L'évolution modulaire des protéines : un point de vue phylogénétique

Anne-Sophie Sertier

N° 153-2011 Année 2011

THÈSE DE L'UNIVERSITÉ DE LYON

Présentée

devant L'UNIVERSITÉ CLAUDE BERNARD LYON1

pour l'obtention

du DIPLÔME DE DOCTORAT

(arrêté du 7 août 2006)

soutenue publiquement le

12 septembre 2011

par

Anne-Sophie SERTIER

L'évolution modulaire des protéines : un point de vue phylogénétique

---

Directeurs de thèse : Daniel KAHN
Vincent DAUBIN

Jury :	Pierre BRÉZELLEC	Rapporteur
	Vincent DAUBIN	Co-directeur de thèse
	Daniel KAHN	Directeur de thèse
	Dominique MOUCHIROUD	Présidente du jury
	Pierre Antoine PONTAROTTI	Rapporteur
	Alain VIARI	Examinateur

UNIVERSITE CLAUDE BERNARD - LYON 1

Président de l'Université	**M. A. Bonmartin**
Vice-président du Conseil d'Administration	M. le Professeur G. Annat
Vice-président du Conseil des Etudes et de la Vie Universitaire	M. le Professeur D. Simon
Vice-président du Conseil Scientifique	M. le Professeur J-F. Mornex
Secrétaire Général	M. G. Gay

COMPOSANTES SANTE

Faculté de Médecine Lyon Est - Claude Bernard	Directeur : M. le Professeur J. Etienne
Faculté de Médecine et de Maïeutique Lyon Sud - Charles Mérieux	Directeur : M. le Professeur F-N. Gilly
UFR d'Odontologie	Directeur : M. le Professeur D. Bourgeois
Institut des Sciences Pharmaceutiques et Biologiques	Directeur : M. le Professeur F. Locher
Institut des Sciences et Techniques de la Réadaptation	Directeur : M. le Professeur Y. Matillon
Département de formation et Centre de Recherche en Biologie Humaine	Directeur : M. le Professeur P. Farge

COMPOSANTES ET DEPARTEMENTS DE SCIENCES ET TECHNOLOGIE

Faculté des Sciences et Technologies	Directeur : M. le Professeur F. Gieres
Département Biologie	Directeur : M. le Professeur F. Fleury
Département Chimie Biochimie	Directeur : Mme le Professeur H. Parrot
Département GEP	Directeur : M. N. Siauve
Département Informatique	Directeur : M. le Professeur S. Akkouche
Département Mathématiques	Directeur : M. le Professeur A. Goldman
Département Mécanique	Directeur : M. le Professeur H. Ben Hadid
Département Physique	Directeur : Mme S. Fleck
Département Sciences de la Terre	Directeur : Mme le Professeur I. Daniel
UFR Sciences et Techniques des Activités Physiques et Sportives	Directeur : M. C. Collignon
Observatoire de Lyon	Directeur : M. B. Guiderdoni
Ecole Polytechnique Universitaire de Lyon 1	Directeur : M. P. Fournier
Ecole Supérieure de Chimie Physique Electronique	Directeur : M. G. Pignault
Institut Universitaire de Technologie de Lyon 1	Directeur : M. le Professeur C. Coulet
Institut de Science Financière et d'Assurances	Directeur : M. le Professeur J-C. Augros
Institut Universitaire de Formation des Maîtres	Directeur : M. R. Bernard

Remerciements

C'est en écrivant les remerciements que l'on se rend compte du nombre impressionnant de personnes qui à un moment donné ont contribué d'une manière ou d'une autre à la réalisation de cette thèse. Il me faudrait plusieurs pages pour être sûre de n'oublier personne et surtout pour remercier chacun à sa juste valeur. Cependant, tous ces gens savent à quel point écrire n'est pas ce que je fais de mieux. Mais comme dirait Janice : "J'ai d'autres qualités". Alors je serais assez brève et si quelqu'un se sent un peu lésé, il lui suffit de me le dire et je me ferai un plaisir de le remercier comme il se doit autour d'une bonne table ou d'une bonne bière !

Je tiens tout d'abord à remercier Daniel et Vincent pour tout ce qu'ils m'ont apporté tant sur le plan scientifique que personnel, pendant ces presque 5 années (et même 6 et demi pour Vincent !). Je tiens particulièrement à les remercier d'avoir cru en moi jusqu'au bout.

J'aimerais remercier toutes les personnes qui ont évaluer et fait avancer mon travail de thèse notamment lors du comité de pilotage : Éric Rivals, Alain Viari, Pierre Brézellec, Pierre Antoine Pontarotti et Dominique Mouchiroud.

Je tiens également à remercier Dominique Mouchiroud et Manolo Gouy qui m'ont chaleureusement accueillie toutes ces années au sein du laboratoire de Biométrie et Biologie Évolutive dans l'équipe de Bioinformatique et Genomique Évolutive. Cela a été un vrai plaisir de travailler dans ce laboratoire où de nombreuses personnes m'ont été d'une grande aide notamment au niveau administratif et informatique. Je tiens également à remercier tous les stagiaires, doctorants, post-doctorants, chercheurs et enseignants-chercheurs de l'équipe pour la qualité de l'environnement scientifique (discussions, collaborations, journal club et autres réunions ad-hoc) et pour mon insertion dans l'équipe pédagogique, mais avouons-le aussi, pour les pauses café où la gastronomie et l'humour (attention, il faut être connaisseur et surtout pas susceptible) ont toujours eu leur place. J'aimerais faire un clin d'œil spécial à Guy, un amateur de bière et un photographe grâce à qui les nuits du badminton (où l'équipe du BBE n'a malheureusement jamais gagné ...) resteront inoubliables !

Je ne sais comment remercier Alexandra qui a été d'une aide précieuse lors de la rédaction de ce manuscrit. Nos séances d'écriture, nos sushis partis, nos balades aux parcs avec Vlad, nos remontages de morales mutuels, et tant d'autres choses ... (parce qu'il y a eu une vie avant ce manuscrit et qu'il y en a une après). Je me dois d'avouer ici que j'ai perdu le pari engager avec Vlad au début du printemps. Il a fait ses premiers pas avant que je termine ma thèse ... Il ira loin ce petit !

Enfin, j'aimerais remercier tous ceux qui ont enrichi ma vie personnelle, contribuant indirectement, mais fondamentalement à l'aboutissement de ce travail. Il y a d'abord tous mes adversaires et coéquipiers du badminton. Un merci spécial à Thierry de m'avoir acceptée jusqu'à la fin de ma thèse dans les cours de bad de la fac (j'ai d'ailleurs pu rencontrer certains de mes étudiants sur le terrain ...). Il y a aussi Lauranne qui m'a recrutée pour la chorale de la Doua (un immense merci, cela a été un pur bonheur), et m'a fait rencontré François, un chef de chœur hors du commun ! Et puis, il y a le Big Ben dit Ben Jefferson ... ça c'est du pote. Je ne sais vraiment pas comment je m'en serais sortie sans lui. Je pourrais citer nos virées au parc (en roller ou en courant), nos apéros spontanés ou Vincent mais ce serait un piètre résumé de tout ce que tu m'as apporté et apportes encore. Alors tout simplement, merci d'être mon ami. Je terminerai en remerciant mes parents, mes frères et sœurs et Ilan mon p'ti neveu. Je vous dois beaucoup et j'aimerais juste vous dire que je suis fière de vous avoir pour famille.

Table des matières

Liste des figures

Annexes

Liste des tables

Chapitre 1

Introduction – Evolution des protéines et architectures en domaines

Les protéines sont les éléments fonctionnels fondamentaux des organismes vivants. Elles sont impliquées dans l'ensemble des processus biologiques des cellules que ce soit dans leur architecture ou leur fonctionnement. Elles peuvent par exemple, avoir une fonction régulatrice, catalytique, motrice ou structurale. L'ensemble des protéines exprimées dans un organisme a un effet sur son phénotype et son adaptation à son environnement. L'évolution des protéines et notamment l'apparition de nouvelles protéines et donc de nouvelles fonctions est au cœur de l'évolution des organismes.

Ce chapitre propose une introduction sur l'évolution des protéines modulaires. Dans un premier temps, nous reviendrons sur la définition des domaines et modules protéiques et sur leur détermination. Ensuite, nous décrirons les différents processus à l'origine de la diversification des protéines. Puis nous présenterons quelques aspects méthodologiques liés à la modularité des protéines et à l'inférence de répertoires ancestraux de protéines. Enfin, nous présenterons le point de vue développé dans cette thèse sur l'évolution des protéines modulaires.

1.1 La protéine : un assemblage d'unités structurales ou évolutives ?

1.1.1 Quelques structures dans un océan de séquences

Les protéines peuvent être définies sur 4 niveaux d'organisation structurale (figure 1.1).

(i) La *structure primaire* correspond à l'enchaînement linéaire des acides aminés reliés par des liaisons peptidiques. On distingue 20 acides aminés universellement distribués.

(ii) La *structure secondaire* correspond à un repliement local de la chaîne d'acides aminés stabilisé par des liaisons hydrogènes. On distingue plusieurs types de structures secondaires : les *hélices* α, les *feuillets* β qui sont généralement connectés par des *coudes* (*turn*) ou des *boucles* (*loop*). Les hélices α et les feuillets β sont des structures périodiques de la chaîne polypeptidique.

(iii) La *structure tertiaire* représente la position dans l'espace 3D de la chaîne polypeptidique et des chaînes latérales.

(iv) La *structure quaternaire* résulte de l'assemblage d'au moins deux chaînes polypeptidiques (identiques ou différentes) par des liaisons non covalentes.

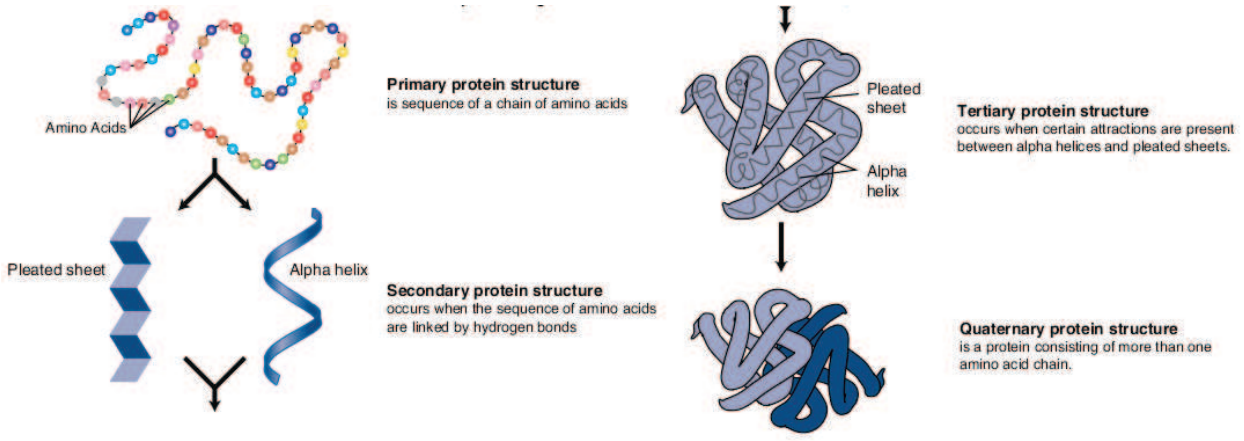


FIGURE 1.1 – **Les quatre niveaux d'organisation structurale des protéines.** On distingue quatre niveaux : primaire, secondaire, tertiaire ou tridimensionnel et quaternaire. [2]

La première séquence peptidique, l'insuline du porc, a été déterminée par Frederick Sanger en 1953 à l'aide d'une technique basée sur la chromatographie. Cette méthode a été largement supplantée par les méthodes de séquençage de l'ADN qui permettent d'obtenir plus rapidement

2. Figure créée par Darryl Leja de l'Institut National de Recherche sur le Génome Humain (NHGRI), accessible à l'adresse `http://www.genome.gov/Glossary/resources/protein.pdf` (légèrement modifiée).

les séquences primaires de protéines à partir de la traduction des séquences nucléiques [3]. C'est de loin le niveau de structure des protéines sur lequel on a le plus d'information, avec des millions de séquences protéiques résolues chez des milliers d'organismes [4] et une croissance exponentielle des bases de données. De ces séquences peuvent être prédites les structures secondaires. Les modèles de prédiction sont basés sur les propriétés biochimiques gouvernant la formation de structures secondaires et sur l'information issue des alignements multiples de séquences homologues. Les méthodes actuelles combinent ces modèles à des algorithmes optimisés d'apprentissage automatique (comme les réseaux de neurones). Certaines méthodes comme celle implémentée dans le programme Porter [Pollastri et McLysaght, 2005] atteignent une précision de 80% dans leurs prédictions de structures secondaires en utilisant les protéines de la PDB (Protein Data Bank) comme jeu de données test (pour une revue complète, voir Pirovano et Heringa [2010]). La détermination des niveaux d'organisation supérieurs des protéines repose sur des techniques plus lourdes à mettre en œuvre. Les deux principales méthodes de résolution des structures 3D sont la cristallographie par rayon X qui a permis de déterminer la première structure 3D, celle de l'hémoglobine de cheval, par Max Perutz et John Kendrew en 1968, et la spectroscopie par résonance magnétique nucléaire (RMN). La croissance des bases de données correspondantes est moins importante malgré les efforts considérables fournis ces dernières années [Grabowski *et al.*, 2007]. On est ainsi passé de 35 000 à 70 000 structures dans la PDB [5] ces 5 dernières années. Ce chiffre est à comparer au 15 millions de séquences disponibles dans UniProtKB fin mai 2011. Même en prenant en compte la redondance au niveau des séquences, le nombre de séquences associées à une structure résolue ne dépasserait pas les quelques pour cent.

1.1.2 Les domaines

Les biochimistes découpent classiquement les protéines en domaines, qui correspondent à des unités structurales capables de se replier indépendamment du reste de la protéine [Wetlaufer, 1973]. Ils ont généralement une forme globulaire dont le cœur est hydrophobe et la surface hydrophile, et ils ne peuvent pas être subdivisés sans les dénaturer [Marsden et Orengo, 2008]. La plupart des protéines caractérisées biochimiquement sont organisées comme des enchaînements de domaines présentant une architecture caractéristique (figure 1.2) : on trouve environ 66% de protéines multidomaines chez les procaryotes et 80% chez les eucaryotes [Chothia *et al.*, 2003; Liu et Rost, 2004]. Certaines peuvent contenir plus d'une centaine de domaines comme par exemple une pro-

3. On considère ici que ces séquences primaires reflètent la séquence en acides aminés de la protéine mature : les modifications post-transcriptionnelles (édition de l'ARNm [Cattaneo, 1990] par exemple) ou post-traductionnelles (permutation circulaire [Grishin, 2001] par exemple) sont donc négligées.

4. La banque de séquences protéiques UniProtKB/TrEMBL [Consortium, 2009] contenait 15 400 876 entrées au 31 mai 2011 `http://expasy.org/sprot/`

5. La Protein Data Bank (PDB) [Kirchmair *et al.*, 2008] contenait 73 656 structures tridimensionnelles le 31 mai 2011 `http://www.pdb.org`

téine structurale du muscle, la titine [Maruyama, 1994]. On associe généralement une propriété à chaque domaine, ainsi une protéine multidomaine possède une combinaison unique de propriétés qui lui confèrent sa fonction. Par exemple, les facteurs de transcription sont des protéines multidomaines dans lesquelles on trouve en général au moins un domaine de fixation à l'ADN et un domaine activateur ou inhibiteur agissant sur le complexe de transcription.

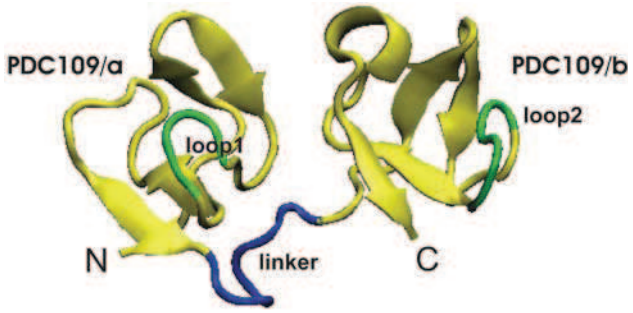


FIGURE 1.2 – **Structure 3D de la protéine multidomaine PDC109.** Cette protéine est constituée de deux domaines Fibronectine de type II (notés PDC109/a etPDC109/b) et d'une région linker indiquée en bleu. La structure a été déterminée par cristallographie par rayon X. Extrait de Kim *et al.* [2010]

1.1.3 Les modules

Pour les évolutionnistes, les protéines sont des combinaisons de modules élémentaires qui représentent des segments de gènes réutilisés plusieurs fois au cours de l'évolution dans des contextes différents. La comparaison systématique des séquences primaires des protéines en utilisant des outils tels que BLAST (Basic Local Alignment Search Tool) [Altschul *et al.*, 1997] permet de mettre en évidence ces fragments de séquences homologues. Cette méthode effectue une recherche de similarité locale d'une séquence requête dans une base de données. BLAST définit une E-value qui correspond au nombre d'alignements ayant un score au moins aussi grand attendu par hasard (on pourrait dire sans ancêtre commun) dans la banque utilisée. Les fragments de séquences sont considérés comme homologues si la E-value est suffisamment faible. Par exemple, dans la figure 1.3, on a détecté dans les protéines FixJ et OmpR des modules homologues en position N-terminale et dans les protéines FixJ et LuxR des modules homologues en position C-terminale. Pourtant ces trois protéines possèdent des architectures de modules différentes et ne peuvent donc pas être considérées comme intégralement homologues. L'homologie entre deux modules n'implique pas automatiquement l'homologie entre les protéines, définie comme un arrangement particulier de modules. En 1994, Hillis a introduit le terme d'homologie partielle entre protéines [Hillis, 1994] pour expliquer ces similarités locales significatives entre protéines non apparentées.

Lorsqu'on se pose la question de l'homologie entre FixJ et OmpR, la réponse est positive localement (les protéines possèdent des fragments homologues) mais négative globalement (leurs architectures sont hétérologues). Cette question des niveaux d'homologie peut être rapprochée

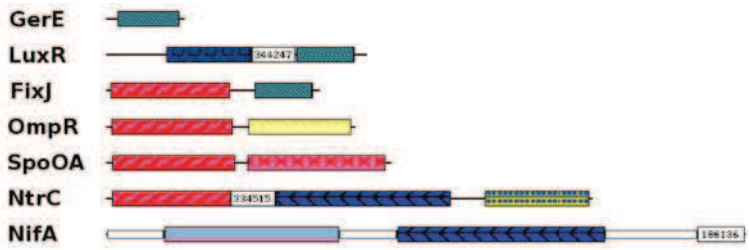


FIGURE 1.3 – **Arrangement en modules de protéines extraites de ProDom.** Ces protéines sont toutes impliquées dans l'activation de la transcription, mais dans différentes voies métaboliques et dans différents organismes. Elles partagent des modules dont les différentes combinaisons sont à l'origine d'une variété de réponses à l'environnement.

de concepts classiques au niveau morphologique (figure 1.4). Le groupe des oiseaux et celui des chauves-souris ont des ailes qui ont des structures différentes et qui sont apparues de manière indépendante au cours de l'évolution. Cependant, toutes deux dérivent indiscutablement du membre antérieur des tétrapodes. Ainsi, le membre antérieur de ces animaux est hétérologue en tant qu'aile, mais homologue en tant que membre. L'intégration de ces différents niveaux d'homologie est essentielle à la compréhension des mécanismes de l'évolution de ces structures.

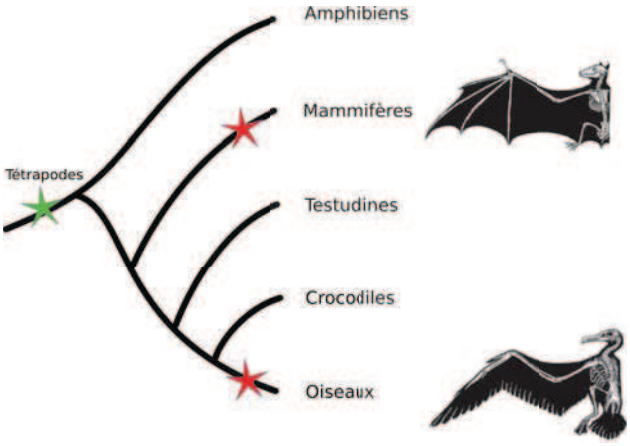


FIGURE 1.4 – **Évolution des ailes chez les tétrapodes.** Les étoiles rouges représentent les apparitions indépendantes des ailes et l'étoile verte correspond à l'ancêtre commun possédant deux paires de membres dont les membres antérieurs transformés en ailes chez les oiseaux et les chiroptères (mammifères volants).

1.1.4 Le domaine est-il un module comme les autres ?

Les protéines peuvent donc être découpées de deux façons différentes : soit en se basant sur leur structure 3D, soit en se basant sur l'analyse comparative de leur séquence primaire. On peut naturellement se demander quelle est la nature du lien entre les modules et les domaines issus de ces découpages.

L'analyse de la structure des gènes dans de nombreux organismes eucaryotes a montré une structure morcelée où les exons, les régions codantes, sont séparées par les introns, des régions non codantes intragéniques. En 1978, Walter Gilbert émet l'hypothèse que les gènes eucaryotes sont le résultat de la recombinaison des exons et que les introns sont les restes de ce processus qui permet le réarrangement des exons pour former de nouvelles protéines [Gilbert, 1978]. Autrement dit, les exons sont des modules qui codent des domaines. Cependant, bien que le réarrangement d'exons ait été invoqué pour expliquer l'apparition de quelques nouvelles protéines dans l'évolution récente des eucaryotes [Liu et Grigoriev, 2004; Orgel et Crick, 1980; Patthy, 1999], et malgré leur rôle dans l'épissage alternatif, la correspondance entre exon et domaine représenterait plutôt des cas particuliers [Kaessmann *et al.*, 2002]. Si l'évolution réarrange indiscutablement d'anciens gènes pour en faire de nouveaux, elle ne semble pas utiliser particulièrement les introns pour ce faire. Au contraire, elle réutilise des fragments de séquences de façon aléatoire. On comprend aisément que tout réarrangement qui se ferait au milieu d'un domaine ne permettrait pas de préserver sa fonction et serait éliminé par la sélection. Les domaines seraient donc généralement des modules [Patthy, 2001].

Cependant, la réciproque n'est pas nécessairement vraie : un module ne correspond pas forcément à un domaine. On peut trouver de nombreuses exceptions. Par exemple, les protéines fibreuses qui contiennent des séquences répétées ou les protéines dont la structure est désordonnée ne présentent pas une structure globulaire : on ne peut donc pas découper ces protéines en domaines. Mais il est tout à fait possible de les découper en modules : certaines séquences répétées sont retrouvées dans différents contextes protéiques attestant de la conservation et de la réutilisation de ces séquences. D'autre part, un module peut correspondre à plusieurs domaines qui resteraient associés pour des contraintes fonctionnelles, comme dans le cas des supradomaines [Vogel *et al.*, 2004b] (discuté dans la section suivante).

La correspondance entre les modules et les domaines a été déterminée sur très peu d'exemples, car peu de structures 3D sont actuellement disponibles comparé à l'immensité de l'univers des séquences disponible. De manière générale, l'univers des domaines est considéré comme assez restreint. Les domaines structuraux sont classiquement regroupés selon une structure hiérarchique définissant plusieurs niveaux d'homologie comme la famille et la superfamille ainsi que des niveaux plus généraux comme les repliements ou la topologie tridimensionnelle pour lesquels l'hypothèse d'homologie est généralement non prouvée. Chothia [1992] estimait à un millier le nombre de domaines, en terme de repliements structuraux. Les différentes estimations réalisées par la suite à partir des données structurales disponibles restent dans le même ordre de grandeur, alors que les estimations du nombre de familles de modules (basée sur la comparaison des séquences protéiques) prédisent entre 8 000 et 60 000 familles [Coulson et Moult, 2002; Schaeffer et Daggett, 2011; Wolf *et al.*, 2000]. La variation des effectifs est en partie due aux méthodes statistiques d'estimation utilisées, mais également à la manière dont les domaines et modules singletons sont pris

en compte. En effet, la plupart des bases de données utilisant les données structurales négligent ces domaines singletons alors que les données des séquences protéiques permettent de les détecter plus facilement [Coulson et Moult, 2002]. Ainsi, l'analyse des séquences prédit un nombre de modules singletons de l'ordre de la centaine de milliers [Bru *et al.*, 2005; Heger et Holm, 2003], alors qu'au niveau des repliements, Lee *et al.* [2003] prédit 10 000 familles de domaines.

Malgré la prudence qui serait requise dans l'utilisation du domaine comme unité d'évolution, les termes de domaine et module sont généralement considérés comme synonymes.

