et les connaissances sur les réseaux,
- de déterminer et analyser les réseaux,
- de définir la notion de DMD permettant de comparer différents systèmes.

Cette méthode a été utilisée pour évaluer des données d'expression de gènes (98). A partir de réseaux connus, une extraction de voies possibles est effectuée suivant certaines contraintes, puis un examen est réalisé pour vérifier si ces voies sont supportées par des données d'expression. Des fonctions de scores ont été définies pour 3 propriétés des voies putatives: pour des schémas d'expression de gènes impliqués dans les voies, pour des schémas de données d'expression synchronisés, et enfin une combinaison des deux. Le calcul de scores de voies et de scores de gènes permet de montrer quelles sont les voies ayant les propriétés recherchées et comment chaque gène a la possibilité d'appartenir à cette voie.

Ma et Zeng (99) ont aussi développé une méthode pour représenter et analyser la structure des voies métaboliques à partir des données génomiques de 80 organismes différents. A partir des données des bases KEGG et LIGAND, une base de données de réactions enzymatiques est extraite en effectuant des corrections pour des incohérences observées entre les fichiers et des erreurs dans les réactions (entre autre le sens des réactions). Après corrections, 1969 réactions irréversibles parmi 4772 réactions sont obtenues.

Les réactions sont reliées à des enzymes. La relation enzyme – réaction n'est pas toujours de type un – un, car certaines réactions sont effectuée par plusieurs enzymes et une enzyme peut réaliser différentes réactions. A partir de la section ENZYME de la base LIGAND, des données sont extraites pour la création d'une matrice représentant la relation gène - enzyme pour 80 organismes complètement séquencés. Ainsi à partir des deux types de relations, une matrice de réactions observées est déduite pour 80 organismes.

La représentation du réseau métabolique est réalisée par un graphe dirigé en construisant une liste d'arcs pour les réactions irréversibles et une liste d'arêtes pour les réactions réversibles. Les métabolites courants (ATP, ADP, NADH, NAD+, H2O, P_i) que l'on peut aussi nommer ubiquitaires car ils sont présents dans de nombreuses réactions

enzymatiques, sont enlevés des listes d'arc et d'arêtes. En effet, si l'on calcul la longueur moyenne des voies en conservant ces métabolites courants, on obtient des petites voies (longueur moyenne de 3.2) alors que les voies métaboliques en réalité sont plus longues. Les métabolites courants définis pour l'ensemble du réseau correspondent aux métabolites externes définis pour un sous-réseau dans le travail de S. Schuster (voir sous-partie suivante).

En calculant la connectivité des métabolites sans considérer les métabolites courants, ils démontrent que le réseau est un réseau petit monde. Les métabolites identifiés avec le plus haut degré de connectivité sont des plaques tournantes dans la plupart des organismes. Ce sont par exemple, des intermédiaires de la glycolyse, de la voie des pentoses, l'acetyl-CoA qui est un métabolite permettant la liaison de la voie de la glycolyse avec le cycle de Krebs, et la voie de synthèse des acides gras. Par contre, les résultats de calculs sur la structure du réseau diffèrent de ceux de Jeong (100). La longueur moyenne des voies de 80 organismes est variée. Ils mettent ainsi en évidence que les bactéries ont en moyenne des voies plus courtes que les eucaryotes et les archées, plus particulièrement pour des bactéries parasites qui ont certainement perdues de nombreux gènes en s'adaptant à leurs hôtes au cours de l'évolution. Ce type d'analyse permettrait d'identifier les courts-circuits utilisés par les bactéries dans certaines voies métaboliques. Les réactions impliquées dans ces voies seront déterminées et utilisées soit pour améliorer la croissance de bactéries facilitant la croissance d'autres organismes, soit pour identifier des fonctions pathogènes pour le développement de nouveaux médicaments.

5.2.2 Analyse de la structure des réseaux métaboliques par la recherche des modes élémentaires

En réalité une cellule à un moment donné est sujette à des contraintes limitant son comportement. Il existe un espace de solutions de voies métaboliques possibles pour un organisme. Cet espace peut être réduit avec la connaissance d'un certain phénotype qui peut-être la liste des enzymes identifiées dans un génome, avec les propriétés biochimiques des composants et avec certaines contraintes imposées par exemple par le milieu de culture.
Des méthodes nommées, analyse de flux (101) ou modes élémentaires (102, 103) cherchent les flux métaboliques ou modes élémentaires possibles pour un génotype métabolique défini. Un mode élémentaire est un ensemble d'enzymes impliquées dans une même fonction

biochimique tel qu'un sous-ensemble des enzymes n'est pas capable de réaliser le mode. La recherche de modes élémentaires comprend 3 étapes.

La 1ère étape consiste à définir le génotype métabolique *in silico* d'un organisme, c'est-à-dire identifier l'ensemble des enzymes détectées dans le génome d'un organisme et établir la liste des réactions et leur sens, correspondant à ces enzymes.

La 2ème étape consiste en la décomposition du grand réseau en sous-systèmes due à l'explosion combinatoire des voies possibles. Ceci est réalisable en se basant sur la connectivité des métabolites permettant de déterminer les métabolites externes et internes aux sous - réseaux. Dans le programme SEPARATOR (104), les métabolites déterminés comme externes sont des métabolites pris ou excrétés par la cellule ainsi que des métabolites participants au moins à un certain nombre de réactions.

La dernière étape consiste à calculer les modes élémentaires. Le programme METATOOL (105) comprend un algorithme qui, à partir de la matrice de stœchiométrie et du sens des réactions, définit un cône convexe et calcule les vecteurs générateurs de ce cône. Les vecteurs générateurs sont les transformations stoechiométriques réalisables entre un substrat et un produit et qui ne sont pas décomposables en d'autres voies.

6 Présentation du sujet de thèse

De nombreux programmes de séquençage de génomes de différents organismes sont en cours ainsi que des expériences sur le transcriptome et le protéome correspondants. Il est nécessaire d'avoir des méthodes automatiques pour annoter tous ces génomes et intégrer les fonctions prédites dans des réseaux comme les voies métaboliques.

Mon travail de thèse a consisté dans un premier temps, à concevoir et développer une méthode automatique et générique pour la prédiction des fonctions enzymatiques d'un génome complet. Cette méthode a été nommée PRIAM (PRofils pour l'Identification Automatique du Métabolisme, Chapitre II). La représentation des résultats sur les graphes des voies métaboliques de la base de données KEGG facilite l'analyse des résultats de cette méthodologie pour la prédiction des voies métaboliques d'un génome complet.

Dans un second temps, cette méthodologie a été appliquée au génome complet de la bactérie *Sinorhizobium meliloti*, récemment séquencé et annoté (106), entre autre par le laboratoire des Interactions Plantes Microorganismes à l'INRA de Toulouse. J'ai pu ainsi valider l'utilisation de cet outil et participer à l'annotation de ce génome. Lors de ma dernière année de thèse, j'ai pu interpréter les résultats obtenus *in silico* avec PRIAM et analyser les voies métaboliques présentes et manquantes chez cette bactérie (Chapitre III, Application au génome de *Sinorhizobium meliloti*).

Cette méthodologie a aussi servi à l'analyse et à la présentation de résultats d'expériences du protéome (107) et du transcriptome de *S. meliloti* cultivé dans différentes conditions (Marcela Davalos, communication personnelle).

II

INFERENCE FONCTIONNELLE DU METABOLISME : PRIAM

La méthodologie d'inférence fonctionnelle du métabolisme à partir d'un génome complet, nommée PRIAM (*PRofils pour l'Identification Automatique du Métabolisme*), développée durant cette thèse comprend deux parties.

Elle comprend tout d'abord un programme de construction et d'évaluation statistique d'une banque de profils spécifiques des fonctions enzymatiques. Puis, le développement d'un programme permettant dans un premier temps d'identifier les fonctions enzymatiques dans un génome complet avec la banque de profils, puis de visualiser les résultats sur les graphes des voies métaboliques de KEGG.

Ces programmes ont été développés en langage PERL, ils intègrent d'autres programmes d'analyse de séquences. L'évaluation statistique a été réalisée grâce à l'utilisation du logiciel S-PLUS. La mise en forme des résultats a été réalisée en HTML et PERL-CGI pour les pages web accueillies par le serveur de la génopôle de Toulouse.

1 Les outils informatiques

1.1 Le langage PERL et l'application CGI

Tous les programmes développés pour réaliser la méthode PRIAM et présenter les résultats sont écrits en langage PERL qui signifie : langage pratique d'extraction et de rapport (Practical Extraction and Report Language).

PERL est un langage optimisé pour extraire des informations de fichiers texte et imprimer des rapports basés sur ces informations. Il est aussi utilisé pour les tâches d'administration système. C'est un langage interprété, il suffit de rendre un programme exécutable, il n'est pas nécessaire de le compiler. Les caractéristiques de PERL sont tout d'abord les tableaux de hachage (ou tableaux associatifs) qui permettent d'associer facilement des valeurs à des clés, et d'autre part la recherche d'expressions régulières en utilisant des techniques sophistiquées de recherche de motif pour pouvoir traiter rapidement de grandes quantités de données. Ces deux caractéristiques sont utiles pour le traitement des données de génomique qui comprennent de nombreuses chaînes de caractères et de nombreuses associations de données.

Par ailleurs, des scripts CGI ont été écrit en PERL pour produire des pages HTML dynamiques à partir de la sortie standard. En effet la CGI (*Common Gateway Interface*) est une interface pour relier les applications externes et les serveurs d'informations (comme ceux sur le World Wide Web).

Ainsi, la CGI fonctionne de la manière suivante pour créer une page HTML suite à la saisie d'un certains nombre d'informations par un utilisateur. Le client (navigateur) saisit les données dans la page HTML. Ces données sont envoyées au serveur qui lance l'application CGI. Le résultat est ensuite retourné au client sous forme d'une page HTML.

1.2 Le logiciel S-PLUS

Le logiciel S-PLUS est un environnement payant pour l'analyse statistique de données et la création de graphiques. Le logiciel R est l'équivalent gratuit de S-PLUS.

Le logiciel S-PLUS permet de faire du calcul, de créer des vecteurs qui sont les structures de base pour traiter des données, des matrices, des tableaux de données. Il contient de nombreuses fonctions permettant d'effectuer des méthodes statistiques classiques, de réaliser des graphiques avec des histogrammes par exemple. C'est aussi un langage de programmation de haut niveau, avec des instructions conditionnelles (if), des boucles (for) permettant l'écriture de nouvelle procédures statistiques ou l'automatisation de tâches.
Ce logiciel a été utilisé pour évaluer statistiquement l'efficacité des profils spécifiques d'enzymes contre la banque de données de protéines Swiss-Prot.

1.3 HTML

Le langage HTML (Hyper Text Markup Language) est le langage universel utilisé pour communiquer sur le Web. C'est un langage hyper texte orienté présentation qui permet d'avoir dans une page un mot à cliquer, ce mot est un lien vers une autre page HTML par exemple. Un document HTML est composé de texte et de balises (tag en anglais). Ces balises permettent de mettre en forme le texte (titre, caractère gras, italique, image, liens, etc...). Les balises HTML ont une marque de début et une marque de fin. Les balises HTML utilisent les caractères < et > comme délimiteurs. La balise de fin est précédée d'un /. Ainsi par exemple les balises pour écrire un titre sont: <title>PRIAM</title> Cette information apparaît dans la barre de titre du client WWW.

2 La caractérisation des fonctions enzymatiques par des profils

Notre objectif a été de concevoir une méthodologie de prédiction des fonctions enzymatiques prenant en compte certaines caractéristiques des enzymes, c'est-à-dire l'existence d'enzymes non homologues, de multi-enzymes, d'enzymes oligomériques ou de complexes enzymatiques. C'est pourquoi nous avons choisi de construire des descripteurs caractéristiques des fonctions enzymatiques en se basant sur la nomenclature EC crée par l'IUBMB (2). En effet, à chaque fonction enzymatique ou numéro EC correspond un ensemble de séquences protéiques qui participent à cette fonction sans être nécessairement similaires.

2.1 Le choix de la base de donnée ENZYME

Parmi les différentes bases de données de fonctions enzymatiques étudiées (ENZYME (3), BRENDA (5), LIGAND (4)) et basées sur la nomenclature EC, la base de données ENZYME a été choisie comme source de données pour l'élaboration d'une méthodologie de prédiction de fonctions enzymatiques, ceci pour deux raisons principales.

Tout d'abord, c'est la base de données avec la base LIGAND, qui contient le plus grand nombre de numéros EC répertoriés : par exemple en mars 2003, il y avait 4136 numéros EC dans la base ENZYME et 4171 numéros EC dans la base LIGAND alors qu'il y avait seulement 3973 numéros EC dans la base BRENDA.

D'autre part, nous avons surtout choisi la base de données ENZYME pour la fiabilité des annotations. En effet, toutes les séquences répertoriées dans la base ENZYME sont issues du travail d'annotation réalisé par l'équipe de Swiss-Prot. Alors que les séquences données pour chaque numéro EC dans les bases de données BRENDA et LIGAND proviennent de bases de données de séquences variées et de différentes qualités. De plus les annotations enzymatiques des séquences de la base LIGAND peuvent provenir de résultats d'analyses de séquences automatiques alors que les annotations de la base ENZYME sont issues d'annotations par des experts.

2.2 La construction des profils spécifiques des fonctions enzymatiques

La première partie de la thèse a donc consisté à développer des descripteurs spécifiques des fonctions enzymatiques, nous avons choisi de construire des matrices de scores position spécifique en se basant sur la nomenclature EC.

Ainsi, dans une première étape (figure 6A), nous avons construit des collections de séquences d'enzymes pour chaque numéro EC, puis (figure 6B) nous avons identifié, dans chaque collection, les régions conservées (appelées modules) et sélectionné les régions représentatives de la collection. Enfin, dans une dernière étape nous avons construit des profils ou matrices de scores positions spécifiques (PSSM) pour chaque région conservée.

Par ailleurs, l'efficacité de ces profils a été évaluée statistiquement et une recherche a été effectuée pour trouver une correspondance entre ces régions conservées et les domaines structuraux de la base de données SCOP.

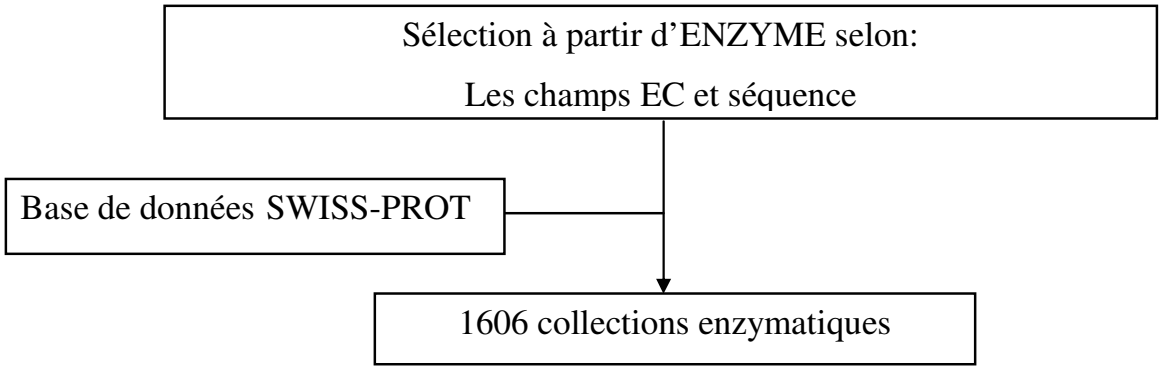


Figure 6A: 1ère étape, construction des collections enzymatiques.

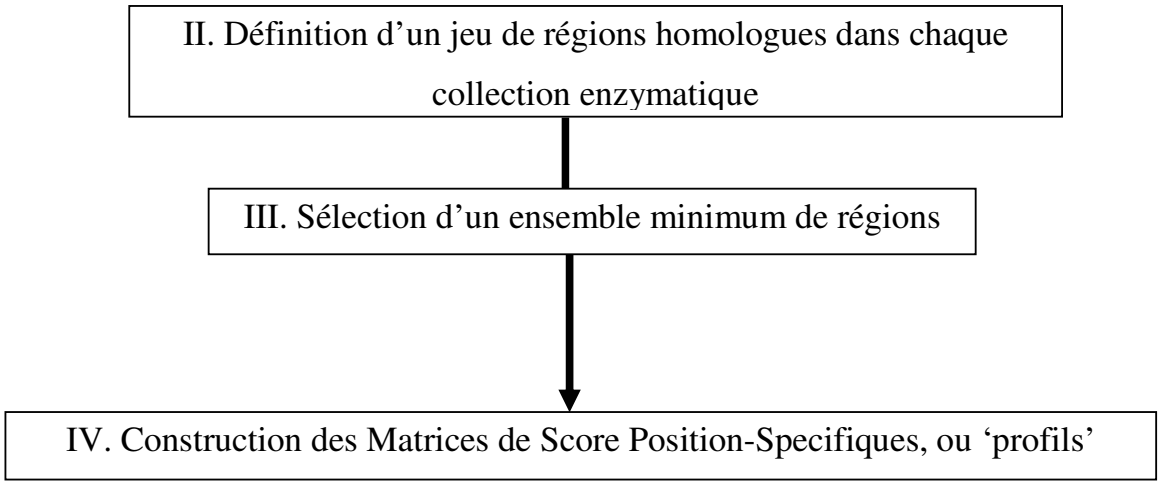


Figure 6B : Étapes suivantes de la construction de la banque de profils spécifiques d'enzymes.

2.2.1 Construction des collections enzymatiques

Nous avons construit des collections de séquences d'enzymes pour diverses activités enzymatiques à partir du champ DR d'une entrée ENZYME (figure 7). Une collection enzymatique est donc un ensemble de séquences protéiques ayant ou participant (pour les enzymes oligomériques et les complexes enzymatiques) à une même fonction enzymatique mais ne possédant pas toujours une similarité de séquences (enzymes analogues, oligomériques et multi-fonctionnelles).

```
//
ID   1.2.1.1
DE   Formaldehyde dehydrogenase (glutathione).
AN   Formic dehydrogenase.
AN   NAD-linked formaldehyde dehydrogenase.
CA   Formaldehyde + glutathione + NAD(+) = S-formylglutathione + NADH.
CC   -!- Some 2-oxoaldehydes are also oxidized.
CC   -!- In the reverse direction, NADPH can replace NADH.
PR   PROSITE; PDOC00058;
DR   P25437, ADH3_ECOLI;  P44557, ADH3_HAEIN;  P39450, ADH3_PASPI;
//
```

Figure 7: Exemple d'une entrée de la base de données ENZYME.

La version 27.0 d'ENZYME a été récupérée par ftp (file transfer protocol) sur le site du serveur expasy (ftp://us.expasy.org/databases/enzyme/). La version 27.0 contient 3870 entrées EC. D'autre part, nous avons installé sur notre machine la version de la base de données de séquences protéiques Swiss-Prot (version 40), contenant 101.602 entrées. Les entrées de la base de données Swiss-Prot (figure 8) ont été formatées au format FASTA en conservant l'information des champs ID (numéro d'identification), AC (numéro d'accession), DE (champ de description de la fonction) et SQ (champ contenant la longueur de la séquence protéique et à la suite la séquence protéique).

```
ID   ADH3_ECOLI     STANDARD;      PRT;   369 AA.
AC   P25437; P75696; Q47533;
DT   01-MAY-1992 (Rel. 22, Created)
DT   01-NOV-1997 (Rel. 35, Last sequence update)
DT   10-OCT-2003 (Rel. 42, Last annotation update)
DE   Alcohol dehydrogenase class III (EC 1.1.1.1) (Glutathione-dependent
DE   formaldehyde dehydrogenase) (EC 1.2.1.1) (FDH) (FALDH).
GN   ADHC OR B0356.
OS   Escherichia coli.
OC   Bacteria; Proteobacteria; Gammaproteobacteria; Enterobacteriales;
OC   Enterobacteriaceae; Escherichia.
OX   NCBI_TaxID=562;
RN   [1]
RP   SEQUENCE FROM N.A.
RC   STRAIN=K12;
RA   Nashimoto H., Saito N.;
RL   Submitted (MAY-1996) to the EMBL/GenBank/DDBJ databases.
RN   [2]
CC   -!- FUNCTION: HAS HIGH FORMALDEHYDE DEHYDROGENASE ACTIVITY IN THE
CC       PRESENCE OF GLUTATHIONE AND CATALYZES THE OXIDATION OF NORMAL
DR   EMBL; D85613; BAA12834.1; ALT_FRAME.
DR   EMBL; D38504; BAA22412.1; -.
KW   Oxidoreductase; Zinc; Metal-binding; NAD; Complete proteome.
FT   METAL        40     40       ZINC 1 (CATALYTIC) (BY SIMILARITY).
FT   METAL        62     62       ZINC 1 (CATALYTIC) (BY SIMILARITY).
SQ   SEQUENCE   369 AA;  39359 MW;  35B59078F8173521 CRC64;
     MKSRAAVAFA PGKPLEIVEI DVAPPKKGEV LIKVTHTGVC HTDAFTLSGD DPEGVFPVVL
```

```
     GHEGAGVVVE VGEGVTSVKP GDHVIPLYTA ECGECEFCRS GKTNLCVAVR ETQGKGLMPD
//
```

Figure 8: Exemple d'un extrait d'entrée de la base de données Swiss-Prot.

Nous avons choisi d'enlever les fragments de protéine pour ne pas fausser plus tard la décomposition des collections enzymatiques en régions similaires puisque nous cherchons à identifier les régions de séquences protéiques les plus informatives pour une activité enzymatique. Ainsi, 1606 collections enzymatiques sont constituées sur les 3870 entrées de la base de données ENZYME. En effet, il y a de nombreuses entrées dans la base ENZYME (plus de 2000) qui ne possèdent pas d'information de séquence. Par ailleurs, quelques collections enzymatiques ne sont pas constituées car elles ne contiennent que des fragments de séquences protéiques.

2.2.2 Identifications des régions similaires dans une collection

Les collections enzymatiques constituées, nous avons recherché dans chaque collection de séquences protéiques, les segments de séquences similaires les plus longs possibles afin de caractériser au mieux les séquences.

Pour effectuer cette recherche de similarité et une décomposition en régions dans un groupe de séquences, nous avons choisi d'utiliser le programme MKDOM2 (44). Ainsi, l'utilisation de l'algorithme MKDOM2 dans chaque collection enzymatique consiste à choisir la plus petite séquence de la collection comme requête pour le programme PSI-BLAST (30) contre l'ensemble des séquences de la collection. Les séquences de la collection obtenues avec une E-value inférieure à 10^{-4} sont conservées et définissent une famille de modules homologues. Nous avons choisi de nommer ces régions modules et non domaines car nous avons constaté que ces régions de protéines sont en moyenne plus grandes qu'un domaine structural tel qu'il est défini dans la base de données de domaines structuraux SCOP (39) (partie 'signification structurale des modules' dans la suite du chapitre II). Le processus de MKDOM2 est réitéré jusqu'à épuisement de la collection enzymatique.
La décomposition en modules obtenue par MKDOM2 est présentée dans trois fichiers de résultats :

- un fichier « numéroEC ».xdom présentant un résumé de la décomposition en modules des protéines de la collection.

- un fichier « numéroEC ».mul, trié selon l'effectif des familles puis selon la longueur des modules. Ce fichier présente les alignements multiples des modules.
- Un fichier db« numéroEC » contient les séquences consensus de toutes les familles de modules. La séquence consensus et l'alignement multiple de chaque module sont calculés avec le programme MULTALIN (24) intégré dans MKDOM2.

2.2.3 Sélection des modules par collection enzymatique

La décomposition en modules d'une collection enzymatique peut-être visualisée avec le programme XDOM (46). La figure 9 présente l'exemple de la collection enzymatique de la fonction purine nucléoside phosphorylase décomposée en 4 modules par le programme MKDOM2. Cette représentation permet d'observer que tous les modules ne sont pas nécessaires pour caractériser toutes les séquences de la collection enzymatique. En effet, si l'on privilégie les modules comprenant le plus grand nombre de séquences et les plus longs, le module colorié en noir et le module colorié en rouge sont suffisants pour caractériser les deux groupes non homologues de séquences. Le premier module caractérisant les enzymes d'organismes procaryotes et le second module les enzymes d'organismes eucaryotes dans cet exemple.

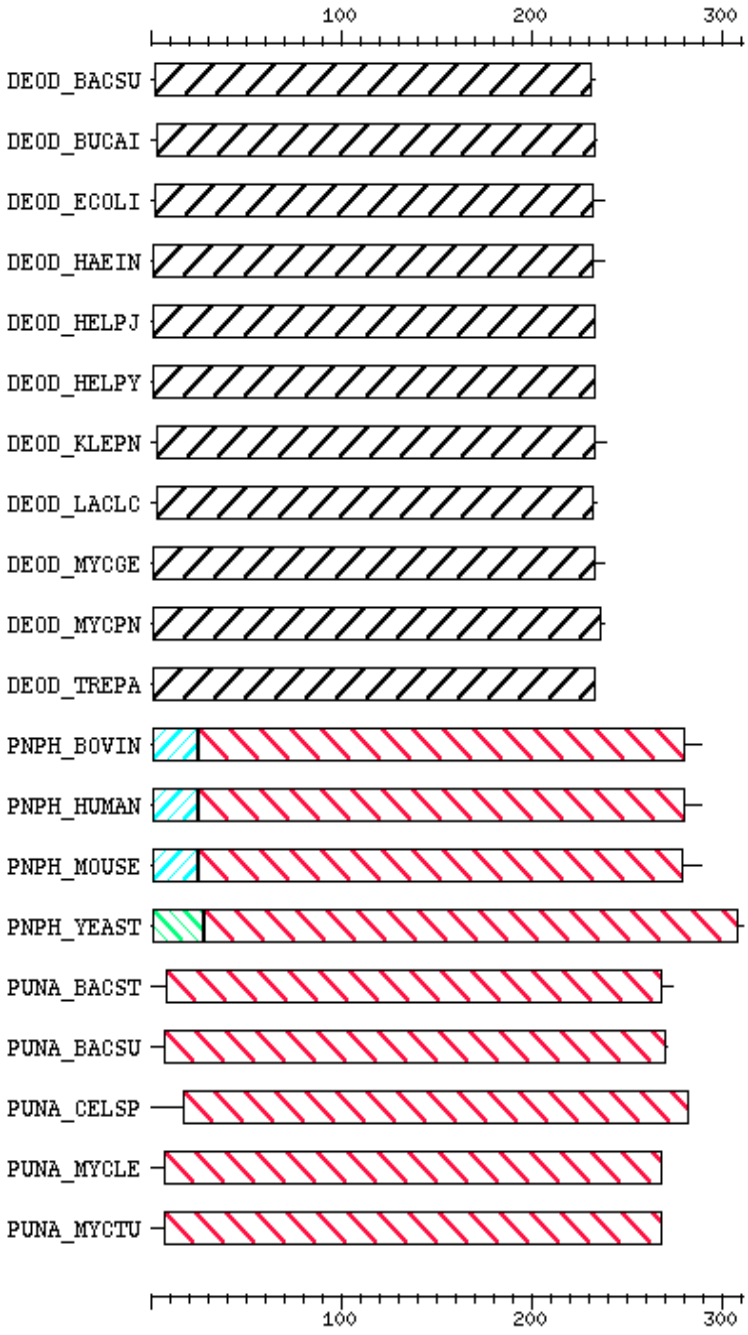


Figure 9 : Décomposition en modules par MKDOM2 de la collection enzymatique Purine nucléoside phosphorylase (EC 2.4.2.1).

En sélectionnant les modules comprenant le plus grand nombre de séquences et les plus longs, le module colorié en noir et le module colorié en rouge sont suffisants pour caractériser les

deux groupes non homologues de séquences. Le premier module caractérisant les enzymes d'organismes procaryotes et le second module les enzymes d'organismes eucaryotes.

Ainsi, nous avons effectué une sélection du nombre minimum de modules pour représenter toutes les séquences d'une collection, en conservant tout d'abord les modules avec le plus grand nombre de séquences puis les modules les plus longs. Cette sélection permet de caractériser les 1606 fonctions enzymatiques par 2435 types de modules.

Le nombre de type de modules dépend de la collection enzymatique. Il varie d'un module pour une collection de séquences similaires à des dizaines de modules pour une collection contenant des complexes enzymatiques. La distribution du nombre de modules sélectionnés par collection enzymatique est présentée figure 10. On observe que plus des trois quarts des collections enzymatiques (1296 numéros EC sur 1606) ont un seul module sélectionné et correspondent à des séquences homologues.

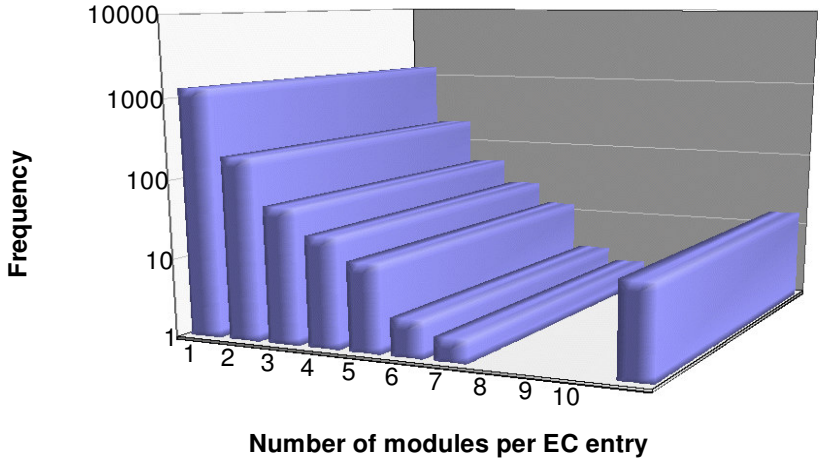


Figure 10: Distribution du nombre de modules sélectionnés pour chaque collection enzymatique de PRIAM.

Les autres collections enzymatiques sont caractérisées par au moins deux modules. Les cas de collections avec au moins deux modules sélectionnés peuvent correspondrent à la présence d'enzymes analogues comme dans le cas de la purine nucléoside phosphorylase (figure 9) ou à une enzyme oligomérique comme pour le cas de l'homocitrate synthase (figure 11).

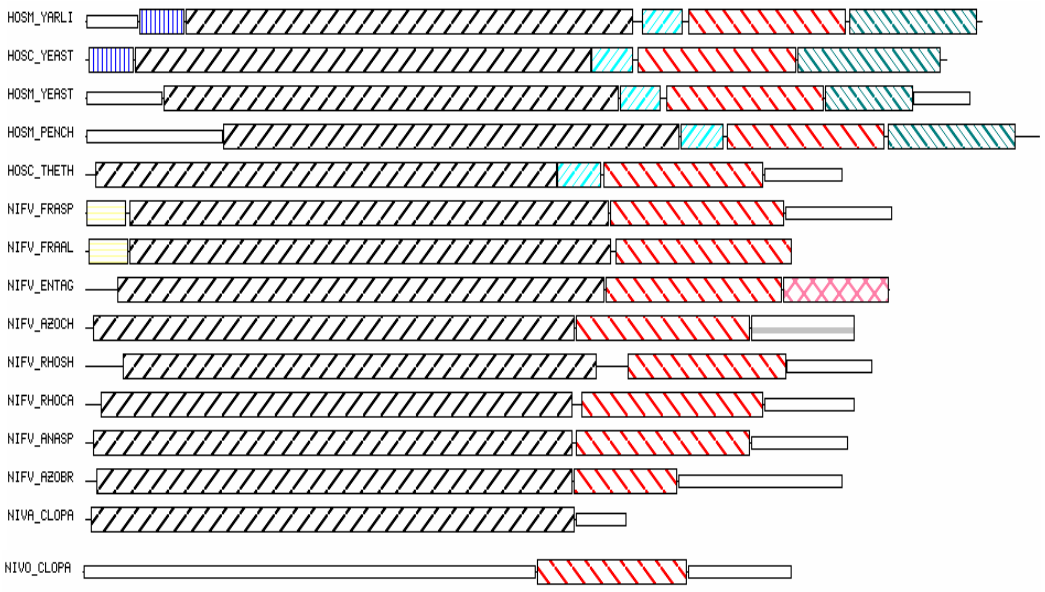


Figure 11: Exemple d'enzymes modulaires avec une règle « AND ».

