Modules réactionnels : un nouveau concept pour étudier l'évolution des voies métaboliques

Matthieu Barba

Thèse

présentée par

Matthieu BARBA

pour obtenir le grade de docteur de l'Université Paris 11

École doctorale « Gènes, Génomes, Cellules »

Modules réactionnels : un nouveau concept pour étudier l'évolution des voies métaboliques

Soutenue le 16 décembre 2011 à l'IGM, Université Paris-Sud XI, devant le Jury composé de :

M[me] MÉDIGUE Claudine	DR CNRS	Rapporteur
M[r] RUBIO Vicente	Investigador Responsable	Rapporteur
M[r] CASANE Didier	PR P7	Examinateur
M[r] GUYONVARCH Armel	PR P11	Examinateur
M[me] LEGRAIN Christianne	CR Principale	Examinatrice
M[r] LABEDAN Bernard	DR CNRS	Directeur de thèse

Travaux réalisés à l'Institut de Génétique et Microbiologie (IGM, UMR 8621)
à l'Université Paris-Sud XI, Orsay

UNIVERSITÉ PARIS SUD

FACULTÉ DES SCIENCES D'ORSAY

IGM Institut de Génétique et Microbiologie

Résumé :

J'ai mis au point une méthodologie pour annoter les superfamilles d'enzymes, en décrire l'histoire et les replacer dans l'évolution de leurs voies métaboliques. J'en ai étudié trois : (1) les amidohydrolases cycliques, dont les DHOases (dihydroorotases, biosynthèse des pyrimidines), pour lesquelles j'ai proposé une nouvelle classification. L'arbre phylogénétique inclut les dihydropyrimidinases (DHPases) et allantoïnases (ALNases) qui ont des réactions similaires dans d'autres voies (dégradation des pyrimidines et des purines respectivement). (2) L'étude de la superfamille des DHODases (qui suivent les DHOases) montre une phylogénie semblable aux DHOases, avec également des enzymes d'autres voies, dont les DHPDases (qui suivent les DHPases). De cette observation est né le concept de module réactionnel, qui correspond à la conservation de l'enchaînement de réactions semblables dans différentes voies métaboliques. Cela a été utilisé lors de (3) l'étude des carbamoyltransférases (TCases) qui incluent les ATCases (précédant les DHOases). J'ai d'abord montré l'existence d'une nouvelle TCase potentiellement impliquée dans la dégradation des purines et lui ai proposé un nouveau rôle en utilisant le concept de module réactionnel (enchaînement avec l'ALNase). Dans ces trois grandes familles j'ai aussi mis en évidence trois groupes de paralogues non identifiés qui se retrouvent pourtant dans un même contexte génétique appelé « Yge » et qui formeraient donc un module réactionnel constitutif d'une nouvelle voie hypothétique. Appliqué à diverses voies, le concept de modules réactionnels refléterait donc les voies métaboliques ancestrales dont ils seraient les éléments de base.

Mots clés :

alignement multiple, arbre phylogénétique, superfamille, amidohydrolase, dihydroorotase, carbamoyltransférase, voie métabolique, réaction chimique, ambigüité de substrat, module réactionnel

Titre en anglais :

Reaction modules: a new concept to study the evolution of metabolic pathways

Résumé en anglais :

I designed a methodology to annotate enzyme superfamilies, explain their history and describe them in the context of metabolic pathways evolution. Three superfamilies were studied: (1) cyclic amidohydrolases, including DHOases (dihydroorotases, third step of the pyrimidines biosynthesis), for which I proposed a new classification. The phylogenetic tree also includes dihydropyrimidinases (DHPases) and allantoinases (ALNases) which catalyze similar reactions in other pathways (pyrimidine and purine degradation, respectively). (2) The DHODases superfamily (after DHOases) show a similar phylogeny as DHOases, including enzymes from other pathways, DHPDases in particular (after DHPases). This led to the concept of reaction module, i.e. a conserved series of similar reactions in different metabolic pathways. This was used to study (3) the carbamoyltransferases (TCases) which include ATCases (before DHOases). I first isolated a new kind of TCase, potentially involved in the purine degradation, and I proposed a new role for it in the light of reaction modules (linked with ALNase). In those three superfamilies I also found three groups of unidentified paralogs that were remarkably part of the same genetic context called “Yge” which would be a reaction module part of an unidentified pathway. The concept of reactions modules may then reflect the ancestral metabolic pathways for which they would be basic elements.

Mots clés en anglais :

multiple sequence alignment, phylogenetic tree, amidohydrolase, dihydroorotase, carbamoyltransferase, metabolic pathway, chemical reaction, substrate ambiguity, reaction module

Remerciements

Je tiens à remercier tout d'abord toute l'équipe de mon laboratoire, en premier lieu Bernard Labedan qui m'a accueilli et accompagné durant toute ma thèse. Il m'a supporté, soutenu et encouragé pendant tout mon travail, même dans des conditions difficiles, et c'est grâce à lui que j'ai pu réaliser ma thèse : je lui en suis très reconnaissant. Je remercie également Olivier Lespinet, toujours présent, et l'ensemble des collègues qui se sont succédé dans le laboratoire : post-docs, doctorants et stagiaires, notamment Stéphane Descorps-Declère. Je remercie également Christianne Legrain qui a gracieusement accepté une collaboration sur mes travaux.

Je remercie l'ensemble du personnel de l'IGM et notamment le groupe des non-statutaires qui m'ont permis de travailler dans un cadre agréable.

Merci aussi à tous les BIBS, enseignants et camarades qui m'ont accompagnés durant mes dernières années d'étude.

Je remercie toute ma famille et en particulier ma sœur Fanny qui a su rester proche de moi, mon frère Thomas qui était toujours là malgré la distance. Je remercie enfin ma mère, qui m'a soutenu pendant toutes mes études, dans tous mes choix, et qui est restée près de moi en toutes circonstances.

Table des matières

Table des annexes

Index des figures

Index des tables

PRÉAMBULE

Le métabolisme est l'ensemble des réactions chimiques catalysées par les organismes vivants. Il s'agit d'un processus hautement intégré et extrêmement régulé de création ou de décomposition de composants organiques afin de produire de l'énergie et/ou des matières premières. Le métabolisme est assuré par un ensemble d'enzymes qui catalysent les étapes successives généralement réunies en voies pour permettre la synthèse ou la dégradation de molécules variées. Ainsi existe-t-il des voies de dégradation du glucose, des voies de biosynthèse des acides aminés, ou encore des voies de recyclage des acides nucléiques (ADN, ARN). Classiquement, on différencie le métabolisme primaire qui permet de produire les constituants essentiels de la cellule (essentiellement les acides nucléiques, les protéines, les constituants membranaires) du métabolisme secondaire qui produit des éléments plus spécifiques permettant une adaptation aux conditions de vie et d'environnement.

Les enzymes sont elles-mêmes le produit de l'expression de l'information génétique de l'organisme. La correspondance un gène = une enzyme = une fonction (par exemple une étape dans une voie métabolique) fut proposée dès 1941 [Beadle & Tatum 1941], mais fut rapidement considérée comme trop simple. Par exemple, chez la bactérie modèle *Escherichia coli* la voie de biosynthèse *de novo* des pyrimidines est constituée de six étapes, catalysées par six enzymes différentes, elles-mêmes codées par six gènes présents en différentes localisations le long du chromosome bactérien. Par contre, chez tous les mammifères (dont l'homme) les trois premières étapes sont assurées par une seule protéine, qui est en fait le produit de la fusion des trois premiers gènes de la voie de biosynthèse. De même, les deux dernières étapes de la voie sont assurées par une seule protéine à deux domaines catalysant respectivement les deux dernières réactions (voir la Figure 1).

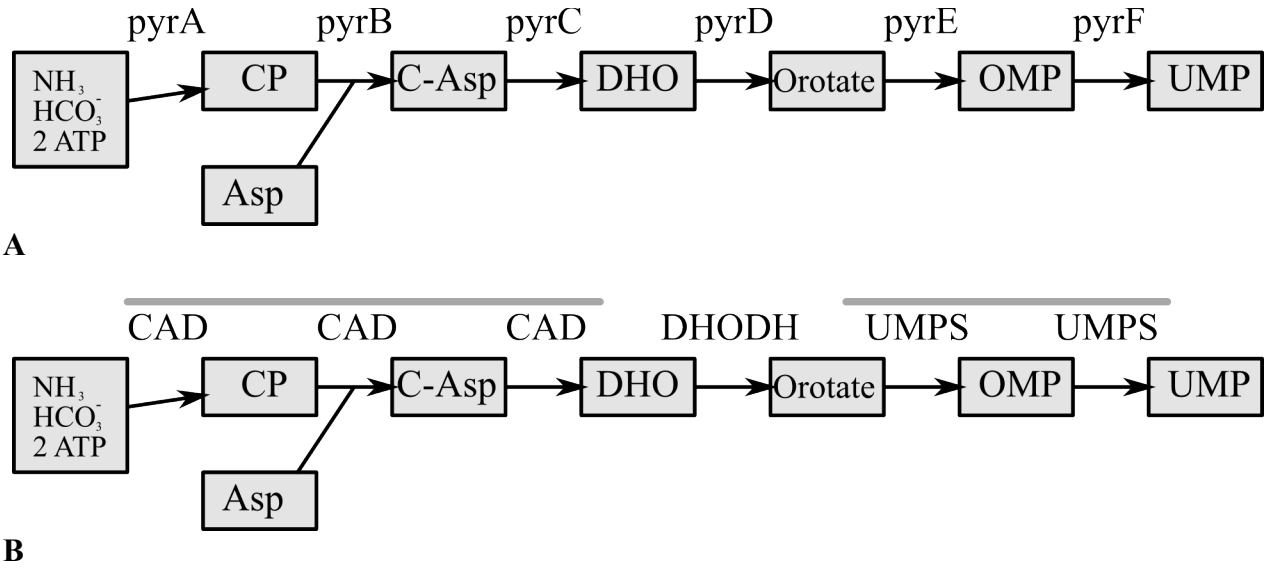


Figure 1 : Représentation simplifiée de l'enchaînement des six enzymes de la voie de biosynthèse des pyrimidines.
A : Les gènes pyrA à pyrF codent chacun une enzyme chez Escherichia coli.
B : Chez les mammifères, les trois premières étapes sont réalisées avec un seul polypeptide codé par un gène unique (CAD), de même pour les deux dernières étapes (UMPS).

Une complexité supplémentaire apparaît chez d'autres organismes où certaines étapes de ces voies métaboliques sont réalisées par des enzymes non homologues, c'est-à-dire qui ne dérivent pas d'un même gène ancestral. Par exemple, chez quelques espèces d'Archées, la première étape de la voie de biosynthèse des pyrimidines pourrait être réalisée par une carbamate kinase au lieu d'une carbamoyl-phosphate synthétase [Durbecq et al. 1997] qui est, elle, présente chez la plupart des autres organismes appartenant aux trois grands domaines du vivant (Archées, Bactéries, Eucaryotes) [Woese, Kandler & Wheelis 1990].

Enfin, d'un organisme à l'autre, les voies métaboliques peuvent emprunter des chemins différents pour arriver au même résultat. Par exemple, il existe actuellement quatre variations connues de la voie de biosynthèse de l'arginine qui partagent certaines étapes en commun, les autres suivant des chemins différents (cf Figure 2).

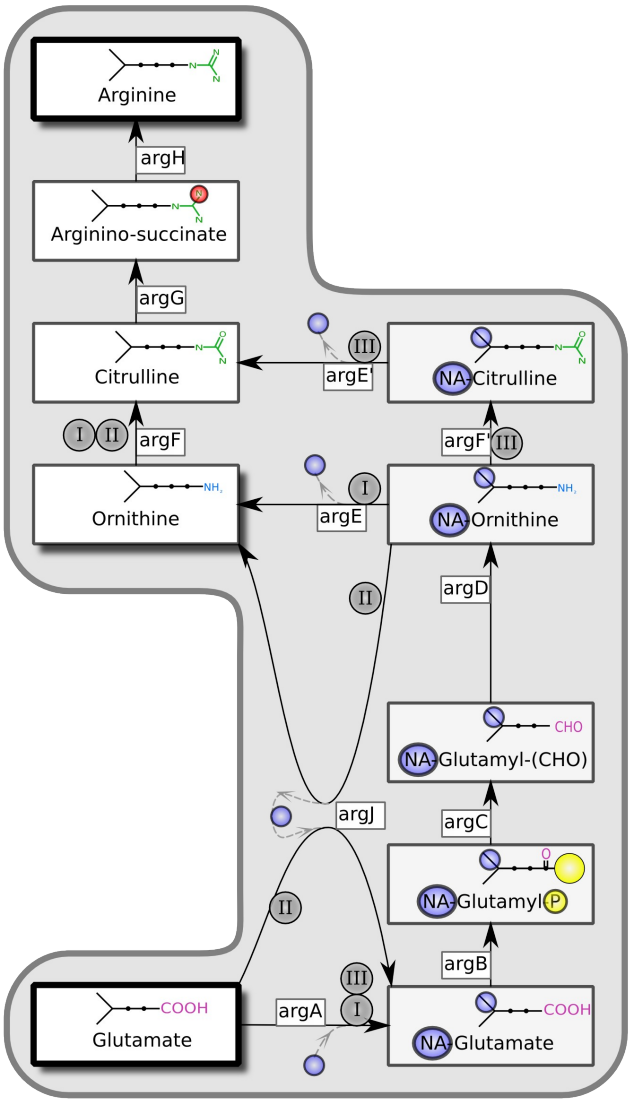


Figure 2 : Trois des variantes de la voie de biosynthèse de l'arginine.
La biosynthèse commence à partir du glutamate et emprunte l'un des trois chemins I, II ou III. Les cadres contiennent les substrats et produits, les flèches désignent les réactions catalysées par les produits des gènes arg (argA à argJ).

Un des premiers objectifs de mon travail de Thèse a été l'étude approfondie de la famille du gène *pyrC* codant la dihydroorotase, l'enzyme qui catalyse la troisième étape de la voie de biosynthèse *de novo* des pyrimidines (voir Figure 1). À cette occasion, j'ai montré que même pour une fonction moléculaire précise, il existait de grandes variations tant phylogénétiques que structurales séparant les enzymes homologues qui la composent. L'étude de cette famille m'a amené

à redéfinir une classification de ces enzymes basée sur la phylogénie, obtenue à partir des données de séquence et de structure, et sur leur contexte génétique (Chapitre 6, p 60). Pour ce faire, j'ai mis au point une nouvelle approche méthodologique d'obtention d'alignement multiple de séquences et d'addition automatisée de nouveaux homologues permettant de reconstruire un arbre phylogénétique exact et toujours à jour (Chapitre 4, p 38).

À partir de cette base solide, l'étude plus globale d'autres enzymes homologues appartenant à la même super-famille que les dihydroorotases m'a amené vers le sujet central de ma thèse, à savoir la définition du concept de module réactionnel afin de retracer les modes de construction et d'évolution des voies métaboliques. Ce concept est né de la réalisation que les enzymes appartenant à cette superfamille d'amidohydrolases cycliques interviennent dans plusieurs voies (voie de biosynthèse des pyrimidines et voies de dégradation des pyrimidines et des purines) qui partagent une succession similaire de réactions chimiquement semblables. Une telle approche conceptuelle permet de comparer l'évolution de diverses voies métaboliques d'un nouveau point de vue, éclairant la manière dont ont été utilisées les briques élémentaires (modules réactionnels) qui constituent ces voies pour les organiser en combinaisons variées correspondant à autant de modules fonctionnels définis dans le cadre de la biologie systémique [Hartwell et al. 1999]. Cette nouvelle approche conceptuelle m'a permis aussi de proposer de nouvelles démarches pour identifier des gènes mal annotés qui pourraient correspondre à des orphelins métaboliques ou coder des activités enzymatiques orphelines (Chapitre 7, p 95).

INTRODUCTION

Chapitre 1 : Fonctionnement dynamique des cellules vivantes et métabolisme

Tout être vivant est constitué d'une ou de plusieurs cellules dont la taille, la morphologie et le mode d'interaction avec l'environnement varient énormément d'une espèce à une autre. Outre les êtres pluricellulaires évidents à nos yeux, il existe de nombreux organismes unicellulaires dont une partie a un rôle essentiel pour l'homme. De nombreux microbes pathogènes sont ainsi responsables de maladies variées, alors que par ailleurs les communautés de microorganismes commensaux constituant la flore intestinale (microbiome des mammifères) jouent un rôle majeur dans la physiologie et le comportement [Qin et al. 2010]. De telles propriétés sont actuellement de plus en plus utilisées en recherche appliquée (chimie, pharmacologie, agronomie, etc...).

Si l'on considère l'ensemble de la biosphère, on trouve en fait des êtres vivants capables de coloniser l'immense majorité des milieux de la Terre, y compris les plus extrêmes : acides, basiques, salés, chauds, froids, irradiés, en altitude, et dans les profondeurs des océans et des roches de la Terre [Pace 1997]. Tous ces êtres vivants ont dû s'équiper différemment pour faire face à toutes ces conditions de vie si différentes. Cependant, ils sont tous formés des mêmes composants essentiels à leur réplication et à leur fonctionnement de base [Pace 2001].

Pour assurer ce fonctionnement, les organismes vivants ont besoin 1) d'énergie, pour assurer l'ensemble des processus de la cellule, et 2) de matières premières, pour construire toutes les molécules nécessaires à leur entretien et à leur survie. L'énergie peut venir de la consommation de matières organiques par modification physico-chimique, ou d'autres processus comme la photosynthèse qui permet de convertir l'énergie lumineuse en énergie chimique. Les matières premières proviennent du milieu extérieur, par exemple l'eau ou le dioxyde de carbone, ou bien de la dégradation de molécules plus complexes qui sont récupérées ou recyclées (cas classique de la glycolyse dégradant le glucose). C'est l'exploitation de l'énergie cellulaire et la production de matière ou sa dégradation qui constitue le métabolisme d'un être vivant. Celui-ci peut être divisé en deux catégories de directions opposées : le catabolisme, qui est la dégradation de molécules pour en extraire leur énergie et récupérer leur matière, et l'anabolisme pour créer les composants cellulaires de base à partir de précurseurs, ce qui nécessite de l'énergie. Ce métabolisme est réalisé dans les cellules vivantes par un ensemble complexe et hautement intégré de voies métaboliques qui sont

soumises à de nombreux mécanismes de régulation au niveau de l'expression des gènes et de l'activité catalytique de leurs produits, les enzymes.

Une voie métabolique peut être vue comme une chaîne de réactions chimiques catalysées par des enzymes synthétisées par un ou plusieurs gènes de l'organisme (Figure 1). Les enzymes travaillent donc à la chaîne, en transformant un substrat A en un produit B, qui devient le substrat de l'enzyme suivante qui le transforme en produit C, et ainsi de suite jusqu'à obtenir le produit final. Chaque réaction est catalysée par une enzyme, mais certaines réactions peuvent également se produire spontanément sans catalyse.

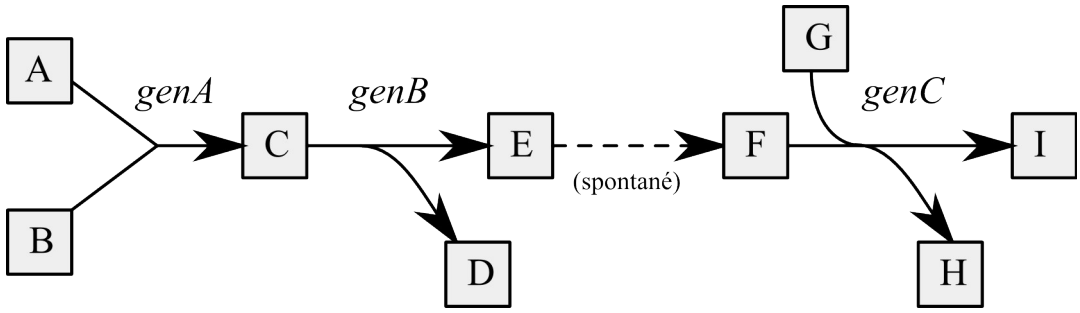


Figure 3 : Schéma d'une voie métabolique hypothétique.

Les molécules A à F prennent part à la voie. A et B sont des substrats qui sont combinés en C par l'enzyme codée par le gène genA. *C est ensuite scindé en E et D (avec* genB*). À la troisième réaction E se transforme spontanément en F. Enfin F et G se combinent en H et I (avec* genC*). Chaque réaction est catalysée par une enzyme, mais certaines réactions peuvent également se produire spontanément sans catalyse.*

À moins de la récupérer telle quelle ou de recycler d'autres composants, les êtres vivants peuvent produire leur matière organique *de novo*, c'est-à-dire en utilisant seulement des éléments de base. La création et la transformation de ces produits chimiques se fait par des voies de biosynthèse présentes dans les organismes en question. Cependant la présence des voies anaboliques peut grandement varier d'une espèce à l'autre.

Par exemple, l'homme peut synthétiser *de novo* dix acides aminés sur les vingt constituant les protéines. Parmi les dix autres, huit sont essentiels (apportés par l'alimentation) et deux semi-essentiels (apport par le lait maternel chez les nourrissons) [Laidlaw & Kopple 1987]. L'homme peut également produire tous les nucléotides nécessaires au maintien de son information génétique. Par contre, certains parasites ou d'autres organismes plus ou moins dépendants de leur hôte (à l'image des *Pasteurellales* qui peuvent être soit commensales soit parasites des voies respiratoires chez de nombreux animaux) ont perdu la voie de biosynthèse des pyrimidines pourtant présente

dans la quasi-totalité des êtres vivants. Ces organismes utilisent alors les composants de l'hôte qu'ils envahissent. Dans d'autres cas exceptionnels, certains organismes en symbiose se partagent le travail : c'est le cas d'un ver des grands fonds marins (*Riftia pachyptila)* et d'une bactérie sulfo-oxydante. La bactérie produit le début de la voie de biosynthèse des pyrimidines dans le trophosome du ver (sac symbiotique), et le ver se charge de la terminer [Minic et al. 2001].

Tous les acides aminés et nucléotides peuvent être produits *de novo* mais nécessitent une matière première élémentaire généralement apportée par le catabolisme (Figure 4), en dehors de l'apport de matières carbonées issues de la photosynthèse.

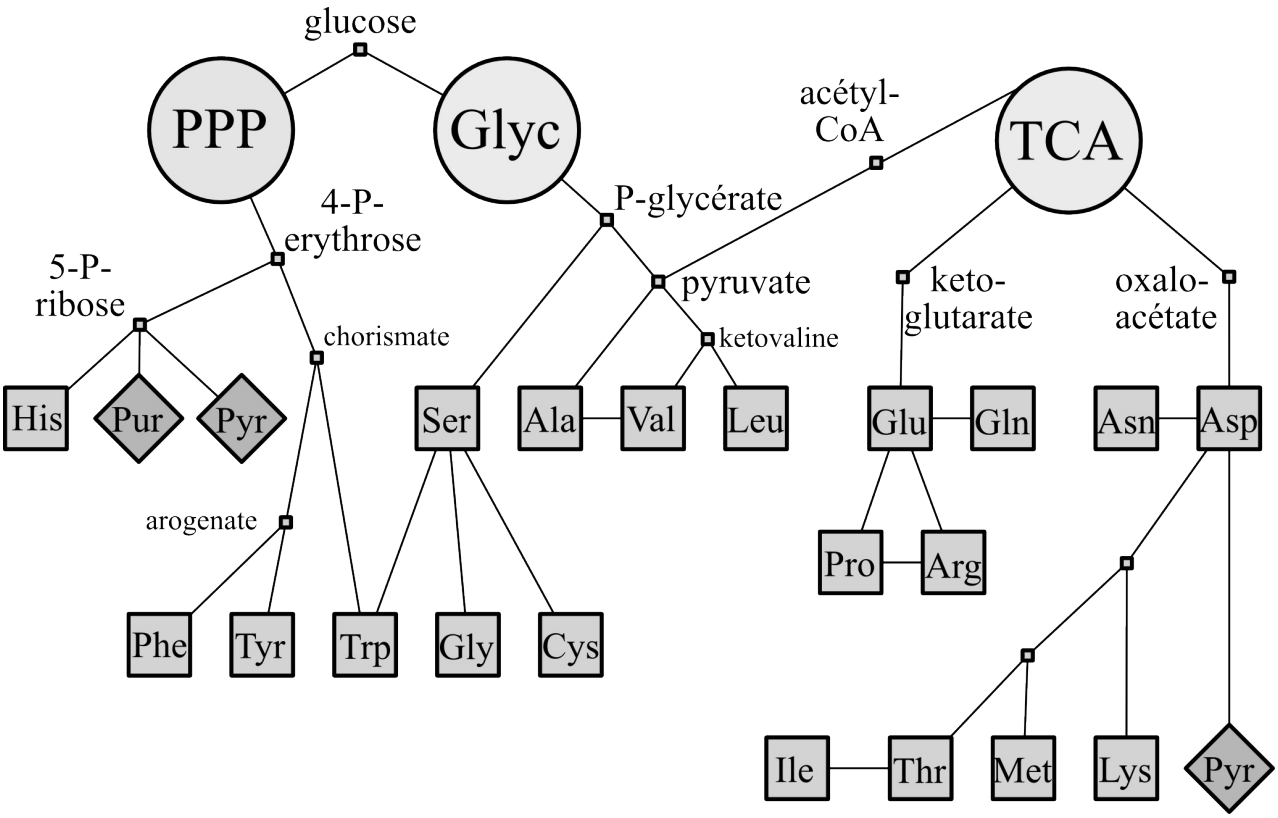


Figure 4: Origine du carbone dans la biosynthèse des acides aminés et des nucléotides.
Les grands cercles correspondent à trois grande voies de dégradation : la voie des pentoses phosphates (PPP), la glycolyse (Glyc) et le cycle de Krebs (TCA) qui dégradent le glucose. Elles fournissent la matière carbonée nécessaire à la synthèse de novo *des acides aminés (carrés) et des nucléotides (losanges Pyr, Pur ; les deux losanges Pur correspondent à deux composés qui sont combinés pour la création des pyrimidines). Les petits carrés sont des intermédiaires notables de ces voies.*

La plupart des composants synthétisés par les êtres vivants peuvent également être dégradés et recyclés pour en retirer énergie et matière première qui seront réutilisés pour le fonctionnement de la cellule. Le glucose par exemple peut être dégradé dans plusieurs voies fondamentales (Figure

4) : la glycolyse, qui donne un peu d'énergie sous forme (finale) d'ATP[1] et produit du pyruvate ; le cycle de Krebs qui transforme ce pyruvate et dont un cycle complet rapporte une grande quantité d'énergie sous forme d'ATP, tout en produisant plusieurs intermédiaires utilisés dans des voies de biosynthèse ; et la voie des pentoses-phosphates qui produit de l'énergie sous forme de pouvoir réducteur ($NADPH,H^+$) mais aussi produit des précurseurs d'acides aminés ainsi que du ribose-5-phosphate, constituant des nucléotides.

Les pyrimidines et purines (nucléotides), ainsi que chaque acide aminé peuvent être dégradés chacun par une voie particulière, mais la présence de ces voies varie énormément selon les espèces.

1 L'ATP est le vecteur de l'énergie chimique des cellules qui fournit l'énergie nécessaire aux réactions chimiques qui en consomment.

Chapitre 2 : Voies métaboliques des pyrimidines et des purines

Le métabolisme des pyrimidines et des purines peut se décomposer en différentes voies que j'ai étudiées au cours de cette Thèse. Ayant réalisé que l'on pouvait montrer que ces voies présentent des similarités remarquables entre certaines de leurs étapes, je me suis intéressé plus précisément aux trois voies suivantes :

- la voie de biosynthèse des pyrimidines ;
- la voie de dégradation réductrice des pyrimidines ;
- la voie de dégradation des purines.

1 Voie de biosynthèse des pyrimidines

Pour créer les pyrimidines (les bases uridine, cytosine et thymidine, cf Annexe p155) *de novo*, tous les êtres vivants – Eucaryotes, Bactéries, Archées – utilisent une voie similaire constituée de 6 étapes successives, catalysées par un nombre variable d'enzymes (Tableau 1, Figure 5).

Étape	Gène	Substrats	Produits	Noms complets (abrégés)
1	*pyrA*	HCO_3^- + NH_3 + 2 ATP	Carbamoyl-phosphate (CP) + H_2O + 2 ADP	carbamoyl-phosphate synthétase (CPSase)
2	*pyrB*	Aspartate + CP	Carbamoyl-aspartate + P	Aspartate carbamoyltransférase (ATCase)
3	*pyrC*	Carbamoyl-aspartate	Dihydroorotate +H_2O	Dihydroorotase (DHOase)
4	*pyrD*	Dihydroorotate	Orotate +$2H^+$ + $2e^-$	Dihydroorotate déshydrogénase (DHODase)
5	*pyrE*	Orotate + PRPP	OMP + PP	Orotate phosphoribosyltransférase (OPRTase)
6	*pyrF*	OMP	UMP +CO_2	OMP décarboxylase (OMPDCase)

Tableau 1 : Les 6 enzymes catalysant la voie de biosynthèse des pyrimidines.
CP = carbamoyl-phosphate ; P=phosphate ; PP=diphosphate ; PRPP=5-phosphoribosyl diphosphate.

1.1 Étape 1 : synthèse du carbamoyl-phosphate

La voie de biosynthèse commence par la création d'un carbamoyl-phosphate par la carbamoyl-phosphate synthétase ou CPSase (PyrA). Elle utilise deux ATP.

Chez la plupart des êtres vivants, cette première étape est catalysée par un ensemble de deux enzymes, CarA et CarB. CarB est constituée de deux domaines homologues, et est parfois séparée en deux chaînes (CarB1 et CarB2) chez quelques espèces : *Aquifex aeolicus* [Ahuja et al. 2001] et de nombreux Firmicutes. Chez beaucoup de Bactéries, les gènes *carA* et *carB* ne sont pas au voisinage des gènes *pyr* mais sont généralement localisés dans l'opéron *arg* au voisinage de *argF* codant l'ornithine carbamoyltransférase, un paralogue de l'aspartate carbamoyltransférase, la prochaine étape de la voie (voir ci-dessous et Chapitre 7). Ceci est un premier exemple de ces nombreuses et complexes inter-relations entre voies métaboliques impliquées dans des fonctions cellulaires différentes ; en effet, chez de nombreuses espèces, cette CPSase joue un rôle dans deux voies en même temps (biosynthèse des pyrimidines, biosynthèse de l'arginine).

Chez quelques rares espèces sans CPSase, essentiellement des Archées, cette première étape pourrait être réalisée par une carbamate kinase, qui catalyse généralement la dégradation du carbamoyl-phosphate dans les voies cataboliques. Cela a notamment été proposé chez *Pyrococcus furiosus* [Durbecq et al. 1997].

Enfin, chez les Métazoaires, les Champignons et d'autres groupes d'Eucaryotes, CarA et CarB sont fusionnés dans une unique chaîne polypeptidique (CAD) qui contient également les domaines des deux étapes suivantes, ATCase et DHOase (voir plus loin en 1.4).

1.2 Étape 2 : aspartate carbamoyltransférase

La seconde étape est l'ajout du groupement carbamoyl à l'aspartate sur son groupement amine. Cette réaction est catalysée par l'aspartate carbamoyltransférase ou ATCase (PyrB) qui a été intensivement étudiée en tant que modèle d'enzyme allostérique [Monod, Wyman & Changeux 1965], c'est-à-dire qu'elle change de conformation en se liant à une molécule effectrice, ce qui modifie sa cinétique [Kantrowitz & Lipscomb 1990]. Dans le cas d'*E. coli*, l'ATCase forme un complexe dodécamérique, constitué de deux trimères de sous-unités catalytiques (PyrB), et de trois dimères de sous-unités régulatrices (PyrI). Dans d'autres cas, comme chez *Aquifex aeolicus*, l'ATCase se présente sous une forme 4D similaire mais avec des dihydroorotases (PyrC) à la place des sous-unités régulatrices.

Les ATCases sont classées en deux familles : ATC I et ATC II [Labedan et al. 1999 ; Labedan et al. 2004]. C'est dans le groupe II que l'on retrouve les complexes PyrB+PyrI (la sous-unité régulatrice), ce qui regroupe la plupart des Archées et des Eucaryotes ainsi que des Bactéries. Dans le groupe I sont retrouvés essentiellement des Bactéries, de sous-types structuraux A (homomère de PyrB, formant un trimère), B (complexe dodécamérique avec PyrC, voir ci-dessous) ou C (monomères).

J'ai étudié la famille des carbamoyltransférases de manière approfondie dans le Chapitre 7.

1.3 Étape 3 : dihydroorotase

La troisième réaction est la fermeture en cycle de la molécule avec l'expulsion d'une molécule d'eau pour créer un dihydroorotate. L'enzyme est la dihydroorotase ou DHOase (PyrC) qui appartient à la superfamille des amidohydrolase, dont j'ai étudié le mode d'évolution de manière exhaustive et détaillée (voir le Chapitre 6).

1.4 Étape 4 : dihydroorotate déshydrogénase

La quatrième réaction est l'étape oxydante de la voie : elle transforme le dihydroorotate en orotate, qui est un cycle aromatique. Elle est catalysée par la dihydroorotate déshydrogénase ou DHODase (PyrD) et nécessite un accepteur d'électrons. J'ai également étudié le mécanisme d'évolution de cette famille (voir le Chapitre 6).

1.5 Étape 5 : orotate phosphoribosyltransférase

La cinquième étape ajoute à l'orotate le ribose phosphorylé (PRPP) qui forme le nucléotide correspondant, l'OMP (en rejetant un diphosphate). Cette réaction est catalysée par l'orotate phosphoribosyltransférase ou OPRTase (PyrE).

1.6 Étape 6 : OMP décarboxylase

La dernière et sixième étape est la transformation de l'OMP en UMP par une décarboxylation, réalisée grâce à l'enzyme OMP décarboxylase ou OMPDCase (PyrF).

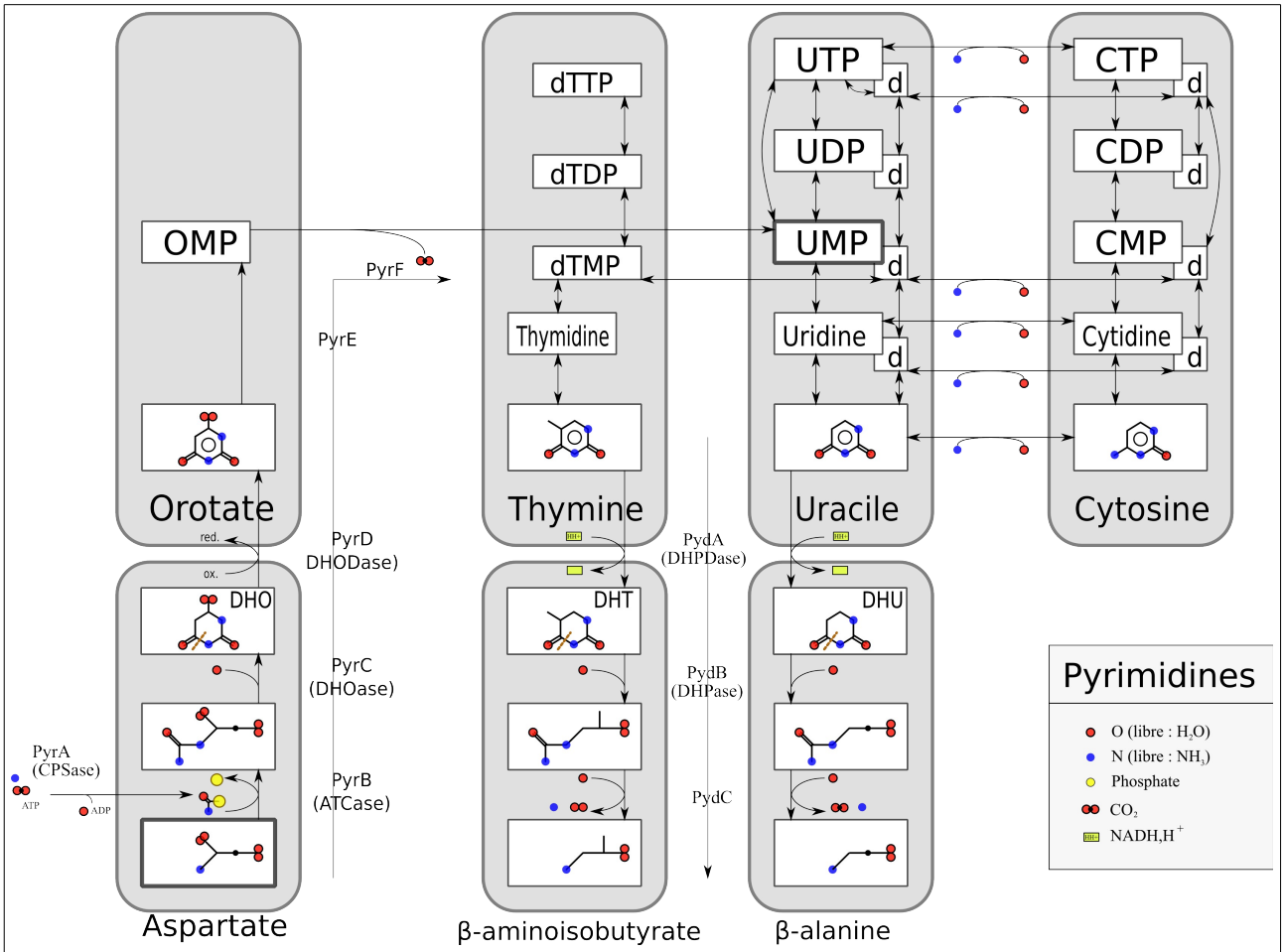


Figure 5 : Représentation de l'ensemble des voies de biosynthèse, de dégradation et d'inter-conversion des pyrimidines.
La voie de biosynthèse (colonne de gauche vers le haut) commence par la création du carbamoyl-phosphate (PyrA) et son ajout à l'aspartate (PyrB) et finit par la décarboxylation de l'OMP en UMP (PyrF).
Les voies de dégradation de la thymine et l'uracile (colonnes 2 et 3 vers le bas) empruntent les mêmes enzymes (PydA, PydB, PydC).

2 Voies de dégradation des pyrimidines

Les pyrimidines peuvent être dégradées par au moins trois voies connues.

La voie réductrice (Figure 5) est présente chez de nombreuses Bactéries et Eucaryotes. Elle dégrade uracile et thymine en des acides β-aminés. Elle utilise trois enzymes (communes aux deux bases) successivement : une dihydropyrimidine déshydrogénase (PydA) qui transforme les deux pyrimidines en dihydropyrimidines ; puis une dihydropyrimidinase (PydB) qui hydrolyse le cycle amide pour donner les acides N-carbamoyl-aminés correspondants. Et enfin une amidohydrolase (PydC) qui hydrolyse le groupement carbamoyl en ammoniac et dioxyde de carbone. J'ai aussi étudié le mécanisme d'évolution de la famille PydA, un des homologues de PyrD (voir le Chapitre

6).

La seconde voie, identifiée chez quelques Bactéries, est la voie oxydante [Hayaishi & Konrberg 1952]. L'uracile est d'abord oxydé en barbiturate, puis dégradé en urée et en malonate par une amidohydrolase et une hydrolase.

Une dernière voie utilisant les produits des gènes *rut* de *Escherichia coli* a récemment été mise en évidence [Parales & Ingraham 2010]. Elle dégrade l'uracile en 3-hydroxy-propionate en relâchant de l'ammoniac et du dioxyde de carbone.

3 Voie de dégradation des purines

La voie de dégradation des purines, présente chez de nombreuses espèces, permet de les utiliser comme source d'azote (Figure 6). Nous avons été amenés à étudier cette voie parce qu'elle contient plusieurs étapes dont les mécanismes réactionnels et les couples substrat/produit sont étonnamment similaires à ceux utilisés dans le métabolisme des pyrimidines (voir Chapitre 6 et Chapitre 7).

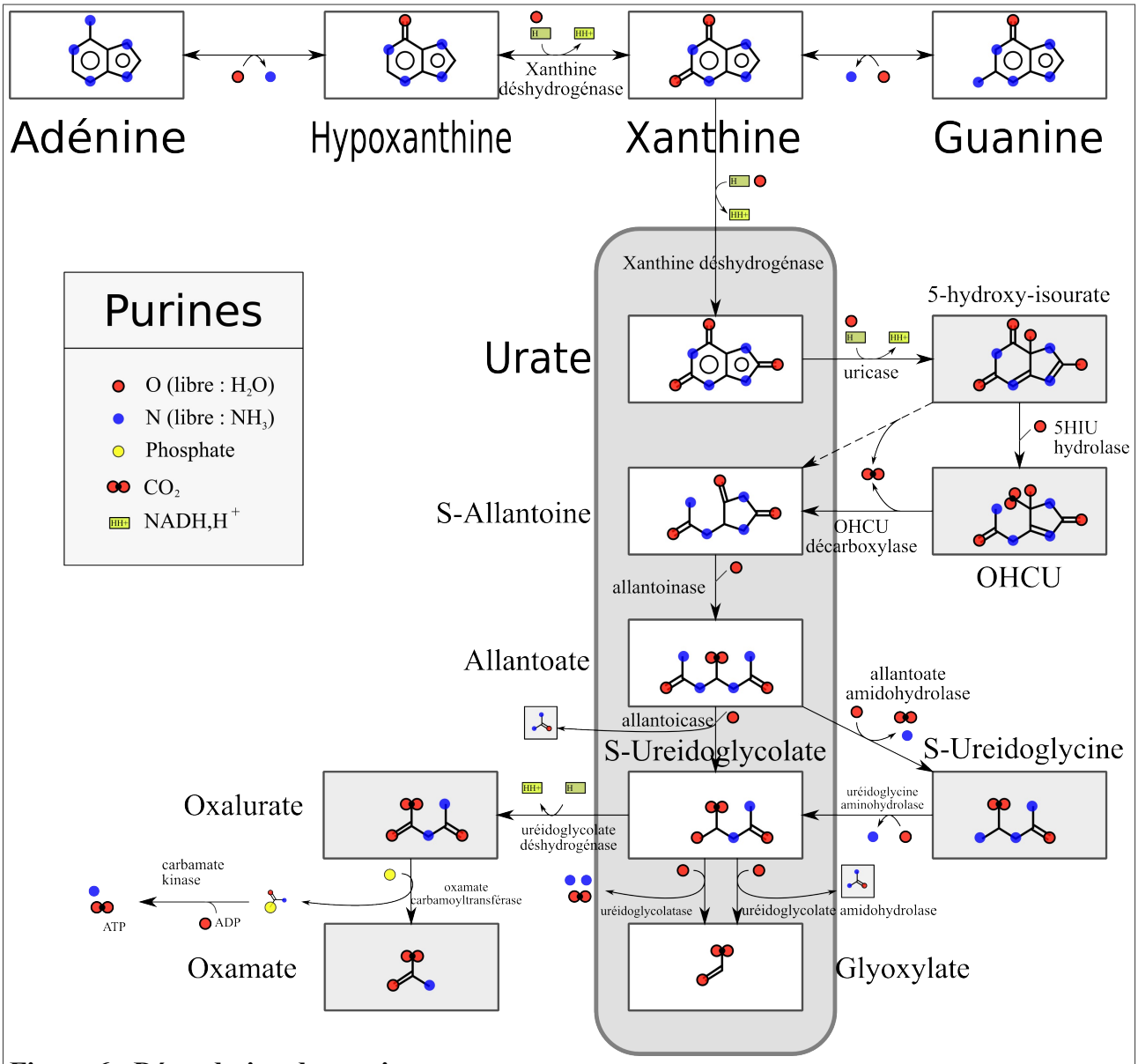


Figure 6 : Dégradation des purines.

Les purines (adénine et guanine) sont transformées en urate, qui est ensuite dégradée via l'allantoïne selon différentes voies selon les organismes, en rejetant de l'urée ou de l'ammoniac selon les voies.

Cette dégradation commence par la transformation des diverses purines en urate, en passant soit par la xanthine (dégradation de la guanine/guanosine en xanthine/xanthosine), soit par l'hypoxanthine (dégradation de l'adénine/adénosine en xanthine/inosine). La xanthine déshydrogénase catalyse à la fois la conversion de l'hypoxanthine en xanthine et de la xanthine vers l'urate. L'urate est ensuite converti en allantoïne avec trois réactions successives : uricase, 5-hydroxyisourate hydrolase et OHCU décarboxylase. Le 5-hydroxyisourate se décompose

spontanément en allantoïne, mais en proportions racémiques (R et S), alors que les deux dernières réactions permettent de créer de la S-allantoïne spécifiquement. Certaines espèces contiennent également une allantoïne racémase (EC 5.1.99.3) qui permet de convertir R-allantoïne et S-allantoïne.

La S-allantoïne est ensuite converti en allantoate par l'allantoïnase (EC 3.5.2.5) qui ouvre le cycle amide restant par hydrolyse.

L'allantoate est ensuite dégradée en uréidoglycolate. Cela peut être réalisé par deux voies, en libérant soit une molécule d'urée, soit de l'ammoniac (2 NH_3 et 1 CO_2). Dans le premier cas, l'allantoate est directement transformé en uréidoglycolate par une allantoïcase en rejetant de l'urée. Dans le second cas, l'allantoate et d'abord hydrolysée par une allantoate amidohydrolase en S-uréidoglycine (libération de 1 NH_3 + 1 CO_2), puis hydrolysée par l'uréidoglycine aminohydrolase en uréidoglycolate, en libérant 1 NH_3. L'uréidoglycine est instable et se dégrade spontanément en urée, ammoniac et glyoxylate, mais le processus est assez lent pour nécessiter une enzyme [Serventi et al. 2010] la dégradant en uréidoglycolate. L'uréidoglycolate est ensuite convertie en glyoxylate, là encore de deux manières possibles, soit en libérant une molécule d'urée (uréidoglycolatase) ou 2 NH_3 et 1 CO_2 (uréidoglycolate amidohydrolase). De l'allantoate au glyoxylate, 4 enchaînements sont possibles (libération d'urée, d'ammoniac, ou les deux). Les différentes espèces peuvent utiliser un de ces quatre chemins pour dégrader les purines.

il existe cependant d'autres voies en conditions anaérobies qui passent par l'oxalurate, soit en oxydant l'uréidoglycolate, soit en utilisant l'uréidoglycine avec du glyoxylate pour donner de l'oxalurate et de la glycine [Ramazzina et al. 2010].

L'oxalurate est hydrolysé en oxamate en rejetant un carbamoyl-phosphate qui sera lui-même dégradé par une carbamate kinase. Dans cette dernière voie, le gène codant l'oxamate carbamoyltransférase (OxTCase) n'a à ce jour toujours pas été découvert (voir les chapitres 6 et 7).

Chapitre 3 : Évolution des enzymes des voies métaboliques

Les voies étudiées, tout comme la plupart des voies métaboliques, sont présentes chez la plupart des êtres vivants actuels, ce qui laisse supposer qu'elles dériveraient d'un petit nombre de voies ancestrales. En comparant les différentes voies métaboliques contemporaines, il est possible de faire remonter l'origine de nombreuses voies essentielles à la période précédant l'apparition de LUCA, acronyme anglais pour désigner l'organisme hypothétique qui était le dernier ancêtre commun et universel à tous les êtres vivants contemporains connus (voir Woese [1998] et le numéro spécial introduit par Lazcano & Forterre [1999]), et d'étudier comment ces voies ont été altérées et remaniées au cours du temps pour comprendre leur variété [Kyrpides, Overbeek & Ouzounis 1999 ; Labedan et al. 1999].

1 Apparition des voies métaboliques

Les premières voies métaboliques seraient apparues quand les premiers organismes vivaient dans une « soupe primordiale ». Selon cette hypothèse formulée indépendamment dans les années 1920 par Oparin [Pennazio 2009] et Haldane [1954], l'environnement prébiotique contenait des matières organiques en abondance en raison de la nature réductrice de l'atmosphère terrestre. L'expérience de Miller & Urey [1959] montra que l'on pouvait obtenir expérimentalement la création d'acides aminés et beaucoup d'autres composés organiques simples en faisant interagir les éléments minéraux et chimiques élémentaires supposés être présents dans l'atmosphère de la jeune Terre il y a quatre milliards d'années. La synthèse spontanée de purines et pyrimidines fut ultérieurement démontrée par le groupe de Miller, confirmant la validité de l'hypothèse d'une riche soupe primordiale.

Plusieurs modèles ont été proposés par la suite pour expliquer la formation des voies métaboliques à partir de ces éléments primordiaux. Le modèle rétrograde anabolique de Horowitz [1945] explique l'apparition des premières voies de biosynthèse en supposant que ces voies se sont formées en partant de la fin de la chaîne. Quand un composant essentiel vient à manquer dans le milieu extérieur, les réserves naturelles ayant été épuisées, l'avantage évolutif revient aux organismes capables d'obtenir ces composants manquants à partir d'autres plus répandus. Le processus se répète si ces autres composants se raréfient. Dans ce modèle, la chaîne métabolique se

construit à partir de la fin, et il nécessite que tous les composants finaux et intermédiaires aient été présents naturellement à un moment donné.