1.2 La diversification des protéines

Les protéines divergent les unes des autres par le remplacement d'acides aminés dans leur séquence mais aussi par des changements plus drastiques comme les insertions, les délétions et les duplications d'une portion plus longue d'acides aminés qui peut éventuellement contenir un ou plusieurs modules. L'écrasante majorité des mutations qui affectent les séquences codantes sont délétères et éliminées par la sélection naturelle, et la plupart de celles qui sont retenues n'altèrent pas fondamentalement la structure de la protéine [Chothia et Lesk, 1986; Russell *et al.*, 1997; Worth *et al.*, 2009]. Cependant, on sait que des substitutions d'acides aminés peuvent avoir un impact considérable sur la structure et la fonction de la protéine [Kinch et Grishin, 2002]. On trouve de multiples exemples chez les enzymes où des mutations dans le site actif, par exemple, peuvent changer la spécificité de la protéine pour un substrat [Bartlett *et al.*, 2003]. L'analyse des superfamilles enzymatiques de CATH [Todd *et al.*, 2001] a montré que 25% d'entre elles étaient composées de membres catalysant des réactions différentes. Des expériences de mutagenèse dirigée ont mis en évidence des modifications structurales drastiques suite au changement de quelques acides aminés [Goldstein, 2008; Meier *et al.*, 2007]. L'exemple de la protéine répresseur Arc présenté dans la figure 1.5 illustre cette idée. La protéine mutante Switch Arc présente deux mutations ponctuelles adjacentes (N11L et L12N) par rapport à la protéine sauvage. Ces deux mutations sont à l'origine d'une modification structurale majeure dans laquelle les brins antiparallèles des feuillets β de la protéine sauvage sont transformées en deux hélices α. Une troisième protéine mutante, Arc-N11L, ne porte qu'une seule des deux mutations et possède donc une Leucine aux positions 11 et 12. Cette protéine intermédiaire est capable d'adopter l'une ou l'autre des conformations de la protéine Arc-répresseur (naturelle ou changée) en fonction de l'environnement (température ou condition de solvant). Cet exemple illustre la complexité de la relation entre l'espace des structures et l'espace des séquences : il existe des raccourcis dans l'espace des séquences qui permettent de faire des sauts importants dans l'espace des structures.

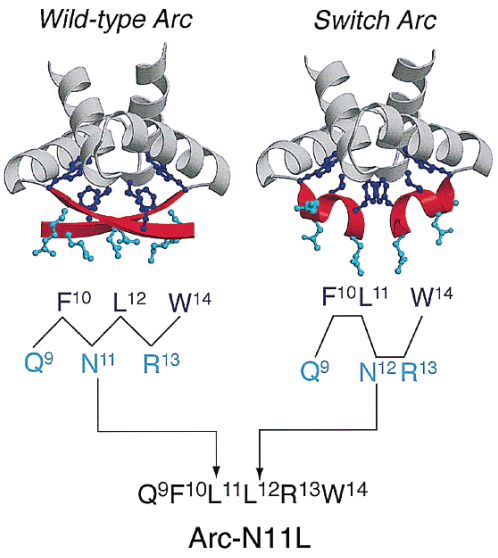


FIGURE 1.5 – **Partition des résidus hydrophobes et polaires entre la surface et l'intérieur des homodimers de la protéine Arc à l'état naturel et modifié.** Les résidus 8 à 46 sont représentés avec MOLSCRIPT [KRAULIS, 1991], avec les régions structurellement différentes en rouge (résidus 8 à 14) et avec les chaînes latérales intérieures et à la surface colorées respectivement en bleu et cyan. La protéine Arc-N11L contient une mutation de surface entre un acide aminé polaire (Asparagine) et un acide aminé hydrophobe (Leucine) par rapport aux deux protéines Arc représentées modifiant le cœur hydrophobe de ces deux protéines. Extrait de Cordes *et al.* [2000].

1.2.1 La combinatoire des domaines

1.2.1.1 Le paradigme de l'évolution des protéines modulaires

Le paradigme de l'évolution des protéines modulaires pourrait être exprimé ainsi : il existe un répertoire limité de domaines à partir duquel l'ensemble des protéines actuelles ont été formées [Apic *et al.*, 2001a; Bashton et Chothia, 2002; Chothia *et al.*, 2003; Patthy, 1999]. La détermination des premiers domaines structuraux ainsi que leur présence dans de nombreux contextes protéiques différents ont été les premiers éléments étayant cette théorie [Bork, 1991], dans laquelle la formation de nouvelles protéines semble s'expliquer par du bricolage d'unités évolutives définies dans les protéines [Jacob, 1977].

L'analyse de la combinaison des domaines montre que la distribution du nombre de partenaires différents suit une loi de puissance : une minorité de domaines sont très versatiles, c'est-à-dire qu'ils se retrouvent associés à de nombreux autres domaines, tandis que la majorité est associée à très peu d'autres domaines [Apic *et al.*, 2001b]. L'abondance d'un domaine est fortement corrélée à sa versatilité [Vogel *et al.*, 2005], c'est-à-dire à sa capacité à être associé à d'autres domaines. Cependant, certains domaines sont fréquemment associés aux mêmes domaines impliquant une surreprésentation de ces combinaisons dans les architectures des protéines. Vogel *et al.* [2004b] les nomment *supradomaines*. L'association des propriétés des domaines serait utile dans différents

contextes protéiques. L'examen de leur structure tridimensionnelle a mis en évidence deux types de relation spatiale qui sont fonctions du degré d'interaction entre les deux domaines : sans interaction, la structuration spatiale des domaines n'est pas cruciale et donc différente d'une protéine à l'autre, tandis que dans le cas d'une interaction fonctionnelle, l'interface entre les domaines est contrainte et donc conservée dans l'ensemble des structures. Ces deux types existent en parts équivalentes dans les structures disponibles [Vogel *et al.*, 2004b]. Les auteurs définissent ces *supradomaines* comme une nouvelle unité évolutive qui peut elle-même se combiner à d'autres domaines et se dupliquer. Cependant, cette proposition pourrait témoigner surtout de la confusion classique entre les notions de domaine et de module : un supradomaine ne serait qu'un module composé de deux domaines.

Dans la majorité des combinaisons, l'ordre des domaines est conservé [Apic *et al.*, 2001a; Kummerfeld et Teichmann, 2009; Vogel *et al.*, 2004a] : la combinaison du domaine A avec le domaine B sera majoritairement retrouvée dans la conformation $A - B$. Cependant, Kummerfeld et Teichmann [2009] mettent en avant que la combinaison $B - A$, bien que rare, est statistiquement surreprésentée parmi les domaines partenaires de A et B. La conservation de cette association peut s'expliquer de deux manières : soit les combinaisons dérivent d'une combinaison ancestrale, soit cette combinaison a été créée plusieurs fois indépendamment de manière convergente, suggérant une contrainte fonctionnelle forte liée à l'association des deux domaines. Il semblerait que dans la majorité des cas la conservation soit la conséquence d'une histoire évolutive commune, la contrainte fonctionnelle apparaît lorsque la connexion entre les domaines est courte, laissant peu de souplesse à la structure globale [Bashton et Chothia, 2002].

1.2.1.2 Évolution des architectures de protéines

Le réarrangement de domaines comprend l'ensemble des mécanismes susceptibles de modifier le contenu en domaines ou l'ordre d'une architecture multidomaine [Pasek *et al.*, 2005]. Les duplications, les insertions, les délétions et les substitutions de domaines sont généralement considérées comme les événements évolutifs majeurs à l'origine des modifications des architectures [Bornberg-Bauer *et al.*, 2005; Moore *et al.*, 2008] (voir figure 1.6). Ils peuvent se combiner pour donner des réarrangements plus complexes comme les permutations circulaires.

Les insertions/délétions de domaines sont une grande source de variabilité dans les architectures de domaines (figure 1.6A). L'analyse des arrangements de domaines dans les chromosomes bactériens [Pasek *et al.*, 2006] a permis de mettre en évidence que les insertions/délétions sont plus fréquentes aux extrémités N et C-terminales des architectures qu'à l'intérieur, et que les fusions et fissions de gènes semblent être des mécanismes dominants. L'acquisition d'un nouveau codon initiateur ou la perte d'un codon stop pourraient expliquer ces événements de fusion et fission [Weiner *et al.*, 2006]. L'analyse des protéines multidomaines chez les eucaryotes a permis

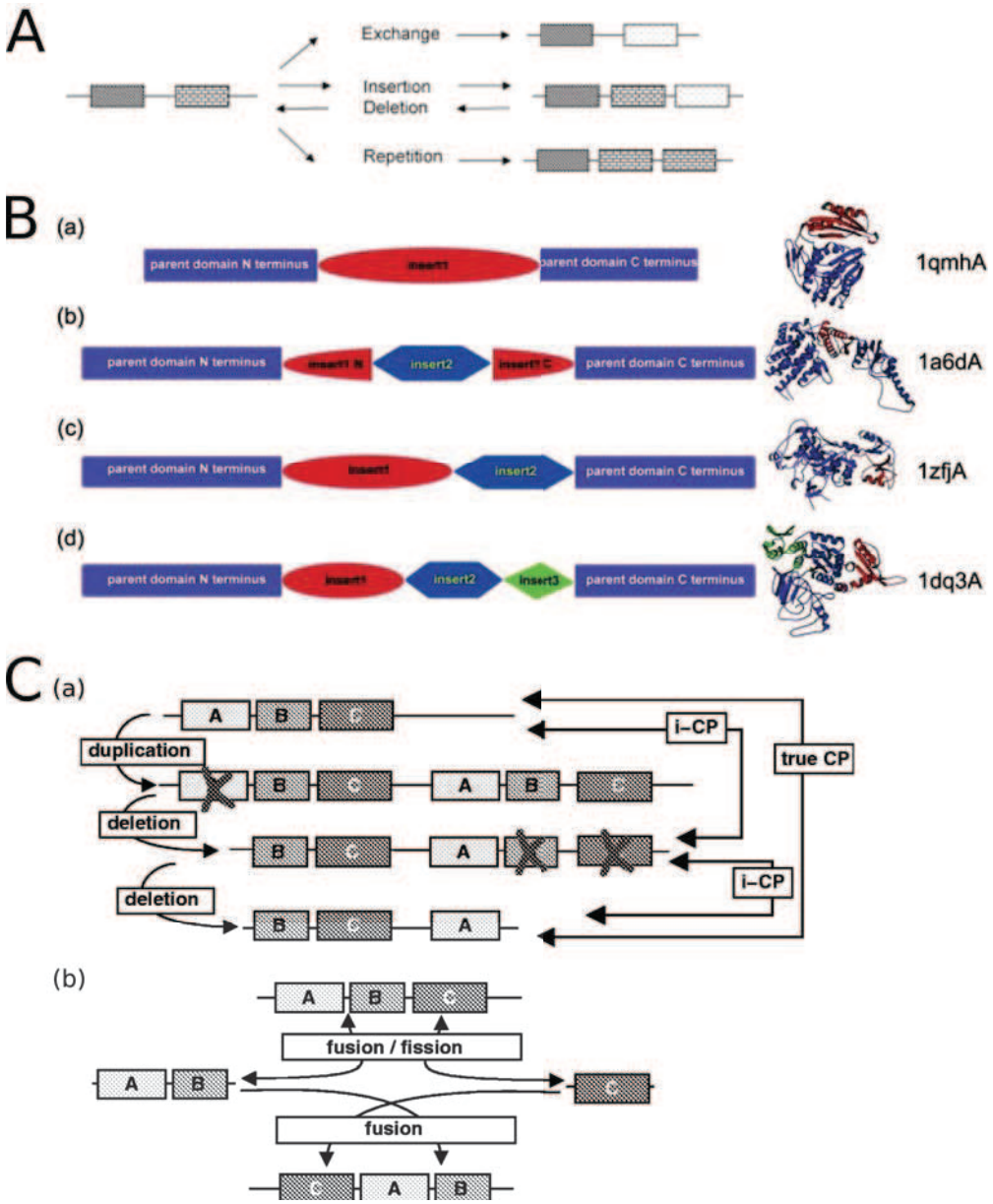


FIGURE 1.6 – **Modifications des architectures de domaines.** (**A**) Les événements élémentaires permettant de créer de nouvelles architectures sont les substitutions de domaines, les insertions/délétions et les duplications, extrait de [Pasek *et al.*, 2006]. (**B**) Insertions emboîtées, extrait de [Aroul-Selvam *et al.*, 2004].(a) Insertion simple (b) Insertion emboîtée (c) Insertion double. (d) Insertion triple. (**C**) Origines des permutations circulaires. Elles peuvent être obtenues à partir d'une duplication de l'architecture suivie de pertes différentielles (a) ou bien de deux événements de fusion indépendants (b). Extrait de Weiner et Bornberg-Bauer [2006].

de mettre en évidence l'importance des insertions/délétions par rapport aux répétitions internes et aux substitutions [Björklund *et al.*, 2005]. Il existe également des cas d'insertions emboîtées de domaines [Aroul-Selvam *et al.*, 2004] (figure 1.6B), c'est-à-dire, des cas où un domaine est inséré à l'intérieur d'un autre domaine. L'analyse des protéines de la PDB [Andreeva *et al.*, 2008] a mis en évidence 40 cas d'insertions emboités. Dans la majorité des cas, une seule insertion est trouvée

(figure 1.6B(a)) mais il existe des cas d'insertions emboîtées multiples. Ces insertions se retrouvent majoritairement dans les structures de type α/β ou $\alpha + \beta$ dans les régions contenant des boucles.

Les évènements de fusion/fission ont été étudiés dans le cadre des prédictions d'interactions fonctionnelles protéines-protéines. Marcotte *et al.* [1999] ont développé une méthode basée sur la génomique comparative dans laquelle l'interaction fonctionnelle entre deux protéines est prédite s'il existe une troisième protéine correspondant à une forme fusionnée des deux premières. Par exemple, les protéines Gyr A et Gyr B, deux sous-unités de l'ADN gyrase d'*Escherichia coli* qui interagissent fonctionnellement sont retrouvées fusionnées dans la protéine Topoisomérase II présente chez *Saccharomyces cerevisiae*. La protéine fusionnée est nommée la Pierre de Rosette en référence à la stèle gravée datant de l'Égypte ancienne qui a permis le déchiffrage moderne des hiéroglyphes ; cette protéine permet de déchiffrer l'interaction entre deux autres protéines. D'un point de vue évolutif, l'existence de ces trois protéines peut être le résultat de fusions ou de fissions. Il est clair que pour les protéines pour lesquelles on trouve des pierres de rosette tendent à participer aux mêmes fonctions. Cependant, certains génomes réduits présentent une proportion importante de Pierres de Rosette comme celui de *Mycoplasma genitalium* dans lequel 15 pierres ont été détectées Enright et Ouzounis [2001]. Elles correspondent à des protéines distinctes chez *Mycoplasma pneumoniae*, une espèce proche dont le génome est 17% plus gros. Dans ce cas, la fusion de ces protéines pourrait aussi bien être le résultat de la pression sélective pour favoriser leur co-fonctionnement que de celle qui tend à rendre le génome plus compact.

L'orientation des événements de fusion et fission sur des arbres phylogénétiques à partir de méthodes de parcimonie a permis de mettre en évidence que les événements de fusion sont quatre fois plus communs que les événements de fission [Kummerfeld et Teichmann, 2005; Snel *et al.*, 2000].

La combinaison de ces différents événements peut aboutir à des réarrangements plus complexes comme les permutations circulaires (CP, figure 1.6C). Elles correspondent à de nouvelles architectures dans lesquelles l'ordre séquentiel des domaines est inversé, c'est-à-dire où le fragment N-terminal d'une protéine est similaire au fragment C-terminal d'une autre et vice-versa. Les premiers exemples de CP ont été décrits pour l'adénine méthyltransférase [Jeltsch, 1999] puis pour différents ABC transporteurs, des déshydrogénases, des chitinases et des protéines de fixation aux oligopeptides [Weiner *et al.*, 2005]. L'une des hypothèses expliquant ces permutations circulaires (et leur fixation) est que la fonctionnalité et la structure 3D des domaines sont plus contraintes que l'ordre des domaines dans l'arrangement. Deux mécanismes majeurs peuvent expliquer ces permutations : un événement de duplication en tandem sur le chromosome suivi de pertes [Jeltsch, 1999] (figure 1.6C.a), ou bien deux événements de fusions indépendants [Bujnicki, 2002] (figure 1.6C.b). Dans le premier cas, après la duplication, il peut apparaitre de nouveaux codons stop et start qui produisent des états intermédiaires de CP. Leur observation permet alors de retracer l'histoire des architectures correspondantes. L'analyse systématique des permutations circulaires détectées dans

les architectures de ProDom version 2004 a permis de mettre en évidence que les événements de fusions indépendants sont les plus fréquents [Weiner *et al.*, 2006].

1.2.2 Diversification des répertoires de protéines

1.2.2.1 L'univers des séquences protéiques : une diversité croissante

Depuis que le premier génome bactérien d'*Haemophilus influenzae* a été séquencé en 1995 [Fleischmann *et al.*, 1995], le nombre de génomes bactériens complètement séquencés double tous les 20 mois environs (et tous les 24 mois environ pour les Archébactéries). De plus, de nombreux projets de métagénomiques visant à échantillonner les populations de microorganismes [Yooseph *et al.*, 2007] contribuent à la croissance exponentielle du nombre de séquences protéiques disponibles. Le regroupement de ces séquences en familles de protéines homologues permet d'appréhender la diversité de l'univers des protéines du monde vivant séquencé. La principale caractéristique de cet univers de famille est sa continuelle croissance avec l'ajout de génomes complètement séquencés (voir figure 1.7a) [Grabowski *et al.*, 2007; Kunin *et al.*, 2003; Marsden *et al.*, 2006]. Avec les quelque 1 700 génomes disponibles actuellement [6], cette croissance ne semble pas ralentir et indiquer une quelconque saturation de l'univers des familles de protéines. Le nombre d'espèces séquencées est relativement élevé (par rapport à la croissance de l'acquisition des données), mais reste très restreint par rapport à celui des espèces du monde vivant dont l'estimation est d'environ 10^7 [Choi et Kim, 2006], suggérant un espace des séquences extrêmement vaste. Le séquençage de génomes eucaryotes est généralement à l'origine d'une hausse significative du nombre de familles de protéines (figure 1.7a). Cependant, la croissance observée n'est pas due aux seuls génomes eucaryotes puisque même en ne considérant que des génomes procaryotes, celle-ci ne sature toujours pas.

L'univers des protéines aujourd'hui disponible a mis en évidence une diversité jusque-là insoupçonnée. La structure tridimensionnelle et la fonction de la plupart des protéines ne sont pas connues. Entre 30 et 40% des protéines sont classées comme "protéines hypothétiques" [Jaroszewski *et al.*, 2009]. De plus, l'univers des protéines est également caractérisé par la présence de séquences orphelines pour lesquelles aucune séquence homologue n'a été détectée. Le nombre de ces séquences croît avec l'ajout de nouveaux génomes (figure 1.7b) à une vitesse moins importante que le nombre de familles mais ne semble pas non plus saturer. L'augmentation de l'échantillonnage taxonomique ne permet pas de résorber ce stock de séquences orphelines, bien que des séquences homologues soient détectées pour certaines d'entre elles. L'analyse de plusieurs souches d'*Escherichia coli* [Lukjancenko *et al.*, 2010] a mis en évidence la présence de séquences orphelines dans chacune des souches appuyant l'idée que l'augmentation de l'échantillonnage taxono-

6. La base de données GOLD contient la liste des génomes complètement séquencés ainsi que les listes des génomes en cours ou en projet de séquençage, www.genomesonline.org/cgi-bin/GOLD/bin/gold.cgi

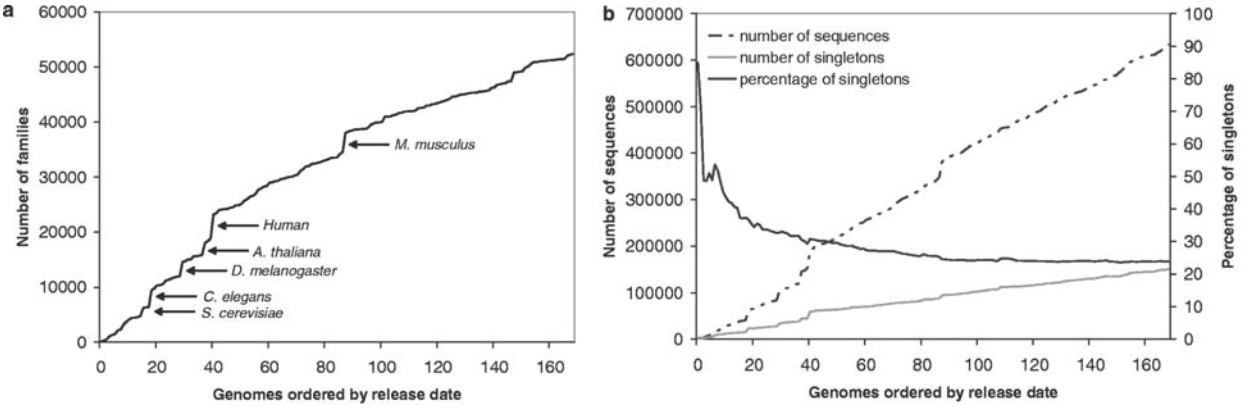


FIGURE 1.7 – **Évolution de la diversité protéique avec le séquençage de nouveaux génomes.**(**a**) Accumulation des familles de protéines dans Gene3D [Lees *et al.*, 2010] avec le temps représenté par l'ordre dans lequel de nouveaux génomes complètement séquencés ont été publiés. Les eucaryotes étiquetés sont impliqués dans de nombreuses nouvelles familles. (**b**) Avec la publication de chaque nouveau génome, le nombre de séquences singletons augmente dans la base Gene3D alors que le pourcentage global de séquences singletons diminue progressivement. Extrait de Marsden *et al.* [2006].

mique ne réglera pas complètement leur sort (figure 1.8). Les séquences orphelines sont le sujet de nombreuses controverses. D'un côté, leur statut de gène a souvent été remis en question notamment chez les eucaryotes où la prédiction des gènes est plus délicate que chez les procaryotes [Wood *et al.*, 2001]. Ainsi les prédictions chez l'homme sont passées de plus de 30 000 gènes [Lander *et al.*, 2001] à un peu moins de 20 000 aujourd'hui [Goodstadt et Ponting, 2006]. D'un autre côté, ces séquences semblent également être une partie intégrante de chaque génome [Marsden *et al.*, 2006] et les premières représentantes de nouvelles familles. Elles peuvent donc nous éclairer sur la manière dont les nouvelles protéines apparaissent et sur la diversification des protéines.

1.2.2.2 L'évolution des répertoires de protéines

Les contenus génomiques sont d'une très grande diversité quantitative : on trouve des répertoires de moins de 500 gènes (ex. *Nanoarchæum equitans*) à 10 000 gènes (ex. *Magnetospirillum magnetotacticum*) chez les procaryotes, et de 2 000 (ex. *Encephalitozoon cuniculi*) à plus de 40 000 (ex. peuplier) chez les eucaryotes. Cependant, la diversité des familles de protéines représentées dans l'ensemble des génomes est plus importante. Alors que les génomes contiennent en général quelques milliers de gènes, on définit actuellement plusieurs centaines de milliers de familles à partir des génomes complètement séquencés.