Les modules détectés pour l'homocitrate synthase (EC 4.1.3.21) sont visualisés avec le programme XDOM. Les modules rayés noir et rouge sont nécessaires pour prédire la présence de l'homocitrate synthase.

2.2.4 Détermination de règles

Si l'on veut effectuer plus tard une recherche de similarité avec ces modules, on remarque que dans le cas d'enzymes analogues pour une activité enzymatique (figure 9), il est nécessaire de trouver une homologie avec l'un ou l'autre des modules alors que dans le cas d'une enzyme oligomérique il est nécessaire de trouver une homologie avec chacun des modules (figure 11). C'est pourquoi, nous avons déterminé une règle pour chaque collection enzymatique, précisant si une homologie doit être trouvée avec l'un ou l'autre des modules (règle « OR ») ou si elle doit être trouvée avec les deux modules (règle « AND »). Cette règle peut être déterminée automatiquement sur l'ensemble des modules d'une collection en tenant compte de l'inventaire des organismes participant à chaque famille de modules. Ainsi, le principe de l'algorithme (figure 13) que nous avons développé est le suivant :

Pour chaque collection, les familles de modules sont triées selon le nombre décroissant d'organismes impliqués. Puis la plus grande liste d'organismes ou liste courante, est comparée à la liste suivante.

- Si les listes d'organismes sont identiques, les modules sont liés par une règle « AND » puisqu'on observe systematiquement la co-occurrence des modules.
- Si au moins un organisme de la nouvelle liste est différent de l'ensemble des organismes observés dans les familles de modules précédentes, alors les modules sont liés par une règle « OR », puisque ce module est requis pour cette nouvelle espèce.
- Enfin, lorsqu'une liste d'organismes courante est incluse dans la liste précédente, cela signifie que la présence du module n'est pas générale, donc non indispensable pour rendre compte de la présence de l'enzyme (figure 12).

Enfin, les listes d'organismes sont fusionnées et le processus est itéré avec la liste suivante.

On obtient ainsi pour une collection enzymatique une règle de modules à trouver pour identifier la fonction enzymatique. Parmi les 1606 collections enzymatiques, 310 sont caractérisées par plusieurs profils, 38 collections ont une règle AND avec deux ou plusieurs profils. Une seule collection est trouvée avec une règle AND et OR, correspondant à la fonction L-serine dehydratase (EC 4.3.1.17), avec deux profils pour les deux sous-unités des enzymes de type procaryote et un profile pour les enzymes de type eucaryote.

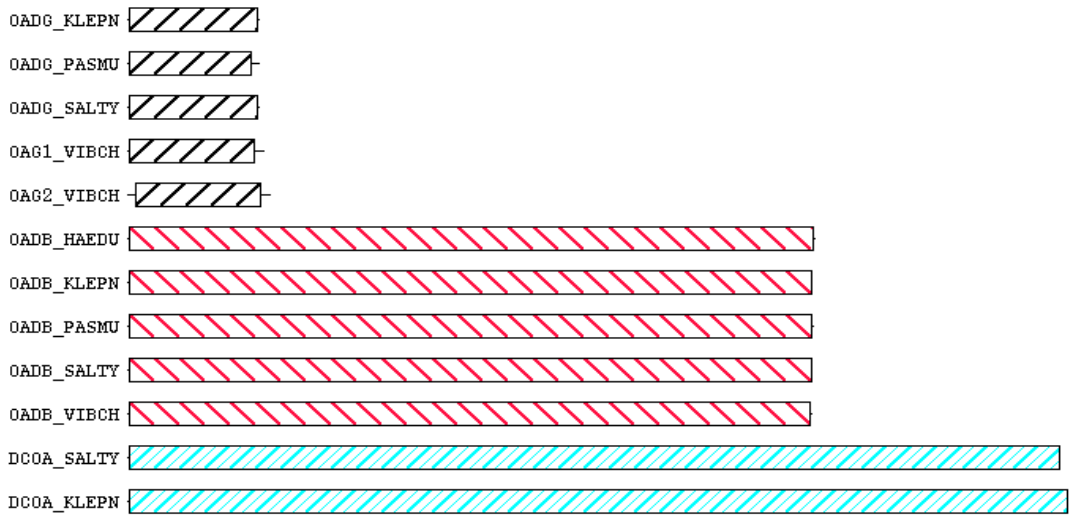


Figure 12: Collection de l'Oxaloacetate decarboxylase (EC 4.1.1.3).

Exemple de collection avec seulement le module rouge sélectionné par l'algorithme de détermination de règles.

Soit *n* le nombre total de modules pour chaque collection enzymatique. Nous appliquons l'algorithme suivant pour sélectionner les modules a utiliser dans la règle *P*:

```
// Initialisation
for i = 1...n  {
            let O_i be the list of organisms represented by module i ;
            let P_i define the logical rule matching module i
            }
sort O_i by decreasing |O_i|
set P = FALSE , O = O_1 ;

// Rule generation
for i = 2...n  {
            if O_i == O_i-1        set P_i = P_i AND P_i-1 ;
            // revise P_i when verified in all species ∈ O_i-1
            elseif O_i ⊄ O set P = P OR P_i-1 ;
            // revise P only when P_i is verified in new species ∉ O
            else set P_i = P_i-1 ;
            // ignore P_i if verified only in subset of O
            O = O ∪ O_i
            }
set P = P OR P_n
```

Figure 13: Pseudo-code générant la règle logique pour chaque collection enzymatique.

2.2.5 Construction des profils

Nous avons choisi de représenter les familles de modules sélectionnées par des profils ou matrices de scores positions spécifiques (PSSM) avec le programme PSI-BLAST (version 2.0.11).

La méthodologie des PSSM a été choisie car elle permet de tenir compte de la diversité de la composition en acides aminés des différentes séquences qui composent la famille de module. Elle permet aussi de faire une recherche d'homologie en tenant compte de la position de certains acides aminés plus représentés dans un alignement multiple. Ces acides aminés

peuvent avoir un rôle important dans la fonction de la séquence, tel que le site actif dans le cas des fonctions enzymatiques. Nous avons choisie les PSSM plutôt que les profils HMM car d'une part les PSSM sont équivalentes à un profil HMM d'ordre 0 et d'autre part parce que l'utilisation d'un profil HMM d'ordre supérieur est très coûteuse en temps lors d'une recherche de similarité.

Enfin, nous avons choisi de construire les profils avec le programme PSI-BLAST car il permet de construire très rapidement une banque de profils à partir d'un ensemble de séquences prédéfini et d'une séquence représentative du groupe de séquences (séquence consensus par exemple).

Pour chaque module d'une collection enzymatique, à partir de deux fichiers produits par le programme MKDOM2, sont construits un fichier de séquences au format fasta avec les segments de séquences appartenant à une famille de module et un fichier avec la séquence consensus du module. Ces deux fichiers sont donnés en entrée au programme PSI-BLAST pour calculer un profil avec l'option –C. Cette option permet de construire un profil en format binaire réutilisable par le même programme pour une recherche de similarité contre une banque de séquences. Il est certain que toutes les séquences du module seront recrutées pour le calcul du profil car elles sont similaires avec une E-value inférieure à 10^{-4} et la E-value par défaut du programme PSI-BLAST est à 10.

Afin de créer une banque de profils spécifiques d'enzymes, utilisable pour une recherche de similarité dans un génome complet, les profils obtenus avec PSI-BLAST sont formatés avec plusieurs programmes du logiciel RPS-BLAST sous lunix et sur solaris. En effet le programme RPS-BLAST utilise une banque de mots comme dans l'heuristique BLAST et plusieurs fichiers qui contiennent, entre autre, une liste de mots précalculés pour les profils permettant une recherche plus rapide. Ces fichiers ne peuvent pas être calculés indépendamment, ils dépendent de l'architecture du système utilisé pour leur calcul.

Le premier programme makemat converti les profils binaires calculés par PSI-BLAST en format ASCII.

Puis le second programme copymat convertit les profils ASCII en une banque dans un fichier pour être lu plus rapidement.

Enfin le programme de formatage classique de banque BLAST formatdb est utilisé.

2.3 Evaluation de l'efficacité des profils

Nous avons évalué l'efficacité de la banque de profils c'est-à-dire que nous avons vérifié la capacité de chaque profil à retrouver toutes les séquences de la collection enzymatique (sensibilité) et peu de similarités avec des séquences de fonctions enzymatiques différentes (spécificité). Pour cela nous avons calculé la spécificité et la sensibilité de chaque profil en réalisant une recherche de similarité avec la base de données Swiss-Prot.

2.3.1 Définitions de la spécificité et de la sensibilité

Les définitions de spécificité et sensibilité correspondent à celles définies par M. Burset et R. Guigo (108). Ainsi, la **spécificité** est égale au nombre des enzymes correctes (Vrais Positifs) sur le nombre total des enzymes prédites par le profil (Vrais Positifs et Faux Positifs) :

$$Sp = VP / (VP + FP)$$

La **sensibilité** est égale au nombre des enzymes correctes (Vrais Positifs) sur le nombre des enzymes réellement correctes (Vrais Positifs et Faux Négatifs) :

$$Sn = VP / (VP + FN)$$

2.3.2 Calcul de la spécificité et de la sensibilité des profils

Chaque profil est utilisé comme entrée du programme PSI-BLAST avec la banque Swiss-Prot, afin d'évaluer son efficacité. Les paramètres du PSI-BLAST utilisés contre la banque Swiss-Prot sont : E-value = 0.1 et j=1, c'est-à-dire une seule itération de PSI-BLAST est effectuée pour vérifier que le profil permet de retrouver directement l'ensemble des séquences qui sont des vrais positifs.

Les séquences dites vrais positifs sont les séquences trouvées par similarité avec le profil avec le même numéro EC que le profil. Les séquences dites faux positifs sont les séquences trouvées par similarité avec un numéro EC différent de celui du profil. Les séquences faux négatifs sont les séquences de la collection enzymatique qui ne sont pas

trouvées par similarité avec le profil. Ces différents cas sont identifiés par la détection des numéros EC et des noms de protéines dans le fichier de résultats du programme PSI-BLAST. La détection des numéros EC pour identifier les vrais et faux positifs posent des problèmes car nous avons observé quelques cas de numéros EC transférés vers un autre numéro entre la version de Swiss-Prot et la version d'ENZYME utilisées. Ainsi, certaines séquences vrais positifs et faux positifs ne sont pas détectés par la recherche d'expressions régulières. Par exemple le profil 146p3.6.3.14 correspond à une des sous-unités de l'ATP synthase (EC 3.6.3.14) et permet de trouver par homologie la sous-unité correspondante. Mais cette sous-unité est annotée avec l'ancien numéro EC (EC 3.6.1.34) dans la base Swiss-Prot, récemment transféré en 3.6.3.14. Par ailleurs des séquences sans numéros EC sont aussi détectées comme des faux positifs alors qu'elles peuvent correspondre à des vrais positifs qui ne sont pas encore annotés avec la nomenclature.

Ainsi 1922 profils sur 2 137 ont une spécificité et une sensibilité différentes de 0, évaluées sur la base de données Swiss-Prot. La distribution du nombre de profils en fonction des valeurs de spécificités et sensibilités est représentée dans la figure 14A. On observe qu'un grand nombre de profils sont très sensibles mais peu spécifiques car il y a de nombreuses séquences de faibles similarités trouvées par le programme PSI-BLAST à cause du seuil de la E-value à 0,1. Cependant nous avons amélioré la spécificité de certains profils en appliquant un seuil à la valeur du score. D'autre part, environ 200 profils sont très spécifiques mais peu sensibles car ils appartiennent à des collections enzymatiques nécessitant plusieurs profils pour caractériser toutes les séquences (exemple des enzymes oligomériques et des complexes enzymatiques).

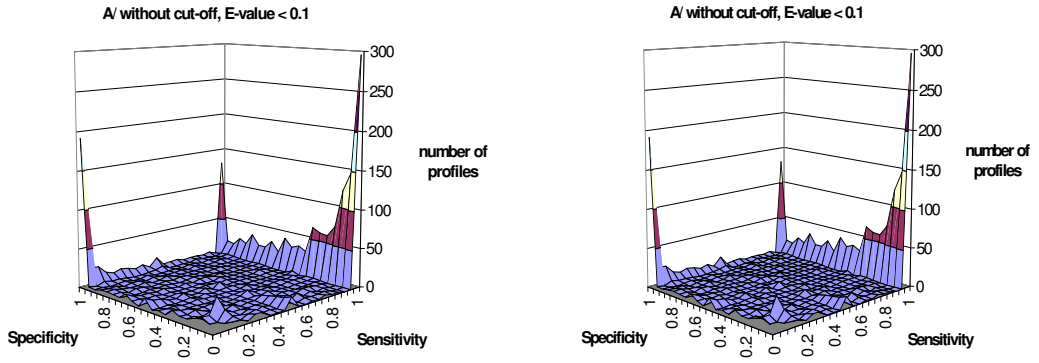


Figure 14 : Distribution de la spécificité et de la sensibilité des profils PRIAM testés contre Swiss-Prot A/ avec E= 0.1 B/ en utilisant des valeurs seuils spécifiques.

2.3.3 Détermination d'un score seuil spécifique de chaque profil

Pour les 1922 profils avec au moins une séquence vrai positif et une séquence faux positif, nous avons pu déterminer une valeur de score afin d'éviter des homologies avec des séquences faux positifs et donc améliorer la spécificité. Nous avons pour cela calculé les distributions cumulées des vrais positifs avec un score de PSI-BLAST inférieur à un seuil (courbe ascendante sur la figure 15) et des faux positifs avec un score supérieur à un seuil (courbe descendante sur la figure 15).

Deux types de distributions peuvent être obtenues (figure 15) :

- soit les courbes sont séparées (à gauche), nous avons déterminé la valeur seuil du score comme étant la médiane du segment entre les deux extrémités des courbes afin d'éliminer tous les faux positifs,
- soit les courbes se coupent (à droite), nous avons déterminé la valeur seuil du score comme l'intersection des deux courbes.

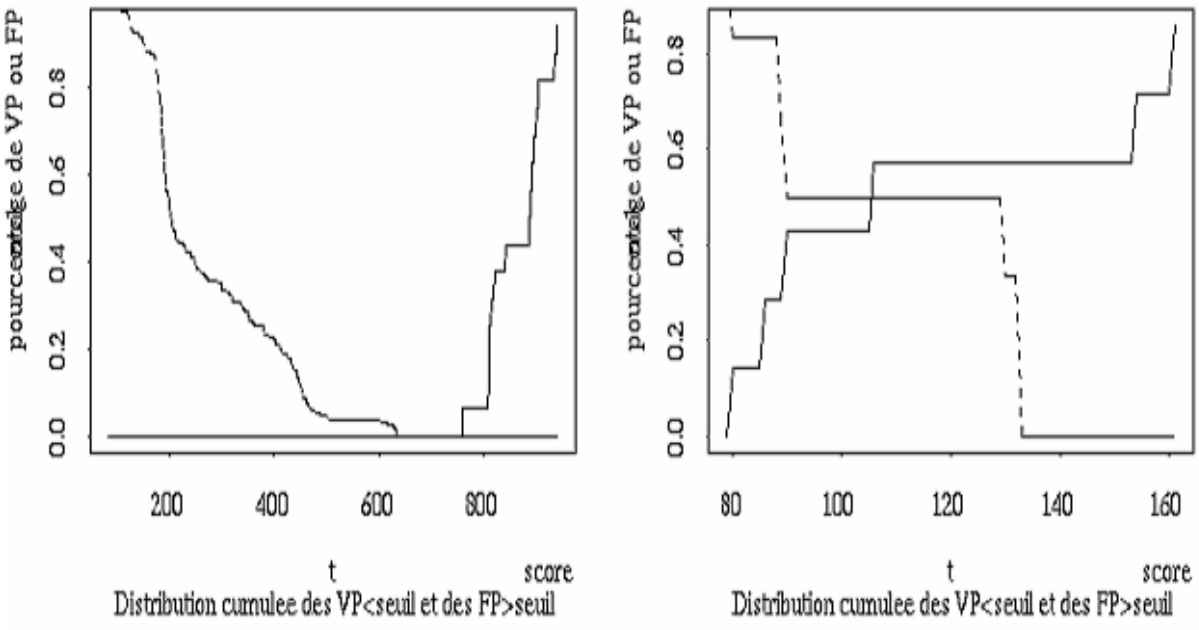


Figure 15 : Deux exemples de distributions cumulées des vrais positifs et faux positifs de deux profils en fonction du score.

Nous avons ainsi déterminé 1922 valeurs seuils du score que nous avons utilisé pour filtrer les résultats de PSI-BLAST avec ces profils contre la base de données Swiss-Prot. La distribution du nombre de profils en fonction des valeurs de spécificités et de sensibilités après seuillage est représentée dans la figure 14B. Nous avons ainsi nettement amélioré la spécificité des profils sans perte de sensibilité.

2.4 Signification structurale des modules

Nous rappelons que nous avons choisi de nommer les régions similaires par le terme module au lieu du terme domaine car nous avons observé que les régions que nous avons définies sont plus grandes que des domaines tels qu'ils sont définis dans les bases de données SCOP (39) ou PRODOM (14).

Par ailleurs, nous avons recherché les modules avec une forte homologie pour des domaines structuraux de la base de données SCOP afin d'avoir une information structurale sur les modules sélectionnés dans les collections enzymatiques.

2.4.1 Choix de la base de données SCOP

Nous avons choisi la base de données de domaines SCOP car la délimitation des domaines se base sur des données de structures tridimensionnelles (base de données PDB) à partir de résultats expérimentaux (RMN ou cristallographie) et sur des données de conservation de ces domaines au cours de l'évolution (domaines phylogénétiques). Le domaine de la base de données SCOP est le plus bas niveau de la hiérarchie de la classification des protéines. Nous avons récupéré les domaines de SCOP au format fasta à partir de la base de données ASTRAL (109). La version 1.55 de SCOP et ASTRAL contiennent 30.867 domaines (correspondant à 16.972 structures de la PDB).

2.4.2 Similarité entre les modules d'enzymes et les domaines SCOP

Nous avons comparé la longueur des modules sélectionnés pour chaque collection avec la longueur des domaines de la base de données SCOP. La figure 16 met en évidence

qu'en moyenne les profils de PRIAM sont plus longs que les domaines de la base de données SCOP. Les modules sont plus longs puisqu'ils sont construits automatiquement sur des sous-ensembles de séquences le plus souvent très similaires. C'est pourquoi, ils correspondent souvent à une grande longueur de la séquence protéique.

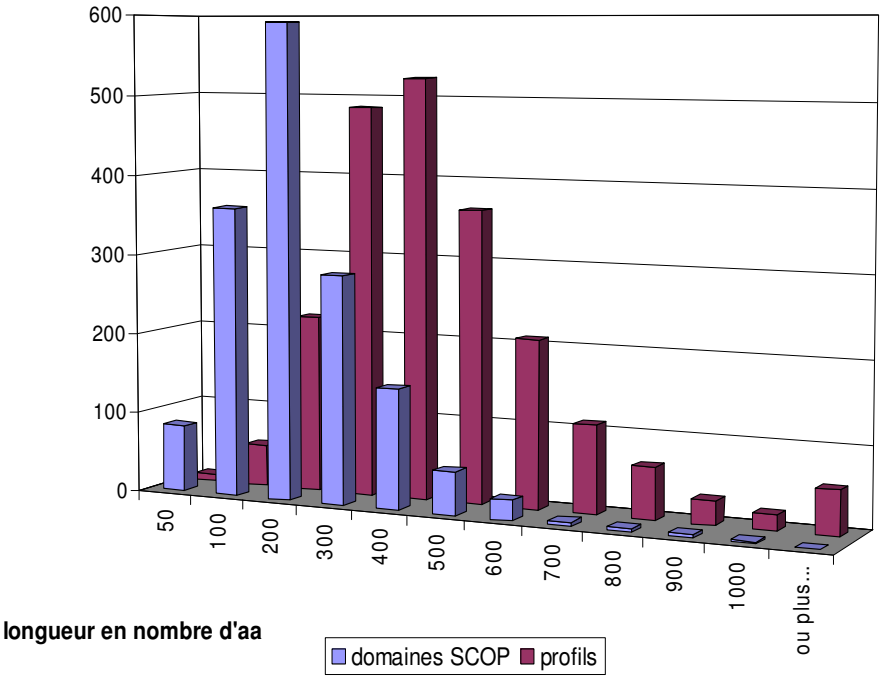


Figure 16 : Distribution de la longueur des domaines de la base de données SCOP et des profils PRIAM.

Par ailleurs, nous avons recherché des similarités entre les profils PRIAM et les domaines de SCOP afin d'avoir une correspondance entre les modules sélectionnés pour les collections enzymatiques et des domaines bien connus. Nous avons exécuté le programme PSI-BLAST avec chaque profil et la banque de domaines de SCOP. Les résultats ont été filtrés de manière à avoir une similarité assez importante entre le module et le domaine SCOP. Nous avons choisi de sélectionner les similarités avec au moins 30% d'identité entre la séquence consensus du profil et le domaine de la base de données SCOP et en sélectionnant sur la longueur de la similarité de deux façons différentes :

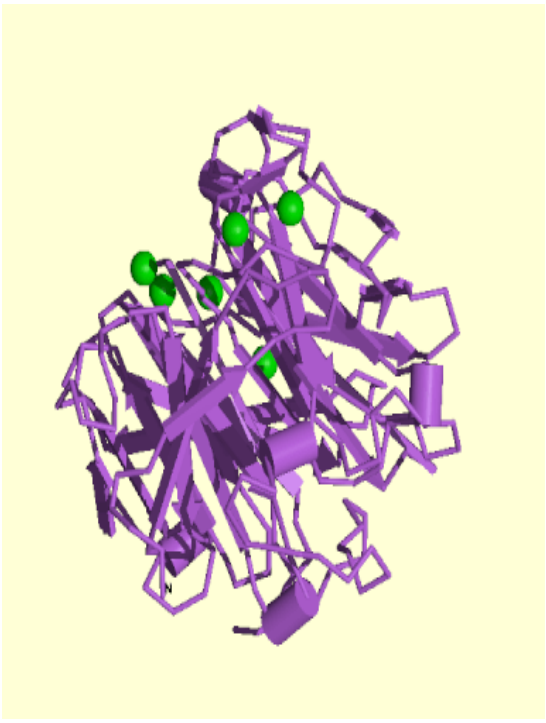
- pour trouver une similarité globale (80% des longueurs du domaine de SCOP et du profil),

- pour trouver une similarité globale sur le domaine de SCOP et locale sur les profils qui sont plus grands en moyenne (80% de la longueur du domaine de SCOP).

Nous avons ainsi obtenu une similarité globale entre les domaines de SCOP et les profils de PRIAM pour 26,5% des profils et une similarité locale pour 47% des profils PRIAM. La base de données SCOP étant basée sur l'ensemble des séquences ayant une structure dans la PDB, les modules pour lesquels nous n'avons pas trouvé de similarité sont des cibles potentielles d'étude pour identifier de nouvelles structures.

Cette analyse nous a permis de trouver par ailleurs 29 cas de numéros EC caractérisés par plusieurs profils possédant des similarités globales pour des domaines avec des repliements différents de SCOP. Les profils correspondant aux repliements différents sont soit des cas de plusieurs sous-unités pour une fonction enzymatique, soit des cas d'enzymes analogues avec des séquences primaires et des repliements différents. Comme par exemple, le cas de la collection enzymatique de la 3-phytase (EC 3.1.3.8) caractérisée par deux profils. Un profil est spécifique d'organismes eucaryotes (profil 1p3.1.3.8) et l'autre de bactéries (profil 2p3.1.3.8). Chaque profil présente une forte similarité pour un repliement différent (figure 17). Le profil 1p3.1.3.8 a une similarité avec le domaine qui a le repliement 'Phosphoglycerate mutase-like' (59 % d'identité, 99 % de la longueur du domaine SCOP et 93 % de la longueur du profil, E value = 0). Le profil 2p3.1.3.8 a une similarité avec le domaine qui a le repliement '6-bladed beta-propeller' (96 % d'identité, 100 % de la longueur du domaine SCOP et 92 % de la longueur du profil, E value = 0).

Fold c.60 : Phosphoglycerate mutase-like
Homologie du profil 1p3.1.3.8 avec la séquence 3-phytase d *'Aspergillus ficuum*.



Fold b.68 : 6-bladed beta-propeller
Homologie du profil 2p3.1.3.8 avec la séquence 3-phytase de *Bacillus amyloliquefaciens*.

Figure 17 : Exemple de deux repliements différents pour deux profils de la collection 3-phytase (EC 3.1.3.8).

3 Prédiction du métabolisme à partir d'un génome complet: le programme PRIAM

La deuxième partie du travail de thèse a été le développement du programme PRIAM réalisant l'annotation automatique des enzymes d'un génome complet avec la banque de profils spécifiques d'enzymes. La figure 18 présente ce processus automatique d'annotation d'un génome complet.

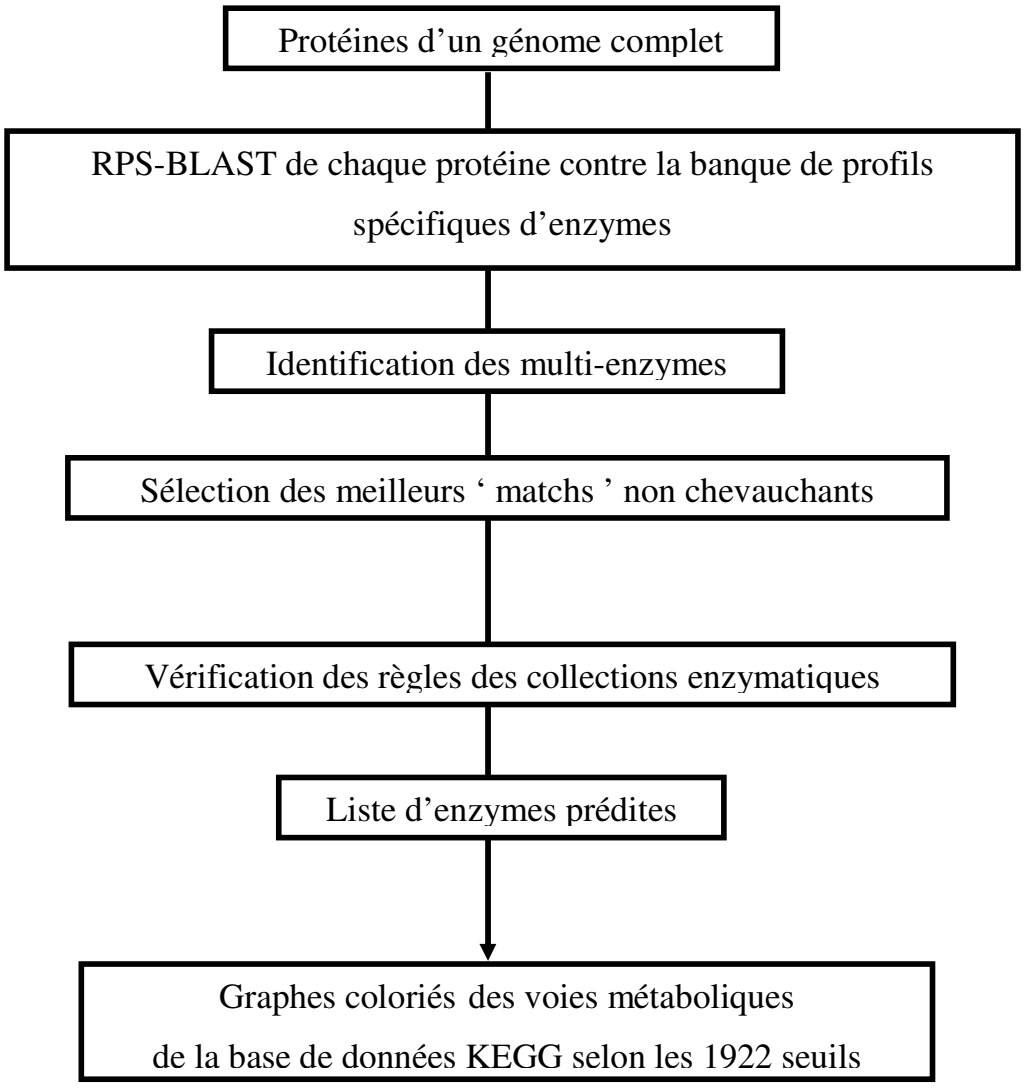


Figure 18 : Processus d'annotation d'un génome complet avec le programme PRIAM.

3.1 Recherche d'homologie contre la banque de profils

La première étape consiste à effectuer une recherche de similarité de chaque protéine prédite à partir d'un génome complet contre la banque de profils. Ainsi chaque protéine au format fasta est utilisée comme requête avec le programme RPS-BLAST contre la banque de profils. Les résultats du programme RPS-BLAST sont filtrés avec une E-value de 10^{-10} afin de trouver de fortes similarités. Les informations sur les similarités obtenues par RPS-BLAST que nous avons conservées pour traiter les résultats sont :

- la E value
- le score brut pour pouvoir le comparer ultérieurement avec le score seuil détecté contre la banque Swiss-Prot.
- le pourcentage d'identité entre la séquence requête et la séquence consensus du profil.

3.2 Identification des protéines muti-enzymatiques

Dans une deuxième étape, les protéines multi-enzymatiques sont identifiées. Le principe de cette recherche est détaillé dans la figure 19. Cette recherche consiste à identifier les meilleures similarités sur des régions non-chevauchantes ou chevauchantes sur une longueur inférieure à 20 aa.

Puis le meilleur 'match', c'est-à-dire la similarité entre la protéine et un profil avec la E value la plus faible, est sélectionné. Plus précisément, le meilleur match est sélectionné pour une protéine mono-enzyme et les meilleurs matchs sont sélectionnés pour les protéines multi-enzymatiques.

```
let n be the total number of PRIAM matches for protein p ;
let M_i be the matches sorted by increasing E value ( i = 1...n );
set M = M_1 ;
for i = 2...n {
        if    M and M_i overlap on p by less than 20 amino acids
        set   M = M ∪ M_i ;
        // M_i corresponds to a new region on query protein p
        }
```

Figure 19: Pseudo-code pour la détection de multi-enzyme, appliqué à chaque protéine p.

3.3 Vérification des règles des collections enzymatique

La règle de chaque collection enzymatique est vérifiée. C'est-à-dire, que dans le cas d'une collection avec une règle « AND », s'il manque une similarité avec l'un des profils, la fonction enzymatique n'est pas prédite. Une liste de fonctions enzymatiques prédites est ainsi obtenue pour le génome complet.

3.4 Représentation des résultats

Les résultats de cette prédiction sont automatiquement représentés sur les graphes des voies métaboliques de la base de données KEGG (70). Nous avons choisi les graphes de la base de données KEGG car ils représentent les voies métaboliques avec l'ensemble des enzymes possibles selon différents organismes. Nous avons choisi ce mode de représentation et avons développé un script coloriant les rectangles, représentant les fonctions enzymatiques, à partir de la liste des numéros EC prédits par PRIAM. Plus précisément, les numéros EC trouvés avec un score en dessous du seuil détecté contre Swiss-Prot sont coloriés en jaune et les numéros EC avec un score au dessus du seuil détecté contre Swiss-Prot sont coloriés en vert.

Cette représentation a été utilisée pour analyser le métabolisme prédit pour deux génomes récemment séquencés : *Sinorhizobium meliloti* (analyse présentée dans le chapitre III) et *Ralstonia solanacearum* (110).