Un autre modèle proposé par Granick [1957] propose la création de voies dans un sens opposé : une première étape permet de créer un composant utile à l'organisme. Vient ensuite une seconde étape qui forme un second composé utile à partir du premier, et ainsi de suite. Ce modèle ne requiert pas de composés prébiotiques préexistants, mais cela suppose que les éléments intermédiaires sont d'une certaine utilité, comme c'est le cas pour la création de la chlorophylle (travail de Granick) ou des hèmes, mais pas pour les voies du type purine comme discuté par Lazcano & Miller [1996]. On rencontre également un modèle prograde catabolique par Cordón [1990] qui est l'inverse du modèle rétrograde anabolique (Horowitz) mais pour les voies de dégradation.

2 Le recrutement des enzymes dans les voies métaboliques

Mais d'où viennent alors les enzymes qui sont recrutées dans ces nouvelles voies ? Une première réponse fut formalisée par Ohno [1970] qui proposa qu'un gène se dupliquant donne deux copies identiques qui divergent par la suite en accumulant des mutations. Si l'une conserve la fonction originelle du gène, l'autre peut être éliminée ou – plus rarement – acquérir une nouvelle fonction (néofonctionnalisation). Des modèles plus développés ont été proposés ensuite, comme le modèle DDC (duplication, délétion, complémentation) de Force et al. [1999] qui distingue trois destins d'un duplicat : une perte rapide (le plus fréquent), une néofonctionnalisation (Ohno), ou une sous-fonctionnalisation. Cette dernière considère que les deux copies se partagent les différentes parties de l'enzyme originelle, en ne gardant chacune que quelques éléments essentiels que l'autre perd, résultant en la nécessaire co-conservation (complémentation) des deux copies. Contrairement à Ohno, les contraintes sont alors immédiates.

Cependant, le modèle de Ohno repose fondamentalement sur l'idée qu'à une enzyme correspond une seule réaction sur un substrat donné dans une certaine voie métabolique (un gène = une fonction), ce qui fut remis en cause par le modèle développé par Ycas [1974] et étendu par Jensen [1976], dans son modèle d'évolution en *patchwork*. Ce modèle considère des voies métaboliques construites en empruntant des enzymes d'autres voies dont l'activité est modifiée par des mutations ponctuelles. L'idée est que les enzymes ancestrales, peu nombreuses, possédaient une activité large, capables d'agir sur des substrats différents. Ces enzymes à large spectre peuvent être

qualifiées d'enzymes plastiques, d'enzymes ambigües, voire d'enzymes « travaillant au noir ». Les anglo-saxons parlent de *promiscuous enzymes* (« enzymes indiscriminantes »). En effet, contrairement au modèle classique (« une clé – une serrure ») de l'enzyme vue comme une molécule hyper spécialisée qui est conservée dans un organisme pour effectuer une fonction bien précise (dans une voie spécifique par exemple), il est maintenant admis que cette enzyme puisse catalyser des réactions chimiques similaires sur d'autres substrats ressemblants, mais dans une plus faible mesure (d'où le travail « au noir »). On parle maintenant d'ambigüité de substrat (Jensen) par opposition au cas beaucoup plus rare où l'enzyme est capable d'effectuer deux types de réaction chimique différentes, comme proposé initialement dans le modèle d'Horowitz [1945].

3 L'histoire évolutive des enzymes et de leurs fonctions

Dès lors, l'histoire évolutive des différents gènes codant des enzymes contemporaines peut nous renseigner sur l'origine des voies qui les utilisent, en remontant jusqu'aux hypothétiques voies ancestrales. Pour ce faire, la manière dont les enzymes ont évolué doit être prise en compte. Cette histoire permet de décrire les relations des différents homologues, c'est-à-dire entre les séquences descendant d'une même séquence ancestrale.

Ces homologues peuvent être distingués selon leurs relations décrites par Fitch [1970] : les orthologues tout d'abord (Figure 7.1), qui sont séparés par un événement de spéciation. Les orthologues descendent donc d'une séquence ancestrale verticalement. L'autre relation définie par Fitch est la paralogie entre deux séquences (Figure 7.2), qui sont séparées par un événement de duplication de gène ancestral selon le modèle proposé la même année par Ohno [1970]. Depuis l'irruption des données génomiques, des descriptions plus fines ont été proposées : la distinction en in-paralogues et out-paralogues [Sonnhammer & Koonin 2002], les premiers (in) étant des paralogues issus d'une duplication récente après un événement de spéciation, résultant en la présence des deux copies proches dans un organisme (Figure 7.3). Les seconds (out) sont au contraire issus d'une duplication plus ancienne précédent la dernière spéciation, ce qui résulte en la présence potentielle des deux copies dans les espèces descendant de cette spéciation (Figure 7.4). Par ailleurs, la mise en évidence de transferts horizontaux entre espèces (notamment Bactéries et Archées, ou dans les cas d'endosymbiose) permet de décrire un autre type de relation, celui de xénologues [Gray & Fitch 1983], dans lequel une des séquences homologues comparées provient d'une espèce plus éloignée (Figure 7.5).

L'exactitude de ces différentes relations ne peuvent être déterminées expérimentalement. On ne peut que les reconstruire avec les données contemporaines disponibles au moment de l'étude. On peut ainsi créer des arbres phylogénétiques qui représentent ces relations en comparant les séquences entre elles.

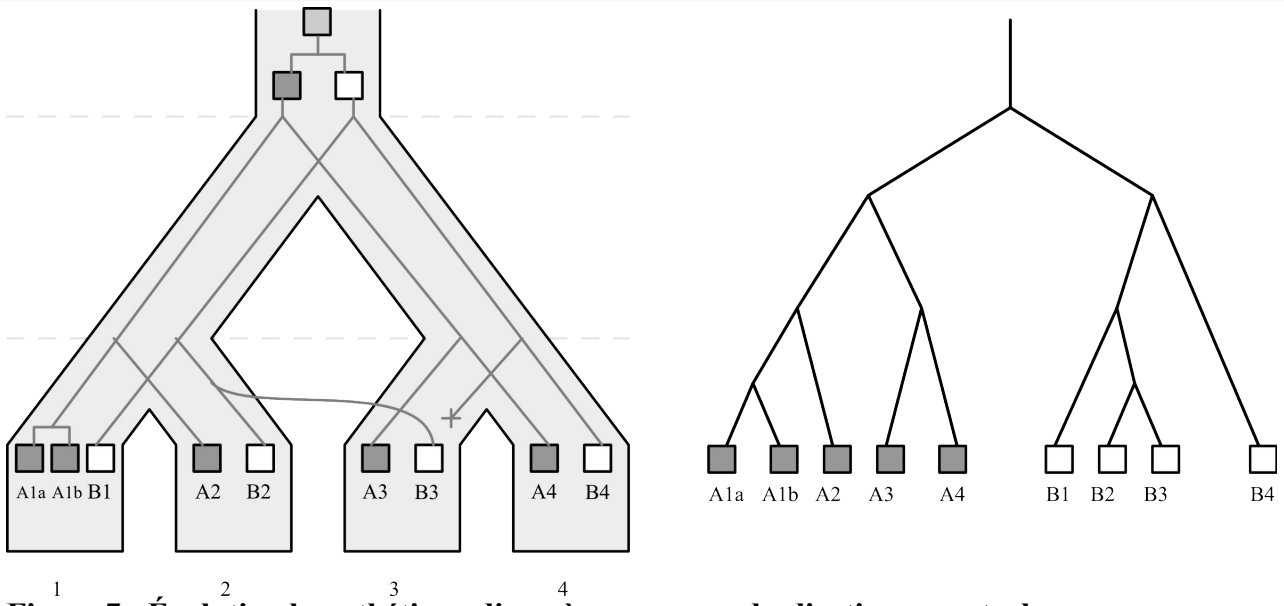


Figure 7 : Évolution hypothétique d'un gène avec une duplication ancestrale.
L'arbre de gauche représente l'évolution du contenu des génomes des différentes espèces au gré des duplications (bifurcations à angle droit) et des évènements de spéciation (lignes en tirets). Les cadres en bas représentent le contenu actuel constaté des génomes. L'arbre de droite représente la même histoire mais telle qu'on la reconstitue avec un arbre phylogénétique (enraciné). 4 espèces sont représentées (1,2,3,4,5) ainsi que deux paralogues ancestraux A et B issus d'une duplication ancestrale.
On observe les relations d'homologie suivantes, avec exemple (la relation est représentée par un tiret) :

1. ***orthologues** : A3-A4 ou A2-A3 [issus verticalement d'un gène ancestral]*
2. ***paralogues** : (chaque A)-(chaque B) [produits d'un événement de duplication ancestral]*
3. *in-paralogues : A1a-A1b [duplication ultérieure à la dernière spéciation de l'espèce 1]*
4. *out-paralogues : A3-B3 [duplication avant tout événement de spéciation]*
5. *xénologue : B3-(autres B) [transfert horizontal du gène B de l'espèce 3]*

4 Conséquences sur l'annotation fonctionnelle

Cette histoire évolutive compliquée rend particulièrement difficile l'annotation fonctionnelle de ces séquences homologues : l'attribution d'une fonction à la protéine codée par un gène donné est un processus complexe. Il faut en particulier différencier le niveau *moléculaire* (quelle est la réaction catalysée, sur quel substrat ?) du niveau *cellulaire*, plus global, qui permet par exemple de replacer une enzyme dans une voie métabolique définie. Mais, comme on l'a vu, les enzymes (en

particulier dans des superfamilles) ont des histoires complexes et la détermination des fonctions doit nécessairement prendre en compte cette phylogenèse, ce qui n'est pas la méthode la plus répandue, en particulier lors de l'annotation automatique pratiquée par les grandes bases de données publiques.

4.1 De nombreuses séquences à annoter

À cause de l'afflux phénoménal de nouvelles séquences disponibles provenant du séquençage de nombreux organismes (dont la biologie est le plus souvent complètement inconnue), favorisé par l'innovation des techniques de séquençage (de plus en plus rapides et de moins en moins coûteuses), on se retrouve à présent avec des millions de séquences à annoter. Même dans une seule famille de gènes homologues, le nombre de séquences se compte désormais souvent en milliers. Seuls quelques organismes modèles, comme la bactérie *E. coli*, la levure *Saccharomyces cerevisiae* ou la mouche *Drosophila melanogaster* (pour ne citer qu'eux) ont été assez étudiés pour disposer d'une bonne quantité de données expérimentales, notamment enzymatiques. Néanmoins, même pour ces organismes dont l'étude a mobilisé des milliers de chercheurs pendant les soixante dernières années, il reste encore beaucoup d'activités déterminées expérimentalement pour lesquelles on n'a pas identifié le gène codant correspondant (activités enzymatiques orphelines, voir Hanson et al. [2010] ; Lespinet & Labedan [2005] ; Lespinet & Labedan [2006] ; Pouliot & Karp [2007]), ou inversement de produits de gènes pour lesquels on n'a pas mis en évidence l'activité enzymatique (orphelins métaboliques, voir Chen & Vitkup [2006]).

Afin d'annoter toutes ces séquences, on est donc obligé d'utiliser des approches automatiques avec des outils bioinformatiques. La méthode la plus généralement utilisée est un transfert d'annotations basé sur la similarité entre deux séquences que l'on suppose être homologues, en utilisant un logiciel comme BLAST [Altschul et al. 1990] qui compare une séquence à tester à un grand jeu de séquences et en extrait les plus ressemblantes. L'annotation de la séquence ayant le meilleur score est alors transférée à la séquence test. L'idée est que des séquences proches sont orthologues et auraient donc une même fonction [Koonin 2005]. Malheureusement, cette approche est trop approximative pour l'annotation fonctionnelle et conduit à la création et à la propagation d'erreurs, car le meilleur score BLAST n'est pas nécessairement le meilleur voisin phylogénétique [Koski & Golding 2001]. C'est ainsi que dans les banques de données GenBank [Benson et al. 2011] ou TrEMBL [UniProt_Consortium 2011] (annotées automatiquement), on estime que l'annotation fonctionnelle est erronée dans 5% à 63% des cas, par opposition aux banques de données annotées manuellement, comme Swissprot [UniProt_Consortium 2011], et dont le taux d'erreur ne dépasserait pas 1% [Schnoes et al. 2009]. Dans la plupart des cas, les erreurs

d'annotation seraient même des sur-prédictions, c'est-à-dire qu'on prédit une fonction précise alors que les données ne permettent de préciser qu'une fonction plus générale, par exemple une fonction amidohydrolase.

4.2 Approches pour l'annotation de superfamilles de protéines

Dans le cas des superfamilles, l'annotation fonctionnelle « par homologie » n'est donc pas du tout adaptée et des outils plus évolués sont nécessaires pour annoter de manière précise les séquences. Ces outils doivent permettent de prendre en compte les structures conservées entre les séquences d'une même famille, la composition génétique des génomes, et bien sûr la phylogénie de la famille, en se basant sur les quelques informations de qualité disponibles (banques de données annotées manuellement et expériences biochimiques publiées dans la littérature scientifique).

4.2.1 Motifs et profils de séquences

Les enzymes possèdent des résidus extrêmement conservés, par exemple ceux qui constituent leur site actif. En comparant les séquences de plusieurs familles, il est alors possible de déterminer quels sont ces résidus et pour quelle famille ils sont spécifiques. Il existe bien des banques de données de motifs conservés, comme PROSITE [Hulo et al. 2006] mais elles s'avèrent trop limitées pour l'étude des superfamilles. De même, il existe des profils de séquence utilisant un modèle de Markov caché (HMM), comme dans la banque Pfam [Finn et al. 2010]. Ces profils sont basés sur des alignements multiples créés automatiquement et représentent l'ensemble des séquences alignées, mais elles sont difficilement utilisables pour l'annotation fine de superfamilles (par exemple dans le cas des amidohydrolases, cf Chapitre 6, il n'y a qu'un profil pour toute la superfamille, qui contient également de lointains parents comme les uréases).

Pour l'étude d'une superfamille, il est donc nécessaire de recenser et de compiler les résidus les plus importants en utilisant des données structurales, notamment les éventuelles structures 3D connues pour reconnaître les résidus structurellement et/ou fonctionnellement conservés. L'utilisation d'un alignement multiple de séquences permet ensuite de comparer la conservation de ces résidus chez toutes les séquences de la superfamille et d'isoler les résidus les plus discriminants permettant de distinguer toutes les familles et d'annoter plus finement les différentes séquences. Un alignement multiple de qualité est donc indispensable à cette étape (voir les prochains chapitres).

4.2.2 Voisinage génétique et profils phylogénétiques

Le contexte génétique d'un gène peut également renseigner sur sa fonction cellulaire s'il est

présent avec d'autres gènes d'une voie connue [Overbeek et al. 1999]. Chez les Bactéries et les Archées, les gènes correspondants à certaines voies sont souvent regroupés au même endroit d'un génome, généralement sous la forme d'opérons, ce qui permet une expression et une régulation homogène de toutes les enzymes de la voie en question. Lorsque ces regroupements sont conservés (au moins deux séquences à la suite ou dans la même groupe), on parle de synténie (Figure 8). La recherche et la mise en évidence de ces groupes de synténie peut permettre de proposer une voie commune, qu'elle soit connue ou pas. Cette approche est donc intéressante pour annoter des gènes qui codent des protéines de fonction même inconnue (par exemple des orphelins métaboliques).

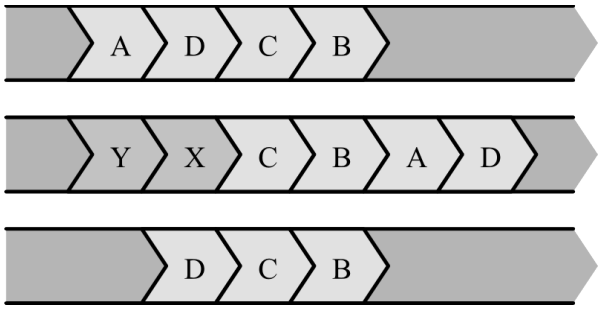


Figure 8 : Contexte génétique hypothétique de 4 gènes A, B, C, D.
Les trois lignes représentent une partie du génome de trois espèces arbitraires différentes. Dans les trois espèces, on retrouve les gènes D, C et B. Le premier et le second génome ont également le A dans le même contexte, alors qu'il est perdu (peut-être ailleurs dans le génome dans la troisième espèce). Les gènes B, C, D sont donc en synténie entre les trois génomes, et les quatre gènes A, B, C, D sont en synténie entre les deux premiers.

Une autre approche plus générale s'affranchit de la proximité des gènes dans le génome en considérant l'ensemble des séquences dans les familles considérées (avec donc un *a priori* sur ces séquences). La co-occurrence de plusieurs gènes dans un même génome permet de donner des informations fonctionnelles sur les produits de ces gènes si ceux-ci sont connus pour être dépendants l'un de l'autre (typiquement quand leurs produits forment un complexe, ou plus généralement s'ils sont présents dans une même voie).

4.2.3 Phylogénie et alignement multiple

Une dernière approche, certainement la plus pertinente mais aussi la plus complexe à réaliser, est bien sûr la reconstruction de la phylogénie des séquences d'une superfamille. Cette approche nécessite néanmoins la construction d'alignements multiples de séquences, qui est un problème en soit (voir le prochain Chapitre 4, p 38).

L'annotation des séquences par phylogénie exploite l'idée d'orthologie et le principe de

monophylie : si des séquences sont orthologues et qu'une même annotation est présente dans plusieurs branches d'un même sous-arbre, alors on transfère l'annotation à toutes les séquences de ce sous-arbre [Eisen 1998], en émettant l'hypothèse que toutes ces séquences partagent un même ancêtre ayant possédé la fonction propagée par spéciations successives (Figure 9).

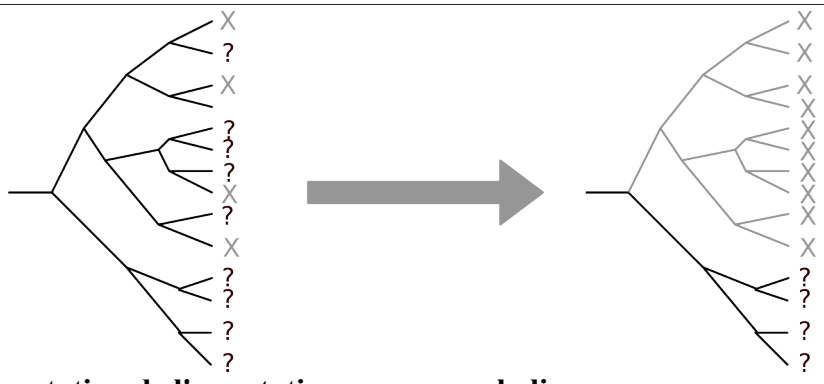


Figure 9 : Représentation de l'annotation par monophylie.
L'arbre de gauche initial contient quatre séquences annotées X réparties dans l'arbre. Si on considère que l'ancêtre de toutes ces séquences était également X, alors on peut propager cette annotation dans l'arbre de droite à tous les descendants de l'ancêtre commun aux séquences initiales.L'arbre doit être enraciné (ici à gauche).

Outre l'annotation des différents groupes d'orthologues partageant une même fonction, un arbre phylogénétique permet aussi potentiellement de mettre en évidence des groupes de paralogues qui ne correspondent à aucune activité connue. En effet, les différents groupes de paralogues témoignent de duplications de gènes ancestraux. La divergence entre les différents paralogues et leur conservation éventuelle dans un même génome est un indice important de l'adoption d'une nouvelle fonction de ces copies.

MÉTHODOLOGIE

Chapitre 4 : Frali, un programme d'alignement de séquence progressif

Les familles des différentes enzymes des voies métaboliques étudiées sont constituées de séquences à l'histoire évolutive complexe, dont on suppose qu'elles sont homologues, et que l'on retrouve dans la plupart des espèces. On est donc amené à comparer des séquences parfois très divergentes, ce qui nécessite l'usage d'outils adéquats pour les aligner de manière exacte, c'est-à-dire en retrouvant l'homologie de site. Un alignement multiple biologiquement pertinent est celui où chaque colonne superpose les résidus qui ont évolué à partir d'un ancêtre commun par simple substitutions au niveau de la position ancestrale [Descorps-Declère et al. 2008 ; Patterson 1988].

1 Problématique des alignements multiples

La mise en place d'un alignement multiple de séquence est donc une tâche complexe et d'autant plus difficile à réaliser que les séquences à aligner sont différentes et nombreuses [Kemena & Notredame 2009]. Un biologiste étudiant une famille de protéines utilise généralement des outils automatiques d'alignement comme ClustalW [Thompson, Higgins & Gibson 1994], Muscle [Edgar 2004], T-Coffee [Notredame, Higgins & Heringa 2000] et beaucoup d'autres (voir les revues récentes de Kemena & Notredame [2009] ; Pei [2008] ; Pirovano & Heringa [2008]). Cependant, une vérification manuelle complémentaire et généralement nécessaire pour s'assurer de la bonne qualité de l'alignement obtenu [Edgar & Batzoglou 2006], ce qui devient très contraignant dès lors que de nouvelles séquences paraissent régulièrement en nombre croissant : tout le travail de réalignement complet et de vérification et correction est à refaire.

2 Principes de Frali

J'ai donc conçu une méthode qui ajoute les nouvelles séquences candidates (parce que trouvées homologues) à un alignement préexistant de qualité et que j'ai implémenté sous la forme d'un programme appelé Frali (pour *frame alignment*, soit « alignement cadre »).

Frali nécessite deux étapes : la première est la conception de l'alignement de base, la graine. Cet alignement de base est fourni par l'utilisateur et est basé sur un petit nombre de séquences représentatives de l'ensemble des séquences de la (super)famille étudiée et pour lesquelles on

dispose – si possible – de leur structure tridimensionnelle. L'utilisateur peut alors utiliser des logiciels d'alignement structuraux (très efficaces mais qui ne peut fonctionner raisonnablement qu'avec un petit nombre de séquences) comme Expresso [Armougom et al. 2006] pour obtenir un bon alignement qu'une vérification manuelle vient parachever. La seconde étape utilise cet alignement qualifié de *graine* comme d'un cadre pour l'ajout de toutes les nouvelles séquences. Les nouvelles séquences sont ainsi ajoutées progressivement à la graine en utilisant les séquences les plus proches de cette graine pour s'y aligner correctement.

3 Publication

Les détails de la méthode implémentée dans Frali ont été publiés dans le papier (en anglais) qui suit :

- Barba M, Lespinet O & Labedan B (**2010**). *Fast and accurate multiple sequence alignment of large and diversified sets of distant homologues.* Actes de Jobim 2010: 81:88.

Fast and accurate multiple sequence alignment of large and diversified sets of distant homologues

Matthieu BARBA, Olivier LESPINET and Bernard LABEDAN

Institut de Génétique et Microbiologie, UMR8621 CNRS,
Université Paris-Sud XI, Bâtiment 400, 91405 Orsay Cedex, France
{matthieu.barba, olivier.lespinet, bernard.labedan}@igmors.u-psud.fr

Abstract *Frali allows delivering an accurate and biologically relevant multiple sequence alignment (MSA) of large and heterogeneous families comprising remote homologues. First, an expert alignment of well-studied representatives of each subfamily is built semi-manually to define a seed alignment that represents the frame of the whole family. Then; the targeted addition of the rest of the parental sequences to this frame is processed after being sampled according to their degree of relatedness to their homologues prealigned in the frame. These new sequences are further clustered before aligning them to this frame using a hidden Markov model based profile-profile approach. This process allows keeping the accuracy gained at the step of building the seed alignment as checked both by benchmarking and by studying a family of distant homologous enzymes involved in various biological functions. Interestingly, this approach further allows a rapid update of a reference MSA as soon as new homologues appear.*

Keywords multiple alignment, remote homologues, HMM profile.

Aligner rapidement et exactement de grands jeux d'homologues distants

Résumé *Pour obtenir un alignement multiple exact et biologiquement valide de séquences homologues distantes appartenant à des grandes familles hétérogènes, une graine formée des représentants caractéristiques de chaque sous-famille est construite pour représenter l'architecture de la famille. Puis, le reste des séquences homologues à cette graine est ajouté progressivement de façon complètement automatisée par une approche profil-profil de modèles cachés de Markov. Cette approche permet de maintenir la qualité optimale de la-graine et (cerise sur le gâteau) de mettre à jour automatiquement à tout moment l'alignement de référence.*

Mots-clés *alignement multiple, homologues distants, profil HMM.*

1 Introduction

Many biologists consistently use completely automatic tools to generate multiple sequence alignment (MSA) without considering their potential flaws. In fact, although many algorithms are now available [1,2,3], constructing a MSA is not a trivial task [4]. Since defining homology is always a hypothesis, only empirical approaches are suitable. Hence, as already underlined [3], MSA are not plain data but models. Therefore, manual construction still remains more appropriate than automated one to get biologically relevant MSA [4], and if an automatic approach is used, a manual check is obligatory to improve the obtained output. It becomes increasingly difficult to meet these requirements as the number of potential homologues increases vertiginously.

Presently, families containing several thousands of homologues have become common, making it mandatory to use a limited number of automated tools, such as Muscle [5], while rendering difficult the required manual check of the output. Moreover, computing such alignments of large sets of sequences in a reasonable time implies a concomitant loss in correctness. Indeed, Kemena and Notredame [3] showed that the present MSA methods lose their accuracy when the number of sequences to multiply align is >100.

The challenge of building accurate MSA becomes even harder when dealing with distantly related homologous proteins. This often occurs in large and diversified families where subfamilies may be very distant from each other, their amino acid sequences sharing very low percentage of sequence identity as exemplified in the test case described below (§ 4.2).

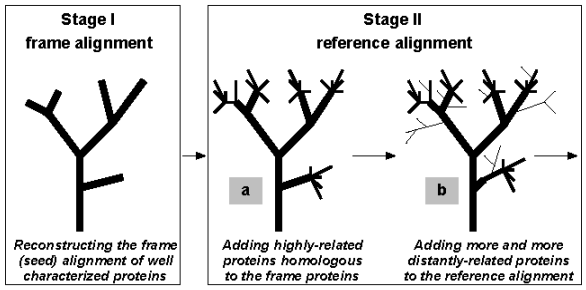


Fig. 1. Stepwise automated generation of an accurate reference alignment of a large and assorted family (stage II) using as a template a biologically relevant frame alignment built semi-manually (stage I).

To cope with these technical limitations, we propose a 2-stage approach based on (i) the construction of a high quality seed alignment, using a small, selected, set of sequences, and (ii) the progressive and targeted addition of the rest of the homologues to this seed (Fig. 1). By definition, a seed alignment would be optimal where each site corresponds to homologous positions, i.e. if each column contains the amino acids believed to have evolved from a common ancestor only through character substitutions [6,7]. For this reason, we call such a biologically relevant seed alignment a *frame alignment,* by analogy with the skeleton of an evolutionary tree, assuming that the topology of its deepest branches is already well defined since residues in each column are supposed to be consistently and correctly aligned (Fig. 1, stage I). In the second, entirely automated, step, all the remaining homologous sequences are sampled along a decreasing gradient of evolutionary distances and further clustered in order to be added selectively and stepwise, using the closest sequences present in this frame alignment as a template at each step. To continue the tree analogy, building such a *reference alignment* with our entirely automated tool, *Frali* corresponds to the gradual addition of more recent twigs and leafs mainly on the existing deep branches (Fig. 1, stage II). Moreover, securing the first stage allows automation of the process of continuously updating the reference MSA when newly published genomes become accessible, while keeping permanent its accuracy and biological relevance.

2 Methodology

2.1 Stage I: Building the seed (frame alignment)

We regard as representative sequences the few proteins that have been experimentally studied and thus are supposed to be correctly annotated. Optimally, at least one representative sequence of each distant subfamily must be included. Moreover, to assess alignment, we preferentially chose experimentally studied proteins that have been crystallized. Accordingly, the primary amino acid sequences were multiply aligned using Expresso [8]. Whenever the number of sequences with known structures was too low, and/or some subfamilies lacked 3D structure data, we used PSI-Coffee since this is ranked as the most accurate program immediately after 3D structure–based algorithms [3].

Although those automated methods are generally efficient, we always had to review manually the frame alignment obtained so that errors – such as the introduction of indels in structural data –could be avoided. This manual check was made by visualizing the aligned 3D structures using ad hoc tools [9].

2.2 Stage II: From the frame to the reference alignment

Once an optimal seed alignment has been obtained, the remaining homologous sequences can be automatically added to the produced frame to build a reference alignment (Fig. 1, Stage II) using Frali. To maintain a high level of accuracy during the whole process, the addition is made stepwise, as summarized in Fig. 2: clustering the homologues, matching them with their closest partners in the frame alignment, and aligning their hidden Markov model (HMM) based profiles [10].

2.2.1 Preparing a high-confidence reference alignment. To facilitate their targeted addition to the reference alignment, we first clustered the homologues sharing >70% sequence identity over >70% of the length of the shorter matching sequence using the fast and alignment-free CD-hit program [11]. In parallel, their closest homologues prealigned in the frame were likewise clustered. Each cluster was processed through the steps given in Fig. 2:

2.2.1.1 Detecting matching clusters. Since the sequences belonging to such clusters are very close by construction, one of them should reasonably be sufficient to search for matching sequences in the corresponding set of prealigned sequences in the frame. Indeed, although the number of new sequences would seem to be huge, a large fraction of them are actually the n^{th} near identical copy of the same sequence, since they are encoded by different strains of the same species or closely related species. Consequently, instead of comparing each sequence of each cluster to every reference alignment

sequence, we defined one sequence representing each kind of clusters. Thus, the number of representatives increases far more slowly than the raw number of sequences. Representative sequences of each cluster were selected as the longest sequence to avoid accidentally using fragments.

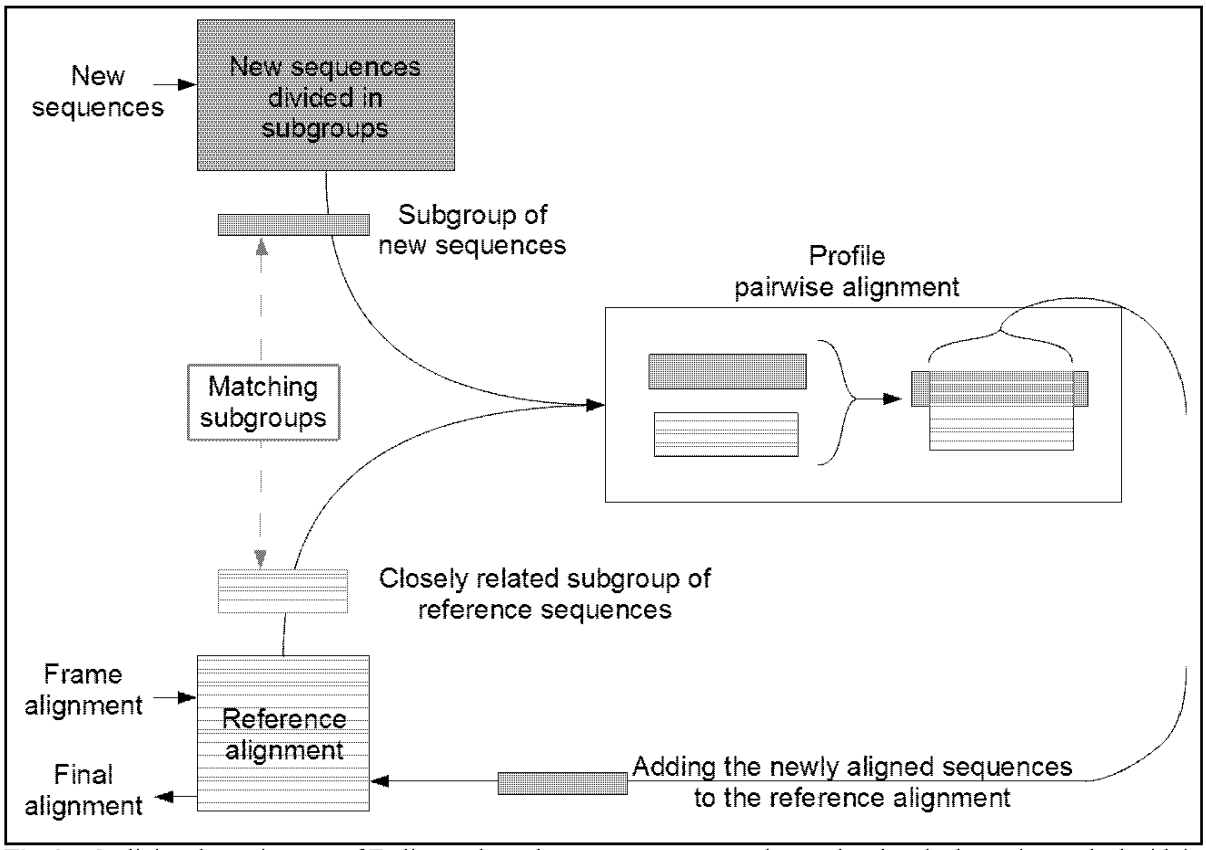


Fig. 2. Outlining the main steps of Frali: new homologous sequences are clustered and each cluster is matched with its closely related subgroup present in the frame alignment by aligning their HMM profiles, allowing their facilitated addition to the reference alignment.

2.2.1.2 Aligning matching clusters. First, the matching frame alignment was stripped of its empty columns before building the HMM profile, and adding them back into the final reference alignment. The new homologues are aligned using Muscle [5], one of the very few methods able to handle large datasets in reasonable times [2]. Although the intrinsic performance of heuristic methods like Muscle is not optimal (discussed in [3]), we ascertained that the elevated level of identity of these highly-related sequences ensures the biological relevance of the obtained MSA. Once matching clusters have been detected using their representative sequences, all their respective sequences were aligned to the prealigned closest frame sequences. Such a progressive addition by subgroups is a crucial trait in our approach. Indeed, it precludes the probable blurring of features specific to each subgroup, such as conserved residues or specific indels (not shown) that would appear in the case of a unique addition of many highly divergent proteins. Noticeably, aligning such groups of closely related homologues allows the further generation of accurate HMM profiles for both the cluster of new sequences under study (HMM_cluster) and the associated cluster of the frame (HMM_frame). The two HMM profiles are then fused using the HHalign program [10]. After addition of each cluster, 2 important points are examined by Frali. (i) Frali extracts selectively the part of HMM_cluster aligning with the HMM_frame by excluding any sequence element located before or after the aligned fraction so as to maximize the efficiency of the HHalign step. This is crucial in discarding the unalignable part (columns absent in the frame) and automatically outlining the homologous segment present in fused proteins. (ii) Frali prevents the misalignment of sequences that are too divergent from the template sequences. Noisy profiles are precluded by impeding the

addition of too distant sequences that will introduce holes of >30 residues. Note that such a safety device does not restrain the addition of sparse natural indels in newly added sequences, since these gaps could be precious phylogenetic markers [27]. Thus, this clusterization step is clearly maintaining the biological relevance when progressively enlarging the frame alignment to the reference alignment, while automated tools would locally damage this relevance (not shown).

2.2.1.3 Improving the HMM alignment and reiterating the whole process for the other clusters. The profile-profile alignment is improved by keeping the accepted indels in the new sequences while reinjecting the common indels that were present in the frame prealignment. This improved cluster alignment is added to the reference alignment. The three steps of the process described above are repeated iteratively for all the other clusters of the set of highly-related sequences, delivering finally a safe reference alignment

2.2.2 Stepwise addition of increasingly distant homologues to the reference MSA. The whole process in Fig. 2 is repeated iteratively while decreasing stepwise the threshold values of sequence identity that are imposed when building the clusters of related sequences to be added, and when matching these clusters to their homologues in the previous reference MSA. These 2 clustering steps are executed once at the beginning of the program and are required for only a few seconds due to the speed of the CD-hit program [11]. Frali progressively processed the homologues found at the 60, 50, 40, and 30% sequence identity cutoffs. Such a stepwise computation of successive new profile-profile alignments is essential in getting a final correct reference MSA, especially when the level of identity becomes too low, while resolving specific problematic cases listed below: (i) Two filters are applied to prevent the introduction of fragments in the reference alignment. First, a maximum length value (which may be defined for each subfamily studied) is imposed as a cutoff before sequence addition to the multiple alignment. A second filter is used after the sequences were aligned, to ensure that the aligned part is complete. This is important, for instance, in the case of multi-domain proteins that are a particular challenge for multiple alignment methods [2]. Since the alignment is done by aligning a query against a template, only the alignable parts of fused proteins will be automatically kept by Frali in the final alignment. The unalignable fragments are set aside in a distinct file that can be read later. (ii) Moreover, we have added a script that detects fused pro-teins where the combined domains of the same protein are homologous to one another (see below, for instance, the case of TrpF and TrpC enzymes). After their detection based on the knowledge of the prealigned frame, these homologous domains are cut and properly aligned during the making of the reference alignment. (iii) Whenever the number of unwanted gaps increases, it might be better to refrain from adding uninformative holes. Keeping the reference alignment as such may prove more stable, since its length would not vary every time an odd sequence appears. Where a significant number of sequences require a common and large gap, the user might consider adding it manually before adding new sequences

3 Implementation of Frali

Frali (http://embg.igmors.u-psud.fr/frali/) is a standalone Perl script package working in a Linux environment with a command-line mode. Frali includes its own modules, and the binary executables needed, such as CD-hit, Blastall, Formatdb, HHalign, and Muscle, provide for both 32 and 64 bit operating systems.

Frali requires 2 main sets of previously computed data: (1) the frame alignment that has been built semi-manually on the basis of expert knowledge (see above), (2) all available homologues that have been collected, as described above. Both inputs are prepared as text files containing FASTA formatted sequences. The output files in FASTA format contain the final updated reference alignment, the leftover sequences that could neither be aligned nor added, and fragments (sequences too small to be added).

Frali can also be used to add directly into the reference alignment new homologous sequences as soon as they are released in public databases. Our choice of defining a representative sequence for each new cluster (see above) allows an acceleration of the process without loss of accuracy. Such a fast and easy update is very helpful for users interested in curation of functional annotation and/or keeping constantly up-to-date phylogenetic trees.

4 Assessing Frali

4.1 Evaluating the accuracy of Frali

We compared the outputs of our 2-step approach

with those of different automated programs, namely ClustalW [13], Dialign [14], Dialign-TX [15], Mafft [16], Muscle [5], Probcons [17], Tcoffee, and 3Dcoffee [18]. Among the benchmark reference alignments described in BAliBASE 3.0 [19] we have utilized the whole package Rev30 made up of 30 aligned families (containing from 24 to 142 sequences), using either the whole sequences or only their homologous regions. Since these RV30 families contain subfamilies with >40% similarity but <20% similarity across the subfamilies, we first applied the psi-CD-hit program [11] to build 10 different seeds for each family by drawing lots among the clusters of its members that share >30% sequence identity. These 600 sets contain 2-15 members (from 2-38% of the total number of family members). Each set was submitted to 2 parallel actions: (i) the sequences extracted from the original BAliBASE alignment were used as a reference seed to which the rest of the homologues were added using Frali; (ii) the full set of all these homologues were submitted to each automated program as unaligned sequences. Since Frali discards the unalignable part of the sequences (Fig. 2 b2), this part (that varies from one seed to the other) was systematically removed before carrying out the reference MSA generated by the automatic tools. This removal was essential to preclude any bias when assessing the obtained reference alignments by measuring the number of correctly aligned residue pairs divided by the number of aligned residue pairs in the true alignment (score SP) and the number of correctly aligned columns divided by the number of columns in the true alignment (score TC), as defined in Thompson et al. (2005).

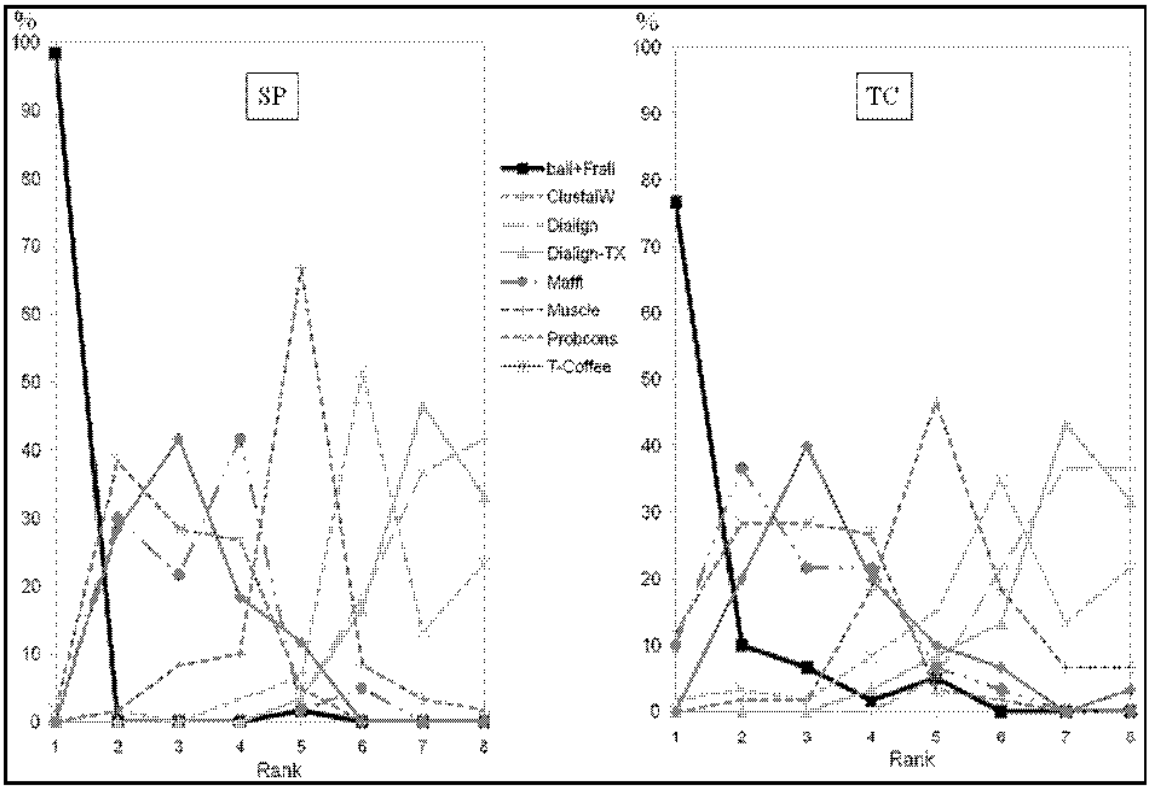


Fig. 3. Ranking the accuracy of Frali using protein alignment benchmarks. For each BAliBASE family 10 different clusters of sequences were built that have been multiply aligned using either our 2-step approach (bali+Frali) or various automated programs listed between panels SP and TC. For comparison, exactly the same portions of each sequence included in each set have been used to build the final MSA. We computed for each set the rank of each program using 2 BAliBASE scores, namely SP (left panel) and TC (right panel), and we computed for each family the median of the ranks in its 10 respective sets. Left and right panels show the percentage of families where each program has been ranked in position 1 to 8.

To gauge each method, we first ranked the SP and TC scores of each program for each set of each family and we further classified each program by computing the median of these 10 ranks for both scores in each family (Fig. 3). Our approach appears to be significantly more accurate than the tested

automated programs (Fig. 3) since it is ranking in the first position in 76.67% of the analyzed families regarding the TC score and in 98% of the tested families when measuring the SP score. Note that when automated programs perform better than Frali, namely 12.67% with Probcons, 10% with Mafft, and 1.67% with ClustalW in the case of the TC score, these BAliBASE families were made of closely related sequences. Moreover, in those cases, Frali is generally ranked second, giving a very slightly lower score.

4.2 Testing the biological relevance of Frali using a challenging family

Two models describing the possible evolution of enzyme activities were experimentally validated a decade ago when a gene encoding the TrpF activity was obtained by transforming either the gene encoding the HisA activity [19] according to the patchwork model [20] or the gene encoding the previous TrpC step [21] according to the retrograde model [22]. Moreover, another retrograde case was previously described since HisA was found to be homologous to its next step HisF [23,24]. Besides, TrpA appears to be distantly related to TrpC and TrpF (unpublished data). Thus, five genes encoding TIM-barrel proteins – TrpA, TrpC, TrpF, HisA, and HisF - are found to form a family of homologues that are probably very ancient. Indeed, the sequence identity separating these exhaustively studied proteins was found to be low (25% separating HisA from HisF) to extremely low (only 11% between HisA and TrpF and 13% between HisF and TrpF according to [25]), but their X-ray structures are superimposable. Thus, these remotely related structural homologues appear to be a challenging test case for analyzing the relevance of Frali.

Since the 3D structures of the majority of these enzymes have been determined, we could build a frame MSA with 19 sequences using either Expresso [8] or Muscle [5]. Unsurprisingly, these two automated programs gave unsatisfactory alignments and the deduced trees built using the FastTree2 program display poor biological relevance (not shown). As described above, we improved this seed alignment to a faithful alignment after manual expert edition using Swiss-PdbViewer 4.0 [9]. The tree reconstructed automatically using Muscle [5] and Expresso [8] were biologically less relevant than the ones obtained from the manually built frame MSA (Fig. 3, left panel) since their HisA and HisF subtrees were not monophyletic and branch with TrpF sequences (not shown). Moreover, their relative branch length and topology were longer than that of the tree built from the frame alignment taken as a reference. Indeed, the K tree scores [26] of Muscle and Expresso trees are 1.25787 and 0.63350, respectively. Fig. 3 further shows how Frali allows building progressively a reference alignment with selected addition to the frame (left panel) of the homologues displaying first at least 40% identity (central panel) and then the rest of the 3229 more distant homologues (right panel). The deduced phylogenetic tree keeps the same skeletal structure already observed in the frame alignment, each subfamily becoming just more and more burgeoning.

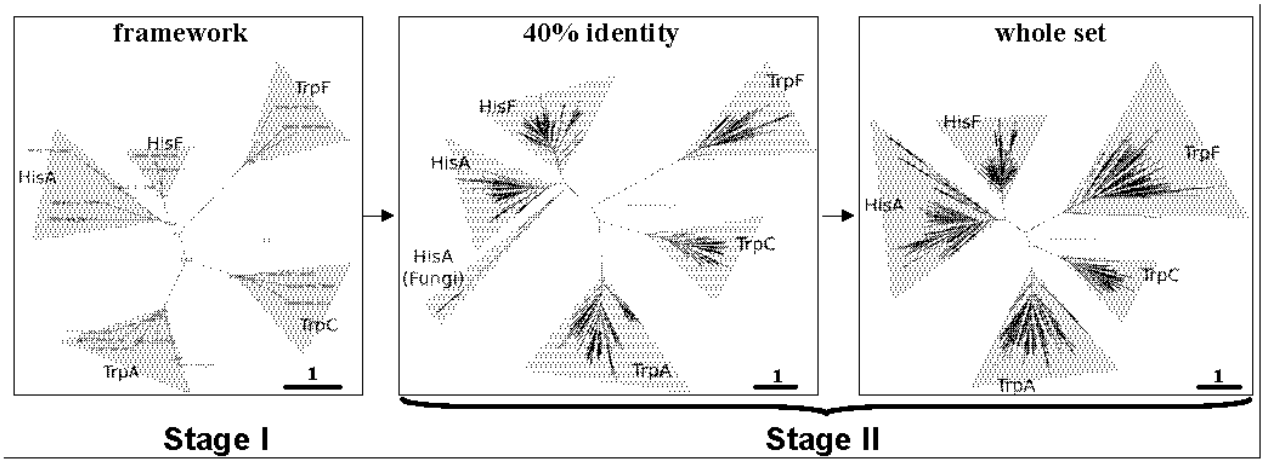


Fig. 4. Progressive addition of newly related sequences to the frame tree (left panel) reconstructed from a manually improved MSA. Trees were built using the FastTree2 program [12]. Central panel shows how the newly added sequences are added nicely as twigs specific to each subfamily on the conserved frame of the whole tree. Right panel confirms this selective addition on an invariant ancient tree skeleton with a concomitant shortening of the deepest branches.

5 Discussion

Frali has been designed to help the biologist escape various methodological and conceptual difficulties when building multiple alignments of large and diverse arrays of homologues that can be very distant. The proposed approach has a cost, since it requires a preliminary manual editing of the MSA of a limited number of experimentally well-characterized proteins that stand for the various subfamilies of such arrays. This limited number of seed sequences could be as low as 5% of the total number of family members. Once such a solid basis is established, the whole alignment can be obtained very rapidly by using the completely automated Frali program. This reasonable effort of manual editing is rewarding in the end since it can guarantee getting a reference MSA that is both accurate and biologically relevant. This is mainly due to our strategy of progressive addition of new homologous proteins that have been sampled by tight clustering, defining a high similarity to a few of the prealigned sequences in the frame alignment. This careful handling of the sequences during the profile-profile step and the strict treatment of the indels helps maintain the accuracy of the obtained reference alignment, as shown in comparative studies with automated programs on the same benchmarking data (Fig. 3). Noticeably, contrarily to the case of completely automated one-step methods, the biologist will keep mastering the intricacies of the process of multiply align complex families of homologous sequences at each step of the Frali approach, even when they are highly dissimilar.