Ces différences de contenus sont également observées pour des espèces très proches et même pour différentes souches d'une même espèce. En génomique comparative, on distingue le core-génome contenant l'ensemble des gènes communs à toutes les souches ou espèces analysées, et le génome accessoire qui contient les gènes présents dans au moins une souche [Medini *et al.*, 2005], mais pas dans toutes. L'union de ces deux ensembles forme le pan-génome. La comparaison de 61

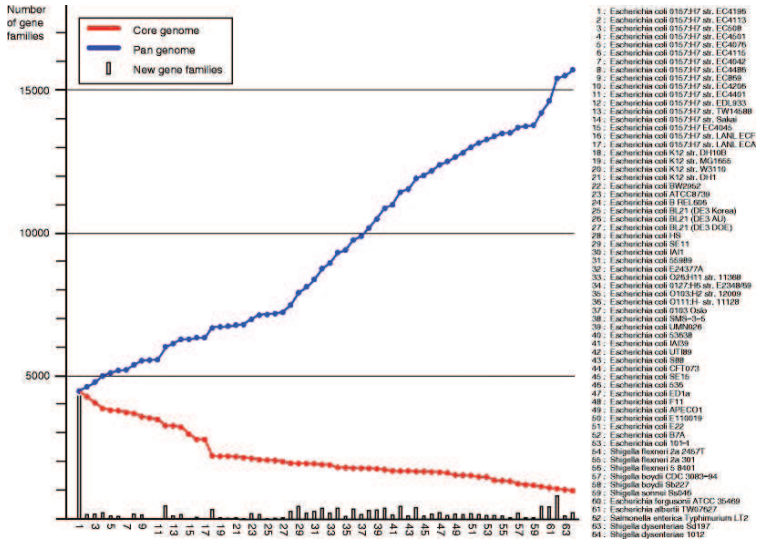


FIGURE 1.8 – **Pan et core génomes des espèces analysées.** La courbe bleue du pan-génome donne le nombre cumulé de familles de gènes présentes dans les génomes pris en compte. La courbe rouge du core génome donne le nombre de familles de gènes conservées dans toutes les souches. Les barres grises montrent le nombre de nouvelles familles de gènes identifiées dans chaque génome. Les 63 génomes utilisés représentent dans l'ordre d'ajout 53 souches d'*Escherichia coli*, 3 souches de *Shiguella flexneri*, 2 souches de *Shiguella boydii*, 1 souche de *Shiguella sonneii*, 1 souche d'*Escherichia fergusonii*, 1 souche d'*Escherichia albertii*, 1 souche de *Salmonella enterica* et 2 souches de *Shiguella dysenteriae*. Extrait de Lukjancenko *et al.* [2010].

souches d'*Escherichia coli* a mis en évidence que le core génome représente seulement 6% du pan-génome [Lukjancenko *et al.*, 2010] (figure 1.8). Cette proportion sera amenée à diminuer de plus en plus avec l'ajout de nouvelles souches. La taille du core génome à l'échelle du monde vivant est de plus en plus restreinte avec l'ajout de nouveaux génomes. Koonin [2003] l'estimait à 60 protéines en considérant 100 génomes, et dans la dernière version d'HOGENOM [Penel *et al.*, 2009], contenant 946 génomes, aucune famille n'est universelle. Cette notion de core génome atteint donc ses limites quelque soit l'échelle évolutive utilisée et traduit une évolution des répertoires dynamique avec un renouvellement important. Dans l'exemple d'*Escherichia coli*, l'ensemble des gènes additionnels spécifiques d'une ou quelques souches, montre à quel point l'acquisition de nouveaux gènes peut être rapide (voir figure 1.8).

La diversité des contenus des génomes actuels que l'on retrouve dans l'ensemble du monde vivant suggère une histoire évolutive complexe dans laquelle les contenus sont sans cesse remodelés. Parmi les mécanismes à l'origine de nouveaux gènes, on peut citer les duplications (par recombinaison par exemple), la rétroposition, les transferts horizontaux et l'apparition *de novo* (figure 1.9A-D). L'analyse de 308 nouveaux gènes dans la lignée de *Drosophila* [Zhou et Wang, 2008] a permis de mettre en évidence l'importance relative de ces différents mécanismes dans cette lignée (figure 1.9E).

- La duplication des gènes représente le mécanisme majeur à l'origine des nouveaux gènes

dans la lignée de *Drosophila*. Il est impliqué dans l'apparition de plus de 80% des nouveaux gènes. Ce mécanisme est considéré depuis près de 40 ans comme l'une des forces majeures dans l'évolution de nouvelles fonctions [Ohno, 1970], bien que dans 50 à 80% des cas, les copies accumulent des mutations jusqu'à leur inactivation (pseudogénisation) [Lynch et Force, 2000]. Dans les autres cas, il a été proposé que deux mécanismes importants permettent la rétention du gène dupliqué : la néofonctionalisation et la subfonctionalisation [Roth *et al.*, 2007]. Lorsque la duplication de gènes créé deux loci redondants, ceux-ci sont libres d'accumuler des mutations qui peuvent conduire à la perte ou au gain d'une fonction tant que l'une des copies assure la fonction ancestrale : c'est la néofonctionalisation [Ohno, 1970]. Dans le cas de la subfonctionalisation [Hughes, 1994; Lynch et Force, 2000], les deux copies accumulent des mutations dégénératives causant la perte complémentaire de sous-fonctions. La duplication de gènes se produit par des événements de recombinaison [Arguello *et al.*, 2006], de rétrotransposition (à l'aide d'éléments transposables comme LINE-1 [Babushok *et al.*, 2007]) ou bien lors de duplications globales de génomes ou de chromosomes.

- La rétroposition ou rétroduplication [Kaessmann, 2009] est un mécanisme dans lequel un ARN messager (ARNm) est rétrotranscrit à un autre locus sur la séquence génomique. Elle peut être considérée comme un cas particulier de duplication où le gène ne possède plus d'introns ni d'éléments régulateurs. Ce mécanisme a été très peu mis en évidence dans la lignée de *Drosophila*.
- L'apparition de nouveaux gènes à partir de séquences non-codantes ou *de novo* est un mécanisme généralement considéré très peu probable comparé aux autres mécanismes présentés ci-dessus [Chothia *et al.*, 2003; Jacob, 1977]. Mais quelques exemples dans la lignée des Drosophiles [Levine *et al.*, 2006; Zhou *et al.*, 2008] et dans celle de l'homme [Knowles et McLysaght, 2009] suggèrent qu'il serait probable. Il serait à l'origine d'un peu moins de 12% des nouveaux gènes dans la lignée de Drosophila (figure 1.9E)

Une manière d'appréhender l'histoire évolutive des répertoires de protéines est d'inférer les contenus des génomes ancestraux à partir de scénarios d'évolution des familles présentes dans les génomes actuels. Alors que les simples profils phylogénétiques décrivent la présence et l'absence dans les espèces contemporaines, les scénarios d'évolution propagent cette information dans l'arbre des espèces correspondant à partir d'un modèle d'évolution. Ces scénarios peuvent être interprétés comme une succession d'événements élémentaires tels que le gain, la duplication, la perte et la transmission verticale. Il est donc possible d'appréhender la diversification des protéines en suivant la diversification des espèces. L'accès aux contenus ancestraux permet de situer précisément les différents événements évolutifs ainsi que leurs fréquences le long de chaque lignée. Il est donc possible de préciser la dynamique du renouvellement des génomes, ce qui n'est pas accessible avec les seuls répertoires des espèces actuelles.

Dans la volonté d'inférer les contenus des génomes ancestraux, il y a certes la possibilité de

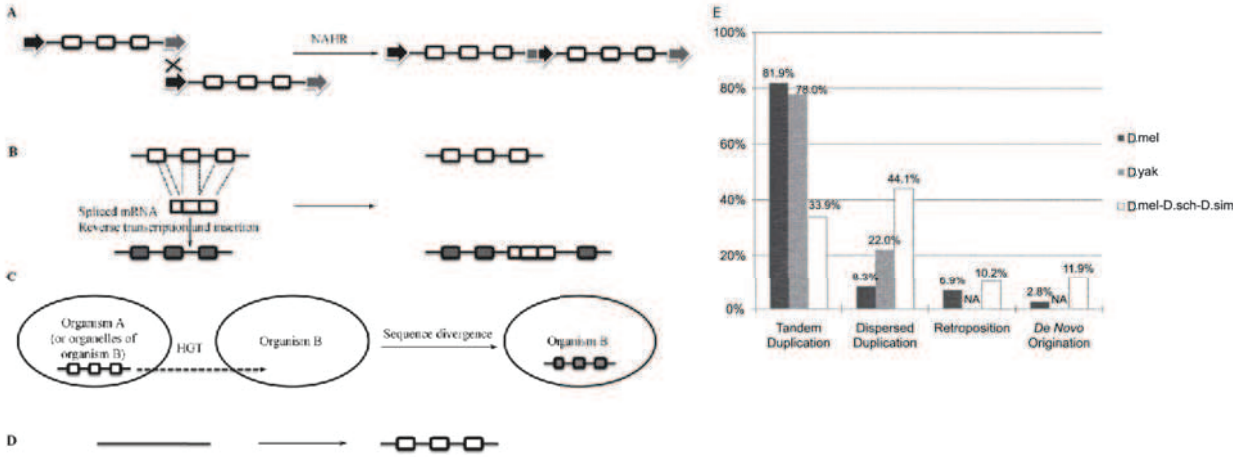


FIGURE 1.9 – **Mécanismes à l'origine de nouveaux gènes.** Les exons sont représentés par les boîtes et les régions intergénique ou intronique sont représentées par des lignes. (**A**) La duplication de gène par recombinaison homologue non allélique (NAHR). Les flèches noires et grises représentent des éléments répétés d'une même famille encadrant le gène dupliqué. Une NAHR entre ces éléments conduirait à une nouvelle copie du gène. (**B**) La rétroposition. Lorsqu'un ARNm épissé d'un gène parental est accidentellement rétrotranscrit et inséré à un autre locus, cela produit une copie du gène parental sans introns. (**C**) Transfert horizontal. De nouveaux gènes peuvent être introduits d'un organisme A vers un organisme B par transfert horizontal. (**D**) Origine *de novo*. Les nouveaux gènes apparaissent à partir de séquences non codantes. (E) La répartition des mécanismes à l'origine des 72 nouveaux gènes de *Drosophila melanogaster*, des 177 nouveaux gènes de *D. yakuba* et des 59 nouveaux gènes apparus chez l'ancêtre de *D. melanogaster*, *D. simulans* et *D. sechellia* est représentée respectivement par les barres gris foncé, gris clair et blanc. Extrait de Zhou et Wang [2008].

comprendre la dynamique de leur évolution, mais il y a aussi la volonté de mettre en relation la composition des génomes des espèces avec leur style de vie et leur adaptation à leur environnement. Différentes analyses de groupes d'espèces spécifiques a permis de mettre en perspective les fonctionnalités gagnées et perdues avec le style de vie des espèces actuelles. Ainsi l'adaptation des espèces à de nouvelles niches ou de nouveaux hôtes corrèle assez bien avec la modification du contenu des génomes, notamment chez les bactéries symbiotiques et parasitaires [Boussau *et al.*, 2004; Kettler *et al.*, 2007; Marri *et al.*, 2007].

Par exemple, l'analyse des contenus des génomes ancestraux des *Alphaprotéobactéries* [Boussau *et al.*, 2004] a mis en évidence l'expansion massive des répertoires des Rhizobiales (symbiotes de plantes) et deux réductions extrêmes indépendantes dans la lignée de *Rickettisa* et *Wolbachia* (bactéries intracellulaires obligatoires) et dans la lignée de *Brucella* et *Bartonella* (bactéries intracellulaires facultatives). La figure 1.10 résume ces extrêmes variations de fréquences. Dans les deux cas, les gènes impliqués dans les modifications des répertoires sont associés aux fonctions liées à l'interaction avec le nouvel environnement des bactéries. On peut noter par exemple la perte indépendante des gènes impliqués dans la mobilité (assemblage du pilus et biosynthèse des flagelles) pour les bactéries intracellulaires, alors que les expansions sont caractérisées par l'acquisition de protéines impliquées dans le transport et le métabolisme de nombreux métabo-

lites secondaires. Ces modifications suggèrent une réponse aux changements environnementaux où l'utilisation des ressources du sol et l'interaction avec les plantes ont augmenté au cours du temps. L'analyse spécifique des *Rickettisa* de Blanc *et al.* [2007] semble globalement en accord puisqu'ils prédisent un ancêtre des *Rickettsia* adapté à la vie cellulaire parasitaire, avec des pertes massives antérieures.

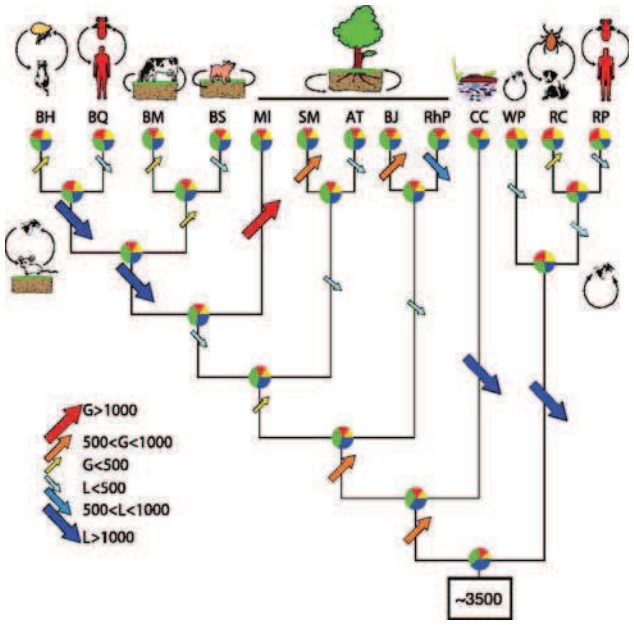


FIGURE 1.10 – **Gain et perte de gènes au cours de l'évolution des Alphaproteobactéries.** Les flèches pointant vers le haut indiquent une expansion des génomes (G) et les flèches pointant vers le bas indiquent une réduction des génomes (L). Leur couleur et leur taille représentent le nombre net de gènes gagnés ou perdus sur chaque branche. Les diagrammes de couleurs représentent la fraction relative des gènes associés à différentes catégories fonctionnelles. Jaune : traitement et stockage de l'information ; vert : métabolisme ; rouge : processus cellulaires et bleu : fonction mal caractérisée. Les abréviations des noms des espèces sont les suivantes : RP, *Rickettsia prowazekii* ; RC, *Rickettsia conorii* ; BQ, *Bartonella quintana* ; BH, *Bartonella henselae* ; BM, *Brucella melitensis* ; BS, *Brucella suis* ; CC, *Caulobacter crescentus* ; AT, *Agrobacterium tumefaciens* ; SM, *Sinorhizobium meliloti* ; ML, *Mesorhizobium loti* ; and BJ, *Bradyrhizobium japonicum* ; WP, *Wolbachia pipientis* ; RhP, *Rhodopseudomonas palustris*. Extrait de Boussau *et al.* [2004].

Cependant, bien que dans certains cas, comme pour les bactéries symbiotiques, de nombreux indices indiquent un lien entre l'évolution des répertoires de gènes et l'adaptation à l'environnement, il reste de nombreux cas où cette relation n'est pas évidente. Par exemple, les *Prochlorococcus* (des Cyanobactéries marines) présentent également dans certaines lignées des réductions du répertoire de gènes [Kettler *et al.*, 2007]. Mais leur mode de vie ainsi que leur environnement ne permettent pas d'expliquer ces réductions, des analyses complémentaires sont donc nécessaires pour expliquer ces variations.

1.3 Méthodes d'analyse de la modularité des protéines

1.3.1 Méthodes de détection et de recherche d'homologie entre protéines

La détection des domaines protéiques est une étape fondamentale pour la caractérisation des architectures de domaines. La définition du domaine est historiquement structurale, cependant, la croissance des données structurales n'est pas suffisante pour annoter entièrement l'univers des protéines, c'est pourquoi la définition complémentaire des modules basées sur l'homologie des séquences protéiques est largement utilisée pour l'annotation modulaire des protéines.

Les méthodes de comparaison de séquences comme BLAST ont des difficultés à retrouver les homologies distantes lorsque la divergence des séquences est trop importante. Lorsque l'identité est inférieur à 30%, (la *Twilight Zone* [Doolittle, 1986]), la reconnaissance de l'homologie est améliorée en utilisant des profils de séquences, construits à partir de l'alignement multiple de séquences connues d'une famille de protéines ou de domaines homologues. La méthode de PSI-BLAST (Position-Specific Iterated BLAST [Altschul *et al.*, 1997]) construit un profil de l'alignement sous la forme d'une PSSM (Position Specific Scoring Matrix) qui calcule un nouveau score pour chaque acide aminée à chaque position de l'alignement. PSI-BLAST peut fonctionner de manière itérative en intégrant au profil les nouvelles séquences recrutées à chaque itération. La méthode HMMer [Finn *et al.*, 2010], utilisant un profil basé sur les modèles de Markov à états cachés, présente une meilleure sensibilité dans la détection des homologies anciennes. Les PSSMs utilisées dans PSI-BLAST sont ici formalisées en modèles probabilistes dans lequel les insertions et les délétions sont modélisées en plus des substitutions.

Lorsque l'identité descend en dessous de 15% (la *Midnight Zone* [Rost, 1998]), il devient nécessaire d'ajouter l'information structurale pour inférer l'homologie. En effet, les structures de protéines homologues sont bien mieux conservées au cours de l'évolution que ne le sont les séquences en acides aminés correspondantes. La définition de la similarité entre deux structures n'est pas unique. Les nombreuses méthodes développées reposent sur des scores de similarités différents. Par exemple, certaines méthodes alignent les résidus dans l'espace comme SSAP (Sequence Structural Alignment Program) [Orengo et Taylor, 1996] ou CORA [Orengo, 1999] et calculent un score associé à cet alignement qui permet de déterminer le degré de similarité entre les structures. D'autres méthodes alignent les séquences codées en structures secondaires pour lesquelles une matrice de dissimilarité est générée comme COMPARER [Sali et Blundell, 1990] ou VAST [Madej *et al.*, 1995]. Cette dernière propose le calcul d'une p-value donnant la signification statistique de l'alignement de façon analogue à la E-value de BLAST.

Les séquences protéiques peuvent être comparées sur l'ensemble de leur longueur et regroupées en familles de protéines homologues. On distingue deux types de bases de données : celles

utilisant une comparaison automatique de toutes les séquences et celles proposant une expertise manuelle des familles. Les premières définissent un critère de similarité entre les séquences puis les regroupent selon des méthodes spécifiques à chaque base de données comme le simple lien pour HOGENOM [Penel *et al.*, 2009] ou l'algorithme MCL [van Dongen, 2000] pour Tribes [Enright *et al.*, 2002]. On peut également citer les bases de données COG [Tatusov *et al.*, 2003] ou SYSTERS [Meinel *et al.*, 2005]. Les secondes définissent les familles de protéines à l'aide de motifs conservés spécifiques qui sont utilisés pour recruter de nouvelles séquences homologues. On peut citer PRINTS [Attwood *et al.*, 2003] et TIGRFAM [Haft *et al.*, 2003] par exemple, ces bases de données ont une couverture de l'univers des protéines assez réduite (quelques milliers de familles sont annotées).

Les bases de données de domaines et modules sont également nombreuses et diffèrent les unes des autres par les données initiales (séquences ou structures) et le protocole d'alimentation de la base de données (automatique, semi-automatique ou manuel). En effet, il existe des méthodes de construction automatique à partir des séquences protéiques comme ProDom [Bru *et al.*, 2005], EVEREST [Portugaly *et al.*, 2007] et ADDA [Heger *et al.*, 2005] qui proposent donc des familles de modules. Les méthodes semi-automatiques offrent une expertise manuelle des familles débouchant sur des modèles de Markov cachés (HMMs), des profils ou des motifs utilisés pour recruter de nouvelles séquences dans les bases de données comme Pfam [Finn *et al.*, 2010], Prosite [Sigrist *et al.*, 2002] ou SMART [Letunic *et al.*, 2009]. En règle générale, elles utilisent les familles de domaines structuraux comme données initiales, ainsi la majorité des modules décrits correspondent à des domaines. Enfin, il existe également des bases de données hiérarchiques de domaines structuraux comme CATH [Cuff *et al.*, 2009b], SCOP [Andreeva *et al.*, 2008] ou Dali [Holm et Rosenström, 2010] utilisant les données disponibles dans la PDB. Ces classifications sont manuelles, mais guidées par des algorithmes de comparaison de structures. Les domaines sont classés en fonction de leur similarité de séquences et de structures. On peut généralement décrire deux niveaux hiérarchiques spécifiques où les domaines sont regroupés en familles (et superfamilles) de domaines homologues. Les niveaux plus généraux regroupent ces familles en fonction des repliements tridimensionnels et du contenu en structures secondaires, l'hypothèse de l'homologie n'est pas vérifiée dans ces niveaux supérieurs.

1.3.2 Méthodes de comparaison des architectures

L'architecture de domaines des protéines est une suite ordonnée de domaines de l'extrémité N-terminale à C-terminale. Une manière de comparer ces architectures est de mesurer leur similarité à partir d'un alignement comme on mesure la similarité de séquences nucléiques ou protéiques. Cependant, l'alphabet correspond à l'ensemble des domaines disponibles, pour lesquels il est impossible de construire une matrice de substitution ou de dissimilarité entre tous les do-

maines. Cependant, des indices ont été créés pour mesurer une distance entre deux architectures. Par exemple, Björklund *et al.* [2005] ont défini la "*Domain distance*" qui compte le nombre de trous dans un alignement simple de deux architectures. Ils montrent que cette distance est corrélée au nombre d'événements évolutifs requis pour passer d'une architecture à l'autre (par insertion, délétion, substitution ou duplication). Pasek *et al.* [2006] mesurent la distance entre deux architectures comme le rapport du nombre de domaines communs et de la longueur de la plus grande architecture.

D'autres méthodes plus spécifiques de comparaison d'architectures ont été mises en place : notamment RASPODOM [Weiner *et al.*, 2005] qui détecte les permutations circulaires entre architectures, et DOMAIN TEAM [Pasek *et al.*, 2005] qui permet une analyse plus fine des mécanismes d'évolution des architectures en étudiant les microsynténies de domaines sur les chromosomes d'espèces procaryotes.

1.4 Méthodes d'analyse de scénarios d'évolution

La dimension évolutive est fondamentale dans la compréhension de la diversification des protéines. L'inférence de scénarios d'évolution permet de reporter sur un arbre des espèces des événements évolutifs comme le gain, la perte ou la duplication d'une famille de protéines pour laquelle on peut déduire une histoire évolutive. L'utilisation des données de génomes complets donne un accès aux répertoires ancestraux qui représentent un point de vue intéressant dans la diversification des espèces comme nous l'avons succinctement présenté précédemment.

De nombreuses méthodes ont été proposées pour réaliser ces inférences ces dernières années. Nous ne présenterons ici que les méthodes développées avant le démarrage de ce projet, les autres seront discutées dans le chapitre suivant en comparaison de la méthode que nous proposons.

1.4.1 Le critère de maximum de parcimonie

Les premiers modèles utilisés se basent sur un critère de maximum de parcimonie qui recherche le scénario expliquant le profil phylogénétique actuel en minimisant un score. Les méthodes développées modélisent essentiellement les événements de gain et de perte, mais les duplications peuvent également être prises en compte. De manière générale, soit les méthodes cherchent à minimiser le nombre de gains en attribuant des pénalités différentes aux événements de gain et de perte [Boussau *et al.*, 2004; Mirkin *et al.*, 2003; Snel *et al.*, 2002], soit le modèle minimise le nombre total d'événements en attribuant des pénalités identiques [Dagan et Martin, 2007; Fong *et al.*, 2007; Kunin et Ouzounis, 2003b].

Les pénalités utilisées pour les événements de gain et de perte sont homogènes sur l'ensemble des branches de l'arbre. Il n'y a donc généralement qu'un seul paramètre à estimer : la pénalité de gain relativement à celle associée à la perte fixée à 1. Le choix de la pénalité est assez différent d'une méthode à l'autre. Il repose généralement sur une analyse empirique des contenus des répertoires ancestraux inférés avec différentes valeurs de pénalités. D'un point de vue quantitatif, lorsque le gain est aussi pénalisé que la perte (ou moins), on augmente les événements de gains multiples qui seront récents et donc dépeupleront les nœuds ancestraux. Lorsqu'au contraire, le gain est extrêmement pénalisé, la plupart des scénarios présentent un unique gain suivi de nombreuses pertes. Les tailles des répertoires ancestraux sont dans ce cas plus grandes que celles des répertoires actuels. Dagan et Martin [2007] choisissent le jeu de pénalités qui permet d'obtenir une distribution des tailles des répertoires ancestraux similaires à celles des répertoires actuels . Mais on peut également construire un intervalle de valeurs excluant les cas extrêmes [Snel *et al.*, 2002]. Ce choix moins risqué permet d'obtenir des bornes inférieures et supérieures aux nombres d'événements inférés, mais les histoires spécifiques des familles ne sont plus interprétables indépendamment. On peut également se baser sur le contenu qualitatif des génomes ancestraux en considérant que le contenu doit permettre à la cellule de vivre et doit donc contenir l'ensemble des voies métaboliques nécessaire à sa survie de base [Boussau *et al.*, 2004; Mirkin *et al.*, 2003].