Par ailleurs les prédictions de fonctions et la représentation des voies métaboliques ont aussi été utilisées pour interpréter les résultats d'expériences sur le protéome (107) et le transcriptome (Communication personnelle de Marcela Davalos, laboratoire des Interactions Plantes Microorganismes) de *S. meliloti* mis dans différentes conditions expérimentales.

4 Evaluation du programme PRIAM sur plusieurs génomes complets et comparaison avec KEGG Orthology

Nous avons évalué l'efficacité du programme PRIAM en l'utilisant pour la prédiction automatique des enzymes de 5 génomes complets dont les gènes sont annotés et presque tous complètement intégrés dans la base de donné Swiss-Prot via le projet HAMAP (111). Nous avons choisi les génomes des bactéries *Haemophilus influenzae* (112) et *Mycoplasma genitalium* (113) tous deux séquencés en 1995, le génome de *Mycoplasma pneumoniae* (114) séquencé en 1996, de l'entérobactérie *Escherichia coli* (115) séquencé en 1997, et de *Buchnera aphidicola* (116) entièrement séquencé en 2000. Cependant, les gènes détectés dans ces cinq génomes n'ont pas tous une fonction biologique attribuée. Ainsi d'après la base EcoCyc (juin 2003), 28 % des gènes d'*E. coli* n'ont pas de fonction attribuée.

4.1 Origine et description des données

Les séquences protéiques des cinq génomes complets : *H. influenzae*, *M. genitalium*, *M. pneumoniae*, *E. coli* et *B. aphidicola* (sous espèce *Acyrthosiphon pisum*) ont été obtenues sur le site ftp expasy du projet HAMAP (High-quality Automated and Manual Annotation of microbial Proteomes) (pour *H. influenzae* : 1 711 protéines, pour *M. genitalium* : 486 protéines, *M. pneumoniae* : 687 protéines, pour *E. coli* : 4 347 protéines et pour *B. aphidicola* : 572 protéines). Tous ces génomes sauf *E. coli* (97%) sont intégrés à 100% dans Swiss-Prot.

Nous avons récupéré les annotations de fonctions enzymatiques en détectant les numéros EC dans le champ DE des entrées Swiss-Prot (version 40). Ainsi, nous avons obtenu pour *H. influenzae* : 367 numéros EC, pour *M. genitalium* : 105 numéros EC, pour *M. pneumoniae* : 110 numéros EC, pour *E. coli* : 623 numéros EC et pour *B. aphidicola* : 214 numéros EC.

4.2 Evaluation de l'efficacité de PRIAM

4.2.1 Comparaison des numéros EC

L'utilisation du programme PRIAM sur les génomes de ces cinq organismes permet d'identifier les numéros EC trouvés en commun par la méthode PRIAM et donné par l'annotation Swiss-Prot (vrais positifs), les numéros EC trouvés que par la méthode PRIAM (faux positifs) et les numéros EC proposés seulement par l'annotation Swiss-Prot (faux négatifs). Le programme PRIAM a été évalué en filtrant les résultats avec une E-value entre 10^{-10} et 10^{-50} sur ces cinq génomes complets afin d'identifier une valeur permettant d'avoir un meilleur compromis entre une bonne spécificité et une bonne sensibilité.

La même comparaison de numéros EC avec Swiss-Prot a été effectuée sur les résultats des annotations obtenues de manière semi-automatique par KEGG Orthology (KO) (117).

Par ailleurs, un test de type Jacknife a été réalisé en construisant cinq jeu de profils sans un des 5 génomes complets, afin d'évaluer les résultats de PRIAM sur un génome complet n'étant pas encore intégré dans ENZYME.

4.2.2 Calcul de la spécificité et de la sensibilité de la méthode PRIAM et de KEGG Orthology

Ainsi nous avons évalué la spécificité (VP / (VP + FP)) et la sensibilité (VP / (VP + FN)) de la méthode PRIAM par rapport à des annotations vérifiées par des experts. La figure 20 présente la moyenne des résultats pour les 5 génomes complets à différentes valeurs de E-value.

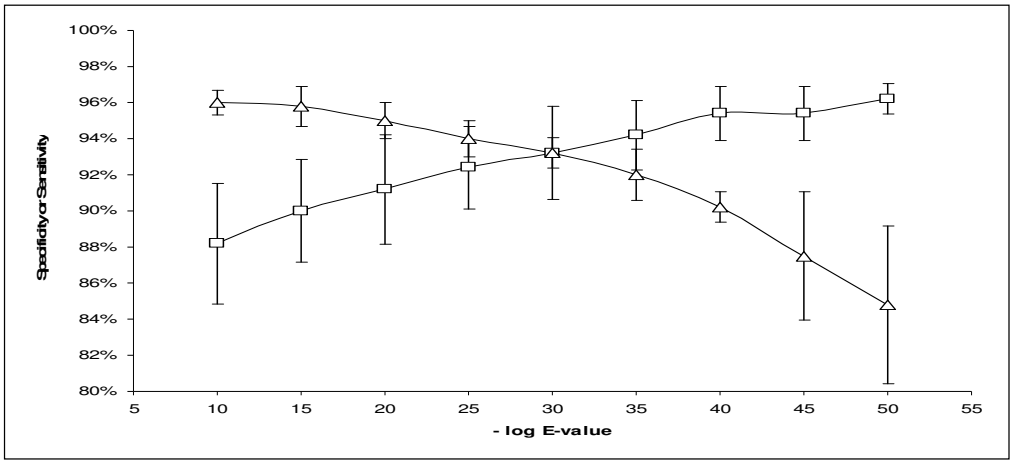


Figure 20: Calibration de la méthodologie PRIAM sur des génomes complets.

Spécificité et sensibilité des prédictions d'enzymes sont calculés en utilisant les ensembles d'enzyme à partir des annotations de génome complets trouvés dans SWISS-PROT pour B. aphidicola, E. coli, H. influenzae, M. genitalium et M. pneumoniae. La moyenne et l'écart type de la spécificité (carrés) de PRIAM et de la sensibilité (triangles) sont calculés pour différentes valeurs de la E-value avec RPS-BLAST.

Ces résultats mettent en évidence une bonne sensibilité et une bonne spécificité de la méthode, en moyenne autour de 93%, avec la E value à 10^{-30}. Les numéros EC qui n'ont pas été trouvés par la méthode PRIAM correspondent le plus souvent à des numéros EC attribués à des protéines possédant plusieurs activités enzymatiques localisées dans une même région protéique.

Le tableau 1 met en évidence que les résultats de PRIAM avec une E-value de 10^{-30} sont meilleurs que les résultats obtenus par KO. Et d'autre part, que les résultats du Jack-nife donnent une bonne spécificité et sensibilité de la méthode sur un nouveau génome.

Genome	**PRIAM**		**PRIAM jacknife**		**KEGG orthology**	
	Specificity	Sensitivity	Specificity	Sensitivity	Specificity	Sensitivity
B. aphidicola	97%	93%	86%	91%	87%	80%
E. coli	93%	93%	92%	88%	89%	91%
H. influenzae	94%	94%	84%	91%	88%	93%
M. genitalium	92%	92%	86%	87%	93%	95%
M. pneumoniae	90%	94%	85%	87%	91%	95%

Tableau 1: Spécificité et sensibilité de la détection des enzymes basées sur PRIAM et sur KO pour cinq génomes complets, en utilisant l'annotation de SWISS-PROT comme standard.

La valeur de la E-value de RPS-BLAST est de 10-30. L'analyse Jacknife a été effectué avec les profiles PRIAM dans lesquels les séquences des génomes correspondants ont été enlevées. La spécificité et la sensibilité des assignements de KEGG orthology ont été calculés de la même manière contre SWISS-PROT pour la comparaison.

4.3 Conclusion

Par l'observation de ces résultats, nous pouvons conclure à une bonne sensibilité de la méthode, même sur un nouveau génome dont les séquences ne sont pas encore intégrées dans la base ENZYME.

La principale difficulté dans cette méthodologie, mais aussi pour toutes les méthodes automatiques d'annotation, est de pouvoir identifier automatiquement un seuil afin d'améliorer la spécificité sans perdre en sensibilité.

Le second point sensible de la méthode est l'identification des enzymes avec différentes spécificités de substrats. En effet, si l'on regarde les meilleures similarités avec des scores très proches pour une protéine et non pas la meilleure similarité, on peut identifier les enzymes avec différentes spécificités de substrats. Une amélioration de la méthode, tout en restant automatique, serait de sélectionner les meilleurs similarités quand elles sont proches (seuil à définir).

5 Site et distribution de PRIAM

L'outil PRIAM e été sujet à publication (118) (Annexe 1) et possède un site de diffusion du programme d'annotation d'un génome complet, ainsi que les résultats obtenus sur deux génomes complets. En effet PRIAM a été utilisé pour l'annotation des génomes des bactéries *S. meliloti* (119) (Annexe 2) et *R. solanacearum* (110) (Annexe 3) étudiés au Laboratoire des Interactions Plantes Microorganismes de l'INRA de Toulouse. Par ailleurs, nous l'avons utilisé pour annoter les résultats des expériences sur le protéome de *S. meliloti*, réaliséés par l'Université Nationale Australienne de Camberra (107) (Annexe 4).

Le site Internet sur l'outil PRIAM est composé de quatre parties (figure 21) : http://genopole.toulouse.inra.fr/bioinfo/priam/ REL_OCT01/index_oct01.html

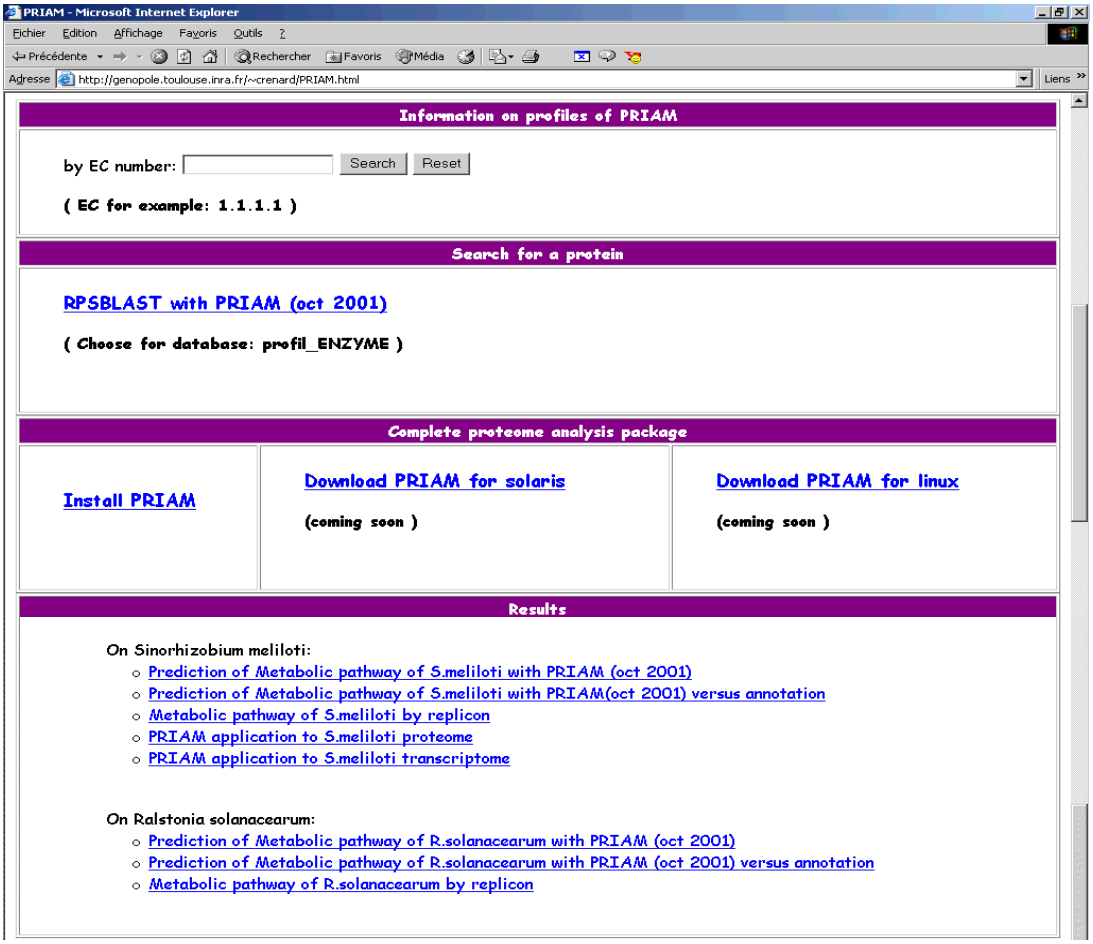


Figure 21: Page d'accueil du site PRIAM.

Un premier cadre permet d'obtenir des informations sur les modules sélectionnés pour chaque collection enzymatique de PRIAM. En donnant le numéro EC de la collection, on obtient la liste des modules sélectionnés pour caractériser la collection (figure 22), nommés par le nom du profil. Chaque profil possède l'information de la spécificité et de la sensibilité après détermination du seuil du profil contre la base de donnée Swiss-Prot, et lorsqu'il est disponible un lien vers le domaine de la base de données SCOP similaire au module PRIAM. Chaque nom de profil est un lien vers une représentation graphique des séquences comportant le module commun sélectionné. Le module d'intérêt est toujours celui observé dans toutes les séquences de la représentation. Cette présentation est faite dynamiquement grâce à une Berkeley DB contenant les images de la décomposition en modules de chaque famille de séquences. La Berkeley DB est une librairie C permettant d'effectuer des appels de fonction qui opèrent directement sur la base de données et sur les enregistrements qu'elle gère. Un programme réalisé par Emmanuel Courcelle (Laboratoire des Interactions Plantes Microorganismes) permet d'aller chercher les images nécessaires à la représentation des protéines participant au profil. Les images sont sauvées dans un cache.

Un second cadre propose d'effectuer une recherche de similarité d'une protéine requête avec le programme RPS-BLAST contre la banque de profils spécifiques des fonctions enzymatiques.

Le troisième cadre permet de récupérer le programme PRIAM compilé sur linux ou solaris avec la documentation d'installation.

Enfin, le dernier cadre présente les résultats sur deux génomes complets étudiés dans le Laboratoire des Interactions Plantes Microorganismes : *S. meliloti* et *R. solanacearum*. Tout d'abord, un lien vers les résultats obtenus avec PRIAM est proposé. Puis un lien est proposé vers les résultats de la comparaison des prédictions de PRIAM avec les annotations obtenues avec l'environnement d'annotation semi-automatique iANT. Enfin, un lien donne les résultats de la présentation des annotations selon les différents réplicons des organismes. Pour la bactérie *S. meliloti*, nous avons aussi présenté les résultats des prédictions obtenues avec PRIAM sur les protéines détectées par les expériences de protéomique dans différentes conditions.

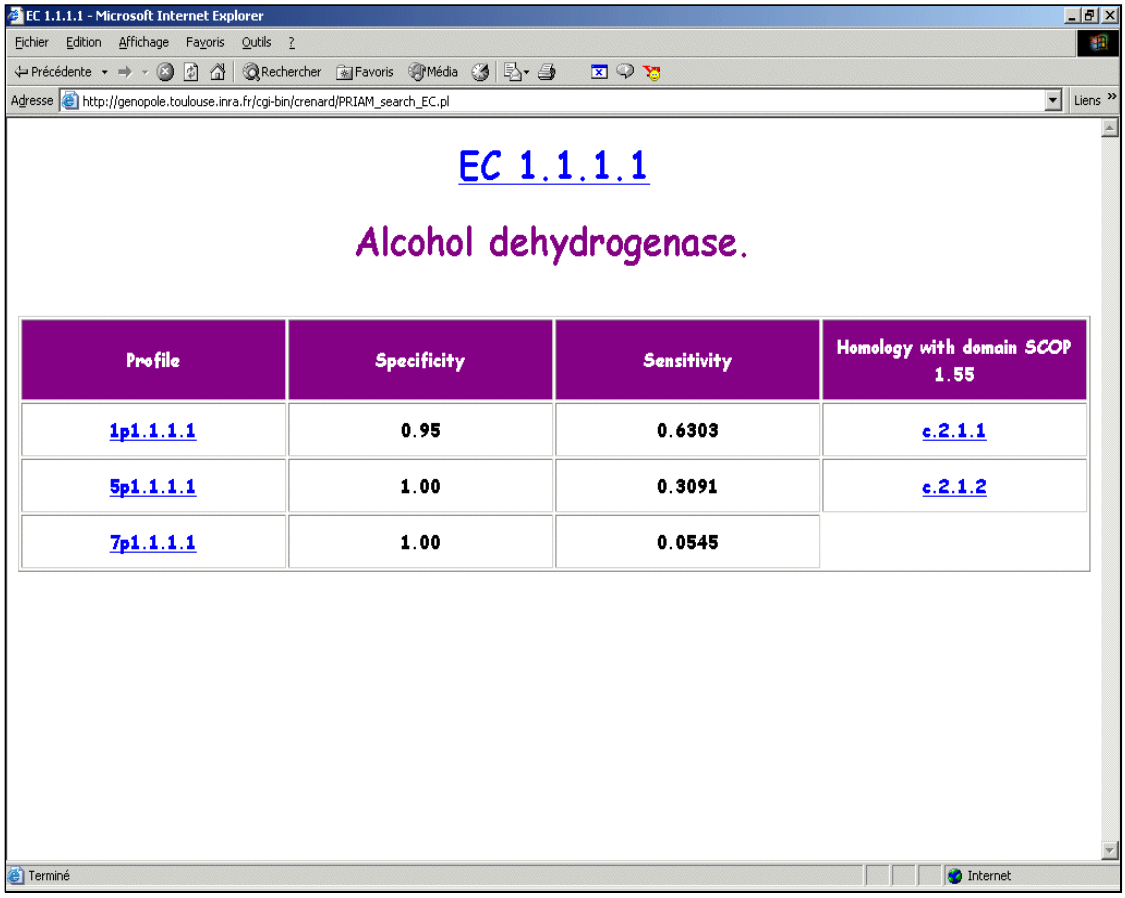


Figure 22 : Présentation de la liste des profils sélectionnés pour la collection de l'alcool déshydrogénase (EC 1.1.1.1).

6 Discussion

La méthodologie PRIAM présente des avantages et des limites spécifiques à sa construction. Par ailleurs, des enzymes de certaines voies métaboliques peuvent être « manquantes » par rapport à d'autres espèces mais pour d'autres raisons qu'un oubli de la méthode. Enfin, nous discuterons pourquoi certaines enzymes prédites par analyse de séquences ne sont pas toujours fonctionnelles dans l'organisme.

6.1 Les avantages de la méthode PRIAM

6.1.1 Une méthode d'annotation automatique des enzymes à partir d'un génome complet

La méthode PRIAM permet la prédiction automatique de l'ensemble des fonctions enzymatiques répertoriées dans la base ENZYME à partir du génome complet d'un organisme et la visualisation de ses voies métaboliques. La meilleure évaluation de l'efficacité de cette méthode est de la tester sur des génomes complets d'organismes récemment séquencés et annotés. C'est pourquoi nous l'avons évaluée sur le génome de la bactérie symbiotique *Sinorhizobium meliloti* et comparée avec les résultats d'annotation obtenus avec l'environnement semi-automatique iANT (75). Ce travail est présenté dans le chapitre III : Application au génome de *Sinorhizobium meliloti*.

6.1.2 L'intégration dans des voies métaboliques, une aide à l'annotation

Le choix de la représentation des résultats par l'intermédiaire des graphes des voies métaboliques de la base de données KEGG permet, pour l'analyse des résultats, de tenir compte du contexte de chaque fonction enzymatique. Ainsi, par exemple, la prédiction d'une fonction avec un score faible peut être confirmée ou infirmée en fonction des autres enzymes trouvées dans la voie métabolique où intervient cette enzyme. Cette évaluation est réalisée automatiquement dans un algorithme comme Pathologic (69) en calculant un score prenant en compte le nombre d'enzymes trouvées dans l'organisme sur le nombre total d'enzymes dans la voie métabolique.

6.1.3 Décomposition des enzymes en modules

Le principal avantage de la méthode PRIAM est d'avoir identifié des modules spécifiques des enzymes en se basant sur la classification des enzymes. Ainsi, par exemple, Shah et Hunter ont effectué la prédiction du numéro EC par une simple recherche de similarité par BLAST ou FASTA et ne permettent de discriminer que 40% des membres des différentes classes par similarité de séquence avec un seuil (120). Cette méthode a obtenu une efficacité de prédiction de fonction plus faible que la méthode PRIAM car ils n'ont pas tenu compte de la modularité des enzymes. Dans PRIAM presque 80% des profils possèdent un seuil permettant de discriminer les vrais positifs des faux positifs. Avec l'utilisation de ces seuils, 75% des profils ont une spécificité et une sensibilité parfaite.

Par ailleurs, la décomposition en module des enzymes nous a permis de rechercher des homologies locales avec une nouvelle protéine. Cette recherche permettant de mettre en évidence des multi-enzymes qui ne sont pas encore répertoriées dans les bases de données de séquences.

6.1.4 Prise en compte des caractéristiques des enzymes

En créant une collection enzymatique pour chaque activité de la nomenclature EC, nous avons pu identifier les collections comportant des enzymes analogues et / ou des enzymes oligomériques. De plus le programme PRIAM possède la spécificité de pouvoir identifier tous les types d'enzymes : les enzymes analogues et les enzymes oligomériques grâce à la vérification des règles, les multi-enzymes par la recherche de 'matchs' non-chevauchants avec les profils.

6.1.5 L'efficacité des profils

Les profils ou matrices de scores positions spécifiques sont de très bons descripteurs de régions protéiques similaires. En effet, ils permettent de caractériser une certaine diversité des séquences pour une même région, tout en mettant en évidence certaines positions plus conservées. Alors qu'une séquence consensus d'une région ne permet qu'une recherche de similarité avec l'acide aminé le plus représenté à une position d'un alignement de séquences.

6.2 Les limites de la méthode PRIAM

6.2.1 La détermination de valeurs seuils du score

La limite principale de cette méthodologie, mais aussi de toutes les méthodes d'analyse de séquences automatiques, est la détermination d'un seuil du score afin d'éliminer les similarités avec les faux positifs. Ainsi nous avons fait le choix de conserver la meilleure similarité pour chaque protéine. Une amélioration de la méthode qui ne sera alors plus totalement automatique mais qui permettra une annotation rapide des enzymes, serait de proposer à l'annotateur les 2 ou 3 meilleurs matchs obtenus avec les profils.

6.2.2 Les limites de la caractérisation des types d'enzymes par des règles

La détermination d'une règle logique pour chaque collection enzymatique dans le cas d'une collection avec une enzymes oligomérique ou un complexe enzymatique repose sur l'hypothèse que les modules composés de séquences observées dans tous les types d'organismes de la collection correspondent à des sous-unités indispensables à la fonction, alors que les autres modules correspondent à des sous-unités accessoires. Ainsi, cette hypothèse a pour conséquence de perdre l'identification de certaines sous-unités.

Par exemple la collection enzymatique de la glutamate synthase (EC 1.4.1.13) est caractérisée dans la base de données de profils PRIAM par un seul profil correspondant à la grande chaîne, sélectionné parmi 2 profils. En effet, l'utilisation de l'algorithme de détermination de règle dans cette collection décide que seul le profil 1p1.4.1.13 permet de caractériser la seule sous-unité observée dans toutes les espèces de la collection. Il semble, en effet, que chez les eucaryotes et chez *Bacillus subtilis*, il existe bien une seule chaîne en homotrimère alors que chez d'autres bactéries il s'agit d'un hétérodimère. Ainsi, avec la méthode PRIAM, il n'est pas possible d'identifier les 2 candidats pour la petite chaîne de la glutamate synthase chez *S. meliloti*. En effet, il y a 3 candidats identifiés avec l'environnement d'annotation iANT. Les protéines SMc01418 et SMc04026 sont annotées petite chaîne de la glutamate synthase et la protéine SMc04028 est annoté grande chaîne de la glutamate synthase.

6.2.3 Les enzymes avec différentes spécificités de substrats

Nous avons observé que certains profils n'ont pas une très bonne spécificité pour l'activité enzymatique recherchée car ils permettent de recruter différents types d'enzymes effectuant la même réaction mais avec des substrats différents. En effet, ces profils correspondent à des régions similaires à différentes enzymes.

Un bon exemple est celui des profils de la malate déshydrogénase et de la lactate déshydrogénase. Ces profils sont très similaires car ils correspondent à des enzymes très semblables. En effet, les sites actifs de ces deux enzymes se trouvent sur quelques acides aminés. De même les sites de spécificité pour les substrats et cofacteurs peuvent être changés par la substitution d'un seul acide aminé (121). Pour caractériser ce type d'enzymes il semble qu'un motif de type PROSITE soit plus spécifique. Une solution a ce problème pourrait être ici aussi la possibilité de donner les deux ou trois profils avec le meilleur score pour une protéine et laisser à l'annotateur le choix d'une fonction ou plusieurs fonctions dans le cas d'une enzyme avec différentes spécificités de substrats.

6.3 Des enzymes manquantes

La prédiction de voies métaboliques à partir d'un génome complet d'une espèce permet de mettre en évidence des enzymes manquantes par rapport à d'autres espèces (122). Ces enzymes manquantes peuvent s'expliquer par différentes causes:

- certaines activités enzymatiques ont été mises en évidence expérimentalement alors que la séquence protéique correspondante n'est pas encore détectée.
- la séquence de l'enzyme de cette espèce est très différente des séquences répertoriées dans les bases de données.
- les enzymes manquantes sont nombreuses dans une voie métabolique et en fait il y a quelques enzymes détectées dans la voie qui ont un rôle dans une autre voie métabolique (enzymes ubiquitaires).
- l'enzyme est vraiment manquante mais n'est pas indispensable pour le fonctionnement de la voie métabolique.

6.4 Des enzymes présentes mais inactives

On observe des cas d'enzymes prédites par l'annotation mais sans aucune activité enzymatique détectable dans les cellules de l'organisme, ceci pour plusieurs raisons:

- l'activité enzymatique n'est pas exprimée lors de l'expérience.
- l'annotation de séquence est correcte mais la protéine n'est pas transcrite ou bien elle a subi des modifications post-traductionnelles la rendant inactive.
- le gène ne possède plus la fonction de son ancêtre ou bien le gène n'est plus nécessaire pour le métabolisme de l'organisme, il pourrait être bientôt éliminé, c'est un cas d'enzyme vestige. Comme par exemple, l'arginine déiminase (ADI) chez *Mycoplasma genitalium* et *pneumoniae* (121). Le gène est trouvé mais ces deux organismes ne possèdent pas cette activité.

Une méthode d'annotation automatique comme celle-ci permet de diminuer fortement le temps d'annotation des enzymes d'un génome complet. Ainsi le génome de *S. meliloti* a été entièrement annoté avec l'environnement iANT en plusieurs mois alors que, l'ensemble des enzymes a été annoté en quelques heures avec l'outil PRIAM.

III
APPLICATION AU GENOME DE *SINORHIZOBIUM MELILOTI*

Ce chapitre présente une application du programme PRIAM au génome de *Sinorhizobium meliloti*, bactérie symbiote étudiée dans le laboratoire des interactions plantes microorganismes de l'INRA de Toulouse.

Dans une première partie, les caractéristiques biologiques de cet organisme sont développées afin de montrer les raisons de son choix comme bactérie symbiotique modèle. Les caractéristiques de son génome ainsi que l'organisation du séquençage de l'ensemble du génome sont décrits.

La deuxième partie est une analyse de la comparaison des prédictions des fonctions enzymatiques des protéines de *S. meliloti* avec la méthode semi-automatique iANT et avec la méthode automatique PRIAM.

La dernière partie est une analyse du métabolisme de *S. meliloti* que l'on peut déduire de ces deux méthodes d'annotation, en tenant compte des connaissances biologiques expérimentales de cette bactérie.

1 *Sinorhizobium meliloti*: bactérie symbiote

S. meliloti est une alpha-proteobactérie du groupe des rhizobia. Cette bactérie peut être soit à l'état libre dans le sol soit en symbiose avec des plantes de la famille des légumineuses, du genre *Medicago*, *Trigonella* et *Melilotus*. La symbiose est une association durable et profitable aux deux partenaires. Ainsi, les bactéries fixatrices d'azote comme *S. meliloti* associées aux légumineuses, réduisent l'azote atmosphérique en ammoniac qui est une source d'azote assimilable par la plante et nécessaire à sa croissance. En échange la plante offre un milieu et les nutriments favorables à la croissance de la bactérie.

La symbiose est possible grâce à l'induction de la formation d'un nouvel organe par la bactérie, le nodule, qui se forme au niveau des racines de la plante. Ce nodule est une niche écologique où la bactérie peut fixer l'azote atmosphérique pour les plantes. L'étude d'une bactérie de type rhizobium comme *S. meliloti* a été choisie pour plusieurs raisons d'intérêt biologique et agronomique. D'une part, pour étudier les différents processus biologiques qui surviennent au cours de l'interaction d'un rhizobium avec une légumineuse. D'autre part, pour comprendre plus spécifiquement la particularité des rhizobia qui est de fixer l'azote pour la plante. Enfin, l'étude de cette symbiose est intéressante du point de vue de l'écologie et de l'agronomie, sachant que les rhizobia permettent la fixation de la moitié de l'azote atmosphérique fixé par voie biologique.

Parmi les rhizobia, *S. meliloti* a été choisi comme organisme modèle pour les analyses moléculaires car c'est une bactérie avec une croissance rapide, utilisée en France comme inoculum. *S. meliloti* a aussi été choisi comme organisme modèle pour les analyses génétiques car il possède un génome de taille moyenne pour un rhizobium (6,8 Mb). Enfin, un de ses hôtes spécifiques, la luzerne (*Medicago*) a été retenu comme légumineuse modèle.

6.5 Bactérie modèle pour l'étude des interactions rhizobium-légumineuse

La symbiose est un processus biologique qui met en jeu des événements de différenciation et des changements métaboliques chez la bactérie et la plante. Ainsi, il existe un mécanisme d'infection de la bactérie par les poils absorbants des racines de la plante. Cette

interaction commence par la détection de signaux moléculaires (les facteurs Nod) par la plante qui induisent l'organogenèse du nodule. Puis, suit le développement d'une organisation particulière du nodule résultant d'une différentiation conjointe des cellules végétales et bactériennes (figure 23).

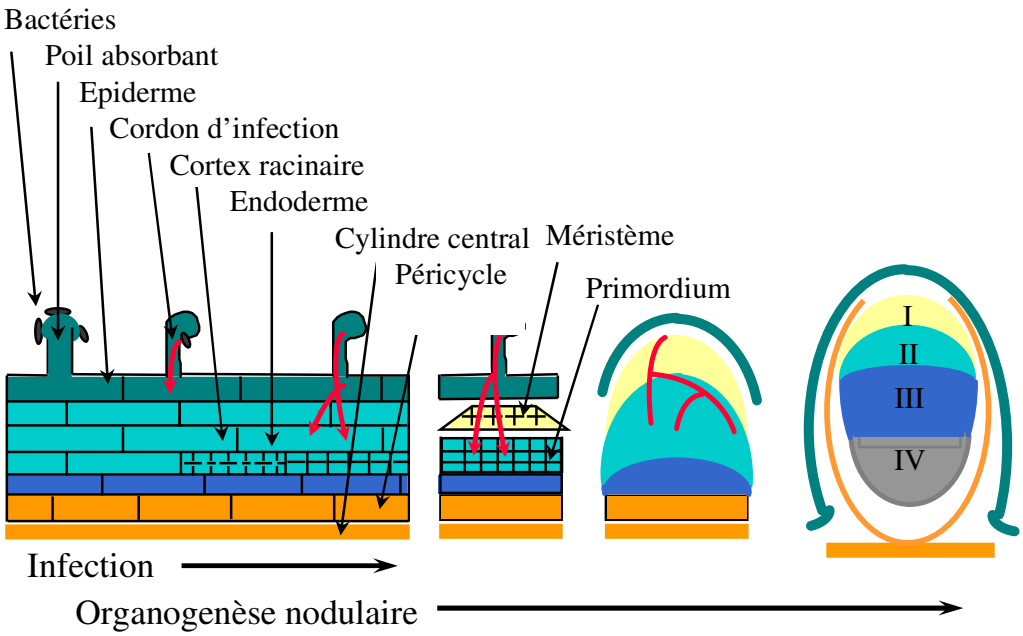


Figure 23 : Organogenèse d'un nodule de légumineuses.