Our tool presents several decisive advantages over other methods. (i) Whatever the present and future level of flooding of newly released genomic sequences, we guarantee the accuracy of the MSA since we start with a high level of truthfulness at the step of the frame alignment, and we keep it unabated when adding stepwise and gradually the whole set of the other homologues. (ii) Our procedure is fast, its rate being linearly proportional to the increase in the total number of sequences to be aligned. (iii) Frali resolves instantaneously difficult cases such as multi-domain and/or fused proteins without any prior detection or treatment. (iv) The opening of too large holes is prevented by our gradual and stepwise procedure, but the possibility of introducing a limited number of gaps is kept since they could be valuable phylogenetic markers [27]. (v) Phylogenetic trees derived from MSA generated with Frali systematically display a better topology and a shorter length than those derived using one-step automated tools.

In addition, the full reference MSA may be updated at any time while keeping its accuracy and biological relevance. Indeed, addition of newly published homologues takes a few seconds and is highly precise. Therefore, Frali allows effortlessly the last update of a phylogenetic tree of a large and complex family to be generated at anytime. Note, however, that the occurrence of representatives of a completely new sub-family could require a supplementary step before their addition to the reference alignment.

Acknowledgements

This work was funded by the CNRS (UMR 8621), the PPF 'Bioinformatique et BioMathématique' of the Université Paris-Sud and the Agence Nationale de la Recherche (ANR-05-MMSA-0009 MDMS_NV_10). M.B. is a PhD student supported by the French Ministry of Research.

References

[1] Pei J. Multiple protein sequence alignment. Curr Opin Struct Biol. 18:382-386, 2008

[2] Pirovano W, Heringa J. Multiple sequence alignment. Methods Mol Biol. 452:143-161, 2008

[3] Kemena C. Notredame C. Upcoming challenges for multiple sequence alignment methods in the high-throughput era. Bioinformatics 25:2455-2465, 2009

[4] Edgar:R.C. and Batzoglou, S. Multiple sequence alignment. Curr. Opin. Struct. Biol.16:368–373, 2006

[5] Edgar R.C. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics 5:113, 2004

[6] Patterson C Homology in classical and molecular biology:Mol. Biol. Evol. 5:603-625, 1988

[7] Descorps-Declère S Lemoine F. Sculo Q Lespinet O and Labedan B. The multiple facets of homology and their use in comparative genomics to study the evolution of genes:genomes:and species. Biochimie 90:595-608, 2008

[8] Armougom F:Moretti S:Poirot O:Audic S Dumas P Schaeli B Keduas V Notredame C. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 34(Web Server issue):W604-608, 2006

[9] Guex N. Peitsch M.C. SWISS-MODEL and the Swiss-PdbViewer: An environment for comparative protein modeling. Electrophoresis 18:2714-2723,

1997

[10] Eddy SR. Profile hidden Markov models. Bioinformatics 14:755-763, 1998

[11] Li W Godzik A. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences. Bioinformatics 22:1658-1659, 2006

[12] Price M.N. Dehal P.S. and Arkin A.P. FastTree: Computing Large Minimum-Evolution Trees with Profiles instead of a Distance Matrix. Mol Biol Evol 26:1641-1650, 2009

[13] Thompson JD, Higgins DG, Gibson TJ. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 2:4673-4680, 1994

[14] Morgenstern,B. DIALIGN 2: improvement of the segment-to-segment approach to multiple sequence alignment. Bioinformatics, 15:211–218, 1999

[15] Subramanian AR, Kaufmann M, Morgenstern B DIALIGN-TX: greedy and progressive approaches for the segment-based multiple sequence alignment. Algorithms Mol. Biol., 3:6, 2008

[16] Do CB, Mahabhashyam MSP, Brudno M, Batzoglou S. PROBCONS: Probabilistic Consistency-based Multiple Sequence Alignment. Genome Research, 15:330-340, 2005

[17] O'Sullivan,O. Suhre K, Abergel C, Higgins DG, Notredame C. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol., 340:385–395, 2004

[18] Thompson,J.D., Koehl,P., Ripp,R. and Poch,O. BAliBASE 3.0: latest developments of the multiple sequence alignment benchmark. Proteins, 61:127–136, 2005

[19] Jürgens C Strom A Wegener D Hettwer S Wilmanns M Sterner R Directed evolution of a (beta alpha)8-barrel enzyme to catalyze related reactions in two different metabolic pathways. Proc Natl Acad Sci USA. 97:9925-9930, 2000

[20] Jensen RA Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409–425, 1976

[21] Altamirano M. M. Blackburn J.M. Aguayo C. Fersht A.R. Directed evolution of new catalytic activity using the alpha/beta-barrel scaffold Nature 403:617–622, 2000

[22] Horowitz:N.H. On the evolution of biochemical syntheses. Proc. Natl. Acad. Sci. USA 31:153–157, 1945

[23] Fani R Mori E Tamburini E Lazcano A. Evolution of the structure and chromosomal distribution of histidine biosynthetic genes. Orig Life Evol Biosph 28:555-570, 1998

[24] Lang D Thoma R Henn-Sax M Sterner R Wilmanns M Structural evidence for evolution of the β/α barrel scaffold by gene duplication and fusion. Science 289:1546–1550, 2000

[25] Leopoldseder S Claren J Jürgens C Sterner R Interconverting the catalytic activities of (ba)8-barrel enzymes from different metabolic pathways: sequence requirements and molecular analysis. J Mol Biol 337:871–879, 2004

[26] Soria-Carrasco V Talavera G Igea J Castresana J. The K tree score: quantification of differences in the relative branch length and topology of phylogenetic trees. Bioinformatics. 23:2954-2956, 2007

[27] Löytynoja A, Goldman N. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632-1635, 2008

Chapitre 5 : Innovations méthodologiques pour l'étude des familles de protéines

1 Mise en place nécessaire d'une base de données *ad hoc*

1.1 Besoins

Afin de gérer toutes séquences des différentes superfamilles étudiées dans ce travail de Thèse et de manipuler les alignements multiples créés avec Frali, j'ai mis en place une base de données relationnelles accompagnée d'un ensemble de scripts qui lui servent d'interface et permettent de traiter les données. Une telle base de données est essentielle compte tenu du nombre de séquences (plusieurs milliers par familles) et de la richesse de leurs attributs (identifiants, annotations...) détaillés ci-dessous. J'ai ainsi créé et maintenu une base de données locale appelée *AP* (pour *aligned proteins*, soit « protéines alignées »).

1.2 Mise en place et fonctionnement de la base de données

Cette base de données a été construite en utilisant le système de gestion de base de données relationnelles PostgreSQL[2] (version 8.4) qui présente l'avantage d'être à la fois libre et puissant. Cette base contient deux grandes tables principales : la table des protéines alignées et la table de la taxonomie des espèces les codant. Plusieurs tables annexes les accompagnent (Figure 10).

La table de taxonomie (*taxonomy*) contient toutes les informations taxonomiques relatives aux taxons représentés dans la table des protéines.

La table des protéines (*proteins*) contient, pour chaque ligne, la région de la protéine alignée extraite de l'alignement multiple réalisé pour une famille donnée . Une même protéine peut être présente dans plusieurs familles différentes (fusions), ce qui résulte en autant de lignes indépendantes. De même, il est possible que plusieurs domaines d'une même protéine soient homologues, et, là encore, il y aura autant de lignes que de domaines homologues. L'ensemble des lignes donne donc l'image de l'alignement multiple de la famille considérée.

La table *proteins* et les tables annexes *(discarded, motifs* et profiles) sont toutes liées à la table *families* qui défini les familles de séquences alignées. Cette dernière table ne contient qu'une seule colonne contenant les identifiants des familles étudiées qui sont utilisés par les autres tables.

2 http://www.postgresql.org

Afin de faciliter les mises à jour récurrentes des données générales de la base de données *AP* des protéines alignées qui nécessite de grandes quantités de données, j'ai créé une copie locale de *Uniprot* adaptée à mes besoins sur le même modèle que ma base de données de familles protéines, à savoir une table d'entrée pour les protéines (*entries*) et de taxonomie (*taxonomy*).

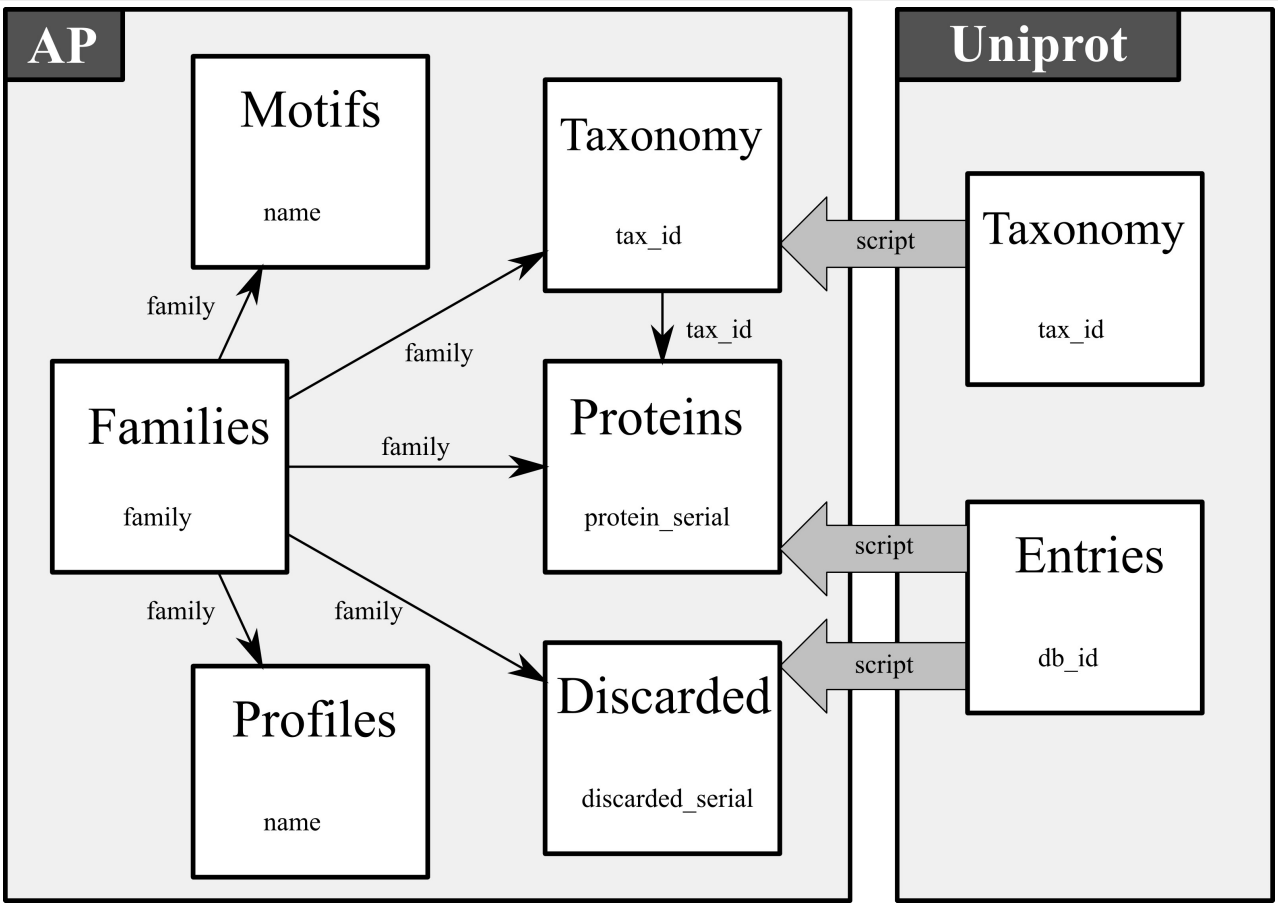


Figure 10 : Schéma simplifié des bases de données *AP* et *Uniprot.*
Les tables sont représentées par des carrés incluant leur titre et leur clé primaire. Elles sont liées par des clés étrangères symbolisées par des flèches fines. Les deux bases AP *(à gauche) et* Uniprot *(à droite) sont indépendantes, mais les tables* taxonomy, proteins *et* discarded d'AP *sont enrichies par script, de manière ponctuelle, avec les données provenant des tables* taxonomy *et* entries d'Uniprot *(symbolisé par des flèches épaisses).*

1.2.1 Table taxonomy *de la base* AP

Colonne	Type	Description
tax_id	Entier non NULL	Identifiant taxonomique de ce taxon défini par le NCBI
parent_id	entier	tax_id du parent direct de ce taxon
newt_id	caractère variant(5)	Identifiant raccourci jusqu'à 5 lettres de l'organisme défini dans NEWT de UniProtKB
tax_name	texte	Nom principal de l'organisme (nomenclature binomiale et nom de souche)
complete_tax	booléen	Si le génome de ce taxon est complet
ranks	texte[]	Ensemble résumé des grands divisions taxonomiques auxquels appartient ce taxon hiérarchiquement
clades	texte[]	Comme rank mais avec des niveaux hiérarchiques personnalisés
ecology	texte	Caractéristiques écologiques de l'organisme en question

Tableau 2 : Schéma de la table *taxonomy* de la base de données *AP*.

Index :

- "taxonomy_pk" PRIMARY KEY, btree (tax_id)
- "taxonomy_parent_id_index" btree (parent_id)
- "taxonomy_tax_id_index" btree (tax_id)

Chaque ligne correspond à un taxon présent dans la table proteins, caractérisé par le tax_id (Tableau 3). Toutes ces données sont mises à jour automatiquement (par script) en se basant sur les tax_id. La colonne rank est créée automatiquement par le script d'import des données en remontant récursivement la hiérarchie (domaine, règne, phylum, classe, ordre, famille, genre).

Colonne	Valeur
parent_id	1305
newt_id	STRSV
tax_name	Streptococcus sanguinis SK36
clades	{Bacteria,Firmicutes,F-LactoBacil,FL-Streptoc}
complete_tax	VRAI
ranks	{Bacteria,Firmicutes,Bacilli,Lactobacillales,Streptococcaceae,Streptococcus}
ecology	Chains;Endocarditis;Yes;2.4;"";Host-associated;No;;Facultative;Human;;Coccus;Mesophilic

Tableau 3 : Exemple d'une ligne de la table taxonomy pour le taxon *Streptococcus sanguinis SK36*

1.2.2 Table proteins *de la base* AP

Colonne	Type	Description
id	texte non NULL	Identifiant unique de la séquence alignée (peut contenir num s'il est défini)
num	entier	Numéro si la séquence a plusieurs domaines homologues alignés
tax_id*	entier	Identifiant défini dans la base Taxonomy du NCBI
db*	texte	Nom de la base de données d'origine (UniProtKB), généralement sp (Swissprot) ou tr (TremBL)
db_id	texte	Identifiant unique (AC) de la séquence dans UniProtKB
db_long_id*	texte	Identifiant long (ID) de la séquence dans UniProtKB (plus descriptif)
alignment	texte	Séquence alignée
alength	entier	Longueur de la séquence alignée (sans les indels)
length*	entier	Longueur de la séquence complète
family	texte	Nom de la famille à laquelle la séquence alignée appartient
type	texte[]	Annotation fonctionnelle, peut contenir plusieurs niveaux
inact	booléen	Si la séquence est alignée mais a été annotée comme inactive (cas particulier de certaines familles, voir les DHOases)
complete*	booléen	Marque si la séquence est notée comme complète dans UniProtKB
location*	texte	Le cas échéant, précise sur quel réplicon (unité de réplication : chromosome, plasmide...) se situe le gène
locus*	entier	Position (ordre) de la séquence dans le génome s'il est complet
direction*	caractère(1)	Direction du gène dans le génome (brin codant), pertinent uniquement dans chaque génome indépendamment
protein_serial	entier (série)	Numéro de série unique

Tableau 4 : Schéma de la table *proteins* de la base de données *AP*.

Index :

- "proteins_pk" PRIMARY KEY, btree (protein_serial)
- "proteins_complete_index" btree (complete)
- "proteins_db_id_index" btree (db_id)
- "proteins_family_index" btree (family)
- "proteins_id_index" btree (id)
- "proteins_location_index" btree (location)
- "proteins_tax_id_index" btree (tax_id)
- "proteins_type_index" gin (type)

Chaque ligne de la table correspond à une séquence alignée dans une famille donnée (Tableau 5). Les colonnes notées avec une astérisque (*) sont des données récupérées

automatiquement (par script) de la base de données annexe *Uniprot* en se basant sur l'identifiant db_id. À noter que plusieurs lignes peuvent porter un même identifiant db_id, mais pour différentes familles, et si nécessaire pour différents domaines homologues dans une même famille (auquel cas le num sera différent).

Colonne	Valeur
id	Q9RV76
num	NULL
tax_id	1299
db	tr
db_id	Q9RVC3
db_long_id	PYRC_DEIRA
alignment	(séquence alignée)
alength	412
length	416
family	PyrC
type	{DHO,III}
inact	Faux
complete	Vrai
location	NULL
locus	1086
direction	-

Tableau 5 : Exemple d'une ligne de la table *proteins* pour la PyrC de *Deinococcus radiodurans.*

1.2.3 Tables annexes

Table *discarded* (séquences exclues)

Colonne	Type	Description
id	texte non NULL	Identifiant unique de la séquence (peut contenir num s'il est défini)
db_id	texte	Identifiant unique (AC) de la séquence dans UniProtKB
db	texte	Nom de la base de données d'origine (UniProtKB), généralement sp (Swissprot) ou tr (TremBL)
family	texte	Nom de la famille à laquelle la séquence alignée devrait appartenir

Tableau 6 : Schéma de la table annexe *discarded* de la base *AP.*

Dans cette table annexe sont listées toutes les séquences exclues parce que fragmentaires ou

de mauvaise qualité (Tableau 7). La présence d'une séquence dans cette table prévient tout import ultérieur dans la table de protéines.

Colonne	Valeur
id	Q9S3S1
db_id	Q9S3S1
db	tr
family	PyrC

Tableau 7 : Exemple d'une ligne de la table *discarded* pour la séquence Q9S3S1 de *Serratia marcescens*.

Table *profiles* (définition des profils)

Colonne	Type	Description
profile	texte	Identifiant du profil
family	texte	Famille à inclure dans le profil
name	texte	Nom donné à la famille
condition	texte	Conditions SQL pour sélectionner les membres de la famille à afficher dans le profil
num	entier	Nombre donnant l'ordre de chaque famille dans le profil

Tableau 8 : Schéma de la table annexe *profiles* de la base *AP*.

Cette table contient les informations de définition de profils pour observer l'absence ou la présence d'une protéine d'une famille donnée selon ses annotations. Chaque ligne contient la description d'un groupe de protéines à afficher dans le profil.

Colonne	Valeur
profile	DHO
family	PyrC
name	PyrC3i
condition	type[1]='DHO' AND type[2]='III' AND inact
num	12

Tableau 9 : Exemple d'une ligne de la table *profiles*.
Cette ligne désigne les DHOases de type III inactives dans la superfamille PyrC en les nommant PyrC3i.

1.2.4 Table motifs

Colonne	Type	Description
name	texte	Nom du motif
family	texte	Famille de séquence utilisant le motif
motif	texte[]	Description du motif en expression rationnelle, plusieurs motifs pouvant être définis séparément
annotation	texte	Conditions SQL pour sélectionner les membres de la famille à afficher dans le profil

Tableau 10 : Schéma de la table annexe motifs de la base *AP*.

Cette table contient les informations de définition de motifs pour annoter certaines séquences qui ont des résidus très conservés caractéristiques.

Colonne	Valeur
name	ATCase
family	ArgF
motif	{291:5:R.Q.E,352:4:HP.P}
annotation	type[1]='ATC'

Tableau 11 : Exemple d'une ligne de la table *motifs*.
Cette ligne donne deux motifs caractéristiques des ATCases, sous la forme « position:taille:motif ». Les points désignent n'importe quel résidu (donc non conservé).

1.2.5 Table entries *de la base* Uniprot

Colonne	Type	Description
db_id	caractère(6)	Identifiant unique (AC) à 6 lettres et chiffres de la séquence défini dans UniProtKB
id	caractère variant(12)	Identifiant long (ID), généralement un court nom de protéine ou l'identifiant court, suivi de l'identifiant raccourci du taxon (newt_id)
tax_id	entier	Identifiant du taxon auquel appartient la séquence
reviewed	booléen	Si VRAI, la séquence provient de Swissprot, sinon de TrEMBL
length	entier	Longueur de la séquence exprimée en nombre d'acides aminés
sequence	texte	Enchainement des acides aminés de la séquence
gene	texte	Nom du gène codant
full_name	texte	Nom complet de la fonction de la protéine telle qu'elle est annotée dans UniProtKB
complete	booléen	Si le génome est complètement séquencé ou non
locus	entier	Position (ordre) dans le génome
location	texte	Position (ordre) de la séquence dans le génome s'il est complet
direction	caractère(1)	Direction du gène dans le génome (brin codant), pertinent uniquement dans chaque génome indépendamment
entry_id	entier	Numéro de série unique

Tableau 12 : Schéma de la table *entries* de la base *Uniprot.*

Index :

- "uniprot_pk" PRIMARY KEY, btree (db_id)
- "complete_index" btree (complete) WHERE complete
- "length_index" btree (length)
- "reviewed_index" btree (reviewed) WHERE reviewed = true
- "tax_id_index" btree (tax_id)
- "tax_locus_index" btree (tax_id, locus)
- "uniprot_id_index" btree (id)

L'ensemble des données provient des bases de données d'UniprotKB[3] [UniProt_Consortium 2011]. Cette table permet de remplir les informations de la table *proteins* de la base *AP.*

3 UniprotKB est accessible à l'adresse : http://www.uniprot.org

1.2.6 Table taxonomy *de la base Uniprot*

Colonne	Type	Description
tax_id	entier	Identifiant numérique du taxon tel que défini par le NCBI
parent_id	entier	tax_id du parent direct de ce taxon
newt_id	caractère variant(5)	Identifiant raccourci jusqu'à 5 lettres de l'organisme
tax_name	texte	Nom principal de l'organisme (nomenclature binomiale et nom de souche)
env	booléen	Si ce n'est pas un génome mais un prélèvement environnemental
complete_tax	booléen	Si le génome de ce taxon est complet
rank	texte	Niveau hiérachique : règne, genre, ordre, etc.

Tableau 13 : Schéma de la table t*axonomy* de la base *Uniprot.*

Index :

- "taxonomy_pk" PRIMARY KEY, btree (tax_id)
- "complete_tax_index" btree (complete_tax)
- "newt_id_index" btree (newt_id)

Chaque ligne correspond à un taxon. Les données de hiérarchie (tax_id, parent_id, tax_name, rank) proviennent de la taxonomie du NCBI[4] [Sayers et al. 2011]. Les newt_id [Phan et al. 2003] et les données de génome complet (complete_tax, env) proviennent de Uniprot[5] [UniProt_Consortium 2011]. Ces données permettent de remplir les informations de la table *taxonomy* de la base *AP.*

2 Scripts d'exploitation des données

Un ensemble de scripts écrits en Perl et utilisables en ligne de commande permettent de gérer les données de la base, de les maintenir à jour et de les exploiter. Ces scripts représentent 9500 lignes de code et ne peuvent être publiées dans ce manuscrit de thèse ; nous prévoyons de créer une page web dédiée pour permettre son téléchargement. Ces scripts sont réunis en quatre groupes : les scripts d'import/export, les scripts de maintenance de la base, les scripts d'annotation et les scripts de visualisation et d'export des résultats.

2.1 Scripts d'import/export

Le script *ap_io_import.pl* permet d'importer un alignement multiple d'un fichier fasta dans

4 Taxonomie du NCBI disponible à l'adresse : http://www.ncbi.nlm.nih.gov/Taxonomy/
5 Taxonomie de Uniprot disponible à l'adresse : http://www.uniprot.org/taxonomy/

l'alignement d'une famille déterminée de la table des protéines. Il est possible en principe d'importer des séquences non alignées, mais les possibilités d'exploitations ultérieures (annotation) sont moindres.

Le script *ap_io_export.pl* permet d'exporter un alignement multiple de séquences d'une famille donnée. Des arguments supplémentaires permettent de restreindre les séquences sélectionnées en exploitant les propriétés de la table des protéines ainsi que de la table de taxonomie. Par exemple, il est possible de n'extraire que les séquences provenant de génomes complets.

2.2 Scripts de maintenance

Il existe deux scripts de maintenance, pour la base *Uniprot* et la base *AP*. Le premier récupère automatiquement les dernières données d'Uniprot si une nouvelle version est disponible et met à jour les tables de protéines et de taxonomie de la base *Uniprot*. Le second permet de compléter et mettre à jour les données des séquences importées dans la base *AP* en se basant sur les différents identifiants de séquence (accession_number) et d'espèce (tax_id).

2.3 Scripts de visualisation

Afin de faciliter la consultation de la base, j'ai également créé des scripts qui listent les séquences selon différentes conditions : les séquences d'une même famille (*ap_view_family.pl*), ou les séquences contenues dans un génome donné (*ap_view_context.pl*). Dans ce cas-là, le script donne différentes informations sur ces séquences, en particulier leur contexte génétique.

2.4 Script d'annotation

Dans la base *AP*, les annotations sont hiérarchiques, c'est-à-dire qu'il y a plusieurs niveaux d'annotation possible, du plus général au plus précis (la variable *type* est une liste). Les niveaux et le type d'annotation varie selon les familles et sont définis librement par l'utilisateur. Plusieurs scripts permettent d'annoter les séquences :

- à partir d'un arbre de distances enraciné donné par l'utilisateur et en annotant par monophylie *(ap_annotation_tree.pl)*. Ce script utilise BioPerl [Stajich et al. 2002] pour analyser les arbres au format Newick. Il parcourt l'arbre à la recherche de nœuds dont les feuilles des deux sous-arbres ont au moins une feuille annotée de la même manière. Si c'est le cas, l'annotation est transférée à toutes les feuilles sous ce nœud ;
- en utilisant des motifs identifiés par l'utilisateur *(ap_annotation_motifs.pl)*. Les motifs correspondent à des positions dans l'alignement multiple de la famille, si bien que des

résidus isolés peuvent servir de motif en utilisant leur position absolue dans la séquence (par exemple une histidine seule). Les motifs peuvent être fournis par l'utilisateur ou définis dans une table annexe à cet effet *(motifs)* ;

- en utilisant le contexte génétique des séquences *(ap_annotation_context.pl)*. La présence dans un même contexte de plusieurs séquences correspondant à des enzymes d'une même voie permet d'annoter les séquences de manière très précise ;
- en utilisant la similarité de séquence *(ap_annotation_similarity.pl)*. Cela nécessite un taux d'identité de séquence élevé. En effet, la fonction varie rarement au dessus de 40% d'identité entre deux séquences [Furnham et al. 2009], mais l'annotation des superfamilles nécessite une plus grande précision, au delà de la simple fonction moléculaire (complexe, voie...). Cela dépend aussi du niveau hiérarchique des annotations.

2.5 Script d'extraction de données

Le script *ap_io_profiles.pl* permet d'extraire des profils des séquences dans tous les génomes contenant au moins un homologue. Les profils phylogénétiques des gènes des voies étudiées (notamment le Tableau 17) ont été créés de cette manière.

RÉSULTATS

Chapitre 6 : Concept de module réactionnel : lien entre l'évolution des amidohydrolases cycliques et la similarité des réactions qu'elles catalysent

1 Résumé

La première famille que j'ai étudiée avec la méthodologie décrite précédemment est celle des dihydroorotases qui catalysent la troisième étape de la voie de biosynthèse des pyrimidines (Chapitre 2, p 22). La superfamille inclut également les hydantoinases, dihydropyrimidinases (voie de dégradation des pyrimidines, Chapitre 2, p 25) et allantoïnases (voie de dégradation des purines, Chapitre 2, p 26).

Les dihydroorotases (DHOases) présentent une grande diversité structurale et sont le résultat d'une histoire évolutive complexe comme le montre leur phylogénie. J'ai donc proposé une nouvelle classification des DHOases basée sur cette phylogénie qui reflète également leurs caractéristiques structurelles et biochimiques. Le premier type I regroupe des DHOases des trois grands domaines du vivant, de structure homodimérique à l'exception des Eucaryotes « supérieurs » dont le domaine DHOase est fusionné dans la CAD. Le type II inclut des DHOases dont tout un domaine a été perdu et qui se présentent sous la forme d'homomères (dimères ou monomères), chez des Bactéries et quelques Eucaryotes. Le type III regroupe des DHOases de Bactéries exclusivement et peuvent être homodimériques ou en complexe dodécamérique avec des ATCases (enzyme précédente dans la voie). Dans ce cas-là, la DHOase peut être catalytiquement inactive, jouant uniquement un rôle structural.

La phylogénie de la superfamille montre remarquablement que les DHOases de type I sont plus proches des autres amidohydrolases cycliques qu'elles ne le sont des autres types de DHOases, d'autant plus qu'elles catalysent des réactions similaires sur des substrats très ressemblants. On retrouve une telle ressemblance réaction/substrat dans l'étape suivante de la voie de biosynthèse des pyrimidines, la dihydroorotate déshydrogénase, qui ressemble à la dihydropyrimidine déshydrogénase (voie de dégradation des pyrimidines). L'étude de leur famille montre que les dihydroorotate déshydrogénases présentent une grande diversité à l'instar des DHOases : un type 2

membranaire et un type 1 cytoplasmique, divisé en sous-types 1A dimérique, 1B et 1S hétérotétramériques (de Bactérie et d'Archées, respectivement). Ces groupes se retrouvent dans la phylogénie de la superfamille, de même qu'un autre groupe, qui s'enracine à la base des 1S, et qui contient non seulement les dihydropyrimidines déshydrogénases, mais aussi deux sous-arbres de protéines non identifiées (notés X1 et X2).

Ce travail m'a amené à décrire le concept de module réactionnel, qui représente un enchaînement des réactions similaires dans des voies différentes, réalisées généralement par des enzymes homologues (DHOase puis DHODase dans la voie de biosynthèse, DHPDase puis DHPase dans la voie de dégradation des pyrimidines). Un tel module représente une brique élémentaire de voie métabolique, issue d'une voie ancestrale, dupliquée et réutilisée pour former des voies plus spécialisées.

2 Publication

Ce travail avait été commencé en collaboration avec le groupe de Nicolas Glansdorff. Bien que Nicolas soit décédé accidentellement en juillet 2009, sa participation initiale avait été décisive, expliquant pourquoi nous l'avons maintenu comme auteur de ce papier :

- Barba Matthieu, Xu Ying, Glansdorff Nicolas & Labedan Bernard. *Exploiting the concept of reaction module relating the evolution of the cyclic amidohydrolases to the similarity of the chemical reactions they catalyze*. In preparation for submission to MBE.

Submitted as a Research Article

EXPLOITING THE CONCEPT OF REACTION MODULE.

1. RELATING THE EVOLUTION OF THE CYCLIC AMIDOHYDROLASES TO THE SIMILARITY OF THE CHEMICAL REACTIONS THEY CATALYZE

Matthieu Barba[1], Ying Xu[2], Nicolas Glansdorff[2#] and Bernard Labedan[1*]

[1]Institut de Génétique et Microbiologie, Université Paris Sud, CNRS UMR 8621,

Bâtiment 400, 91405 Orsay Cedex, France

[2]Microbiology, Free University of Brussels (VUB) and J.M. Wiame Research Institute

1, ave E. Gryzon, B-1070, Brussels, Belgium

[#]Deceased

[*]Corresponding author:

Tel : +33 1 69 15 35 60

Fax : +33 1 69 15 72 96

E-mail: bernard.labedan@igmors.u-psud.fr

Running head: *Evolution of cyclic amidohydrolases*

Key words: amidohydrolase, dihydroorotase, HMM profile, structure-function relationship, multiple sequence alignment, phylogenetic tree, chemical reactions, reaction module

ABSTRACT

Dihydroorotases (DHOases) are universal proteins catalyzing the third step of pyrimidine biosynthesis. These zinc metalloenzymes belong to the superfamily of cyclic amidohydrolases, comprising also allantoinases (ALL) and hydantoinases (HYD) / dihydropyrimidinases (DHP). The evolutionary relationships between these homologous enzymes were estimated after designing a two-step approach to build an accurate multiple sequence alignment in order to get a trustworthy phylogenetic tree. Such a rigorous approach helped us to propose a new classification of DHOases in three major types that are rather distant. ALL, HYD and DHP enzymes that are involved in catabolic pathways (degradation of purines and pyrimidines) define three monophyletic subtrees that share a common ancestor with the DHOases type I. Thus, these other cyclic amidohydrolases are phylogenetically closer of one class of DHOases although they differ in their catalytic mechanism more than the three DHOases classes do between them. To rationalize the evolutionarily close proximity of these cyclic amidohydrolases that catalyze rather different cellular functions, we introduced a new concept based on the following facts. (i) These enzymes catalyze similar chemical reactions on very similar substrates. (ii) We further demonstrate that enzymes dihydroorotate dehydrogenase (PyrD) and dihydropyrimidine dehydrogenase (PydA), that catalyze the next step of their respective pathways form another family of homologous proteins. We call reaction module the connection of sets of homologous proteins catalyzing successive steps in parallel pathways using similar chemical mechanisms on similar substrates. These reaction modules could be seen as the elementary constituents of the so-called functional modules in system biology. Interestingly, two families of uncharacterized proteins are defined in the superfamily tree of DHODases/DHPDases as subtrees having recently diverged from the DHPDases subtree. Based on their gene context, these uncharacterized proteins appeared to be involved in purine degradation. The companion paper will further explore the evolutionary links between purine and pyrimidine metabolisms and will show

how this concept of reaction module is useful to assign a molecular function to metabolic orphans or a coding gene to an orphan enzymatic activity.

INTRODUCTION

Recent progress in genome sequencing and systems biology allow now to deal with the emergence and evolution of modern molecular function of proteins using global approaches (Caetano-Anollés et al., 2009 - Kim and Caetano-Anollés, 2010). In this context, studying the evolution of metabolic pathways involves to trace back how the enzymes that catalyze their successive steps have adapted to perform specific chemical reactions. One of the main experimental approaches used to achieve such an analysis requires studying families of homologous enzymes that perform the same or a similar molecular function (see, for example, Engelhardt et al., 2005, 2006, and 2009 and references inside). An increasingly prevailing model postulates that present-day enzyme families and superfamilies are the result of the progressive divergence of ancestral proteins endowed with promiscuous function (for a recent review, see Khersonsky and Tawfik, 2010). Contrarily to the classical model proposed by Ohno (1970), it is now anticipated that innovation (enzyme promiscuity) preceded gene duplication and divergence of the paralogous copies by descent with modification (Hughes, 2005). To explain the appearance of many closely related families grouping into mechanistically diverse superfamilies, Glasner et al. (2006) have proposed to distinguish two degrees of promiscuity: shared chemistry (substrate ambiguity) and substrate binding (catalytic promiscuity). More and more data suggest that substrate ambiguity, first defined in the classical patchwork model of Jensen (1976) rather than catalytic promiscuity (O'Brien and Herschlag, 1999) is the main road to allow divergence of most enzyme families (Khersonsky et al., 2006, Khersonsky et al., 2011). As already underlined, many enzyme superfamilies are mechanistically diverse because of the importance of chemistry in the evolution of catalysis (reviewed in Glasner et al 2006).

In this paper, we are focusing on a protein family belonging to amidohydrolases, one of the most structurally and catalytically diverse superfamilies (Seibert and Raushel, 2005 and references

inside). Dihydroorotases (DHOases) are universal proteins catalyzing the third step of pyrimidine biosynthesis corresponding to the reversible cyclization of carbamoyl L-aspartate to L-dihydroorotate. These zinc metalloenzymes appear remarkably diverse at the levels of their tertiary and quaternary structures. In eukaryotes, DHOase is found as part of a large (1.5 MDa) hexameric multifunctional protein CAD (Shoaf and Jones 1971 - Evans and Guy, 2004 and references inside) that also contains the enzymes catalyzing the first two steps of pyrimidine biosynthesis, namely carbamoyl-phosphate synthase II (CPS, EC 6.3.5.5) and aspartate carbamoyltransferase (ATC, EC 2.1.3.2). In prokaryotes, DHOases are often monofunctional homodimers but may be complexed as multifunctional oligomers such as the DHO-ATC (DAC) complex in *Aquifex aeolicus* (Zhang et al., 2009 and references inside). In such complexes, DHOase subunit may have a catalytic role or a structural one. In this last case, organisms must own a second, active, DHOase (Berg and Evans, 1993).

The former group of O'Donovan (see Fields et al., 1999 - Brichta, 2002 – Brichta et al., 2004) already tried to cope with this diversity of DHOases by proposing a classification. They defined two main types (I and II) of DHOases according to their subunit mass size. The larger type I was separated in four classes (a to d) defined as a combination of the nature of their association with other interacting proteins and of their role, either active (enzymatic role) or inactive (structural role).

Moreover, DHOases (EC 3.5.2.3) share a common ancestor with allantoinases (EC 3.5.2.5) and hydantoinases (HYD)/dihydropyrimidinases (DHP) that catalyze the hydrolysis of cyclic C-N bonds (Holm and Sander, 1997 - Seibert and Raushel, 2005 - Lohkamp et al., 2006). This cyclic amidohydrolases superfamily is itself only a minor part of the huge superfamily of *sensu lato* amidohydrolases that probably stands as the paradigm of large structural and catalytic diversity (Seibert and Raushel, 2005) resulting from a long and remarkable evolutionary history (Holm and

Sander, 1997). For instance, although DHP and HYD have the same E.C. number (EC 3.5.2.2), they are different enzymes and have different substrate specificities (Lohkamp et al., 2006 - Liu et al, 2007). DHP enzymes (including L-hydantoinases) catalyze the reversible hydrolytic ring opening of six- or five-membered cyclic diamides such as dihydropyrimidines and 5'-monosubstituted hydantoins to the corresponding 3-ureido acids and carbamoyl amino acids, respectively. On the other hand, HYD are widely used in the production of D-amino acids which are precursors for synthesis of antibiotics, peptides and pesticides (Liu et al, 2007) but these enzymes could not hydrolyze dihydropyrimidines (Lohkamp et al., 2006), and their natural substrates are often unknown (Seibert and Raushel, 2005).

This paper first reviews the classification of DHOases (Fields et al., 1999) by examining the rather complex and intricate phylogeny of all cyclic amidohydrolases obtained after setting a new approach to get an accurate multiple sequence alignment (MSA) of the whole superfamily. The reconstructed phylogenetic tree of cyclic amidohydrolases helped to propose a new, simpler, classification of DHOases in three classes while disentangling a significant part of their complex structure - function relationships. Moreover, we tried to understand why the other cyclic amidohydrolases are phylogenetically closer of one class of DHOases than are the three classes between them. We examine the respective importance of shared chemistry (substrate ambiguity) versus substrate binding (catalytic promiscuity).when comparing these homologous cyclic amidohydrolases as well as the respective enzymes catalyzing the next steps in the relevant pathways. This led us to propose the new concept of reaction modules that link successive homologous enzymes as part of the evolutionary process that progressively built functional modules.

MATERIALS AND METHODS

Building a reference multiple sequence alignment (MSA) of cyclic amidohydrolases

To build a reliable MSA reflecting the structural and functional diversity of cyclic amidohydrolases, we used a two-step approach. In a first step, we collected (March 2011) the 20 homologous sequences (Table 1) that have been labeled experimentally studied as indicated in UniProtKB/Swiss-Prot (The UniProt Consortium, 2008) and published in the Protein Data Bank (Dutta et al., 2009). These sequences were multiply aligned using the Expresso update (Armougom et al., 2006) of the 3D-Coffee program (O'Sullivan et al., 2004 – Poirot et al., 2004). 3D-Coffee has been benchmarked as optimal when sequence identity between target and template falls below 50% (Dalton and Jackson, 2007). The obtained alignment was further improved by hand to define a seed MSA.

In a second step, a HMM profile of the seed MSA was created to screen UniProtKB (The UniProt Consortium, 2008) using HMMsearch (Eddy, 1998) in order to find out suitable homologs sharing sequence identity over at least 67% of the length of the shorter matching protein. We used a threshold of $E = 10\ e^{-15}$ to retrieve all close homologs. These homologs were further clustered using Cd-hit (Li and Godzik, 2006). For each cluster, an automated MSA was built with MUSCLE (Edgar, 2004) and a HMM profile (*HMM_cluster*) was computed (Eddy, 1998). In parallel, another HMM profile was computed for their closest homologous sequences present in the seed alignment (*HMM_seed*). Then, the two profiles *HMM_cluster* and *HMM_seed* were aligned using the HHalign program (Söding, 2005). A stepwise approach allows adding progressively each aligned cluster to the seed alignment. To make this step-up more efficient and safer, we started with highly matching sequences (at least 70% identity), and the whole process was repeated while the identity threshold was progressively decreased at 60, 55, 50, 45, and 40%, respectively. This allowed to exclude a few

unreliable distant sequences and to assort the individual tribes that are part of each aligned cluster. In particular, such a control limited the addition of too many new indels to the reference alignment, preventing a too large increase of the MSA length in the successive outputs. This is a crucial point since many alignment methods are known to make systematic errors in the placement of insertions and deletions that impact further evolutionary analyses (Löytynoja and Goldman, 2008). At that step, we excluded the addition of any amino acid segment that would be responsible of an unbearable lengthening of the initial seed alignment of experimentally studied sequences except if this sequence appears bringing essential biological information.

A further script has been designed to detect the emergence of new homologs each time a new version of the UniProtKB database was published. These probable homologs were assessed and further added to the reference alignment using the HMM stepwise approach described above. Accordingly, we worked at any time with MSA that were always brought up to date.

Reconstructing phylogenetic trees

These last updated MSA were used to derive phylogenetic trees with direct (PhyML version 3.0 (Guindon and Gascuel, 2003)) or approximate (FastTree version 2.1 (Price et al., 2010)) maximum likelihood approaches. The substitution matrix used in PhyML was LG (Le and Gascuel, 2008 – Le et al., 2008), the adaptation of the WAG one (Whelan and Goldman, 2001). Both approaches estimate the likelyhood under a discrete gamma model to account for the different rates of evolution at different sites of the alignement. Moreover, robustness of the reconstructed PhyML trees topologies were further assessed using either a bootstrap approach or a much faster alternative the approximate likelihood-ratio test (aLRT) (Anisimova and Gascuel, 2006). The obtained trees (written in Newick format) were visualized using either MEGA version 5.0 (Tamura et al., 2011) or Dendroscope version 2.0 (Huson et al., 2007) programs.

Annotation by monophyly

Phylogenetic trees contain a few experimentally characterized proteins branching with many unknown sequences. Let's consider two monophyletic subtrees sharing a recent common ancestor. Each time a group of sequences belonging to the first monophyletic subtree contained at least one reliably annotated leaf, while the other one did not, we transferred its functional annotation uniquely to the other leaves of this first subtree. Such a cautionary approach prevents any damaging overinterpretation of functional proximity.

RESULTS

Establishing a reference multiple sequence alignment of cyclic amidohydrolases

With the deluge of sequences released by the ceaseless progress of genomics, the availability of more and more distant homologs makes increasingly difficult to build an accurate MSA and thus a robust phylogeny. To meet this challenge, we set up a two-stage procedure. First, we call for an equilibrated sampling of a small set of representative sequences (seed) that are viewed as sufficiently consistent and biologically significant to reflect the structural and functional diversity of the whole family of cyclic amidohydrolases. Requested seed gene products must have been experimentally studied and their 3D structures published. In a second stage, a controlled automated process adds progressively all detected homologs - including distant ones - to this seed MSA without modifying the initial alignment of well-grounded homologous positions.

As detailed in Material and Methods, this seed alignment was made in two steps. First, the 20 representative sequences were automatically multiply aligned using the program 3D-Coffee (O'Sullivan et al., 2004 – Poirot et al., 2004). Since several important structures were misaligned in the C-terminal part, it appeared necessary to manually curate the automatic MSA to reach an optimal version that was taken as our reference seed alignment. Fig. 1 shows the phylogenetic tree

derived from this seed. This seed tree already shows the large diversity of cyclic amidohydrolases both at the levels of their structures and functions. In total, we have three molecular functions (defined by their respective EC number) that define at least five different cellular functions defined by their respective protein name as listed in Table 1. For instance, a monophyletic subtree contains all sequences labeled with the EC 3.5.2.2, i.e. D-hydantoinases and dihydropyrimidinases and the related DPYL protein devoid of most of the conserved residues that are essential for binding the metal cofactor, strongly suggesting that they act as collapsin response mediator proteins (Goshima et al., 1995). Moreover, the same cellular functions may have rather different evolutionary stories. For instance, the L-hydantoinase appears closer to allantoinases than to the D-hydantoinases. Likewise, as already described by Lohkamp et al. (2006), fungal dihydropyrimidinase appears to be very distant from the other dihydropyrimidinases.

Reconstructing the phylogeny of cyclic amidohydrolases and assessing their annotation

A HMM profile of the resulting seed MSA was further created to find out all suitable homologs in UniProtKB using HMMsearch (Eddy, 1998) and add them to the seed alignment using a progressive step by step approach as detailed in Material and Methods. This process allows the ordered introduction of more and more distantly related sequences into the reference alignment.

Fig. 2 shows the tree obtained after aligning some 2600 homologous sequences belonging to 559 different genera. Since only a very small proportion of these homologs was endowed with a sound functional annotation (e.g. found as experimental evidence as registered in SwissProt), we further reannotated all homologous sequences that have never been experimentally studied by using an approach based on monophyly as described in M&M. Fig. 2 shows a simplified view of the phylogenetic distribution of the different types and classes of all cyclic amidohydrolases after

reannotation by monophyly. The obtained unrooted tree is made of three main monophyletic groups that define three classes of DHOases.

Class II corresponds to the more distant subtree from the trifurcation point (Fig. 2, down) put together all small DHOases that were previously classified as Type IIa byFields et al. (1999). These DHOases II are active and form dimers. Many of them are found in organisms that also contain a larger inactive subunit PyrC associated in multimers with subunit PyrB (the preceding step of the pathway): these inactive paralogous copies are belonging to either class III (dodecamers in many bacteria, mainly proteobacteria), or class I (CAD-like complex in fungi).

The second subtree defining class III (Fig. 2, left) contains a large assortment of DHOases that are found exclusively in Bacteria and belong to various 4D structures. All the homomeric PyrC (dimers classified as Ia by Fields et al. (1999)) are clustered there and they are mixed with the active or inactive subunit PyrC that form multimeric complex with PyrB and that have been classified as Ib by Fields et al. (1999). Interestingly, the large majority of class III *pyrC* genes are in synteny with *pyrB* genes. Thus, this neighborhood does not influence the way the encoded PyrC product will adopt its specific tertiary and/or quaternary structures. In particular, it is not possible to define an ancient unique event separating the choice between building homomers or multimers. Nevertheless, most of the quaternary structures associating PyrB with an inactive PyrC (open circles in Fig. 2) are located in the most recent branches.

The third main monophyletic group (Fig. 2, right) gathers a large array of sequences that cluster in different subtrees endowed with different molecular functions and displaying various 3D and 4D structures. Here, we have four main subdivisions. The first three subtrees correspond to class I DHOases found in archaea (type Iarc), in eukaryotes (types Ieuk) and bacteria (type Ibac), respectively, unifying previous type Ia, Ic and Id separated in the classification of Fields et al.

(1999). Archaeal sequences are mainly paraphyletic: Crenarchaeota are present in a separate branch emerging close to the trifurcation node while Euryarchaeota are diverging at the basis of the monophyletic subtree grouping also Bacteria and Eukaryotes. Remarkably, the newly-defined Thaumarchaeota phylum (Brochier-Armanet et al., 2009) seems devoid of any homologous PyrC sequences. Note also that a few spirochaetes (e.g. Treponema, Leptospira) sequences are branching in the subtree containing Euryarchaeota sequences. Bacterial sequences, including the other spirochaetes, are distributed into monophyletic groups corresponding to Gammaproteobacteria (mainly Pseudomonadaceae), Bacteroidetes, Planctomycetes, and alphaproteobacteria, respectively. Inside Eukaryotes, the dimers (mainly Euglenozoa) are well separated from the CAD composed of two monophyletic groupings corresponding to inactive (Fungi and Choanoflagellata) and active (Metazoa and Amoebozoa) PyrC. Again, inactive structures are found only in the most recent branches.