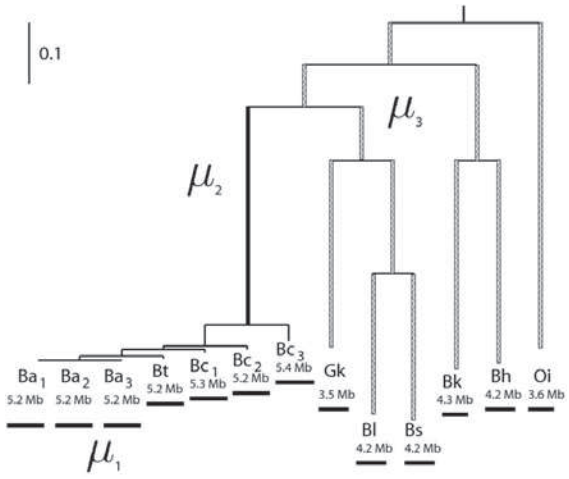
Les méthodes utilisant le critère de parcimonie ont l'avantage d'être rapides et le modèle évolutif utilisé est relativement simple : peu de paramètres sont utilisés. Cependant, quelques inconvénients méthodologiques et biologiques restent difficiles à améliorer avec ces approches. Par exemple, les pénalités des différents événements sont fixées a priori par l'expérimentateur en fonction des résultats obtenus avec différentes pénalités : cela dirige les résultats vers ce qui est attendu par l'expérimentateur et permet difficilement de complexifier le modèle en ajoutant des paramètres qu'il faudrait estimer de la même manière. De plus, les cas ambigus où plusieurs scénarios sont possibles sont résolus de manière arbitraire. L'ensemble de ces choix a un impact important sur les résultats obtenus.

1.4.2 Le critère de maximum de vraisemblance

Le critère de maximum de vraisemblance est une méthode statistique pour inférer les paramètres d'une distribution de probabilité d'un échantillon donné. Les événements évolutifs de gain, de perte et de duplication sont, dans ce type de modèle, associés à une probabilité que le modèle estime en fonction du jeu de données disponible en maximisant une fonction de vraisemblance donnée. Le scénario évolutif retenu est celui de probabilité maximum. L'utilisation de ce critère dans l'inférence des scénarios d'évolution permet de pallier certaines limitations des méthodes utilisant le critère de parcimonie notamment sur l'estimation des paramètres. Elles permettent également d'augmenter le nombre de paramètres du modèle. En effet, les résultats obtenus avec les

modèles de parcimonie suggèrent une dynamique différente des répertoires le long de l'arbre avec des événements de gain et de perte plus ou moins importants. Cette variabilité peut ainsi être prise en compte en supposant que le gain et la perte varient sur les branches de l'arbre [Hao et Golding, 2006; Iwasaki et Takagi, 2007], c'est-à-dire que les variations dans les génomes actuels s'expliquent par des taux d'évolution propres aux génomes. La méthologie d'inférence de caractères ancestraux avec des paramètres hétérogènes le long d'un arbre phylogénétique a été introduit par Pagel [1994]; Pagel *et al.* [2004] à l'aide d'un modèle de Markov continu.

(a)



Analyse de l'évolution dans le groupe des *Bacillus*. Différents taux d'insertion/délétion sont utilisés sur la phylogénie. Le taux μ_1 correspond aux branches du groupe *Bacillus cereus* (Bc), le taux μ_2 correspond à la branche (en noir) menant au groupe Bc et le taux μ_3 est appliqué aux branches restantes. Extrait de Hao et Golding [2008].

(b)

$$R_n = \begin{array}{c} \\ 0 \\ 1 \\ 2 \\ 3 \end{array} \begin{array}{c} \begin{array}{cccc} 0 & 1 & 2 & 3 \end{array} \\ \begin{bmatrix} -\alpha_n & \alpha_n & 0 & 0 \\ \beta_n & -\alpha_n - \beta_n & \alpha_n & 0 \\ 0 & \beta_n & -\alpha_n - \beta_n & \alpha_n \\ 0 & 0 & \beta_n & -\beta_n \end{bmatrix} \end{array}$$

Modélisation de l'évolution des familles par un processus de Markov, R_n est la matrice de transition d'un état à l'autre. Il y a 4 états possibles : absence (0), présence d'1, 2 ou 3 représentants de la famille. α_n correspond à la probabilité de gain et β_n à la probabilité de perte. Les duplications sont modélisées comme des gains (plafonnés à 3) et les événements multiples ne sont pas autorisés (passage de 1 à 3 par exemple). Extrait de Iwasaki et Takagi [2007].

FIGURE 1.11 – **Description de 2 modèles utilisant le critère de maximum de vraisemblance.** (a) Modèle prenant partiellement en compte la variabilité entre lignées. (b) Modèle prenant en compte les duplications et une variabilité sur toutes les branches. Les abréviations des noms des espèces de l'arbre sont les suivantes : Ba_1, *Bacillus anthracis* Ames ; Ba_2, *Bacillus anthracis* "Ames Ancestor" ; Ba_3, *Bacillus anthracis* Sterne ; Bt, *Bacillus thuringiensis* ; Bc_1, *Bacillus cereus* ZK ; Bc_2, *Bacillus cereus* ATCC 10 987 ; Bc_3, *Bacillus cereus* ATCC 14 579 ; Gk, *Geobacillus kaustophilus* ; Bl, *Bacillus licheniformis* ; Bs, *Bacillus subtilis* ; Bk, *Bacillus clausii* ; Bh, *Bacillus halodurans* ; Oi, *Oceanobacillus iheyensis* .

Hao et Golding [2006] ont étudié spécifiquement l'évolution des taux d'évolution des répertoires de gènes dans le sous-arbre des *Bacillus* où ils ont distingué jusqu'à 3 taux d'évolution différents à partir des longueurs de branches inférées sur l'arbre des espèces. Ils ont comparé plusieurs modèles emboîtés : les probabilités de gain et de perte sont homogènes sur l'arbre, elles sont différentes dans les espèces proches du groupe *Bacillus cereus*, et enfin un troisième taux est introduit pour distinguer la longue branche menant au groupe *Bacillus cereus* (figure 1.11a). La comparaison des vraisemblances montre que le modèle avec trois catégories de branches explique significativement mieux les données. Cette variabilité inter-espèce peut être analysée plus précisé-

ment en laissant chaque branche évoluer à sa propre vitesse. Iwasaki et Takagi [2007] ont mis en place un processus de Markov pour modéliser les gains, les pertes et jusqu'à 3 duplications (il y a donc 4 états possibles : absence, 1, 2 et au moins 3 copies présentes, figure 1.11b). L'utilisation du critère de maximum de vraisemblance permet de comparer efficacement (à partir d'un modèle probabilisé) différents modèles évolutifs pour décrire l'évolution des familles de protéines sans a priori sur les résultats à obtenir, ce qui est un avantage utile par rapport au critère de parcimonie.

1.4.3 Des données biologiques pour compléter les modèles

Dans l'ensemble des méthodes présentées précédemment, les données biologiques utilisées sont les familles de protéines (ou de domaines) homologues à partir desquelles on déduit un profil phylogénétique. Celui-ci code la présence et l'absence de chaque famille dans les espèces considérées à partir du contenu de la famille. D'un point de vue biologique, la perte d'une famille de gènes peut s'expliquer par une pseudogénisation du gène qui correspond à l'accumulation de mutations ou d'insertion/délétion jusqu'à inactivation du gène. Ils peuvent être détectés à l'aide de méthodes comme PSI-BLAST. Blanc *et al.* [2007] utilise les pseudogènes pour distinguer les gains multiples des pertes sur le sous-arbre des *Rickettsia*. La reconstruction des génomes ancestraux se base sur la présence dans les espèces contemporaines de gènes entiers ou sous forme de pseudogènes sans utiliser de critère de parcimonie ou de vraisemblance explicite. Cette méthodologie présente un point intéressant avec la recherche des pseudogènes qui sont de bons indicateurs des pertes, cependant leur détection devient difficile lorsque les espèces sont trop éloignées.

1.5 Réconciliation des approches modulaires et phylogénétiques de l'évolution des protéines

L'évolution des protéines a été abordée de deux manières complémentaires. D'un côté, l'univers des protéines est décrit comme extrêmement vaste et dynamique. Les répertoires de familles de protéines sont sans cesse remodelés au cours de l'évolution. Cependant, l'origine des nouvelles protéines inférée au cours de l'histoire n'est pas connue et les mécanismes à l'origine de ces gains ne sont pas détaillés, car inaccessible à partir des seuls répertoires de protéines. D'un autre côté, les protéines sont décrites comme modulaires et leur diversification est directement liée aux réarrangements des domaines composant leur architecture. L'univers des domaines est décrit comme restreint et statique avec une origine ancienne pour la majorité des domaines. La diversité des protéines s'explique principalement par la recombinaison de ces domaines à travers les mécanismes d'insertion, délétion, substitution et duplication. Très peu d'analyses ont restitué ces arrangements dans un contexte phylogénétique explicite pour proposer une origine aux architectures [Ekman

et al., 2007; Fong *et al.*, 2007; Jiang et Blouin, 2007]. Ces travaux reconstruisent les répertoires ancestraux en architectures de domaines et proposent une origine en comparant l'ensemble des architectures disponible au cours de l'évolution. Les travaux de Ekman *et al.* [2007] sur les eucaryotes ont permis de retracer l'apparition des architectures de protéines et des domaines protéiques dans l'histoire des eucaryotes. Les innovations en domaines sont majoritairement anciennes et correspondent à des protéines monodomaines, tandis que les architectures de domaines sont apparues majoritairement par réarrangement d'anciens domaines et architectures. La comparaison des architectures ancestrales a montré que le mécanisme principal de réarrangement est l'insertion de domaines. Cependant, dans cette analyse, la couverture des séquences protéiques par les domaines est d'un peu moins de 40%, ce qui peut être une source d'erreur quant à l'origine de certaines architectures.

Dans cette thèse, nous avons cherché à mettre en place une méthodologie réconciliant ces deux aspects de l'évolution des protéines. En effet, l'aspect modulaire des protéines est essentiel à la compréhension de la diversification des protéines, mais l'étude de l'évolution nécessite le positionnement de cette modularité sur un arbre des espèces. Ainsi, nous proposons une approche dans laquelle l'évolution des protéines est appréhendée du point de vue de leur décomposition en domaines ou modules en utilisant trois bases de données : HOGENOM pour les familles de protéines, Pfam pour les familles de domaines expertisés et ProDom pour les familles de modules protéiques construites automatiquement. Nous avons modélisé l'évolution de ces familles par un réseau Bayésien basé sur l'arbre phylogénétique des espèces. Les scénarios d'évolution les plus probables, qui reflètent la présence ou l'absence de chaque protéine, domaine ou module dans les espèces ancestrales, ont été inférés dans le cadre de ce modèle. La mise en relation de ces scénarios permet d'analyser l'émergence de nouvelles protéines en fonctions de domaines ou modules ancestraux.

Le chapitre 2 présente le modèle mis en place pour inférer les scénarios d'évolution des familles de protéines, de domaines et de modules. Celui-ci prend en compte l'hétérogénéité des probabilités de gain et de perte sur l'ensemble de l'arbre. L'estimation des paramètres du modèle ainsi que l'inférence des scénarios ont été réalisées par une méthode au maximum de vraisemblance basée sur le modèle de réseau Bayésien. Le chapitre 3 présente les résultats obtenus sur l'évolution des protéines et des domaines, mais aussi sur les prédictions de l'origine des nouvelles protéines du point de vue de Pfam et de celui de ProDom. Ces deux visions de l'évolution des protéines modulaires sont discutées en prenant en compte les différents biais associés à chacune des bases de données. Enfin, le chapitre 4 propose une discussion de ces résultats et de leurs implications générales.

Chapitre 2

Inférence des scénarios d'évolution avec les Réseaux Bayésiens

Notre approche consiste à mettre en parallèle l'évolution des familles de protéines et de modules dans le but d'expliquer l'histoire des protéines à partir de celle des modules qui composent leur architecture. L'histoire d'une famille est représentée par la succession des gains, des pertes et des transmissions verticales [Snel *et al.*, 2002] qui ont eu lieu au cours de l'évolution. Seul le profil phylogénétique des familles est disponible, il décrit la présence et l'absence d'une famille donnée dans les génomes actuels. L'inférence d'un scénario d'évolution d'une famille consiste à propager cette information dans les espèces ancestrales d'un arbre phylogénétique.

Nous avons mis en place un modèle d'évolution des familles dans lequel les probabilités de gain et de perte sont hétérogènes le long de l'arbre. Ce modèle permet de modéliser plus finement les différentes dynamiques évolutives des répertoires de protéines mises en évidence dans de nombreuses analyses [Boussau *et al.*, 2004; Kettler *et al.*, 2007]. Ce modèle est utilisé dans le contexte des réseaux Bayésiens qui réalisent l'estimation des paramètres et l'inférence des scénarios au maximum de vraisemblance. Cette structure est parfaitement adaptée à la problématique d'inférence et dispose d'algorithmes d'estimation des paramètres et d'inférence précis et optimisés. Cette méthodologie très générale s'applique déjà dans de nombreux domaines : la physique, la santé (notamment pour le diagnostic), la sociologie et de nombreuses problématiques biologiques comme l'inférence de réseaux cellulaires [Friedman, 2004], la prédiction d'interactions protéines-protéines [Jansen *et al.*, 2003] ou encore l'analyse de l'expression des gènes [Friedman *et al.*, 2000; Gevaert *et al.*, 2006].

L'objectif de ce chapitre est de présenter le modèle de réseau Bayésien utilisé pour inférer les scénarios d'évolution des familles de protéines et de modules. Après une présentation théorique de la méthodologie utilisée et la manière dont nous l'avons appliquée à notre problématique, nous présenterons les différentes analyses qui nous ont permis de la valider.

2.1 Construction du réseau bayésien : définitions et méthodologies

2.1.1 Un réseau Bayésien : définition et application

Un réseau bayésien est un modèle probabiliste qui décrit de manière graphique les dépendances conditionnelles entre des variables aléatoires d'intérêt. C'est un mélange de la théorie des graphes et de celle des probabilités afin de représenter une distribution de probabilités jointes sur un ensemble de variables aléatoires [1]. Cette structure est idéale pour formaliser un problème, acquérir de l'information et en extraire des connaissances. Elle permet d'effectuer des inférences dans un contexte d'incertitude.

Notre approche modélise les gains et les pertes le long de l'arbre des espèces. Nous travaillons avec des variables discrètes, ainsi l'ensemble des définitions et algorithmes présentés sont valides dans ce cas précis. Le cas de variables continues n'est pas présenté.

Définition 2.1. *Un **réseau Bayésien** $\mathcal{B}$ est défini par :*

- *un ensemble de variables aléatoires $\mathcal{U} = \{X_1, X_2, \ldots, X_n\}$ définies sur un espace probabilisé $(\Omega,\mathbf{P})$,*
- *un graphe dirigé et acyclique (DAG, pour Directed Acyclic Graph) $\mathcal{G}(V,E)$, où V est l'ensemble des nœuds associés aux variables aléatoires et E est l'ensemble des arcs de G représentant les dépendances conditionnelles entre elles,*
- *une distribution de probabilités conditionnelles associée à chaque variable X_i et déterminée en fonction des parents de X_i, $pa(X_i)$ dans $\mathcal{G}$. Si X_i n'a pas de parent alors on définit la distribution $\mathbf{P}(X_i)$.*

Théorème 2.1 (La règle de chaîne)**.** *Soit $\mathcal{B}$ un réseau Bayésien sur $\mathcal{U} = \{X_1,X_2,\ldots,X_n\}$. $\mathcal{B}$ définit une unique distribution de probabilités jointes $\mathbf{P}(\mathcal{U})$ donnée par le produit de toutes les tables de probabilités conditionnelles définies dans $\mathcal{B}$:*

$$\mathbf{P}(\mathcal{U}) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid pa(X_i)) \tag{2.1}$$

1. L'annexe A (page 103) présente des notions de base sur les probabilités associées aux réseaux Bayésiens ainsi que sur leur relation avec les graphes.

avec $pa(X_i)$ *l'ensemble des parents de* X_i *dans* $\mathcal{B}$,

et $\mathbf{P}(X_i \mid pa(X_i)) = \mathbf{P}(X_i)$ *lorsque* $pa(X_i) = \emptyset$

La définition d'un réseau Bayésien nécessite la spécification des variables aléatoires et des liens dirigés entre elles. Pour chacune des variables, la loi de probabilité conditionnelle est spécifiée. Dans le cas de variables discrètes, cette loi peut être représentée par une matrice spécifiant la probabilité des états de chaque variable en fonction des combinaisons possibles des états de ses parents dans le graphe.

2.1.2 Estimation des paramètres du modèle

Les probabilités conditionnelles ne sont en général pas connues, il est donc nécessaire de les estimer. Les réseaux Bayésiens proposent différentes approches d'estimation des paramètres en fonction du jeu de données disponible (complet ou incomplet). Dans les applications pratiques, les bases de données sont souvent incomplètes. Certaines variables ne sont observées que partiellement ou même jamais. Rubin [1976] distingue trois types de données manquantes :

- MCAR (*Missing Completely At Random*) : la probabilité qu'une observation soit manquante ne dépend pas des données, c'est-à-dire le fait de ne pas avoir de valeur pour une variable Y est indépendant des autres variables X.
- MAR (*Missing At Random*) : la probabilité qu'une observation soit manquante ne dépend pas de la valeur qu'elle prend, c'est-à-dire le fait de ne pas avoir une valeur pour une variable Y dépend uniquement d'autres variables X observées.
- NMAR (*Non Missing At Random*) : la probabilité qu'une observation soit manquante dépend de la valeur qu'elle prend (non observée), c'est-à-dire le fait de ne pas avoir la valeur pour une variable Y est dépendant de la valeur non observée de celle-ci.

Dans notre jeu de données, seuls les profils phylogénétiques sont disponibles, ainsi les valeurs des variables associées aux espèces ancestrales sont toujours manquantes, on dispose donc de données MCAR.

Dans le cas des données MCAR (ou MAR), l'estimation des paramètres peut être effectuée de différentes manières. Une première approche possible et la plus simple consiste à estimer les paramètres à partir de l'ensemble des données complètement observées. Dans le cadre des réseaux Bayésiens, il suffit d'utiliser tous les exemples où X_i et $pa(X_i)$ sont complètement mesurés pour l'estimation de $\mathbf{P}(X_i \mid pa(X_i))$. Cependant, lorsque le nombre de variables est élevé, il devient difficile d'avoir suffisamment d'exemples pour chaque variable et pour que la qualité de l'estimation soit bonne. Une deuxième approche consiste à remplir les données manquantes dans un premier temps puis à réaliser l'estimation des paramètres sur le jeu de données complet créé. Notre choix s'est porté sur l'algorithme *Espérance-Maximisation* (EM) [Dempster *et al.*, 1977]

qui est implémenté dans le contexte des réseaux Bayésiens [Heckerman, 1999; Spiegelhalter et Lauritzen, 1990]. Cette méthode est plus efficace en terme de coût de calcul que la plupart des algorithmes existants, et la vitesse de convergence de cet algorithme est plus rapide que celle du *Gibbs Sampling* [Geman et Geman, 1984] pour lequel la vitesse diminue avec la croissance des données manquantes.

Cet algorithme, décrit ci-dessous, part de valeurs de paramètres aléatoires pour estimer les données manquantes (étape Espérance) à partir desquelles de nouveaux paramètres seront calculés (étape Maximisation). Ces deux étapes sont réitérées jusqu'à ce que la vraisemblance des paramètres converge.

Définition 2.2. *Soit $\mathcal{B} = (\mathcal{G},\theta)$, un réseau Bayésien défini par une structure $\mathcal{G}$ et un ensemble de paramètres θ. Soit $\mathcal{D}$ un jeu de données.*

$\forall d \in \mathcal{D}$, $\mathbf{P}(d \mid \mathcal{B}) = \mathcal{L}(\mathcal{B} \mid d)$ *est la vraisemblance de $\mathcal{B}$ sachant d*

Si les d sont indépendants alors

$$\mathcal{L}(\mathcal{B} \mid \mathcal{D}) = \prod_{d \in \mathcal{D}} \mathbf{P}(d \mid \mathcal{B})$$

On utilise couramment la log-vraisemblance :

$$\mathcal{LL}(\mathcal{B} \mid \mathcal{D}) = \sum_{d \in \mathcal{D}} log(\mathbf{P}(d \mid \mathcal{B}))$$

Dans l'étape de maximisation, on utilise l'estimation au maximum a posteriori (MAP) au lieu de l'estimation au maximum de vraisemblance. Cette estimation donne de meilleurs résultats lorsque certaines observations n'existent pas ou sont très rares dans le jeu d'apprentissage, ce qui résulte en une observation impossible ($N_{ijk}^{\theta^t} = 0$). La différence réside dans l'ajout d'un a priori α_{ijk} pour chacun des paramètres qui peuvent être interprétés comme des pseudo comptages. Nous avons utilisé l'équivalent Bayésien de Dirichlet avec des a priori uniformes (appelé BDeu) [Buntine, 1991; Heckerman *et al.*, 1995] dont les valeurs par défaut sont données par $\alpha_{ijk} = \frac{1}{|X_i|*|pa(X_i)|} = \frac{1}{4}$. Par exemple, l'estimation de $\hat{\theta}_{i00}^{t+1}$ est :

$$\hat{\theta}_{i00}^{t+1} = \frac{N_{i00}^{\theta^t} + 0{,}25}{N_{i00}^{\theta^t} + N_{i01}^{\theta^t} + 0{,}5}$$

L'étape d'espérance utilise les paramètres courants pour inférer les états manquants. Par conséquent, le jeu de paramètre doit être initialisé au début de l'algorithme, généralement de manière aléatoire. L'algorithme EM se termine lorsque la vraisemblance des paramètres n'augmente plus. Cependant, l'algorithme ne garantit pas d'obtenir le maximum global lorsque l'on part d'une initialisation aléatoire, il est donc nécessaire de réaliser plusieurs estimations avec des initialisations différentes.

Algorithme 1 Algorithme EM-MAP pour les réseaux Bayésiens

Soit $\mathcal{B}$ le réseau Bayésien sur les variables $\mathcal{U} = \{X_1,X_2,\ldots,X_n\}$.
Soit θ_{ijk} la probabilité de X_i d'être dans l'état j sachant que la configuration de ses parents est k, c'est-à-dire la probabilité conditionnelle $\mathbf{P}(X_i = j \mid pa(X_i) = k)$.
Une estimation au *maximum a posteriori* (MAP) $\hat{\theta}_{ijk}$, du paramètre θ_{ijk} sachant le jeu de données $\mathcal{D} = d_1, ..., d_m$ est calculée de la manière suivante :

Choisir $\epsilon > 0$ comme seuil du critère d'arrêt
Soit $t = 0$
Soit θ^0 une initialisation aléatoire de l'ensemble des paramètres θ.
Pour tout θ_{ijk}, associer un a priori α_{ijk}
Répéter
Étape E : Pour chaque $i \in$[1 , n], calculer la table des effectifs attendus en fonction des paramètres courants :

$$\begin{aligned} N_{ijk}^{\theta^t} &= \mathbb{E}\left[N(X_i = j, pa(X_i) = k) \mid \mathcal{D}\right] \\ &= \sum_{s=1}^{m} \mathbf{P}(X_i = j, pa(X_i) = k \mid d_s, \theta^t) \end{aligned}$$

avec $N(X_i = j, pa(X_i) = k)$ le nombre de cas où $X_i = j$ et $pa(X_i) = k$

Étape M : Utiliser les effectifs attendus comme s'ils étaient les effectifs réels pour estimer chaque $\hat{\theta}_{ijk}$ au *maximum a posteriori* :

$$\hat{\theta}_{ijk}^{t+1} = \frac{N_{ijk}^{\theta^t} + \alpha_{ijk}}{\sum_{k'} \left(N_{ijk'}^{\theta^t} + \alpha_{ijk'}\right)} \tag{2.2}$$

L'estimation au *maximum de vraisemblance* étant :

$$\hat{\theta}_{ijk}^{t+1} = \frac{N_{ijk}^{\theta^t}}{\sum_{k'} N_{ijk'}^{\theta^t}} \tag{2.3}$$

$t = t + 1$
Jusqu'à $|log(\mathbf{P}(\mathcal{D} \mid \theta^{t+1})) - log(\mathbf{P}(\mathcal{D} \mid \theta^t))| \leq \epsilon$

2.1.3 Méthode d'inférence des scénarios

Les relations de cause à effet entre les variables définies dans un réseau Bayésien sont probabilisées. Ainsi l'observation d'une ou plusieurs causes n'entraîne pas systématiquement l'effet ou les effets qui en dépendent, mais modifie la probabilité de les observer. Cette actualisation des probabilités en fonction des connaissances introduites dans le modèle permet de réaliser des inférences. Celle-ci consiste à propager une ou plusieurs informations (états connus pour certaines variables) à travers le réseau pour en déduire comment sont modifiées les probabilités des états des autres variables. La distribution de probabilité est modifiée selon le théorème 2.2.