La première étape de l'infection est l'attachement et l'agglutination de la bactérie aux poils absorbants de la racine de la plante. Cette interaction induit une déformation du poil absorbant à 360° appelée « crosse de berger ». Les bactéries pénètrent au niveau de la courbure grâce à la dégradation de la paroi végétale et à l'invagination de la membrane plasmique végétale formant un cordon d'infection. Ce cordon pénètre dans la racine et se différencie progressivement en primordium nodulaire. Les cellules de ce primordium se divisent activement pour former le méristème nodulaire. Un gradient de différentiation cellulaire se met en place du méristème jusqu'à la base du nodule. Les bactéries sont relâchées dans le cytoplasme des cellules du primordium nodulaire où elles se différencient en bactéroïdes fixateurs d'azote.

De nombreux gènes bactériens interviennent dans ces différentes étapes, dans le mécanisme d'infection (gènes exo), dans la nodulation de la plante (gènes *nod*) et dans le

fonctionnement du nodule avec tous les gènes impliqués dans la fixation de l'azote (gènes *nif* et *fix*).

6.6 Bactérie modèle pour la fixation symbiotique de l'azote

La fixation de l'azote atmosphérique (N_2) est une caractéristique de certaines bactéries et archéobactéries appelées organismes diazotrophes. Ces organismes peuvent fixer l'azote moléculaire pour leur propre croissance. Ils sont nommés fixateurs libres. Ou bien ils fixent l'azote pour la croissance de leur hôte, ce sont alors des fixateurs symbiotiques comme la bactérie *S. meliloti*.

La capacité à fixer l'azote est liée à l'existence chez la bactérie du complexe enzymatique nitrogénase permettant l'assimilation de l'azote en ammoniac. La plante assimile l'ammoniac comme source d'azote et fournit en échange des substrats carbonés issus de la photosynthèse pour la croissance de la bactérie (figure 24). La réaction enzymatique réalisée par la nitrogénase requiert beaucoup d'énergie. La plante fournit cette énergie à la bactérie par ses composés carbonés. Il faut 16 molécules d'ATP pour réduire une molécule d'azote. L'oxygène est indispensable pour la formation de l'ATP, cependant la nitrogénase est une enzyme dénaturée par l'O_2. C'est pourquoi, il existe toute une stratégie cellulaire et moléculaire pour contrôler la concentration en oxygène. Il existe d'une part, une barrière de diffusion à l'oxygène au niveau du cortex interne qui maintient la concentration entre 5 et 30 nM. Par ailleurs une hémoprotéine végétale, la leghémoglobine, permet une concentration élevée d'oxygène dans la zone de fixation du nodule (figure 25, d'après (123)). D'autre part, *S. meliloti* possède une chaîne respiratoire avec une haute affinité pour l'oxygène codée par l'opéron *fix*NOQP, permettant aux bactéroïdes de respirer dans un environnement pauvre en O_2.

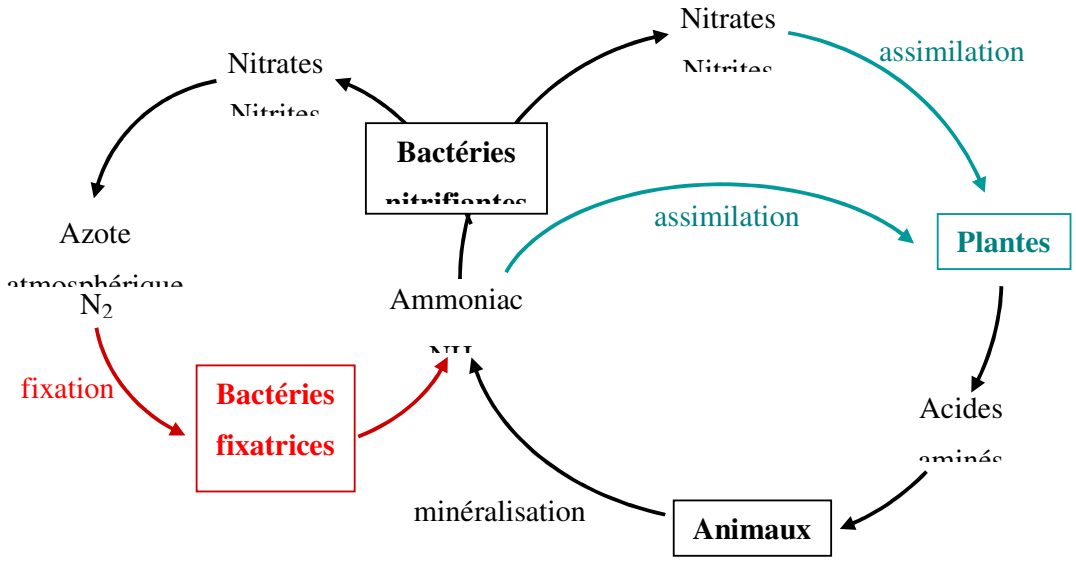


Figure 24 : Le cycle de l'azote, d'après Lehninger, 1993

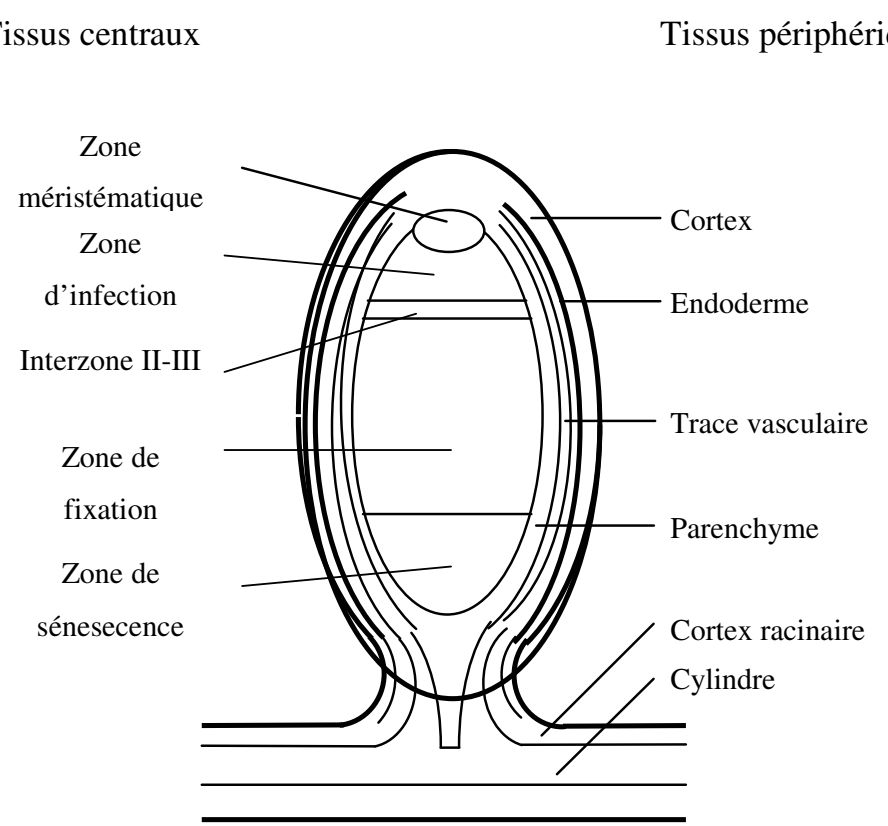


Figure 25: Organisation d'un nodule de luzerne, d'après Vasse et al., 1990

La deuxième phase de la fixation de l'azote est l'assimilation de l'ammonium par la plante. Bien que *S. meliloti* possède toutes les enzymes pour assimiler l'ammonium, celles-ci sont réprimées lors de la symbiose pour permettre l'exportation de l'ammonium vers la plante.

6.7 Symbiose d'intérêt écologique et agronomique

L'azote est un facteur limitant pour la production agricole. La principale source d'azote se trouve dans l'atmosphère alors que la plupart des organismes récupèrent l'azote sous forme d'ammoniac, de nitrate, de nitrite ou dans les acides aminés pour les animaux. Les seuls organismes capables de fixer l'azote atmosphérique sont des procaryotes tel que les rhizobia.

En effet, les rhizobia réduisent la moitié de l'azote atmosphérique fixé par voie biologique, l'équivalent de ce que fixent les industries et 30% de l'azote total fixé dans le cycle. Il est intéressant de cultiver des légumineuses, ceci sans ajout d'engrais azotés. Les légumineuses sont des végétaux riches en protéines, avec un fort intérêt alimentaire et pourraient remplacer l'utilisation des farines animales maintenant interdites sachant les conséquences épidémiologiques de cette source de protéines.

S. meliloti a la capacité d'avoir une relation symbiotique avec un nombre limité de légumineuses, mais d'autres Sinorhizobium tel que *Sinorhizobium* sp. NGR234 peuvent interagir avec plus de 110 légumineuses et une non-légumineuse, *Parasponia* (124). Des recherches pour acquérir un transfert d'une capacité de fixation symbiotique de l'azote à d'autres végétaux comme les céréales favoriseraient aussi la diminution de l'utilisation des engrais azotés comme les nitrates, polluant l'eau.

6.8 Organisation du génome de *Sinorhizobium meliloti*

Le génome de *S. meliloti* est formé de trois réplicons : un chromosome de 3,7 Mb et deux mégaplasmides de 1,4 et 1,7 Mb appelés pSyma et pSymb. Le grand réplicon est appelé chromosome car il porte les gènes dits de ménage et les opérons portant les ARNr.

Le génome de *S. meliloti* possède un grand nombre de régions répétées de longueur variable. Elles peuvent avoir une centaine de paires de bases comme les éléments mosaïques RIME, ou les palindromes « A, B et C ». Ces séquences sont spécifiques des rhizobia mais leur rôle biologique reste à élucider. Les régions répétées peuvent aussi faire plusieurs kilobases comme les séquences d'insertions, les duplications de gènes ou les clusters de gènes. Le génome de *S. meliloti* est aussi caractérisé par un pourcentage en GC élevé, de 60 à 62%.

6.9 Projet international de séquençage du génome de *Sinorhizobium meliloti*

Les approches de génétique ou de biologie moléculaire pour l'identification des fonctions des gènes régulés en symbiose ont révélé leurs limites. Pour mieux comprendre à l'échelle moléculaire les changements survenant lors d'une interaction symbiotique, il a été décidé de déterminer l'ensemble des gènes de la bactérie symbiote. C'est pourquoi, l'analyse

globale du génome de *S. meliloti* a été réalisée en effectuant le séquençage complet de son génome.

Un projet international de séquençage de la totalité du génome regroupant des laboratoires européens et américains a été mis en place. L'organisation de la répartition du séquençage s'est basée sur l'organisation du génome en trois réplicons (figure 26).

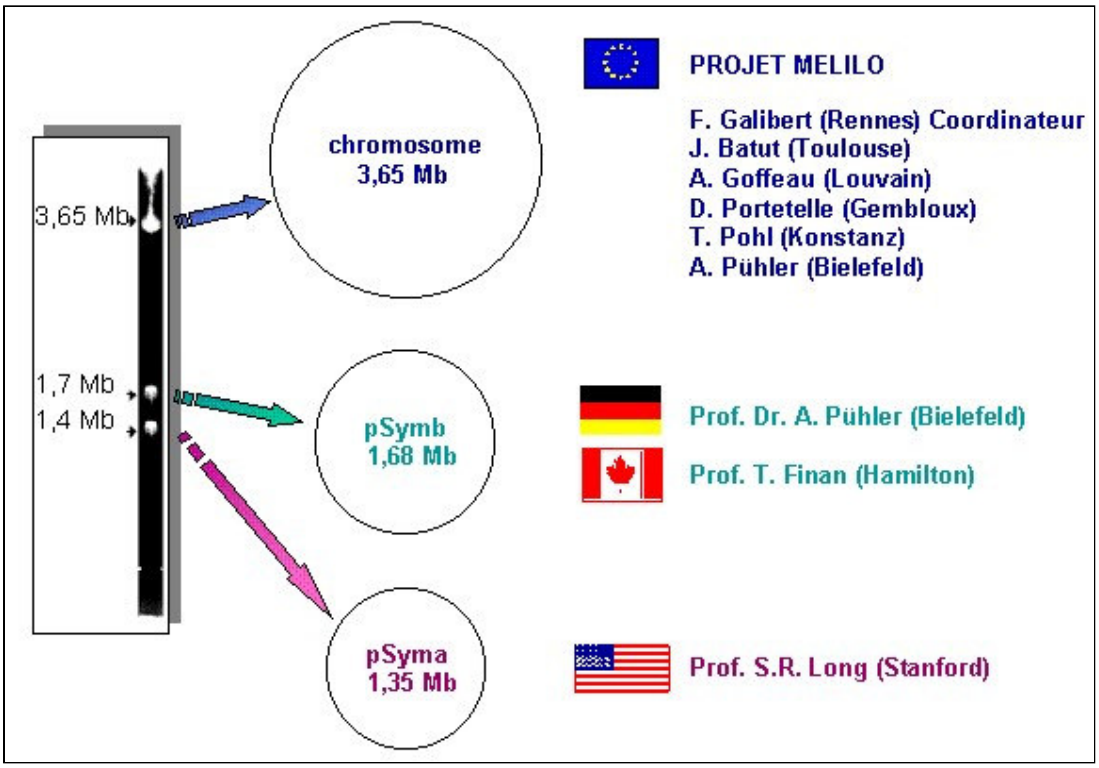


Figure 26: Projet international de séquençage du génome de S. meliloti.

Ainsi, le séquençage du chromosome (3,7 Mb) a été pris en charge par le consortium européen MELILO comprenant 6 laboratoires coordonnés par le Pr. F. Galibert (UPR 41 CNRS Recombinaisons Génétiques à Rennes). La stratégie d'un séquençage en deux étapes a été retenue. Celle ci consiste tout d'abord à construire un contig de clones recombinants portant de larges inserts, puis à entreprendre le séquençage par fragmentation aléatoire d'un minimum de clones. Pour faciliter le séquençage, en sélectionnant un nombre minimum mais suffisant de clones pour couvrir tout le chromosome, une cartographie fine du chromosome a été réalisée se basant sur la construction d'un contig de BACs ordonnés (125).

Le séquençage du mégaplasmide pSymb (1,7 Mb) a été réalisé en parallèle par une équipe canadienne (Prof. Turlough Finan, Université d'Hamilton, Canada) et une équipe allemande (Prof. Dr. Alfred Pühler, Université de Bielefeld, Allemagne). L'équipe canadienne a séquencé par fragmentation aléatoire cinq grandes régions non chevauchantes sur une longueur de 550 kb, clonées dans un vecteur BAC. L'équipe allemande avec la collaboration de la société LION Bioscience a séquencé l'intégralité du mégaplasmide pSymb, en utilisant la stratégie en deux étapes comme pour le chromosome, avec au préalable une cartographie fine du réplicon (126).

Le séquençage du mégaplasmide pSyma (1,4 Mb) a été effectué par une équipe américaine (Prof. Sharon R. Long, Department of Biological Sciences & Howard Hughes Medical Institute, Stanford). Une stratégie de séquençage en une étape a été retenue. Elle consiste à linéariser l'ADN du réplicon par une enzyme puis à le purifier par champ pulsé et enfin à le fragmenter de façon aléatoire en fragments de l'ordre de 1 à 2 kb. Les fragments sont alors séquencés à partir du fragment cloné dans un vecteur. De même que pour les deux autres réplicons, une carte physique de BAC couvrant l'ensemble du pSyma a été réalisée afin de faciliter l'assemblage des fragments de séquences (127).

Les trois réplicons ont été entièrement séquencés à la fin de l'année 2000 et l'ensemble des séquences obtenues a été annoté par les différents membres du projet international (106, 119) (128) (129) en utilisant l'environnement d'annotation semi-automatique iANT développé à Toulouse par Patricia Thebault et Jérôme Gouzy (75).

7 L'annotation du génome de *Sinorhizobium meliloti*

L'annotation d'un génome complet nécessite de réaliser de nombreuses analyses de séquences allant de la détection des gènes, des ARN, à la prédiction des fonctions. De nombreux outils et méthodologies existent pour annoter les génomes. Les méthodologies sont de trois types : manuelles, semi-automatiques et entièrement automatiques. Nous avons ici réaliser une analyse de l'annotation des fonctions enzymatiques de *S. meliloti* avec un environnement d'annotation semi-automatique iANT et un outil automatique PRIAM.

7.1 Annotation semi-automatique du génome avec iANT

Les 6,8 Mb du génome de *S. meliloti* ont été annotés avec l'environnement d'annotation semi-automatique iANT développé par Patricia Thebault et Jérôme Gouzy (75) dans le laboratoire des interactions plantes microorganismes (outil développé dans le chapitre I). L'outil iANT est un environnement d'annotation comportant une automatisation des analyses de séquences et une interface d'annotation conviviale permettant aux annotateurs de choisir facilement l'annotation de chaque gène. Il contient des outils de détection de gènes et de décalage de lecture (FRAMED (130)), des outils d'analyse de séquences nucléiques (Blastn, tRNAscan-SE, …) et de séquences protéiques (Blastp, ProDom, PROSITE, Tmpred, PSORT, …). Les trois réplicons ont été annotés respectivement par les trois groupes qui les ont séquencés. L'expérience des annotateurs permet de mesurer qu'il faut 3 jours pour annoter 100 kb de séquence génomique bactérienne avec iANT. Dans le génome de *S. meliloti*, 6204 gènes ont été détectés. Parmi les gènes prédits, 10,7 % ne présentent aucune similarité avec des gènes présents dans les bases de données. L'ensemble de la séquence annotée est accessible sur la page web suivante :

http://sequence.toulouse.inra.fr/meliloti.html

Parmi les différents champs d'annotation remplis dans l'environnement iANT, le champ activité enzymatique donne l'indication d'un numéro EC quand la similarité de séquence semblait significative pour l'annotateur. Il y a 808 protéines annotées avec un ou plusieurs numéros EC complets et 221 protéines avec un numéro EC tronqué de type 1.1.-.-. Ce choix de précision a été fait car il était difficile de préciser le substrat ou le cofacteur de la

réaction enzymatique. Ainsi 532 numéros EC différents ont été trouvés avec l'outil iANT dans le génome de *S. meliloti*.

7.2 Annotation automatique du génome avec PRIAM

Le programme PRIAM a été utilisé pour annoter de façon automatique l'ensemble des 6204 protéines de *S. meliloti*. Cette opération met 1h40 sur un pentium III 699 MHz sous Linux. Nous avons obtenu une prédiction de fonction enzymatique avec un ou plusieurs numéros EC attribués pour 1460 protéines. Quatorze protéines sont annotées comme étant bifonctionnelles. Ces prédictions correspondent à 660 numéros EC différents. Parmi ces numéros EC, 353 sont prédits avec un score de RPS-BLAST supérieur au seuil défini contre la banque SWISS-PROT (numéros EC coloriés en vert dans les graphes de KEGG) et 307 numéros EC avec un score de RPS-BLAST inférieur au seuil (numéros EC coloriés en jaune dans les graphes de KEGG). Les résultats de PRIAM sur *S. meliloti* sont visualisés sur les graphes des voies métaboliques de la base de donnée KEGG et sont consultables sur la page web :
http://genopole.toulouse.inra.fr/bioinfo/priam/RES_PACKAGE/RES_MELILO/pred_metabo_melilo.html

Les graphes des voies métaboliques de la base de données KEGG sont des images fixes représentant les voies métaboliques les plus connues. Toutes les enzymes de toutes les espèces sont représentées par des rectangles contenant un numéro EC. C'est pourquoi pour une réaction enzymatique transformant un substrat en un produit on peut observer plusieurs numéros EC correspondant à différentes enzymes capables de catalyser la réaction.

Sur chaque voie métabolique les numéros EC prédits par l'outil PRIAM sont identifiés en jaune ou vert, alors que les numéros EC décrits dans ENZYME mais ne possédant pas de séquences correspondantes et donc pas de profil sont en gris. Chaque numéro EC jaune ou vert est cliquable et permet d'accéder à une liste de prédictions des séquences protéiques trouvées pour ce numéro EC. Ainsi cette liste contient les 1460 protéines avec le ou les numéros EC prédits. Chaque protéine avec sa prédiction fonctionnelle est un lien vers l'entrée du gène prédit par iANT. Il est donc possible de comparer l'annotation obtenue avec PRIAM avec l'annotation trouvée avec iANT. Cependant, cette observation n'est possible que par un travail fastidieux de consultation de chaque numéro EC et de chaque protéine correspondante.

C'est pourquoi nous avons décidé de réaliser une comparaison de ces annotations avec une visualisation graphique des résultats.

Par ailleurs, la représentation avec les graphes de KEGG a aussi été utilisée pour représenter les résultats de l'annotation iANT selon la localisation sur les réplicons. Les résultats sont consultables sur cette page :
http://genopole.toulouse.inra.fr/bioinfo/priam/RES_PACKAGE/RES_MELILO/metabo_melilo_parep.html

Un code couleur est utilisé pour différencier les trois réplicons et une autre couleur (gris) est utilisée lorsque des gènes de différents réplicons sont trouvés pour une même fonction enzymatique. Cette représentation permet une analyse rapide de la localisation des gènes d'une voie métabolique.

7.3 Comparaison des annotations iANT avec les prédictions PRIAM

7.3.1 Visualisation de la comparaison des résultats

Nous avons réalisé une comparaison des numéros EC obtenus pour le génome de *S.meliloti* avec les méthodes iANT et PRIAM. Pour représenter les résultats de cette comparaison nous avons aussi utilisé les graphes des voies métaboliques de KEGG en représentant les accords et les désaccords entre les deux méthodes par un code couleur. Ainsi les numéros EC prédits par les deux méthodes sont en verts (vert clair = en dessous du seuil évalué contre Swiss-Prot et vert foncé = au dessus du seuil), les numéros EC prédits seulement par la méthode PRIAM sont en jaune (= en dessous du seuil) et orange (= au dessus du seuil) et enfin les numéros EC prédits seulement par l'annotation iANT sont en rouge et violet. Les numéros EC en violet sont des collections enzymatiques sans séquences protéiques dans la version 27.0 d'ENZYME mais pour lesquels des prédictions ont étés obtenues grâce à des homologies avec des séquences d'autres banques comme SP-TrEMBL. Les résultats de cette comparaison sont visualisés sur la page web :
http://genopole.toulouse.inra.fr/bioinfo/priam/RES_PACKAGE/RES_MELILO/pred_metabo_melilocomp.html

De même que pour la représentation des résultats de l'outil PRIAM, chaque numéro EC est un lien hypertexte. Nous proposons un lien hypertexte des numéros EC prédits en accord par les deux méthodes et des numéros EC prédits seulement par l'annotation iANT

vers la ou les entrées des gènes dans l'interface iANT. Les numéros EC prédits seulement par la méthode PRIAM sont des liens hypertexte vers la liste de prédiction.

7.3.2 Analyse de la comparaison des résultats

La comparaison des résultats obtenus par les deux méthodes (Tableau 2) permet de remarquer que le nombre de numéros EC prédits en commun par les deux méthodes (493 EC) est proche du nombre de numéros EC prédits par la méthode iANT (532 EC).

	iANT	**PRIAM avec E = 10^{-10}**
EC prédits	532	660
Protéines avec un EC prédit	808	1460
EC prédits que par iANT	39	
EC prédits que par PRIAM		167
EC prédits en commun	493	

Tableau 2 : Comparaison des résultats de l'environnement d'annotation iANT et de la méthode PRIAM sur S. meliloti.

Trente huit numéros EC n'ont pas été trouvés par l'outil PRIAM, parmi eux 7 numéros EC (violet) ne pouvaient effectivement pas être trouvés car aucun profil ne pouvait être construit à partir de la base ENZYME. Les 31 numéros EC restants n'ont pas pu être attribués pour plusieurs raisons. Le plus souvent ces numéros EC correspondent à des activités enzymatiques avec une activité pour un substrat particulier alors que les profils correspondants ne sont pas très spécifiques des substrats. Ainsi, 14 numéros EC n'ont pas été

trouvés car les protéines ont été annotées avec une activité similaire pour un autre substrat. De même, 10 numéros EC n'ont pas été attribués car ils correspondent à des protéines avec plusieurs spécificités de substrat donc plusieurs numéros EC ont été donnés par l'annotation alors que la méthode PRIAM annote selon la meilleure homologie (un seul numéro EC). Enfin 7 numéros EC ne sont pas trouvés par l'outil PRIAM car les candidats proposés par l'annotation iANT ne présentent aucune similarité significative contre la banque de profils utilisés par la méthode PRIAM.

Il y a par ailleurs 167 numéros EC proposés par la méthode PRIAM dont 16 numéros EC avec une similarité dont le score est supérieur au seuil détecté contre Swiss-Prot (numéros EC coloriés en orange dans les graphes de KEGG). Dans le tableau 3 nous avons représenté une synthèse de ces 16 cas. On observe 4 numéros EC (en jaune dans le tableau 3) différents avec les deux méthodes. Mais ces fonctions correspondent à des activités enzymatiques similaires. Ils ne diffèrent que par le substrat ou le cofacteur. L'outil PRIAM permet de préciser l'activité enzymatique pour trois cas (en bleu dans le tableau 3) et sur les 9 cas de numéros EC prédits seulement par la méthode PRIAM, 6 semblent être très probables (en vert dans le tableau 3).

Les autres numéros EC proposés par la méthode PRIAM permettent le plus souvent de préciser les numéros EC partiels des 221 protéines annotées de cette manière par l'annotation iANT. Cependant, il est nécessaire d'analyser chaque proposition. Quelques exemples de prédictions de la méthode PRIAM développées dans la partie suivante semblent très probables dans le contexte d'une voie métabolique.

EC PRIAM	fonction	% Identité	E value	EC iANT	fonction
1.1.1.27	L-lactate dehydrogenase	35	2.10-95	1.1.1.37	Malate dehydrogenase
1.1.1.94	Glycerol-3-phosphate dehydrogenase (NAD(P)+)	49	2.10-94	1.1.1.8	Glycerol-3-phosphate dehydrogenase (NAD+)
1.6.1.2	NAD(P)(+) transhydrogenase (AB-specific)	61	6.10-23	1.6.1.1	NAD(P)(+) transhydrogenase (B-specific)
2.6.1.76	Diaminobutyrate-2-oxoglutarate transaminase	78	0	2.6.1.46	Diaminobutyrate--pyruvate aminotransferase

1.2.1.27	Methylmalonate-semialdehyde dehydrogenase (acylating)	48	10-162	1.2.1.-	putative malonic semialdehyde oxidative decarboxylase
2.7.1.95	Kanamycin kinase	42	2.10-39	2.7.1.-	putative aminoglycoside 3'-phosphotransferase
5.1.3.12	UDP-glucuronate 5'-epimerase	100	0	5.1.3.-	UDP-glucuronic acid epimerase
1.8.4.6	Protein-methionine-S-oxide reductase	56	1.10-66		probable petide methionine sulfoxide reductase
2.4.1.15	Trehalose-6-phosphate synthase	41	0		probable Trehalose-6-phosphate synthase
2.7.1.24	Dephospho-CoA kinase	38	1.10-41		conserved hypothetical protein
3.1.4.14	[Acyl-carrier protein] phosphodiesterase	36	5.10-42		probable [Acyl-carrier protein] phosphodiesterase
3.2.1.73	Endo-beta-1,3-1,4 glucanase	38	2.10-44		Endo-beta-1,3-1,4 glycanase
3.6.3.12	Potassium-transporting ATPase	52	0		probable Potassium-transporting ATPase A chain
3.1.5.1	dGTPase	27	2.10-47		conserved hypothetical protein
3.3.2.3	Epoxide hydrolase	27	4.10-13		putative Epoxide hydrolase
3.6.1.13	ADP-ribose pyrophosphatase	30	9.10-12		conserved hypothetical protein

Tableau 3: Comparaison des annotations obtenues avec iANT et PRIAM.

en jaune: numéros EC différents.

en bleu: précision du numéro EC par PRIAM.

en vert: prédiction du numéro EC par PRIAM très probable.

en blanc: prédiction du numéro EC par PRIAM.

8 Analyse des voies métaboliques de *S. meliloti* obtenues avec la méthode PRIAM

La prédiction des fonctions enzymatiques à partir du génome de *S. meliloti* nous permet de déduire l'ensemble des voies métaboliques possibles de cet organisme. Cette connaissance du métabolisme est un outil important pour interpréter les résultats d'expériences obtenues avec le transcriptome ou le protéome de la bactérie. En effet, la bactérie opère des changements considérables de son métabolisme lorsqu'elle entre en symbiose avec la plante et qu'elle se transforme en bactéroïde. Ainsi, des gènes sont spécifiquement induits lorsque le bactéroïde est dans le nodule. C'est pourquoi, il est nécessaire d'identifier les fonctions des gènes et situer dans quelles voies métaboliques ils agissent pour comprendre les changements physiologiques qui surviennent dans la bactérie.

Dans cette section, nous présentons les prédictions des voies métaboliques avec les résultats de l'annotation semi-automatique et de l'annotation automatique afin de décrire le métabolisme de *S. meliloti* de la façon la plus complète.

8.1 Métabolisme des carbohydrates

8.1.1 Les grandes voies de dégradation des carbohydrates

8.1.1.1 La glycolyse

La glycolyse est une chaîne de réactions enzymatiques qui permet à une cellule d'obtenir de l'énergie à partir du glucose. Cette voie est connue chez de très nombreux organismes.

Chez *S. meliloti* la voie de la glycolyse (figure 27) ou voie d'Embden-Meyerhof-Parnas, voie de dégradation des carbohydrates en pyruvate est presque complète, seule l'activité 6-phosphofructokinase (EC 2.7.1.11) semble manquante. Cependant, une enzyme homologue de la 6-phosphofructokinase a été identifiée chez *S. meliloti*, elle fonctionne avec le pyrophosphate inorganique (EC 2.7.1.90 dans le graphe du métabolisme du fructose et du mannose (figure 30)) au lieu de l'ATP (EC 2.7.1.11). Des expériences de mutants d'*E. coli*

montrent que cette enzyme ne possède pas l'activité dans le sens de la glycolyse (131). Il serait donc nécessaire de déterminer le ou les sens de cette enzyme chez *S. meliloti*. Cette étape de la glycolyse est aussi réalisée par une voie passant par le cycle des pentoses (figure 28). Cette voie permet d'obtenir le D-glyceraldehyde-3P de la glycolyse. En effet, l' α-D-glucose-6P est transformé en D-ribulose-5P puis en D-xylulose-5P, ce dernier est ensuite transformé grâce à une transketolase (EC 2.2.1.1) en D-glyceraldehyde-3P.

Toutes les enzymes de la glycolyse de *S. meliloti* sont codées par des gènes sur le chromosome. Seulement 2 enzymes, la phosphoglucomutase (EC 5.4.2.2) et la fructose biphosphate aldolase (EC 4.1.2.13) sont codées par des gènes présents sur le réplicon pSymb. La fonction fructose biphosphate aldolase est en effet représentée par 2 classes de protéines. La classe I présente sur le chromosome et la classe II sur le réplicon. Nous avons dans la collection enzymatique de cette fonction, un profil pour chacune de ces classes.

La suite de réactions du cycle des pentoses phosphate utilisée par la glycolyse comporte des enzymes sur le chromosome et une seule enzyme, une transketolase (EC 2.2.1.1) sur pSymb de même type que deux protéines sur le chromosome.