The last major monophyletic division associated to the class I subtree is made of the other cyclic amidohydrolases that define two main monophyletic subtrees grouping allantoinases and hydantoinases/dihydropyrimidinases, respectively. Interestingly, at the basis of the allantoinases subtree are emerging a few sequences close to the L-hydantoinase (as already found in the seed tree, see Fig. 1). The other subtree contains a mixture of D-hydantoinases (including a monophyletic subgroup of phenylhydantoinases such as HyuA described in E. coli (Kim et al., 2000)) and dihydropyrimidinases mainly found in eukaryotes. Note however that, as observed already for the inactive DHOases, the inactive DHP-like sequences (open circles in Fig. 2) are appearing in the most recent branches of this tree confirming the topology shown in Fig. 1.

Fig. 3 shows that the phylogeny of cyclic amidohydrolases summarized in Fig. 2 is supported by a comparison of their tertiary structure. All cyclic amidohydrolases are made of a small domain

and a TIM barrel, a very common alpha/beta protein fold (Wierenga, 2001 – Nagano et al., 2003 and references inside) with an active site usually involving one or two zinc ions with several highly conserved residues (mainly histidines). Fig. 3 shows a simplified view (adapted from Holm and Sander, 1997) of these 3D structures, where the 8 strands of the TIM barrel are schematized as an octagon and the small domain as a tail. Cyclic amidohydrolases that correspond to the subtype I of the nomenclature defined by Seibert and Raushel (2005) are found to diverge by a limited number of differences in the conserved positions necessary to their molecular function. Inside DHOase type I, bacteria may diverge from archaea and eukaryotes at the level of the first histidine in strand 1 where it is substituted with a glutamine. Note also substitutions at the level of the aromatic amino acid position close of the carboxylated lysine (Kcx) in strand 4, which can be either tyrosine or phenylalanine. The main (and irremediable) difference separating type II from the two other DHOase types is the loss of the small domain. Type III displays two differences: a loss of one of the two Zn ions in part of the species, a substitution of the Kcx (strand 4) for aspartic acid associated with a concomitant presence of glycine as an immediate neighbor and the loss of the nearby aromatic amino acid (tyrosine or phenylalanine) present in all other types. Finally, Fig. 3 confirms the structural proximity of allantoinases and dihydropyrimidinases/hydantoinases with DHOases belonging to class I although they display different molecular and cellular functions. We further look how one can rationalize such phylogenetic vicinity.

Similarities in the chemistry used by successive homologous enzymes belonging to parallel metabolic pathways

While dihydroorotases catalyze the third step of the biosynthesis of pyrimidines, allantoinases and dihydropyrimidinases are involved in the catabolism of purines and pyrimidines, respectively. Fig. 4 shows a high similarity in the chemistry of the different cyclic amidohydrolases analyzed in Figs 2 and 3: the chemical structures of the substrates look comparable, the reaction mechanisms

appear similar and they produce related components. For instance, the carbamoylaspartate (substrate of DHOases) is highly similar to the N-carbamoyl-beta-aminoisobutyrate (product of thymine degradation) and the N-carbamoyl-beta-amino acid (product of uracil degradation). Moreover, Fig. 5 shows a strong similarity in the chemical reactions made by the enzymes working in the next step of pyrimidine metabolism in both directions (synthesis and degradation, respectively) but not in the case of purine degradation. Although the transformation of allantoin to ureidoglycine via allantoate is mechanistically similar to that of dihydropyrimidine to N-carbamoyl-beta-amino acid, the previous step, degradation of xanthine to allantoin via uric acid is chemically unrelated. In the case of antiparallel ways of pyrimidine metabolism, the dihydroorotate is transformed in orotate by the dihydroorotate dehydrogenase DHODase (EC 1.3.3.1) in a very similar process of the transformation of uracil or thymine in dihydrouracil or dihydrothymine by their respective dihydropyrimidine dehydrogenase DHPDase (EC 1.3.1.2). Since the consecutive steps of carbamoylation (PyrC/PydB) and of redox reaction (PyrD/PydA) are made by homologous enzymes, we call *reaction module* such a set of successive enzymes involved in parallel pathways, sharing a comparable chemical reaction on similar compounds.

To improve our knowledge of the evolutionary mechanisms leading to such modules, we further looked at the evolutionary relationships between DHODases and DHPDases.

The same methodological approach described above in the case of cyclic amidohydrolases, was used to perform an accurate seed alignment of the amino acid sequences of PyrD (EC 1.3.3.1), and PydA (EC 1.3.1.2) and PreA (EC 1.3.1.1) that have been crystallized, before adding progressively all sound homologs found in the present (summer 2011) public databases, to get an optimal MSA. Fig. 6 shows the topology of the phylogenetic tree obtained from this MSA. First, the obtained tree confirms that PyrD homologs are clustering in two main subtrees, corresponding to

the cytoplasmic type DHOD 1 and the membrane-bound type DHOD 2 (Björnberg et al., 1997). Moreover, DHOD 1 sequences can be further separated in two monophyletic subclasses. First, we have the emergence of a minority of PyrD subunits structured as homodimers defining a subtree containing all DHOD 1A. The majority of subunits PyrD forming heterotetramers with the subunit PyrK define the subtree DHOD 1B that share a common ancestor with four other subtrees: (i) the variant DHOD 1S made of PyrD forming heterotetramers with a subunit analog to PyrK (but with very low sequence similarity), first described in the archaeon *Sulfolobus solfataricus* (Sørensen and Dandanell, 2002); (ii) its sister subtree contains three monophyletic groups corresponding to all the known PydA and PreA forming heterotetramers with PydX and PreT respectively; (iii) diverging before these dihydropyrimidinases (DHPDases), we found two groups of unknown dehydrogenases provisionally called DHase X1 and DHase X2, respectively. Thus, this tree confirms the close relationships between DHODases and DHPDases and supports our introduced concept of reaction module.

We further considered the putative function of these unknown DHases by looking at the gene context of their respective encoding genes. DHases X2 form a large group of homologs that diverged more recently than DHase X1 from DHPDases. We found that its encoding gene present in 69 species belonging to nearly all bacterial phyla is often close to a gene annotated as encoding a pyruvate-ferredoxin oxidoreductase. Moreover, in 13 out of these 69 species, we found also as immediate neighbor to this pyruvate-ferredoxin oxidoreductase a gene homolog to *preT*, encoding the ferredoxin part of the complex PreA-PreT of the *E. coli* DHPase (Hidese et al., 2011). In the other X2 species, this preT-like gene is present but not in the same transcriptional unit containing DHase X2 encoding gene. *E coli* contains four *preT* paralogs but only one copy of *preA*. By analogy, one can guess that X2 could be the partner of one of these *preT* paralogs and form a

complex with the pyruvate-ferredoxin oxidoreductase to work in a dehydrogenation of an unknown component that may be similar to dihydropyrimidines.

Fig. 7 summarizes the gene context of DHase X1 homologs in various organisms that are often thermophiles (T) and sometimes pathogens (P). Inset A shows as an example the full context in *Petrotoga mobilis*. Inset B details the homology relationships between these neighbouring genes in the different species by taking as reference the *E. coli* data although there is – paradoxically – no DHase X1 homolog in this model organism. Interestingly, the genes encoding these DHase X1 are belonging to transcriptional units containing the homologs of *hyuA (ygeZ), ygeW, ygeY*, *ygfL, xdhA, xdhB, xdhC, ygfU* and *yqeA*, coding a hydantoinase (appearing in the tree of cyclic amidohydrolases in Fig. 2), the hypothetical carbamoyltransferase YgeW recently studied and crystallized (Li et al., 2011), the uncharacterized peptidase M20 YgeY, the uncharacterized metal dependent hydrolase SsnA, the three subunits (XdhA, XdhB, and XdhC) of xanthine deshydrogenase, a xanthine/uracil permease and the carbamate kinase-like protein YgeA, respectively.

Thus, many of the genes products transcribed with DHase X1are involved more or less directly in various steps of purine catabolism. This gene context suggests that we found an indirect link between the evolution of enzymes involved in pyrimidine and purine metabolism.

DISCUSSION

The superfamily of cyclic amidohydrolases displays several important features detailed below: their specific analyses required meeting several challenges that could help to improve our general understanding of the mode of evolution of enzymes and the way metabolic pathways have evolved. The first challenge to meet was a methodological one. Since such superfamilies contain both significant proportions of very close and of very distant amino acid sequences, we devised a way to

optimize the multiple alignment of all homologous sequences in order to reconstruct a phylogenetic tree that accurately reflect their actual evolutionary relationships. As previously discussed by Beiko et al (2005), there is no existing algorithm allowing building a MSA where mathematical or statistical optimality and biological optimality would be equivalent. In order to align distant homologous proteins with highest possible biological optimality, we started from a small set of homologs limited to those where the 3D structure is known and we manually curated it. Then, we devised a stepwise approach to progressively add all the homologs (mainly obtained from never studied organisms) while maintaining the high quality of the seed alignment (Fig. 1). Such a gradual approach helped to detect – and correct - numerous errors in the functional annotation of proteins that are based uniquely on sequence identity as recorded in public databases (Furnham et al., 2009 – Schnoes et al., 2009). Accordingly, it also improved the quality of the structure-function relationships of the different homologs (Cantarel et al., 2006).

The obtained phylogenetic tree (Fig. 2) identifies three main subtrees of dihydroorotases and helps to simplify their previous classification (Fields et al., 1999) while clarifying their evolutionary relationships. Whereas DHO II contains uniquely dimers of the small-size dihydroorotases, DHO III and DHO I are made of sequences that differ in their structures 3D and 4D. For instance, DHO III is a mixture of homodimers of PyrC subunits, and of dodecamers made of PyrB and PyrC subunits. Likewise, DHO I contains homodimers in archaea and in bacteria, while in complex eukaryotes we have the evolved hexameric multifunctional fusion CAD.

Strikingly, there is no direct relationship between the ranges of sequence identity and of divergence in cellular functions played by homologous cyclic amidohydrolases. The allantoinases that are involved in purine degradation share a rather recent common ancestor with the dihydropyrimidinases that catalyze one step of pyrimidine degradation. This monophyletic subtree (Fig. 2) is branching inside the main subtree containing also dihydroorotases of class I. Although all

dihydroorotases are involved in step 3 of pyrimidine biosynthesis and display the same activity (same EC number), the evolutionary distances separating the three classes are far greater than the distance separating class 1 DHO from the other cyclic amidohydrolases that are doing different catalyses.

To explore this rather surprising result we further analyzed the different constituents of the biochemical process used by these different homologs in their respective pathways. The importance of chemistry in the evolution of catalysis has been previously reviewed by Glasner et al. (2006). Indeed, it has long been recognized that many proteins retain the ability to catalyze similar chemical transformations on numerous substrates, defined as substrate ambiguity by Jensen (1976). Moreover, shared chemistry (substrate ambiguity) seems to be a major mechanism to allow divergence of most enzyme families (Khersonsky et al., 2006, Khersonsky et al., 2011). We further introduced in this paper the concept of reaction module that combines the importance of shared chemistry with the homology of the enzymes defining at least two successive steps that are chemically similar in parallel pathways.

Indeed, we first noticed (Fig. 5) that cyclic amidohydrolases use similar chemical reactions (illustrated by close EC numbers). Moreover, their respective substrates appear to be also chemically similar, as well as their respective products (Fig. 4). We further look at the enzymes using these respective products made by the cyclic amidohydrolases as their substrates and found that at least in the case of pyrimidine metabolism, they are themselves homologs and doing strongly similar catalysis (Fig. 5). As shown on Fig. 6, these redox enzymes DHODases (anabolic way) and DHPDases (catabolic way) are also forming another multifunctional family where DHPDases are closer to one class of DHODases than are the different classes of DHODases.

We interpret this finding as being the evidence of the existence of reaction modules and as demonstrating its importance in the analysis of the evolution of metabolism. In biology, the concept

of modularity has a long history (see for instance Caetano-Anollés et al., 2009 and numerous references inside). In comparative anatomy, structural modules representing the parts of an organism, usually at the adult stage, have been discussed since Cuvier and Saint-Hilaire two centuries ago (as summarized in Pereira-Leal et al., 2006). More recently, Riley and Labedan (1997) have underlined the importance of identifying structural segments of homology to trace back protein evolution and better assess their functional annotation. Finally, functional modules have been proposed to be repeatedly used to build metabolic pathways (Hartwell et al., 1999 - Ravasz et al., 2002). We further propose that reaction modules would be the elementary bricks used to build these functional modules during the evolution of metabolic pathways when connections are established by the product of one enzyme being the substrate of the next enzyme in the cascade. Duplicating and reusing such elementary bricks would make possible to create a large array of possible pathways as already proposed by Jensen (1976). James and Tawfik (2003) have proposed that primordial enzymes were promiscuous, with high conformational diversity allowing a large functional diversity. As they further noticed, this reminds of ancient models proposed independently by Landsteiner (1936) and Pauling (1940) suggesting that functional diversity could go far beyond sequence diversity.

Our results suggest that primordial DHOase was a very promiscuous enzyme able to work with a large array of components (Fig. 4) and to interact with primordial DHODase that took delivery of its product. This primordial element of pathway (i.e. primordial reaction module) would have been duplicated, the successive copies of the respective ancestral genes maintaining their mechanistic links while progressively adapting to more and more specific substrates. Since such copies of ancestral cyclic amidohydrolases must remain connected with copies of the neighboring step of the same reaction module, one expect that they were less free to diverge that the homologous copies that were evolved to other kind of protein interactions. This could explain why the distance

separating DHO I from ALL, HYD and DHP, is shorter than the distances between the three DHO classes (Fig. 2). Indeed, DHO I that is still present in the three domains of life would have been the ancestral form from which DHO II and DHO III diverged significantly after adapting to two different major events: the loss of the small domain in the case of DHO II, and the interaction of PyrC subunit with PyrB in the case of DHO III. Finally, the loss of enzymatic activity of PyrC in some dodecameric PyrB-PyrC complexes is probably even more recent since it appears in the most lately diverging branches (Fig. 2).

This modular approach helped us to rationalize at least the evolutionary links between DHOases and DHPases. The relations with allantoinases are more indirect but the occurrence of the X1 homologs in the PyrD – PydA tree (Fig. 6) and their location in a genetic context specific to purine degradation (Fig. 7) are already an indirect evidence of intricate relationships between purine and pyrimidine metabolisms. This last point will be developed in the companion paper.

ACKNOWLEDGEMENTS. This work was funded by the CNRS (UMR 8621). Matthieu Barba is supported by a doctoral grant from the French Ministère de la Recherche.

LITERATURE CITED

Anisimova M., O. Gascuel, 2006. Approximate likelihood ratio test for branchs: A fast, accurate and powerful alternative. Systematic Biology, 55:539-552

Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V, Notredame C. 2006. Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee. Nucleic Acids Res. 34:W604-608

Berg, S. T., and D. R. Evans. 1993. Subunit structure of class A aspartate transcarbamoylase from Pseudomonas fluorescens. Proc. Natl. Acad. Sci. USA 90:9819–9822

Björnberg, O., Rowland, P., Larsen, S., and Jensen, K. F. 1997. Active site of dihydroorotate dehydrogenase A from *Lactococcus lactis* investigated by chemical modification and mutagenesis. Biochemistry 36, 16197–16205.

Brichta DM, Ph. D. Dissertation, Univ. of. North Texas, 2003

Brichta DM, Azad KN, Ralli P, O'Donovan GA. 2004. *Pseudomonas aeruginosa* dihydroorotases: a tale of three pyrCs. Arch Microbiol. 182:7-17

Brochier-Armanet C, Boussau B, Gribaldo S, Forterre P. 2008. Mesophilic Crenarchaeota: proposal for a third archaeal phylum, the Thaumarchaeota. Nat Rev Microbiol. 6:245-252

Caetano-Anollés G, Yafremava LS, Gee H, Caetano-Anollés D, Kim HS, Mittenthal JE. 2009. The origin and evolution of modern metabolism. Int J Biochem Cell Biol. 41:285-297.

Cantarel BL, Morrison HG, Pearson W. 2006. Exploring the relationship between sequence similarity and accurate phylogenetic trees. Mol Biol Evol. 23:2090-2100

Dalton JA, Jackson RM. 2007 An evaluation of automated homology modelling methods at low target template sequence similarity. Bioinformatics 23:1901-1908.

Dutta S, Burkhardt K, Young J, Swaminathan GJ, Matsuura T, Henrick K, Nakamura H, Berman HM. 2009. Data deposition and annotation at the worldwide protein data bank. Mol Biotechnol. 42:1-13

Eddy SR. 1998. Profile hidden Markov models. Bioinformatics 14:755-763

Edgar, R.C. 2004. MUSCLE: a multiple sequence alignment method with reduced time and space complexity. BMC Bioinformatics, 5:113

Engelhardt BE, Jordan MI, Muratore KE, Brenner SE. 2005. Protein Molecular Function Prediction by Bayesian Phylogenomics. PLoS Comput Biol 1:e45

Engelhardt BE, Jordan MI, Brenner SE. 2006. A graphical model for predicting protein molecular function. Proc 23rd Intl Conf Machine Learning 2006:038.1–8

Engelhardt BE, Jordan MI, Repo ST, Brenner SE. 2009. Phylogenetic molecular function annotation. J Phys. 180:12024

Evans, D. R., and Guy, H. I. 2004. Mammalian pyrimidine biosynthesis: fresh insights into an ancient pathway. J. Biol. Chem. 279:33035–33038.

Fields, C., Brichta, D., Shepherdson, M., Farinha, M. and O'Donovan, G.A. 1999. Phylogenetic analysis and classification of dihydroorotases: a complex history for a complex enzyme. Paths to Pyrimidines. An International Newsletter 7:49-63.

Furnham N, Garavelli JS, Apweiler R, Thornton JM. 2009. Missing in action: enzyme functional annotations in biological databases. Nat Chem Biol. 5:521-5

Glasner ME, Gerlt JA, Babbitt PC. 2006. Evolution of enzyme superfamilies. Curr Opin Chem Biol. 10:492-497

Goshima Y., Nakamura F., Strittmatter P., Strittmatter S.M. 1995. Collapsin-induced growth cone collapse mediated by an intracellular protein related to UNC-33. Nature 376:509-514

Guindon S, Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol. 52:696–704.

Hartwell LH, Hopfield JJ, Leibler S, Murray AW. 1999. From molecular to modular cell biology. Nature. 402:C47-52

Holm, L. and Sander, C. 1997. An Evolutionary Treasure: Unification of a broad set of amidohydrolases related to urease. Proteins 28:72-82.

Hughes AL 2005. Gene duplication and the origin of novel proteins. Proc Natl Acad Sci U S A. 102:8791-8792

Huson DH, Richter DC, Rausch C, Dezulian T, Franz M, Rupp R. 2007. Dendroscope: An interactive viewer for large phylogenetic trees. BMC Bioinformatics 8:460

James LC, Tawfik DS. 2003. Conformational diversity and protein evolution--a 60-year-old hypothesis revisited. Trends Biochem Sci. 28:361-368

Jensen RA: 1976. Enzyme recruitment in evolution of new function. Annu Rev Microbiol 30:409-425.

Khersonsky O, Tawfik DS. 2010. Enzyme promiscuity: a mechanistic and evolutionary perspective. Annu Rev Biochem. 79:471-505

Khersonsky O, Malitsky S, Rogachev I, Tawfik DS. 2011. Role of chemistry versus substrate binding in recruiting promiscuous enzyme functions. Biochemistry. 50:2683-2690

Kim KM, Caetano-Anollés G. 2010. Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. Mol Biol Evol. 27:1710-1733.

Kim GJ, Lee DE, Kim HS. 2000. Functional expression and characterization of the two cyclic amidohydrolase enzymes, allantoinase and a novel phenylhydantoinase, from Escherichia coli. J Bacteriol. 182:7021-7028

Koski, L.B. and Golding, G.B. 2001. The closest BLAST hit is often not the nearest neighbor. J. Mol. Evol. 52:540–542.

Kumar S, Dudley J, Nei M, Tamura K. 2008. MEGA: A biologist-centric software for evolutionary analysis of DNA and protein sequences. Briefings in Bioinformatics 9:299-306.

Landsteiner, K. 1936. The Specificity of Serological Reactions, reprinted 1962, Dover Publications

Le SQ, Gascuel O. 2008. An Improved General Amino Acid Replacement Matrix. Mol Biol Evol, 25:1307-1320.

Le SQ, Lartillot N, Gascuel O. 2008. Phylogenetic mixture models for proteins. Phil. Trans. R. Soc. B 363:3965–3976

Li W & Godzik A. 2006. Cd-hit: a fast program for clustering and comparing large sets of protein or nucleotide sequences Bioinformatics 22:1658-1659

Li Y, Jin Z, Yu X, Allewell NM, Tuchman M, Shi D. 2011. The ygeW encoded protein from Escherichia coli is a knotted ancestral catabolic transcarbamylase. Proteins. 79:2327-2334

Liu A, Li T, and Fu R. 2007. Amidohydrolase Superfamily. In: Encyclopedia of Life Sciences. John Wiley & Sons, Ltd: Chichester

Lohkamp B, Andersen B, Piskur J, Dobritzsch D. 2006. The crystal structures of dihydropyrimidinases reaffirm the close relationship between cyclic amidohydrolases and explain their substrate specificity. J Biol Chem. 281:13762-13776

Löytynoja A, Goldman N. 2008. Phylogeny-aware gap placement prevents errors in sequence alignment and evolutionary analysis. Science. 320:1632-1635

McNeely K, Xu Y, Ananyev G, Bennette N, Bryant DA, Dismukes GC. 2011. Synechococcus sp. strain PCC 7002 nifJ mutant lacking pyruvate:ferredoxin oxidoreductase. Appl Environ Microbiol. 77:2435-2444

Nagano N, Orengo CA, Thornton JM. 2003. One fold with many functions: the evolutionary relationships between TIM barrel families based on their sequences, structures and functions. J Mol Biol. 321:741-65.

Nam SH, Park HS, Kim HS. 2005. Evolutionary relationship and application of a superfamily of cyclic amidohydrolase enzymes. Chem Rec.5:298-307.

O'Brien PJ, Herschlag D: 1999. Catalytic promiscuity and the evolution of new enzymatic activities. Chem Biol 6:R91-R105.

O'Sullivan,O. Suhre K, Abergel C, Higgins DG, Notredame C. 2004. 3DCoffee: combining protein sequences and structures within multiple sequence alignments. J. Mol. Biol., 340:385–395.

Ohno, S. 1970. Evolution by gene duplication. Springer-Verlag

Pauling, L. 1940. A theory of the structure and process of formation of antibodies. J. Am. Chem. Soc. 62, 2643–2657

Pereira-Leal JB, Levy ED, Teichmann SA. 2006. The origins and evolution of functional modules: lessons from protein complexes. Philos Trans R Soc Lond B Biol Sci. 361:507-517.

Poirot O, Suhre K, Abergel C, O'Toole E, Notredame C. 2004. 3DCoffee@igs: a web server for combining sequences and structures into a multiple sequence alignment. Nucleic Acids Res. 32:W37-40

Price MN, Dehal PS, Arkin AP. 2010. FastTree 2—approximately maximum-likelihood trees for large alignments. PLoS ONE 5:e9490.

Ravasz E, Somera AL, Mongru DA, Oltvai ZN, Barabási AL. 2002. Hierarchical organization of modularity in metabolic networks. Science **297**:1551-1555

Riley M, Labedan B. 1997. Protein evolution viewed through Escherichia coli protein sequences: introducing the notion of a structural segment of homology, the module. J Mol Biol. **268**:857-868.

Schnoes AM, Brown SD, Dodevski I, Babbitt PC. 2009. Annotation error in public databases: misannotation of molecular function in enzyme superfamilies. PLoS Comput Biol. 5:e1000605

Seibert CM, Raushel FM. 2005. Structural and catalytic diversity within the amidohydrolase superfamily. Biochemistry **44**:6383-6391

Shannon P, Markiel A, Ozier O, Baliga NS, Wang JT, Ramage D, Amin N, Schwikowski B, Ideker T. 2003. Cytoscape: a software environment for integrated models of biomolecular interaction networks. Genome Res. **13**:2498-2504.

Shimodaira H, and Hasegawa M. 1999. Multiple Comparisons of Log-Likelihoods with Applications to Phylogenetic Inference Mol Biol Evol **16**: 1114-1116

Shoaf, W. T. & Jones, M. E. 1971. Initial steps in pyrimidine synthesis in Ehrlich ascites carcinoma. Biochem. Biophys. Res. Commun. **45**:796-802.

Söding J. 2005. Protein homology detection by HMM-HMM comparison. Bioinformatics 21:951-960

Sørensen G, Dandanell G. 2002. A new type of dihydroorotate dehydrogenase, type 1S, from the thermoacidophilic archaeon Sulfolobus solfataricus. Extremophiles, 6, 245-251

Tamura K, Peterson D, Peterson N, Stecher G, Nei M, and Kumar S (2011) MEGA5: Molecular Evolutionary Genetics Analysis using Maximum Likelihood, Evolutionary Distance, and Maximum Parsimony Methods. Molecular Biology and Evolution

The UniProt Consortium. 2008. The Universal Protein Resource (UniProt). Nucleic Acids Res. 36:D190-D195

Wierenga R K. 2001. The TIM-barrel fold: a versatile framework for efficient enzymes. FEBS Lett. 492:192–198

Whelan S, Goldman N. 2001. A general empirical model of protein evolution derived from multiple protein families using a maximum-likelihood approach. Mol Biol Evol. 18:691–699.

Zhang P, Martin PD, Purcarea C, Vaishnav A, Brunzelle JS, Fernando R, Guy-Evans HI, Evans DR, Edwards BF. 2009. Dihydroorotase from the hyperthermophile Aquifex aeolicus is

activated by stoichiometric association with aspartate transcarbamoylase and forms a one-pot reactor for pyrimidine biosynthesis. Biochemistry 48:766-778.

Protein name	UniProt	EC	PDB	Species name	Length
Dihydroorotase	P05020 (PYRC_ECOLI)	3.5.2.3	2EG6	*Escherichia coli*	348
Dihydroorotase	P06204 (PYRC_SALTY)	3.5.2.3	3JZE	*Salmonella typhimurium*	348
Dihydroorotase	Q0PBP6 (PYRC_CAMJE)	3.5.2.3	3PNU	*Campylobacter jejuni*	335
Dihydroorotase	O66990 (PYRC_AQUAE)	3.5.2.3	3D6N	*Aquifex aeolicus*	422
Dihydroorotase	P65907 (PYRC_STAAW)	3.5.2.3	3GRI	*Staphylococcus aureus*	424
Dihydroorotase	P96081 (PYRC_THEAQ)	3.5.2.3	2Z00	*Thermus aquaticus*	426
Dihydroorotase	Q81WF0 (PYRC_BACAN)	3.5.2.3	3MPG	*Bacillus anthracis*	428
Dihydroorotase	Q7MVW1_PORGI)	3.5.2.3	2GWN	*Porphyromonas gingivalis*	449
Allantoinase	P77671 (ALLB_ECOLI)	3.5.2.5	3E74	*Escherichia coli*	453
Allantoinase	Q9KAH8 (ALLB_BACHD)	3.5.2.5	3HM7	*Bacillus halodurans*	438
L-hydantoinase	P81006 (HYDL_ARTAU)	3.5.2.2	1GKR	*Arthrobacter aurescens*	458
D-hydantoinase	Q45515 (HYDA_BACST)	3.5.2.2	1K1D	*Bacillus stearothermophilus*	471
D-hydantoinase	Q8VTT5 (HYDA_BURPI)	3.5.2.2	1NFG	*Burkholderia pickettii*	457
D-hydantoinase	Q0PQZ5 (Q0PQZ5_RHIME)	3.5.2.2	3DC8	*Bacillus sp. AR9*	484
D-hydantoinase	Q7SIE9 (Q7SIE9_THESP)	3.5.2.2	1GKP	*Thermus sp.*	458
Dihydropyrimidinase	Q55DL0 (DPYS_DICDI)	3.5.2.2	2FTW	*Dictyostelium discoideum*	503
Dihydropyrimidinase	Q14117 (DPYS_HUMAN)	3.5.2.2	2VR2	*Homo sapiens*	519
Dihydropyrimidinase	Q9P903 (DPYS_SACKL)	3.5.2.2	2FVK	*Saccharomyces kluyveri*	542
DHP related protein	Q16555 (DPYL2_HUMAN)	unsure	2GSE	*Homo sapiens*	572
DHP related protein	P97427 (DPYL1_MOUSE)	unsure	1KCX	*Mus musculus*	572

Table 1. The list of the crystallized cyclic amidohydrolases that have been multiply aligned to create the seed of the exhaustive and updated MSA of all homologs.

FIGURE LEGENDS

Figure 1: The phylogenetic tree derived from the seed alignment of cyclic amidohydrolases.
The sequences listed in Table 1 have been multiply aligned as described in the text and a tree has been derived from this MSA using PhyML. The confidence limits for each node were further estimated using the approximate likelihood-ratio test (aLRT) (Anisimova and Gascuel, 2006).

Figure 2: The phylogenetic tree of cyclic amidohydrolases.
This is the simplified view of the tree obtained with FastTree. The confidence limits for the roots of each monophyletic subtree were computed using the Shimodaira-Hasegawa test (1999). Squared roman numbers define the three classes of dihydroorotases defined in this work. The scheme in the left corner gives the correspondence between our new classification and the previous one proposed by Fields et al. (1999).The 4D structures determined for the members belonging to the different monophyletic subtrees are schematized in front of each subtree.

Figure 3: Comparison of the tertiary structure of cyclic amidohydrolases.
Tertiary structures have been simplified as in Holm and Sander (1997) and drawn on the schematized topology of their phylogenetic tree. The substitutions of important residues in typical motifs in each strand of the TIM barrel and the loss of the small domain by DHO II are indicated by thick black arrows. The variations in the aromatic residue in strand 4 of DHO I class are shown by dotted arrows. Gray arrow points to the sometimes missing Zn ion.

Figure 4: Comparing the chemistry of the chemical reactions catalyzed by the cyclic amidohydrolases.
The chemical structures of the substrate and product of each enzyme are superposed to underline their respective similarities.

Figure 5: Comparing the parallel steps catalyzed by the cyclic amidohydrolases.
The chemical structures of the substrate and product of each enzyme are aligned to underline their respective similarities in the respective step catalyzed by the set of homologous enzymes. The reaction modules described in the text are framed in a gray rectangle.

Figure 6: The phylogenetic tree of DHODases/DHPDases.
This is the simplified view of the tree obtained with FastTree.

Figure 7: The context of genes encoding X1 unknown sequences.
A. Genetic neighborhood in Petrotoga mobilis as schematized in MetaCyc.
B. Table of the X1 neighborhood in various organisms using the E. coli nomenclature.

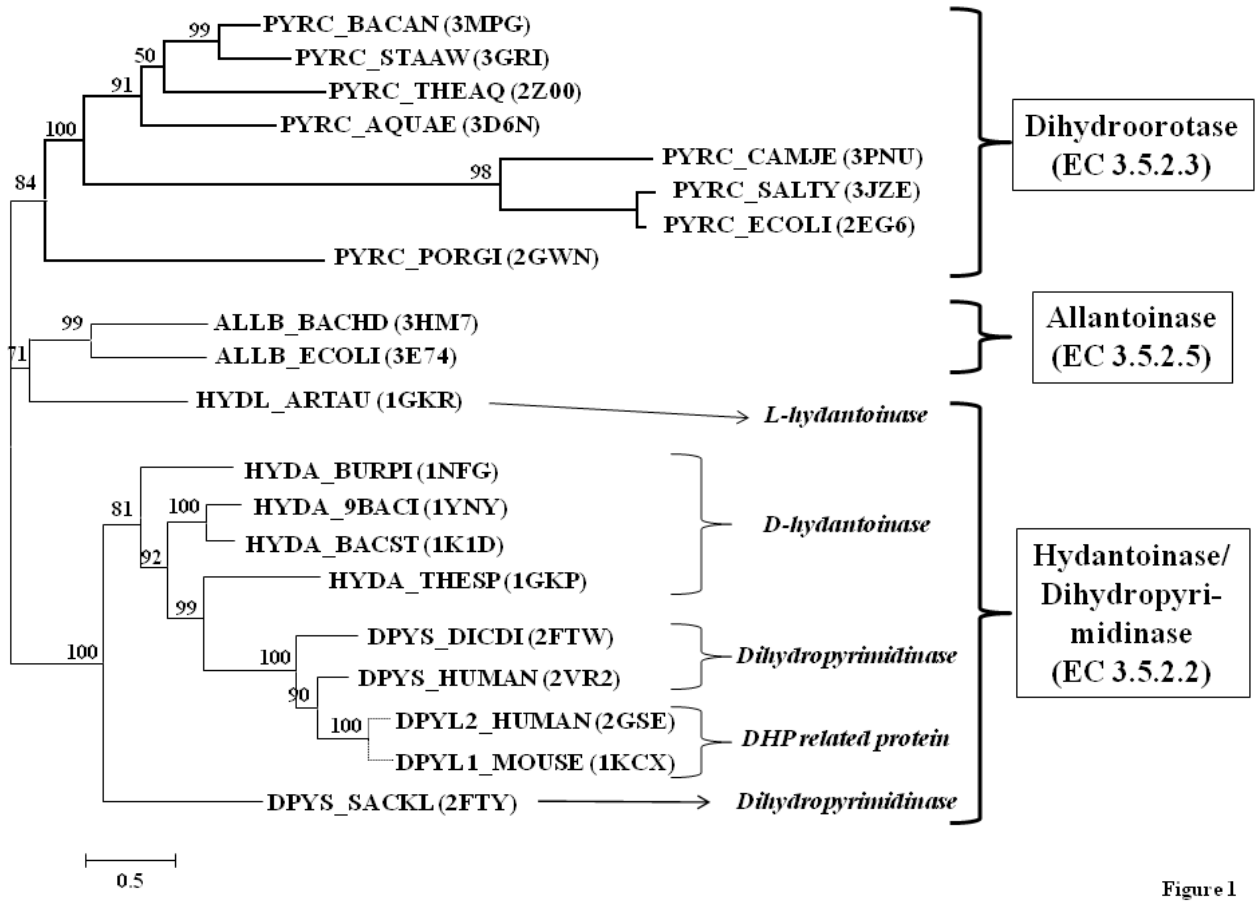

PYRC_BACAN (3MPG)
PYRC_STAAW (3GRI)
PYRC_THEAQ (2Z00)
PYRC_AQUAE (3D6N)
PYRC_CAMJE (3PNU)
PYRC_SALTY (3JZE)
PYRC_ECOLI (2EG6)
PYRC_PORGI (2GWN)
Dihydroorotase (EC 3.5.2.3)
ALLB_BACHD (3HM7)
ALLB_ECOLI (3E74)
Allantoinase (EC 3.5.2.5)
HYDL_ARTAU (1GKR)
L-hydantoinase
HYDA_BURPI (1NFG)
HYDA_9BACI (1YNY)
HYDA_BACST (1K1D)
HYDA_THESP (1GKP)
D-hydantoinase
DPYS_DICDI (2FTW)
DPYS_HUMAN (2VR2)
Dihydropyrimidinase
DPYL2_HUMAN (2GSE)
DPYL1_MOUSE (1KCX)
DHP related protein
DPYS_SACKL (2FTY)
Dihydropyrimidinase
Hydantoinase/ Dihydropyri-midinase (EC 3.5.2.2)
0.5

Figure 1

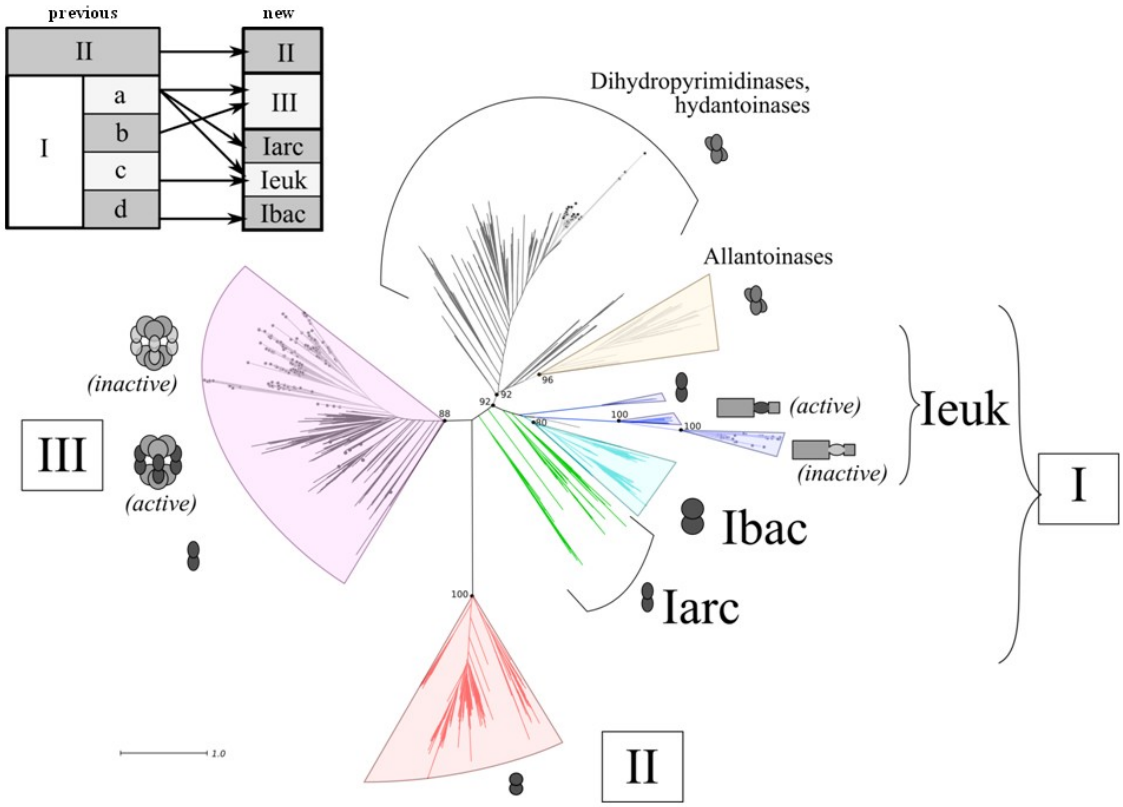


Figure 2

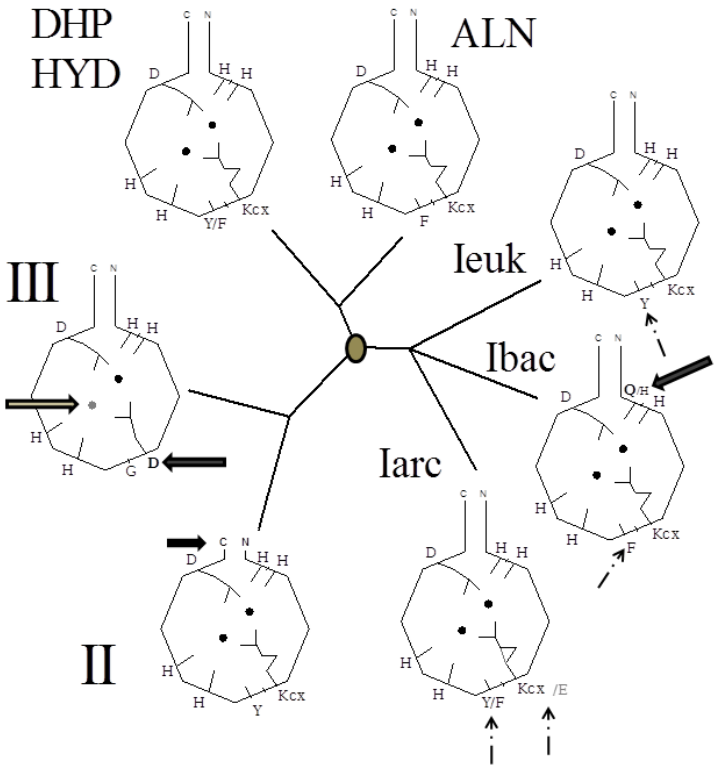


Figure 3

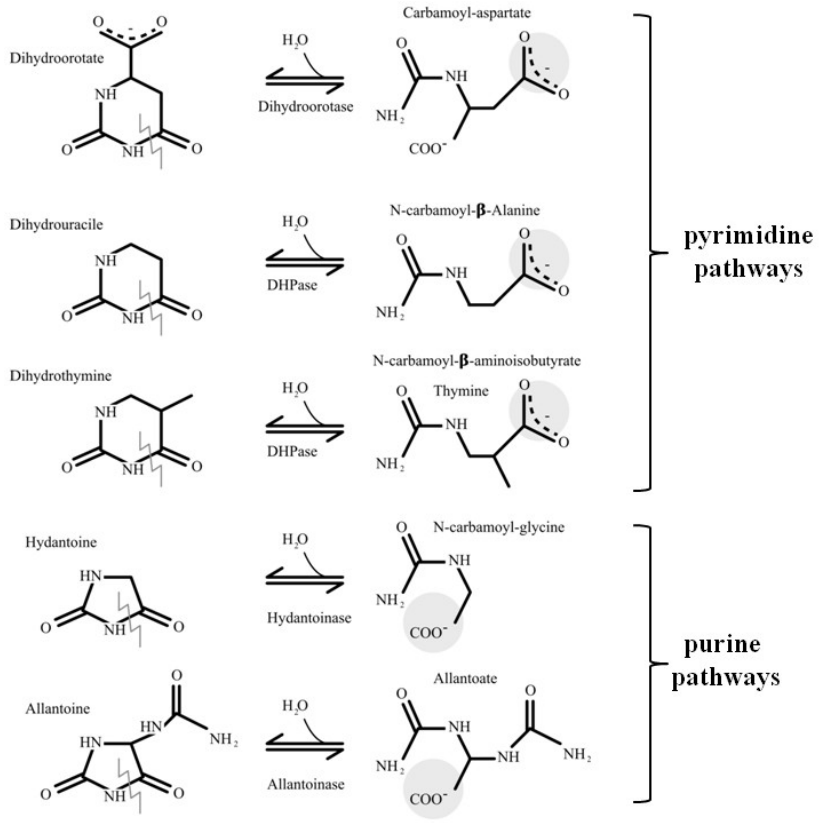


Figure 4

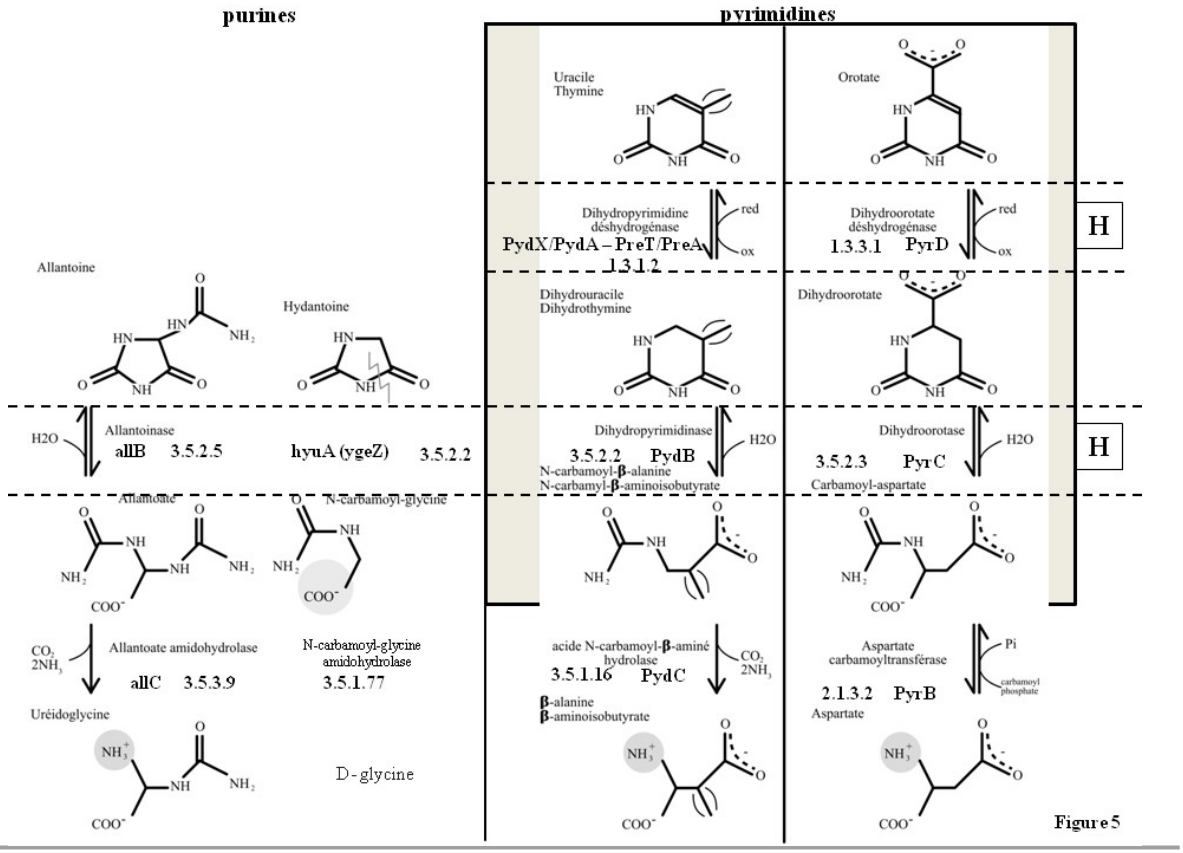


Figure 5

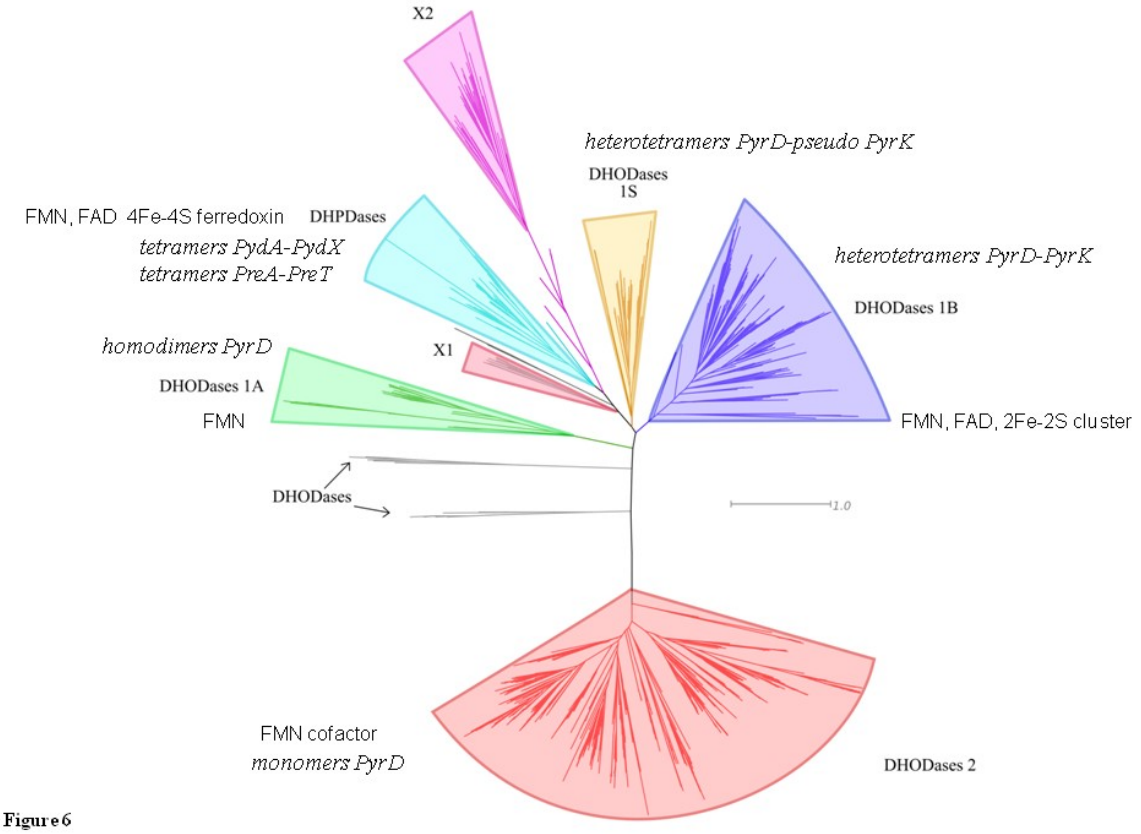


Figure 6

A

B

	DHase	YgeZ	YgeW	CKase	YgeY	SsnA	XDH	Perméase
Escherichia coli	-	hyuA	ygeW	yqeA	ygeY	ssnA	yes	ygfU
Petrotoga mobilis	**A9BIU8**	A9BIT5	A9BIT6	-	A9BIT4	A9BIU2	yes	A9BIT1
Kosmotoga olearia	**C5CHS6**	C5CHS4	C5CHS3	-	C5CHS1	C5CHS5	yes	-
Spirochaeta smaragdinae (1)	**E1R8T3**	E1R8T2	E1R8S8	-	-	-	yes	
Spirochaeta smaragdinae (2)	-	-	E1R4A0	-	E1R493	E1R492	yes	-
Clostridium difficile (1)	**C9XR99**	C9XR98	-	-	-	C9XRA1	-	C9XRA0
Clostridium difficile (2)	**C9XJU2**	C9XJU8	-	-	-	-	yes	C9XJU0
Halanaerobium hydrogeniformans	**E4RJU5**	E4RK16?	E4RJV6	-	E4RJV7	E4RJV1	yes	E4RJU3
Thermosphaera aggregans	**D5U0C9**	D5U113?	D5U0D6	D5U0D3	-	D5U0D1	-	-
Staphylothermus marinus	**A3DL41**	A3DKS9?	A3DL30	A3DL38	-	A3DL39	-	-
Moorella thermoacetica	-	Q2RGZ6	Q2RGZ7	-	-	-	yes	-
Nocardioides sp.	-	A1SH63	A1SH60	-	-	A1SH57	yes	-
Sebaldella termitidis	-	D1AIF8	D1AID8	D1AID9	D1AID7	D1AIE6	-	D1AIE1

Figure 7

Chapitre 7 : Retour sur les carbamoyltransférases et identification de nouvelles activités

1 Comprendre les relations d'homologie entre les enzymes impliquées dans le métabolisme des purines et des pyrimidines

Dans le chapitre précédent (Chapitre 6), nous avons introduit le concept de module réactionnel pour proposer une explication logique à l'homologie trouvée entre des enzymes homologues effectuant des réactions semblables successives dans les voies antiparallèles de biosynthèse et dégradation des pyrimidines. Bien que l'on trouve des relations d'homologie entre au moins deux enzymes de voies d'utilisation des purines et des pyrimidines, nous n'avons pas trouvé de relation logique directe mais avons identifié des relations indirectes, en particulier en étudiant le module DHOase/DHPase – DHODase/DHPDase. Nous observons que certains homologues à fonction inconnue ont un contexte génétique homologue au contexte des gènes *yge* chez *E. coli* qui code en particulier pour une carbamoyltransférase associée à une carbamate kinase et qui seraient impliquées dans la dégradation de l'allantoïne.