Théorème 2.2. *Soit le réseau Bayésien $\mathcal{B}$ sur les variables $\mathcal{U} = \{X_1,X_2,\ldots,X_n\}$ et soient $\mathcal{E} =$*

$\{X_{obs}\}$ les observations (connaissances des états de certaines variables). Alors :

$$\mathbf{P}(\mathcal{U},\mathcal{E}) = \prod_{i=1}^{n} \mathbf{P}(X_i \mid pa(X_i)) \prod_{X \in \mathcal{E}} \mathbf{P}(X)$$

Et pour tout $X_i \in \mathcal{U}$, on a

$$\mathbf{P}(X_i \mid \mathcal{E}) = \frac{\sum_{\mathcal{U}\setminus\{X_i\}} \mathbf{P}(\mathcal{U},\mathcal{E})}{\mathbf{P}(\mathcal{E})} \tag{2.4}$$

De nombreux algorithmes sont disponibles pour réaliser l'inférence dont les principales différences viennent de leur compromis entre vitesse, complexité, généralité (structure du réseau particulière ou pas) et précision (exact ou approché). Nous avons utilisé l'algorithme de l'arbre de jonction [Jensen *et al.*, 1990; Lauritzen et Spiegelhalter, 1988] qui est un algorithme exact et généraliste.

L'inférence réalisée a pour objectif d'obtenir deux types d'informations pour chaque famille : les états ancestraux les plus probables et les probabilités marginales associées aux inférences.

- *Le scénario le plus probable :* l'inférence assigne à chaque variable non connue l'état le plus probable sachant le profil. Elle cherche à maximiser la vraisemblance totale du scénario : on a donc une optimisation globale de l'arbre.
  Définition 2.3. *Le **scénario le plus probable** correspond à l'affectation complète x de $X \in \mathcal{U}\setminus\mathcal{E}$ avec $\mathcal{E} = \{X_{observées}\}$ pour lesquelles $\mathbf{P}(X = x \mid \mathcal{E} = e)$ est maximal, avec e le profil phylogénétique correspondant.*
- *Les probabilités marginales :* on peut également calculer les probabilités marginales de chaque variable dont l'état est inconnu, sommées sur tous les scénarios possibles. Elles peuvent être interprétées comme le soutien de l'inférence de chaque état par le modèle.
  Définition 2.4. *La **probabilité marginale** de $X_i \in \mathcal{U}\setminus\mathcal{E}$ est $\mathbf{P}(X_i \mid \mathcal{E} = e)$, obtenue à l'aide de l'équation 2.4.*

2.2 Application et implémentation

2.2.1 Utilisation de BNT avec Matlab

De nombreuses boîtes à outils ont été développées pour utiliser les réseaux Bayésiens dans différents langages de programmation (C++, R, Matlab, Java), une liste assez exhaustive est maintenue par K.P. Murphy [Murphy, 2005]. Nous avons utilisé BNT (Bayesian Network Toolkit version 1.0.4)[Murphy, 2001], qui est une boîte à outils développée pour Matlab (version R14, The MathWorks,Inc.).

2.2.2 Données utilisées pour valider le modèle

Nous avons travaillé sur un jeu de données de 170 espèces complètement séquencées (voir annexe C, page 137) comprenant 9 eucaryotes, 19 archées et 142 bactéries. Le sous-ensemble de familles de protéines extraites d'HOGENOM version 3 [Penel *et al.*, 2009] et restreintes aux 170 espèces a été utilisé pour valider le modèle. Il contient 194 844 familles dont 43 452 sont présentes dans au moins deux espèces différentes. Les profils phylogénétiques représentent la présence et l'absence de ces familles dans les espèces considérées.

2.2.3 Modèle utilisé pour les scénarios d'évolution.

La structure du réseau Bayésien est représentée par l'arbre phylogénétique des espèces analysées [1]. Chaque nœud interne de l'arbre représente une espèce ancestrale et chaque feuille représente les espèces contemporaines sélectionnées. L'arbre des espèces utilisé peut ne pas être complètement résolu (chaque espèce ancestrale peut avoir plus de 2 espèces filles), il contient 269 nœuds. Les variables aléatoires de $\mathcal{U}$ modélisent la présence ou l'absence des familles dans chacune des espèces. Ce sont donc des variables aléatoires discrètes à deux états que nous modélisons par 0 pour l'absence et 1 pour la présence.

Cette structure d'arbre indique que chaque variable a un unique parent, excepté la racine qui n'en a pas. Ainsi les distributions de probabilités conditionnelles peuvent être représentées simplement par une table conditionnelle (figure 2.1a) qui spécifie la probabilité de chaque état de la variable X_i en fonction de son parent, $\mathbf{P}(X_i \mid pa(X_i))$. La table permet de définir deux paramètres libres : le gain et la perte. La racine de l'arbre n'ayant pas de parent, la variable est définie par un seul paramètre : la probabilité de présence. Dans notre approche, les probabilités de gains et de pertes peuvent varier indépendamment le long de l'arbre. Ainsi, le nombre de paramètres du modèle est fonction du nombre n de nœuds de l'arbre : $2n - 1$.

2.2.4 Estimation des paramètres

Pour réaliser l'estimation des paramètres avec l'algorithme EM, il faut définir un jeu d'apprentissage, initialiser les paramètres du modèle et choisir un critère d'arrêt. Les paramètres du modèle sont initialisés de manière aléatoire dans les intervalles $[10^{-4}, 5.10^{-3}]$ pour les probabilités de gain et $[10^{-2}, 0{,}5]$ pour les probabilités de perte. Ils correspondent à une gamme de valeurs réalistes : de l'ordre de 20 à 1 000 gains et entre 1% et 50% des protéines parentales perdues. Les différentes itérations de l'algorithme EM leur permettent de dévier de cette valeur initiale. L'algorithme s'arrête

1. L'arbre des espèces utilisé dans ce chapitre est celui extrait du NCBI et est présenté en annexe C page 140.

lorsque la vraisemblance ($\mathcal{LL}$, LogLikelihood) a convergé, avec pour critère d'arrêt :

$$\frac{\mid \mathcal{LL}_t - \mathcal{LL}_{t-1} \mid}{\frac{1}{2}(\mid \mathcal{LL}_t \mid + \mid \mathcal{LL}_{t-1} \mid)} < 10^{-5}$$

Le jeu d'apprentissage est extrait de manière aléatoire du jeu de données complet et contient 10 000 familles. En règle général, on choisit un jeu d'apprentissage indépendant des données pour lesquelles on veut faire de l'inférence. Cependant, comme aucune donnée n'est accessible pour les espèces ancestrales, il ne sera pas possible de valider les paramètres obtenus sur un jeu de données test. De plus, l'objectif est d'obtenir des paramètres qui correspondent le mieux au jeu de données, il est donc préférable d'en utiliser un sous-ensemble, voire l'ensemble des données lorsque les coûts de calcul le permettent.

Plusieurs estimations de paramètres avec différentes initialisations des réseaux Bayésiens ont été réalisées. Elles sont indépendantes les unes des autres, nous avons donc réalisé les différentes estimations en parallèle. Les calculs ont été réalisés sur la grille de calcul du centre de l'IN2P3 [1]. Trois types de processeurs sont disponibles : Intel Xeon 5345 Quad core, 2,33GHz ; Intel Xeon Single core, 28GHz et AMD Opteron 2,0GHz avec pour chacun entre 2 et 16 GB de RAM. Chaque itération (étapes E et M) dure en moyenne 1h30 avec 10 000 familles dans le jeu d'apprentissage. Le temps d'exécution dépend essentiellement de la topologie du réseau utilisée dans l'étape E, qui détermine la distribution de probabilité jointe. Cette étape est la plus couteuse puisqu'elle infère les scénarios des familles du jeu d'apprentissage à partir des paramètres courants. Quant à l'étape M, les paramètres sont mis à jour grâce à l'équation 2.2. La mémoire nécessaire est de 250MB et elle dépend essentiellement du nombre de familles dans le jeu d'apprentissage.

2.2.5 Inférence des scénarios

Le jeu de données total contient 10 575 profils phylogénétiques différents. L'inférence du scénario de probabilité maximale ainsi que des probabilités marginales associées aux états inférés a donc été réalisée sur chacun d'eux (figure 2.1c). Le temps d'inférence est de 2h50 et la mémoire utilisée est d'environ 555 MB. Il est possible de paralléliser les inférences par les données puisque chaque profil est indépendant des autres.

1. Centre de calcul de l'Institut national de physique nucléaire et de physique des particules (http://cc.in2p3.fr)

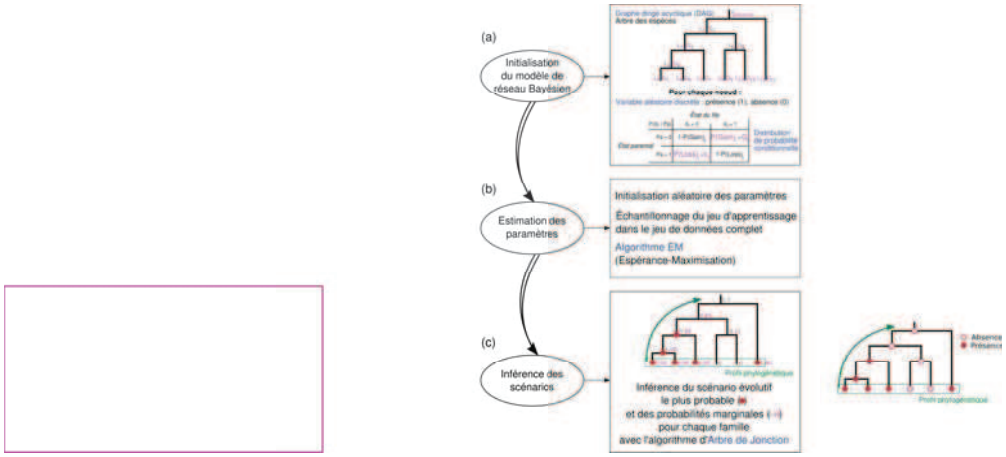


FIGURE 2.1 – **Étapes de la création du modèle de réseau Bayésien.** (**a**) Le modèle est défini par l'arbre des espèces et des probabilités de gains et de pertes qui définissent les paramètres du modèle. (**b**) L'estimation des paramètres est réalisée à partir de réseaux initialisés aléatoirement et d'un jeu d'apprentissage extrait du jeu de données complet avec l'algorithme EM. (**c**) Les profils phylogénétiques sont propagés dans les espèces ancestrales en fonction du modèle avec l'algorithme d'arbre de jonction. Dans l'exemple schématisé, les ronds rouges représentent la présence de la famille de protéines inférée au nœud correspondant et la probabilité marginale associée à la présence est indiquée à chaque nœud.

2.3 Validation de l'estimation des paramètres

Les 537 paramètres du modèle sont estimés avec l'algorithme EM, qui est le plus efficace dans le cas de données incomplètes. Cet algorithme garantit la convergence de la vraisemblance vers un maximum local qui peut dépendre de la valeur initiale des paramètres ce qui constitue une source de variabilité dans l'estimation des paramètres. Il existe une deuxième source de variabilité provenant du jeu d'apprentissage. En effet, le jeu de données total est trop important pour être utilisé entièrement, ainsi seul un sous-ensemble de ces données est utilisé pour l'apprentissage. Il est donc important d'évaluer la stabilité de l'estimation des paramètres et la robustesse des scénarios d'évolution qui en dérivent.

Les probabilités de gain et de perte pourraient théoriquement varier dans l'intervalle $]0\,,1[$. Cependant, d'un point de vue biologique les nombres de gains et de pertes attendus se situent dans une fourchette qui va d'une dizaine à quelques milliers d'événements. En conséquence, nous avons restreint les intervalles pour l'initialisation des paramètres dans l'algorithme EM : $[10^{-4}\,, 5.10^{-3}]$ pour les probabilités de gain et $[10^{-2}\,, 0{,}5]$ pour les probabilités de perte.

Des analyses préliminaires, dans lesquelles l'algorithme EM effectuait une estimation au maximum de vraisemblance (équation 2.3) ont été effectuées. Ces analyses ont montré que les vraisemblances obtenues lorsqu'on échantillonne les valeurs initiales des paramètres dans l'intervalle complet sont plus dispersées (coefficient de variation CV=7,7%) que lorsqu'on les échantillonne sur l'intervalle restreint (CV=$6{,}9.10^{-4}$). De plus, la vraisemblance maximale obtenue est nettement supérieure avec l'intervalle restreint ($\mathcal{LL}$ $\simeq$-80 670 contre $\mathcal{LL}$ $\simeq$-110 400 en moyenne sur 20 estimations).

L'utilisation des a priori de Dirichlet dans l'algorithme EM (EM-MAP, équation 2.2) permet de diminuer l'impact des événements rares ou non observés pouvant induire une probabilité nulle. Nous avons échantillonné 200 initialisations avec lesquelles nous avons estimé les paramètres à partir du même jeu d'apprentissage en utilisant la méthode de Dirichlet. Les 200 vraisemblances convergent vers la même valeur maximale ($\mathcal{LL}$ $\simeq$ -72 278) avec une dispersion de $1{,}6.10^{-5}$. Les valeurs de paramètres obtenues sont globalement très similaires en chacun des nœuds de l'arbre. En effet, leur dispersion est en moyenne de $2{,}2\%$ pour les probabilités de gain et de $4{,}6\%$ pour les probabilités de perte. La distribution des dispersions (voir figure 2.2) montre que seuls quelques paramètres présentent une dispersion importante (CV > 25%) : 5 probabilités de gain et 10 probabilités de perte.

Ces résultats sont indépendants du jeu d'apprentissage utilisé. En effet, cette analyse a été effectuée avec 5 jeux d'apprentissages différents et les résultats sont identiques (voir tableau 2.1) : les vraisemblances maximales obtenues sont voisines d'un jeu d'apprentissage à l'autre et la dispersion des paramètres reste faible. Cela permet de conclure que l'initialisation des paramètres a

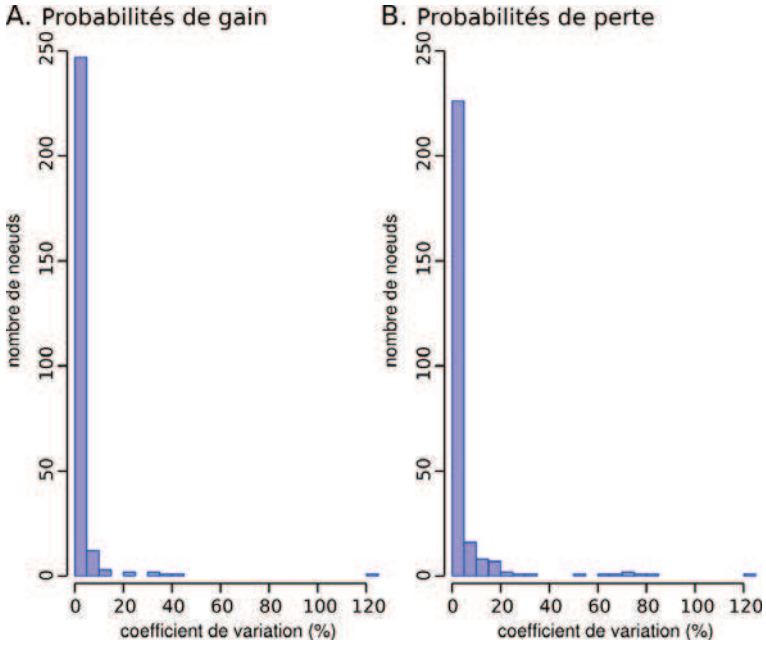


FIGURE 2.2 – **Distributions de la dispersion des estimations des paramètres avec différentes initialisations.** Le CV a été calculé à partir des 200 estimations réalisées à partir de différentes initialisations pour les probabilités de gain (**A**) et les probabilités de perte (**B**) avec le même jeu d'apprentissage.

un impact faible sur les valeurs de paramètres obtenues.

Jeux	Vraisemblance		Probabilités de gain		Probabilités de perte	
d'apprentissage	Valeur	CV $\times 10^{-5}$	CV moyen	Nb > 25%	CV moyen	Nb > 25%
1	-72 279	1,6	2,2%	5	4,6%	10
2	-72 919	2,6	3,0%	10	7,4%	27
3	-71 439	1,0	1,1%	2	1,6%	5
4	-71 815	1,4	0,96%	2	1,8%	7
5	-71 726	1,1	0,74%	1	1,5%	2

TABLEAU 2.1 – **Dispersion des estimations avec 5 jeux d'apprentissage différents**. Pour chaque jeu d'apprentissage, 200 estimations des probabilités de gain et de perte ont été réalisées à partir de différentes initialisations. La vraisemblance maximale obtenue sur les 200 estimations est donnée ainsi que la dispersion des valeurs. Pour les probabilités de gain et de perte, le CV moyen correspond à la moyenne des dispersions calculées pour chaque probabilité de gain et de perte. Le nombre de nœuds pour lesquels la dispersion de l'estimation est supérieure à 25% est également donné.

Les nœuds les plus variables ne sont en général pas les mêmes selon le jeu d'apprentissage utilisé. Cela nous amène à penser que celui-ci pourrait être une source de variabilité moins négligeable que les paramètres initiaux. Les jeux d'apprentissage utilisés sont un échantillon aléatoire de 10 000 familles, soit environ 5% du jeu de données global. Des analyses préliminaires réalisées sur le sous-arbre des Gammaprotéobactéries ont montré que passer de 5% à 10% du jeu de

données globale pour le jeu d'apprentissage réduit significativement l'écart-type des estimations [1] mais la moyenne des estimations reste la même [2]. Ainsi augmenter la taille du jeu d'apprentissage permet de diminuer la variance des estimations, mais n'a pas d'impact sur leur moyenne. Le jeu d'apprentissage de 5% représente un bon compromis entre la qualité des estimations et les coûts de calcul engendrés par la taille du jeu d'apprentissage.

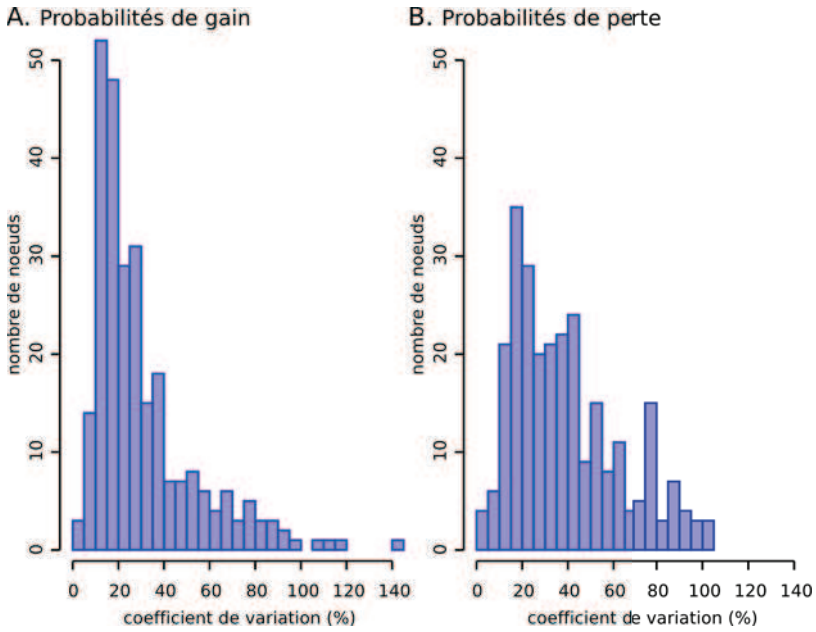


FIGURE 2.3 – **Distributions de la dispersion des estimations des paramètres avec différents jeux d'apprentissage.** Le CV a été calculé sur 100 estimations réalisées à partir de différents jeux d'apprentissage pour les probabilités de gain (**A**) et les probabilités de perte (**B**).

Nous avons comparé les paramètres estimés avec 100 jeux d'apprentissage. Pour chacun d'eux, 200 initialisations ont été effectuées. L'estimation avec la meilleure vraisemblance après 6 itérations est gardée pour mener l'estimation jusqu'à la convergence de la vraisemblance. Ainsi une seule estimation est faite par jeu d'apprentissage. La vraisemblance maximale obtenue est très peu variable d'un jeu d'apprentissage à l'autre, en effet, la dispersion est de seulement $0{,}91\%$ ($\mathcal{LL} \in$ [-73 270 , -70 130]). Pour chaque probabilité de gain et de perte, la distribution des estimations a été analysée (figure 2.3). La variabilité des estimations est plus importante en fonction du jeu d'apprentissage utilisé. En effet, la dispersion est en moyenne de 30% pour les probabilités de gain et de 39% pour les probabilités de perte. Les nœuds les plus variables se retrouvent dans les sous-arbres dont les espèces contemporaines ont de petits répertoires de protéines. Parmi les 20 probabilités de gain les plus variables, on retrouve les espèces ancestrales telles que *Helicobacter*, *Rickettsia*-

1. Test de Wilcoxon sur données appariées : comparaison des écarts-types des estimations des probabilités de perte et de gain à chaque nœud calculés avec deux jeux d'apprentissage de taille différente (sur 20 initialisations initiales différentes). P-value = 1.10^{-3} pour la perte et p-value = 7.10^{-6} pour le gain.

2. Test de Wilcoxon sur données appariées : comparaison des moyennes des estimations des probabilités de perte et de gain à chaque nœud calculés avec deux jeux d'apprentissage de taille différente (sur 20 initialisations initiales différentes). P-value = $0{,}36$ pour la perte et p-value = $0{,}22$ pour le gain.

ceae, *Neisseriaceae*, *Chlamydia*, *Chlamydophila* ou *Spirochaetales*. Et parmi les 20 probabilités de perte les plus variables, 14 sont des espèces actuelles comme les deux souches d'*Helicobacter pilori*, *Rickettsia conorii*, les deux souches de *Neisseiria meningitidis*, *Chlamydia muridarum* et *Chl. trachomatis*, *Chlamydophila caviae* et *Chl. pneumoniae* ou encore les 2 souches de *Leptospira interrogans* [1]. Ces observations peuvent s'expliquer par la rareté des familles informatives disponibles pour estimer ces paramètres. Ainsi, en fonction du jeu d'apprentissage, l'information disponible peut être limitante pour estimer précisément les probabilités.

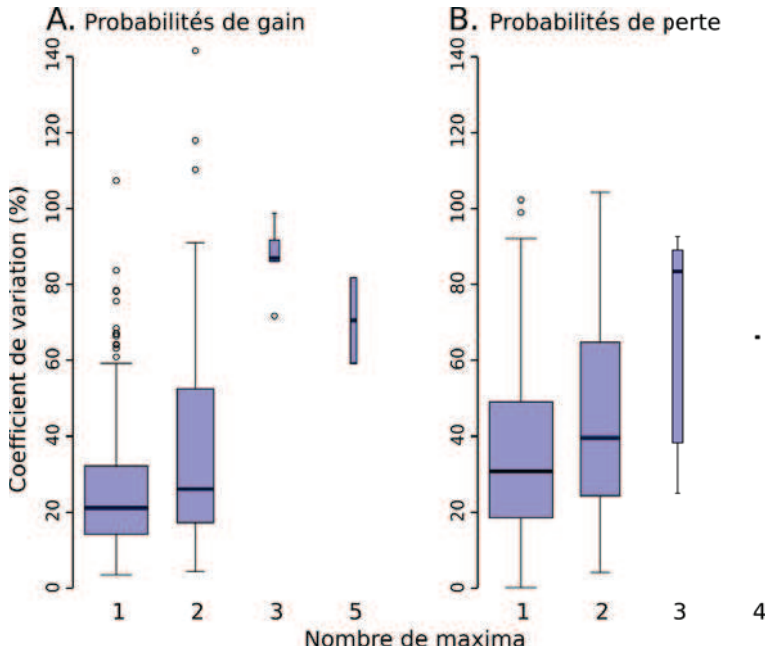


FIGURE 2.4 – **Corrélation entre la dispersion des estimations et le nombre de maxima de la distribution correspondante.** Les distributions de gain (**A**) et de perte (**B**) ont été créées à partir des 100 estimations de paramètres. La dispersion est mesurée à l'aide du coefficient de variation et le nombre de maxima comprend le maximum global et l'ensemble des maxima locaux, dont la densité est > 10% de celle du maximum global. La largeur des boîtes est proportionnelle à la racine carrée du nombre de paramètres dans chaque catégorie.

Il existe également une deuxième raison qui peut expliquer la variabilité due aux jeux d'apprentissage : il s'agit de l'information biologique portée par les familles. Les familles informatives peuvent être suffisamment nombreuses, mais en partie contradictoires. On peut imaginer par exemple qu'une certaine proportion de familles ont un scénario soutenant le gain au nœud X et qu'une autre proportion de familles induit un gain dans certains enfants de X plutôt qu'au nœud X lui-même. Ainsi, selon l'échantillonnage, une catégorie de familles ou l'autre pourrait être aléatoirement enrichie, ce qui influerait sur les probabilités de gain. Si cette hypothèse est exacte, les distributions de gain et de perte devraient présenter plusieurs maxima locaux. Ceux-ci ont été dénombrés lorsque leur densité était supérieure à 10% de celle du maximum globale de la distribution. Le nombre de maxima locaux détecté dans les distributions des probabilités est fortement

1. La liste des espèces est disponible dans l'annexe C page 137.

corrélé à la dispersion de ces distributions : le coefficient de corrélation est de $0{,}56$ pour le gain et $0{,}45$ pour la perte (figure 2.4). La présence de maxima locaux est relativement fréquente puisque 27% des probabilités de gain et 29% des probabilités de perte estimées présentent au moins un maximum local en plus du maximum global dans la distribution obtenue.