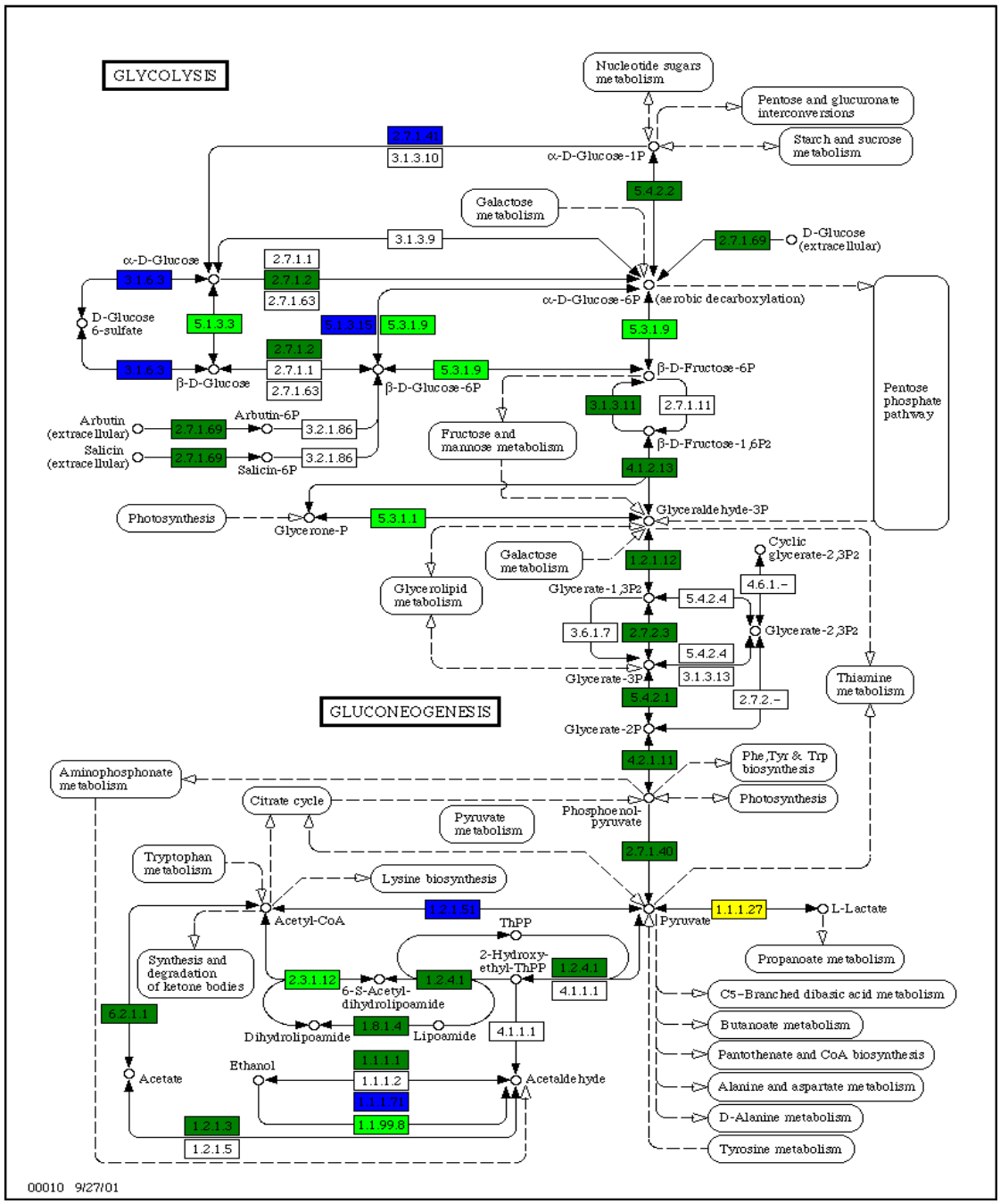


Figure 27: Voie de la glycolyse chez S.meliloti.

Utilisation du code couleur de comparaison des annotations iANT et PRIAM.

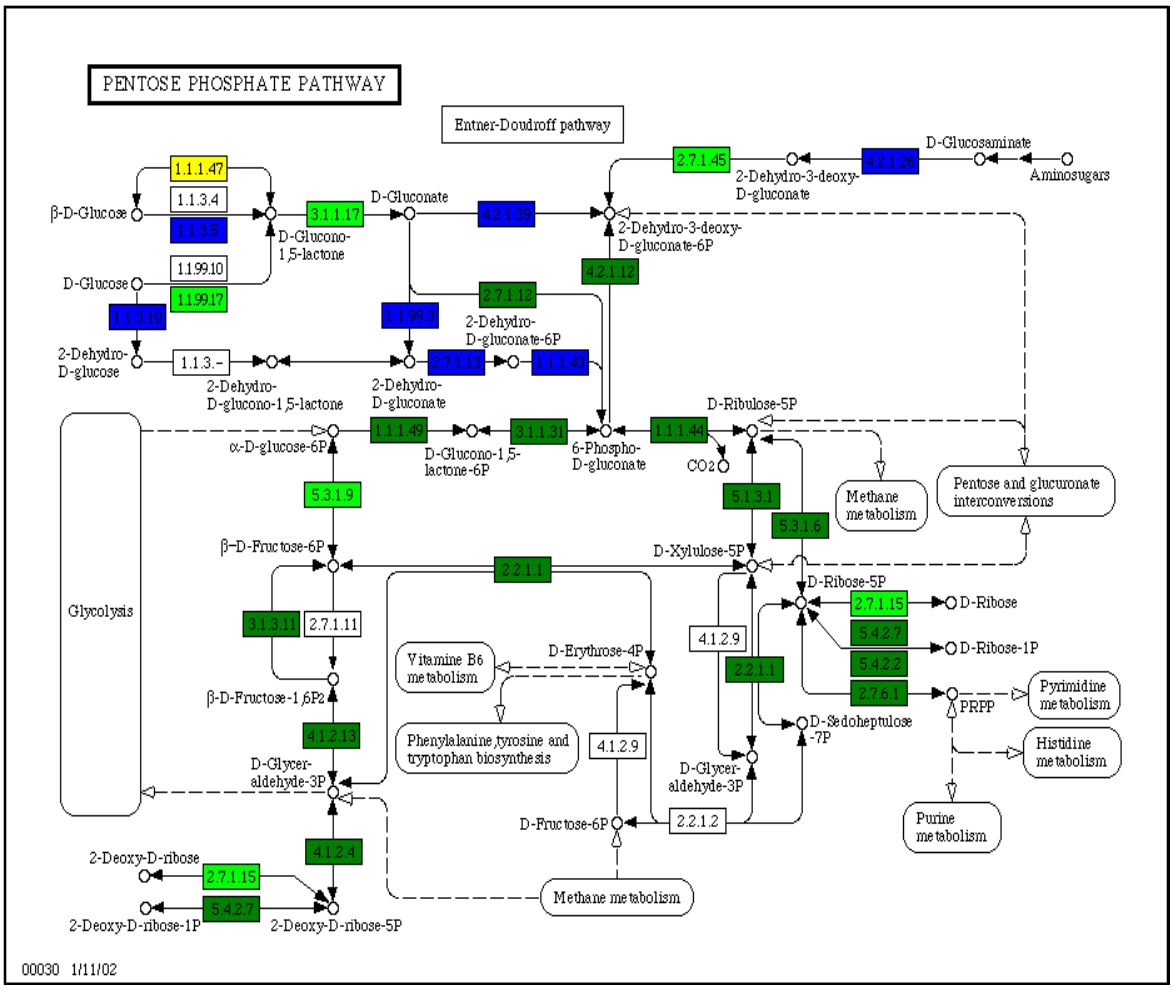


Figure 28 : Cycle des pentoses phosphate et voie d'Entner-Doudoroff.

8.1.1.2 Le cycle des pentoses phosphate et la voie d'Entner-Doudoroff

Le cycle des pentoses (figure 28) est un cycle indispensable pour les organismes prototrophes pour la synthèse des acides nucléiques, des chaînes latérales de l'histidine et du tryptophane. Ce cycle est complet chez *S.meliloti*. La voie d'Entner-Doudoroff**,** voie alternative pour l'utilisation du glucose et du gluconate par le 6-phosphogluconate est aussi complète.

Dans ces 2 voies, 3 enzymes sont codées par des gènes localisés sur pSymb. Quatre enzymes sont codées par des gènes localisés à la fois sur le chromosome et pSymb et la gluconolactonase (EC 3.1.1.17) possède une copie de son gène sur les réplicons pSyma et pSymb. Des analyses d'enzymes et de radio-respirométrie (132) ont permit d'établir que le

catabolisme du glucose et du gluconate dans la bactérie à l'état libre dans le sol s'effectue par la voie d'Entner-Doudoroff plutôt que par la glycolyse. De même en symbiose, la glycolyse semble peu importante alors que la gluconéogenèse doit être essentielle dans le bactéroïde. Celui-ci reçoit de son hôte des composés en C4, intermédiaires du cycle de Krebs, tel que le fumarate, le malate et le succinate comme source principale de carbone donc d'énergie pour la fixation de l'azote.

8.1.2 La gluconéogenèse

La gluconéogenèse, permettant la resynthèse du glucose à partir du pyruvate est complète chez *S. meliloti* (figure 27 et 30). En effet, on identifie les enzymes qui se substituent aux 2 réactions irréversibles, fortement exothermiques de la glycolyse : la pyruvate kinase (EC 2.7.1.40) et la phosphofructokinase (EC 2.7.1.11).

Ainsi, la réaction catalysée par la pyruvate kinase (EC 2.7.1.40) est une réaction difficile dans le sens de la gluconéogenèse. Quatre voies sont connues pour transformer le pyruvate en phosphoenol-pyruvate. Trois sont présentes chez *S. meliloti* par prédiction de fonction.

8.1.2.1 La voie des enzymes maliques

La gluconéogenèse passant par la voie des enzymes maliques est présente chez *S. meliloti* (figure 29). On identifie 2 types d'enzymes maliques, l'enzyme malique NAD dépendante (EC 1.1.1.39) et l'enzyme malique NADP dépendante (EC 1.1.1.40). Ces enzymes sont capables d'agir de façon réversible pour transformer le pyruvate en L-malate. Cependant, expérimentalement chez *S. meliloti* il n'a pas pu être mis en évidence un niveau détectable de formation de L-malate (133). Ces enzymes ont une action de décarboxylation oxydative du malate en pyruvate. Les enzymes maliques trouvées chez *S. meliloti* ont la particularité d'être chimériques. La partie N-terminale est similaire aux autres enzymes maliques connues alors que la partie C-terminale est similaire à des phosphates acetyltransférases (EC 2.3.1.8). Nous avons en effet identifié avec MKDOM2, deux régions dans les classes enzymatiques 1.1.1.39 et 1.1.1.40, dues à ce type d'enzymes chimériques. Seules les régions portant soit l'activité 1.1.1.39 soit 1.1.140 sont sélectionnées par l'algorithme de sélection des profils suffisants pour caractériser la fonction. Dans le graphe du métabolisme du pyruvate (figure 29), on remarque que PRIAM n'a pas prédit l'enzyme

malique NAD dépendante (EC 1.1.1.39). En effet la protéine SMc00169 (gène *dme*) prédite par iANT, est trouvée avec une meilleure similarité pour le profil de l'enzyme malique NADP dépendante (EC 1.1.1.40). Expérimentalement, l'équipe de Voegele (133) a trouvé que le gène *dme* code pour une protéine avec une activité avec NAD+ et avec NADP+, caractéristique de certaines enzymes maliques 1.1.1.39.

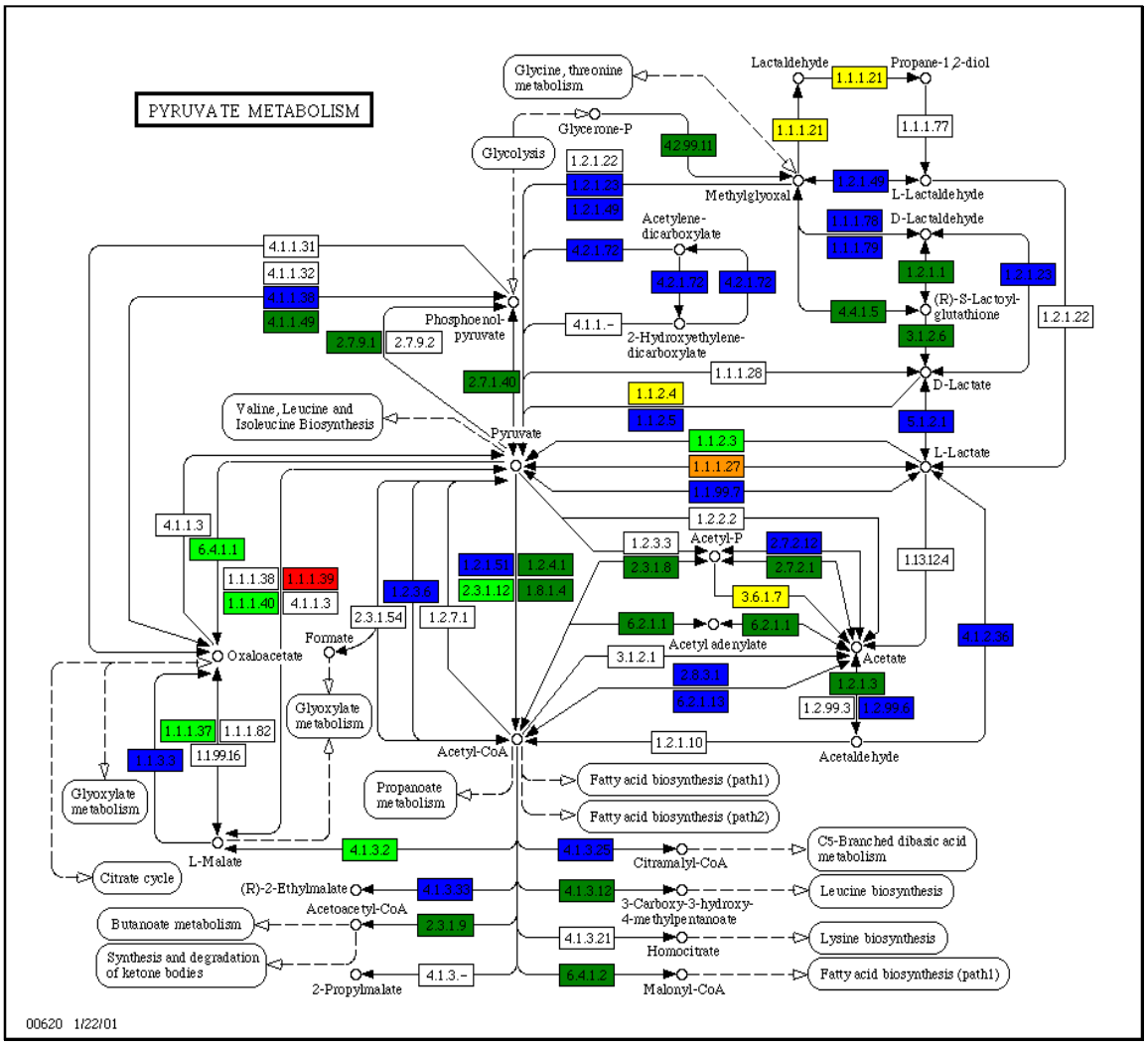


Figure 29: Le métabolisme du pyruvate.

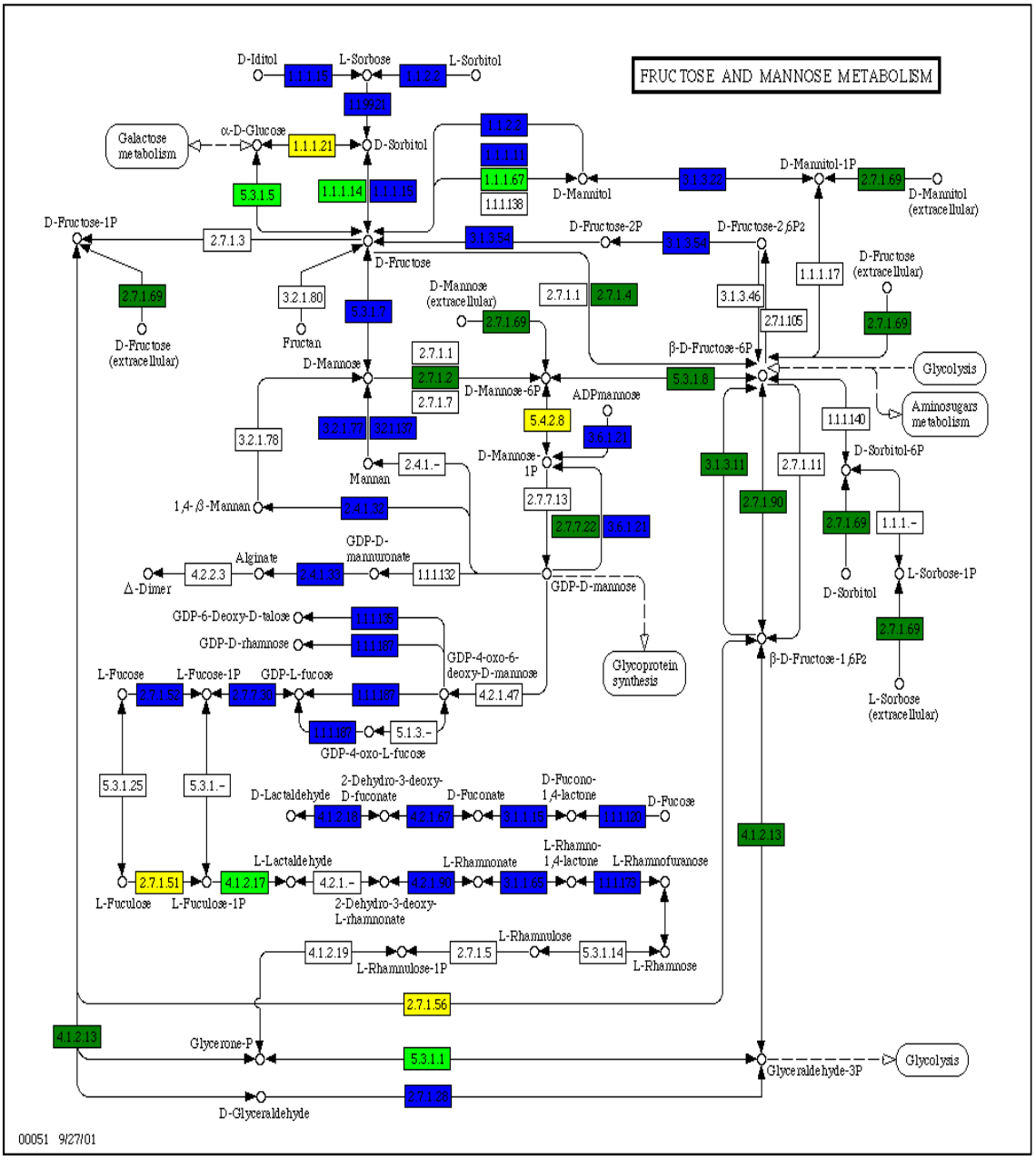


Figure 30 : Voie du métabolisme du fructose et du mannose.

8.1.2.2 La phosphoenol-pyruvate synthétase ou la pyruvate phosphate dikinase

Un autre type de synthèse est connu chez les bactéries pour produire du phosphoenol pyruvate à partir du pyruvate. Deux voies avec une enzyme unique sont connues : soit la phosphoenol-pyruvate synthétase (EC 2.7.9.2) qui n'a pas été identifiée par homologie de

séquence chez *S. meliloti*, soit la pyruvate phosphate dikinase (PPDK) (EC 2.7.9.1). Pour cette dernière un très bon candidat a été identifié avec 63% d'identité pour la séquence consensus de l'unique profil qui caractérise l'enzyme pyruvate phosphate dikinase (EC 2.7.9.1).

8.1.2.3 La pyruvate carboxylase et la phosphoenolpyruvate carboxykinase

Enfin, un cycle futile passant par 2 enzymes successives est identifié chez *S. meliloti*, la pyruvate carboxylase (EC 6.4.1.1) transformant le pyruvate en oxaloacetate (rôle anaplérotique), suivie de la phosphoenolpyruvate carboxykinase (PCK) (EC 4.1.1.49) synthétisant le phosphoenol-pyruvate. Des expériences de génétique avec des mutants (134) montrent que l'enzyme PCK est la voie principale chez *S. meliloti* mais que l'enzyme PPDK avec l'activité combinée des enzymes maliques est une voie alternative possible pour la formation de phosphoenol-pyruvate. Toutes ces enzymes sont codées par des gènes localisés sur le chromosome.

8.1.2.4 La fructose-1,6-biphosphatase et la 6-phosphofructokinase

La deuxième étape de la glycolyse qui n'est pas réversible est la réaction réalisée par la phosphofructokinase (EC 2.7.1.11). Cette étape dans le sens de la gluconéogenèse est réalisée par la fructose-1,6-biphosphatase (EC 3.1.3.11) ou par la 6-phosphofructokinase fonctionnant avec le pyrophosphate inorganique (EC 2.7.1.90) qui sont identifiées chez *S. meliloti*.

8.1.3 Métabolisme des autres substrats carbonés

Nous allons maintenant nous intéresser à l'identification des autres substrats carbonés métabolisés par *S. meliloti*.

Dans la voie du métabolisme du fructose et du mannose (figure 30), on trouve 3 hexoses qui sont connus comme pouvant être dédradés par *S. meliloti* : le fructose, le mannose et le D-mannitol. Le fructose est transformé par une fructokinase (EC 2.7.1.4) en β-D-fructose-6P qui est un composant de la glycolyse et de la voie d'Entner-Doudoroff. Le mannose est lui aussi transformé en β-D-fructose-6P en passant par l'intermédiaire du D-mannose-6P. De même pour le D-mannitol en passant par le D-fructose.

Dans la voie du métabolisme du galactose (figure 31), on trouve un hexose (galactose), 2 disaccharides (lactose, sucrose) et un trisaccharide (raffinose) qui peuvent être métabolisés. Tous ces substrats carbonés sont utilisés par la voie d'Entner-Doudoroff en passant par le D-glucose puis l'α-D-glucose-6P.

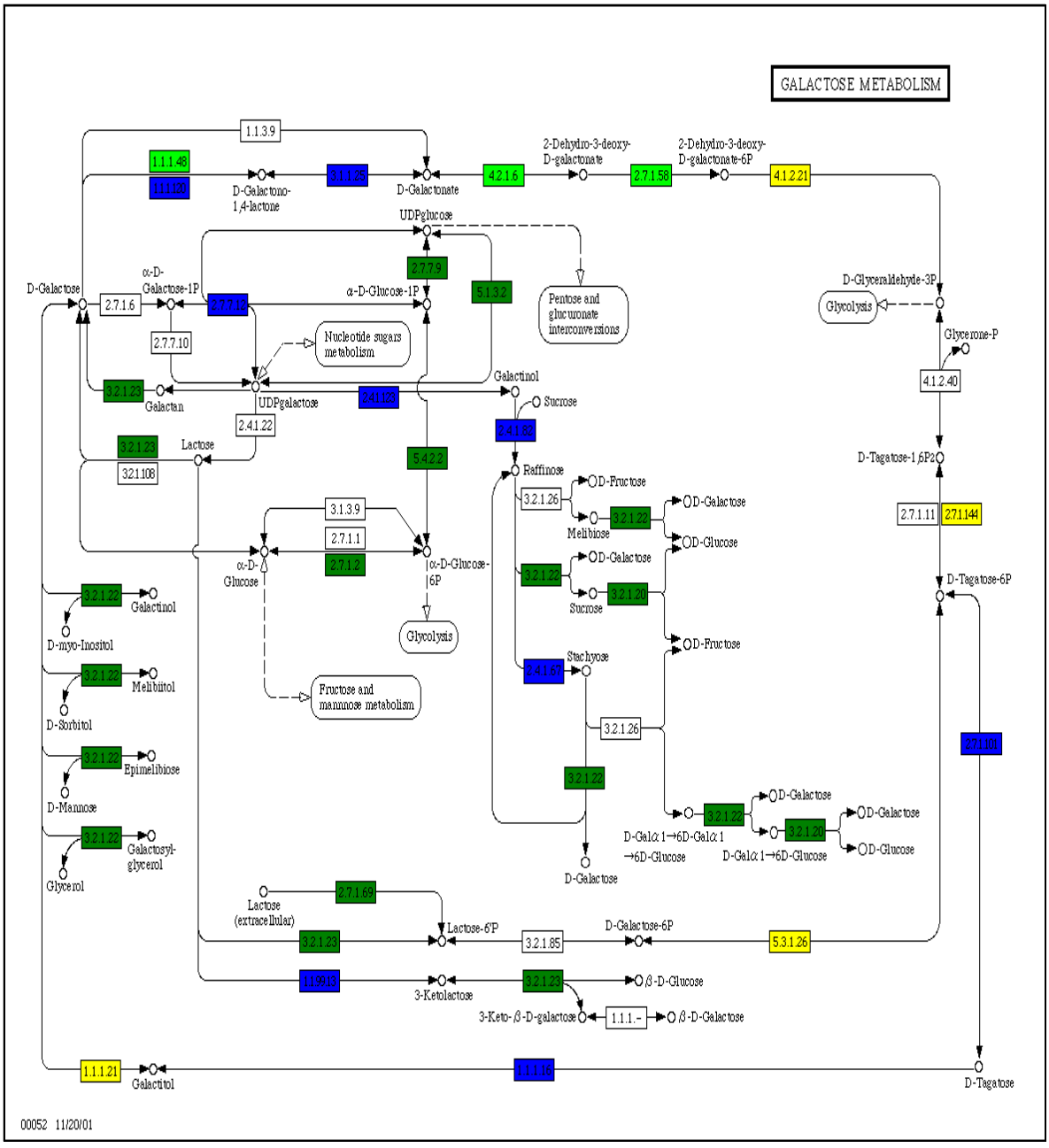


Figure 31: Voie du métabolisme du galactose.

Le galacticol ou dulcitol pourrait être aussi utilisé chez *S. meliloti* mais l'aldéhyde réductase (EC 1.1.1.21) identifiée ne possède que 31% d'identité avec la séquence consensus du profil de cette enzyme.

Dans la voie des interconversions du pentose et du glucuronate, le pentose xylose serait métabolisé en D-ribulose-5P, composé de la voie des pentoses phosphate.

Le L-arabinose est connu pour être métabolisé dans toutes les espèces de rhizobium (132), il est en effet trouvé comme transformé en L-arabinonate dans la voie du métabolisme de l'ascorbate et de l'aldarate. Cependant les enzymes réalisant les réactions suivantes, formant du pyruvate et de l'acetaldéhyde ne sont pas trouvées par l'annotation car il n'y a pas de séquences correspondantes.

8.1.4 La plaque tournante du métabolisme intermédiaire, le cycle de Krebs

La deuxième phase de la respiration aérobie comprend 2 parties, le cycle des acides tricarboxyliques puis la phosphorylation oxydative.

Le cycle des acides tricarboxyliques (cycle TCA) ou cycle de Krebs (figure 32) couplé à la respiration permet de générer l'énergie indispensable pour la réduction de l'azote atmosphérique par le bactéroïde. Il est aussi utilisé pour produire les précurseurs pour la biosynthèse des acides aminés, des purines, des pyrimidines et des vitamines. Tous les gènes codant pour les enzymes du cycle ont été trouvés chez *S. meliloti*. Durant la symbiose, le bactéroïde reçoit une grande quantité d'acides dicarboxyliques produits par la plante, tel que le succinate et la malate, métabolisés dans le cycle de Krebs. Les enzymes du cycle de Krebs sont nécessaires pour l'établissement d'une symbiose efficace.

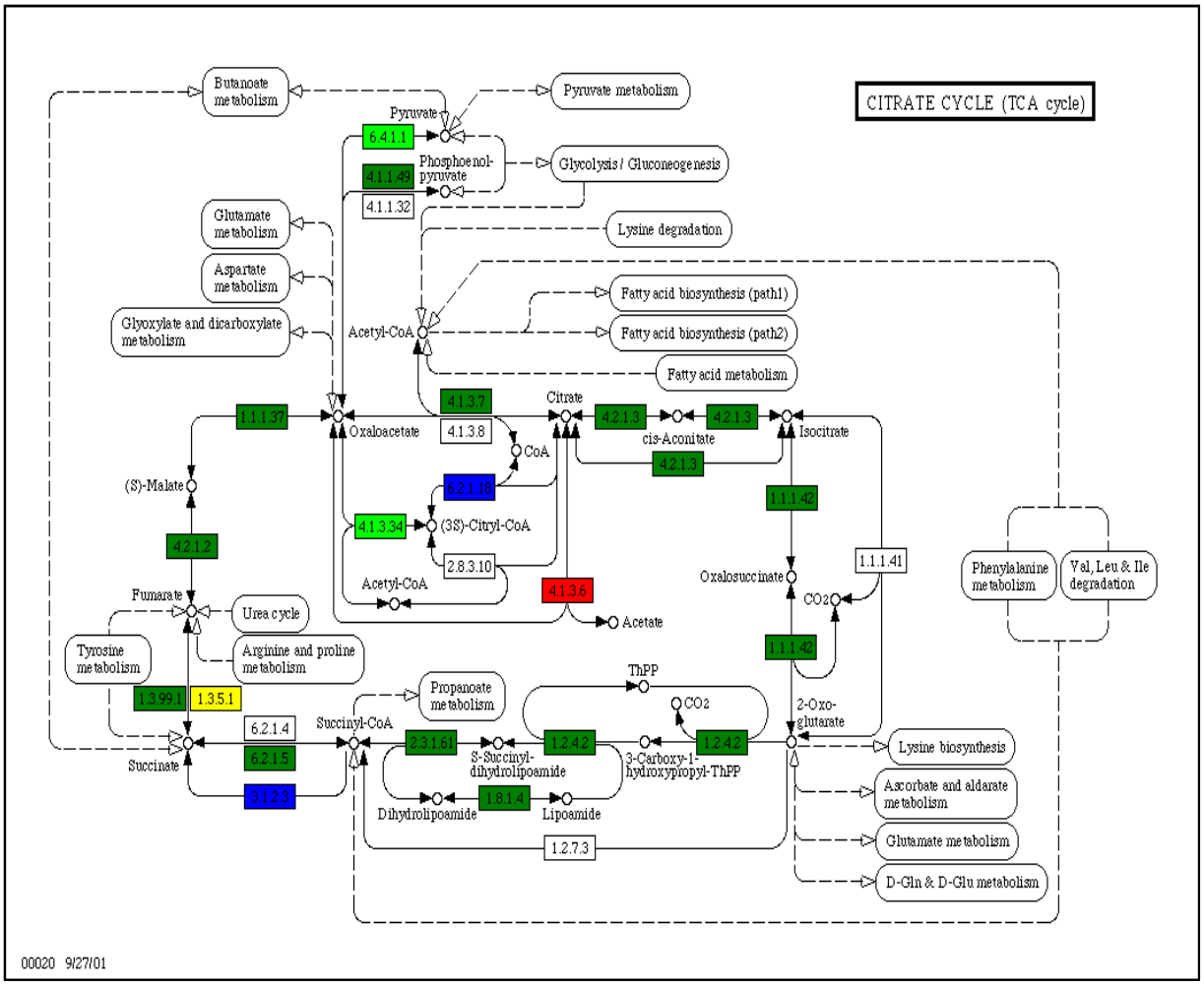


Figure 32: Le cycle des acides tricarboxyliques.

Le premier substrat du cycle de Krebs n'est pas le pyruvate mais l'acétyl CoA. L'étape précédant le cycle consiste donc à transformer le pyruvate en acétyl CoA (réaction visualisée dans le graphe de la glycolyse figure 27).

Ceci est réalisé par l'intermédiaire de la pyruvate déshydrogénase (EC 1.2.4.1), contenant trois composants enzymatiques : la pyruvate déshydrogénase (E1), la dihydrolipoamide acetyltransferase (E2) et la lipoamide déshydrogénase (E3). Chez *S. meliloti*, on trouve 2 sous-unités pour la pyruvate déshydrogénase (E1), une protéine pour la dihydrolipoamide acetyltransferase (E2, EC 2.3.1.12) et 3 exemplaires pour la lipoamide déshydrogénase (E3, EC 1.8.1.4).