Dans ce chapitre (qui sera ultérieurement écrit pour soumission au journal MBE de manière conjointe avec le manuscrit du précédent chapitre), nous regardons plus en détail le métabolisme des purines et nous réexaminons la phylogénie de l'étape précédente ATCase – DHOase en se focalisant sur les carbamoyltransférases à fonction inconnue trouvée dans des organismes peu ou pas connus expérimentalement.

2 Une carbamoyltransférase parmi les autres ?

La famille des N-carbamoyltransférases (TCases) qui catalysent toutes le transfert d'un groupe carbamoyl d'un carbamoyl-phosphate sur le groupement amine d'un composé chimique, généralement un acide aminé, a été très étudiée, en particulier dans notre groupe en collaboration avec le groupe de Nicolas Glansdorff ([Labedan et al. 1999 ; Labedan et al. 2004]).

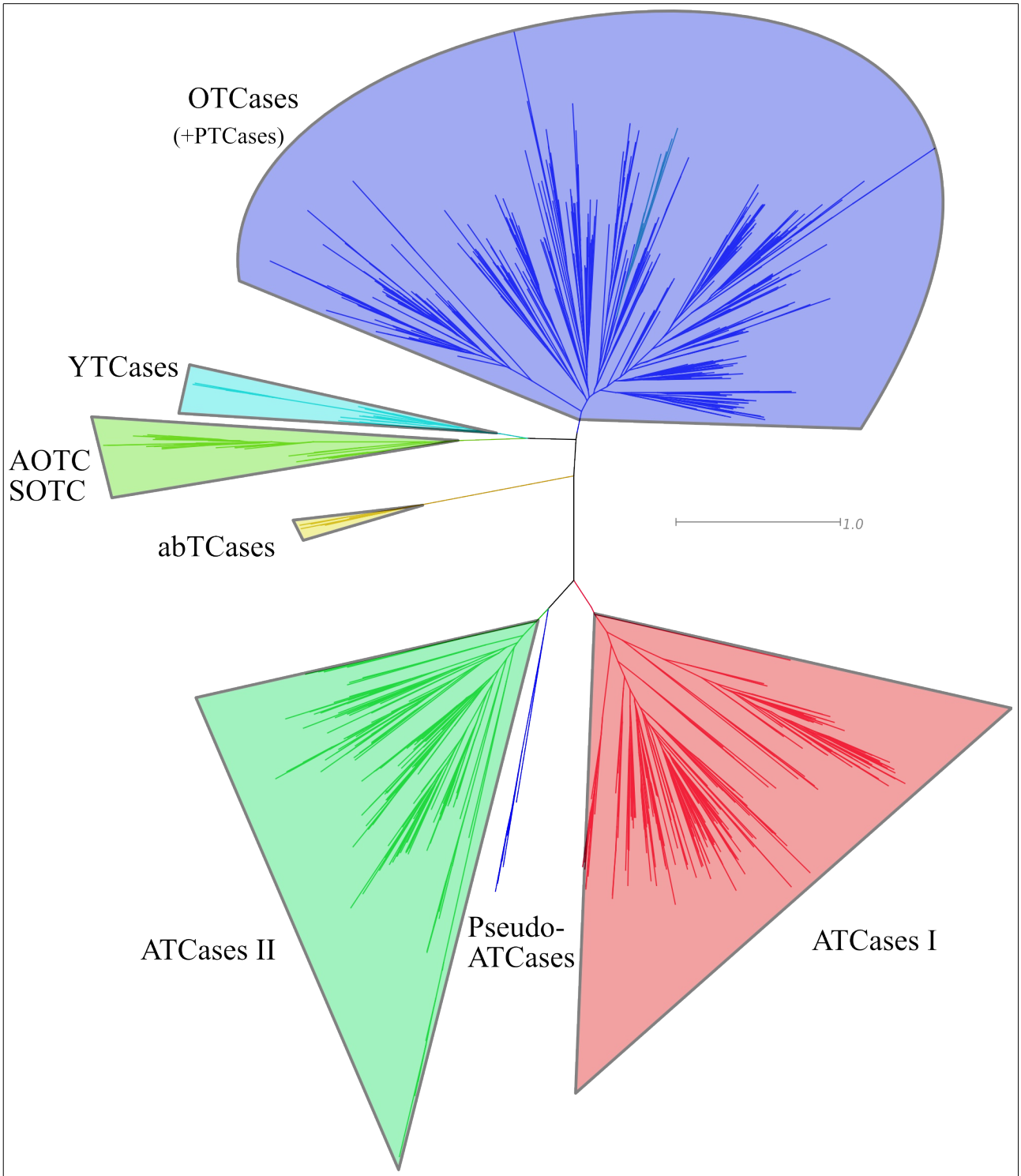


Figure 11 : Arbre phylogénétique non enraciné des carbamoyltransférases.
ATCases : aspartate carbamoyltransférases (types I et II)
Pseudo-ATCases : TCases paralogues proches des ATCases
AOTCases/SOTCases : acétyl/succinyl-ornithines carbamoyltransférases
abTCases : carbamoyltransférases dans des voies de biosynthèse d'antibiotiques
OTCases : ornithines carbamoyltransférases
PTCases : putrescine carbamoyltransférases
YTCases : carbamoyltransférases inconnues (également notées UTCases)

On peut distinguer plusieurs groupes de TCases, aux niveaux biochimiques et phylogénétiques (Figure 11) : 1) les ATCases qui catalysent la deuxième étape de la voie de biosynthèse des pyrimidines. 2) Les OTCases (et PTCases) qui interviennent dans le métabolisme de l'arginine (biosynthèse et dégradation). 3) Les AOTCases et SOTCases qui sont aussi impliquées dans le métabolisme de l'arginine, mais phylogénétiquement éloignées des OTCases. 4) Les protéines produites par les gènes de type *zwa6* (appelées ici abTCases) qui interviennent dans la biosynthèse de la Zwittermicine A (un antibiotique produit par des *Bacillus*) et d'autres antibiotiques [Emmert et al. 2004]. 5) Un dernier groupe de carbamoyltransférases (notées YTCases), à fonction cellulaire non identifiée, se trouve à la base du groupe A/SOTCases. L'une d'entre elles est codée par le gène *ygeW* d'*E. coli* (d'où le Y).

De plus, dans le tout dernier arbre construit par notre groupe, nous avons mis en évidence un nouveau sous-groupe de carbamoyltransférases jusque là annotées comme des ATCases, et que nous allons étudier plus en détail.

3 Les ATCases et les pseudo-ATCases

Précédemment, nous avions montré que les ATCases se distinguent en deux grands groupes : I et II qui correspondent en particulier à des structures quaternaires différentes [Labedan et al. 2004]. Elles partagent cependant des motifs spécifiques au niveau des résidus qui se lient au substrat, notamment les quatre résidus His265, Pro266 et Pro268 (numérotation de l'ATCase de *E. coli*). Or, dans un sous-arbre, proche de la racine des ATCases II (Figure 11), on trouve des homologues que nous allons appeler des pseudo-ATCases. En effet, elles sont toutes annotées comme ATCases dans les bases de données publiques mais possèdent des motifs structuraux différents des « vraies » ATCases (Tableau 14). Les résidus R230 et Q232 qui permettent notamment la fixation du groupe carboxyl du dihydroorotate : sont substitués chez les pseudo-ATCases par des alanine, glycine, serine ou cystéine. Les résidus spécifiques à l'activité carbamoyltransférase (présents dans toute la famille) sont par contre conservés dans tous ces groupes (STRT 53-56 *E. coli*, HPxQ 135-138 *E. coli* notamment).

Les organismes contenant ces pseudo-ATCases possèdent également dans leur génome un gène codant une copie d'ATCase « vraie », ce qui conforte leur statut hypothétique de nouvelles TCases à fonction inconnue.

Le sous-arbre de pseudo-ATCases montre que l'on peut les différencier en trois groupes

correspondant à trois motifs différents (Tableau 14).

	His265-P268 Ecoli	R230-Q232 Ecoli (motif ATCase)
AOTCase, SOTCase	HCLP	
YTCase	HCLP, HVLP, HALP	
OTCase	HCLP	
abTCase	HDLP	
ATCase (I & II)	HP[LG]P	R . Q
pseudo-ATCase 1	HPLA	AIA, AIS, SIA
pseudo-ATCase 2	H[ST]LP	G . [SC]
pseudo-ATCase 3	HSLP	V . P

Tableau 14: Motifs conservés chez les carbamoyltransférases.
*Les nombres des résidus correspondent à l'ATCases d'*E. coli*. Le second motif (R . Q) est caractéristique des ATCases. Les résidus entre crochets sont plusieurs possibilités pour un même emplacement, et un point désigne un résidu non conservé.*

4 Description des sous-familles de pseudo ATCases

L'arbre de ces pseudo-ATCases enraciné par la plus proche ATC II (celle de *Pyrococcus abyssi*) permet de distinguer trois sous-types qui diffèrent par leurs motifs structuraux Tableau 14) et leur contexte génétique (Figure 12). Voyons-les plus en détail.

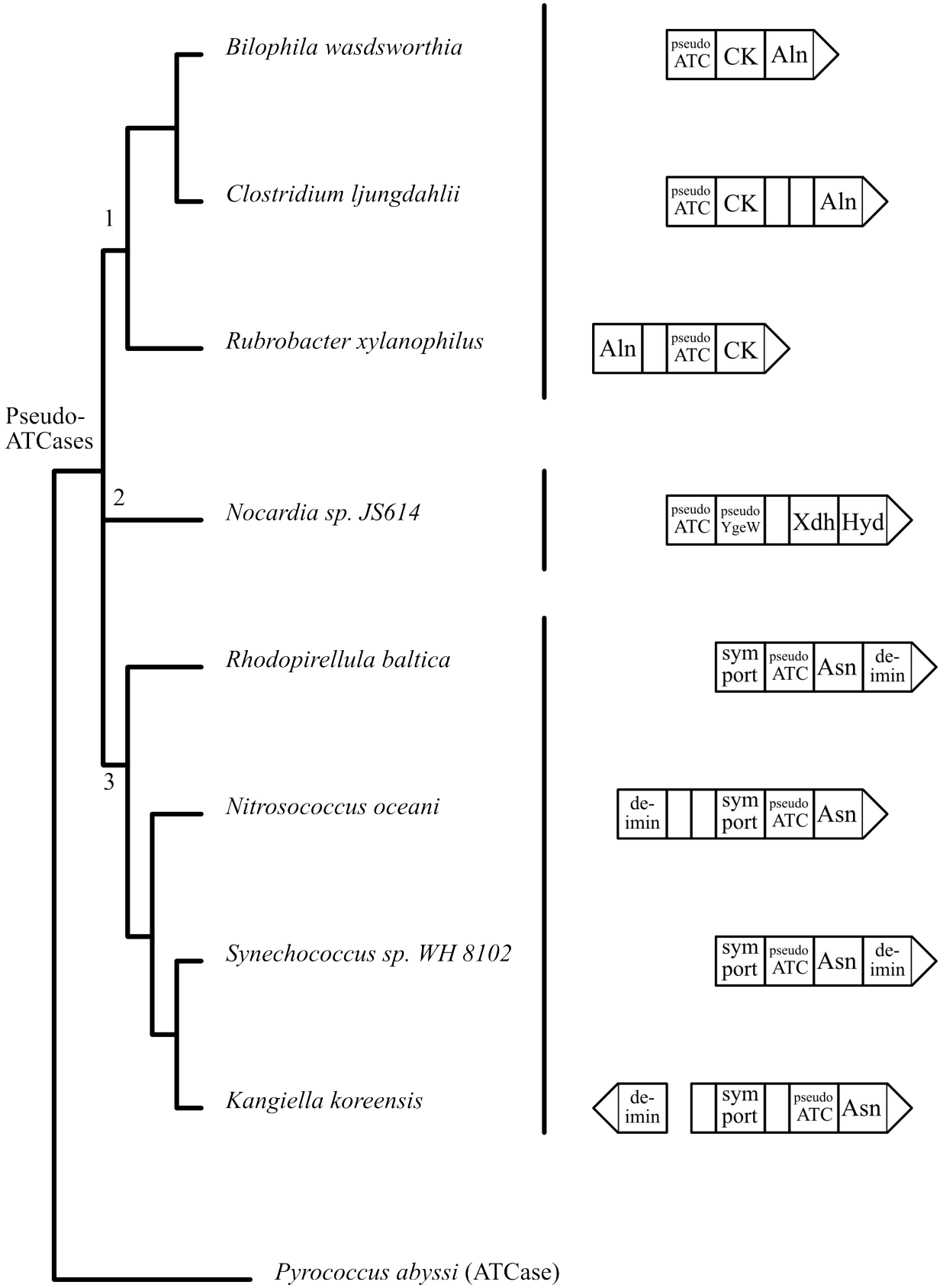


Figure 12 : Arbre des pseudo-ATCases et leurs contextes génétiques.

4.1 Pseudo-ATCases 1

6 séquences sont disponibles sur Uniprot, dont 2 issues de génomes complets. La pseudo-ATCase de *Rubrobacter xylanophilus* (Q1AS69) est présente dans un contexte génétique très intéressant, car il contient tous les gènes codant les premières étapes de la voie de dégradation des purines jusqu'à la production de l'allantoate (Figure 13, Rxyl_2840 à Rxyl_2850), ainsi qu'une perméase permettant d'importer des purines, en formant une même unité de transcription.

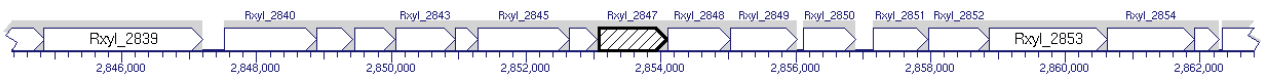


Figure 13 : Contexte génétique de la pseudo ATCase de *Rubrobacter xylanophilus*.
Liste des annotations fonctionnelles :
Rxyl_2840 : uracil/xanthine perméase
Rxyl_2841 : OHCU décarboxylase
Rxyl_2842 : CMP/dCMP déaminase
Rxyl_2843 : uricase
Rxyl_2844 : 5HIU hydrolase
Rxyl_2845 : allantoïnase
Rxyl_2846 : non identifié
Rxyl_2847 : pseudo-ATCase
Rxyl_2848 : carbamate kinase
Rxyl_2849 : asparaginase/glutaminase
Rxyl_2850 : régulateur transcriptionnel
(D'après MetaCyc [Karp et al. 2002])

Dans cette unité de transcription, le gène Rxyl_2847 codant la pseudo-ATCase est suivi immédiatement de Rxyl_2848 codant une carbamate kinase. Un tel voisinage suggère un rôle catabolique pour cette TCase, contrairement à ce qu'on attendrait d'une ATCase (anabolique) qui est généralement couplée à une CPSase. En amont de cette unité se trouve des gènes de la xanthine déshydrogénase (Rxyl_2836-2839) et une hypoxanthine phosphoribosyltransférase (Rxyl_2835), alors qu'en aval de cette unité de transcription on trouve également un autre groupe contenant des gènes de la voie de dégradation du glyoxylate en D-glycérate (Rxyl_2851 à Rxyl_2854). Les gènes codant les enzymes qui dégradent l'allantoate en glyoxylate n'ont par contre pas été détectés dans le génome de *R. xylanophilus*, que ce soit par les voies rejetant de l'urée ou de l'ammoniac : (allantoïcase, allantoate amidohydrolase...).

Dans tous les autres génomes appartenant à cette sous-famille, (*Bilophila wadsworthia, Clostridium ljungdahlii*), on trouve aussi cette pseudo-ATCase en compagnie d'une CKase et d'une allantoïnase, mais pas avec les groupes de gènes responsables de la dégradation des purines en

allantoïne.

4.2 Pseudo-ATCases 2

Une autre pseudo-ATCase est trouvée uniquement chez l'Actinobactérie *Nocardioides sp. JS614*. Elle présente des motifs de liaison au substrat (Tableau 14), et un contexte génétique différents de toutes les autres pseudo-ATCases et branche entre les deux autres sous-familles (Figure 12). De plus et remarquablement, son gène codant est au voisinage immédiat d'un homologue de *ygeW*) et de gènes codants les sous-unités de la xanthine déshydrogénase. Les *Nocardioides* ont la particularité de dégrader de nombreux produits de synthèse qui sont polluants, comme l'atrazine ou le chlorure de vinyle (un précurseur de PVC). Il est donc plausible que les gènes codant les deux TCases et leurs voisins fassent partie d'une voie de dégradation de *Nocardioides* non identifiée.

4.3 Pseudo-ATCases 3

Dans le troisième sous-type, qui possède des motifs de liaison au substrat qui diffère des ATCases classiques, toutes les espèces partagent un même voisinage génétique : pseudo-ATCase 3 + Asparagine synthétase + N-carbamoyl-L-acide aminé amidohydrolase + symport Na+ (qui est quelquefois annoté comme symport pour l'urée). Les espèces qui contiennent ces ATCases appartiennent à des groupes taxonomiques divers : Planctomycetes (*Rhodobacter baltica*) cyanobactéries, et divers ordres de Gammaprotéobactéries (Chromatiales, Oceanospirillales, Alteromonadales). Remarquablement, tous ces organismes ont la particularité d'avoir adopté un habitat marin.

5 À la recherche des gènes codant l'oxamate carbamoyltransférase

Jusqu'à récemment on pensait que le gène codant l'enzyme OxTCase était du type *ygeW* d'*E. coli*. L'activité OxTCase a en effet été mesurée chez *E. coli* et quelques autres espèces, et le génome d'*E. coli* code cette TCase hypothétique en plus d'une ATCase et d'une OTCase qui ont chacune été bien caractérisées structurellement et fonctionnellement (activité, structures 3D et 4D, etc...). On la trouve dans un groupe monophylétique de l'arbre des TCases distinct des autres TCases connues (le plus proche étant les AOTCases). Cette YgeW est présente chez différentes souches d'*Escherichia coli*, de *Clostridium* et d'*Enterococcus faecalis*. Le gène *ygeW* lui-même est dans un contexte génétique qui contient notamment une carbamate kinase, compagnon essentiel de l'activité OxTCase, et les trois gènes codant la xanthine déshydrogénase. De nombreuses recherches se sont

donc focalisées sur ce gène *ygeW* mais sans succès. Très récemment, le groupe de Tuchman [Li et al. 2011] a cristallisé cette protéine YgeW et confirmé qu'elle a la structure typique d'une TCase. Cependant, ils n'ont pu détecter aucune activité enzymatique en prenant l'oxalurate ou de nombreux autres composés comme substrat. De même, Polo, Fita & Rubio [2010] ont décrit dans une communication au dernier colloque ICAP (Washington, 2010) la cristallisation d'une autre TCase inconnue présente chez *E. faecalis*. Bien que cette protéine présente une structure typique de carbamoyltransférase, aucun des nombreux substrats potentiels testés n'a permis d'identifier son activité moléculaire. La fonction de ces YTCases demeure inconnue [Li, Weinstock & Murray 1995].

Vu ses propriétés putatives, nous avons alors postulé que la pseudo-ATCase de *Rubrobacter* pourrait exprimer cette activité OxTCase. Le gène candidat Rxyl_2847 possède en effet des propriétés intéressantes comme le montre la Figure 12 : Son plus proche voisin code une carbamate kinase (comme dans le cas de *ygeW*), et son unité de transcription contient l'ensemble des gènes de dégradation des purines, de l'hypoxanthine à l'allantoate. De plus, d'un point de vue chimique, les ATCases se distinguent des autres TCases connues par la carbamoylation de son amine en α, alors que pour les OTCases, PTCases, AOTCases, SOTCases elle se fait en δ (bout de groupe latéral de l'acide aminé). La carbamoylation de l'oxamate se fait lui aussi sur l'amine en α, ce qui justifierait la position phylogénétique de cette TCase parmi les ATCases bien identifiées.

Afin de valider expérimentalement cette hypothèse, nous avons entamé une collaboration avec le laboratoire de Christianne Legrain (Institut de Recherches Microbiologiques J.-M. Wiame, Bruxelles) qui a une longue expérience des enzymes extrêmophiles pour étudier les propriétés du produit du gène Rxyl_2847 de *Rubrobacter xylanophilus DSM 9941*. Cette bactérie est remarquable sur plusieurs points : elle se branche à la base du sous-arbre des Actinobactéries, elle est extrêmement résistante aux radiations, tout en étant thermophile (optimum de croissance à 60 °C) ; enfin, elle peut dégrader de l'hémicellulose et du xylane (d'où son nom). Son génome a été séquencé en 2006 mais sa biologie n'est pas connue. Ses propriétés thermophiles et le fait qu'elle a été peu étudiée rendent les expérimentations plus compliquées, d'une part pour la mise en culture, et d'autre part pour l'étude des réactions dont certains produits de la voie qui nous intéresse sont instables à haute température (en particulier le carbamoyl-phosphate). Les nombreuses mesures d'activité enzymatique sur des extraits bruts de *Rubrobacter* ayant cru en présence d'allantoïne n'ont pas été concluantes, les faibles activités détectées disparaissant après dialyse des extraits. De plus, l'incubation en présence de carbamoyl-phosphate pendant 10 min à 55°C, d'un filtrat d'extrait brut

non-dialysé, obtenu par ultrafiltration sur une membrane filtrant au seuil de 10 kDa, produit un composé carbamoylé qui n'est pas observé avec les extraits de la bactérie cultivée en absence d'allantoïne. Enfin, pour contourner les problèmes dus à l'instabilité du carbamoyl-phosphate (dégradation à 50% environ après 5 min à 50°C), des mesures ont été effectuées dans le sens catabolique (arsénolyse de l'oxalurate). De nouveau, aucune activité significative n'a pu être enregistrée dans ces conditions qui permettent des temps d'incubation beaucoup plus longs et l'utilisation d'extraits plus concentrés.

En fait, ces résultats négatifs, bien que décevants à première vue, nous ont permis de progresser. La structure 3D de la protéine Rxyl_2847 a été modélisée par Bauvois et Legrain (communication personnelle) en se basant sur la structure expérimentalement obtenue (PDB : 1ML4) de son homologue le plus proche, l'ATCase de l'archée hyperthermophile *Pyrococcus abyssi*. Cette ATCase a été cristallisée avec un analogue de substrat N-Phosphonacetyl-L-aspartate (PALA) à la place de l'aspartate / carbamoyl-aspartate [Boxstael et al. 2003]. La comparaison des structures montre que les résidus se liant au substrat sont tous conservés à l'exception des R229 et Q231 (de *Pyrococcus*). Ceux-ci se lient aux deux oxygènes du groupe carboxyl en β de l'aspartate (Figure 14), mais ils sont remplacés par une glycine et une sérine, respectivement, ce qui ouvre la poche et permet l'accès à un substrat plus volumineux que l'aspartate ou que l'oxamate (qui est encore moins volumineux).

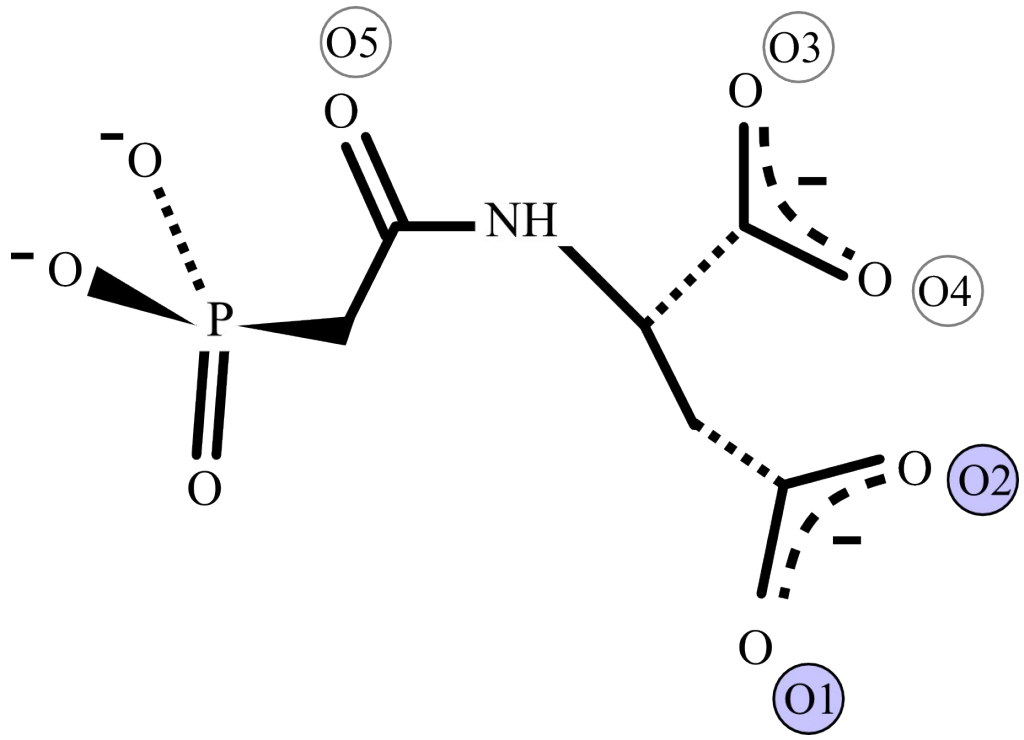


Figure 14 : Structure de la N-Phosphonacetyl-L-aspartate (PALA).
Les oxygènes 1 et 2 mis à évidence sont liés à l'ATCase.

Le modèle semble également confirmer que cette TCase est structurellement une ATCase puisqu'elle agit sur le groupement amine du carbone α, comme pour l'aspartate, et contrairement aux autres TCases expérimentalement étudiées (en δ pour OTCase, PTCase, A/SOTCases (cf Annexe p 156)).

6 Une nouvelle carbamoyltransférase dans la voie de dégradation des purines

Sur la base de ces résultats d'identification de l'activité de la protéine Rxyl_2847 , nous avons réexaminé le contexte génétique de la voie de dégradation des purines chez *Rubrobacter* sous un nouvel angle. En effet, des pseudo-ATCases du type de *Rubrobacter* sont toutes en association avec une carbamate kinase d'une part, ce qui laisse supposer un rôle catabolique, et proches du gène codant l'allantoïnase d'autre part, qui suggère un rôle dans la voie de dégradation des purines. En outre, la présence, chez *Rubrobacter xylanophilus*, des gènes responsables des étapes de dégradation conduisant au glyoxylate, suggère que le groupe de gènes dont fait partie la pseudo-ATCase de *Rubrobacter xylanophilus* (TCase + CKase + allantoïnase) permettrait de dégrader l'allantoïne en glyoxylate. De plus, les différents gènes connus codant les étapes de décomposition

de l'allantoate vers le glyoxylate n'ont par ailleurs pas été trouvés dans le génome de *R, xylanophilus* : les recherches par BLAST dans le génome ne détectent aucune séquence suffisamment proche pour les qualifier d'homologues, que ce soit pour les étapes effectuées par l'allantoïcase ou l'allantoate amidohydrolase d'une part, ni celles effectuées par une uréidoglycolatase et une uréiodoglycolate amidohydrolase, d'autre part. Examinons plus en détail la validité de cette hypothèse en reprenant notre concept de module réactionnel.

7 Des étapes parallèles dans les voies métaboliques des purines et des pyrimidines

Afin de proposer une voie qui intégrerait ces réactions, nous avons comparé la voie de dégradation des purines avec les voies de biosynthèse des pyrimidines et de dégradation (réductrice) des pyrimidines (Tableau 15) avec qui elle possède des similarités de réaction et de substrat, ainsi que des enzymes homologues pour ces réactions semblables.

L'enchaînement des réactions de la voie de biosynthèse des pyrimidines est bien conservé chez toutes les espèces, avec la création d'un carbamoyl-phosphate par une carbamoyl-phosphate synthétase (CPSase), parfois remplacée par une carbamate kinase, suivie du transfert de ce groupement carbamoyl sur l'aspartate par l'ATCase. Vient ensuite la dihydroorotase (DHOase) qui cyclise la carbamoyl-aspartate en dihydroorotate. Enfin, le dihydroorotate est réduit en orotate par la dihydroorotate déshydrogénase (DHODase).

Il existe par contre trois voies connues de dégradation des pyrimidines qui ont la particularité de dégrader tant l'uracile et la thymine avec les mêmes enzymes : une voie oxydante (via le barbiturate, utilisée par quelques Bactéries) [Soong et al. 2002], une voie Rut (chez *E. coli* et d'autres Protéobactéries) découverte récemment [Parales & Ingraham 2010] et une voie réductrice, la mieux étudiée, qui dégrade l'uracile et la thymine en leurs béta-acide aminés respectifs. Cette dernière voie possède des réactions en sens contraire aux étapes similaires de la voie de biosynthèse : d'abord une déshydrogénase (DHPDase), homologue de la DHODase (voir le chapitre précédent), transforme thymine et uracile en dihydro-thymine et dihydro-uracile, respectivement. Puis une amidohydrolase (DHPase), homologue de la DHOase et de l'allantoïnase, ouvre le cycle des dihydropyrimidines en leur N-carbamoyl-béta acide aminé associé. La dernière étape de décarbamoylation diffère en revanche de la voie de biosynthèse : le groupement carbamoyl est hydrolysé irréversiblement en CO_2 et NH_3. Cette étape est réalisée par une déiminase, qui est elle

même homologue de l'allantoate amidohydrolase. Cette dernière fait justement partie des activités manquantes chez *Rubrobacter xylanophilus* pour la dégradation des purines.

Activité	Dégradation des purines	Dégradation des pyrimidines	Biosynthèse des pyrimidines
déshydrogénase	-	DHPDase	DHODase
Amidohydrolase cyclique	ALNase	DHPase	DHOase
(dé)carbamoylation	Allantoate amidohydrolase	N-carbamoyl-β-alanine déiminase	ATCase + CKase/CPSase

Tableau 15 : Similarité des réactions des voies de dégradation des purines et de dégradation et biosynthèse des pyrimidines.
Les déshydrogénases et amidohydrolases sont homologues entre elles, tandis que seules l'allantoate amidohydrolase et la N-carbamoyl-β-alanine déiminase sont homologue pour l'étape de décarbamoylation.

Outre la présence de différentes enzymes homologues qui catalysent des réactions similaires, l'enchaînement de ces différentes réactions est également conservé, de sorte qu'on observe par exemple l'enchaînement d'une CKase avec une TCase, aussi bien dans un sens anabolique (ATCase) que catabolique (OxTCase). Cet enchaînement TCase + CKase se retrouve également dans les voies métaboliques de l'arginine, dans laquelle on retrouve la plupart des autres TCases dont l'OTCase qui peut jouer un rôle aussi bien biosynthétique que catabolique, avec cependant des copies de gènes différentes, une pour chaque rôle.

8 L'hypothèse d'une uréidoglycine carbamoyltransférase

Nous proposons qu'une TCase hypothétique catalyserait l'étape suivant directement la production d'allantoïnase, en remplaçant l'allantoate amidohydrolase, c'est-à-dire :

Allantoate + Pi → Uréidoglycine + carbamoyl-phosphate (uréidoglycine carbamoyltransférase)

au lieu de :

Allantoate + H_2O → Uréidoglycine + CO_2 + 2 NH_3 (allantoate amidohydrolase)

Cette enzyme serait fonctionnellement une uréidoglycine carbamoyltransférase. Cette réaction s'inscrirait dans l'enchaînement de la voie de dégradation des purines, en remplacement de l'allantoate amidohydrolase, directement après l'allantoïnase avec laquelle cette pseudo-ATCase est

voisine (dans le génome). Au lieu de rejeter du CO_2 et 2 NH_3, la réaction permettrait de récupérer du carbamoyl-phosphate qui sera par la suite dégradé par la carbamate kinase.

Cet enchaînement serait l'équivalent de celui en sens opposé des premières étapes de la voie de biosynthèse des pyrimidines (Figure 16). La première étape de dégradation à partir de l'allantoïne hydrolyse cette dernière en allantoate comme la troisième étape de biosynthèse ferme le cycle de la carbamoyl-aspartate en dihydroorotate. La seconde étape transfère le groupement carbamoyl de l'allantoate sur un phosphate, en donnant de l'uréidoglycine, comme la seconde étape de biosynthèse qui transfert le groupement carbamoyl d'un carbamoyl-phosphate sur le groupe aminé d'une aspartate. Enfin, la troisième étape de dégradation hydrolyserait le carbamoyl-phosphate en rejetant CO_2, NH_3 et un ATP, ce qui est l'exact opposé de la voie de biosynthèse qui produit un carbamoyl-phosphate au prix de l'hydrolyse d'un ATP.

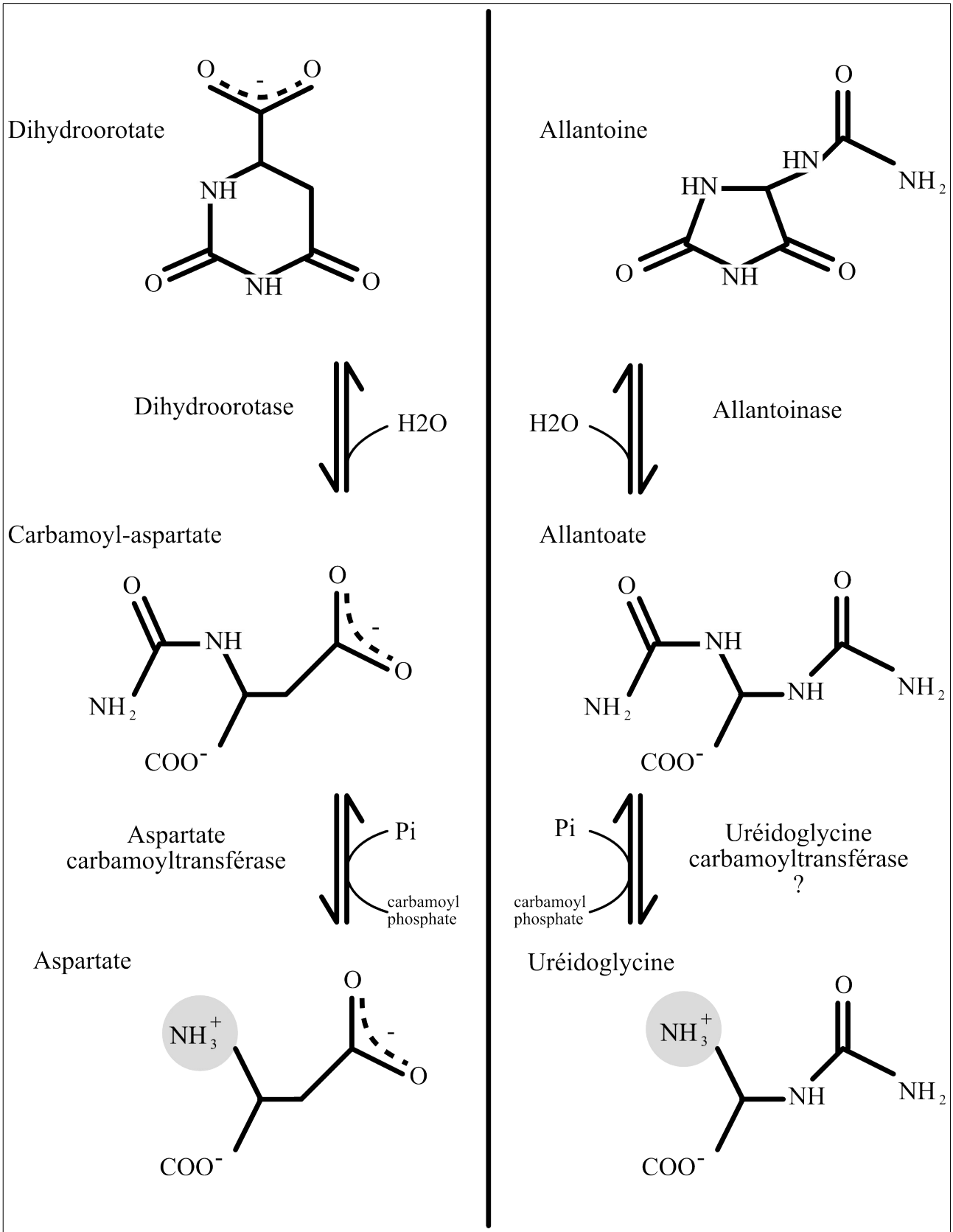


Figure 15 : Comparaison des réactions de carbamoylase et de l'amidohydrolase cyclique associée dans les voies de biosynthèse des pyrimidines (gauche) et la voie hypothétique de dégradation des purines (droite).

La décarbamoylation de l'allantoate présente de fortes similitudes avec celle du carbamoyl-aspartate : le groupe aminé sur lequel se fait la carbamoylation est en α du groupe carboxyl, alors qu'elle se fait en δ sur les autres TCases (du métabolisme de l'arginine) : OTCase, PTCase, AOTCase, SOTCase (Figure 12).

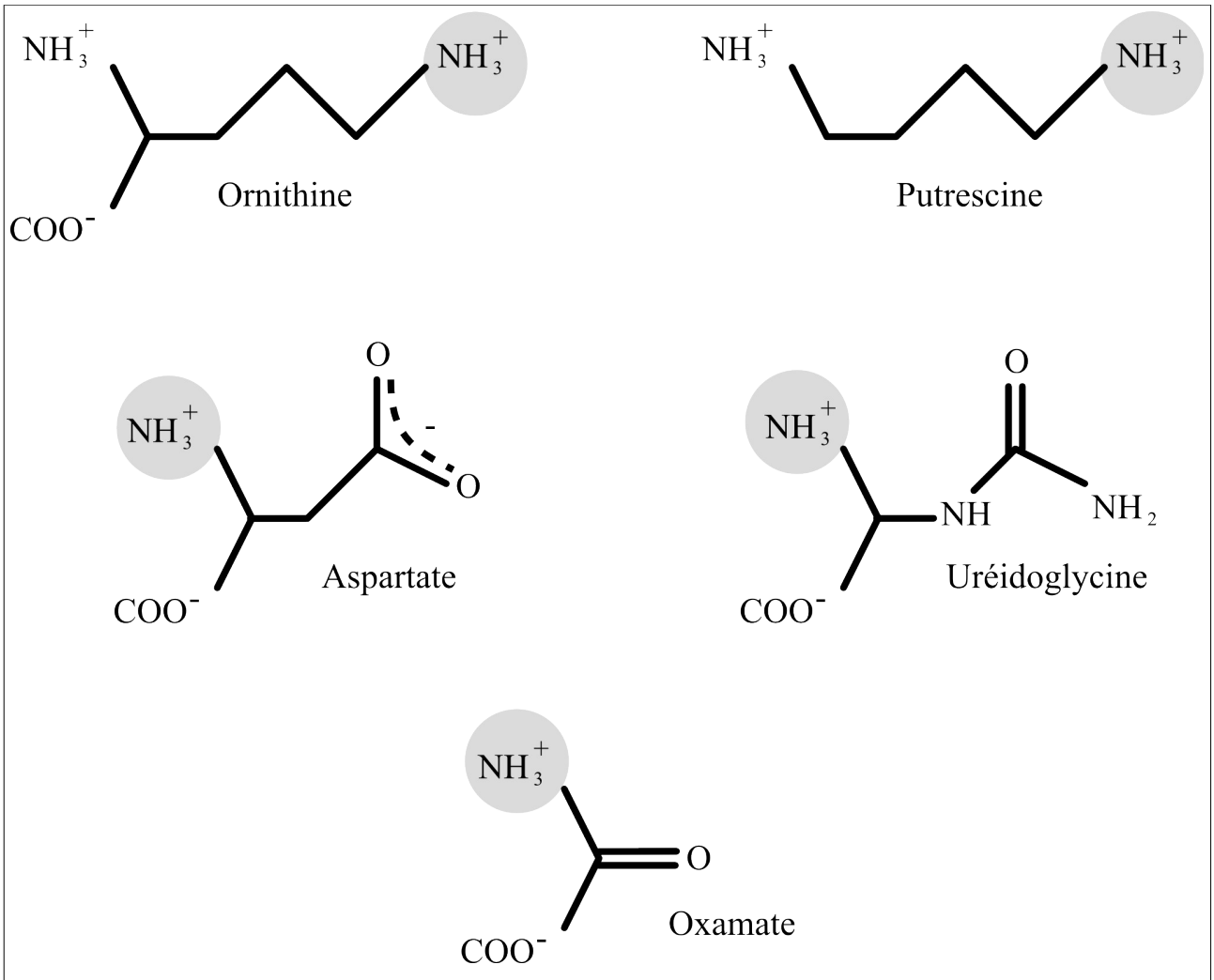


Figure 16 : Comparaison de différents substrats de carbamoyltransférases.
Le groupe -NH_3^+ en gris est celui sur lequel se fait la carbamoylation. L'ornithine et la putrescine sont toutes deux carbamoylées sur leur groupe aminé en bout de chaine (en δ), de même pour l'acétylornithine et le succinylornithine qui sont tous deux acétylé/succinylés sur le groupe aminé en α. La carbamoylation a lieu sur le groupement en α de l'aspartate. Sur l'uréidoglycine, cible potentielle de la pseudo-ATCase de Rubrobacter, la carbamoylation est très similaire. Sur l'oxamate la carbamoylation a également lieu dans une configuration similaire, mais la molécule est beaucoup plus petite.

9 Proximité de substrat, proximité phylogénétique

L'allantoate/uréidoglycine semble donc être un bon candidat de substrat/produit pour cette TCase. Il existe cependant plusieurs substrats dans la voie des purines qui pourraient aussi être la cible de cette carbamoyltransférase, outre l'allantoate : l'allantoïne, l'uréidoglycine et l'uréidoglycolate, ainsi que l'oxalurate (qui donne de l'oxamate, via l'OxTCase).

J'ai comparé la similarité de ces différentes molécules en utilisant l'outil ChemMine [Backman, Cao & Girke 2011] qui permet de construire des dendrogrammes des différentes molécules chimiques et d'en comparer les caractéristiques. La comparaison des différentes structures des molécules potentielles de la voie, carbamoylées d'une part (Figure 17A) et décarbamoylées d'autre part (Figure 17B), montre que l'allantoate et l'uréidoglycine sont les plus similaires aux carbamoylaspartate et aspartate, respectivement. Les molécules des voies de dégradation des pyrimidines (N-carbamoyl-β-alanine, N-carbamoyl-β-aminoisobutyrate) forment un groupe frère qui permet d'enraciner les voies purines. Cela est en accord avec la proximité phylogénétique des pseudo-ATCases avec les ATCases, ainsi qu'avec les motifs structuraux conservés, très ressemblant avec l'ATCase, notamment avec la carbamoylation sur le groupe aminé en α, mais divergents au niveau de la liaison au groupe carboxyl, remplacé par un groupe amide dans les pseudo-ATCases.

La modélisation de la structure 3D de cette TCase avait suggéré que le substrat devait être plus volumineux que l'aspartate et l'oxamate, or l'allantoate (168,068 g/mol) est le plus volumineux des différents composés comparés, en étant légèrement plus massif que le carbamoyl-aspartate (168,064 g/mol).

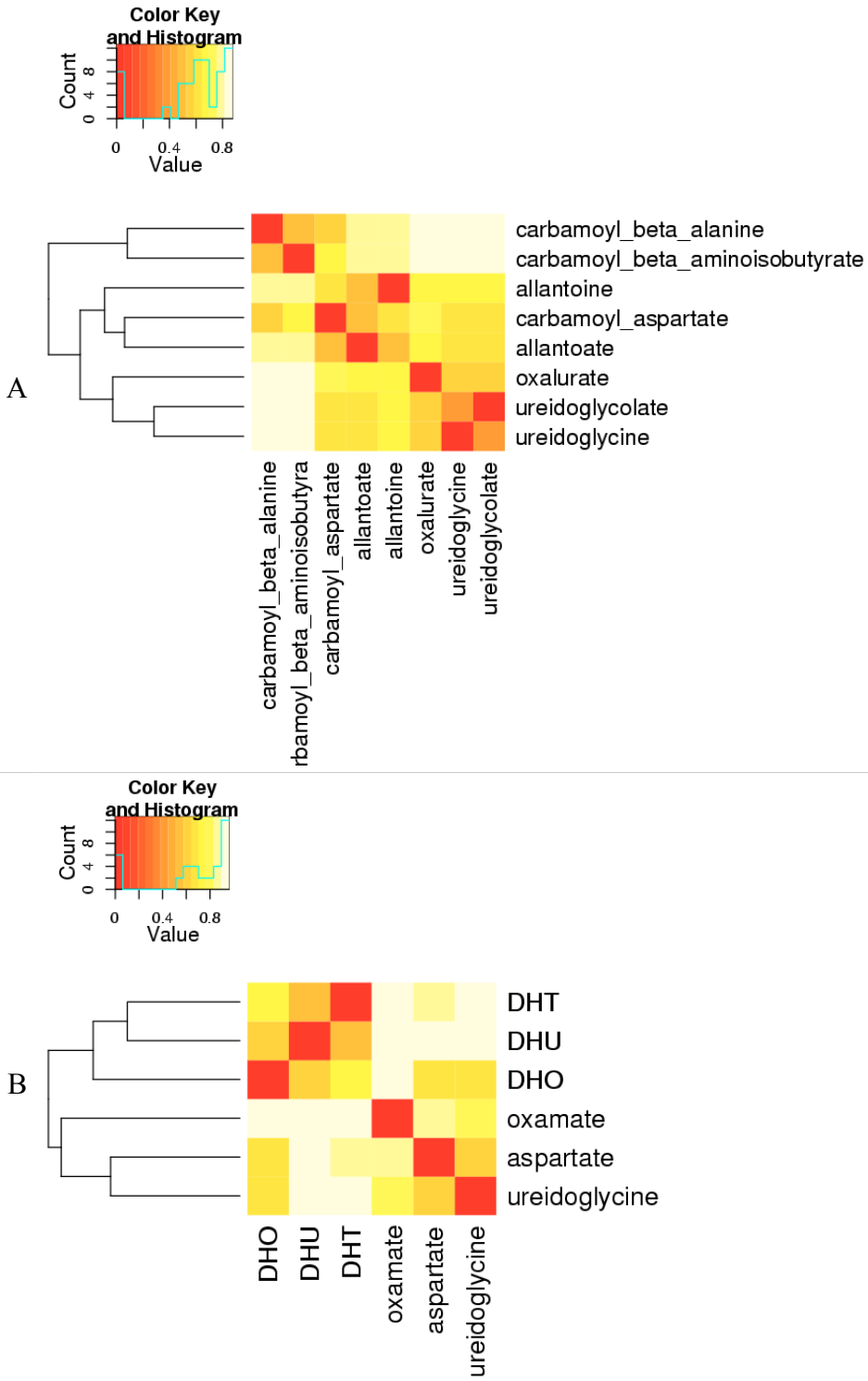


Figure 17 : Dendrogramme représentant la similarité des molécules carbamoylées et décarbamoylées.

Dendrogrammes obtenus par clustering hiérarchique avec lien moyen avec l'outil ChemMine. Le score utilisé est le coefficient de Tanimoto (cf ChemMine). Les molécules carbamoylées sont en A et les décarbamoylées en B.