L'estimation des paramètres est une étape fondamentale dans la création du modèle puisque ces paramètres représentent une première estimation de la dynamique des génomes en ce qui concerne les gains et les pertes de familles. L'algorithme EM repose sur des initialisations aléatoires des paramètres et un échantillonnage du jeu d'apprentissage. Si les résultats montrent que l'initialisation des paramètres n'influence pas les paramètres estimés, le contenu du jeu d'apprentissage peut quant à lui avoir une incidence non négligeable sur certains paramètres notamment dans les sous-arbres où les effectifs sont petits. Ainsi certains paramètres peuvent avoir des valeurs très différentes en fonction du jeu d'apprentissage utilisé bien que la vraisemblance maximale obtenue soit similaire d'un jeu de données à l'autre. Ces paramètres pour lesquels une valeur unique ne peut pas être attribuée à travers l'estimation sont des paramètres non identifiables. Si l'hypothèse des événements rares semble expliquer une partie de la variabilité observée notamment pour les espèces ayant de petits protéomes, l'hypothèse d'une information contradictoire portée par les familles semble avoir un impact plus fort. L'augmentation de la taille du jeu d'apprentissage diminue la variance des estimations, cependant le compromis entre la précision des paramètres et les ressources de calcul disponibles ne permet d'utiliser qu'un sous-ensemble des familles disponibles.

2.4 Robustesse des scénarios d'évolution

La variabilité de certains paramètres estimés, discutée dans la section précédente, pose la question de la robustesse des scénarios d'évolution inférés à partir de ces paramètres. Pour mesurer l'influence de ces variations de paramètres sur les scénarios inférés, nous avons comparé les scénarios d'évolution inférés avec 2 jeux d'apprentissage différents que nous nommerons M_{full1}, M_{full2}. Les scénarios d'évolution ont été inférés pour les 43 452 familles de protéines présentes dans au moins deux espèces contemporaines.

La comparaison des patrons de présence / absence dans l'arbre pour chaque famille montre que 92% des scénarios sont identiques entre les 2 modèles. Les 3 485 scénarios différents présentent en moyenne 1,9 état différent, mais plus de 60% n'ont qu'un état différent. L'analyse des probabilités marginales des états différents entre les deux modèles montre que moins de 4% de ces différences sont soutenues dans les deux modèles à la fois (avec une probabilité $\geq 0{,}95$), et moins de 10% sont soutenues dans au moins l'un des modèles. Le contenu des espèces ancestrales est peu modifié par ces différences (figure 2.5, parties bleues). En effet, 96% des présences inférées et soutenues sont communes aux deux modèles, $1{,}5\%$ sont soutenues seulement avec M_{full1} et $2{,}7\%$ seulement avec

M_{full2}. La plupart des différences se situent dans les espèces les plus anciennes.

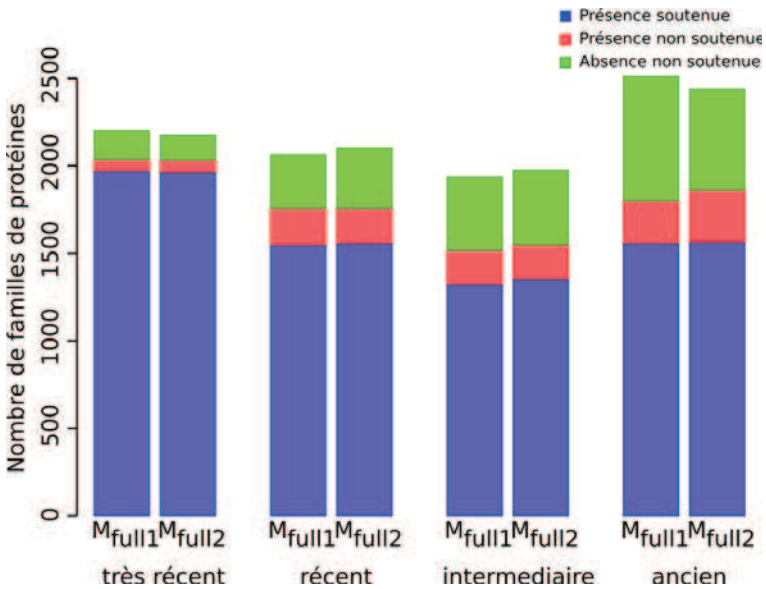


FIGURE 2.5 – **Contenu moyen en familles de protéines des espèces ancestrales inférées à partir de 2 jeux d'apprentissage différents.** Les diagrammes comparent les valeurs moyennes du nombre de familles inférées dans les répertoires de différentes catégories d'espèces ancestrales avec les modèles M_{full1} et M_{full2}. Les espèces ancestrales sont regroupées en fonction du nombre d'espèces contemporaines présentes dans leur sous-arbre : *très récentes* (2 espèces), *récentes* (3-4 espèces), *intermédiaires* (5-9) ou *anciennes* (plus de 10 espèces). Chaque diagramme est divisé en 3 catégories de familles : les familles présentes et soutenues par le modèle (la probabilité marginale est $\geq 0{,}95$), les familles présentes mais non soutenues et enfin les familles prédites absentes et non soutenues par le modèle, respectivement représentées par les sections bleues, rouges et vertes.

Les familles dont les états ne sont pas soutenus ont été dénombrées dans chaque espèce ancestrale (la figure 2.5). Le patron obtenu est assez clair, les nœuds présentant le plus de familles non soutenues sont LUCA (Last Universal Common Ancestor), et les espèces ancestrales des groupes protéobactériens, bactériens et eucaryotes comme *Alpha-Beta-Gamma-Delta/Epsilon proteobacteria*, *Actinobacteridae*, *Bacilli*, *Bacteroidetes*, *Deinococci*, *Spirochaetales*, *Bilateria* ou *Coelomata* (retrouvés essentiellement dans la catégorie de nœud *ancien*). Les ancêtres bactériens correspondent à de grandes polytomies où la présence d'une protéine dans l'un des sous-arbres peut influencer sa présence dans les espèces sœurs pour lesquelles le soutien diminue. Cela explique la plus forte fréquence des absences non soutenues. Au contraire, les nœuds dans lesquels le soutien est le meilleur correspondent aux ancêtres les plus récents. Ce sont généralement les ancêtres directs des espèces contemporaines qui ont un contenu en familles assez petit comme *Rickettsia*, *Buchnera aphidicola*, *Tropherima wipplei*, *Chlamydia*, *Chlamydophila* ou *Streptococcus pyogenes*. Les prédictions sont meilleures parce que ces nœuds sont proches des observations. De plus, on ne trouve aucune corrélation entre le nombre de familles pour lesquelles les états inférés sont différents à un nœud donné et la variabilité des paramètres à ce nœud [1] ou la variabilité des paramètres

1. La corrélation est de $0{,}15$ pour le gain (p-value = $0{,}012$) et de $-0{,}027$ pour la perte (p-value = $0{,}66$).

dans le parent [1]. En conclusion, ces résultats suggèrent que les scénarios inférés sont globalement robustes face aux variations de certains paramètres.

2.5 Modélisation explicite des variations de contenus en gènes

Comme on l'a vu précédemment, le modèle proposé autorise les probabilités de gain et de perte à varier le long de l'arbre des espèces. Cela génère un grand nombre de paramètres qui dépend du nombre de nœuds dans l'arbre. Cette hypothèse est justifiée d'un point de vue biologique puisque l'ensemble des résultats démontre effectivement que les fréquences de gain et de perte sont variables sur l'ensemble de l'arbre. Nous avons comparé le modèle complet à un modèle plus simple dans lequel les probabilités de gains et de pertes sont identiques le long de l'arbre, analogue de ce point de vue aux méthodes de parcimonie.

Nous avons construit deux modèles : M_{full} le modèle complet et M_0 le modèle à 3 paramètres où les probabilités de gain et de perte sont identiques le long de l'arbre, le troisième paramètre étant la probabilité de présence à la racine. Les paramètres ont été estimés à l'aide du même jeu d'apprentissage, et en utilisant seulement 20 initialisations. Les paramètres estimés obtenant la meilleure vraisemblance sont gardés pour réaliser l'inférence des scénarios d'évolution.

La vraisemblance des paramètres du modèle M_{full} (-71 946) est beaucoup plus grande que celle obtenue avec le modèle M_0 ($\mathcal{LL} \simeq$-81 767). Pour savoir si cette augmentation de vraisemblance est significative, nous avons appliqué le test LRT (Likelihood Ratio Test, Neyman et Pearson [1928]), qui permet de comparer les vraisemblances obtenues pour deux modèles emboîtés.

Likelihood Ratio Test (LRT)

> Soit $\mathcal{L}_0$, la vraisemblance maximale obtenue pour le modèle M_0.
> Soit $\mathcal{L}_{full}$, la vraisemblance maximale obtenue pour le modèle M_{full}.
> Soit k, le nombre de paramètres (libre) ajoutés dans le modèle le plus complexe (M_{full}) par rapport au modèle le plus simple (M_0).
> Sous l'hypothèse de M_0, la statistique $-2\ln\left(\frac{\mathcal{L}_0}{\mathcal{L}_{full}}\right)$ suit asymptotiquement une distribution de χ^2 à k degrés de liberté.

Si l'on applique ce test à nos données, χ^2 = 19 641 et k = 534, on obtient une P-value $< 2{,}2.10^{-16}$. Ce résultat valide statistiquement la prise en compte de paramètres variables le long de l'arbre.

La figure 2.6 présente les distributions des probabilités de gain et de perte estimées avec le modèle M_{full} et les valeurs de la probabilité de gain et de la probabilité de perte du modèle M_0.

1. La corrélation est de $-0{,}0052$ pour le gain (p-value = 0,93) et de $-0{,}053$ pour la perte (p-value = 0,38).

Cette figure montre clairement la grande amplitude des probabilités obtenues avec M_{full} que ce soit pour le gain ($[2{,}5.10^{-5}\,, 6{,}5.10^{-2}]$) ou pour la perte ($[1{,}8.10^{-3}\,, 0{,}88]$), alors que les probabilités estimées pour M_0 correspondent à la moyenne de ces distributions (respectivement $4{,}3.10^{-3}$ et $1{,}8.10^{-1}$). Cette large gamme de valeurs qui reflète l'évolution spécifique de chaque lignée ne peut donc pas être capturée avec seulement 2 paramètres.

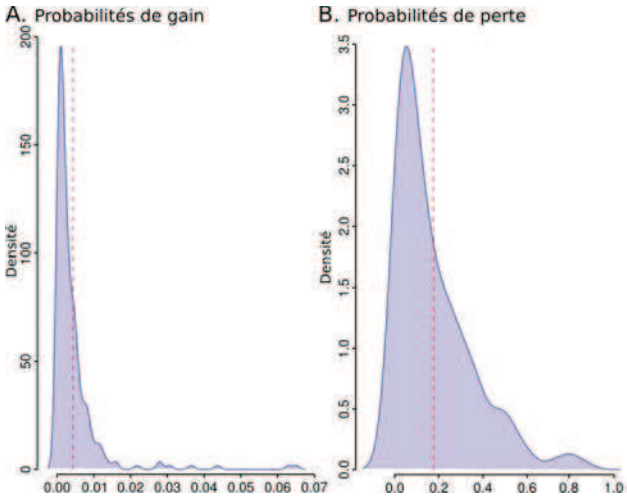


FIGURE 2.6 – **Distributions des probabilités de gain et de perte.** Les distributions des probabilités de gain (**A**) et de perte (**B**) estimées pour chaque nœud sont représentées en bleu pour le modèle M_{full} et par un ligne pointillée rouge pour le modèle M_0.

Les scénarios obtenus sont identiques dans la plupart des cas (89%) (voir tableau 2.2). Cependant, les probabilités marginales montrent en général un soutien beaucoup plus faible avec le modèle M_0. Ainsi seuls 30% des scénarios d'évolution sont complètement soutenus avec M_0 contre plus de 70% pour M_{full} et le nombre d'états spécifiquement non soutenus par M_0 est 2,5 fois plus important que ceux spécifiquement non soutenus avec M_{full}. Les scénarios différents ont en moyenne 2,7 états différents et 92% d'entre eux ne sont pas soutenus par au moins l'un des modèles. Ils ont significativement moins d'événements de gain et plus d'événements de perte avec le modèle complet (test de Wilcoxon sur données appariées, P-value ~ 0 dans les 2 cas). Cela a un impact direct sur le contenu des génomes des espèces ancestrales. En effet, le modèle complet a tendance à prédire une seule origine plus ancienne (avec des événements de pertes plus récents) alors que le modèle à 3 paramètres prédit plus de gains multiples (que l'on peut interpréter comme des transferts horizontaux). Ainsi, les contenus des génomes ancestraux sont plus importants avec le modèle complet (voir figure 2.7).

La comparaison des modèles M_{full} et M_0 peut paraître un peu triviale. En effet, ils représentent les deux modèles extrêmes que l'on peut mettre en place : tous les nœuds ont les mêmes probabilités ou bien tous les nœuds ont des probabilités différentes. Il est possible de concevoir des modèles intermédiaires où les nœuds sont regroupés en classes. Hao et Golding [2006] ont été les premiers à proposer de tels modèles sur le sous-arbre des *Bacillus*. Les différentes classes étant définies à

	Total	M_{full} et M_0	M_{full} seulement	M_0 seulement
Scénarios identiques	38 624			
↪ complètement soutenus	28 387	9 338	17 296	1 753
Scénarios différents	4 828			
↪ complètement soutenus	796	6	466	324
États inférés non soutenus	94 856	25 961	19 959	48 936

TABLEAU 2.2 – **Comparaison des 43 452 scénarios inférés avec le modèle complet M_{full} et le plus simple M_0.** Les scénarios complètement soutenus n'ont aucun état inféré avec une probabilité marginale inférieure à $0{,}95$.

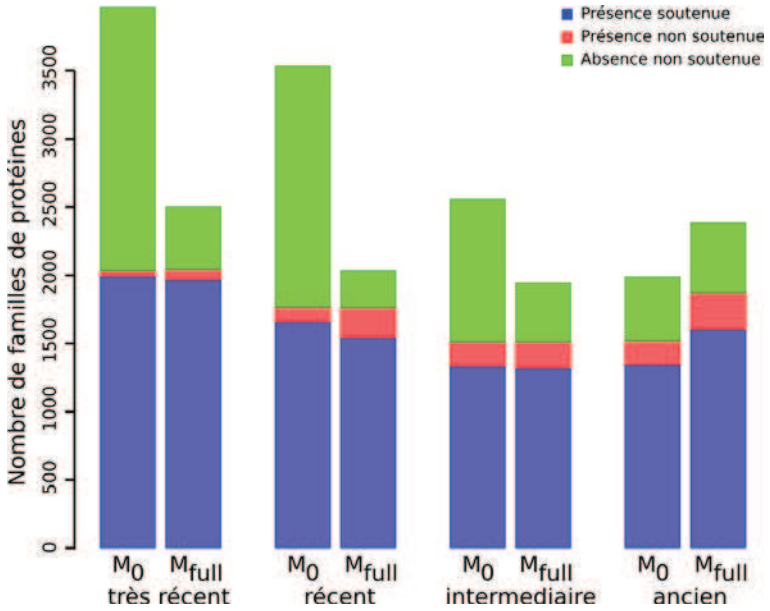


FIGURE 2.7 – **Contenu moyen en familles de protéines des espèces ancestrales inférées avec les modèles M_0 et M_{full}.** Les diagrammes comparent les valeurs moyennes des contenus dans différentes catégories d'ancêtres. Leur classification ainsi que la légende des couleurs sont les mêmes que celles de la figure 2.5.

partir des longueurs de branches et de la topologie de l'arbre (regroupement de sous-arbres). Ce choix reste globalement assez arbitraire. En effet, notre modèle prédit des probabilités de gain et de perte assez variable sur l'ensemble du sous-arbre de *Bacillus* : $[9{,}4.10^{-4}\,, 5{,}8.10^{-3}]$ pour le gain et $[3{,}5.10^{-2}\,, 2{,}5.10^{-1}]$ pour la perte. Les distributions des probabilités estimées (figure 2.6) montrent une grande amplitude des valeurs possibles, ainsi définir un nombre de classes capturant l'ensemble de cette variabilité peut être assez arbitraire.

2.6 Les réseaux Bayésiens : une méthode probabiliste plus générale que les méthodes de parcimonie

Le modèle d'évolution des familles de protéines présenté ici prend en compte des probabilités de gain et de perte différentes le long de l'arbre. Leur estimation ainsi que l'inférence des scénarios ont été réalisées avec un critère de maximum de vraisemblance. L'ensemble de ces choix est crucial pour l'inférence de scénarios d'évolution. De nombreuses méthodologies ont été développées ces dernières années, elles sont présentées dans le tableau 2.3. Les principales différences entre ces modèles viennent des paramètres considérés (gain, perte, duplication ou taux d'évolution), de leur homogénéité ou hétérogénéité sur l'arbre et entre les familles, et enfin du critère de parcimonie ou de vraisemblance utilisé.

L'hypothèse de variabilité des probabilités le long de l'arbre se base sur de nombreuses observations de réduction et d'expansion des répertoires de précédentes analyses [Blanc *et al.*, 2007; Boussau *et al.*, 2004; Kettler *et al.*, 2007; Snel *et al.*, 2002]. Les contenus des génomes sont dynamiques et sans cesse renouvelés. Ainsi, la grande variation des fréquences empiriques de gain et de perte contredit l'utilisation d'un poids relatif uniforme le long de l'arbre des espèces. La comparaison de l'ajustement aux données d'un modèle homogène et d'un modèle hétérogène montre sans ambiguïté que le modèle hétérogène est justifié par les données.

La variabilité entre lignées semble donc importante : la dynamique des gains et des pertes (expansions et réductions) des génomes s'inscrit dans l'évolution globale des espèces, dans leur constante adaptation à l'environnement. Cependant, certains auteurs ont également cherché à modéliser une variabilité entre familles de protéines. Par exemple, les gènes qui servent aux processus informationnels (réplication, transcription, traduction) sont moins sujets aux transferts horizontaux ou aux pertes massives que les gènes dits opérationnels (impliqués dans le métabolisme et les interactions avec le milieu) [Jain *et al.*, 1999]. Ainsi, Hao et Golding [2008] ont modifié leur modèle de base pour prendre en compte la variabilité entre familles de gènes. Cette modélisation se fait à l'aide d'une loi Γ qui est discrétisée pour former plusieurs classes de gènes, ce qui peut permettre d'augmenter significativement la vraisemblance (Hao et Golding [2008], mais voir également Csurös et Miklós [2009]).

Enfin, une autre méthode a été développée, prenant en compte les relations de coévolution entre familles de protéines. Tuller *et al.* [2010] reprochent aux critères de parcimonie et de vraisemblance d'obtenir parfois des résultats ambigus (scénarios différents ayant le même score ou des probabilités très proches) et de choisir de manière arbitraire. En effet, les méthodes de parcimonie utilisent différents critères pour choisir entre différents cas équiparcimonieux : Fong *et al.* [2007] choisissent la présence à la racine dans les cas ambigus, ce qui va biaiser les scénarios en faveur d'une origine ancienne ; Boussau *et al.* [2004] utilisent une méthode qui favorise les gains anciens

en favorisant les réversions par rapport aux convergences (gains multiples) ; Mirkin *et al.* [2003] choisissent le scénario qui minimise le nombre total d'événements puis le nombre de gains en cas d'égalités. Tuller *et al.* [2010] proposent de considérer la coévolution des protéines pour résoudre les ambigüités de certains scénarios.

Afin d'illustrer l'impact de ces différences de méthode sur l'inférence des scénarios d'évolution, nous choisissons de montrer ici les scénarios d'évolution des archées obtenus par trois méthodes différentes : une méthode de parcimonie [Makarova *et al.*, 2007] et deux méthodes de vraisemblance (Csurös et Miklós [2009] et la nôtre). Le modèle d'évolution développée par Csurös et Miklós [2009] est plus fin que le nôtre puisque les duplications sont prises en compte, et la variabilité de ce paramètre entre les familles de protéines est également modélisée (à l'aide d'une loi Gamma discrétisée en quatre classes). Les scénarios respectifs sont résumés dans la figure 2.8. Une comparaison rigoureuse de ces trois méthodes nécessiterait une inférence sur le même jeu de données ainsi que la même topologie de l'arbre des espèces, ce qui n'est pas le cas ici. Le nombre d'espèces considérées ainsi que la topologie de l'arbre des Archées est différent dans les trois analyses et les familles de protéines sont extraites de la base de données arCOGs [Makarova *et al.*, 2007] pour les deux autres modèles. La méthode de construction de cette base est assez différente de celle d'HOGENOM. Ainsi, le jeu de données de Csurös et Miklós [2009] contient 7 461 familles présentes dans au moins 2 espèces et 6 755 familles spécifiques d'espèces, alors que nos données sont réparties en 4 372 familles présentes dans au moins 2 espèces et 12 652 familles spécifiques d'espèces. Ainsi, 3 000 familles supplémentaires sont distribuées dans les espèces ancestrales. Nous ne comparerons donc que les résultats inférés pour LACA (Last Archaeal Common Ancestor). Notre méthode prédit 1 137 familles de protéines à LACA, alors que les méthodes de parcimonie et de vraisemblance de Csurös et Miklós [2009] en prédisent respectivement 996 et 2 050. Cependant, l'un des avantages des méthodes de vraisemblance est l'accès aux probabilités postérieures des états inférés, qui permettent de quantifier leur robustesse. Dans notre méthode 1 036 familles de LACA ($\sim$91% des familles) sont soutenues avec une probabilité $p > 0{,}90$, contre environ 1 300 par le méthode de Csurös et Miklós [2009]($\sim$63% des familles). Dans ce cas précis il apparaît que notre méthode produit des prédictions plus robustes.

En conclusion, la méthodologie décrite ici présente des propriétés intéressantes pour l'inférence des scénarios d'évolution. Le modèle d'évolution est en accord avec les résultats publiés ces dernières années dans lesquels la variabilité entre lignées est incontestable. De plus, l'utilisation d'un critère de maximum de vraisemblance permet d'ajuster les paramètres aux données sans a priori sur les résultats et d'obtenir des probabilités marginales validant localement les états ancestraux inférés. Les réseaux Bayésiens proposent une structure théorique robuste sur de grands jeux de données et parfaitement adaptée à l'inférence.