L'acétyl CoA est aussi synthétisé à partir de l'acétate par une acétate kinase (EC 2.7.2.1 dans le graphe du métabolisme du pyruvate, figure 29) puis par une phosphate acetyltransférase (EC 2.3.1.8) ou par une acétyl CoA synthétase (EC 6.2.1.1 dans le graphe de la glycolyse figure 27).

Enfin, une dernière voie permet de synthétiser l'acétyl CoA, visualisée dans le graphe du métabolisme du butanoate. L'acétyl CoA est produit à partir de poly-beta-hydroxy-butyrate, un polyester synthétisé par les bactéroïdes. Le polyester est transformé par une acétoacétyl-CoA réductase (EC 1.1.1.36) en acétoacétyl-CoA puis en acétyl-CoA par une acétyl-CoA acétyltransférase (EC 2.3.1.9).

Toutes les enzymes du cycle de Krebs sont codées par des gènes situés sur le chromosome.

On remarque sur le graphe du cycle TCA que la protéine Smc03793 avec l'activité citrate lyase (EC 4.1.3.6 en rouge) trouvée avec l'environnement d'annotation iANT, n'est pas détectée par la méthode PRIAM. En fait, l'activité citrate lyase (EC 4.1.3.6) est composée de 2 sous-unités pour lesquelles la nomenclature EC a créé 2 autres numéros EC. La chaîne alpha porte l'activité Citrate CoA-transferase (EC 2.8.3.10) et la chaîne bêta l'activité Citryl-CoA lyase (EC 4.1.3.34). PRIAM possède un profil pour chaque chaîne mais en 2 exemplaires du fait de la redondance dans la description de ces activités. La chaîne alpha est représentée par les profils 1p2.8.3.10 et 2p4.1.3.6. La chaîne bêta est représentée par les profils 1p4.1.3.34 et 1p4.1.3.6. Seule la sous-unité bêta est trouvée chez *S. meliloti* avec PRIAM et iANT. Par ailleurs, l'EC 4.1.3.6 suit une règle « AND » avec les 2 profils : il faut obtenir un match significatif avec les 2 profils. C'est pourquoi, seul l'EC 4.1.3.34 de la sous-unité bêta est identifié par PRIAM. La citrate lyase n'est pas un élément du cycle TCA, elle produit de l'acétate qui est aussi produit par la voie du métabolisme du pyruvate (figure 29). L'étape suivante de la respiration aérobie, la phosphorylation oxydative est décrite dans la partie sur le métabolisme énergétique.

8.1.5 Les réactions anaplérotiques du cycle de Krebs

Comme chez d'autres organismes, *S. meliloti* possède des voies permettant de court-circuiter certaines réactions du cycle des acides tricarboxyliques et de régénérer les composés carbonés du cycle. Ces voies sont appelées anaplérotiques.

8.1.5.1 Le métabolisme des composés à 4 carbones

Il existe 3 types d'enzymes anaplérotiques permettant de régénérer les acides dicarboxyliques du cycle TCA.

1/ Les enzymes maliques : elles permettent la synthèse de malate à partir du pyruvate. Elles sont présentées dans l'analyse de la gluconéogénèse.

2/ L'aspartase (EC 4.3.1.1) : c'est une enzyme qui permet la production de fumarate à partir d'aspartate dans la voie du métabolisme de l'alanine et de l'aspartate (figure 33). Aucun candidat n'est trouvé pour cette enzyme chez *S. meliloti*.

3/ l'aspartate amino-transferase (EC 2.6.1.1) : elle permet la synthèse de glutamate et d'oxaloacetate à partir d'aspartate. Deux gènes candidats ont été localisés sur le chromosome par les deux méthodes iANT et PRIAM. En effet, ces gènes avaient été précédemment identifiés et caractérisés chez *S. meliloti* (135), et *S. meliloti* est capable de transporter l'aspartate (136).

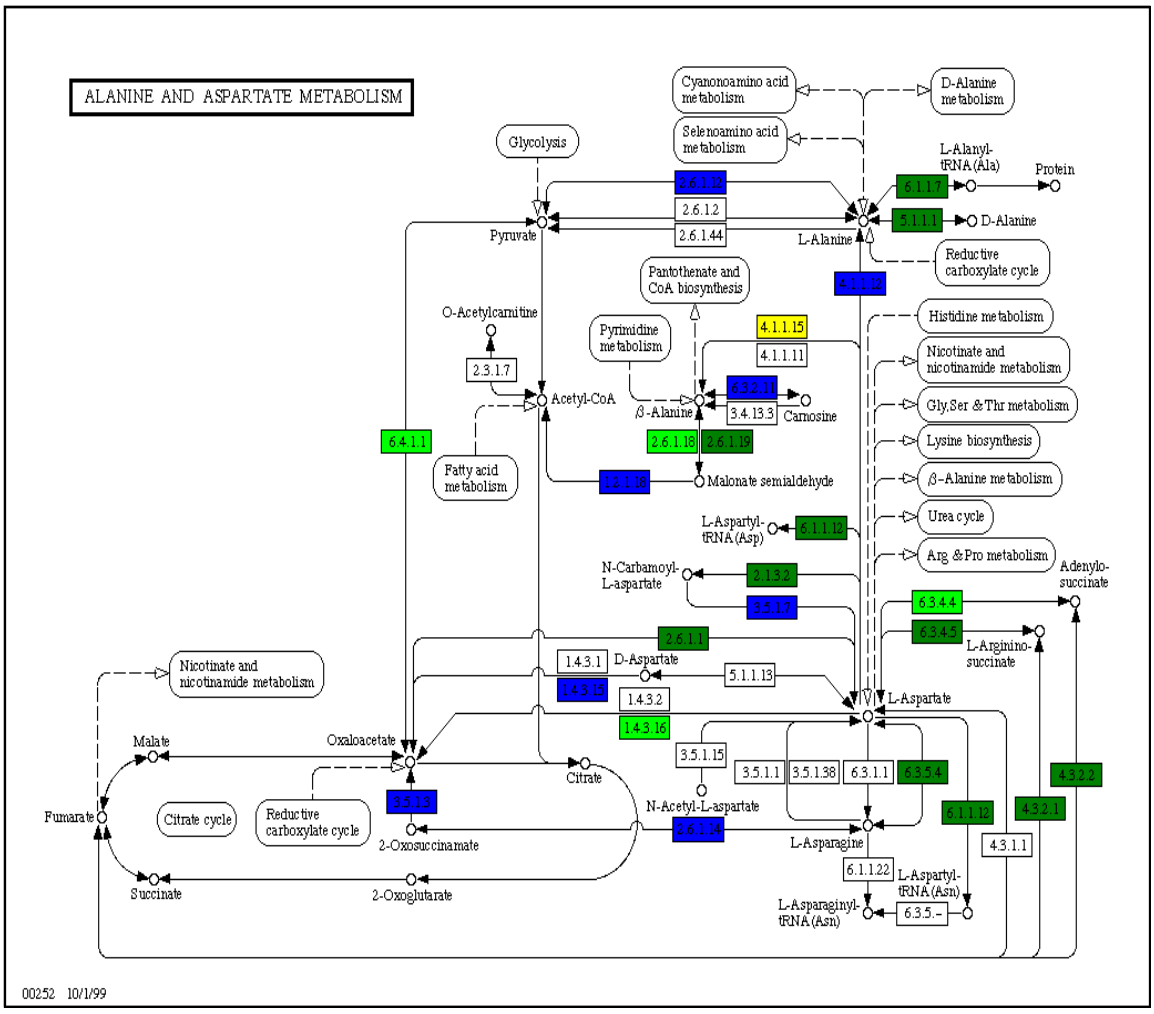


Figure 33 : Métabolisme de l'alanine et de l'aspartate.

8.1.5.2 Le métabolisme du 2-oxoglutarate par le court-circuit du γ-aminobutyrate

Chez *E. coli* ce court-circuit est utilisé pour métaboliser l'oxoglutarate endogène pendant la croissance anaérobie quand l'activité de la 2-oxoglutarate déshydrogénase (EC 1.2.4.2) est réprimée. Chez *S. meliloti* cette voie est identifiée avec la méthode PRIAM dans le graphe du métabolisme du glutamate (figure 34). En effet, des expériences de génétiques basées sur l'utilisation de mutants de la 2-oxoglutarate déshydrogénase chez *S. meliloti* (137) prouvent l'existence de ce court-circuit.

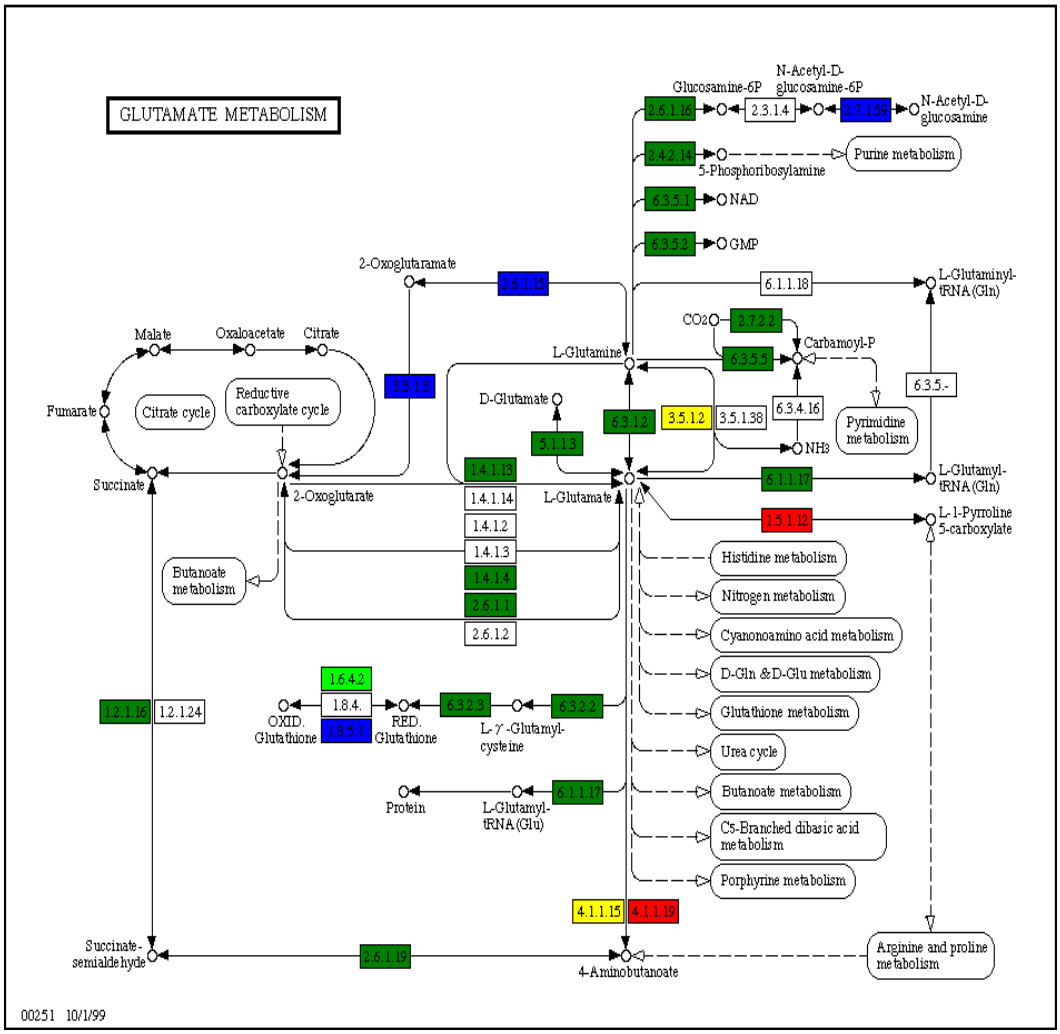


Figure 34 : Métabolisme du glutamate.

Le glutamate précurseur du γ-aminobutyrate, est synthétisé à partir du 2-oxoglutarate du cycle TCA par 3 types d'enzymes : la glutamate synthase (EC 1.4.1.13), la glutamate déshydrogénase (EC 1.4.1.4) et l'aspartate aminotransférase (EC 2.6.1.1) citée dans la voie anaplérotique de l'aspartate. Pour toutes ces enzymes, un ou plusieurs gènes candidats ont été trouvés par l'annotation iANT. Ces enzymes sont aussi trouvées avec la méthode PRIAM.

La 1ère étape du court-circuit dans la voie du métabolisme du glutamate est la décarboxylation du glutamate en γ-aminobutyrate par la glutamate décarboxylase (EC 4.1.1.15). Un candidat est proposé par la méthode PRIAM sur le mégaplasmide pSyma (SMa2402), avec une faible similarité pour la séquence consensus de l'un des 2 profils qui

caractérisent la collection 4.1.1.15. Or, une forte activité de cette enzyme a été trouvée dans le bactéroïde (138).

La deuxième étape de ce court-circuit est la réaction réalisée par la γ-aminobutyrate aminotransferase (EC 2.6.1.19) trouvée sur le mégaplasmide pSymb.

La dernière étape est l'oxydation du succinate semi-aldéhyde en succinate par la succinate semi-aldéhyde dehydrogénase (EC 1.2.2.16). Il y a quatre candidats proposés par l'annotation iANT et 6 bons candidats identifiés par la méthode PRIAM sur le chromosome et les deux mégaplasmides. Cette enzyme semble être très importante chez *S. meliloti* et suggère que le court-circuit du γ-aminobutyrate est utilisé dans cette bactérie.

8.1.5.3 Le métabolisme des composés à 3 carbones

Les composés à 3 carbones tel que le pyruvate et le propanoate sont des substrats de voies anaplérotiques du cycle tricarboxylique. Cependant les enzymes métabolisant le pyruvate tel que la pyruvate carboxylase, la pyruvate orthophosphate dikinase et la phosphoénolpyruvate carboxykinase ont aussi un rôle dans la gluconéogénèse (voir le paragraphe 3.1.2).

Des études sur le bactéroïde et sur la bactérie libre dans le sol (139), montrent que le propanoate peut être utilisé et transformé en succinyl CoA, intermédiaire du cycle TCA. En effet, toutes les enzymes de cette voie sont identifiées par iANT et PRIAM sur le chromosome et le mégaplasmide pSymb. L'enzyme clé de cette voie est la propionyl CoA carboxylase (EC 6.4.1.3) qui transforme le propionyl CoA en methyl-malonyl CoA. Le propionyl CoA est aussi un produit de la dégradation de certains acides aminés tels que la valine, la leucine et l'isoleucine. La propionyl CoA carboxylase est une enzyme composée de deux chaînes. La méthode PRIAM ne permet que d'identifier la chaîne beta car c'est l'unique chaîne observée chez tous les organismes présents dans la collection enzymatique.

8.1.5.4 Le métabolisme de l'acetyl CoA par la voie du glyoxylate

Le cycle tricarboxylique peut être court-circuité au niveau de l'isocitrate par l'isocitrate lyase (EC 4.1.3.1), catalysant la formation du glyoxylate et du succinate, puis par la malate synthase (EC 4.1.3.2) formant du malate (figure 35). Les gènes codant pour ces deux enzymes ont été localisés sur le chromosome. Par ailleurs, le glyoxylate est un produit des voies de dégradation des purines.

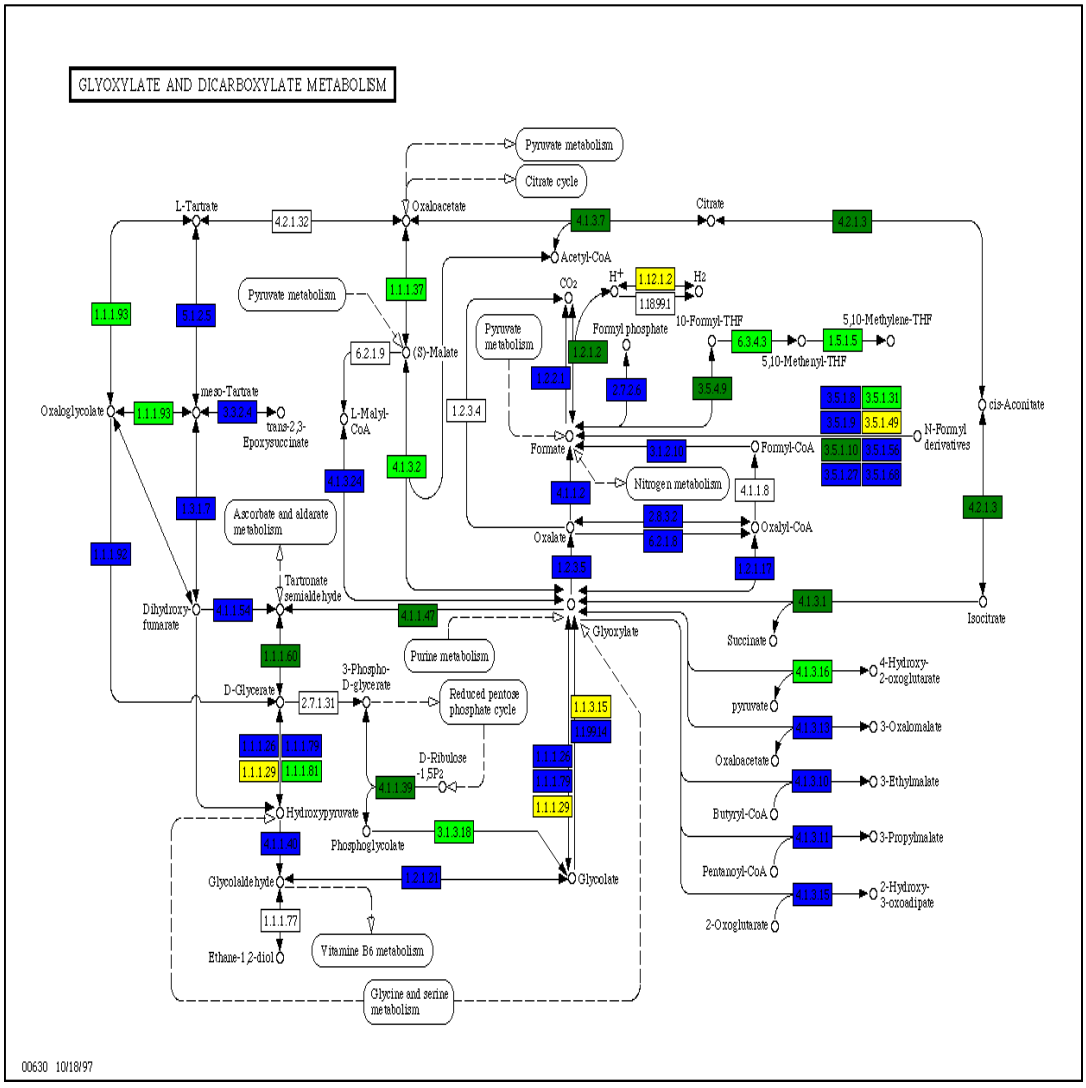


Figure 35 : Métabolisme du glyoxylate.

Le cycle de Krebs de *S. meliloti* est une voie complète avec de nombreux courts-circuits permettant de le régénérer. En effet, ce cycle est une voie métabolique très importante pour la bactérie lors de la symbiose puisqu'elle reçoit de nombreux composés dicarboxyliques de la plante. D'autres organismes comme *Helicobacter pylori* présentent une voie non cyclique (140). Chez cette bactérie pathogène le cycle TCA est moins important car la bactérie est capable par ailleurs de générer de l'énergie par une respiration anaérobie en utilisant le fumarate comme accepteur.

La plupart des enzymes du métabolisme des carbohydrates se trouvent sur le chromosome de *S. meliloti* confirmant son rôle principal pour la survie de l'organisme

(tableau 4). Cependant, on trouve sur le mégaplasmide pSymb quelques enzymes de la voie des pentoses phosphate, voie qui semble plus utilisée que la glycolyse et quelques enzymes impliquées dans des voies anaplérotiques, très utiles pour régénérer le cycle de Krebs.

Voie métabolique	complète	candidat PRIAM	chromosome	pSyma	pSymb
Synthèse du glutamate	X		X		
Synthèse de l'aspartate	X		X		
Synthèse de la thréonine	X		X		
Synthèse de la cystéine	X		X		
Synthèse de la lysine	0	X	X		
Synthèse de l'arginine	X		X		
Synthèse de la tyrosine	X		X		
Synthèse de la phénylalanine	X		X		
Synthèse du tryptophane	X		X		
Synthèse de la glutamine	X		X		
Synthèse de la méthionine	X		X		
Synthèse de l'histidine	0		X		
Synthèse de la sérine	X		X	X	X
Synthèse de la glycine	X		X	X	X
Synthèse de la valine	X		X		X
Synthèse de la leucine	X		X		X
Synthèse de l'isoleucine	X		X		X
Synthèse de la proline	X		X		X
Synthèse de la D-alanine	X		X	X	
Synthèse de l'asparagine	X				X

Tableau 4 : Localisation sur les réplicons de S. meliloti des voies de synthèse des acides aminés.

En jaune : localisation majoritaire, en rouge : voie incomplète

En bleu : voie complète grâce au candidat PRIAM

8.2 Métabolisme énergétique

S. meliloti est une bactérie aérobie stricte utilisant la chaîne respiratoire pour générer son énergie. Les enzymes caractéristiques des voies de fermentation ne sont pas présentes dans son génome. Ainsi, la pyruvate décarboxylase (EC 4.1.1.1), en cause dans la fermentation alcoolique est absente, ainsi que la L-lactate déshydrogénase (EC 1.1.1.27). En effet le gène candidat mdh proposé par la méthode PRIAM pour la L-lactate déshydrogénase est une malate déshydrogénase (EC 1.1.1.37). Le cas de ces profils très similaires est présenté dans la discussion du chapitre II, Inférence fonctionnelle du métabolisme : PRIAM.

8.2.1 La chaîne respiratoire

La 2ème étape de la respiration aérobie, la phosphorylation oxydative est une cascade d'oxydoréductions permettant de produire des molécule d'ATP à partir des molécules de NADH et FADH2. Cette chaîne respiratoire (figure 36) est normalement réalisée par cinq complexes protéiques : la NADH déshydrogénase (EC 1.6.5.3), la succinate déshydrogénase (EC 1.3.99.1), l'ubiquinol-cytochrome c réductase (EC 1.10.2.2), la cytochrome c oxydase (EC 1.9.3.1) et l'ATP synthase (EC 3.6.1.34 dans le graphe, transféré vers l'EC 3.6.3.14).

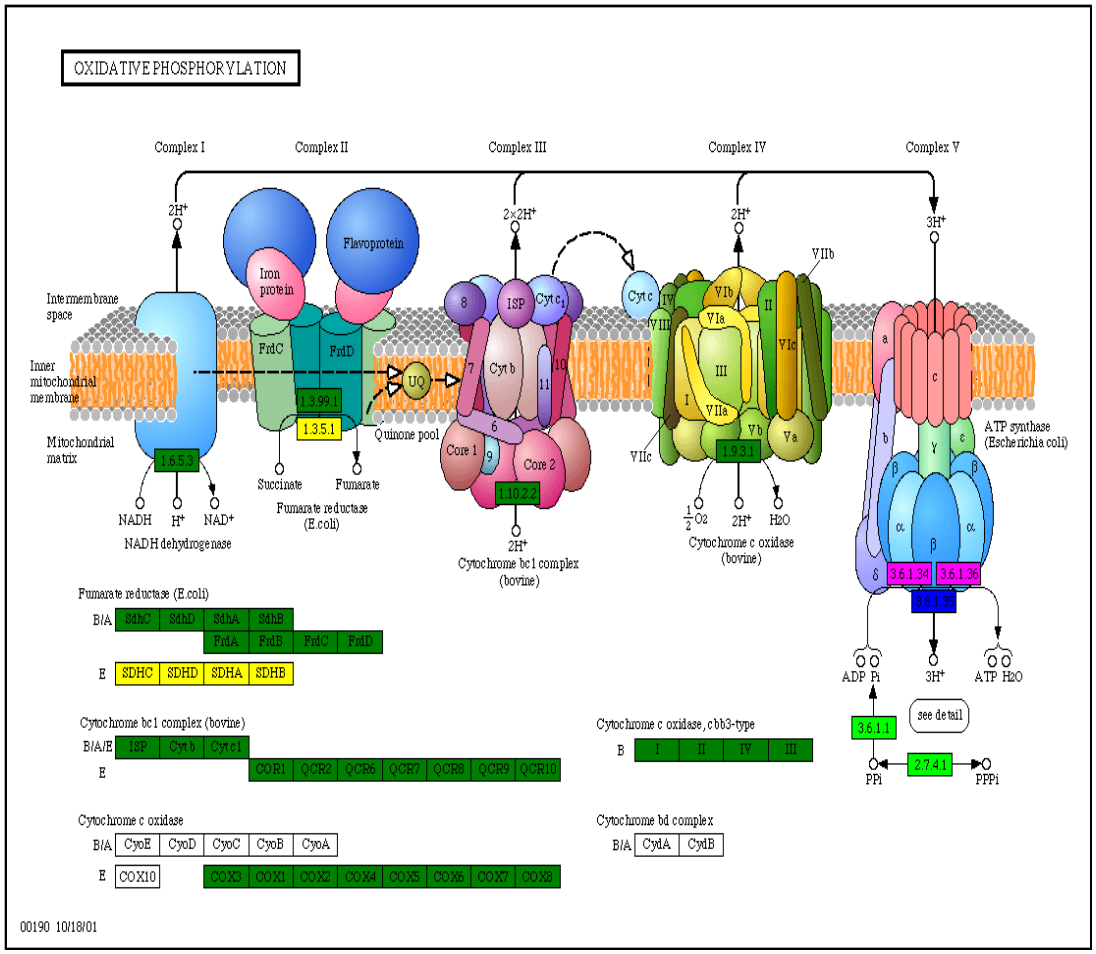


Figure 36 : La chaîne respiratoire.

Ces complexes sont connus chez *S. meliloti* et ont été trouvés par l'annotation avec iANT (figure 36) alors que certains gènes n'ont pas été identifiés par la méthode PRIAM. Ainsi, deux sous-unités (*nuo*H et *nuo*I) du complexe de la NADH dehydrogenase ne sont pas trouvés avec les profils de la collection 1.6.5.3. Par ailleurs, 2 gènes du complexe III n'ont pas été identifiés et ne sont pas visualisés sur le graphe car ce sont des sous-unités qui n'ont pas d'activité enzymatique et pour lesquelles aucun numéro EC n'a été attribué. On retrouve le même phénomène pour plusieurs gènes du complexe IV. Cependant avec l'utilisation de la méthode PRIAM, on met en évidence des sous-unités appartenant à différents complexes oxydases. En effet, *S. meliloti* possède un système de chaînes respiratoires branchées avec des oxydases finales ayant des affinités différentes pour l'O_2 (141). Ainsi, il existe deux groupes de gènes (*cta*CDBGE et *cox*MNOP) connus pour avoir une action en condition aérobie. Ils

s'apparentent au complexe cytochrome *aa3*. Alors qu'en condition de micro-oxie, condition existant dans le nodule lors de la symbiose et nécessaire pour éviter l'inactivation de la nitrogénase (EC 1.18.6.1) (142), le complexe IV est un cytochrome *cbb3* codé par le groupe de gènes *fix*NOPQ. Enfin, le complexe V, l'ATP synthase (EC 3.6.1.34) ne donne aucun résultat sur le graphe de la chaîne respiratoire car dans la version de la base de données ENZYME utilisée, cet EC a été transféré vers l'EC 3.6.3.14. L'ATP synthase possède 10 sous-unités qui sont identifiées par l'annotation avec iANT. Trois des ces sous-unités ne sont pas identifiées en utilisant la méthode PRIAM.

8.2.2 Métabolisme azoté

S. meliloti est une bactérie capable d'assimiler l'azote réduit de l'environnement ainsi que l'azote sous forme de nitrate ou nitrite. Lors de la symbiose, elle transforme ces composés en ammoniac qui est alors assimilable par la plante.

L'assimilation du nitrate (figure 37) peut se faire par deux types de nitrates réductases :

- la nitrate réductase (NADH) (EC 1.6.6.1 transféré vers l'EC 1.7.1.1) dite assimilatrice,
- la nitrate réductase respiratoire (EC 1.7.99.4).

Un candidat est proposé par PRIAM pour la fonction nitrate réductase assimilatrice, la protéine SMb20584 qui présente 24% d'identité avec la séquence consensus du profil de la collection enzymatique 1.6.6.1. C'est une similarité assez faible, d'autant plus que la protéine est plus courte que le profil caractérisant cette fonction.

Par contre, la nitrate réductase respiratoire présente un candidat trouvé sur le mégaplasmide pSymb avec les deux méthodes iANT et PRIAM. Un autre candidat (SMa1236) est proposé par la méthode PRIAM sur le pSyma avec 66% d'identité pour la séquence consensus du profil. Ce dernier candidat semble être très probable. Cette analyse informatique met en évidence un oubli d'EC lors de l'annotation. Cette activité a été trouvée expérimentalement dans la bactérie (143).

L'assimilation du nitrite en ammoniac se fait par la voie de la nitrite réductase (NADPH) (EC 1.6.6.4) qui permet de transformer le nitrite en ammoniac en une seule réaction. Les deux sous-unités de l'enzyme sont bien identifiées sur le pSymb.

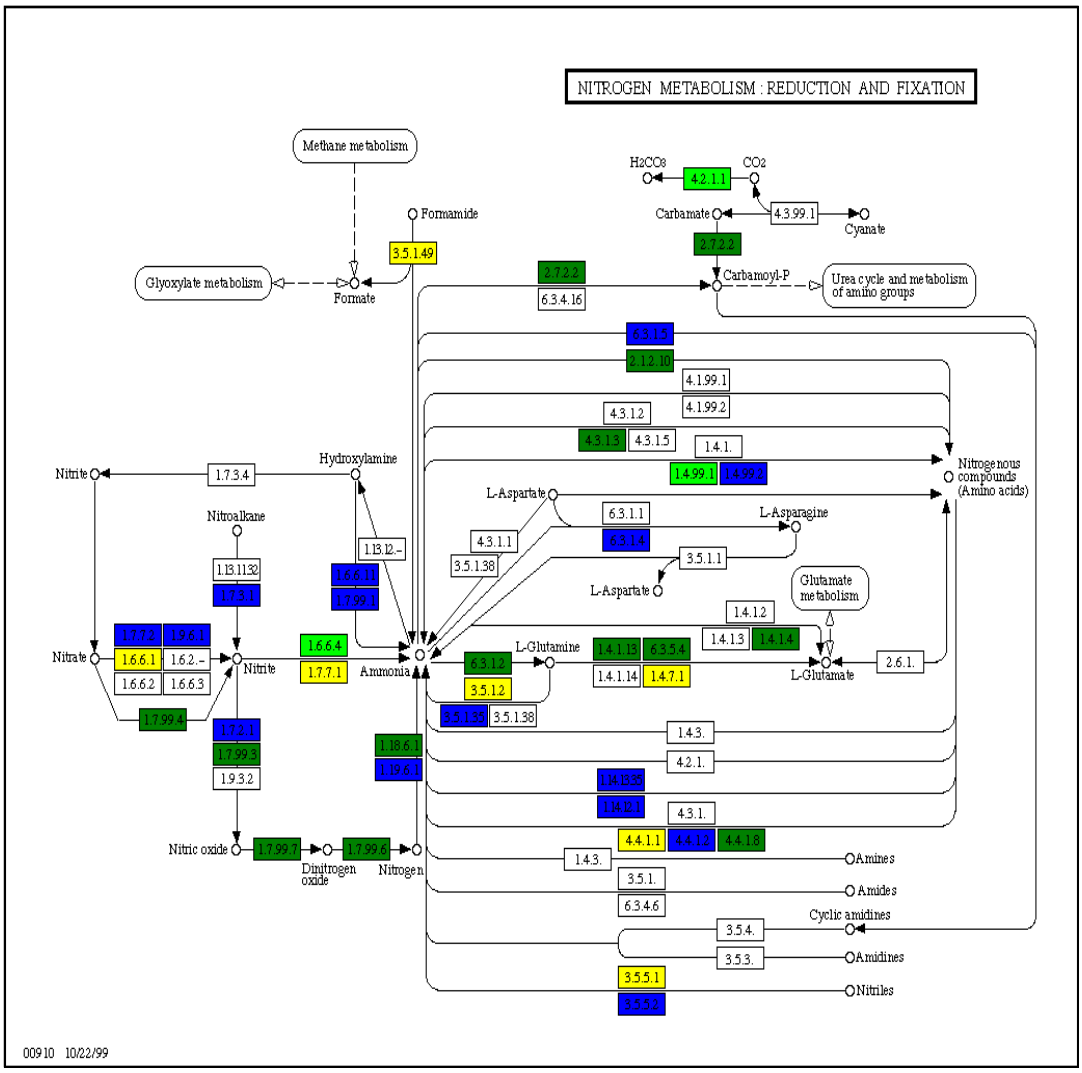


Figure 37 : Réduction et fixation de l'azote.