Discussion

1 Au delà du réductionnisme

Pendant de nombreuses années, les biochimistes et les généticiens se sont efforcés de comprendre le fonctionnement d'une enzyme dans tous ses détails. Cette approche réductionniste a longtemps été indispensable et a apporté une quantité considérable d'informations. Cependant, en particulier avec l'irruption des données génomiques, on s'est rendu compte qu'il fallait replacer cette enzyme unique dans son contexte cellulaire, c'est-à-dire essayer de comprendre le fonctionnement du système dont elle n'est que l'un des éléments.

Cette nouvelle manière de voir les choses s'est accompagnée de la découverte récente que l'ancien modèle clé/serrure ne rendait pas compte de la réelle nature des enzymes qui sont capables d'une certaine plasticité en pouvant travailler sur des substrats similaires mais différents et participer ainsi à des voies métaboliques diverses [Jensen 1976].

C'est pourquoi les enzymes se retrouvent souvent réunies dans de grandes superfamilles multigéniques dont les fonctions moléculaires et/ou cellulaires sont variées, tout en demeurant similaires. Nous avons vu ainsi dans ce travail de Thèse que les dihydroorotases, les allantoïnases, les dihydropyrimidinases et les hydantoinases forment la superfamille des amidohydrolases cycliques en catalysant toutes l'hydrolyse d'un cycle amide (de type pyrimidine ou hydantoine). Ces enzymes ont des fonctions moléculaires différentes et sont utilisées dans des voies diverses (biosynthèse des pyrimidines, dégradation des pyrimidines, dégradation des purines...). La détermination exacte des relations d'homologie (orthologie, paralogie) entre les membres de ces différentes familles, bien que particulièrement difficile à réaliser, reste cruciale si l'on veut comprendre l'évolution des voies métaboliques dans lesquelles interviennent ces enzymes et retracer leur origine hypothétique dans des voies ancestrales remontant parfois avant l'apparition de LUCA, le dernier ancêtre commun à l'ensemble des organismes contemporains.

Plutôt que de chercher à reconstruire l'histoire évolutive d'une seule enzyme indépendamment, il apparaît pertinent de considérer plutôt l'évolution des enzymes homologues et de leurs partenaires dans les différentes voies métaboliques dans lesquelles elles interviennent. En effet, les différentes protéines d'une voie donnée peuvent être vues comme un cas particulier de

modules fonctionnels [Hartwell et al. 1999] qui réunissent tous les intervenants nécessaires à une fonction précise.

L'étude des familles d'enzymes, cependant, repose sur des séquences en acides aminés identifiées lors de l'annotation structurale d'organismes séquencés qui, pour la plupart, n'ont pas ou peu été étudiés – mis à part quelques organismes modèles, comme *Escherichia coli* dont pourtant un tiers des gènes codant des protéines demeurent orphelins, sans annotation fonctionnelle [Hu et al. 2009]. L'annotation fonctionnelle des protéines codées par un organisme nouvellement séquencé se fait, dans l'immense majorité des cas, de manière automatique essentiellement par homologie, c'est-à-dire en comparant les séquences à identifier avec des séquences déjà annotées expérimentalement et en admettant que leur ressemblance est suffisamment forte pour témoigner d'une parenté. Or cette approche systématique de comparaison de séquences deux à deux peut souvent créer et propager des erreurs d'annotation (voir par exemple Furnham et al. [2009] et Schnoes et al. [2009]) et les séquences trouvées comme étant les plus ressemblantes ne sont pas forcément les meilleurs voisins phylogénétiques [Koski & Golding 2001]. Cette approche automatique n'est en effet qu'une approximation des relations d'homologie et ne peut remplacer une réelle étude évolutive. C'est pourquoi la reconstruction d'un arbre phylogénétique devient nécessaire dès lors que l'on veut annoter pertinemment des séquences supposées être homologues.

2 De nouvelles approches sont nécessaires

2.1 Maîtriser l'afflux irrépressible des données génomiques

Une stratégie judicieuse d'annotation fonctionnelle et d'analyse comparée des familles doit donc être utilisée. J'ai mis au point de nouvelles approches méthodologiques pour permettre l'étude détaillée des familles multifonctionnelles et leur histoire évolutive en relation avec leurs partenaires dans les voies métaboliques.

Le Chapitre 4 décrit comment j'ai créé le programme Frali qui permet de construire et de maintenir à jour un alignement de séquences multiple (MSA) basé sur un alignement de base (appelé une *graine*) de grande qualité. Couplée à une recherche d'homologues à partir d'un profil basé sur un modèle de chaînes de Markov cachées (HMM [Eddy 2009]) et utilisant l'alignement préexistant, cette approche permet de se passer d'une recherche manuelle de nouvelles séquences et de l'ajout manuel de ces séquences à l'alignement. On évite ainsi de recréer l'ensemble de l'alignement à chaque fois qu'une nouvelle séquence est disponible (c'est-à-dire pratiquement

chaque jour au rythme actuel des publications de génomes nouveaux). Cette méthode permet également de n'aligner que la partie expérimentalement justifiée des séquences (domaine homologue), et donc d'éviter d'avoir à pré-découper manuellement les longues séquences fusionnées pour en extraire la seule partie alignable. Il est même possible de prendre en compte automatiquement les cas où plusieurs domaines d'une même séquence sont homologues entre eux (provenant de produits de duplication de gènes qui ont ensuite fusionné) ; dans ce cas la séquence est scindée et les multiples domaines alignés. L'alignement de qualité régulièrement mis à jour peut être utilisé pour reconstruire des arbres phylogénétiques, mais également pour extraire des motifs structuraux conservés. Dans les deux cas, un tel alignement de bonne qualité est indispensable.

Afin de gérer efficacement l'utilisation et la mise à jour des alignements multiples des familles étudiées, j'ai mis en place une base de données relationnelles appelée AP (*Aligned Proteins*) accompagnée d'un ensemble de scripts qui en permettent l'exploitation. Ces scripts permettent d'importer et d'exporter les données, d'annoter les séquences avec divers outils *ad hoc*, basés sur l'interprétation des arbres phylogénétiques (annotation par monophylie), la conservation de motifs structuraux, ou encore la synténie avec d'autres gènes caractérisés. Enfin, d'autres scripts permettent de visualiser les résultats, notamment en extrayant différents profils phylogénétiques des différents génomes représentés, accompagnés de détails précieux comme la proximité génétique des autres gènes d'intérêt.

2.2 Les modules réactionnels : un nouveau concept pour étudier l'évolution du métabolisme

L'analyse comparée de réactions chimiques successives réalisées par des enzymes homologues dans des voies différentes m'a amené à créer le concept de module réactionnel. Un module réactionnel est formé d'une succession de réactions biochimiques retrouvée dans diverses voies métaboliques, et assurées – aux étapes comparables – par des enzymes généralement homologues. L'idée sous-jacente est qu'un module réactionnel refléterait un élément de base d'une voie ancestrale qui se serait ensuite différencié pour créer des voies de plus en plus spécialisées permettant de transformer un substrat donné, tout en conservant des processus de réactions similaires. De manière plus globale, un module réactionnel serait une brique élémentaire d'un module fonctionnel défini par la biologie des systèmes.

Chaque superfamille de protéines d'un tel module est donc constituée d'un ensemble d'enzymes qui effectueraient une même réaction biochimique, mais sur des substrats différents quoique similaires et agissant dans différentes voies. C'est avec cette vision que j'ai abordé l'étude

de différentes superfamilles impliquées dans les voies de biosynthèse et de dégradation des pyrimidines, à commencer par les amidohydrolases cycliques.

3 Application à la superfamille des amidohydrolases cycliques

Comme rappelé dans le Chapitre 6, la superfamille des amidohydrolases cycliques contient les dihydroorotases (DHOases), les allantoïnases (ALNases), les hydantoinases (HYDases) et les dihydropyrimidinases (DHPases).

3.1 Une nouvelle classification des dihydroorotases

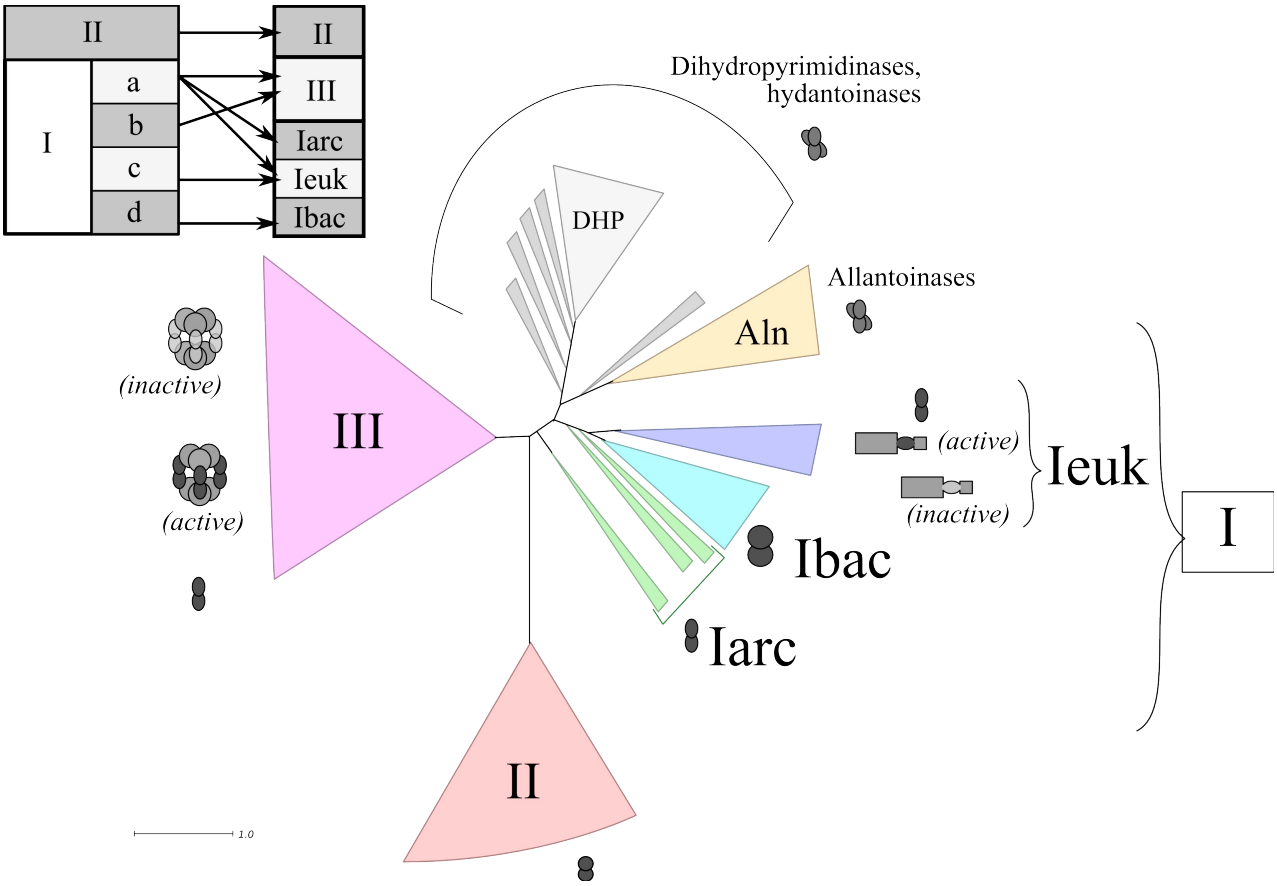


Figure 18 : Arbre phylogénétique de la famille des DHOases et nouvelle classification.

Les DHOases qui catalysent la troisième étape de la voie de biosynthèse des pyrimidines constituent la majorité des membres de cette superfamille et se distinguent en plusieurs types dont j'ai proposé une nouvelle classification basée sur des caractéristiques phylogénétiques, structurelles

et synténiques et qui est une modification de l'ancienne classification proposée par Fields et al. [1999]. Ces types sont appelés I, II et III :

- Les séquences de type I sont retrouvées dans les trois grands domaines du vivant (Archées, Bactéries, Eucaryotes) et sont phylogénétiquement plus proches des autres amidohydrolases cycliques que des autres types de DHOases. Elles partagent d'ailleurs avec elles une même structure élémentaire en homodimères, bien que les autres amidohydrolases soient surtout trouvées en tétramères. Chez quelques Eucaryotes par contre, la DHOase I est fusionnée dans une grande protéine CAD avec les enzymes des deux étapes précédentes dans la voie.
- Les séquences de type II sont très éloignées phylogénétiquement de toutes les autres, ayant perdu en particulier le domaine en sandwich β initialement présent dans les régions N- et C-terminales. Ce type II se retrouve essentiellement chez des Bactéries, mais est présent aussi chez quelques Eucaryotes qui les ont probablement acquis par endosymbiose.
- Les séquences de type III, enfin, sont essentiellement bactériennes et se retrouvent dans la plupart des grands groupes taxonomiques de Bactéries, dont ceux qui sont considérés les plus anciens (ou plus exactement ayant divergé le plus tôt). Ces sous-unités PyrC sont généralement en complexe multimérique avec l'unité catalytique PyrB des aspartates carbamoyltransférases (ATCases) catalysant l'étape précédente de la biosynthèse. Selon les organismes, les sous-unités PyrC ont une fonction catalytique ou seulement structurale (la perte d'activité catalytique est survenue plus récemment). Dans ce dernier cas, les organismes contiennent aussi un paralogue actif de type I ou II. Les données phylogénétiques actuelles ne permettent pas de trancher pour savoir si les organismes ancestraux possédaient plusieurs paralogues perdus ou non par leurs descendants, ou si le paralogue catalytique a été réintroduit suite à un événement récent pour pallier la transformation du type III en sa configuration structurale et non catalytique.

3.2 Une répartition taxonomique qui témoigne de l'histoire évolutive complexe des dihydroorotases

Bien que les trois grands types de DHOases décrits ci-dessus soient clairement séparés phylogénétiquement, leur réparition dans les espèces vivantes témoigne d'une histoire évolutive complexe.

3.2.1 DHOases d'Archées

Chez les Archées, toutes les DHOases sont du même type I (Iarc, Figure 18). On notera cependant que ces DHOases sont branchées dans différents groupes à la base des DHOases de type I (Iarc).

Bien que la majorité des Archées contiennent une DHOase I, il existe de rares exceptions. La première est la DHOase de *Thermococcus sibiricus MM 739* (THESM). *T. sibiricus* est une hyperthermophile qui a été isolé dans un puis de pétrole de Sibérie à près de 2000 m de profondeur par une température de 60°C à 84°C et qui aurait survécu depuis le Jurassique en métabolisant la matière organique des sédiments avec lesquels il a été enfoui [Miroshnichenko et al. 2001]. Sa DHOase est de type III, seule exception de ces DHOases qui sont toutes chez des Bactéries. Remarquablement, le gène *pyrC* de THESM se trouve en compagnie des autres gènes *pyrB, pyrI, pyrD, pyrK* de la voie de biosynthèse des pyrimidines, mais tous ces autres gènes sont proches phylogénétiquement des autres gènes orthologues chez les autres Archées. En particulier, l'ATCase de THESM est du type II B qui correspond à une structure quaternaire où PyrB est accompagné d'une PyrI dans un complexe dodécamérique. N'ayant pas d'ATCase de type I dans son génome, la DHOase III de THESM est donc probablement sous la forme d'un dimère libre.

Autre exception, les Thaumarchées *Cenarchaeum symbiosum A 414004* (CENSY) et *Nitrosopumilus maritimus SCM1* (NITMS) ne possèdent pas de DHOase d'aucune sorte, bien qu'elles possèdent toutes deux les gènes de toutes les autres enzymes de la voie : *pyrB* + *pyrI*, *pyrD* + *pyrK, pyrE, pyrF.* Elles possèdent cependant une amidohydrolase dans la même famille des DHOases mais qui ne contient aucun des motifs universels des DHOases connues (Uniprot : A0RTM1_CENSY, A9A2Y2_NITMS). Cette amidohydrolase, isolée des autres gènes de la voie dans le génome, pourrait être une DHOase d'un nouveau genre avec une configuration du site actif différent de toutes les autres DHOases. Mais on ne peut pas non plus exclure que la fonction de DHOase soit remplie par une enzyme d'une toute autre famille.

3.2.2 DHOases d'Eucaryotes

DHO I	CAD	CAD inact + DHOase II	DHO II	Aucune
Kinetoplastida (Euglenozoa)	Metazoa Dictyosteliida	Fungi Choanoflagellida *Cyanidioschyzon merolae* (Rhodophyta)	Apicomplexa Bacillariophyta, Peronosporales, Phaeophyceae (Stramenopiles) Perkinsida (Alveolata) Chlorophyta, Streptophyta (Viridiplantae) Paulinella (Rhizaria)	Entamoeba (Amobozoa) Trichomonas (Parabasalia)

Tableau 16 : Répartition des DHOases chez les Eucaryotes.

Les DHOases d'Eucaryotes ont plusieurs DHOases possibles (Tableau 16), à commencer par les CAD qui sont une fusion de DHOases de type I avec une CPSase et une ATCase II, dans l'ordre CPSase – DHOase – ATCase. Les trois enzymes sont donc réunies en un même polypeptide ce qui permet une plus grande efficacité des réactions, qui ont lieu l'une après l'autre. On notera que l'ordre des domaines ne correspond pas à l'ordre des réactions (PyrA – PyrC – PyrB). Il existe en fait entre les domaines DHOase et ATCase une chaine d'une centaine d'acides aminés qui permet au domaine ATCase de se placer correctement lors du repliement de la protéine.

Cette CAD se trouve classiquement chez les Animaux (Metazoa) et les Dictyosteliida avec un domaine DHOase actif, ainsi que chez les champignons où le domaine DHOase est inactif (CAD-like). La fonction DHOase est alors jouée par une DHOase II dans le même génome. On trouve également une CAD-like + DHOase II chez les Choanoflagellida. Cette triple fusion remarquable est un des critères pour réunir tous ces taxons dans le clade des Unicontes (= 1 flagelle), par opposition aux Bicontes (= 2 flagelles).

Chez les Bicontes la plupart des espèces ont un type II seul, notamment les plantes et les algues vertes, les Chromalveolata (Stramenopiles, Alveolata). À l'exception des Plasmodium, la plupart des DHOases II de Bicontes sont proches de DHOases II de Protéobactéries et de Cyanobactéries, ce qui suggère qu'elles auraient acquis leur DHOase par endosymbiose. Une algue

rouge, *Cyanidioschyzon merolae* contient cependant une CAD-like avec une DHOase II, comme une Uniconte.

On trouve également une exception à ces répartitions chez *Naegleria gruberi* (NAEGR) qui possèderait une DHOase de type I bactérienne, proche de DHOases de Bacteroidetes.

3.2.3 DHOases de Bactéries

Chez les Bactéries, on retrouve les trois types de DHOase (Tableau 17 et Figure 19), parfois jusqu'aux trois en même temps dans un même génome (*Pseudomonas aeruginosa*). La répartition des différents types est loin d'être homogène : la plupart des Bactéries possèdent une DHOase de type III, dont une bonne partie où celle-ci est inactivée. Dans ce cas, l'activité DHOase est réalisée par une DHOase I ou une DHOase II. Remarquablement, chaque génome n'a jamais deux DHOases d'un même type ; on ne retrouve en particulier aucune DHOase III active avec une DHOase III inactive, sauf dans quelques cas très rares comme chez *Actinosynnema mirum* où une deuxième DHOase III est présente en plus de la DHOase III que contiennent toutes les Actinobacteria. Mais dans ce cas-là, cette seconde DHOase n'est plus en synténie avec les autres gènes de la voie.

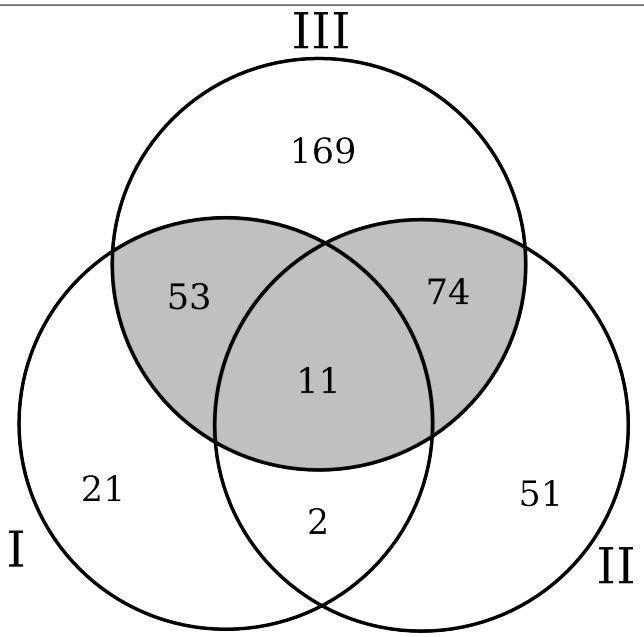


Figure 19 : Diagramme de Venn représentant les différentes combinaisons des types I, II et III de DHOases dans les génomes bactériens.
Les données ont été limitées en ne représentant que le nombre de genre afin d'éviter de surestimer les génomes d'espèces aux nombreuses souches séquencées (comme Escherichia coli*). Dans les régions grisées la DHOase de type III est catalytiquement inactive. Par exemple, il y a 53 génomes contenant une combinaison I + III inactive.*

Nom d'espèce	PyrB1	PyrB2	PyrI	PyrC1	PyrC2	PyrC3	PyrC3i	Rang taxonomique
Candidatus Koribacter versatilis Ellin345	1	1				1		Bacteria,Acidobacteria,Candidatus_Koribacter
Corynebacterium glutamicum	1					1		Bacteria,Actinobacteria,Actinobacteria_class_,Actinomycetales,Corynebacteriaceae,Corynebacterium
Mycobacterium smegmatis str. MC2 155	1					1		Bacteria,Actinobacteria,Actinobacteria_class_,Actinomycetales,Mycobacteriaceae,Mycobacterium
Mycobacterium tuberculosis	1					1		Bacteria,Actinobacteria,Actinobacteria_class_,Actinomycetales,Mycobacteriaceae,Mycobacterium
Streptomyces coelicolor	1					1		Bacteria,Actinobacteria,Actinobacteria_class_,Actinomycetales,Streptomycetaceae,Streptomyces
Bifidobacterium longum		1	1			1		Bacteria,Actinobacteria,Actinobacteria_class_,Bifidobacteriales,Bifidobacteriaceae,Bifidobacterium
Aquifex aeolicus	1					1		Bacteria,Aquificae,Aquificae_class_,Aquificales,Aquificaceae,Aquifex
Bacteroides thetaiotaomicron		1	1	1				Bacteria,Bacteroidetes,Bacteroidia,Bacteroidales,Bacteroidaceae,Bacteroides
Flavobacterium psychrophilum JIP02/86	1			1			1	Bacteria,Bacteroidetes,Flavobacteria,Flavobacteriales,Flavobacteriaceae,Flavobacterium
Salinibacter ruber DSM 13855	1					1		Bacteria,Bacteroidetes,Sphingobacteria,Sphingobacteriales,Rhodothermaceae,Salinibacter
Chlorobaculum tepidum	1					1		Bacteria,Chlorobi,Chlorobia,Chlorobiales,Chlorobiaceae,Chlorobaculum
Chloroflexus aurantiacus J-10-fl	1				1		1	Bacteria,Chloroflexi,Chloroflexi_class_,Chloroflexales,Chloroflexaceae,Chloroflexus
Gloeobacter violaceus	1			1			1	Bacteria,Cyanobacteria,Gloeobacteria,Gloeobacterales,Gloeobacter
Synechocystis sp. PCC 6803	1				1		1	Bacteria,Cyanobacteria,Chroococcales,Synechocystis
Thermosynechococcus elongatus BP-1	1			1			1	Bacteria,Cyanobacteria,Chroococcales,Thermosynechococcus
Prochlorococcus marinus	1				1		1	Bacteria,Cyanobacteria,Prochlorales,Prochlorococcaceae,Prochlorococcus
Thermus thermophilus HB8	1					1		Bacteria,Deinococcus-Thermus,Deinococci,Thermales,Thermaceae,Thermus
Deinococcus radiodurans	1					1		Bacteria,Deinococcus-Thermus,Deinococci,Deinococcales,Deinococcaceae,Deinococcus
Dictyoglomus turgidum DSM 6724	1					1		Bacteria,Dictyoglomi,Dictyoglomia,Dictyoglomales,Dictyoglomaceae,Dictyoglomus
Bacillus cereus ATCC 14579	1					1		Bacteria,Firmicutes,Bacilli,Bacillales,Bacillaceae,Bacillus
Bacillus subtilis	1					1		Bacteria,Firmicutes,Bacilli,Bacillales,Bacillaceae,Bacillus
Listeria monocytogenes	1					1		Bacteria,Firmicutes,Bacilli,Bacillales,Listeriaceae,Listeria
Staphylococcus aureus subsp. aureus NCTC 8325	1					1		Bacteria,Firmicutes,Bacilli,Bacillales,Staphylococcaceae,Staphylococcus
Clostridium acetobutylicum		1	1			1		Bacteria,Firmicutes,Clostridia,Clostridiales,Clostridiaceae,Clostridium
Clostridium botulinum A str. Hall		1	1			1		Bacteria,Firmicutes,Clostridia,Clostridiales,Clostridiaceae,Clostridium
Enterococcus faecalis	1					1		Bacteria,Firmicutes,Bacilli,Lactobacillales,Enterococcaceae,Enterococcus
Lactobacillus plantarum	1					1		Bacteria,Firmicutes,Bacilli,Lactobacillales,Lactobacillaceae,Lactobacillus
Lactococcus lactis subsp. lactis	1					1		Bacteria,Firmicutes,Bacilli,Lactobacillales,Streptococcaceae,Lactococcus
Streptococcus pneumoniae R6	1					1		Bacteria,Firmicutes,Bacilli,Lactobacillales,Streptococcaceae,Streptococcus
Moorella thermoacetica ATCC 39073	1					1		Bacteria,Firmicutes,Clostridia,Thermoanaerobacterales,Thermoanaerobacteraceae,Moorella
Fusobacterium nucleatum subsp. nucleatum	1					1		Bacteria,Fusobacteria,Fusobacteria_class_,Fusobacteriales,Fusobacteriaceae,Fusobacterium
Thermodesulfovibrio yellowstonii DSM 11347	1					1		Bacteria,Nitrospirae,Nitrospira_class_,Nitrospirales,Nitrospiraceae,Thermodesulfovibrio
Rhodopirellula baltica	1			1			1	Bacteria,Planctomycetes,Planctomycetacia,Planctomycetales,Planctomycetaceae,Rhodopirellula
Caulobacter vibrioides	1			1			1	Bacteria,Proteobacteria,Alphaproteobacteria,Caulobacterales,Caulobacteraceae,Caulobacter
Rhodobacter sphaeroides 2.4.1	1				1		1	Bacteria,Proteobacteria,Alphaproteobacteria,Rhodobacterales,Rhodobacteraceae,Rhodobacter
Rhodospirillum rubrum ATCC 11170	1			1			1	Bacteria,Proteobacteria,Alphaproteobacteria,Rhodospirillales,Rhodospirillaceae,Rhodospirillum
Bradyrhizobium japonicum	1			1			1	Bacteria,Proteobacteria,Alphaproteobacteria,Rhizobiales,Bradyrhizobiaceae,Bradyrhizobium
Agrobacterium tumefaciens str. C58	1				1		1	Bacteria,Proteobacteria,Alphaproteobacteria,Rhizobiales,Rhizobiaceae,Agrobacterium
Bordetella pertussis	1	1			1		1	Bacteria,Proteobacteria,Betaproteobacteria,Burkholderiales,Alcaligenaceae,Bordetella
Burkholderia pseudomallei	1	1			1		1	Bacteria,Proteobacteria,Betaproteobacteria,Burkholderiales,Burkholderiaceae,Burkholderia
Neisseria gonorrhoeae FA 1090		1	1		1			Bacteria,Proteobacteria,Betaproteobacteria,Neisseriales,Neisseriaceae,Neisseria
Neisseria meningitidis serogroup B		1	1		1			Bacteria,Proteobacteria,Betaproteobacteria,Neisseriales,Neisseriaceae,Neisseria
Desulfovibrio vulgaris str. Hildenborough	1					1		Bacteria,Proteobacteria,Deltaproteobacteria,Desulfovibrionales,Desulfovibrionaceae,Desulfovibrio
Geobacter sulfurreducens	1					1		Bacteria,Proteobacteria,Deltaproteobacteria,Desulfuromonadales,Geobacteraceae,Geobacter
Campylobacter jejuni	1				1		1	Bacteria,Proteobacteria,Epsilonproteobacteria,Campylobacterales,Campylobacteraceae,Campylobact
Helicobacter pylori	2				1		1	Bacteria,Proteobacteria,Epsilonproteobacteria,Campylobacterales,Helicobacteraceae,Helicobacter
Shewanella oneidensis		1			1			Bacteria,Proteobacteria,Gammaproteobacteria,Alteromonadales,Shewanellaceae,Shewanella
Buchnera aphidicola _Acyrthosiphon pisum_		1	1		1			Bacteria,Proteobacteria,Gammaproteobacteria,Enterobacteriales,Enterobacteriaceae,Buchnera
Escherichia coli K-12		1	1		1			Bacteria,Proteobacteria,Gammaproteobacteria,Enterobacteriales,Enterobacteriaceae,Escherichia
Salmonella enterica subsp. enterica serovar Typhi		1	1		1			Bacteria,Proteobacteria,Gammaproteobacteria,Enterobacteriales,Enterobacteriaceae,Salmonella
Shigella dysenteriae Sd197		1	1		1			Bacteria,Proteobacteria,Gammaproteobacteria,Enterobacteriales,Enterobacteriaceae,Shigella
Yersinia pestis		1	1		1			Bacteria,Proteobacteria,Gammaproteobacteria,Enterobacteriales,Enterobacteriaceae,Yersinia
Coxiella burnetii	1				1			Bacteria,Proteobacteria,Gammaproteobacteria,Legionellales,Coxiellaceae,Coxiella
Haemophilus influenzae Rd KW20								Bacteria,Proteobacteria,Gammaproteobacteria,Pasteurellales,Pasteurellaceae,Haemophilus
Pseudomonas aeruginosa	1			1	1		1	Bacteria,Proteobacteria,Gammaproteobacteria,Pseudomonadales,Pseudomonadaceae,Pseudomonas
Francisella tularensis subsp. tularensis		1		1				Bacteria,Proteobacteria,Gammaproteobacteria,Thiotrichales,Francisellaceae,Francisella
Vibrio fischeri ES114		1	1		1			Bacteria,Proteobacteria,Gammaproteobacteria,Vibrionales,Vibrionaceae,Aliivibrio
Vibrio cholerae		1	1		1			Bacteria,Proteobacteria,Gammaproteobacteria,Vibrionales,Vibrionaceae,Vibrio
Xanthomonas campestris pv. campestris	1			1				Bacteria,Proteobacteria,Gammaproteobacteria,Xanthomonadales,Xanthomonadaceae,Xanthomonas
Leptospira interrogans		1		1				Bacteria,Spirochaetes,Spirochaetes_class_,Spirochaetales,Leptospiraceae,Leptospira
Borrelia burgdorferi								Bacteria,Spirochaetes,Spirochaetes_class_,Spirochaetales,Spirochaetaceae,Borrelia
Thermanaerovibrio acidaminovorans DSM 6589	1					1		Bacteria,Synergistetes,Synergistia,Synergistales,Synergistaceae,Thermanaerovibrio
Mesoplasma florum								Bacteria,Tenericutes,Mollicutes,Entomoplasmatales,Entomoplasmataceae,Mesoplasma
Thermotoga maritima		1	1			1		Bacteria,Thermotogae,Thermotogae_class_,Thermotogales,Thermotogaceae,Thermotoga

Tableau 17 : Présence et type de pyrB, pyrI et pyrC dans les génomes de Bactéries.

3.2.4 Histoire des DHOases : l'histoire évolutive d'un groupe de gènes

Si on veut reproduire l'histoire évolutive des DHOases plus pertinemment en évitant d'éventuels transferts horizontaux qui auraient eu lieu (des échanges de types I, II et III), il est

important de se concentrer non pas sur la présence dans le génome des différents types de DHOases et d'ATCase, mais bien de leur covoisinage. En effet, *pyrC* et *pyrB* se retrouvent fréquemment dans un même voisinage génétique.

Si on se réfère aux contexte génétique de ces PyrB et PyrC, on peut distinguer deux cas bien distincts (Tableau 18) : 1) un groupe de synténie entre une DHOase III (active ou inactive) et une ATCase I, et un groupe de synténie entre une ATCase de type II et une pyrI, parfois avec une DHOase de type I. Dans le cas *pyrB* I + *pyrC* III, seules des Bactéries sont représentées, et ce parmi la plupart des groupes bactériens connus. Dans le second cas *pyrB* II + *pyrI*, on trouve d'une part des Archées, dans le voisinage desquelles on trouve également leur gène *pyrC* de type I, mais aussi bon nombre de bactéries. On notera que, mis à part une poignée de Firmicutes, quand *pyrC* est en synténie avec *pyrB* II et *pyrI,* elle est systématiquement de type I. L'ordre de ces groupes est également conservé : pyrB_I>pyrC_III>, de même que pyrB_II>pyrI>pyrC_I. On pourra également rajouter à cela les CAD d'Eucaryotes, qui contiennent des domaines pyrB II et pyrC I, comme certaines Archées.

pyrB I> *pyrC* III>	*pyrB* II> *pyrI*>*(+pyrC I>)*
	Thermoprotei *(+pyrC* I) Korarchaea (+*pyrC* I) et les autres Archaea (sauf Thaumarchaea)
• Acidobacteria • Actinobacteria • Aquificaea • Chloroflexi (active ou inactive) • Deinococcus-Thermus • Deferribacteres • Dictyoglomi • Elisimicrobia • Fibrobacteres (inactive) • Firmicutes • (Fusobacteria) *Fusobacterium* • Gemmatimonadetes • Nitrosopiraea • Planctomycetes (inactive) • Protéobacteria : ◦ Alpha (inactive) ◦ Burkholderia (inactive) ◦ Quelques Gamma (inactive) ◦ Delta (active) ◦ Epsilon • Magnetococcus • Synergestes • Thermobaculum • Verrucomicrobia (active ou inactive)	• (Actinobacteria) *Bifidus* (*pyrC* III) • Bacteroidia • Chlamidyaea (+ *pyrC* I) • (Firmicutes) Quelques Clostridiales et des Clostridia *(+pyrC* III) • Quelques Fusobacteria (+*pyrC* I) • Protéobactéria : ◦ (Beta) Neisseriales ◦ Quelques Gamma • Siprochaetes (parfois +*pyrC* I) • Thermotogae (fusion)

Tableau 18: Voisinage des gènes pyrB d'ATCase avec le type de DHOase pyrC associé.
Lorsque pyrC *est présente mais inactive elle est notée comme telle. Si* pyrC *est présent dans le voisinage du second groupe, il est* noté+pyrC. *Sinon, il est simplement noté* pyrC *entre parenthèses.*

Les données des différents groupes taxonomiques suggère qu'il y aurait eu un ensemble ancestral de trois gènes, *pyrB, pyrI et pyrC*, proches les uns des autres. ATCase et PyrI formant un complexe, tandis que les DHOases étaient en simples dimères, structure quaternaire ancestrale de la famille. Après divergence entre les trois domaines, les Archées ont gardé cet ensemble de trois gènes ancestraux. Chez les Eucaryotes, la PyrI a été perdue, au profit d'une fusion chez les Unikontes entre les gènes PyrA, PyrB et PyrC. Chez les Bactéries, des duplications ont permis l'apparition de deux protéines complémentaires, PyrB (I) et PyrC (III), capables de former un complexe de deux trimères de PyrB et de trois dimères de PyrC, en remplacement de la PyrI.

Dans certains cas, l'activité DHOase de la PyrC III a été perdue plus tard (Proteobacteria), complémentée par l'autre copie restante de PyrC I ; la copie a néanmoins été conservée pour son rôle dans le complexe formé avec PyrB. Il n'est pas clair si cette inactivation a eu lieu une ou plusieurs fois. On retrouve en effet de grands clades où la DHOase est inactive (Alphaproteobacteria) mais aussi des genres où on trouve les deux sortes, comme chez les Campylobacteres où on trouve une DHOase III active chez les *C. concisus*, et une inactive chez les *C. jejuni*, complémentée alors par une DHOase de type II.

L'histoire des DHOases de type II est plus difficile a reconstituer. En effet, la structure de ces DHOases est très différentes des types I et III, la renvoyant loin dans l'arbre phylogénétique et rendant les déductions difficiles. La présence d'une DHOase de type II chez les plantes et de nombreux Eucaryotes unicellulaires est aisément explicable par une endosymbiose récente avec des plastes, comme le suggère la phylogénie des types II. Ce transfert aurait remplacé une DHOase ancestrale de type I, comme on n'en trouve plus que chez certaines Kinetoplastida (*Trypanosoma*, *Leishmania*). Cette hypothèse est compatible avec la présence d'une ATCase de type II native chez ces Eucaryotes, que l'on trouve dans des gènes non fusionnés.

L'origine même des DHOases II est également intrigante : si on considère sa structure, on constate que l'interface des deux sous-unités de ses dimères n'est pas la même que celle des autres amidohydrolases de la famille. En effet, cette interface correspondrait par analogie à l'interface entre la DHOase et l'ATCase dans le complexe dodécamérique des DHO III + ATC I (en comparant celles d'*Escherichia coli* et d'*Aquifex aeolicus* par exemple) : elle est constituée de trois boucles dans la structure en tonneau (α-β)x8 formées entre les feuillets 5 et 6, 7 et 8, et une boucle suivant directement le dernier feuillet 8. On peut alors suggérer que cette DHOase II descendrait d'une ancestrale DHOase III, qui aurait remplacé l'interface avec l'ATCase par une interface avec une autre sous-unité. Dès lors, l'interface « classique » du dimère aurait été perdue, menant à la perte de régions inutiles de la protéine, en particulier le domaine sandwich-β.

4 La complexité des relations structure – fonction au sein des amidohydrolases cycliques et leurs conséquences

La phylogénie de la famille révèle que les autres amidohydrolases cycliques sont apparemment plus proches des DHOases de type I que des deux autres types, laissant penser que la forme ancestrale des DHOases serait ce type I et que les formes II et III en seraient de lointains

dérivés. Cette hypothèse est appuyée par le fait qu'on retrouve des DHOases I chez les trois grands domaines du vivant, alors que le type III est exclusivement bactérien, de même que le type II dont la présence chez certains Eucaryotes proviendrait de transferts horizontaux récents via des endosymbioses.

Remarquablement, les autres amidohydrolases apparaissent donc moins distantes phylogénétiquement les unes des autres que ne le sont les différents types de DHOases entre elles. De plus, elles ont une fonction moléculaire similaire (hydrolyse d'un cycle amide) et interviennent dans des voies parallèles à celle de la voie de biosynthèse des pyrimidines (Figure 20). Le rôle de ces enzymes demeure donc similaire entre toutes ces voies, et l'idée de modules réactionnels m'a amené à rechercher si les autres familles de ces voies adjacentes à ces réactions présentaient une même configuration phylogénétique.

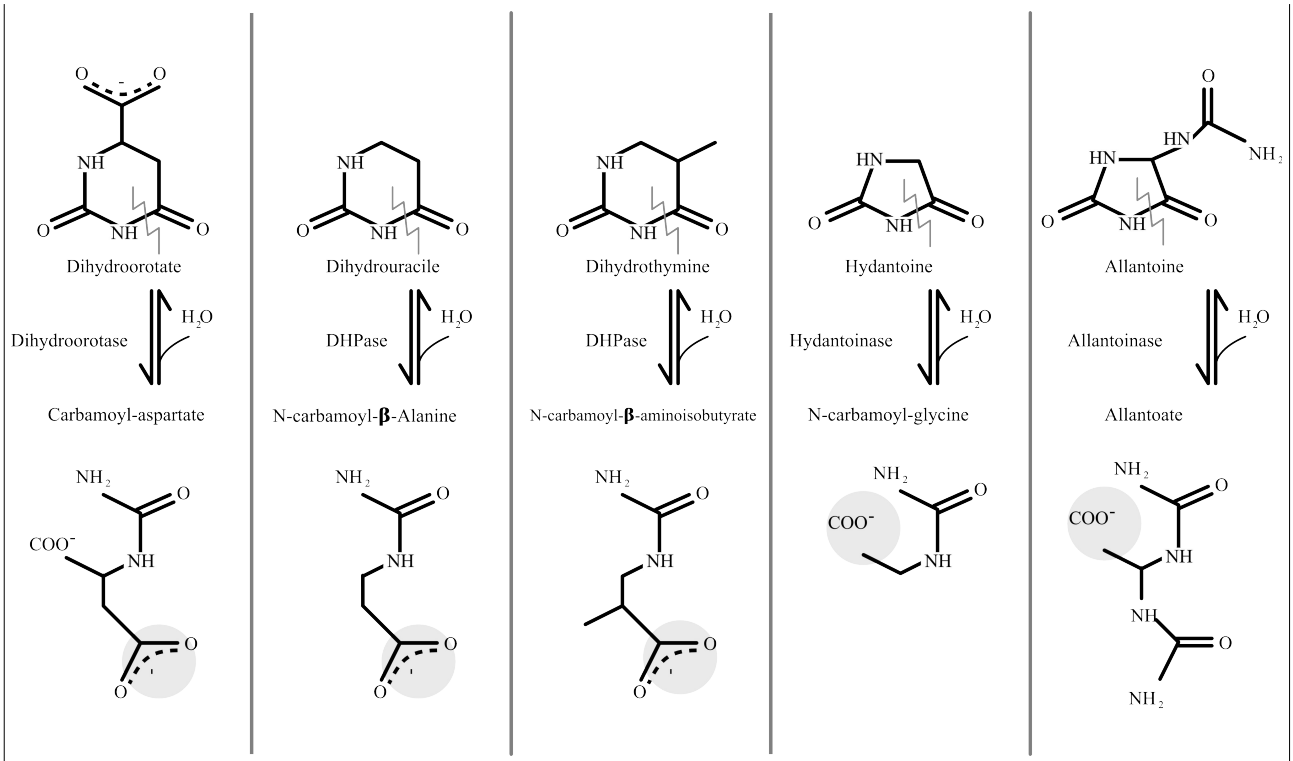


Figure 20 : comparaison des différentes réactions catalysées par les amidohydrolases cycliques.

Dans la voie de dégradation des pyrimidines, on trouve une dihydropyrimidine déshydrogénase (DHPDase) précédant la dihydropyrimidinase (DHPase). Ces deux étapes successives apparaissent antiparallèles. des étapes de DHOase, et de dihydroorotate déshydrogénase de la voie de biosynthèse (Figure 21). Les DHOases et DHPases appartiennent à la superfamille des

amidohydrolases cycliques précédemment décrite, tandis que les DHODases et DHPDases sont homologues entre elles. Il est donc intéressant de retracer l'évolution de ces déshydrogénases et de comparer leur phylogénie avec celles des amidohydrolases cycliques.

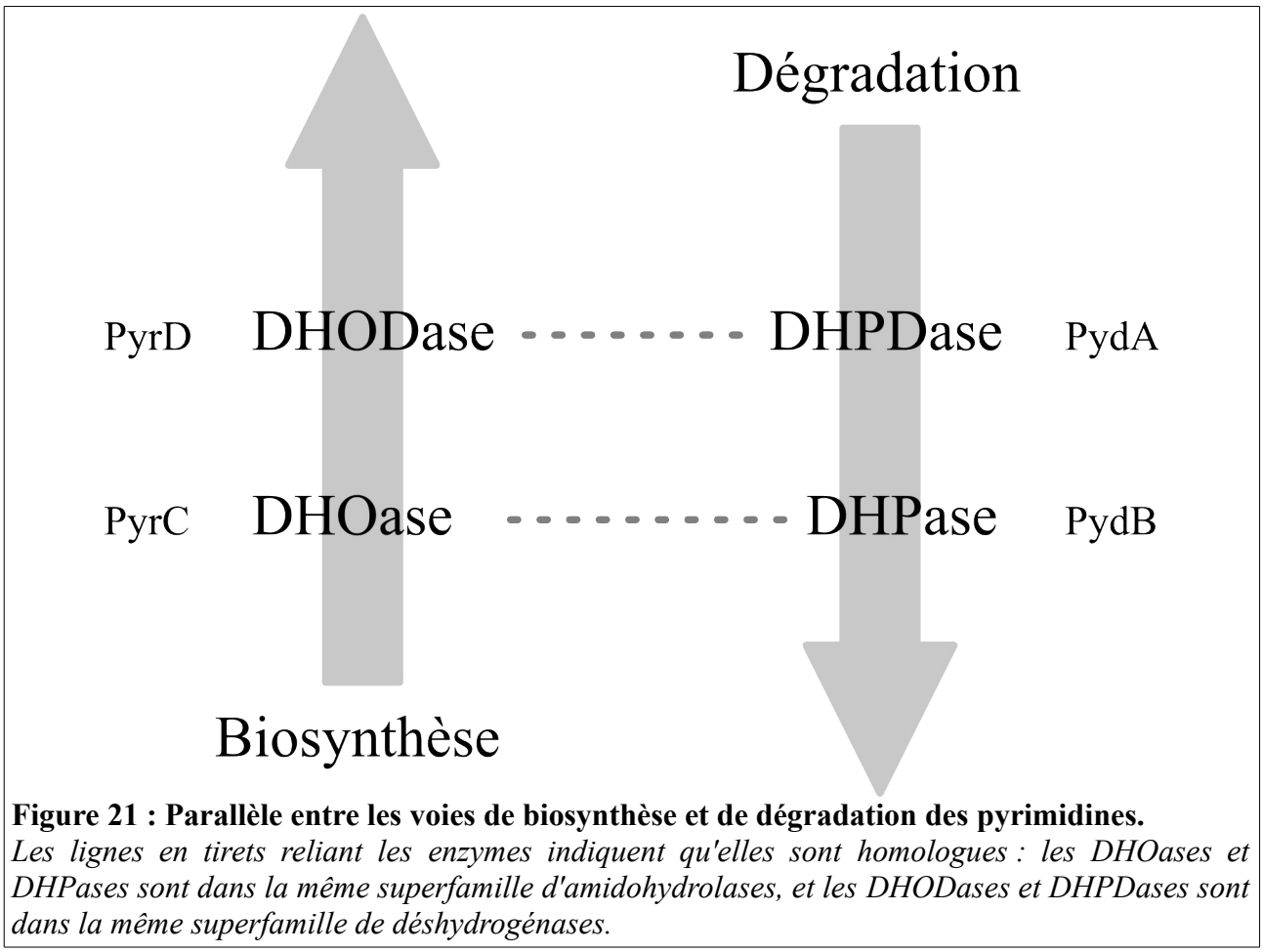
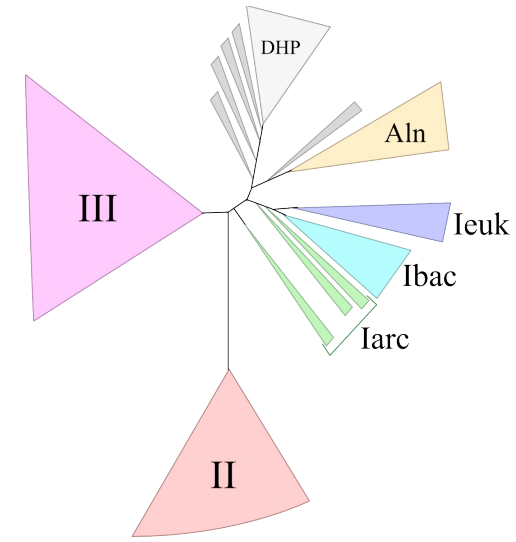
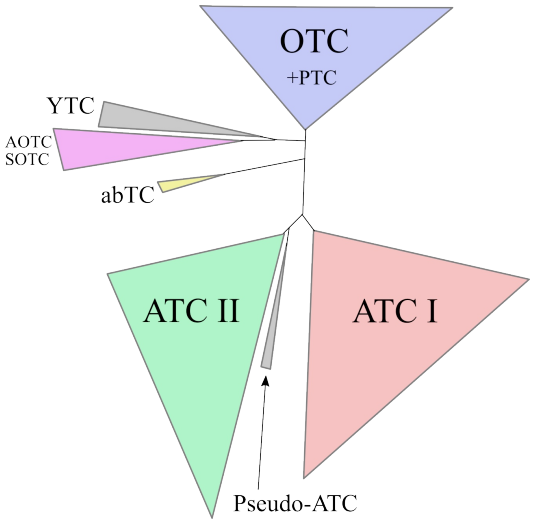


Figure 21 : Parallèle entre les voies de biosynthèse et de dégradation des pyrimidines.
Les lignes en tirets reliant les enzymes indiquent qu'elles sont homologues : les DHOases et DHPases sont dans la même superfamille d'amidohydrolases, et les DHODases et DHPDases sont dans la même superfamille de déshydrogénases.