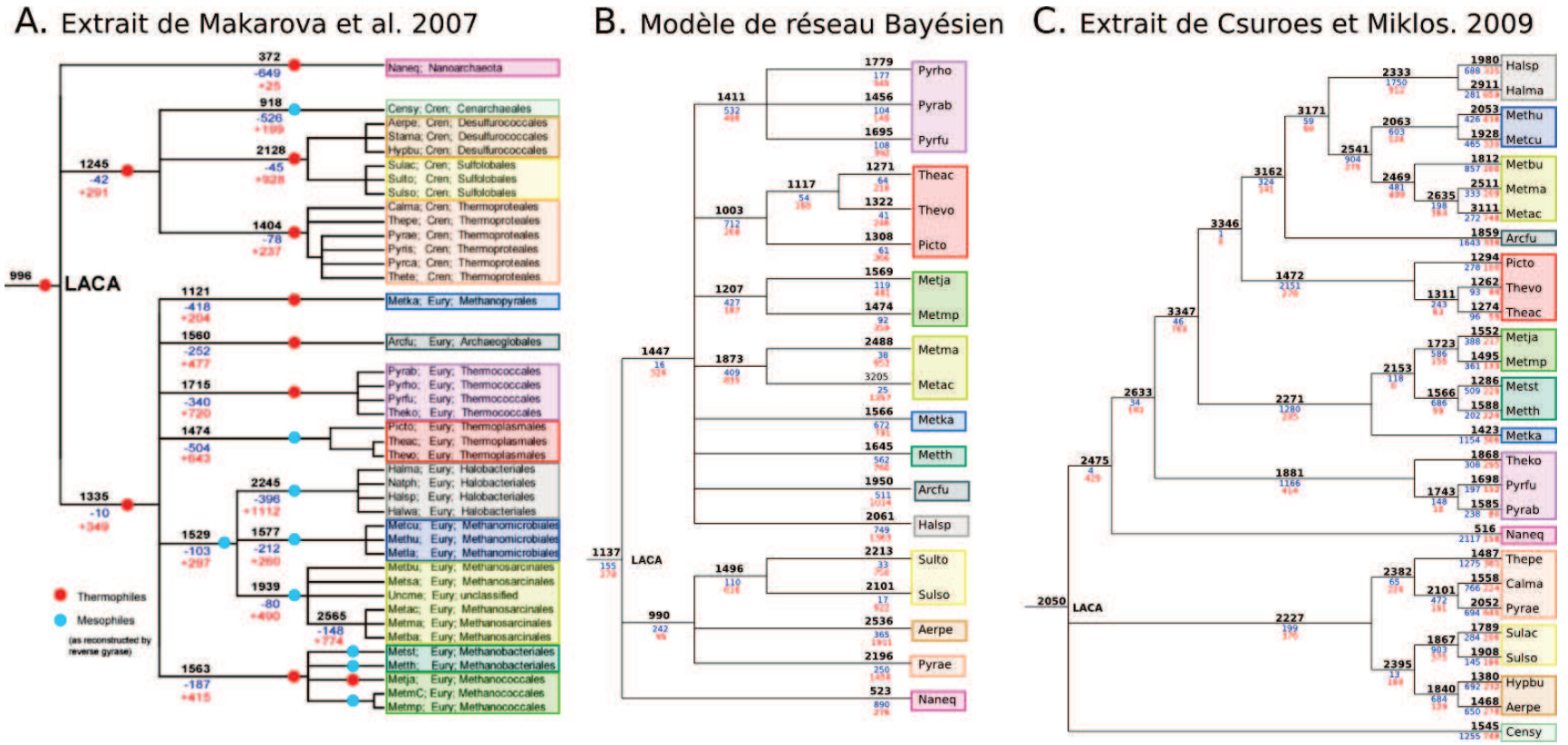


FIGURE 2.8 – **Évolution du contenu en protéines des Archées en fonction de trois méthodes d'inférence.** (**A**) Méthode de parcimonie de Makarova *et al.* [2007]. (**B**) Méthode de réseau Bayésien de cette thèse. (**C**) Méthode de vraisemblance de Csurös et Miklós [2009]. Chaque branche est étiquetée de la manière suivante : en noir, le nombre inféré de familles de protéines présentes dans le nœud auquel mène la branche ; en rouge, le nombre de familles gagnées ; en bleu, le nombre de familles perdues le long de la branche. Les boîtes colorées regroupent les espèces des différents clades archéens indiqués dans le panneau **A**. Les abréviations des noms d'espèces sont les suivantes (dans l'ordre d'apparition de l'arbre en A) : *Nanoarchaeum equitans* (Naneq), *Cenarchaeum symbiosum* (Censy), *Aeropyrum pernix* (Aerpe), *Staphylothermus marinus* (Stama), *Hyperthermus butylicus* (Hypbu), *Sulfolobus acidocaldarius* (Sulac), *Sulfolobus tokodaii* (Sulto), *Sulfolobus solfataricus* (Sulso), *Caldivirga maquilingensis* (Calma), *Thermofilum pendens* (Thepe), *Pyrobaculum aerophilum* (Pyrae), *Pyrobaculum islandicum* (Pyris), *Pyrobaculum calidifontis* (Pyrca), *Thermoproteus tenax* (Thete), *Methanopyrus kandleri* (Metka), *Archaeoglobus fulgidus* (Arcfu), *Pyrococcus abyssi* (Pyrab), *Pyrococcus horikoshii* (Pyrho), *Pyrococcus furiosus* (Pyrfu), *Thermococcus kodakaraensis* (Theko), *Picrophilus torridus* (Picto), *Thermoplasma acidophilum* (Theac), *Thermoplasma volcanium* (Thevo), *Haloarcula marismortui* (Halma), *Natronomonas pharaonis* (Natph), *Halobacterium sp.* (Halsp), *Haloquadratum walsbyi* (Halwa), *Methanoculleus marisnigri* (Metcu), *Methanospirillum hungatei* (Methu), *Methanocorpusculum labreanum* (Metla), *Methanococcoides burtonii* (Metbu), *Methanosaeta thermophila* (Metsa), *Methanosarcina acetivorans* (Metac), *Methanosarcina mazei* (Metma), *Methanosarcina barkeri fusaro* (Metba), *Methanosphaera stadtmanae* (Metst), *Methanothermobacter thermoautotrophicus* (Metth), *Methanocaldococcus jannaschii* (Metja), *Methanococcus maripaludis C5* (MetmC), *Methanococcus maripaludis S2* (Metmp).

Références	Inf.	Gain - Perte	Dupl.	Lg branche	Variation lignées	Variation familles	Particularités
Koonin *et al.* [2004]	MP	oui	non	non	non	non	Transferts horizontaux et regains refusés.
Mirkin *et al.* [2003]; Yang et Bourne [2009], Itoh *et al.* [2007]; Kettler *et al.* [2007]	MP	oui	non	non	non	non	Minimise un score.
Snel *et al.* [2002], Boussau *et al.* [2004]	MP	oui	oui	non	non	non	Minimise un score.
Dagan et Martin [2007]; Fong *et al.* [2007]; Kunin et Ouzounis [2003a,b]; Makarova *et al.* [2007], Ouzounis *et al.* [2006]	MP	non	non	non	non	non	Minimise un nombre d'événements.
Hahn *et al.* [2005]	ML	non	oui	phylo	non	non	Taux d'évolution.
Hao et Golding [2006], Marri *et al.* [2007]	ML	non	non	phylo	oui	non	Taux d'évolution et regroupement a priori des lignées.
Hao et Golding [2008]	ML	non	non	phylo	oui	9	Taux d'évolution et regroupement a priori des lignées.
Spencer *et al.* [2006]	ML	oui	oui	model	non	non	Processus de Markov continu (21 états).
Cohen *et al.* [2008], Cohen et Pupko [2010]	ML	oui	non	phylo	non	16	Processus de Markov continu (2 états).
Iwasaki et Takagi [2007]	ML	oui	oui	non	oui	non	Processus de Markov continu (4 états).
Méthode RB	ML	oui	non	non	oui	non	Réseaux Bayésiens (modèle graphique).
Didelot *et al.* [2009]	ML	oui	non	phylo	oui	non	Processus de Markov continu (2 états).
Csurös et Miklós [2009]	ML	oui	oui	model	oui	4	Processus de Markov continu.
Blanc *et al.* [2007]; Ekman *et al.* [2007]; Ogura *et al.* [2005]; Putnam *et al.* [2007]; Sakarya *et al.* [2008]	Autre	non	non	non	non	non	Pas de modèle évolutif. La répartition chez les ancêtres est fonction du contenu des espèces filles. Les relations d'orthologie/paralogie ou les pseudogènes peuvent être utilisés.
Tuller *et al.* [2010]	Autre	non	oui	non	oui	non	Taux d'évolution. La co-évolution est prise en compte.
Lagomarsino *et al.* [2009]; Molina et van Nimwegen [2008]; Qian *et al.* [2001]	Autre	non	non	non	non	oui/ non	Analyse des distributions des familles (regroupées en catégories fonctionnelles ou pas) dans les génomes.

TABLEAU 2.3 – **Récapitulatif des différentes méthodes d'inférence existantes et de leurs principales caractéristiques.** Les références notées en gras décrivent une méthode d'inférence ayant les caractéristiques présentées dans le tableau, les autres utilisent (avec ou sans modification) l'une de ces méthodes. *Inf.* : méthode d'inférence, *MP* : Maximum de parcimonie, *ML* : Maximum de vraisemblance. Les 5 colonnes suivantes précisent les paramètres du modèle comme la prise en compte du gain, de la perte, de la duplication et des longueurs de branches (estimées par une reconstruction phylogénétique, *phylo*, ou par le modèle, *model*), mais aussi l'hétérogénéité de ces paramètres entre les lignées et entre les familles. Cette dernière est prise en compte à l'aide d'une loi Gamma discrétisée dont le nombre de classes est donné le cas échéant.

Chapitre 3

L'expansion de l'univers des protéines à travers l'évolution des modules protéiques

Dans ce projet, nous avons cherché à comprendre l'apparition de nouvelles protéines en combinant les aspects évolutifs et modulaires des protéines. Pour cela, nous avons inféré des scénarios d'évolution, selon un principe de maximum de vraisemblance, pour les familles de protéines d'HOGENOM [Penel *et al.*, 2009] et les familles de modules de Pfam [Finn *et al.*, 2010] et ProDom [Bru *et al.*, 2005]. Les modules Pfam sont expertisés manuellement, ce qui garantit une certaine qualité des familles de modules. Les modules ProDom sont construits automatiquement, ce qui permet d'obtenir une couverture maximale de l'univers des protéines. La détermination d'une architecture en modules Pfam et ProDom caractéristique de chaque famille de protéines a permis de lier les histoires évolutives des protéines et des modules. Ainsi l'émergence de chaque famille de protéines peut être mise en perspective par rapport aux histoires des modules composant leur architecture.

Ce chapitre présente l'ensemble des résultats obtenus sur l'évolution des répertoires de protéines et des modules composant leur architecture, ainsi que sur la mise en parallèle de ces deux niveaux d'évolution. Après avoir détaillé les différents jeux de données utilisés et les outils d'inférence et d'analyse des scénarios d'évolution, nous décrirons la dynamique évolutive des répertoires de familles de protéines. Ensuite, nous présenterons le point de vue des modules Pfam, puis celui des modules ProDom sur l'évolution des répertoires de modules et sur l'émergence des nouvelles protéines. Enfin, nous comparerons plus précisément les résultats obtenus avec ProDom et Pfam pour caractériser leurs points forts et leurs points faibles.

3.1 Méthodes et données pour l'inférence des contenus en protéines et modules ancestraux

Cette section présente les différents jeux de données des familles de protéines et des familles de modules utilisés ainsi que les modifications apportées aux familles de modules ProDom qui ont été regroupées une deuxième fois. Les procédures associées à la détermination des architectures en modules ProDom et Pfam des familles de protéines sont ensuite développées. Puis nous décrivons brièvement l'application de la méthode des réseaux Bayésiens présentée dans le chapitre précédent, pour l'inférence des scénarios d'évolution des familles de protéines et de modules. Ces scénarios sont interprétés à l'aide de différents événements évolutifs qui nous permettront d'analyser la dynamique évolutive des répertoires de protéines et de modules. Enfin, nous décrivons l'outil web développé pour l'analyse et la visualisation des données obtenues.

3.1.1 Préparation des jeux de données

Nous avons travaillé sur le même jeu de données de 170 espèces, dont les génomes sont complètement séquencés [1], que dans le chapitre précédent. L'arbre des espèces [1] utilisé dans ce chapitre est mieux résolu que celui de la taxonomie du NCBI, dont les principales modifications sont présentées en annexe C (page 142).

3.1.1.1 Description des bases de données

Les familles de protéines ont été extraites de la base de données HOGENOM version 3 [Penel *et al.*, 2009]. Les protéines entières sont regroupées selon un critère de similarité assez stringent : au moins 50% de similarité sur au moins 80% de la longueur des protéines détectées avec le programme BLASTP [Altschul *et al.*, 1997] dont le seuil de E-value est fixé à 10^{-4}. Un critère de simple lien permet de regrouper les protéines similaires. Le critère de similarité permet d'obtenir des familles dont les alignements sont de bonne qualité tout en maximisant le nombre d'espèces représentées. De plus, les familles de protéines sont relativement homogènes en terme d'architecture de domaines, il est donc possible de faire l'hypothèse qu'une famille de protéines correspond à un arrangement en domaines particulier. Le sous-ensemble des familles restreintes aux espèces sélectionnées contient 194 844 familles dont 151 392 sont spécifiques de l'une des espèces considérées.

Les familles de modules protéiques ont été extraites des bases de données ProDom (version 2005.1) [Bru *et al.*, 2005] et Pfam (version 17) [Finn *et al.*, 2010]. La définition des mo-

1. La liste des espèces est disponible en annexe C, page 137 et l'arbre phylogénétique des espèces modifié est page 141

dules est basée sur la conservation des séquences protéiques, on préfère donc parler de modules protéiques plutôt que de domaines.

ProDom est construite automatiquement à partir des séquences protéiques extraites de Swiss-Prot et TrEmbl [Consortium, 2009] en utilisant l'algorithme MKDOM2 [Gouzy *et al.*, 1999]. La procédure itérative part de l'hypothèse que la plus petite séquence non fragmentaire de la banque est monodomaine et peut donc être utilisée comme séquence requête pour rechercher ses domaines homologues dans la base de séquences. La recherche de similarité locale est faite à l'aide du programme PSI-BLAST [Altschul *et al.*, 1997] avec un seuil pour la E-value fixé à 10^{-6}. Les premiers domaines recherchés sont issus des domaines structuraux expertisés de SCOP [Andreeva *et al.*, 2008].

Pfam est une base de familles de modules expertisés manuellement (Pfam-A). Chaque famille est caractérisée par un alignement expertisé et contenant relativement peu de séquences représentatives de la famille. Cet alignement est ensuite utilisé pour recruter l'ensemble des modules de la famille à l'aide d'un modèle de Markov caché (profile HMM) et du programme HMMER3. [1]

3.1.1.2 Regroupement des familles de modules ProDom

fLa procédure de recherche de similarité de ProDom n'est pas assez sensible, ce qui tend à subdiviser les familles en sous-familles de modules. Or pour l'analyse de l'évolution des modules à l'échelle du vivant, il faut prendre en compte les relations d'homologies plus distantes. Il a donc été nécessaire d'effectuer un nouveau regroupement des familles capturant les relations d'homologie plus distantes. Le regroupement des familles de modules s'est effectué en deux étapes : la première repose sur la détection des relations d'homologie entre les familles de modules, et la deuxième effectue le regroupement des familles en fonction de ces relations d'homologie (figure 3.1, partie 1 et 2).

Recherche des relations d'homologies

Les relations entre familles de modules ont été obtenues à partir de deux recherches de similarité indépendantes. Dans la première, les résultats de la comparaison de tout ProDom contre lui-même effectuée avec le programme BLASTP et un seuil de la E-value fixé à 10^{-2} ont été utilisés (ceux-ci sont disponibles dans ProDom). Une relation entre deux familles de modules est détectée lorsque le chevauchement entre les deux modules est supérieur à 80% de la longueur de chacun des modules. Ce seuil relativement élevé permet de garantir une homologie globale et donc une certaine homogénéité des longueurs des modules appartenant à une même famille (figure 3.1,1a). Dans la deuxième analyse, les profils des alignements des familles de protéines ont été comparés

1. La version HMMER2 a été utilisée avec la version 17 de la base de données, ces programmes sont disponibles à l'adresse suivante : `http://hmmer.janelia.org/software/archive`

à la banque de séquences consensus des familles de modules ProDom à l'aide du programme PSI-BLAST. Cette procédure a permis d'obtenir une association entre les familles de protéines et les familles de modules (figure 3.1, 1b).

Pour réaliser le PSI-BLAST, nous avons utilisé les deux bases de données complètes, soit 736 449 familles de modules extraites de ProDom version 2005.1 et 262 865 familles de protéines extraites d'HOGENOM version 3. Les séquences consensus des familles de modules ProDom ont été utilisées comme banque de recherche.

Les alignements des familles de protéines ont été utilisés comme alignement de départ dans la construction du profil de PSI-BLAST. L'alignement de la plupart des familles est disponible dans HOGENOM, excepté pour les 83 plus grosses familles. Leur alignement a été calculé avec MUSCLE [Edgar, 2004] sur un échantillon aléatoire de 300 séquences. Les séquences consensus des alignements ont été utilisées comme séquences requêtes. Elles ont été calculées à l'aide de la formule suivante appliquée à chaque position i de l'alignement :

$$cons_i = \arg\max_{a \in \mathcal{A}} \left(\sum_{b \in \mathcal{B}_i} subst(a,b) \right) \tag{3.1}$$

où $cons_i$ représente l'acide aminé en i-ème position de la séquence consensus, $\mathcal{A}$ représente l'ensemble des 20 acides aminés plus le gap, $\mathcal{B}_i$ contient la colonne i de l'alignement et $subst(a,b)$ correspond au score de la matrice de substitution Blosum62 si a et b sont des acides aminés, sinon $subst(a,gap) = subst(gap,b) = -8$ et $subst(gap,gap) = +2$. Le PSI-BLAST a été réalisé avec les options suivantes : le seuil de la E-value pour la conservation d'un résultat a été fixé à 10^{-2}, la matrice Blosum62 a été utilisée et une seule itération a été réalisée.

Toutes les correspondances trouvées n'ont pas été gardées. En effet, on recherche des modules entiers ayant une correspondance sur les familles de protéines. Ainsi, les résultats pour lesquels la portion de la séquence requête ou de la séquence du module impliquée dans l'alignement a une longueur inférieure à 80% de la longueur du module n'ont pas été gardés.

Les relations d'homologies entre familles de modules ont été déterminées en analysant leur chevauchement sur les séquences consensus des familles de protéines. Lorsque deux modules se chevauchent sur plus de 80% de la longueur du plus grand module, alors la relation est gardée. Par exemple, dans la figure 3.1 partie 1b, une relation est détectée entre PD1 et PD4 mais pas entre PD1 et PD2.

Procédure de regroupement

L'algorithme MCL (Markov CLustering) a été utilisé pour regrouper les familles. Cet algorithme [van Dongen, 2000] construit un graphe de relations pondérées et le partitionne en simulant des marches aléatoires à l'intérieur de celui-ci à l'aide de matrices stochastiques (appelées matrices

de Markov). Cet algorithme a un paramètre appelé Inflation qui influence la granularité des regroupements : une forte inflation (de l'ordre de $I = 4$) augmente le nombre de regroupements détectés qui réuniront moins de nœuds, alors qu'une faible inflation (de l'ordre de $I = 1.1$) diminue le nombre de regroupements dont les effectifs seront plus grands.

Pour le regroupement des familles de modules, nous avons utilisé un poids uniforme pour toutes les relations, puisqu'il n'était pas possible de définir un poids (comme la E-value ou un score) à partir des relations issues du PSI-BLAST contre HOGENOM (la comparaison des familles de modules est indirecte). Le paramètre d'inflation par défaut ($I = 2$) a été utilisé.

Cette procédure a permis de regrouper 133 165 familles de modules ProDom en 37 760 nouvelles familles. La plus grosse nouvelle famille regroupe 600 familles de modules. Ces regroupements de familles ProDom constituent les nouvelles familles de modules ProDom considérées dans l'ensemble des résultats de ce chapitre. Par simplification, l'identifiant de chaque nouvelle famille correspond à l'identifiant ProDom de la famille regroupée la plus représentée dans ProDom.

3.1.2 Découpage des familles de protéines en architectures de modules

Pour faire le lien entre les familles de protéines et de modules, nous avons déterminé une architecture de modules spécifique pour chacune des familles de protéines avec les modules ProDom d'une part et Pfam d'autre part. La première stratégie envisagée a été d'aligner les architectures de modules des séquences protéiques composant une famille dans HOGENOM. Avec les données de ProDom, les architectures d'une famille sont identiques pour 36% des familles de protéines ayant au moins deux architectures disponibles. Les différences observées dans les autres familles ont essentiellement deux explications. La première concerne la méthode de regroupement d'HOGENOM qui peut par exemple ajouter par transitivité des séquences ayant un module supplémentaire. La deuxième implique les subdivisions de certaines familles de modules ProDom qui peuvent conduire à des architectures de modules différentes. Cependant, même en prenant en compte ces deux phénomènes, il n'a pas été possible de déterminer une architecture pour toutes les familles (données non montrées). Nous avons donc mis en place une procédure basée sur la recherche de similarité par PSI-BLAST entre les familles de protéines et les familles de modules. Cette procédure a permis d'améliorer le regroupement des familles de modules ProDom en nouvelles familles comme nous l'avons montré dans la section précédente, et de déterminer une architecture en modules ProDom spécifique pour chaque famille de protéines.

Les arrangements en modules ProDom ont été déterminés à partir des résultats de recherche de similarité des familles de protéines sur la banque de familles de modules ProDom à l'aide du programme PSI-BLAST (figure 3.1, partie 3). Les modules identifiés ont été traduits en nouvelles familles de ProDom d'après le regroupement réalisé précédemment. Lorsque deux occurrences

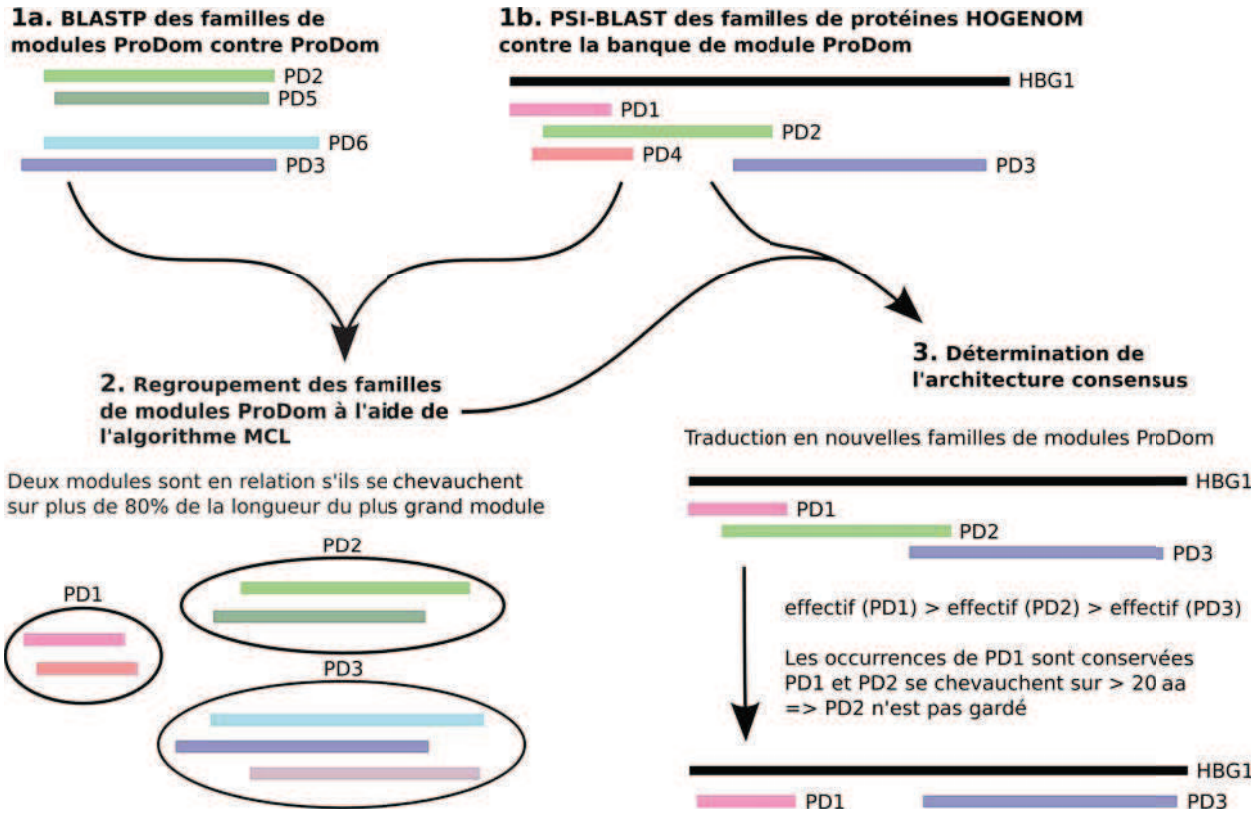


FIGURE 3.1 – **Procédure d'obtention des architectures en modules ProDom. 1.** Les homologies entre familles de modules ont été détectées à l'aide d'une recherche de similarité des familles de ProDom contre ProDom avec BLASTP (**1a**), et d'une comparaison des familles de protéines d'HOGENOM et de la banque des séquences consensus de ProDom avec PSI-BLAST (**1b**). **2.** Les familles de modules similaires ont été regroupées en nouvelles familles à l'aide de l'algorithme MCL. **3.** Les architectures de modules ProDom sont ensuite déduites des résultats du PSI-BLAST. Les familles de modules détectées sont traduites en nouvelles familles obtenues à l'étape 2. La famille PD4 devient donc PD1 d'après le regroupement effectué. Les familles de modules sont ensuite triées par ordre décroissant de leur effectif dans ProDom. Dans cet exemple, toutes les occurrences de la famille PD1 sont conservées en premier, puis toutes les occurrences des autres familles qui ne chevauchent aucun module conservé sur plus de 20 acides aminés : PD2 n'est donc pas conservée.

d'une famille de modules se chevauchent sur plus de 20 acides aminés, c'est celle ayant obtenu la plus petite E-value avec le PSI-BLAST qui est conservée. Les familles de modules sont ensuite triées en fonction de leur effectif dans la base de données ProDom dans l'ordre décroissant. Les familles les plus représentées sont conservées prioritairement lorsque deux familles différentes se chevauchent sur plus de 20 acides aminés. Une architecture de modules ProDom a été déterminée pour 190 648 familles. Ces architectures sont constituées à partir d'un jeu de données de 267 595 familles de modules ProDom dont 176 891 sont spécifiques de l'une des espèces sélectionnées. Le tableau 3.1 récapitule les caractéristiques de ces architectures comme la couverture en acides aminés ou la modularité moyenne des architectures.