Une voie de dénitrification fait intervenir plusieurs enzymes : la nitrite réductase (EC 1.7.99.3), la NO réductase (EC 1.7.99.7), la N_2O réductase (EC 1.7.99.6).

Par ailleurs, le complexe enzymatique de la nitrogénase (EC 1.18.6.1), est l'enzyme capable d'assimiler l'azote atmosphérique. Toutes ces enzymes ont été identifiées par les deux méthodes d'annotation sur le pSyma.

L'ammoniac obtenu peut être assimilé chez la bactérie en glutamate par deux voies possibles. L'utilisation de l'une ou l'autre de ces voies se fait selon la concentration en $NH4^+$ dans la cellule. Lorsque la concentration en $NH4^+$ est élevée, l'ammoniac est assimilé en glutamate par la glutamate déshydrogénase (GDH) (EC 1.4.1.4) dont le gène est localisé sur

le pSyma. Quand la concentration en $NH4^+$ est faible, l'ammoniac est assimilé en glutamate par la glutamine synthétase (GS) (EC 6.3.1.2) puis par la glutamate synthase (NADPH) (nommée GOGAT) (EC 1.4.1.13). Il existe trois formes de la glutamine synthétase (GS) : GS I et III dont les gènes sont présents sur le chromosome et GS II dont le gène est sur le pSymb. Par ailleurs, 2 autres séquences sont proposées sur le chromosome par les méthodes iANT et PRIAM, comme étant des GS putatives avec 33 et 24% d'identité pour le profil. Deux autres séquences codées par des gènes présents sur le chromosome ont été identifiées par la méthode PRIAM avec 32% d'identité. L'enzyme glutaminase (EC 3.5.1.2 en jaune) capable de réaliser la réaction inverse a été trouvée avec 54% d'identité pour la séquence consensus du profil de la collection sur le chromosome.

L'ammoniac peut aussi être assimilé chez la bactérie en acides aminés par plusieurs enzymes tel que l'histidine ammonia-lyase (EC 4.3.1.3) localisée sur les trois réplicons, l'aminométhyltransferase (EC 2.1.2.10) avec un bon candidat sur le chromosome et enfin, en carbamoyl phosphate par une carbamate kinase (EC 2.7.2.2) dont le gène est localisé sur le pSymb. Le carbamoyl phosphate peut être incorporé dans le cycle de l'urée, intermédiaire du métabolisme de nombreux acides aminés.

En conclusion, toutes les enzymes de la voie de l'assimilation de l'azote en ammoniac, voie caractéristique de la symbiose de *S. meliloti*, se trouvent sur les mégaplasmides pSyma et pSymb. Cette observation, met en évidence le rôle important de ces mégaplasmides lors de la symbiose par rapport au chromosome qui porte l'ensemble des gènes de ménage de la bactérie.

8.2.3 Métabolisme des composés en C1

Le métabolisme des composés en C_1 n'a pas été étudié chez *S. meliloti*. Cependant, on trouve dans son génome des gènes avec des homologies significatives pour des enzymes impliquées dans le métabolisme du méthanol. Le méthanol peut être transformé en formaldéhyde par trois types d'enzymes (figure 38). Une catalase (EC 1.11.1.6) qui présente trois bons candidats sur les trois réplicons, une méthanol dehydrogenase (EC 1.1.99.8) et une peroxydase (EC 1.11.1.7) avec chacune un candidat sur le pSymb.

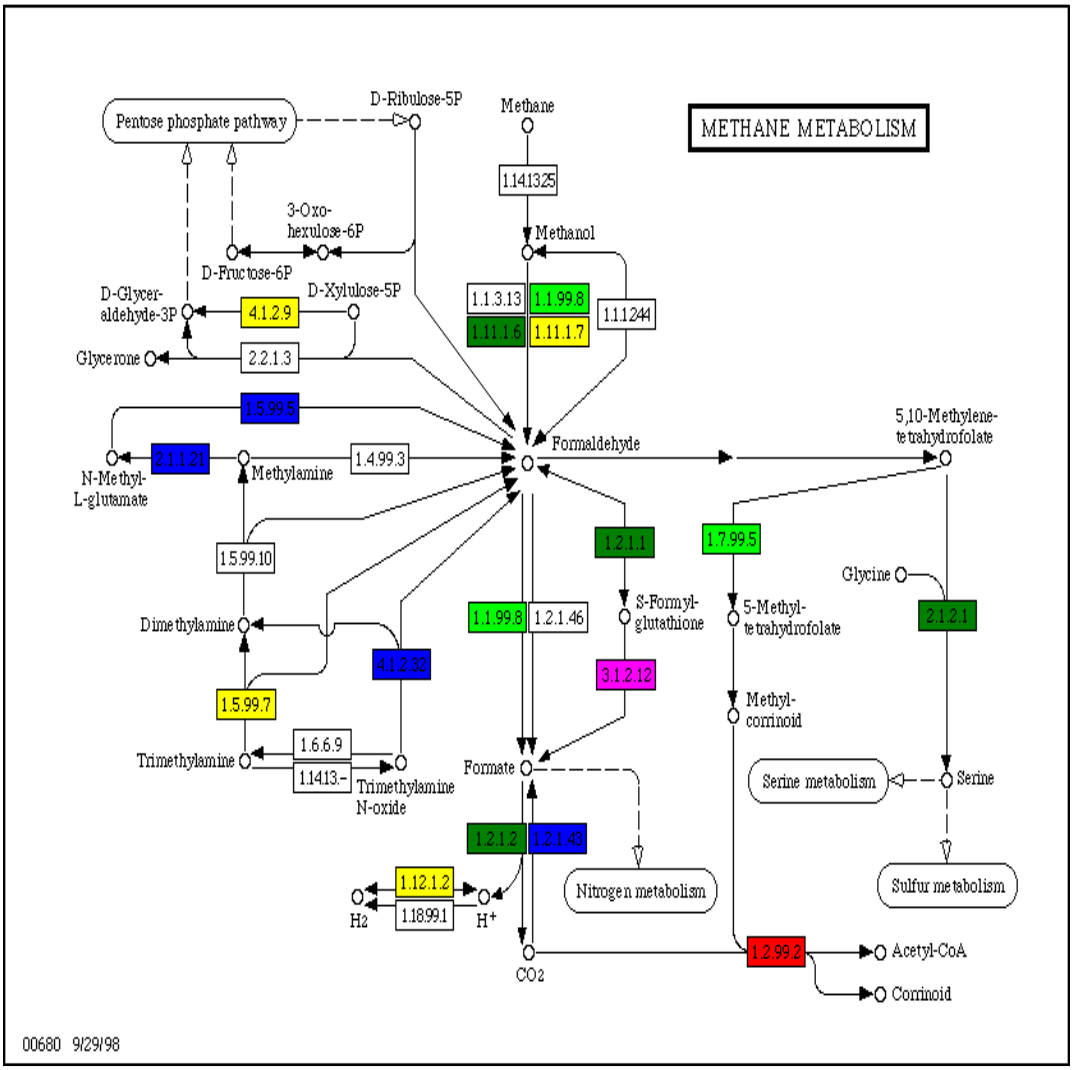


Figure 38: Métabolisme des composés en C1.

Le formaldéhyde est un composé intermédiaire central du métabolisme des bactéries méthylotrophes. C'est un composé toxique pour les bactéries qui doivent l'oxyder en CO_2. Chez les bactéries méthylotrophes, quatre systèmes sont connus comme voies d'oxydation du formaldéhyde (144). Trois de ces systèmes semblent être présents chez *S. meliloti*.

La voie connue chez plusieurs α–protéobactéries autotrophes, passant par la formaldéhyde déshydrogénase (glutathion) (EC 1.2.1.1), la S-formylglutathion hydrolase (EC 3.1.2.12) et la formate déshydrogénase (EC 1.2.1.2) a été trouvée chez *S. meliloti*. La S-formylglutathion hydrolase (colorié en violet dans les graphes de KEGG) n'a pas été identifiée avec PRIAM car aucune information de séquence n'était disponible dans la version

27.0 d'ENZYME. Les sous-unités du complexe enzymatique formate déshydrogénase ont été trouvées, trois ont des gènes localisés sur pSyma et 4 sous-unités ont des gènes localisés sur le chromosome avec l'annotation iANT. Avec la méthode PRIAM, les gènes de deux sous-unités seulement ont été obtenus sur le pSyma.

La deuxième voie impliquant des enzymes dépendantes du tétrahydrofolate a été identifiée (EC 2.1.2.1 : glycine hydroxymethyltransferase). Elle permet la formation de sérine.

La troisième voie, allant vers la voie des pentoses phosphate a été identifiée par la méthode PRIAM avec la prédiction de la Xylulose-5-phosphate phosphoketolase (EC 4.1.2.9) (53% d'identité avec la séquence consensus du profil).

Enfin, une quatrième voie a été décrite chez *Methylobacterium extorquens* AM1 (145) (146) (147), impliquant trois enzymes : la méthényl H_4MPT cyclohydrolase (EC 3.5.4.27), la méthényl H_4F cyclohydrolase (EC 3.5.4.9) et la méthylène H_4MPT déshydrogénase NADP dépendante (EC 1.5.99.9). Seul le gène de la méthényl H_4F cyclohydrolase (EC 3.5.4.9) a été localisé sur le chromosome, codant pour une enzyme bi-fonctionnelle. Cette enzyme a un rôle anabolisant dans le métabolisme des composés en C1 chez *M. extorquens* AM1. Il semble que les enzymes H_4MPT dépendantes soient très spécifiques des bactéries méthylotrophes et des archeo- bactéries méthanogènes (146).

S. meliloti semble posséder les enzymes nécessaires au métabolisme du méthanol. Par ailleurs, il a été récemment découvert que la réaction spontanée transformant le formaldéhyde en S-hydroxymethylglutathion (réaction non visualisée sur la figure 38, avant l'EC 1.2.1.1) est en fait une réaction catalysée par une enzyme d'activation du formaldéhyde glutathion-dépendente (Gfa) (148). Cette enzyme découverte chez *Paracoccus denitrificans* présente une similarité de séquence (75% d'identité) avec une ORF de pSymb.

8.2.4 Métabolisme des composés soufrés

L'assimilation du soufre par l'intermédiaire des sulfates ou des sulfures est un métabolisme indispensable, pour la production des acides aminés soufrés tel que la cystéine, la méthionine, pour la synthèse des clusters Fer-Soufre et la formation des facteurs Nod, signaux moléculaires permettant la réponse symbiotique. Le métabolisme soufré comprend la formation d'une molécule, la 3'-phosphoadenosine 5'-phosphosulfate (PAPS) (figure 39) intermédiaire de la synthèse de la cystéine et précurseur des oligosaccharides N-acétylés, les facteurs Nod. La formation des PAPS est réalisée par les gènes *nod*PQ (149) codant une

enzyme bifonctionnelle avec les activités sulfate adénylyltransférase (EC 2.7.7.4) et adénylylsulfate kinase (EC 2.7.1.25) se trouvant sur les mégaplasmides pSyma et pSymb. Il existe aussi des gènes codant des protéines avec une activité sulfate adénylyltransférase (EC 2.7.7.4) trouvés sur le chromosome. Par ailleurs, il existe un ensemble d'enzymes pour la synthèse de cystéine à partir de la L- sérine : la sérine O-acétyltransférase (EC 2.3.1.30) et la cystéine synthase (EC 4.2.99.8) dont les gènes ont été localisés sur le chromosome. Cependant, les enzymes pour passer du sulfite à l'H_2S permettant de faire le lien entre PAPS et la formation de cystéine n'ont pas été identifiées.

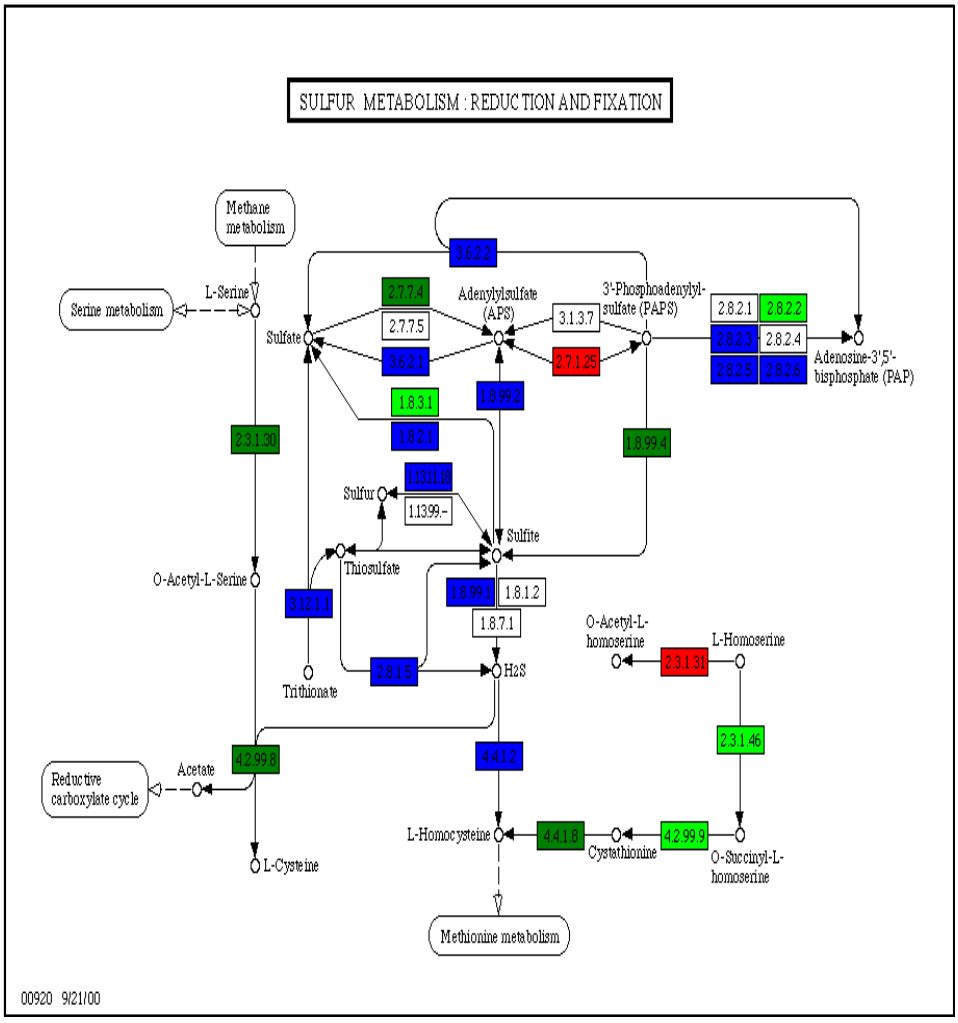


Figure 39 : Métabolisme des composés soufrés.

8.3 Métabolisme des acides aminés

La synthèse des acides aminés est importante pour la bactérie *S. meliloti* en symbiose. En effet, des mutants auxotrophes pour la leucine, l'asparagine, la méthionine, la tyrosine sont incapables de fixer l'azote (150).

Les voies de synthèse de 18 acides aminés sur 20 ont été complètement identifiées d'après l'analyse des résultats. La voie de synthèse de l'histidine est incomplète car il manque l'histidinol phosphatase (EC 3.1.3.15). D'autre part, la voie de synthèse de la lysine à partir de l'aspartate est incomplète car il manque la succinyldiaminopimelate aminotransférase (EC 2.6.1.17). Cette enzyme ne possédait pas de séquences annotées avec ce numéro EC dans la version d'ENZYME utilisée. A ce jour, il y a un ensemble de séquences ARGD annotées avec cette fonction ainsi qu'avec la fonction acétylornithine aminotransférase (EC 2.6.1.11). En effet, dans la collection enzymatique 2.6.1.11, il y a des séquences bifonctionnelles 2.6.1.11 / 2.6.1.17. Un seul profil caractérise cette collection et donc une seule région protéique permet de caractériser ces deux fonctions. Un bon candidat a été obtenu avec le profil de la collection 2.6.1.11, qui pourrait aussi être annoté succinyldiaminopimelate aminotransférase (EC 2.6.1.17) d'après les résultats obtenus pour la voie de synthèse de la lysine.

La plupart des voies de biosynthèse des acides aminés possèdent des enzymes dont les gènes sont essentiellement sur le chromosome (tableau 4). Seulement deux voies ne sont pas codées par le chromosome, le pSyma possède le gène pour une D-alanine aminotransférase (EC 2.6.1.21) permettant de synthétiser la D-alanine qui est un composé nécessaire à la synthèse de la paroi de la bactérie, à partir du pyruvate et le pSymb porte la voie de biosynthèse de l'asparagine.

8.4 Métabolisme des nucléotides

Les voies de biosynthèse des purines et pyrimidines sont essentielles pour la synthèse des acides nucléiques, de molécules énergétiques comme l'ATP, le GTP, de certains acides aminés comme le glutamate, la sérine et de cofacteurs comme la thiamine, la riboflavine, le folate. Les voies du métabolisme des purines et des pyrimidines sont complètes et sont présentent essentiellement sur le chromosome. Les nucléotides sont synthétisés à partir du PRPP (5-Phospho-alpha-D-ribose 1-diphosphate) produit de la voie des pentoses phosphate. Dans ces voies, plusieurs cas de numéros EC coloriés en orange dans les graphes de KEGG (EC 3.1.5.1, EC 5.1.3.12) permettent de rajouter un numéro EC oublié ou manqué par

l'annotation manuelle. Et quelques cas de numéros EC coloriés en rouge dans les graphes de KEGG (EC 2.1.2.3, EC 3.2.1.37), mettent en évidence des homologies avec des enzymes bifonctionnelles.

8.5 Métabolisme des cofacteurs et vitamines

Un cofacteur est un complément inorganique essentiel à la réalisation d'une réaction enzymatique. Chez *S. meliloti* l'ajout de cofacteurs comme la biotine, la thiamine, et la riboflavine permet une meilleure colonisation de la rhizosphère (151). Ils peuvent aussi avoir un rôle important lors de l'interaction symbiotique.

D'après les résultats de prédiction de fonction des gènes, deux cofacteurs et un groupement prosthétique ont des voies métaboliques complètes chez *S. meliloti*. A partir du GTP, les trois cofacteurs sont :

- le folate, cofacteur essentiel dans le métabolisme des composés en C_1.
- la riboflavine, précurseur du FAD.
- la molybdoptérine (dans la voie de synthèse du folate) groupement prosthétique soufré impliqué dans des réactions de réduction du nitrate.

Deux voies métaboliques de synthèse sont presque complètes. La synthèse du nicotinate, une des formes de la vitamine B3, est présente sur le chromosome de *S. meliloti* mais la nicotinate nucleotide adenylyltransferase (EC 2.7.7.18) n'a pas été trouvée avec iANT. Un candidat présent sur le chromosome est proposé avec une faible similitude globale pour la séquence consensus du profil de cette fonction enzymatique. De même la voie de synthèse du pantothénate à partir du Coenzyme A est complète, si on tient compte des candidats trouvés par PRIAM sur le chromosome (tableau 5), pour les activités acyl-carrier protein phosphodiesterase (EC 31.4.14) et dephospho-CoA kinase (EC 2.7.1.24). Par ailleurs, les voies de synthèse d'autres cofacteurs ou groupements prosthétiques comme le pyridoxal ou vitamine B6, l'ubiquinone, la porphyrine, l'hème et la thiamine, ne sont pas complètes. Enfin, la voie de synthèse de la biotine est absente même si deux candidats sont proposés avec de faibles homologies. Il est connu que la souche de *S. meliloti* séquencée est auxotrophe pour ce cofacteur (152).

Voie métabolique	complète	candidat PRIAM	chromosome	pSyma	pSymb
Métabolisme des purines	X		X		X
Métabolisme des pyrimidines	X		X	X	
Métabolisme du folate	X		X		
Métabolisme de la riboflavine	X		X		
Métabolisme du molybdoptérine	X		X		
Métabolisme du nicotinate	0	X	X		
Métabolisme du pantothenate	0	X	X		X
Métabolisme du pyridoxal	0		X		
Métabolisme de l'ubiquinone	0		X	X	
Métabolisme de la porphyrine	0		X		
Métabolisme de l'hème	0		X		
Métabolisme de la thiamine	0		X		X
Métabolisme de la biotine	0	X	X		X
Métabolisme des lipides	X		X		X
Métabolisme des corps cétones	X		X		X
Métabolisme des glycerolipides	X		X		X

Tableau 5 : Localisation sur les réplicons de S. meliloti des voies métaboliques.

En jaune : localisation majoritaire

En rouge : voie incomplète

En bleu: voie complète grâce au candidat PRIAM.

8.6 Métabolisme des lipides et lipides complexes

Les gènes impliqués dans la voie métabolique des lipides sont entièrement identifiés sur le chromosome et le mégaplasmide pSymb. De même, les voies de synthèse et de dégradation des corps cétoniques et des glycérolipides sont complètes et les gènes sont trouvés sur ces deux réplicons.

9 Discussion

L'analyse du métabolisme de *S. meliloti* a été réalisée à travers la comparaison d'un environnement semi-automatique avec une méthode automatique d'annotation. La comparaison de deux méthodes d'annotation a été possible grâce à l'existence de la nomenclature EC des fonctions enzymatiques. La nomenclature est essentielle pour une annotation homogène et pour réaliser rapidement une mise à jour de l'annotation au fur et à mesure de l'arrivée de données de séquences. Cette comparaison met en évidence les avantages et les défauts de ces deux types de méthodes.

La méthode automatique PRIAM permet de faire une recherche systématique de toutes les fonctions enzymatiques répertoriées avec un numéro EC. Ainsi très peu de numéros EC ne sont pas trouvés par rapport à la méthode semi-automatique iANT. Cependant, un grand nombre de fonctions enzymatiques de la nomenclature EC ne possèdent pas encore d'informations de séquences. Les fonctions enzymatiques oubliées s'expliquent par le fait que dans la méthode PRIAM une seule fonction est attribuée à chaque région protéique alors que certaines régions peuvent correspondre à des activités pour différents substrats. Par ailleurs, on observe des difficultés pour annoter automatiquement les gros complexes enzymatiques puisque il existe des sous-unités sans activité enzymatique et donc sans numéro EC attribué. De nombreuses fonctions enzymatiques sont proposées en plus par rapport à la méthode semi-automatique. Mais le plus souvent il est nécessaire d'avoir l'expertise d'un annotateur pour interpréter la similitude.

Enfin, cette analyse des prédictions de fonctions enzymatiques dans le contexte des voies métaboliques permet de mettre en évidence des enzymes manquantes dans certaines voies comme l'histidinol phosphatase (EC 3.1.3.15) dans la voie de biosynthèse de l'histidine. Deux raisons peuvent expliquer ces absences : soit ces fonctions ne sont pas prédictibles par analyse de séquence à ce jour, soit elles sont réellement absentes dans l'organisme.

L'environnement d'annotation semi-automatique iANT a l'avantage de pouvoir prédire des fonctions enzymatiques qui ne sont pas encore intégrées dans la nomenclature EC. D'autre part, l'expertise des annotateurs permet d'attribuer des fonctions pour des faibles similarités sur la base de connaissances biologiques ou bibliographiques de l'organisme. Cependant, en réalisant cette comparaison de méthodes, on observe que l'annotateur peut se

tromper dans le choix d'un numéro EC, voir même omettre son attribution. Enfin, la durée pour annoter des génomes complets par ce type de méthode est un facteur limitant.

Il est en général difficile de prédire les enzymes avec différentes spécificités de substrats. La spécificité de substrats est une caractéristique biologique qui est mise en évidence expérimentalement. Certaines enzymes prédites par analyse de séquence ne possèdent pas toujours une activité bidirectionnelle. C'est le cas de la malate déshydrogénase de *S. meliloti*.

Par ailleurs, il existe parfois plusieurs voies parallèles permettant de transformer un substrat en un produit comme par exemple la transformation du pyruvate en phosphoénolpyruvate. Seule l'expérimentation biologique permet d'identifier les voies les plus utilisées. Enfin, les expériences permettent de caractériser des séquences atypiques, non encore répertoriées dans les banques de séquences.

L'analyse des prédictions des fonctions enzymatiques chez *S. meliloti* nous permet d'avoir une vue globale de son métabolisme intermédiaire, avec l'identification complète de certaines voies comme le cycle de Krebs, la voie des pentoses phosphate plutôt que la glycolyse. Mais aussi, la mise en évidence de voies plus spécifiques telle que l'assimilation de l'azote atmosphérique par la nitrogénase.

L'observation des localisations des gènes sur les différents réplicons met en évidence le rôle majeur du chromosome pour le métabolisme de la bactérie. Cependant, les deux mégaplasmides portent aussi quelques gènes codant pour des enzymes. Ainsi, le réplicon pSyma code pour un gène de la synthèse de la D-alanine et pour les gènes de la nitrogénase (Tableau 6) et la synthèse de l'asparagine est codée par des gènes sur le pSymb (Tableau 4).

Voie métabolique	**complète**	**candidat PRIAM**	**chromosome**	**pSyma**	**pSymb**
Glycolyse	X	X	X	0	X
Voie des pentoses phosphate	X		X	0	X
Voie d'Entner-Doudoroff	X		X	X	X
Gluconéogénèse	X		X	0	X

Cycle des acides tricarboxyliques	X		X	0	0
Voies anplérotiques	X		X	X	X
Chaîne respiratoire	X		X	X	X
Métabolisme azoté	X		X	X	X
Métabolisme des composés en C1	X		X	X	X
Métabolisme des composés soufrés	0		X	X	X

Tableau 6: Localisation des voies métaboliques sur les réplicons de S. meliloti.

En jaune: localisation majoritaire

En rouge: voie incomplète

IV
CONCLUSIONS ET PERSPECTIVES

Cette thèse avait pour objectif de développer un outil générique de prédiction de fonctions enzymatiques et des voies métaboliques à partir d'un génome complet et de l'appliquer au génome de la bactérie fixatrice d'azote *Sinorhizobium meliloti*.
Dans cette perspective, nous avons tout d'abord développé le programme PRIAM (*PRofils pour l'Identification Automatique du Métabolisme*), pour la prédiction automatique des fonctions enzymatiques et la visualisation des voies métaboliques d'un organisme.
Puis nous avons appliqué cette méthodologie sur le génome complet de *S. meliloti*, annoté de manière semi-automatique, afin d'analyser le métabolisme de cette bactérie symbiote de légumineuses.

- Programme de prédiction de fonctions enzymatiques et visualisation des voies métaboliques à partir d'un génome complet (PRIAM).

Les résultats croissants de la génomique acquis par le séquençage de nombreux génomes, par des expériences sur les transcriptomes ou les protéomes, nécessitent des outils automatiques de prédiction des fonctions des gènes. La méthode PRIAM permet une annotation automatique des enzymes car elle comprend des descripteurs spécifiques (PSSM) de certaines régions protéiques couplés à des règles permettant d'identifier les enzymes analogues, les enzymes oligomériques et les multi-enzymes. Par ailleurs, elle permet aussi une première analyse du métabolisme possible pour un organisme par la visualisation graphique des voies métaboliques.

- Application et analyse du métabolisme de la bactérie *Sinorhizobium meliloti*.

S. meliloti est une bactérie vivant en symbiose avec des légumineuses. La symbiose est un processus biologique qui met en jeu des événements de différenciation et des changements métaboliques chez les deux partenaires. C'est pourquoi, à partir du séquençage du génome de la bactérie, il est intéressant d'en déduire toutes ses voies métaboliques possibles. Cette analyse permet de mettre en évidence le rôle central du cycle de Krebs et l'utilisation

importante du cycle des pentoses phosphate plutôt que la voie de la glycolyse pour l'assimilation des sucres. Ainsi que la présence de voies spécifiques à cette bactérie comme l'assimilation de l'azote atmosphérique. L'utilisation de PRIAM sur les résultats d'expériences avec le protéome et le transcriptome de *S. meliloti* dans différentes conditions de croissance a montré son utilité pour comprendre la physiologie de cet organisme.

- Comparaison des métabolismes

L'utilisation d'un outil automatique de prédiction de fonctions comme PRIAM sur de nombreux génomes complets est une première étape pour envisager de comparer les métabolismes de nombreuses espèces. En effet l'analyse comparée de voies métaboliques dans différents génomes apporte des informations importantes sur l'évolution de la physiologie des espèces, permet d'identification de cibles pharmacologiques spécifiques à un organisme (153).

- Vérifier la fonctionnalité des voies métaboliques prédites

L'ensemble des voies métaboliques identifiées à partir des connaissances générales sur les enzymes trouvées dans d'autres espèces n'est pas suffisant pour affirmer que ses voies sont effectivement fonctionnelles. L'ensemble des enzymes identifiées avec PRIAM pour un génome d'intérêt peut être par la suite traité *in silico* pour décrire les chemins pour aller d'un substrat à un produit par la recherche des modes élémentaires. Cependant, il est bien sûr nécessaire de réaliser des expériences pour valider leur expression dans l'organisme.

En conclusion, nous proposons une méthode automatique d'annotation des fonctions enzymatiques à partir d'un génome complet. Méthode que nous avons validée sur plusieurs génomes complets et qui nous a permis une analyse approfondie du métabolisme de la bactérie *S .meliloti*.

BIBLIOGRAPHIE

1. Clamp, M., Andrews, D., Barker, D., Bevan, P., Cameron, G., Chen, Y., Clark, L., Cox, T., Cuff, J., Curwen, V. *et al.* (2003) Ensembl 2002: accommodating comparative genomics. *Nucleic Acids Res*, **31,** 38-42
2. Tipton, K. and Boyce, S. (2000) History of the enzyme nomenclature system. *Bioinformatics*, **16,** 34-40.
3. Bairoch, A. (2000) The ENZYME database in 2000. *Nucleic Acids Res*, **28,** 304-305
4. Goto, S., Okuno, Y., Hattori, M., Nishioka, T. and Kanehisa, M. (2002) LIGAND: database of chemical compounds and reactions in biological pathways. *Nucleic Acids Res*, **30,** 402-404
5. Schomburg, I., Chang, A. and Schomburg, D. (2002) BRENDA, enzyme data and metabolic information. *Nucleic Acids Res*, **30,** 47-49
6. Davidson, J.N. and Peterson, M.L. (1997) Origin of genes encoding multi-enzymatic proteins in eukaryotes. *Trends Genet*, **13,** 281-285.
7. Fitch, W.M. (2000) Homology a personal view on some of the problems. *Trends Genet*, **16,** 227-231
8. Nahum, L.A. and Riley, M. (2001) Divergence of function in sequence-related groups of Escherichia coli proteins. *Genome Res*, **11,** 1375-1381.
9. Galperin, M.Y. and Koonin, E.V. (1999) Functional genomics and enzyme evolution. Homologous and analogous enzymes encoded in microbial genomes. *Genetica*, **106,** 159-170
10. Ohno, S., Wolf, U. and Atkin, N.B. (1968) Evolution from fish to mammals by gene duplication. *Hereditas*, **59,** 169-187
11. Galperin, M.Y., Walker, D.R. and Koonin, E.V. (1998) Analogous enzymes: independent inventions in enzyme evolution. *Genome Res*, **8,** 779-790.
12. Smith, T.F. and Waterman, M.S. (1981) Identification of common molecular subsequences. *J Mol Biol*, **147,** 195-197
13. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. and Lipman, D.J. (1990) Basic local alignment search tool. *J Mol Biol*, **215,** 403-410.