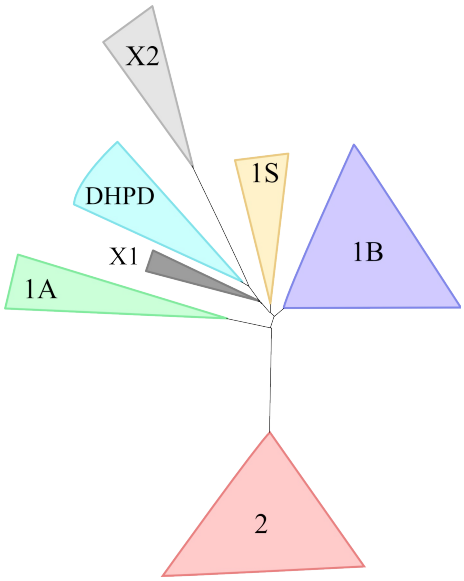


A : Arbre simplifié des amidohydrolases cycliques.
Ieuk = DHOase de type I, Eucaryotes
Ibac = DHOase de type I, Bactéries
Iarc = DHOase de type I, Archées
II = DHOase de type II
III = DHOases de type III
Aln = allantoïnases
DHP = dihydropyrimidinases et hydantoinases

Les petits sous-arbres sans étiquette sont des amidohydrolases cycliques dont la fonction reste inconnue. Certaines sont dans un contexte de type Yge.



B : arbre simplifié des carbamoyltransférases.
ATC I = ATCases de type I
ATC II = ATCases de type II
Pseudo-ATC = Pseudo-ATCases à la base des ATCases qui s'en différencient
OTC = OTCases anaboliques et cataboliques, et PTCases
abTC = TCases dans des voies de biosynthèse d'antibiotiques
AOTC/SOTC = sous-arbre des AOTCases et SOTCases
YTC = YTCases, groupe de TCases inconnues contenant YgeW d'*E. coli*



C : arbre simplifié des dihydroorotates déshydrogénases.
1A : DHODase de type 1A homodimérique
1B : DHODase de type 1B hétérotétramérique de Bactéries
1S : DHODase de type 1B hétérotétramérique d'Archées
2 : DHODases de type 2 membranaire (mitochondrie chez les Eucaryotes)
DHPD = DHPDase
X1 = groupe de séquences inconnues dans un contexte de type Yge
X2 = groupe de séquences inconnues

Figure 22 : Arbres phylogénétiques simplifiés des trois superfamilles étudiées.
Les triangles représentent des sous-arbres, proportionnels au nombre de séquences.

5 Réexamen de l'évolution des familles des partenaires des amidohydrolases

5.1 Famille des dihydroorotate déshydrogénases et leur intégration dans une superfamille

À l'instar de la famille des amidohydrolases cyclique, la superfamille des DHODases/DHPDases est dominée par les DHODases de la voie de biosynthèse des pyrimidines, que l'on peut également séparer en plusieurs types sur des bases phylogénétiques et biochimiques (Figure 22C) : les types 2 membranaires et constituants de la chaîne respiratoire (retrouvés chez bon nombre de Bactéries et dans la mitochondrie de la plupart des Eucaryotes aérobies) sont très éloignés des types 1 cytoplasmiques. Ces derniers sont eux-mêmes subdivisés en trois sous-types : 1A qui forment des homodimères ; 1B qui sont des hétérotétramères des PyrD bactériens avec un partenaire PyrK ; et 1S qui sont des hétérotétramères présents chez les Archées et dont l'homologie du partenaire de PyrD (codé par le gène *orf1* chez *Sulfolobus solfataricus*) n'est pas clairement établie avec PyrK.

Dans un sous-arbre qui prend sa racine parmi les DHODases de types 1, à la base des DHODases de type 1S d'Archées, on trouve un ensemble de séquences que l'on peut diviser en au moins trois grands groupes monophylétiques (Figure 22C : DHPD, X1, X2). Le premier est celui comprenant les séquences PydA (associée à la sous-unité PydX) et PreA (associée à la sous-unité PreT) de DHPDases. Les DHPDases ont la particularité de contenir deux clusters 4Fe-4S (pour le transport d'électrons), alors que les PreT/PydX en contiennent un seul.

Les deux autres groupes de ce sous-arbre, notés X1 et X2 ne correspondent à aucune séquence connue, mais elles ont aussi ce domaine supplémentaire avec deux clusters 4Fe-4S. De manière intéressante leurs contextes génétiques respectifs nous apportent des informations indirectes mais importantes. Certaines X2 sont systématiquement présentes dans un contexte contenant des pyruvate:flavodoxine oxydoréductases, ce qui suggère un rôle oxydoréducteur dans la voie de fixation de l'azote atmosphérique. Les X1 par contre sont notamment présentes dans un contexte qui contient à la fois des amidohydrolases cycliques proches des DHPases (Figure 22A), mais aussi des carbamoyltransférases d'activité inconnue (YTCases, Figure 22B). C'est l'une des principales raisons qui m'ont amené à réanalyser les complexités de la phylogénie des carbamoyltransférases (Figure 22B).

5.2 Les carbamoyltransférases

Dans l'arbre phylogénétique de la superfamille (Figure 22B), les ATCases sont très distinctes des autres TCases et sont clairement divisées en deux grands groupes I et II. Les ornithine TCases (OTCases) constituent un autre groupe majeur et interviennent dans la voie de biosynthèse de l'arginine (OTCases anabolique), de même que dans une voie de dégradation de l'arginine (OTCases cataboliques). Les putrescines TCases (PTCases) interviennent également dans une voie de dégradation de l'arginine et sont phylogénétiquement très proches des OTCases (leurs séquences n'en sont pas distinguées dans l'arbre de la Figure 22B). Ce dernier point est à rapprocher du fait que les PTCases possèdent également une faible activité OTCase comme chez *Enterococcus faecalis* [Llácer et al. 2007], voire dans certains cas comme chez *Listeria monocytogenes* une double fonction d'OTCase et de PTCase en intervenant alors véritablement dans deux voies de dégradation de l'arginine [Chen et al. 2011].

En plus de ces deux grands groupes, on notera la présence de quelques séquences dont celles que j'appelle abTCases qui participent à la biosynthèse d'antibiotiques et qui forment un petit groupe à part. Dans un dernier sous-arbre de TCases, on trouve deux grands groupes de séquences : dans le premier sont réunies les acétylornithine TCases (AOTCases) et les succinylornithine TCases (SOTCases), qui interviennent dans des voies de biosynthèse de l'arginine (voies alternatives à celle utilisant l'OTCase anabolique). Dans le second on observe un un groupe de séquences qui ne correspondent à aucune enzyme de fonction connue, que j'ai appelé YTCases (quelquefois dénommées UTC dans des versions précédentes et dans des publications qui citent ce travail plus ancien[6]) parce que l'on y trouve entre autres la protéine YgeW qui semblait le candidat adéquat pour exprimer l'activité oxamate carbamoyltransférase.

L'étude plus approfondie que j'ai réalisée de cette grande famille (voir le Chapitre 7 précédent) a mis en évidence l'existence d'un autre groupe de TCases inconnues à la base des ATCases de type II (Figure 22B). Malgré la proximité phylogénétique de ces séquences avec les ATCases, plusieurs indices plaident pour une tout autre activité : des indices structurels d'abord, car la plupart des motifs spécifiques des ATCases correspondant à l'attachement du substrat à l'enzyme diffèrent des motifs trouvés chez toutes les ATCases (plus précisément l'attachement au groupe carboxyl), alors que tous les autres motifs caractéristiques des TCases sont conservés (voir Tableau 14 du Chapitre 7 précédent). Des indices génomiques ensuite, car tous les génomes présentant ces

6 UTC signifie *unknown TCase* (« TCase inconnue ») et n'est donc pas un nom adapté pour désigner les TCases particulières que j'ai préféré nommer « YTCases ».

TCases contiennent également une ATCase « normale ». Afin de déterminer la fonction hypothétique de ces pseudo-ATCases, j'ai étudié à nouveau le contexte génétique de chacune des séquences comme je l'ai fait avec les X1 et X2 dans la superfamille des dihydroorotates déshydrogénases. Cette analyse a montré que ces pseudo-ATCases pouvaient être divisées en trois sous-groupes, correspondant à trois sous-arbres (voir Figure 12 du Chapitre 7 précédent).

L'un des groupes de pseudo-ATCases en particulier contenait des séquences systématiquement en synténie avec une allantoïnase et une carbamate kinase. L'allantoïnase est une amidohydrolase cyclique, apparentée aux DHOases (Figure 22A), qui intervient dans une toute autre voie, la voie de dégradation des purines. Dans un des génomes en particulier, celui de l'Actinobactérie thermophile *Rubrobacter xylanophilus*, le gène de cette pseudo-ATCase est même voisin des tous les gènes successifs codant la dégradation des purines jusqu'à l'allantoïnase. Or depuis de nombreuses années, l'activité oxamate TCase (OxTCase) présente dans cette voie est recherchée sans que l'on puisse trouver son gène codant. Le candidat le plus prometteur jusqu'alors était le produit du gène *ygeW* d'E.coli qui fait partie des inconnues YTCases (Figure 22B). Cependant, les derniers travaux publiés [Li et al. 2011] et non publiés (présentation au colloque ICAP juillet 2010, Washington [Polo, Fita & Rubio 2010]) excluent que ce gène code une OxTCase, nous avons alors proposé que cette pseudo-ATCase de *R. thermophilus* code une activité oxamate TCase.

Nous avons donc lancé une collaboration avec Christianne Legrain pour étudier la pseudo-ATCase de *Rubrobacter thermophilus* afin d'en tester les propriétés. Les résultats actuels n'ont pas permis de conclure de manière décisive à l'existence d'une activité OxTCase. Cependant, ce résultat en apparence négatif m'a permis de proposer une nouvelle hypothèse. Afin de prédire la fonction de cette TCase d'activité inconnue, j'ai employé l'ensemble de mes outils et concepts précédemment décrits :

- Le contexte génétique laisse penser que cette pseudo-ATCase, à défaut d'être une OxTCase, jouerait un rôle dans une voie de dégradation des purines. En considérant le problème à la lumière du module réactionnel constitué de l'enchaînement d'une TCase et d'une amidohydrolase cyclique (ATCase + DHOase dans la voie de biosynthèse des pyrimidines), j'ai émis l'hypothèse que cette pseudo-ATCase catalysait l'étape suivant l'amidohydrolase correspondante (c'est-à-dire l'allantoïnase) dans la voie de dégradation des purines, puisque leurs deux gènes codants sont en synténie.

- Dans la voie classique de dégradation des purines, l'allantoate produite par l'allantoïnase est décarbamoylé en uréiodoglycine par une enzyme homologue de celle qui décarbamoyle le produit des DHPases dans les voies de dégradation des pyrimidines. Cette étape correspond à une décarbamoylation irréversible du substrat, contrairement aux TCases.
- Dans le modèle proposé, l'allantoate serait décarbamoylé de la même manière mais par la pseudo-ATCase. Les substrats et produits de cette « uréidoglycine carbamoyltransférase » putative sont également proches de ceux de l'ATCase : une comparaison des différentes molécules pouvant être la cible de TCases montre que l'uréidoglycine ressemble le plus à l'aspartate en terme de forme et de taille, ce qui vient justifier la proximité phylogénétique des ATCases et des pseudo-ATCases.

5.3 Voie de dégradation des purines chez Rubrobacter xylanophilus

Nous proposons que la pseudo-ATCase codée par *R. xylanophilus* serait une uréidoglycine carbamoyltransférase effectuant la dégradation de l'allantoïne en glyoxylate au cours de la voie catabolique des purines. Cette activité remplacerait l'allantoate amidohydrolase trouvée dans les autres espèces où l'uréidoglycine est ensuite transformée en uréiodoglycolate qui est lui-même dégradé en glyoxylate ou en oxamate. Chez d'autres espèces, comme *Bacillus subtilis* et *Klebsiella pneumoniae*, l'uréidoglycine réagit avec du glyoxylate pour former de l'oxalurate et de la glycine. Le glyoxylate de cette réaction provient de la décomposition spontanée de l'uréidoglycine. L'uréidoglycine est en effet instable : elle se décompose spontanément en ammoniac, urée et glyoxylate (Figure 14). Aucun des gènes pour les enzymes dégradant l'uréidoglycine précédemment décrits n'ont pu être trouvés dans le génome de *R. xylanophilus* (uréidoglycine aminohydrolase ou uréidoglycine-glyoxylate aminotransférase) avec une recherche par BLASTp en utilisant les séquences identifiées dans d'autres génomes. Il y a donc deux possibilités : soit il existe des gènes additionnels non identifiés qui permettraient de dégrader l'uréidoglycine en glyoxylate, soit la décomposition spontanée de l'uréidoglycine est suffisante pour produire du glyoxylate.

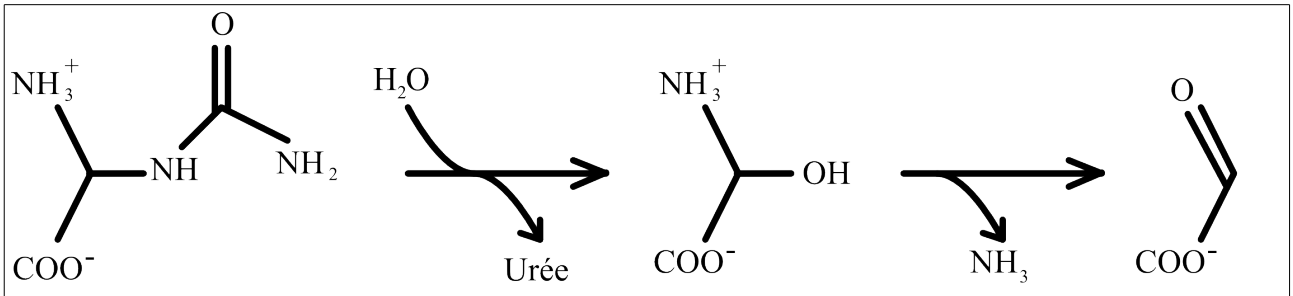


Figure 23 : Dégradation spontanée de l'uréidoglycine en urée, ammoniac et glyoxylate.

5.4 Origine des pseudo-ATCases

Les pseudo-ATCases sont à la base de l'arbre des ATCases, proches des ATCases de type II. Les trois types qui différent sur plusieurs critères détaillés ci-dessus sont potentiellement utilisables dans des voies cataboliques, par opposition aux ATCases qui sont exclusivement anaboliques. On retrouve ici le même scénario que chez les OTCases, qui peuvent être anaboliques ou cataboliques, les deux rôles étant joués par des OTCases distinctes. Il est donc possible que le gène ancestral codant cette TCase catabolique n'aurait été conservé que chez quelques espèces.

5.5 Multiples voies de dégradations

Les *Nocardiodes* sont connues pour pouvoir décomposer des composés chimiques industriels jugés polluants. Cela implique la présence de plusieurs voies de dégradation, dont la voie contenant une pseudo-ATCase et une YTCase peuvent faire partie. On notera la présence d'une amidohydrolase cyclique (probablement pas une ATCase) dans le même contexte de ces gènes, qui pourrait encore une fois être associé avec une des TCases.

Les gènes de la xanthine déshydrogénase (XDHase, gènes *xdhA, xdhB, xdhC*) étaient un des indices laissant penser que *ygeW* faisait partie de la voie de dégradation des purines. Or *Nocardioides sp. JS614* ne contient pas moins de 6 groupes de gènes homologues à ceux de la XDHase (homologues de *xdhB* : A1SD82, A1SD88, A1SE96, A1SEA6, A1SH64, A1SNT0 sur Uniprot), dont celui proche des TCases sus-mentionnées, ce qui laisse supposer qu'il y aurait autant de voies de dégradation potentielles associées. Les homologues de la XDHase peuvent avoir diverses fonctions d'oxydases, comme par exemple sur la caféine [Yu et al. 2008].

5.6 Les enzymes inconnues des trois familles dans l'hypothétique voie Yge

Les membres des trois superfamilles étudiées (Figure 22) se retrouvent donc dans plusieurs voies où les réactions qu'ils catalysent se succèdent. Bien que la grande majorité des séquences

puisse être annotée et dotée d'une fonction dans au moins une de ces voies métaboliques, il reste dans ces différentes superfamilles des séquences non identifiées dont le rôle biologique demeure inconnu. Remarquablement, une proportion conséquente des gènes codant ces séquences se retrouvent dans un même contexte génétique que j'ai surnommé Yge par utilisation des noms de gènes présents chez *E. coli*. On retrouve en effet les séquences de YTCases (YgeW de *E. coli*, Figure 22B) dans un même voisinage que des amidohydrolases cycliques (YgeZ, triangles sans labels proches des DHPases dans la Figure 22A) inconnues et des déshydrogénases X1 (Figure 22C). Les X1 ne sont cependant pas systématiquement présentes, et peuvent être remplacées par un homologue du partenaire de la DHPase (YgfK chez *E. coli*). Les deux contenant de clusters 4Fe-4S, l'idée est qu'une même fonction de transfert d'électron serait assurée par l'une ou par l'autre.

Nom du gène chez *E. coli*	Gène spécifique	Fonction proposée
ygeS, ygeT, ygeU	*xdhA, xdhB, xhdC*	Xanthine déshydrogénase
ygeW		**YTCase**
ygeX	*dpaL*	Diaminopropionate ammonia lyase
ygeY		Déacétylase/désuccinylase
ygeZ	***hyuA***	**Phénylhydantoinase**
ygfK / **X1**		**Ferredoxine**
ygfL	*ssnA*	?
ygfO	*xanQ*	Perméase de purines (ou pyrimidines)

Tableau 19 : Quelques une des gènes dans le contexte Yge et leur fonction hypothétique. Les noms en gras correspondent aux trois familles d'intérêt.

Dans ce contexte Yge se retrouvent d'autres familles de gènes de réactions diverses (Tableau 19). Il est donc probable que cet ensemble de gènes « Yge » forme une ou plusieurs voies métaboliques encore inconnues et qu'elles réutilisent l'enchaînement des gènes de TCase, amidohydrolase cyclique et de déshydrogénase des trois superfamilles que j'ai étudiées au cours de ma Thèse.

6 Puissance et pertinence de l'approche par modules réactionnels

L'approche d'une étude intégrée d'enzymes catalysant des étapes similaires successives de différentes voies à la lumière du concept de module réactionnel peut s'appliquer à d'autres voies

métaboliques, comme celles de l'arginine qui partagent des éléments communs avec la voie des pyrimidines. La Figure 24 montre comment les voies de l'arginine partagent un nombre significatif de gènes homologues codant des enzymes qui catalysent des réactions chimiquement équivalentes. En alignant l'ensemble des réactions similaires des différentes voies de biosynthèse et de dégradation connues de l'arginine (chez différents organismes) en se basant simplement sur les similarités de réaction et de substrat, on peut mettre en évidence des copies différentes d'enchaînement de réactions similaires (Figure 24B). Par exemple, ArgD, RocD, AstC, SpuC sont toutes des aminotransférases définissant la même superfamille. Elles sont suivies des carbamoyltransférases ArgF (OTCase), ArgF' (AOTCase), ArgF'' (SOTCase) et PtcA (PTCase). De nombreux autres parallèles sont possibles, y compris horizontalement (par exemple les décarboxylases SpeA et SpeC).

La mise en évidence de ces enchaînements parallèles ne s'arrête pas aux voies de l'arginine : en comparant la voie de biosynthèse de l'arginine avec la voie de biosynthèse AAA de la lysine (Figure 25), on constate un parallèle frappant entre l'enchaînement des réactions des enzymes ArgA à ArgE et celui regroupant les enzymes LysX à LysK. Les substrats sont similaires, la lysine ne se distinguant de l'ornithine que par un carbone en plus dans la chaîne du groupe latéral. Dans certains cas même, les deux voies de biosynthèse partagent une même enzyme pour une même réaction (par exemple ArgD qui est aussi LysJ chez certains organismes comme *E. coli* [Ledwidge & Blanchard 1999]).

Ces différentes observations sont en accord avec l'idée que ces modules réactionnels seraient le reflet des éléments de base d'une voie ancestrale qui auraient divergé par duplications successives et divergence des copies pour créer les voies spécialisées et indépendantes que l'on retrouve dans les organismes contemporains.

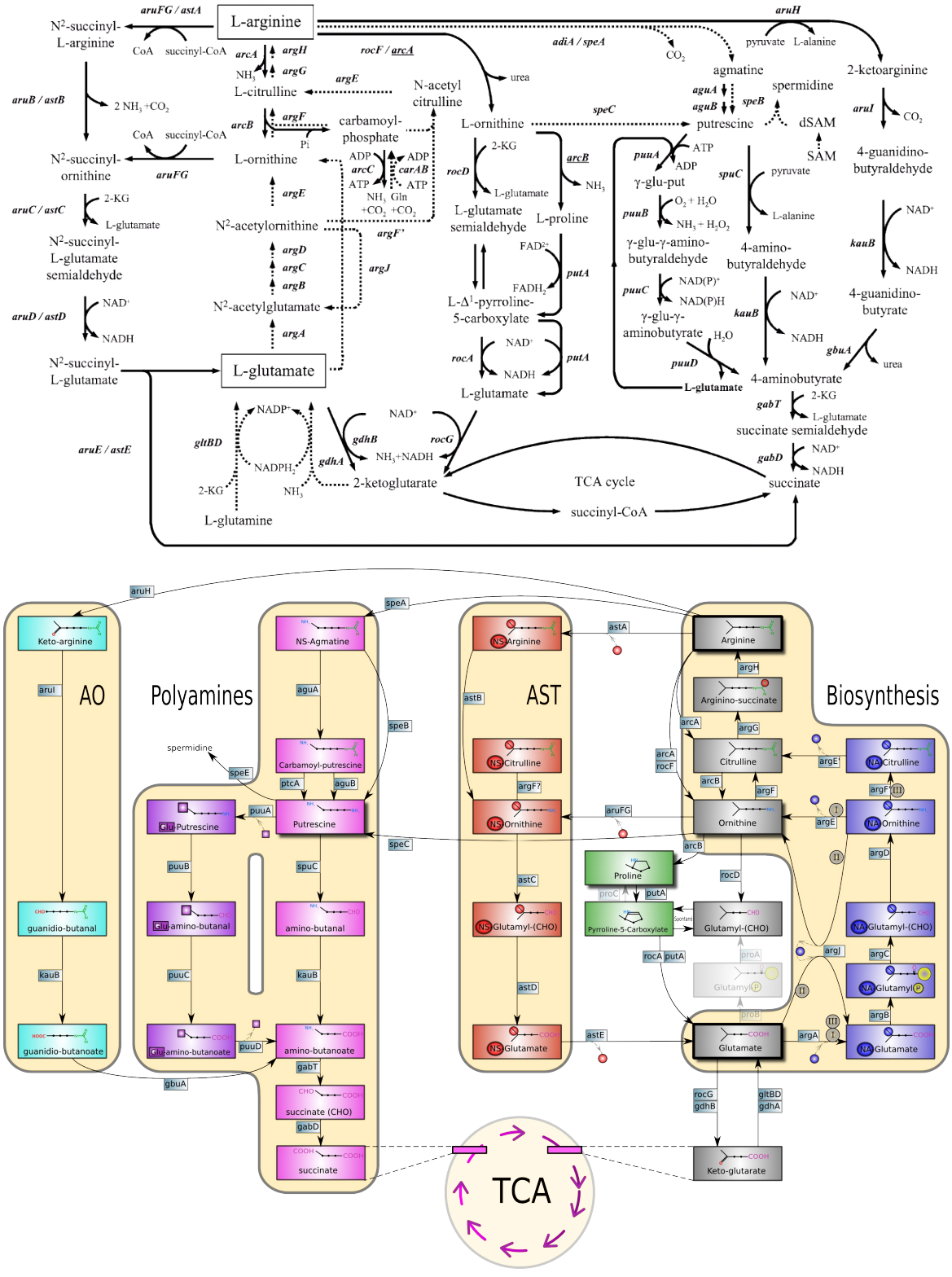


Figure 24 : Représentations des voies métaboliques de l'arginine, en considérant l'enchaînement des réactions et leurs enzymes (A), et l'enchaînement et la similitude des réactions et des substrats (B).

*(Légende de la Figure 24) le premier schéma par Lu [2006] représente l'enchaînement des différentes réactions. C'est un graphe dont les sommets sont les produits/substrats et les arêtes sont les réactions/enzymes. Le second schéma reprend le même graphe mais j'ai organisé les réactions et substrats pour qu'ils soient alignés avec d'autres réactions et substrats similaires. On retrouve donc sur une même ligne des substrats similaires (par exemple ornithine et acétylornithine) ou des réactions similaires (*ArgF *et* ArgF').

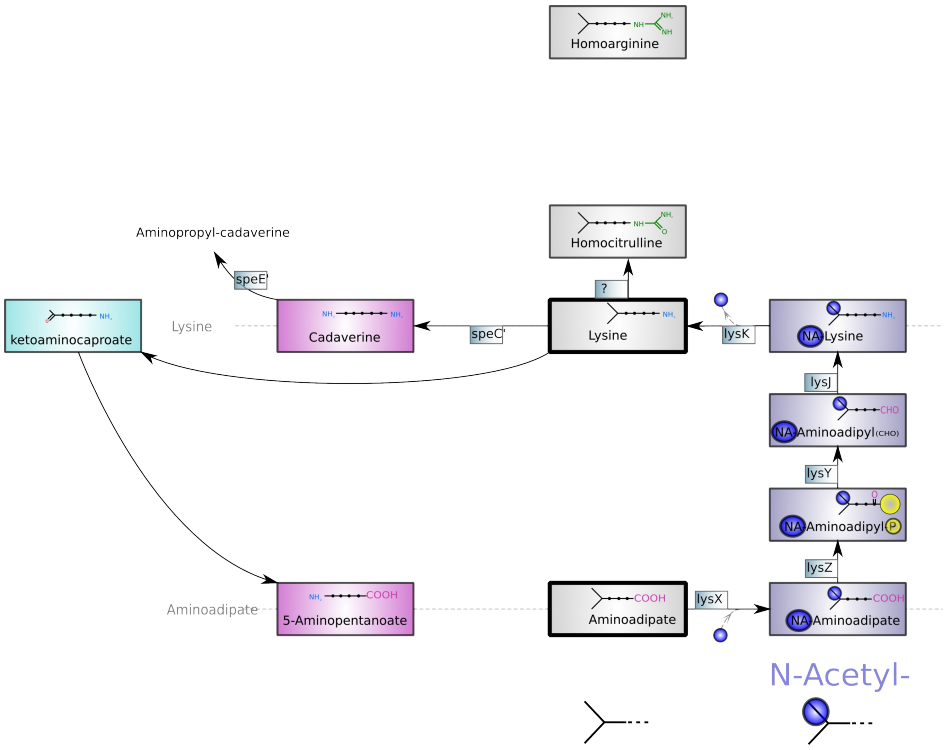


Figure 25 : Représentation d'une partie de la voie de biosynthèse de la lysine (groupe diaminoadipate procaryote) contenant un module réactionnel similaire à ceux des voies métaboliques de l'arginine.
L'arrangement des réactions et substrats de ces voies est calquée sur celui des voies de l'arginine. Les réactions de la partie acétylée sont similaires à celles de la voie de biosynthèse de l'arginine.

CONCLUSION

Ces dernières années le nombre de séquences protéiques publiées à augmenté de manière phénoménale, qu'elles soient issues de génomes entiers ou de métagénomes. Ils n'y a cependant que quelques séquences pour lesquelles on dispose de preuves expérimentales de leurs fonctions, alors qu'il y a désormais des millions de séquences disponibles. L'annotation fonctionnelle automatique, basée sur de simples comparaison de séquences deux à deux, a vite montré ses limites. Une annotation correcte nécessite en effet de prendre en compte deux concepts importants :

Le premier est la plasticité des enzymes et leur histoire évolutive complexe. Plutôt que de considérer simplement qu'une enzyme catalyse une réaction spécifique, il faut prendre en compte le fait que nombre d'enzymes peuvent utiliser différents substrats (avec une plus ou moins bonne efficacité) pour une même réaction. Ceci est d'autant plus vrai que l'on remonte vers des séquences de plus en plus ancestrales. Cette ambigüité de substrat doit être prise en compte tant pour l'annotation des séquences que pour leur histoire évolutive et l'analyse de la biochimie de ces différentes réactions. En effet, une enzyme peut être suffisamment ambigüe pour intervenir dans plusieurs voies métaboliques.

Cette plasticité ouvre sur un second concept, qui est la capacité des systèmes biologiques à être modulaires. Les protéines ne doivent plus être considérées isolément (1 enzyme = 1 fonction) mais dans un contexte biologique et cellulaire. Elles font partie de voies métaboliques dans lesquelles elles s’enchaînent et interagissent pour assurer la fonction de la voie : l'ensemble des enzymes joue alors un rôle commun et constitue un module fonctionnel. Cette modularité que l'on retrouve dans tout le vivant à différents niveaux (métamères des animaux, modules des plantes, voies de signalisation, voies métaboliques...) se caractérise par un ensemble d'éléments formant un tout difficile à dissocier. Les modules réactionnels que j'ai définis sont encore un nouveau type de modules, composants élémentaires de voies métaboliques et vestiges de voies ancestrales, dans lesquelles les réactions similaires catalysées par des enzymes homologues s’enchaînent.

Une annotation fonctionnelle significative nécessite donc des approches poussées qui prennent en compte le contexte phylogénétique (l'histoire évolutive), modulaire (les voies métaboliques) et biochimique (la similarité des réactions et des substrats) des enzymes étudiées.

De tels concepts et approches enrichissent la compréhension des mécanismes d'évolution du métabolisme, Ils sont également partie prenante de la biologie des systèmes qui considère ces voies métaboliques comme autant de réseaux plus ou moins interconnectés dans un ensemble hautement intégré et régulé.

BIBLIOGRAPHIE

1. Ahuja A, Purcarea C, Guy HI & Evans DR (**2001**). *A novel carbamoyl-phosphate synthetase from Aquifex aeolicus*. J Biol Chem, 276: 45694-45703.
2. Altschul SF, Gish W, Miller W, Myers EW & Lipman DJ (**1990**). *Basic local alignment search tool*. J Mol Biol, 215: 403-410.
3. Armougom F, Moretti S, Poirot O, Audic S, Dumas P, Schaeli B, Keduas V & Notredame C (**2006**). *Expresso: automatic incorporation of structural information in multiple sequence alignments using 3D-Coffee*. Nucleic Acids Res, 34: W604-W608.
4. Backman TWH, Cao Y & Girke T (**2011**). *ChemMine tools: an online service for analyzing and clustering small molecules*. Nucleic Acids Res, 39: W486-W491.
5. Beadle GW & Tatum EL (**1941**). *Genetic Control of Biochemical Reactions in Neurospora*. Proc Natl Acad Sci U S A, 27: 499-506.
6. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J & Sayers EW (**2011**). *GenBank*. Nucleic Acids Res, 39: D32-D37.
7. Boxstael SV, Cunin R, Khan S & Maes D (**2003**). *Aspartate transcarbamylase from the hyperthermophilic archaeon Pyrococcus abyssi: thermostability and 1.8A resolution crystal structure of the catalytic subunit complexed with the bisubstrate analogue N-phosphonacetyl-L-aspartate*. J Mol Biol, 326: 203-216.
8. Chen J, Cheng C, Xia Y, Zhao H, Fang C, Shan Y, Wu B & Fang W (**2011**). *Lmo0036, an ornithine and putrescine carbamoyltransferase in Listeria monocytogenes, participates in arginine deiminase and agmatine deiminase pathways and mediates acid tolerance*. Microbiology, 157: 3150-3161.
9. Chen L & Vitkup D (**2006**). *Predicting genes for orphan metabolic activities using phylogenetic profiles*. Genome Biol, 7: R17.
10. Cordón F, (**1990**). *Tratado Evolucionista de Biologia*. Anthropos, Editorial del Hombre. Aguilar, Madrid, Spain.
11. Descorps-Declère S, Lemoine F, Sculo Q, Lespinet O & Labedan B (**2008**). *The multiple facets of homology and their use in comparative genomics to study the evolution of genes, genomes, and species*. Biochimie, 90: 595-608.
12. Durbecq V, Legrain C, Roovers M, Piérard A & Glansdorff N (**1997**). *The carbamate kinase-like carbamoyl phosphate synthetase of the hyperthermophilic archaeon Pyrococcus furiosus, a missing link in the evolution of carbamoyl phosphate biosynthesis*. Proc Natl Acad Sci U S A, 94: 12803-12808.
13. Eddy SR (**2009**). *A new generation of homology search tools based on probabilistic inference*. Genome Inform, 23: 205-211.
14. Edgar RC (**2004**). *MUSCLE: multiple sequence alignment with high accuracy and high throughput*. Nucleic Acids Res, 32: 1792-1797.
15. Edgar RC & Batzoglou S (**2006**). *Multiple sequence alignment*. Curr Opin Struct Biol, 16: 368-373.

16. Eisen JA (**1998**). *Phylogenomics: improving functional predictions for uncharacterized genes by evolutionary analysis*. Genome Res, 8: 163-167.

17. Emmert EAB, Klimowicz AK, Thomas MG & Handelsman J (**2004**). *Genetics of zwittermicin a production by Bacillus cereus*. Appl Environ Microbiol, 70: 104-113.

18. Fields C, Brichta D, Shepherdson M, Farinha M & O'Donovan G (**1999**). *Phylogenetic Analysis and Classification of Dihydroorotases: A Complex History for a Complex Enzyme*. Paths to pyrimidines - An international newsletter: 49-63.

19. Finn RD, Mistry J, Tate J, Coggill P, Heger A, Pollington JE, Gavin OL, Gunasekaran P, Ceric G, Forslund K, Holm L, Sonnhammer ELL, Eddy SR & Bateman A (**2010**). *The Pfam protein families database*. Nucleic Acids Res, 38: D211-D222.

20. Fitch WM (**1970**). *Distinguishing homologous from analogous proteins*. Syst Zool, 19: 99-113.

21. Force A, Lynch M, Pickett FB, Amores A, Yan YL & Postlethwait J (**1999**). *Preservation of duplicate genes by complementary, degenerative mutations*. Genetics, 151: 1531-1545.

22. Furnham N, Garavelli JS, Apweiler R & Thornton JM (**2009**). *Missing in action: enzyme functional annotations in biological databases*. Nat Chem Biol, 5: 521-525.

23. Granick S (**1957**). *Speculations on the origins and evolution of photosynthesis*. Ann N Y Acad Sci, 69: 292-308.

24. Gray GS & Fitch WM (**1983**). *Evolution of antibiotic resistance genes: the DNA sequence of a kanamycin resistance gene from Staphylococcus aureus*. Mol Biol Evol, 1: 57-66.

25. Haldane JBS (**1954**). *The origin of life*. New Biology, 16: 12-27.

26. Hanson AD, Pribat A, Waller JC & de Crécy-Lagard V (**2010**). *'Unknown' proteins and 'orphan' enzymes: the missing half of the engineering parts list--and how to find it*. Biochem J, 425: 1-11.

27. Hartwell LH, Hopfield JJ, Leibler S & Murray AW (**1999**). *From molecular to modular cell biology*. Nature, 402: C47-C52.

28. Hayaishi O & Konrberg A (**1952**). *Metabolism of cytosine, thymine, uracil, and barbituric acid by bacterial enzymes*. J Biol Chem, 197: 717-732.

29. Horowitz NH (**1945**). *On the Evolution of Biochemical Syntheses*. Proc Natl Acad Sci U S A, 31: 153-157.

30. Hu P, Janga SC, Babu M, Díaz-Mejía JJ, Butland G, Yang W, Pogoutse O, Guo X, Phanse S, Wong P, Chandran S, Christopoulos C, Nazarians-Armavil A, Nasseri NK, Musso G, Ali M, Nazemof N, Eroukova V, Golshani A, Paccanaro A, Greenblatt JF, Moreno-Hagelsieb G & Emili A (**2009**). *Global functional atlas of Escherichia coli encompassing previously uncharacterized proteins*. PLoS Biol, 7: e96.

31. Hulo N, Bairoch A, Bulliard V, Cerutti L, Castro ED, Langendijk-Genevaux PS, Pagni M & Sigrist CJA (**2006**). *The PROSITE database*. Nucleic Acids Res, 34: D227-D230.

32. Huson DH, Richter DC, Rausch C, Dezulian T, Franz M & Rupp R (**2007**). *Dendroscope: An interactive viewer for large phylogenetic trees*. BMC Bioinformatics, 8: 460.

33. Jensen RA (**1976**). *Enzyme recruitment in evolution of new function*. Annu Rev Microbiol, 30: 409-425.

34. Kantrowitz ER & Lipscomb WN (**1990**). *Escherichia coli aspartate transcarbamoylase: the molecular basis for a concerted allosteric transition*. Trends Biochem Sci, 15: 53-59.

35. Karp PD, Riley M, Paley SM & Pellegrini-Toole A (**2002**). *The MetaCyc Database*. Nucleic Acids Res, 30: 59-61.

36. Kemena C & Notredame C (**2009**). *Upcoming challenges for multiple sequence alignment methods in the high-throughput era*. Bioinformatics, 25: 2455-2465.

37. Koonin EV (**2005**). *Orthologs, paralogs, and evolutionary genomics*. Annu Rev Genet, 39: 309-338.

38. Koski LB & Golding GB (**2001**). *The closest BLAST hit is often not the nearest neighbor*. J Mol Evol, 52: 540-542.

39. Kyrpides N, Overbeek R & Ouzounis C (**1999**). *Universal protein families and the functional content of the last universal common ancestor*. J Mol Evol, 49: 413-423.

40. Labedan B, Boyen A, Baetens M, Charlier D, Chen P, Cunin R, Durbeco V, Glansdorff N, Herve G, Legrain C, Liang Z, Purcarea C, Roovers M, Sanchez R, Toong TL, de Casteele MV, van Vliet F, Xu Y & Zhang YF (**1999**). *The evolutionary history of carbamoyltransferases: A complex set of paralogous genes was already present in the last universal common ancestor*. J Mol Evol, 49: 461-473.

41. Labedan B, Xu Y, Naumoff DG & Glansdorff N (**2004**). *Using quaternary structures to assess the evolutionary history of proteins: the case of the aspartate carbamoyltransferase*. Mol Biol Evol, 21: 364-373.

42. Laidlaw SA & Kopple JD (**1987**). *Newer concepts of the indispensable amino acids*. Am J Clin Nutr, 46: 593-605.

43. Lazcano & Forterre (**1999**). *The molecular search for the last common ancestor*. J Mol Evol, 49: 411-412.

44. Lazcano A & Miller SL (**1996**). *The origin and early evolution of life: prebiotic chemistry, the pre-RNA world, and time*. Cell, 85: 793-798.

45. Ledwidge R & Blanchard JS (**1999**). *The dual biosynthetic capability of N-acetylornithine aminotransferase in arginine and lysine biosynthesis*. Biochemistry, 38: 3019-3024.

46. Lespinet O & Labedan B (**2005**). *Orphan enzymes?*. Science, 307: 42.

47. Lespinet O & Labedan B (**2006**). *ORENZA: a web resource for studying ORphan ENZyme activities*. BMC Bioinformatics, 7: 436.

48. Li X, Weinstock GM & Murray BE (**1995**). *Generation of auxotrophic mutants of Enterococcus faecalis*. J Bacteriol, 177: 6866-6873.

49. Li Y, Jin Z, Yu X, Allewell NM, Tuchman M & Shi D (**2011**). *The ygeW encoded protein from Escherichia coli is a knotted ancestral catabolic transcarbamylase*. Proteins, 79: 2327-2334.

50. Llácer JL, Polo LM, Tavárez S, Alarcón B, Hilario R & Rubio V (**2007**). *The gene cluster for agmatine catabolism of Enterococcus faecalis: study of recombinant putrescine transcarbamylase and agmatine deiminase and a snapshot of agmatine deiminase catalyzing its reaction*. J Bacteriol, 189: 1254-1265.

51. Miller SL & Urey HC (**1959**). *Organic compound synthesis on the primitive earth*. Science, 130: 245-251.

52. Minic Z, Simon V, Penverne B, Gaill F & Hervé G (**2001**). *Contribution of the bacterial endosymbiont to the biosynthesis of pyrimidine nucleotides in the deep-sea tube worm Riftia pachyptila*. J Biol Chem, 276: 23777-23784.

53. Miroshnichenko ML, Hippe H, Stackebrandt E, Kostrikina NA, Chernyh NA, Jeanthon C, Nazina TN, Belyaev SS & Bonch-Osmolovskaya EA (**2001**). *Isolation and characterization of Thermococcus sibiricus sp. nov. from a Western Siberia high-temperature oil reservoir*. Extremophiles, 5: 85-91.

54. Monod J, Wyman J & Changeux JP (**1965**). *On the nature of allosteric transitions: a plausible model*. J Mol Biol, 12: 88-118.

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

56. Ohno S, (**1970**). *Evolution by Gene duplication*. Springer-Verlag. New-York.

57. Overbeek R, Fonstein M, D'Souza M, Pusch GD & Maltsev N (**1999**). *The use of gene clusters to infer functional coupling*. Proc Natl Acad Sci U S A, 96: 2896-2901.

58. Pace NR (**1997**). *A molecular view of microbial diversity and the biosphere*. Science, 276: 734-740.

59. Pace NR (**2001**). *The universal nature of biochemistry*. Proc Natl Acad Sci U S A, 98: 805-808.

60. Parales RE & Ingraham JL (**2010**). *The surprising Rut pathway: an unexpected way to derive nitrogen from pyrimidines*. J Bacteriol, 192: 4086-4088.

61. Patterson C (**1988**). *Homology in classical and molecular biology*. Mol Biol Evol, 5: 603-625.

62. Pei J (**2008**). *Multiple protein sequence alignment*. Curr Opin Struct Biol, 18: 382-386.

63. Pennazio S (**2009**). *Alexandr Oparin and the origin of life on Earth*. Riv Biol, 102: 95-118.

64. Phan IQH, Pilbout SF, Fleischmann W & Bairoch A (**2003**). *NEWT, a new taxonomy portal*. Nucleic Acids Res, 31: 3822-3823.

65. Pirovano W & Heringa J (**2008**). *Multiple sequence alignment*. Methods Mol Biol, 452: 143-161.

66. Polo LM, Fita I & Rubio V (**2010**). *A Structure in search for a function: the crystal structure of an orphan transcarbamoylase*. The International Conference on Arginine and Pyrimidines: 24.

67. Pouliot Y & Karp PD (**2007**). *A survey of orphan enzyme activities*. BMC Bioinformatics, 8: 244.

68. Pruitt KD, Tatusova T & Maglott DR (**2007**). *NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins*. Nucleic Acids Res, 35: D61-D65.

69. Qin J, Li R, Raes J, Arumugam M, Burgdorf KS, Manichanh C, Nielsen T, Pons N, Levenez F, Yamada T, Mende DR, Li J, Xu J, Li S, Li D, Cao J, Wang B, Liang H, Zheng H, Xie

Y, Tap J, Lepage P, Bertalan M, Batto J-M, Hansen T, Paslier DL, Linneberg A, Nielsen HB, Pelletier E, Renault P, Sicheritz-Ponten T, Turner K, Zhu H, Yu C, Li S, Jian M, Zhou Y, Li Y, Zhang X, Li S, Qin N, Yang H, Wang J, Brunak S, Doré J, Guarner F, Kristiansen K, Pedersen O, Parkhill J, Weissenbach J, Consortium MIT, Bork P, Ehrlich SD & Wang J (**2010**). *A human gut microbial gene catalogue established by metagenomic sequencing*. Nature, 464: 59-65.

70. Ramazzina I, Costa R, Cendron L, Berni R, Peracchi A, Zanotti G & Percudani R (**2010**). *An aminotransferase branch point connects purine catabolism to amino acid recycling*. Nat Chem Biol, 6: 801-806.

71. Sayers EW, Barrett T, Benson DA, Bolton E, Bryant SH, Canese K, Chetvernin V, Church DM, DiCuccio M, Federhen S, Feolo M, Fingerman IM, Geer LY, Helmberg W, Kapustin Y, Landsman D, Lipman DJ, Lu Z, Madden TL, Madej T, Maglott DR, Marchler-Bauer A, Miller V, Mizrachi I, Ostell J, Panchenko A, Phan L, Pruitt KD, Schuler GD, Sequeira E, Sherry ST, Shumway M, Sirotkin K, Slotta D, Souvorov A, Starchenko G, Tatusova TA, Wagner L, Wang Y, Wilbur WJ, Yaschenko E & Ye J (**2011**). *Database resources of the National Center for Biotechnology Information*. Nucleic Acids Res, 39: D38-D51.

72. Schnoes AM, Brown SD, Dodevski I & Babbitt PC (**2009**). *Annotation error in public databases: misannotation of molecular function in enzyme superfamilies*. PLoS Comput Biol, 5: e1000605.

73. Serventi F, Ramazzina I, Lamberto I, Puggioni V, Gatti R & Percudani R (**2010**). *Chemical basis of nitrogen recovery through the ureide pathway: formation and hydrolysis of S-ureidoglycine in plants and bacteria*. ACS Chem Biol, 5: 203-214.

74. Sonnhammer ELL & Koonin EV (**2002**). *Orthology, paralogy and proposed classification for paralog subtypes*. Trends Genet, 18: 619-620.

75. Soong C-L, Ogawa J, Sakuradani E & Shimizu S (**2002**). *Barbiturase, a novel zinc-containing amidohydrolase involved in oxidative pyrimidine metabolism*. J Biol Chem, 277: 7051-7058.

76. Stajich JE, Block D, Boulez K, Brenner SE, Chervitz SA, Dagdigian C, Fuellen G, Gilbert JGR, Korf I, Lapp H, Lehväslaiho H, Matsalla C, Mungall CJ, Osborne BI, Pocock MR, Schattner P, Senger M, Stein LD, Stupka E, Wilkinson MD & Birney E (**2002**). *The Bioperl toolkit: Perl modules for the life sciences*. Genome Res, 12: 1611-1618.

77. Sterk P, Kulikova T, Kersey P & Apweiler R (**2007**). *The EMBL Nucleotide Sequence and Genome Reviews Databases*. Methods Mol Biol, 406: 1-22.

78. Thompson JD, Higgins DG & Gibson TJ (**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.

79. UniProt Consortium (**2011**). *Ongoing and future developments at the Universal Protein Resource*. Nucleic Acids Res, 39: D214-D219.

80. Woese C (**1998**). *The universal ancestor*. Proc Natl Acad Sci U S A, 95: 6854-6859.

81. Woese CR, Kandler O & Wheelis ML (**1990**). *Towards a natural system of organisms: proposal for the domains Archaea, Bacteria, and Eucarya*. Proc Natl Acad Sci U S A, 87: 4576-4579.

82. Ycas M (**1974**). *On earlier states of the biochemical system*. J Theor Biol, 44: 145-160.

83. Yu CL, Kale Y, Gopishetty S, Louie TM & Subramanian M (**2008**). *A novel caffeine dehydrogenase in Pseudomonas sp. strain CBB1 oxidizes caffeine to trimethyluric acid*. J Bacteriol, 190: 772-776.

ANNEXES

1 Corbank

Au début de mon travail de préparation au Doctorat je me suis heurté au problème de déterminer les équivalences des identifiants de séquences entre deux bases de génomes complets. Ce travail m'a permis de mettre en évidence les problèmes d'annotation structurale existant entre ces bases publiques.

Il existe deux grandes bases de génomes complets : RefSeq[7] (NCBI) [Pruitt, Tatusova & Maglott 2007] et GenomeReviews[8] (EBI) [Sterk et al. 2007]. Le but de ces bases est de fournir des annotations homogènes pour tous les génomes publiés en utilisant les mêmes méthodes systématiquement, et ce afin de servir de support solide à l'annotation des nouveaux génomes publiés. Cependant, nous avons montré que ces deux bases de données publiques présentent des différences non triviales, si bien qu'il n'est pas simple de trouver la correspondance entre les CDS[9] d'une base à l'autre.

Pour résoudre cette question, j'ai développé (avec Stéphane Descorps-Declère) un programme autonome appelé *Corbank* qui permet de lister immédiatement les identifiants correspondants entre les deux bases pour chaque génome complet publié. La méthode développée permet de trouver les séquences qui correspondent exactement (utilisation de tables de hachage), mais aussi les cas où celles-ci diffèrent au niveau de leur séquence (distances euclidiennes). Dans ce deuxième cas, le programme détecte le type de différence (séquences de taille différente, différences au début ou en fin de séquence, etc.).

Appliquée à l'ensemble des génomes disponibles, *Corbank* permet de créer très rapidement les correspondances entre identifiants et de quantifier les différences au niveau des génomes et au niveau des bases en général. Il a ainsi été montré entre autres que 1% des séquences correspondantes diffèrent, et que ces différences sont majoritairement dues à un décalage entre les différents codons initiateurs possibles.

Le programme et les résultats sont consultables sur un site web dédié et régulièrement

7 RefSeq est accessible à l'adresse suivante : http://www.ncbi.nlm.nih.gov/RefSeq/
8 GenomeReviews est accessible à l'adresse suivante : http://www.ebi.ac.uk/GenomeReviews/
9 CDS : *Coding sequence*, « séquence codante ». Partie d'une séquence nucléotidique qui code une protéine.

mis à jour : http://www.corbank.u-psud.fr.

1.1 Résumé

Le détails de la méthode implémentée dans *Corbank* a été publié dans le papier (en anglais) qui suit :

- Descorps-Declère S*, Barba M* & Labedan B **(2008)**. *Matching curated genome database: a non trivial task.* BMC Genomics, 2008: 9:501. [* : contributions égales]
    - Ce papier a été noté comme 

(« très consulté » sur le site de BMC : http://www.biomedcentral.com/1471-2164/9/501)

BMC Genomics

BioMed Central

Research article

Open Access

Matching curated genome databases: a non trivial task

Stéphane Descorps-Declère†, Matthieu Barba† and Bernard Labedan*

Address: Institut de Génétique et Microbiologie, Université Paris Sud XI, CNRS UMR 8621, Bât. 400, 91405 Orsay Cedex, France

Email: Stéphane Descorps-Declère - stephane.descorps-declere@igmors.u-psud.fr; Matthieu Barba - matthieu.barba@igmors.u-psud.fr; Bernard Labedan* - bernard.labedan@igmors.u-psud.fr

* Corresponding author †Equal contributors

Published: 24 October 2008

BMC Genomics 2008, **9**:501 doi:10.1186/1471-2164-9-501

Received: 12 June 2008
Accepted: 24 October 2008

This article is available from: http://www.biomedcentral.com/1471-2164/9/501

Abstract

Background: Curated databases of completely sequenced genomes have been designed independently at the NCBI (RefSeq) and EBI (Genome Reviews) to cope with non-standard annotation found in the version of the sequenced genome that has been published by databanks GenBank/EMBL/DDBJ. These curation attempts were expected to review the annotations and to improve their pertinence when using them to annotate newly released genome sequences by homology to previously annotated genomes. However, we observed that such an uncoordinated effort has two unwanted consequences. First, it is not trivial to map the protein identifiers of the same sequence in both databases. Secondly, the two reannotated versions of the same genome differ at the level of their structural annotation.

Results: Here, we propose CorBank, a program devised to provide cross-referencing protein identifiers no matter what the level of identity is found between their matching sequences. Approximately 98% of the 1,983,258 amino acid sequences are matching, allowing instantaneous retrieval of their respective cross-references. CorBank further allows detecting any differences between the independently curated versions of the same genome. We found that the RefSeq and Genome Reviews versions are perfectly matching for only 50 of the 641 complete genomes we have analyzed. In all other cases there are differences occurring at the level of the coding sequence (CDS), and/or in the total number of CDS in the respective version of the same genome.

CorBank is freely accessible at http://www.corbank.u-psud.fr. The CorBank site contains also updated publication of the exhaustive results obtained by comparing RefSeq and Genome Reviews versions of each genome. Accordingly, this web site allows easy search of cross-references between RefSeq, Genome Reviews, and UniProt, for either a single CDS or a whole replicon.

Conclusion: CorBank is very efficient in rapid detection of the numerous differences existing between RefSeq and Genome Reviews versions of the same curated genome. Although such differences are acceptable as reflecting different views, we suggest that curators of both genome databases could help reducing further divergence by agreeing on a minimal dialogue and attempting to publish the point of view of the other database whenever it is technically possible.

Background

Public genomic databanks are inexorably inundated by newly sequenced genomes. The number of complete sequence of prokaryotic genomes that are published per year has increased more than tenfold in the last seven years with a present rate close to four newly published prokaryotic genomes per week. One of the main challenges encountered by genome databanks is that complete genomic sequences are submitted with a heterogeneous and (too) often crude gene annotation [1-4]. To cope with these major problems and to improve the representation of genomic information, NCBI and EBI are proposing curated versions, the Reference Sequence (RefSeq) [5] and Genome Reviews [6], respectively. Each database team is working independently but they share the same main goal of delivering an up-to-date, standardized and comprehensive view of the completely sequenced genomes that are present in the International Nucleotide Sequence Database (INSD) repository (GenBank/EMBL/DDBJ),

To facilitate the use of these standardized genomic data in comparative genomics studies, both RefSeq and Genome Reviews include manually curated information. Noticeably, RefSeq and Genome Reviews provide cross-references to public databases to facilitate database searches. Interestingly, many of these cross-references (/db_xref) are specific to the curated database: for instance, RefSeq has/db_xref to Entrez [7] and often to CDD [8], whereas Genome Reviews has/db_xref to Gene Ontology [9], InterPro [10], and UniProt [11], and occasionally to HOGENOM [12], and PDB [13].

Thus, it would be advantageous to work with both curated databases since they look more complementary than concurrent. However, there is no immediate way to match the respective sequence identifiers listed by either RefSeq or Genome Reviews for the same gene of the same reannotated genome, although the knowledgebase UniProt [11] began to add links to both genome databases as this paper was in preparation. Moreover, the independent efforts of NCBI and EBI curators in improving the structural annotation of a few CDS, lead to increasingly different genomic versions of the same organism. Three different instances are expected when comparing the structural annotations made independently by RefSeq and Genome Reviews curators: (i) the amino acid sequences are exactly identical, (ii) both CDS share an overlapping identical segment but differ in length, (iii) a few CDS are found exclusively in one genome database. This last instance corresponds often to the redefinition of a putative CDS as being a pseudogene on the basis of structural features.

We aimed to obtain immediate and exhaustive cross-references of each protein-coding gene when dealing with such possible divergences that reflect different points of view between RefSeq and Genome Reviews. Accordingly, we designed CorBank, a software (see [14]) that detects not only perfect identities but also any differences between RefSeq and Genome Reviews databases.

Results

Complete sequences of each replicon of each prokaryotic organism endowed with the same Taxonomy ID in both RefSeq and Genome Reviews were downloaded from each database and mapped by their common INSD identification numbers. Then, as schematized on Fig. 1, we compared both database versions of the same genomic data to identify the cross-references for each gene and to measure their level of matching. Accordingly, the different scripts that make up the CorBank program [14] were applied to these mapped data in two successive steps in order first to find exact matches and then to identify the nature and location of any difference in imperfect matches.

Matching gene sequences in independently curated genome databases

To be as fast as possible, we did not compare the sequence partners by using efficient but slow programs such as BLASTClust [15]. Rather, we used the Perl language to build hash tables where each amino acid sequence is a key that indexes its encoding CDS. Matching is straightforward when the same key is found for the two versions of the same gene sequence – one in RefSeq and the other in Genome Reviews (Fig. 1, yellow part). In rare instances, more than two identical sequences were found for the two versions of the same genome. This occurred for example with strictly identical insertion sequences present at different locations on the analyzed genome. Moreover, we could not dismiss the hypothesis that in very very rare cases pairs of completely conserved paralogues could form bidirectional best matches that may be erroneously interpreted. To handle these problems, we further used the respective gene positions to identify the pertinent couples of corresponding sequences (Fig. 1, yellow part).

Using this approach based on hash tables, we found that 98% of copies of the 1,983,258 genes described in both databases are matching, allowing instantaneous retrieval of their respective cross-references (see, for instance, Fig. 2 Table C).

However, the view was more contrasted when comparing complete genome annotation instead of looking at each individual gene. Table 1 shows that only 50 of the 641 complete genomes we have analyzed are perfectly matching at the level of their structural annotation. The other ones differ in terms of their respective total number of sequences and/or distribution of perfect matching sequences (Table 1). The copies in both curated databases of 260 genomes differ by their total numbers of genes and

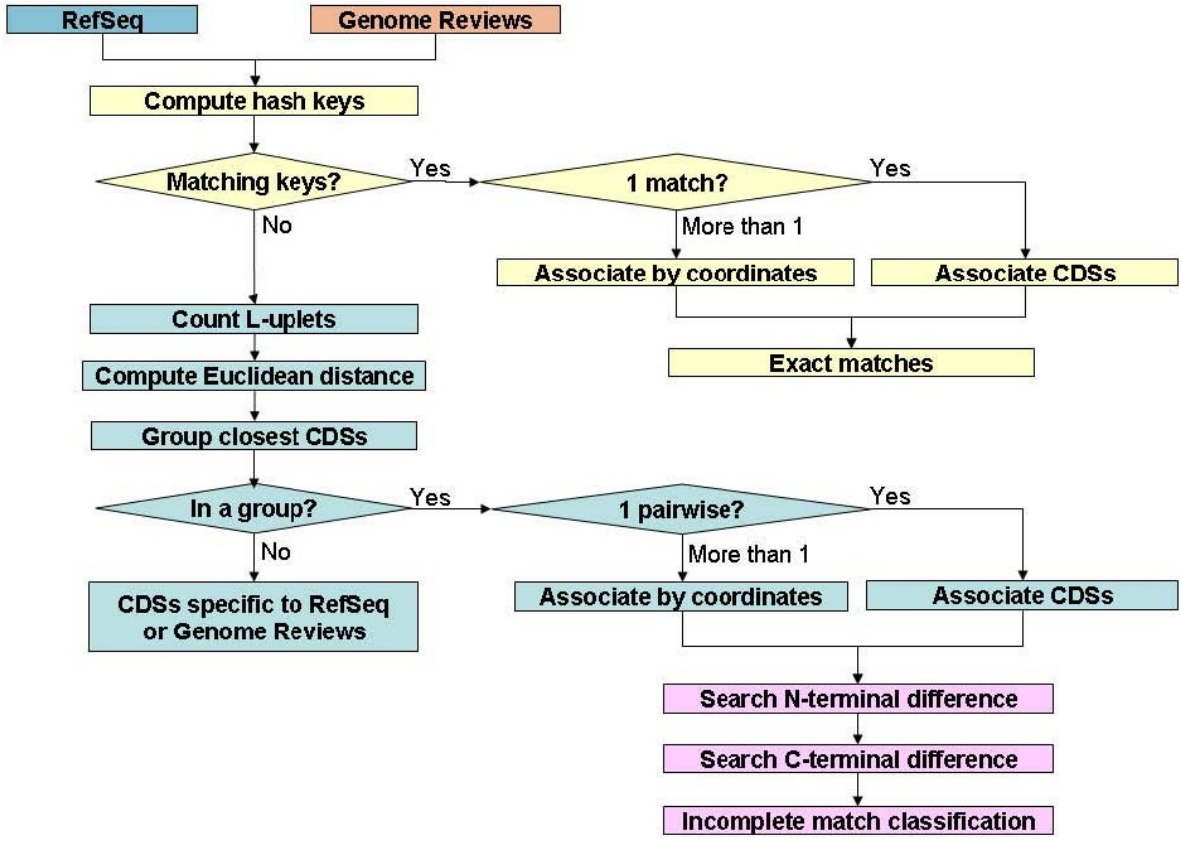
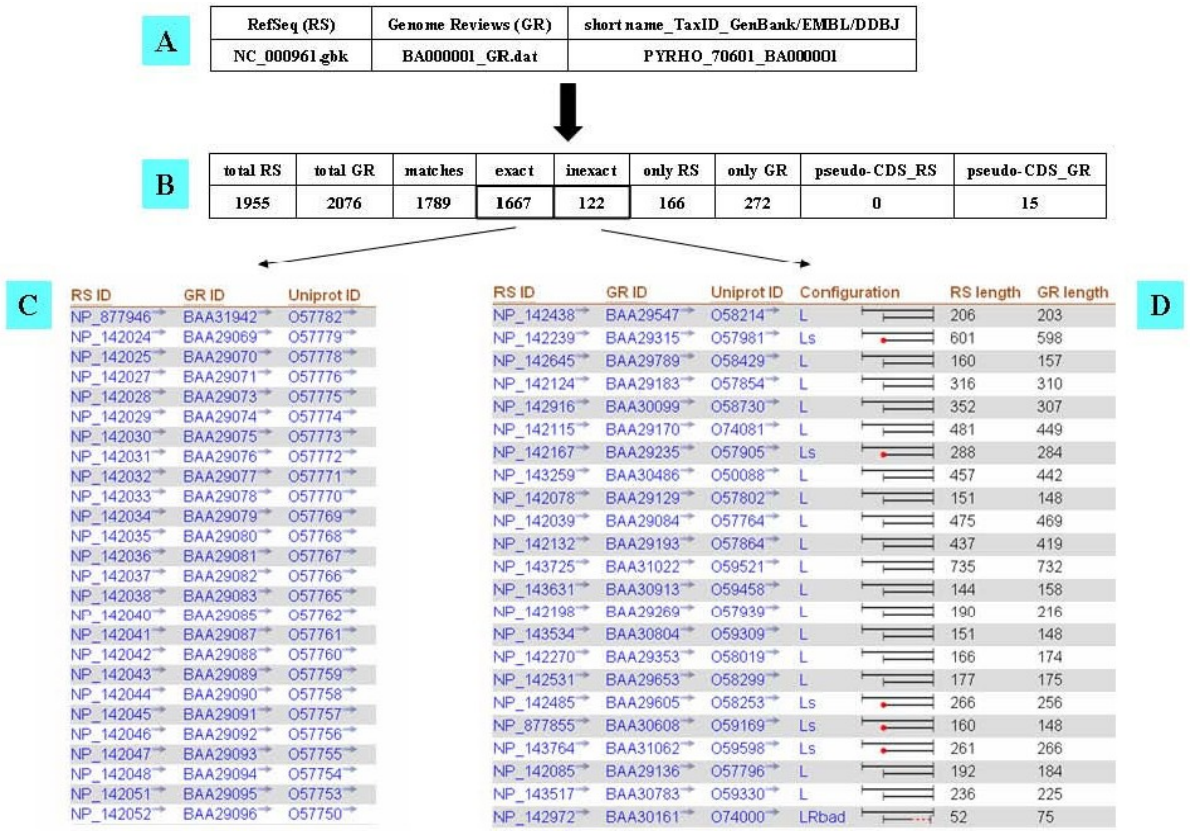


Figure 1
The different steps of the CorBank program. The main steps of the pipeline of Perl scripts are distinguished by different colors. The process of cross-referencing exact matching of the RefSeq and Genome Reviews versions of the same gene is indicated in yellow. The identification of inexact matches of genes that display a different structural annotation in both databases is made by the blue steps. Finally, disclosing the nature of the detected structural differences is made by the pink steps.

by a significant proportion (up to 12.5%, see below *Xanthomonas oryzae pv. oryzae KACC10331* in Table 3) of inexact matching of individual genes. The two versions of 321 species differ by their respective total numbers of genes but their corresponding CDS are matching exactly. For instance, *Bordetella petrii DSM 12804* has 5004 CDS that are matching exactly but RefSeq contains 23 CDS that are absent from Genome Reviews, whereas Genome Reviews display four additional CDS and 24 pseudo-CDS (amino acid sequence without a protein_id) that are not present in the RefSeq file. Finally, only 10 genomes have the same total numbers of genes but up to 7.4% of their corresponding genes display inexact matching. For instance, *Xanthomonas campestris pv. campestris str. 8004* displays 4273 CDS in both genomic databases but the respective amino acid sequence of the product of 310 of them differ between RefSeq and Genome Reviews. Complete data are available in Additional file 1 and on the CorBank site [14]).

Defining peculiarities of gene sequences that are partially identical between independently curated genome databases

We further studied these imperfectly matching sequences by measuring their similarity using an alignment-free approach (for a review and references inside, see [16]). Indeed, such an approach is fast and well-adapted to comparison of varying versions of the same sequence that share a significant common part. As detailed in Methods, we calculated the Euclidean distance that separates the distributions of words of length L (= 10) for each copy of the same gene in RefSeq and Genome Reviews, respectively (Fig. 1, blue part). This allows finding the cross-references between the respective imperfectly matching



A

RefSeq (RS)	Genome Reviews (GR)	short name_TaxID_GenBank/EMBL/DDBJ
NC_000961.gbk	BA000001_GR.dat	PYRHO_70601_BA000001

B

total RS	total GR	matches	exact	inexact	only RS	only GR	pseudo-CDS_RS	pseudo-CDS_GR
1955	2076	1789	1667	122	166	272	0	15

C

RS ID	GR ID	Uniprot ID
NP_877946	BAA31942	O57782
NP_142024	BAA29069	O57779
NP_142025	BAA29070	O57778
NP_142027	BAA29071	O57776
NP_142028	BAA29073	O57775
NP_142029	BAA29074	O57774
NP_142030	BAA29075	O57773
NP_142031	BAA29076	O57772
NP_142032	BAA29077	O57771
NP_142033	BAA29078	O57770
NP_142034	BAA29079	O57769
NP_142035	BAA29080	O57768
NP_142036	BAA29081	O57767
NP_142037	BAA29082	O57766
NP_142038	BAA29083	O57765
NP_142040	BAA29085	O57762
NP_142041	BAA29087	O57761
NP_142042	BAA29088	O57760
NP_142043	BAA29089	O57759
NP_142044	BAA29090	O57758
NP_142045	BAA29091	O57757
NP_142046	BAA29092	O57756
NP_142047	BAA29093	O57755
NP_142048	BAA29094	O57754
NP_142051	BAA29095	O57753
NP_142052	BAA29096	O57750

D

RS ID	GR ID	Uniprot ID	Configuration	RS length	GR length
NP_142438	BAA29547	O58214	L	206	203
NP_142239	BAA29315	O57981	Ls	601	598
NP_142645	BAA29789	O58429	L	160	157
NP_142124	BAA29183	O57854	L	316	310
NP_142916	BAA30099	O58730	L	352	307
NP_142115	BAA29170	O74081	L	481	449
NP_142167	BAA29235	O57905	Ls	288	284
NP_143259	BAA30486	O50088	L	457	442
NP_142078	BAA29129	O57802	L	151	148
NP_142039	BAA29084	O57764	L	475	469
NP_142132	BAA29193	O57864	L	437	419
NP_143725	BAA31022	O59521	L	735	732
NP_143631	BAA30913	O59458	L	144	158
NP_142198	BAA29269	O57939	L	190	216
NP_143534	BAA30804	O59309	L	151	148
NP_142270	BAA29353	O58019	L	166	174
NP_142531	BAA29653	O58299	L	177	175
NP_142485	BAA29605	O58253	Ls	266	256
NP_877855	BAA30608	O59169	Ls	160	148
NP_143764	BAA31062	O59598	Ls	261	266
NP_142085	BAA29136	O57796	L	192	184
NP_143517	BAA30783	O59330	L	236	225
NP_142972	BAA30161	O74000	LRbad	52	75

Figure 2
Differentiating exact and inexact matches. A partial view of the output of the CorBank program obtained when comparing the two versions of the genome of the archeon *Pyrococcus horikoshii* OT3 is detailed in several tables. Table A recapitulates the respective database information about this species and its computed label. Table B shows a summary of the data obtained using CorBank to find what is either common to both databases or specific of each one. Table C illustrates a few instances of exact matches. Table D exemplifies a few inexact matches with detailed configuration of the difference in the structural annotations of each copy of the same gene. The definitions of these inexact configurations are given in the Additional file 1.

copies of the same gene (see, for instance, Fig. 2 Table D). A large variety of differences explaining these imperfect matches have been found using the CorBank program (Fig. 1, pink part). All of these differences – including the very rare ones – have been categorized as summarized in the Additional file 1 and on the page http://www.corbank.u-psud.fr/help.html. CorBank is able to filter any differences in any sequence locations (see, for instance, Fig. 2 Table D).

We found that the differences between matching sequences that have unequal lengths were predominantly

Table 1: The reannotated copies of the same genome in independently curated databases[a] are predominantly divergent

copies of the same genome sequence in both curated databases[a] with		all CDS matching exactly	
		NO	*YES*
identical number of genes	*NO*	260 (40.5%)	321 (50%)
	YES	10 (1.5%)	50 (8%)

[a] RefSeq (Release 30) and Genome Reviews (Release 94.0) of July 2008

(98.7%) located at the N-terminal part. Indeed, it is often difficult to identify the start codon, especially when several methionines are found in this N-terminal region (see, for example, [17]).

Identifying the whole differences separating independently curated copies of a genome

Scanning paired versions of the same genome with CorBank allows computing the statistics of similarities and differences between genome databases. Figs. 2 and 3 detail the results obtained with the archaeon *Pyrococcus horikoshii*. The genomes of three *Pyrococcus* have been published ten years ago: *P. horikoshii* in 1998 [18], *P. abyssi* in 1999 [19] and *P. furiosus* in 2000 [20]. Since then, these genomes, sequenced and annotated by independent groups, have been curated several times. Fig. 2 shows that many differences have accumulated between the curated versions of the *P. horikoshii* genome in RefSeq and Genome Reviews (Fig. 2 Table B). First, the respective total numbers of genes are strikingly different. Among the 1955 sequences published in RefSeq and the 2076 ones listed in Genome Reviews, only 1789 are matching. Secondly, we have only 1667 of these matches that are exact (Fig. 2 Table C), while 122 display various differences. Fig. 2 (Table D) details a few instances of these differences in length and location of the start and end of each gene. Thirdly, Fig. 3 shows that there are a significant number of sequences putatively encoded by the *P. horikoshii* genome that are found in uniquely one genome database: 166 genes in RefSeq (Fig. 3 Table E) and 272 in Genome Reviews (Fig. 3 Table F), respectively. However, Genome Reviews classifies as pseudo-CDS a list of 15 amino acid

E

RS ID	Start	End	Locus	Product
NP_14202E	1,281	1,096	PH0002a	preprotein translocase subunit SecE
NP_87772E	3,258	2,965	PH0005.1n	hypothetical protein
NP_14204E	26,382	26,107	ΠH0026b	elongation factor 1-beta
NP_14205C	26,574	26,383	PH0026a	
NP_877727	27,507	27,884	PH0027.1n	
NP_14206C	37,724	37,437	PH0041a	DNA-directed RNA polymerase subunit L
NP_87772E	52,311	52,562	PH0059.1n	hypothetical protein
NP_87772E	55,985	56,197	PH0065.1n	hypothetical protein

F

GR ID	Uniprot ID	Start	End	Locus	Product
BAA29072	O57777	1,883	2,194	PH0004	Putative uncharacterized protein PH0004
BAA29086	O57763	15,743	16,264	PH0018	Putative uncharacterized protein PH0018
BAA29097	O57752	28,374	28,718	PH0029	Putative uncharacterized protein PH0029
BAA29098	O57751	29,049	29,387	PH0030	Putative uncharacterized protein PH0030
BAA29101	O57748	30618	30941	PH0033	Putative uncharacterized protein PH0033
BAA29106	O57743	35,561	36,028	PH0038	Putative uncharacterized protein PH0038
BAA29108	O74077	37,046	37,480	PH0040	Putative uncharacterized protein PH0040
BAA29109	O74078	37,177	37,599	PH0041	Putative uncharacterized protein PH0041

B

total RS	total GR	matches	exact	inexact	only RS	only GR	pseudo-CDS_RS	pseudo-CDS_GR
1955	2076	1789	1667	122	166	272	0	15

H

total RS	total GR	matches	exact	inexact	only RS	only GR
1955	2076	1806 (15)	1680 (13)	124 (2)	149	270

G

protein_id	uniprot	start	end	locus	product
none	P58199	1281	1096	PH0002.1	Preprotein translocase subunit secE
none	P58748	26382	26107	PH0026.1	Elongation factor 1-beta
none	P57671	37724	37437	PH0041.1	DNA-directed RNA polymerase subunit L
none	P84142	270758	271033	PH0305.1	Acylphosphatase
none	P58189	475976	475689	PH0529.1	50S ribosomal protein L31e
none	P59734	1091926	1091675	PH1212.1	UPF0248 protein PH1212.1
none	P59937	1138710	1138958	PH1266.1	50S ribosomal protein L14e
none	P58503	1185334	1185537	PH1316.1	30S ribosomal protein S17e
none	P61124	1334400	1334200	PH1495.1	50S ribosomal protein L24e
none	P62015	1446425	1446252	PH1631.1	DNA-directed RNA polymerase subunit K
none	P58193	1545510	1545211	PH1771.1	Protein translation factor SUI1 homolog
none	P60463	1647744	1647574	PH1895.1	Preprotein translocase subunit secG
none	P58746	1661808	1662107	PH1909.1	30S ribosomal protein S24e
none	P61239	1662109	1662261	PH1910.1	30S ribosomal protein S27ae
none	P58078	1682393	1682196	PH1939.1	30S ribosomal protein S27e

Figure 3
Differentiating exact and inexact matches, following. Table E illustrates a few instances of genes found uniquely in RefSeq. Table F exemplifies a few genes specific to Genome Reviews. Table G lists the pseudo-CDS specific to Genome Reviews. Table H re-evaluates the data presented in Table B after identifying by their positions the pseudogenes and pseudo-CDS specific to RefSeq and Genome Reviews, respectively and assessing their exactitude.

sequences which have no protein_id. Since these pseudo-CDS are found as standard coding sequences among the 166 sequences that are specific to RefSeq (Fig. 3 Table G), we ascertained this point. CorBank was further used to match these 15 pseudo-CDS using uniquely the position information that have been kept in both databases. As a result, Fig. 3 Table B was improved in Fig. 3 Table H after matching 13 of the 15 Genome Reviews pseudo-CDS as exact and two ones as inexact. Thus, it appears that RefSeq and Genome Reviews are producing increasingly divergent views of the same genome.

Table 2 shows the same trend for the two other *Pyrococcus* species, although the divergence is less marked. Such a discrepancy is strongly diminished when looking at the genome of the related *Thermococcus kodakarensis*, belonging to the same family (Thermococcaceae), which has been published more recently (in 2004 [21]). However, this example does not reflect a general (statistical) trend between the amount of divergences and time elapsed since the completion of sequence that would be true for all analyzed genomes (see below Tables 3 and 4 and accompanying text).

Discussion

CorBank is fulfilling two complementary goals: (1) to deliver immediate cross-references between each copy of each gene published in both RefSeq and Genome Reviews genome databases; (2) to identify any differences between both independently curated structural annotations. The first objective is achieved almost immediately: e.g. cross-referencing the two databases versions of a 3000 CDS genome is completed in less than 1 second on a basic home computer. Exhaustive comparison of the 641 prokaryotic species present in both databases at the end of July 2008 (Genome Reviews Release 94.0, 22nd July 2008 – RefSeq Release 30, July 11, 2008) has been completed in less than 60 min. Thus, the efficiency of CorBank is largely equivalent to that of the PICR tool that is described in a paper [22] that appeared as we were writing a first version of this manuscript. PICR, a web service allowing matching a large variety of protein sequence identifiers, is restricted to 100% identity matches and cannot discriminate the correct pair when recovering more than two identical sequences since it does not exploit information about genomic locations, contrarily to Corbank. Thus, this PICR tool and a previous one, MagicMatch [23], are not as efficient as CorBank to match exhaustively genome databases. This quality is especially true of our second goal that is achieved uniquely by CorBank. Its exhaustive comparison of the species currently present in both RefSeq and Genome Reviews shows dramatic differences in the structural annotations of a large portion of their copies of the same genomes (Tables 1 to 3, Figs. 2 and 3). Of the 641 compared genomes, 581 differ in their total numbers of CDS and 270 have from 1 to 781 coding sequences per genome that differ in length.

The large majority of the 50 perfectly matching genomes correspond to newly sequenced species where the manual curation has not been started. However, there is no direct correlation between the sequencing age and the level of divergence between the lastly curated versions of the same genome as shown on Tables 3 and 4 that list the top ten database-specific organisms in both RefSeq and Genome Reviews, respectively. Actually, a Spearman test failed to show any correlation of the different parameters computed by CorBank with the time elapsed since the completion of sequence (not shown).

Surprisingly, even the two versions of a model organism such as *Escherichia coli* K12 (substrain MG1655) that has been recently extensively reannotated in cooperative works [24,25] display significant differences. Of the 4295 gene-encoding proteins, only 4130 are matching (including 10 inexact matches), and both databases differ in their interpretation of some genes as being described as pseudo-CDS: 23 in RefSeq versus 24 in Genome Reviews. In fact, the structural identification of putative pseudogenes in *E. coli* K12 has been previously described (see [26] and references inside) but it is surprising that there is still disagreement even for these *E. coli* K12 pseudogenes.

As we were writing this paper, UniProtKB began to add/db_xref to RefSeq and Genome Reviews protein_id. However, we observed that rather often the same SwissProt file

Table 2: Complete distributions of the divergences of curated databases[a] in the case of closely related species

analyzed species	Comparing CDS in RefSeq Release 30 (RS) and Genome Reviews Release 94.0 (GR) databases[a]							
	total number		*matches*				*specific to*	
	RS	*GR*	*total*	*exact*	*inexact*	*by location*	*RS*	*GR*
P. horikoshii	1955	2076	1806	1680	124	0	149	270
P. furiosus	2125	2065	2065	1942	115	8	60	0
P. abyssi	1896	1786	1783	1715	68	0	113	3
T. kodakarensis	2306	2306	2306	2303	2	1	0	0

[a] versions of May 2008

Table 3: Top ten organisms having the highest number of CDS specific to RefSeq (RS) database

rank	organism	Total		matches			specific to	
		RS	*GR*	*total*	*inexact*	*by location*	*RS*	*GR*
RS1/GR7	Pyrococcus horikoshii OT3	1955	2076	1806	124	0	**149**	270
RS2	Neisseria meningitidis Z2491	2049	1991	1897	37	26	**120**	68
RS3/GR4	Xanthomonas oryzae pv. oryzae KACC10331	4144	4540	4030	497	2	**114**	510
RS4	Pyrococcus abyssi GE5	1896	1796	1783	68	0	**113**	3
RS5/GR6	Shewanella oneidensis MR-1	4467	4779	4364	34	1	**103**	415
RS6/GR8	Escherichia coli O157:H7 str. Sakai	5318	5461	5227	391	2	**87**	232
RS7	Deinococcus radiodurans R1	3181	1303	3099	91	1	**82**	4
RS8	Pyrococcus furiosus DSM 3638	2125	2065	2065	115	8	**60**	0
RS9	Lactococcus lactis subsp. lactis 111403	2321	2266	2263	68	0	**58**	3
RS10	Thermoplasma volcanium GSS1	1499	1526	1444	351	1	**55**	82

The organisms are sorted by their respective rank that is computed as the number of CDS that are found only in RefSeq database (Release 30). The organism names standing in the top ten list of both databases (Tables 3 and 4) are in bold.

has cross-references to multiple RefSeq and Genome Reviews protein_id. This is why we think that CorBank is – presently – the only software publishing unambiguous mapping of RefSeq, Genome Reviews, and UniProt identifiers of a protein.

Conclusion

Data dependencies inherent to the annotation process by homology make genome data predestined for propagated errors [1-4]. Thus, data cleansing is a necessity for genome data after the data is produced. However, such cleansing is uneasy since it is often impossible to find the correct solution right away. Instead, there often exists a set of alternative solutions. Accordingly, RefSeq and Genome Reviews appear to have diverged in looking for correct solutions when performing credibility checking on the INSD crude data. Credibility checking is a very important step for genome data production since the correctness of data is crucial before it is used within other processes such as annotation of newly sequenced genomes by homology to previously annotated genomes. However, such independent efforts made by both automatic and manual procedures [5,6] led to increasingly divergent reannotated data as shown in this work. Clearly, the time has come to enable curators of both genome databases to establish a minimum of dialogue. Whenever it would be technically possible, a useful compromise may be found where each database publishes the point of view of the other one. We acknowledge that such a harmonization effort looks rather complicated to be done. However, it would be very helpful for the whole community.

Table 4: Top ten organisms having the highest number of CDS specific to Genome Reviews (GR) database

rank	organism	Total		matches			without sequence or specific to	
		RS	*GR*	*total*	*inexact*	*by location*	*RS*	*GR*
GR1	Mycobacterium leprae TN	1605	2723	1605	77	1	0	1118
GR2	Orientia tsutsugamushi str. Boryong (Seoul National University)	1182	2143	1182	3	0	0	961
GR3	Orientia tsutsugamushi str. Boryong (Kitasato University)	1562	2085	1562	6	0	0	523
GR4/RS3	Xanthomonas oryzae pv. oryzae KACC10331	4144	4540	4030	497	2	114	510
GR5	Acinetobacter baumannii ATCC 17978	3368	3807	3368	77	0	0	439
GR6/RS5	Shewanella oneidensis MR-1	4467	4779	4364	34	1	103	415
GR7/RS1	Pyrococcus horikoshii OT3	1955	2076	1806	124	0	149	270
GR8/RS6	Escherichia coli O157:H7 str. Sakai	5318	5461	5227	391	2	87	232
GR9	Prochlorococcus marinus subsp. pastoris str. CCMP1986	1717	1935	1714	4	2	3	221
GR10	Prochlorococcus marinus str. MIT 9312	1810	1962	1810	10	0	0	152

The organisms are sorted by their respective rank that is computed as the number of CDS that are found only in Genome Reviews database (Release 94.0). The organism names standing in the top ten list of both databases (Tables 3 and 4) are in bold.

Methods

Comparing copies of the same genomes in curated databases

The whole genomic sequences present in RefSeq [5] and Genome Reviews [6] were downloaded at their respective FTP sites [27,28]. A first script allows matching respective downloaded files for the same genome. This script creates a mapping list between the replicons (chromosomes and plasmids) of the genome databases RefSeq and Genome Reviews. It links each respective genome identifier by using their common INSD identifier. A recognizable label, based on the association of its short name, NCBI tax_id and its INSD identifier, is associated to each matched replicon, e.g. PYRHO_70601_BA000001 for *Pyrococcus horikoshii* OT3 [17]. CorBank further compiles for each analyzed species the respective number of perfect and imperfect matches, and the sequences that are specific to a genome database as detailed below and in Fig. 1.

Detecting perfect matches between copies of the same gene in RefSeq and Genome Reviews

We built hash tables where each amino acid sequence is a key that indexes its encoding CDS (Fig. 1, yellow part). Each time the same key is found for the two versions of the same gene sequence made possible to cross-reference the respective protein identifiers in RefSeq [5], Genome Reviews [6], and UniProt [11] as shown on Fig. 2 (Table C).

Estimating similarity of partially identical sequences

In a second step (Fig. 1, blue part), CorBank is detecting all imperfect matches using an alignment-free comparison [for a review, see [16]]. We used a word approach as initially proposed by Blaisdell [29] and further documented by Zharkikh and Rzhetsky [30] to measure the similarity between sequences without any alignment. The distribution of the frequency of words of length L (L = 10 residues) in each amino acid sequence was computed for both copies of the same gene. These *L-uplets* are the respective signature of the sequence. The measure of the similarity between both copies of the same sequence is based on the Euclidean distance d^E that separates them:

$$d_L^E(X,Y) = \sum_{i=1}^{K} (c_{L,i}^X - c_{L,i}^Y)^2$$

The vectors c_L^X and c_L^Y represent word counts for the versions X and Y of the amino acid sequences encoded by the same gene in the respective RefSeq and Genome Reviews versions and K is the number of different *L- uplets* possible for the L-length. These X and Y copies are expected to share a largely common part but are of unequal sizes, one copy having an extension of variable size. To exclude any bias due to too large extensions, we stated that the maximum value of the distance d that separates two unequal copies of the same sequence could not be less than the difference between their respective numbers of amino acids.

In a third step (Fig. 1, pink part), CorBank is further analyzing all imperfect matches to define the location of the difference between both paired copies of the same gene. CorBank is first searching if the difference takes place on either the N-terminal side or the C- terminal one. In rare cases, the difference is located elsewhere, including in the common segment of both copies that could differ for only one residue. The Additional file 1 details all encountered cases, including the very rare ones.

Authors' contributions

SDD inspired using a word approach. SDD and MB developed together the CorBank program. MB set up the present CorBank website and also made in-depth analysis of the data obtained with CorBank. BL initiated the work, participated in the data analysis and wrote the draft manuscript. All authors read and finalized the whole version of the manuscript.

Acknowledgements

We thank Olivier Lespinet and Frédéric Lemoine for helpful discussions and critical reading of the manuscript and the three Reviewers for their constructive and useful comments. This work was funded by the CNRS (UMR 8621) and the Agence Nationale de la Recherche (ANR-05-MMSA-0009 MDMS_NV_10).

References

1. Bork P, Bairoch A: **Go hunting in sequence databases but watch out for the traps.** *Trends in Genetics* 1996, **12**:425-427.
2. Brenner SE: **Errors in genome annotation.** *Trends Genet* 1999, **15**:132-133.
3. Janssen P, Goldovsky L, Kunin V, Darzentas N, Ouzounis CA: **Genome coverage, literally speaking. The challenge of annotating 200 genomes with 4 million publications.** *EMBO Rep* 2005, **6**:397-399.
4. Ouzounis CA, Karp PD: **The past, present and future of genome-wide re-annotation.** *Genome Biology* 2002, **3**:. comment2001.1-2001.6
5. Pruitt KD, Tatusova T, Maglott DR: **NCBI reference sequences (RefSeq): a curated non-redundant sequence database of genomes, transcripts and proteins.** *Nucleic Acids Res* 2007, **35**:61-65 [http://www.ncbi.nlm.nih.gov/RefSeq/].
6. Sterk P, Kersey PJ, Apweiler R: **Genome Reviews: Standardizing Content and Representation of Information about Complete Genomes.** *OMICS* 2006, **10**:114-118 [http://www.ebi.ac.uk/GenomeReviews/].
7. Maglott D, Ostell J, Pruitt KD, Tatusova T: **Entrez Gene: gene-centered information at NCBI.** *Nucleic Acids Res* 2007, **35**:D26-31 [http://www.ncbi.nlm.nih.gov/sites/gquery].
8. Marchler-Bauer A, Anderson JB, Derbyshire MK, DeWeese-Scott C, Gonzales NR, Gwadz M, Hao L, He S, Hurwitz DI, Jackson JD, Ke Z, Krylov D, Lanczycki C, Liebert CA, Liu C, Lu F, Marchler GH, Mullokandov M, Song JS, Thanki N, Yamashita RA, Yin JJ, Zhang D, Bryant SH: **CDD: a conserved domain database for interactive domain family analysis.** *Nucleic Acids Res* 2007, **35**:D237-40 [http://www.ncbi.nlm.nih.gov/Structure/cdd/cdd.shtml].
9. The Gene Ontology Consortium: **Gene Ontology: tool for the unification of biology.** *Nature Genet* 2000, **25**:25-29 [http://www.geneontology.org/index.shtml].

10. Mulder NJ, Apweiler R: **The InterPro database and tools for protein domain analysis.** *Curr Protoc Bioinformatics* 2008, **Chapter 2:** [http://www.ebi.ac.uk/interpro/]. Unit 27
11. The UniProt Consortium: **The Universal Protein Resource (UniProt).** *Nucleic Acids Res* 2007, **35**:D193-197 [http://www.expasy.org/sprot/].
12. **HOGENOM** [http://pbil.univ-lyon1.fr/databases/hogenom.php]
13. Berman HM, Henrick K, Nakamura H: **Announcing the worldwide Protein Data Bank.** *Nature Structural Biology* 2003, **10**:980 [http://www.wwpdb.org/].
14. **CorBank** [http://www.corbank.u-psud.fr/]
15. **BLASTClust** [http://www.ncbi.nlm.nih.gov/blast/docs/blastclust.html]
16. Vinga S, Almeida J: **Alignment-free sequence comparison-a review.** *Bioinformatics* 2003, **19**:513-523.
17. Frishman D, Mironov A, Mewes H-W, Gelfand M: **Combining diverse evidence for gene recognition in completely sequenced bacterial genomes.** *Nucleic Acids Research* 1998, **26**:2941-2947.
18. Kawarabayasi Y, Sawada M, Horikawa H, Haikawa Y, Hino Y, Yamamoto S, Sekine M, Baba S, Kosugi H, Hosoyama A, *et al.*: **Complete sequence and gene organization of the genome of a hyper-thermophilic archaebacterium, *Pyrococcus horikoshii* OT3.** *DNA Res* 1998, **5**:55-76.
19. Cohen GN, Barbe V, Flament D, Galperin M, Heilig R, Lecompte O, Poch O, Prieur D, Querellou J, Ripp R, *et al.*: **An integrated analysis of the genome of the hyperthermophilic archaeon *Pyrococcus abyssi*.** *Mol Microbiol* 2003, **47**:1495-1512.
20. Robb FT, Maeder DL, Brown JR, DiRuggiero J, Stump MD, Yeh RK, Weiss RB, Dunn DM: **Genomic sequence of hyperthermophile, Pyrococcus furiosus: implications for physiology and enzymology.** *Meth Enzymol* 2001, **330**:134-157.
21. Fukui T, Atomi H, Kanai T, Matsumi R, Fujiwara S, Imanaka T: **Complete genome sequence of the hyperthermophilic archaeon *Thermococcus kodakaraensis* KOD1 and comparison with Pyrococcus genomes.** *Genome Res* 2005, **15**:352-363.
22. Côté RG, Jones P, Martens L, Kerrien S, Reisinger F, Lin Q, Leinonen R, Apweiler R, Hermjakob H: **The Protein Identifier Cross-Referencing (PICR) service: reconciling protein identifiers across multiple source databases.** *BMC Bioinformatics* 2007, **8**:401 [http://www.ebi.ac.uk/Tools/picr/].
23. Smith M, Kunin V, Goldovsky L, Enright AJ, Ouzounis CA: **MagicMatch – crossreferencing sequence identifiers across databases.** *Bioinformatics* 2005, **21**:3429-3430.
24. Riley M, Abe T, Arnaud MB, Berlyn MK, Blattner FR, Chaudhuri RR, Glasner JD, Horiuchi T, Keseler IM, *et al.*: ***Escherichia coli* K-12: a cooperatively developed annotation snapshot – 2005.** *Nucleic Acids Res* 2006, **34**:1-9.
25. Karp PD, Keseler IM, Shearer A, Latendresse M, Krummenacker M, Paley SM, Paulsen I, Collado-Vides J, Gama-Castro S, *et al.*: **Multidimensional annotation of the *Escherichia coli* K-12 genome.** *Nucleic Acids Res* 2007. doi:10.1093/nar/gkm740
26. Ochman H, Davalos LM: **The nature and dynamics of bacterial genomes.** *Science* 2006, **311**:1730-1733.
27. **FTP NCBI** [ftp://ftp.ncbi.nih.gov/refseq/]
28. **FTP EBI** [ftp://ftp.ebi.ac.uk/pub/databases/genome_reviews]
29. Blaisdell BE: **A measure of the similarity of sets of sequences not requiring sequence alignment.** *Proc Natl Acad Sci USA* 1986, **83**:5155-5159.
30. Zharkikh AA, Rzhetsky A: **Quick assessment of similarity of two sequences by comparison of their L-tuple frequencies.** *Biosystems* 1993, **30**:93-111.

2 Abréviations

- abTCase : carbamoyltransférase d'antibiotiques
- AOTCase : acétylornithine carbamoyltransférase
- ATCase : aspartate carbamoyltransférase
- ATP : adénosine triphosphate
- CP : carbamoyl-phosphate
- CPSase : Carbamoyl-phosphate synthétase
- TCase : Carbamoyltransférase or transcarbamoylase
- DHOase : Dihydroorotase
- DHODase : Dihydroorotate déshydrogénase
- DHPase : Dihydropyrimidinase
- DHPDase : Dihydropyrimidine déshydrogénase
- OMP : orotidine monophosphate
- OTCase : ornithine carbamoyltransférase
- PTCase : putrescine carbamoyltransférase
- Pi : Phosphate inorganique
- PRPP : 5-phosphoribosyl diphosphate
- SOTCase : succinylornithine carbamoyltransférase
- UTC(ase) : carbamoyltransférase non identifiée
- YTC(ase) : carbamoyltransférase non identifiée contenant les YgeW

2.1 Bases azotées, nucléosides et nucléotides

Base	Base azotée	Nucléoside	Abréviation
Purine	Adénine	Adénosine	A
Pyrimidine	Cytosine	Cytidine	C
Purine	Guanine	Guanosine	G
Purine	Hypoxanthine	Inosine	I
Pyrimidine	Orotate	Orotidine	O
Pyrimidine	Thymine	Thymidine	T
Pyrimidine	Uracile	Uridine	U
Purine	Xanthine	Xanthosine	X

Figure 26: Liste des bases azotées et leurs dérivés

Les ribonucléotides sont abrégés, par exemple avec l'adénosine :

- A**MP** (adénosine monophosphate)
- A**DP** (adénosine diphosphate)
- A**TP** (adénosine triphosphate)

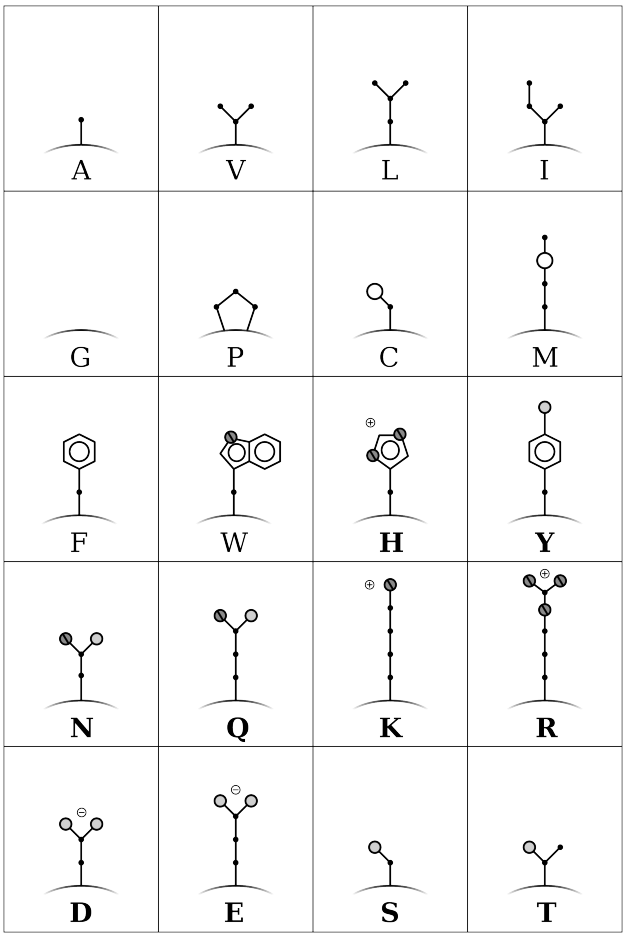
Les désoxyribo-nucléotides sont notés de la même manière en préfixant avec d :

- **d**AMP (désoxyadénosine monophosphate)

2.2 Acides aminés

Nom complet	3 lettres	1 lettre
alanine	Ala	A
arginine	Arg	R
asparagine	Asn	N
aspartate	Asp	D
cystéine	Cys	C
glutamate	Glu	E
glutamine	Gln	Q
glycine	Gly	G
histidine	His	H
isoleucine	Ile	I
leucine	Leu	L
lysine	Lys	K
méthionine	Met	M
phénylalanine	Phe	F
proline	Pro	P
sérine	Ser	S
thréonine	Thr	T
tryptophane	Trp	W
tyrosine	Tyr	Y
valine	Val	V

A



B

Figure 27: Liste des acides aminés, abréviations et représentation simplifiée.
A : table des acides aminés protéiques et leurs abréviations à trois et une lettre.
B : représentation simplifiée des vingt acides aminés protéiques. Au dessus de chaque lettre, l'arc correspond à la partie commune des acides aminés qui constitue le squelette des protéines. Pour chacun la chaîne latérale est représentée avec un carbone (petit point noir), un azote (cercle gris barré), un oxygène (cercle gris pâle) ou un soufre (cercle blanc), sauf la glycine qui n'a pas de chaîne latérale.

- La chaîne latérale des acides aminés (Figure 27B) leur donne des propriétés chimiques différentes. On distingue plusieurs classes :

- apolaires : AVLI GPCM FW
- polaires : Y NQ ST
- aliphatiques : AVLI GP M
- aromatiques : FWHY
- chargés basiques : H KR
- chargés acides : DE

- Les acides aminés naturels des protéines sont des acides α-aminés, c'est-à-dire que leur formule est de type : $^{+}H_3N$-CHR-COO^{-} où R est le groupe latéral (Figure 27B). Le groupe amine ($-NH_3^{+}$) est donc lié au premier carbone après le groupe carboxylique ($-COO^{-}$), noté α. Dans les acides β-aminés, le groupe amine est lié au second carbone sur la chaîne latérale. Lorsqu'un autre groupe amine est présent dans un acide aminé, sa place est donnée par une lettre grecque : équivalente. Par exemple, le groupe amine de la lysine (K) est en ε (cinquième carbone).

3 Index

4 Divers

- Les figures d'arbres ont été créés à l'aide de l'outils de visualisation d'arbres phylogénétiques Dendroscope version 2 [Huson et al. 2007]. Site Internet : http://ab.inf.uni-tuebingen.de/software/dendroscope/

- La plupart des figures ont été créées avec le logiciel de dessin vectoriel libre Inkscape version 0.47. Site Internet : http://inkscape.org