14. Servant, F., Bru, C., Carrere, S., Courcelle, E., Gouzy, J., Peyruc, D. and Kahn, D. (2002) ProDom: automated clustering of homologous domains. *Brief Bioinform*, **3,** 246-251

15. Bateman, A., Birney, E., Cerruti, L., Durbin, R., Etwiller, L., Eddy, S.R., Griffiths-Jones, S., Howe, K.L., Marshall, M. and Sonnhammer, E.L. (2002) The Pfam protein families database. *Nucleic Acids Res*, **30,** 276-280.

16. Dayhoff, M.O., Schwartz, R.M. and Orcutt, B.C. (1978) In Dayhoff, M. O. (ed.), *Atlas of protein sequence and structure*. National biomedical research foundation Washington DC, Vol. 5, pp. 345-358.

17. Henikoff, S. and Henikoff, J.G. (1992) Amino acid substitution matrices from protein blocks. *Proc Natl Acad Sci U S A*, **89,** 10915-10919

18. Risler, J.L., Delorme, M.O., Delacroix, H. and Henaut, A. (1988) Amino acid substitutions in structurally related proteins. A pattern recognition approach. Determination of a new and efficient scoring matrix. *J Mol Biol*, **204,** 1019-1029

19. Altschul, S.F., Boguski, M.S., Gish, W. and Wootton, J.C. (1994) Issues in searching molecular sequence databases. *Nat Genet*, **6,** 119-129

20. Karlin, S. and Altschul, S.F. (1990) Methods for assessing the statistical significance of molecular sequence features by using general scoring schemes. *Proc Natl Acad Sci U S A*, **87,** 2264-2268

21. Needleman, S.B. and Wunsch, C.D. (1970) A general method applicable to the search for similarities in the amino acid sequence of two proteins. *J Mol Biol*, **48,** 443-453

22. Pearson, W.R. (1990) Rapid and sensitive sequence comparison with FASTP and FASTA. *Methods Enzymol*, **183,** 63-98

23. Thompson, J.D., Higgins, D.G. and Gibson, T.J. (1994) CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res*, **22,** 4673-4680

24. Corpet, F. (1988) Multiple sequence alignment with hierarchical clustering. *Nucleic Acids Res*, **16,** 10881-10890

25. Morgenstern, B. (1999) DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. *Bioinformatics*, **15,** 211-218

26. Notredame, C., Higgins, D.G. and Heringa, J. (2000) T-Coffee: A novel method for fast and accurate multiple sequence alignment. *J Mol Biol*, **302,** 205-217

27. Notredame, C. and Higgins, D.G. (1996) SAGA: sequence alignment by genetic algorithm. *Nucleic Acids Res*, **24,** 1515-1524

28. Gribskov, M., McLachlan, A.D. and Eisenberg, D. (1987) Profile analysis: detection of distantly related proteins. *Proc Natl Acad Sci U S A*, **84,** 4355-4358.

29. Gribskov, M., Luthy, R. and Eisenberg, D. (1990) Profile analysis. *Methods Enzymol*, **183,** 146-159

30. Altschul, S.F., Madden, T.L., Schaffer, A.A., Zhang, J., Zhang, Z., Miller, W. and Lipman, D.J. (1997) Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res*, **25,** 3389-3402.

31. Eddy, S.R. (1998) Profile hidden Markov models. *Bioinformatics*, **14,** 755-763

32. Krogh, A., Brown, M., Mian, I.S., Sjolander, K. and Haussler, D. (1994) Hidden Markov models in computational biology. Applications to protein modeling. *J Mol Biol*, **235,** 1501-1531

33. Hofmann, K. (2000) Sensitive protein comparisons with profiles and hidden Markov models. *Brief Bioinform*, **1,** 167-178.

34. Bucher, P., Karplus, K., Moeri, N. and Hofmann, K. (1996) A flexible motif search technique based on generalized profiles. *Comput Chem*, **20,** 3-23

35. Henikoff, S. and Henikoff, J.G. (1994) Position-based sequence weights. *J Mol Biol*, **243,** 574-578.

36. Tatusov, R.L., Altschul, S.F. and Koonin, E.V. (1994) Detection of conserved segments in proteins: iterative scanning of sequence databases with alignment blocks. *Proc Natl Acad Sci U S A*, **91,** 12091-12095

37. Schaffer, A.A., Aravind, L., Madden, T.L., Shavirin, S., Spouge, J.L., Wolf, Y.I., Koonin, E.V. and Altschul, S.F. (2001) Improving the accuracy of PSI-BLAST protein database searches with composition-based statistics and other refinements. *Nucleic Acids Res*, **29,** 2994-3005.

38. Muller, A., MacCallum, R.M. and Sternberg, M.J. (1999) Benchmarking PSI-BLAST in genome annotation. *J Mol Biol*, **293,** 1257-1271.

39. Lo Conte, L., Brenner, S.E., Hubbard, T.J., Chothia, C. and Murzin, A.G. (2002) SCOP database in 2002: refinements accommodate structural genomics. *Nucleic Acids Res*, **30,** 264-267.

40. Schaffer, A.A., Wolf, Y.I., Ponting, C.P., Koonin, E.V., Aravind, L. and Altschul, S.F. (1999) IMPALA: matching a protein sequence against a collection of PSI-BLAST-constructed position-specific score matrices. *Bioinformatics*, **15,** 1000-1011

41. Marchler-Bauer, A., Anderson, J.B., DeWeese-Scott, C., Fedorova, N.D., Geer, L.Y., He, S., Hurwitz, D.I., Jackson, J.D., Jacobs, A.R., Lanczycki, C.J. *et al.* (2003) CDD: a curated Entrez database of conserved domain alignments. *Nucleic Acids Res*, **31,** 383-387

42. Sonnhammer, E.L. and Kahn, D. (1994) Modular arrangement of proteins as inferred from analysis of homology. *Protein Sci*, **3,** 482-492

43. Park, J. and Teichmann, S.A. (1998) DIVCLUS: an automatic method in the GEANFAMMER package that finds homologous domains in single- and multi-domain proteins. *Bioinformatics*, **14,** 144-150

44. Gouzy, J., Corpet, F. and Kahn, D. (1999) Whole genome protein domain analysis using a new method for domain clustering. *Comput Chem*, **23,** 333-340.

45. Heger, A. and Holm, L. (2001) Picasso: generating a covering set of protein family profiles. *Bioinformatics*, **17,** 272-279.

46. Gouzy, J., Eugene, P., Greene, E.A., Kahn, D. and Corpet, F. (1997) XDOM, a graphical tool to analyse domain arrangements in any set of protein sequences. *Comput Appl Biosci*, **13,** 601-608.

47. Etzold, T., Ulyanov, A. and Argos, P. (1996) SRS: information retrieval system for molecular biology data banks. *Methods Enzymol*, **266,** 114-128

48. Schuler, G.D., Epstein, J.A., Ohkawa, H. and Kans, J.A. (1996) Entrez: molecular biology database and retrieval system. *Methods Enzymol*, **266,** 141-162

49. Fujibuchi, W., Goto, S., Migimatsu, H., Uchiyama, I., Ogiwara, A., Akiyama, Y. and Kanehisa, M. (1998) DBGET/LinkDB: an integrated database retrieval system. *Pac Symp Biocomput***,** 683-694.

50. Discala, C., Benigni, X., Barillot, E. and Vaysseix, G. (2000) DBcat: a catalog of 500 biological databases. *Nucleic Acids Res*, **28,** 8-9

51. Benson, D.A., Karsch-Mizrachi, I., Lipman, D.J., Ostell, J. and Wheeler, D.L. (2003) GenBank. *Nucleic Acids Res*, **31,** 23-27

52. Stoesser, G., Baker, W., Van Den Broek, A., Garcia-Pastor, M., Kanz, C., Kulikova, T., Leinonen, R., Lin, Q., Lombard, V., Lopez, R. *et al.* (2003) The EMBL Nucleotide Sequence Database: major new developments. *Nucleic Acids Res*, **31,** 17-22

53. Miyazaki, S., Sugawara, H., Gojobori, T. and Tateno, Y. (2003) DNA Data Bank of Japan (DDBJ) in XML. *Nucleic Acids Res*, **31,** 13-16

54. Boeckmann, B., Bairoch, A., Apweiler, R., Blatter, M.C., Estreicher, A., Gasteiger, E., Martin, M.J., Michoud, K., O'Donovan, C., Phan, I. *et al.* (2003) The SWISS-PROT

protein knowledgebase and its supplement TrEMBL in 2003. *Nucleic Acids Res*, **31,** 365-370

55. O'Donovan, C., Martin, M.J., Gattiker, A., Gasteiger, E., Bairoch, A. and Apweiler, R. (2002) High-quality protein knowledge resource: SWISS-PROT and TrEMBL. *Brief Bioinform*, **3,** 275-284
56. Wu, C.H., Yeh, L.S., Huang, H., Arminski, L., Castro-Alvear, J., Chen, Y., Hu, Z., Kourtesis, P., Ledley, R.S., Suzek, B.E. *et al.* (2003) The Protein Information Resource. *Nucleic Acids Res*, **31,** 345-347
57. Westbrook, J., Feng, Z., Chen, L., Yang, H. and Berman, H.M. (2003) The Protein Data Bank and structural genomics. *Nucleic Acids Res*, **31,** 489-491
58. Lo Conte, L., Ailey, B., Hubbard, T.J., Brenner, S.E., Murzin, A.G. and Chothia, C. (2000) SCOP: a structural classification of proteins database. *Nucleic Acids Res*, **28,** 257-259
59. Pearl, F.M., Lee, D., Bray, J.E., Sillitoe, I., Todd, A.E., Harrison, A.P., Thornton, J.M. and Orengo, C.A. (2000) Assigning genomic sequences to CATH. *Nucleic Acids Res*, **28,** 277-282.
60. Orengo, C.A., Brown, N.P. and Taylor, W.R. (1992) Fast structure alignment for protein databank searching. *Proteins*, **14,** 139-167
61. Pearl, F.M., Bennett, C.F., Bray, J.E., Harrison, A.P., Martin, N., Shepherd, A., Sillitoe, I., Thornton, J. and Orengo, C.A. (2003) The CATH database: an extended protein family resource for structural and functional genomics. *Nucleic Acids Res*, **31,** 452-455
62. Sigrist, C.J., Cerutti, L., Hulo, N., Gattiker, A., Falquet, L., Pagni, M., Bairoch, A. and Bucher, P. (2002) PROSITE: a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, **3,** 265-274
63. Attwood, T.K., Bradley, P., Flower, D.R., Gaulton, A., Maudling, N., Mitchell, A.L., Moulton, G., Nordle, A., Paine, K., Taylor, P. *et al.* (2003) PRINTS and its automatic supplement, prePRINTS. *Nucleic Acids Res*, **31,** 400-402
64. Mulder, N.J., Apweiler, R., Attwood, T.K., Bairoch, A., Barrell, D., Bateman, A., Binns, D., Biswas, M., Bradley, P., Bork, P. *et al.* (2003) The InterPro Database, 2003 brings increased coverage and new features. *Nucleic Acids Res*, **31,** 315-318
65. Ellis, L.B., Hou, B.K., Kang, W. and Wackett, L.P. (2003) The University of Minnesota Biocatalysis/Biodegradation Database: post-genomic data mining. *Nucleic Acids Res*, **31,** 262-265

66. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Collado-Vides, J., Paley, S.M., Pellegrini-Toole, A., Bonavides, C. and Gama-Castro, S. (2002) The EcoCyc Database. *Nucleic Acids Res*, **30,** 56-58.

67. Karp, P.D., Riley, M., Saier, M., Paulsen, I.T., Paley, S.M. and Pellegrini-Toole, A. (2000) The EcoCyc and MetaCyc databases. *Nucleic Acids Res*, **28,** 56-59.

68. Karp, P.D., Riley, M., Paley, S.M. and Pellegrini-Toole, A. (2002) The MetaCyc Database. *Nucleic Acids Res*, **30,** 59-61.

69. Paley, S.M. and Karp, P.D. (2002) Evaluation of computational metabolic-pathway predictions for Helicobacter pylori. *Bioinformatics*, **18,** 715-724

70. Kanehisa, M. (2002) The KEGG database. *Novartis Found Symp*, **247,** 91-101

71. Kuffner, R., Zimmer, R. and Lengauer, T. (2000) Pathway analysis in metabolic databases via differential metabolic display (DMD). *Bioinformatics*, **16,** 825-836.

72. Wittig, U. and De Beuckelaer, A. (2001) Analysis and comparison of metabolic pathway databases. *Brief Bioinform*, **2,** 126-142.

73. Ouzounis, C.A. and Karp, P.D. (2002) The past, present and future of genome-wide re-annotation. *Genome Biol*, **3**

74. Stephens, R.S., Kalman, S., Lammel, C., Fan, J., Marathe, R., Aravind, L., Mitchell, W., Olinger, L., Tatusov, R.L., Zhao, Q. *et al.* (1998) Genome sequence of an obligate intracellular pathogen of humans: Chlamydia trachomatis. *Science*, **282,** 754-759

75. Thebault, P., Servant, F., Schiex, T., Kahn, D. and Gouzy, J. (2000) L'environnement iANT: integrated ANnotation Tool. *Journées Ouvertes Biologie Informatique Mathématiques 2000*

76. Medigue, C., Rechenmann, F., Danchin, A. and Viari, A. (1999) Imagene: an integrated computer environment for sequence annotation and analysis. *Bioinformatics*, **15,** 2-15

77. Kretschmann, E., Fleischmann, W. and Apweiler, R. (2001) Automatic rule generation for protein annotation with the C4.5 data mining algorithm applied on SWISS-PROT. *Bioinformatics*, **17,** 920-926.

78. Andrade, M.A., Brown, N.P., Leroy, C., Hoersch, S., de Daruvar, A., Reich, C., Franchini, A., Tamames, J., Valencia, A., Ouzounis, C. *et al.* (1999) Automated genome sequence analysis and annotation. *Bioinformatics*, **15,** 391-412

79. Galperin, M.Y. and Koonin, E.V. (1998) Sources of systematic error in functional annotation of genomes: domain rearrangement, non-orthologous gene displacement and operon disruption. *In Silico Biol*, **1,** 55-67

80. Brenner, S.E. (1999) Errors in genome annotation. *Trends Genet*, **15,** 132-133
81. Devos, D. and Valencia, A. (2001) Intrinsic errors in genome annotation. *Trends Genet*, **17,** 429-431.
82. Kelley, L.A., MacCallum, R.M. and Sternberg, M.J. (2000) Enhanced genome annotation using structural profiles in the program 3D- PSSM. *J Mol Biol*, **299,** 499-520.
83. Karp, P.D. (2000) An ontology for biological function based on molecular interactions. *Bioinformatics*, **16,** 269-285.
84. Camon, E., Magrane, M., Barrell, D., Binns, D., Fleischmann, W., Kersey, P., Mulder, N., Oinn, T., Maslen, J., Cox, A. *et al.* (2003) The Gene Ontology Annotation (GOA) project: implementation of GO in SWISS-PROT, TrEMBL, and InterPro. *Genome Res*, **13,** 662-672
85. Consortium, T.G.O. (2001) Creating the gene ontology resource: design and implementation. *Genome Res*, **11,** 1425-1433
86. Iliopoulos, I., Tsoka, S., Andrade, M.A., Enright, A.J., Carroll, M., Poullet, P., Promponas, V., Liakopoulos, T., Palaios, G., Pasquier, C. *et al.* (2003) Evaluation of annotation strategies using an entire genome sequence. *Bioinformatics*, **19,** 717-726
87. Boneca, I.G., de Reuse, H., Epinat, J.C., Pupin, M., Labigne, A. and Moszer, I. (2003) A revised annotation and comparative analysis of Helicobacter pylori genomes. *Nucleic Acids Res*, **31,** 1704-1714
88. Tatusov, R.L., Galperin, M.Y., Natale, D.A. and Koonin, E.V. (2000) The COG database: a tool for genome-scale analysis of protein functions and evolution. *Nucleic Acids Res*, **28,** 33-36.
89. Pellegrini, M., Marcotte, E.M., Thompson, M.J., Eisenberg, D. and Yeates, T.O. (1999) Assigning protein functions by comparative genome analysis: protein phylogenetic profiles. *Proc Natl Acad Sci U S A*, **96,** 4285-4288.
90. Marcotte, E.M., Pellegrini, M., Thompson, M.J., Yeates, T.O. and Eisenberg, D. (1999) A combined algorithm for genome-wide prediction of protein function. *Nature*, **402,** 83-86.
91. Dandekar, T., Snel, B., Huynen, M. and Bork, P. (1998) Conservation of gene order: a fingerprint of proteins that physically interact. *Trends Biochem Sci*, **23,** 324-328.
92. Wolf, Y.I., Rogozin, I.B., Kondrashov, A.S. and Koonin, E.V. (2001) Genome alignment, evolution of prokaryotic genome organization, and prediction of gene function using genomic context. *Genome Res*, **11,** 356-372.

93. Enright, A.J., Iliopoulos, I., Kyrpides, N.C. and Ouzounis, C.A. (1999) Protein interaction maps for complete genomes based on gene fusion events. *Nature*, **402,** 86-90.

94. Snel, B., Bork, P. and Huynen, M. (2000) Genome evolution. Gene fusion versus gene fission. *Trends Genet*, **16,** 9-11.

95. Marcotte, E.M., Pellegrini, M., Ng, H.L., Rice, D.W., Yeates, T.O. and Eisenberg, D. (1999) Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285,** 751-753.

96. Karp, P.D., Krummenacker, M., Paley, S. and Wagg, J. (1999) Integrated pathway-genome databases and their role in drug discovery. *Trends Biotechnol*, **17,** 275-281

97. Goesmann, A., Haubrock, M., Meyer, F., Kalinowski, J. and Giegerich, R. (2002) PathFinder: reconstruction and dynamic visualization of metabolic pathways. *Bioinformatics*, **18,** 124-129.

98. Zien, A., Kuffner, R., Zimmer, R. and Lengauer, T. (2000) Analysis of gene expression data with pathway scores. *Proc Int Conf Intell Syst Mol Biol*, **8,** 407-417

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

100. Jeong, H., Tombor, B., Albert, R., Oltvai, Z.N. and Barabasi, A.L. (2000) The large-scale organization of metabolic networks. *Nature*, **407,** 651-654.

101. Schilling, C.H., Letscher, D. and Palsson, B.O. (2000) Theory for the systemic definition of metabolic pathways and their use in interpreting metabolic function from a pathway-oriented perspective. *J Theor Biol*, **203,** 229-248.

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

103. Schuster, S., Fell, D.A. and Dandekar, T. (2000) A general definition of metabolic pathways useful for systematic organization and analysis of complex metabolic networks. *Nat Biotechnol*, **18,** 326-332

104. Schuster, S., Pfeiffer, T., Moldenhauer, F., Koch, I. and Dandekar, T. (2002) Exploring the pathway structure of metabolism: decomposition into subnetworks and application to Mycoplasma pneumoniae. *Bioinformatics*, **18,** 351-361.

105. Pfeiffer, T., Sanchez-Valdenebro, I., Nuno, J.C., Montero, F. and Schuster, S. (1999) METATOOL: for studying metabolic networks. *Bioinformatics*, **15,** 251-257

106. Galibert, F., Finan, T.M., Long, S.R., Puhler, A., Abola, P., Ampe, F., Barloy-Hubler, F., Barnett, M.J., Becker, A., Boistard, P. *et al.* (2001) The composite genome of the legume symbiont Sinorhizobium meliloti. *Science*, **293,** 668-672

107. Djordjevic, M.A., Chen, H.C., Natera, S., Van Noorden, G., Menzel, C., Taylor, S., Renard, C., Geiger, O. and Weiller, G.F. (2003) A global analysis of protein expression profiles in Sinorhizobium meliloti: discovery of new genes for nodule occupancy and stress adaptation. *Mol Plant Microbe Interact*, **16,** 508-524

108. Burset, M. and Guigo, R. (1996) Evaluation of gene structure prediction programs. *Genomics*, **34,** 353-367.

109. Chandonia, J.M., Walker, N.S., Lo Conte, L., Koehl, P., Levitt, M. and Brenner, S.E. (2002) ASTRAL compendium enhancements. *Nucleic Acids Res*, **30,** 260-263

110. Salanoubat, M., Genin, S., Artiguenave, F., Gouzy, J., Mangenot, S., Arlat, M., Billault, A., Brottier, P., Camus, J.C., Cattolico, L. *et al.* (2002) Genome sequence of the plant pathogen Ralstonia solanacearum. *Nature*, **415,** 497-502.

111. Gattiker, A., Michoud, K., Rivoire, C., Auchincloss, A.H., Coudert, E., Lima, T., Kersey, P., Pagni, M., Sigrist, C.J., Lachaize, C. *et al.* (2003) Automated annotation of microbial proteomes in SWISS-PROT. *Comput Biol Chem*, **27,** 49-58

112. Fleischmann, R.D., Adams, M.D., White, O., Clayton, R.A., Kirkness, E.F., Kerlavage, A.R., Bult, C.J., Tomb, J.F., Dougherty, B.A., Merrick, J.M. *et al.* (1995) Whole-genome random sequencing and assembly of Haemophilus influenzae Rd. *Science*, **269,** 496-512

113. Fraser, C.M., Gocayne, J.D., White, O., Adams, M.D., Clayton, R.A., Fleischmann, R.D., Bult, C.J., Kerlavage, A.R., Sutton, G., Kelley, J.M. *et al.* (1995) The minimal gene complement of Mycoplasma genitalium. *Science*, **270,** 397-403

114. Himmelreich, R., Hilbert, H., Plagens, H., Pirkl, E., Li, B.C. and Herrmann, R. (1996) Complete sequence analysis of the genome of the bacterium Mycoplasma pneumoniae. *Nucleic Acids Res*, **24,** 4420-4449

115. Blattner, F.R., Plunkett, G., 3rd, Bloch, C.A., Perna, N.T., Burland, V., Riley, M., Collado-Vides, J., Glasner, J.D., Rode, C.K., Mayhew, G.F. *et al.* (1997) The complete genome sequence of Escherichia coli K-12. *Science*, **277,** 1453-1474

116. Shigenobu, S., Watanabe, H., Hattori, M., Sakaki, Y. and Ishikawa, H. (2000) Genome sequence of the endocellular bacterial symbiont of aphids Buchnera sp. APS. *Nature*, **407,** 81-86

117. Bono, H., Ogata, H., Goto, S. and Kanehisa, M. (1998) Reconstruction of amino acid biosynthesis pathways from the complete genome sequence. *Genome Res*, **8,** 203-210

118. Claudel-Renard, C., Chevalet, C., Faraut, T. and Kahn, D. (2003) Enzyme-specific profiles for genome annotation: PRIAM. *Nucleic Acids Res*, **31,** 6633-6639

119. Capela, D., Barloy-Hubler, F., Gouzy, J., Bothe, G., Ampe, F., Batut, J., Boistard, P., Becker, A., Boutry, M., Cadieu, E. *et al.* (2001) Analysis of the chromosome sequence of the legume symbiont Sinorhizobium meliloti strain 1021. *Proc Natl Acad Sci U S A*, **98,** 9877-9882

120. Shah, I. and Hunter, L. (1997) Predicting enzyme function from sequence: a systematic appraisal. *Proc Int Conf Intell Syst Mol Biol*, **5,** 276-283

121. Pollack, J.D. (1997) Mycoplasma genes: a case for reflective annotation. *Trends Microbiol*, **5,** 413-419.

122. Cordwell, S.J. (1999) Microbial genomes and "missing" enzymes: redefining biochemical pathways. *Arch Microbiol*, **172,** 269-279

123. Vasse, J., de Billy, F., Camut, S. and Truchet, G. (1990) Correlation between ultrastructural differentiation of bacteroids and nitrogen fixation in alfalfa nodules. *J Bacteriol*, **172,** 4295-4306

124. Denarie, J., Debelle, F. and Prome, J.C. (1996) Rhizobium lipo-chitooligosaccharide nodulation factors: signaling molecules mediating recognition and morphogenesis. *Annu Rev Biochem*, **65,** 503-535

125. Capela, D., Barloy-Hubler, F., Gatius, M.T., Gouzy, J. and Galibert, F. (1999) A high-density physical map of Sinorhizobium meliloti 1021 chromosome derived from bacterial artificial chromosome library. *Proc Natl Acad Sci U S A*, **96,** 9357-9362

126. Barloy-Hubler, F., Capela, D., Batut, J. and Galibert, F. (2000) High-resolution physical map of the pSymb megaplasmid and comparison of the three replicons of Sinorhizobium meliloti strain 1021. *Curr Microbiol*, **41,** 109-113

127. Barloy-Hubler, F., Capela, D., Barnett, M.J., Kalman, S., Federspiel, N.A., Long, S.R. and Galibert, F. (2000) High-resolution physical map of the Sinorhizobium meliloti 1021 pSyma megaplasmid. *J Bacteriol*, **182,** 1185-1189

128. Barnett, M.J., Fisher, R.F., Jones, T., Komp, C., Abola, A.P., Barloy-Hubler, F., Bowser, L., Capela, D., Galibert, F., Gouzy, J. *et al.* (2001) Nucleotide sequence and predicted functions of the entire Sinorhizobium meliloti pSymA megaplasmid. *Proc Natl Acad Sci U S A*, **98,** 9883-9888

129. Finan, T.M., Weidner, S., Wong, K., Buhrmester, J., Chain, P., Vorholter, F.J., Hernandez-Lucas, I., Becker, A., Cowie, A., Gouzy, J. *et al.* (2001) The complete sequence of the 1,683-kb pSymB megaplasmid from the N2-fixing endosymbiont Sinorhizobium meliloti. *Proc Natl Acad Sci U S A*, **98,** 9889-9894

130. Schiex, T., Gouzy, J., Moisan, A. and de Oliveira, Y. (2003) FrameD: A flexible program for quality check and gene prediction in prokaryotic genomes and noisy matured eukaryotic sequences. *Nucleic Acids Res*, **31,** 3738-3741

131. Kemp, R.G. and Tripathi, R.L. (1993) Pyrophosphate-dependent phosphofructo-1-kinase complements fructose 1,6-bisphosphatase but not phosphofructokinase deficiency in Escherichia coli. *J Bacteriol*, **175,** 5723-5724

132. Stowers, M.D. (1985) Carbon metabolism in Rhizobium species. *Annu Rev Microbiol*, **39,** 89-108

133. Voegele, R.T., Mitsch, M.J. and Finan, T.M. (1999) Characterization of two members of a novel malic enzyme class. *Biochim Biophys Acta*, **1432,** 275-285

134. Osteras, M., Driscoll, B.T. and Finan, T.M. (1997) Increased pyruvate orthophosphate dikinase activity results in an alternative gluconeogenic pathway in Rhizobium (Sinorhizobium) meliloti. *Microbiology*, **143 (Pt 5),** 1639-1648

135. Irigoyen, J.J., Sanchez-Diaz, M. and Emercich, D.W. (1990) Carbon metabolism enzymes of Rhizobium meliloti cultures and bacteroids and their distribution within alfalfa nodules. *Appl. Environ. Microbiol*, **56,** 2587-2589

136. McRae, D.G., Miller, R.W., Berndt, W.B. and Joy, K. (1989) Transport of C4-dicarboxylates and amino acids by Rhizobium meliloti bacteroids. *Molecular Plant-Microbe Interactions*, **2,** 273-278

137. Dunn, M.F. (1998) Tricarboxylic acid cycle and anaplerotic enzymes in rhizobia. *FEMS Microbiol Rev*, **22,** 105-123

138. Miller, R.W., McRae, D.G. and Joy, K. (1991) Glutamate and gamma-aminobutyrate metabolism in isolated Rhizobium bacteroids. *Molecular Plant-Microbe Interactions*, **4,** 37-45

139. De Hertogh, A.A., Mayeux, P.A. and Evans, H.J. (1964) The relationship of cobalt requirement to propionate metabolism in Rhizobium. *The Journal of Biological Chemistry*, **239,** 2446-2453

140. Marais, A., Mendz, G.L., Hazell, S.L. and Megraud, F. (1999) Metabolism and genetics of Helicobacter pylori: the genome era. *Microbiol Mol Biol Rev*, **63,** 642-674

141. Delgado, M.J., Bedmar, E.J. and Downie, J.A. (1998) Genes involved in the formation and assembly of rhizobial cytochromes and their role in symbiotic nitrogen fixation. *Adv Microb Physiol*, **40,** 191-231

142. Appleby, C.A. (1984) Leghemoglobin and Rhizobium respiration. *Annu Rev Plant Physiol*, **35,** 443-478

143. Kereszt, A., Slaska-Kiss, K., Putnoky, P., Banfalvi, Z. and Kondorosi, A. (1995) The cycHJKL genes of Rhizobium meliloti involved in cytochrome c biogenesis are required for "respiratory" nitrate reduction ex planta and for nitrogen fixation during symbiosis. *Mol Gen Genet*, **247,** 39-47

144. Vorholt, J.A., Chistoserdova, L., Stolyar, S.M., Thauer, R.K. and Lidstrom, M.E. (1999) Distribution of tetrahydromethanopterin-dependent enzymes in methylotrophic bacteria and phylogeny of methenyl tetrahydromethanopterin cyclohydrolases. *J Bacteriol*, **181,** 5750-5757

145. Chistoserdova, L., Vorholt, J.A., Thauer, R.K. and Lidstrom, M.E. (1998) C1 transfer enzymes and coenzymes linking methylotrophic bacteria and methanogenic Archaea. *Science*, **281,** 99-102

146. Vorholt, J.A., Chistoserdova, L., Lidstrom, M.E. and Thauer, R.K. (1998) The NADP-dependent methylene tetrahydromethanopterin dehydrogenase in Methylobacterium extorquens AM1. *J Bacteriol*, **180,** 5351-5356

147. Pomper, B.K., Vorholt, J.A., Chistoserdova, L., Lidstrom, M.E. and Thauer, R.K. (1999) A methenyl tetrahydromethanopterin cyclohydrolase and a methenyl tetrahydrofolate cyclohydrolase in Methylobacterium extorquens AM1. *Eur J Biochem*, **261,** 475-480

148. Goenrich, M., Bartoschek, S., Hagemeier, C.H., Griesinger, C. and Vorholt, J.A. (2002) A glutathione-dependent formaldehyde-activating enzyme (Gfa) from Paracoccus denitrificans detected and purified via two-dimensional proton exchange NMR spectroscopy. *J Biol Chem*, **277,** 3069-3072

149. Schwedock, J.S. and Long, S.R. (1992) Rhizobium meliloti genes involved in sulfate activation: the two copies of nodPQ and a new locus, saa. *Genetics*, **132,** 899-909

150. Kerppola, T.K. and Kahn, M.L. (1988) Symbiotic phenotypes of auxotrophic mutants of Rhizobium meliloti 104A14. *J Gen Microbiol*, **134 (Pt 4),** 913-919

151. Streit, W.R., Joseph, C.M. and Phillips, D.A. (1996) Biotin and other water-soluble vitamins are key growth factors for alfalfa root colonization by Rhizobium meliloti 1021. *Mol Plant Microbe Interact*, **9,** 330-338

152. Streit, W.R. and Phillips, D.A. (1996) Recombinant Rhizobium meliloti strains with extra biotin synthesis capability. *Appl Environ Microbiol*, **62,** 3333-3338

153. Dandekar, T., Schuster, S., Snel, B., Huynen, M. and Bork, P. (1999) Pathway alignment: application to the comparative analysis of glycolytic enzymes. *Biochem J*, **343,** 115-124.