Pour déterminer les arrangements en modules Pfam, nous avons utilisé l'outil HMMER2. Cet outil permet de comparer une séquence protéique à la banque de profils HMM (Hidden Markov

Model) des familles de modules Pfam. Nous avons utilisé les séquences consensus des alignements des familles de protéines (formule 3.1) comme séquences requêtes, et les profiles HMM qui permettent de rechercher la similarité globale des modules comme banque de recherche. Le seuil de la E-value a été fixé à 10^{-2}. La détermination des architectures consensus en modules Pfam a été effectuée de la même manière que pour les modules ProDom : les familles les plus représentées dans Pfam sont conservées prioritairement lorsque deux familles se chevauchent sur plus de 20 acides aminés. Une architecture de modules Pfam a été déterminée pour 76 337 familles de protéines. Sur les 7 868 familles de modules présentes dans Pfam-A, 6 659 sont présentes dans au moins l'une des architectures inférées, et parmi elles 705 sont spécifiques de l'une des espèces sélectionnées. La couverture des familles de protéines ainsi que leur modularité moyenne, selon ProDom ou Pfam, sont résumées dans le tableau 3.1.

	ProDom	Pfam
Familles avec architecture	98%	39%
Couverture en acides aminés	85%	21%
↪ familles avec architecture	87%	55%
Modularité	2,2	1,7

TABLEAU 3.1 – **Couverture et modularité des familles de protéines.** Les chiffres sont donnés pour les architectures de modules reconstruites avec les modules ProDom et Pfam. La couverture en acides aminés correspond à la proportion d'acides aminés couverts par un module sur la séquence consensus de chaque famille de protéines. La modularité est le nombre moyen de modules par architecture.

3.1.3 Identification des paramètres des modèles

L'inférence des scénarios d'évolution a été réalisée à l'aide du modèle de réseau Bayésien décrit dans le chapitre précédent [1]. Le modèle d'évolution utilisé définit des probabilités de gain et de perte hétérogènes le long de l'arbre. Les probabilités de gain et de perte ont été estimées indépendamment pour chacun des trois jeux de données (HOGENOM, ProDom et Pfam). Les estimations ont été réalisées avec 20 initialisations différentes et un jeu d'apprentissage de 10 000 familles extraites aléatoirement des jeux de données des familles de protéines et des familles de modules de ProDom. Les paramètres de Pfam ont été estimés avec l'ensemble des familles puisqu'il y en a moins de 10 000. Pour chaque jeu de données, les jeux de paramètres avec la meilleure vraisemblance ont été gardés pour calculer les scénarios d'évolution optimaux et les probabilités marginales associées à chaque état des scénarios d'évolution. Les profils phylogénétiques utilisés comme données de départ représentent la présence et l'absence des familles de protéines ou de modules dans chacune des 170 espèces contemporaines sélectionnées.

1. Voir la section 2.2, page 30

3.1.4 Typologie de l'histoire des familles

Les scénarios d'évolution inférés retracent l'histoire évolutive des familles dans laquelle cinq types d'événement évolutif peuvent être distingués (figure 3.2). Tout d'abord, on peut distinguer deux origines possibles pour une famille :

1. La famille est spécifique du clade dans lequel elle est gagnée. Dans ce cas, on parle d'*innovation* : une nouvelle protéine ou un nouveau module est apparu.
2. La famille est gagnée au moins deux fois dans deux sous-arbres distincts. Dans ce cas, on peut postuler que l'un des gains correspond à une innovation et le ou les autres sont le résultat de *transferts horizontaux*. Cependant, notre approche ne permet pas d'identifier l'événement d'innovation parmi les différents gains et donc l'origine de la famille puisque le modèle utilisé ne prend pas explicitement en compte les transferts horizontaux. Ainsi, tous les gains sont interprétés comme des transferts horizontaux. L'hypothèse alternative de convergence n'est pas retenue ici puisqu'on fait l'hypothèse que les familles sont constituées de modules ou de protéines homologues.

La famille peut être *transmise verticalement* à sa descendance (héritage) ou bien être *perdue* au cours de son histoire. On peut enfin distinguer un dernier type d'événement lorsque la famille a été gagnée une première fois, puis perdue avant d'être *regagnée* dans la même lignée. Ce dernier type d'événement constitue un cas particulier de transfert horizontal.

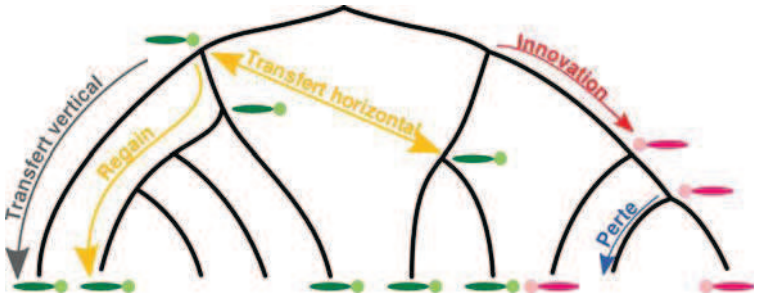


FIGURE 3.2 – **Schéma représentant les différentes origines possibles pour une famille présente à un nœud donné.** On distingue cinq types d'événement : le transfert vertical lorsque la famille est également présente dans le parent ; l'innovation lorsqu'une nouvelle famille est spécifique du sous-arbre ; le transfert horizontal lorsque plusieurs événements de gain sont inférés dans différents sous-arbres ; le regain lorsque dans une même lignée une famille est gagnée, perdue puis regagnée ; la perte.

Le contenu des répertoires ancestraux en protéines et en modules a été déduit des scénarios d'évolution sans prendre en compte les probabilités marginales associées aux états inférés. Les familles de chaque répertoire ont ensuite été classées en fonction de leur provenance : transférées verticalement, innovées ou transférées horizontalement (catégorie comprenant les événements de regain).

3.1.5 EvolProDom : un site web de visualisation et d'analyse des scénarios d'évolution

Ce projet nécessite la gestion de différents jeux de données à l'interface des bases de données de protéines HOGENOM et de modules ProDom et Pfam : les architectures consensus des familles de protéines, et les scénarios d'évolution des familles de protéines et de modules. Deux outils facilitant l'accès, la navigation, ainsi que la visualisation de l'ensemble de ces informations ont été développés : des services web et une interface web. Ces deux outils reposent sur une librairie commune de fonctions d'accès aux données (scénarios, événements, architectures de protéines, etc.) dont l'avantage est la facilité d'ajout de fonctions supplémentaires lorsque de nouvelles données sont produites. L'association des différentes fonctionnalités permet de faire des requêtes complexes, notamment pour rechercher l'origine d'une famille de protéines en fonction du répertoire parental en modules.

3.1.5.1 Implémentation

Les services web ont été développés en collaboration avec Lauranne Duquenne (dans le cadre du projet EMBRACE et du développement de ProDom sur la grille). Certains sont accessibles en ligne sur le site de ProDom et ont été publiés sur le site EMBRACE Service Registry [EMBRACE, 2008] , qui regroupe une collection de services web pour les sciences de la vie. Ils permettent d'accéder aux scénarios d'évolution des familles de domaines de ProDom CG267 [Bru *et al.*, 2005] (ProDom version 2006 restreinte aux génomes complets) et des familles de protéines d'HOGENOM version 4 [Penel *et al.*, 2009]. Des associations de différents services web (workflows) sont également disponibles. Ils permettent de faire le lien entre les familles de protéines et les familles de domaines, par exemple pour rechercher l'origine d'une nouvelle architecture en fonction du répertoire en modules du parent.

L'interface web est avant tout un outil de visualisation des scénarios d'évolution et des arrangements en modules des familles de protéines et de modules. Cette interface permet de réaliser des requêtes sur les différentes données disponibles, de naviguer dans l'ensemble de ces données et d'accéder à différentes bases de données externes. Le développement de cette interface a été réalisé en partie par Idris Galbert, une étudiante en master professionnel Compétence Complémentaire en Informatique, que j'ai encadrée.

3.1.5.2 Fonctionnalités

Cette interface propose un formulaire permettant à l'utilisateur de choisir : (1) les données sur lesquelles il veut travailler, avec la possibilité de comparer les résultats obtenus avec les fa-

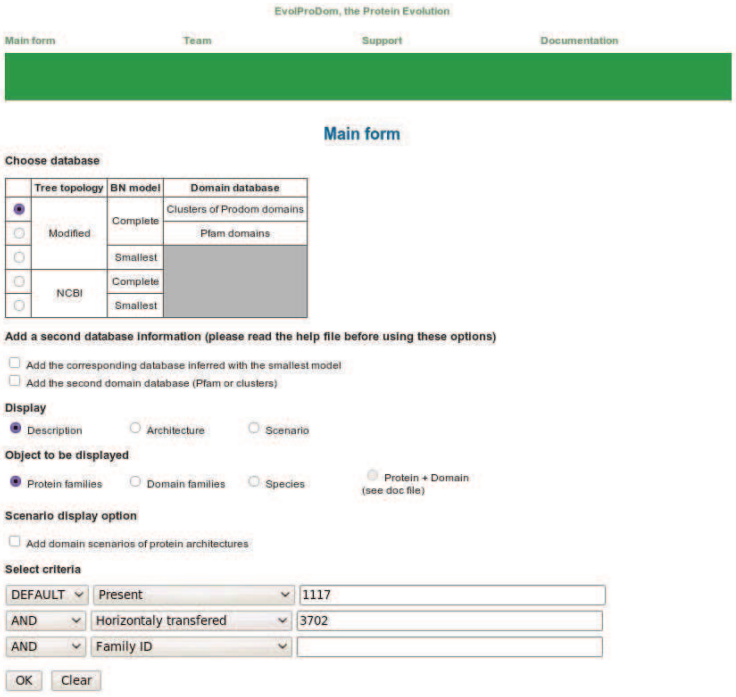


FIGURE 3.3 – **Exemple de requête.** Cette requête recherche l'ensemble des familles de protéines présentes dans l'ancêtre des cyanobactéries (identifiant taxonomique 1117) et transférées horizontalement chez *Arabidopsis thaliana* (identifiant taxonomique 3702). Elle renvoie la description des 185 familles satisfaisant ces critères. Les arrangements en modules et les scénarios d'évolution de chacune des familles sont également disponibles à partir de la page de description.

milles de modules ProDom et Pfam ou bien les résultats obtenus avec différents modèles évolutifs (implémentés dans les réseaux Bayésiens) ; (2) l'objet de son étude, les familles de protéines, de modules ou bien les espèces ; (3) la visualisation de la description de l'objet, des arrangements en modules ou des scénarios d'évolution ; (4) les caractéristiques de l'objet à travers un système de requêtes combinant opérateurs logiques et mots-clés auxquels l'utilisateur associe une valeur (voir l'exemple de la figure 3.3).

La visualisation des scénarios d'évolution est la partie la plus importante de cette interface. En effet, l'objectif était de pouvoir visualiser plusieurs scénarios sur le même arbre, par exemple le scénario d'une famille de protéines et des familles de modules composant son architecture. Pour cela, nous avons modifié certains aspects de l'affichage de la librairie TreeFam [Ruan *et al.*, 2008] qui permet d'afficher des arbres phylogénétiques "dynamiques". Le ou les scénarios d'évolution sont représentés par des carrés (pour les protéines) ou des ronds (pour les domaines) colorés aux nœuds où la famille est présente (voir figure 3.4). Pour une question de facilité de lecture, le nombre de familles par arbre est au maximum de trois. Les probabilités marginales sont affichées lorsqu'elles sont inférieures à $0{,}9$, ce qui permet de visualiser rapidement un faible soutien du modèle.

L'exemple de la figure 3.4 présente le scénario d'évolution de la famille du cytochrome f (fa-

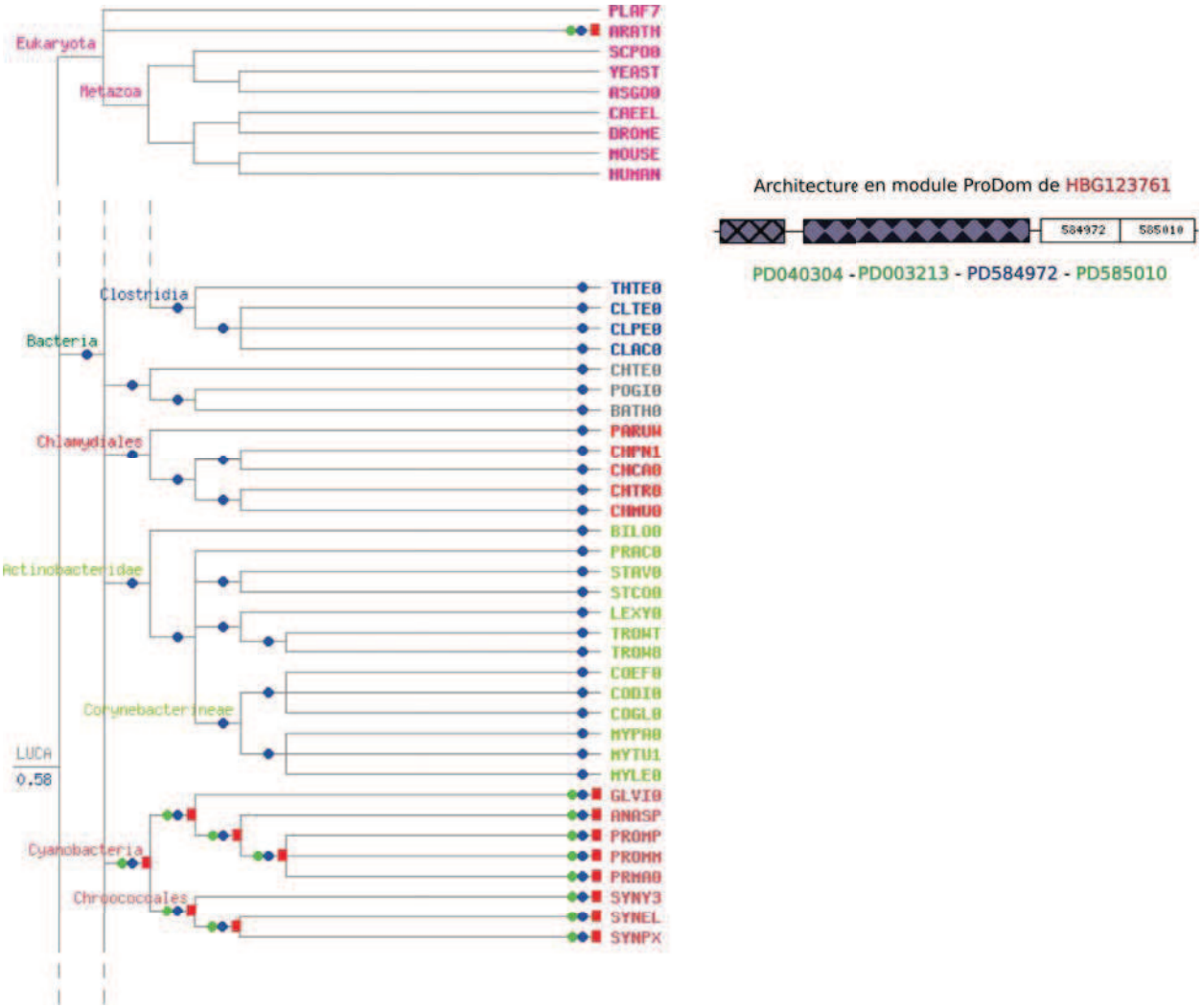


FIGURE 3.4 – **Scénarios d'évolution de la famille du cytochrome f et des familles de modules PD584972 et PD003213 présentes dans son architecture.** La présence de HBG123761 est représentée par des carrés rouges, celle de PD584972 par des ronds bleus et celle de PD003213 par des ronds verts. Les probabilités marginales inférieures à 0,9 sont affichées en dessous du nœud de la couleur de la famille correspondante et indiquent les prédictions moins bien soutenues localement par le modèle. L'architecture en modules ProDom représentée compte deux modules supplémentaires, PD040304 et PD585010, qui ont le même scénario d'évolution que PD003213, comme symbolisé par la couleur de leur identifiant.

mille HBG123761). Cette famille fait partie des résultats trouvés avec la requête de la figure 3.3. Cette protéine participe à la formation du complexe cytochrome b6-f qui effectue des transferts d'électron entre le photosystème II et le photosystème I impliqués dans la chaîne de réaction de la photosynthèse. Le scénario d'évolution de la famille de protéines indique qu'elle est apparue à la fois chez les Cyanobactéries et chez les plantes, dans lesquelles la protéine est codée dans le chloroplaste. Dans cet exemple, on déduit un transfert horizontal des Cyanobactéries vers les plantes lors de l'événement d'endosymbiose à l'origine des chloroplastes. L'architecture en modules de cette famille est composée de 4 modules ProDom dont 3 sont spécifiques des cytochromes f et le quatrième est retrouvé dans différents contextes protéiques. Le premier module défini sur la séquence, PD040304, correspond à la séquence signal de la protéine. Quant aux deux autres nouveaux do-

maines, ils semblent être des maillons importants dans la mise en place de la photosynthèse en étant à l'origine du cytochrome F qui est essentiel dans la chaîne de réaction photosynthétique. Cette interface fournit ainsi une vue d'ensemble des données disponibles (scénarios, arrangements en modules, annotations, etc.) qui permet une interprétation des scénarios d'évolution.

3.2 Dynamique évolutive des familles de protéines

L'inférence de scénarios d'évolution pour chacune des familles de protéines représentées dans les 170 espèces analysées nous permet de décrire leur histoire évolutive le long de l'arbre des espèces, et d'en déduire des répertoires ancestraux. Ces scénarios sont le reflet des événements de gain et de perte qui permettent d'apprécier la dynamique évolutive à la fois des familles et des répertoires de protéines. La dynamique de ces répertoires est très hétérogène sur l'ensemble de l'arbre du vivant. Elle montre une variation importante des tailles de répertoires d'une lignée à l'autre, comme illustré dans la figure 3.5A. Dans cette figure, les diagrammes associés à chaque espèce représentent les répertoires en familles de protéines, classées en fonction de leur origine probable. La surface des diagrammes est proportionnelle au nombre de familles inférées présentes et permet de visualiser les événements d'expansion et de réduction des répertoires de gènes qui ont eu lieu tôt au cours de l'évolution.

Ces variations s'observent également à plus petite échelle comme illustré dans le sous-arbre des Alphaprotéobactéries (figure 3.5B). À partir d'un répertoire de taille moyenne, un peu moins de 2 000 familles, les répertoires de la lignée des *Rickettsiaceae* se sont réduits jusqu'à moins de 650 familles de protéines chez *Rickettsia prowazekii*. Au contraire, les lignées des *Bradyrhizobiaceae* et des *Rhizobiaceae* ont vu leur répertoire s'accroître jusqu'à plus de 5 600 familles chez *Bradyrhizobium japonicum*. Ces observations sont en accord avec les résultats antérieurs [Blanc *et al.*, 2007; Boussau *et al.*, 2004]. La réduction des génomes d'organismes endosymbiotiques tels que les *Rickettsia* (parasites intracellulaires obligatoires), les Mollicutes (parasites des eucaryotes), les *Chlamydiaceae* (agents pathogènes des animaux) ou encore les *Buchnera* (endosymbiotes obligatoires des pucerons) [1] s'expliquent en partie par une adaptation à la vie intracellulaire et parasitaire, dans laquelle la sélection ne maintient plus de nombreux gènes devenus inutiles dans l'environnement de l'hôte. La réduction du génome est également liée à une dégénérescence du génome dans lequel les gènes de réparation de l'ADN sont en majorité perdus [Moran *et al.*, 2008; Ochman et Moran, 2001]. L'accumulation de mutations délétères est le résultat d'une taille de population efficace réduite et de la réduction des échanges de matériels génétiques (phénomène de dérive génétique).

1. Les diagrammes des répertoires associés aux différents sous-arbres sont disponibles en annexe B (pages 115 à 135)

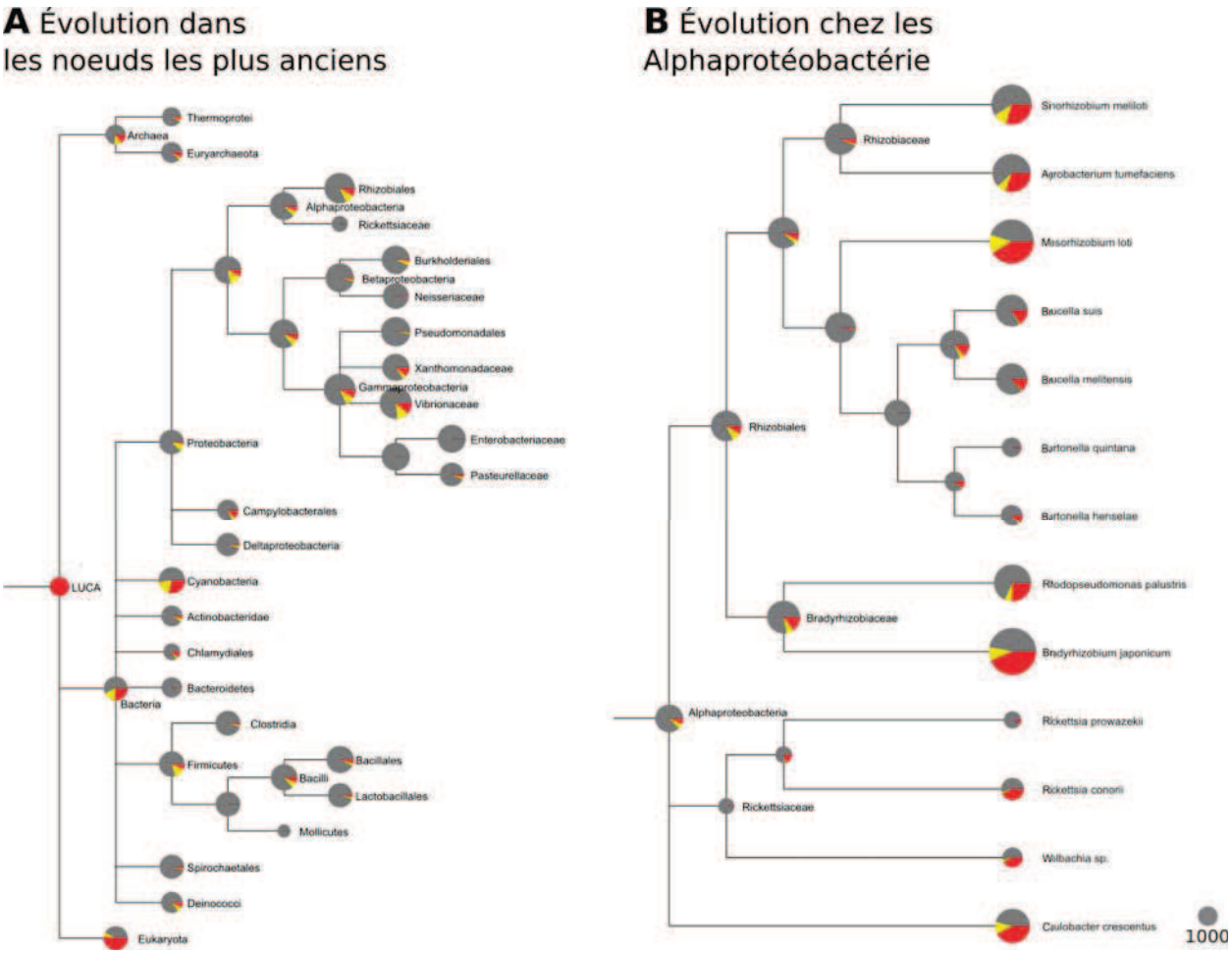


FIGURE 3.5 – **Évolution des répertoires des familles de protéines dans les nœuds les plus anciens de la taxonomie (A) et chez les Alphaprotéobactéries (B).** Les diagrammes circulaires reflètent le nombre de familles inférées présentes à chaque nœud de l'arbre des espèces. Les familles héritées de leur parent sont représentées par un secteur gris. Les familles gagnées pouvant être retrouvées dans d'autres clades sont inférées transférées horizontalement et indiquées par un secteur jaune. Les familles regagnées sont indiquées en noir. Toutes les autres familles sont prédites innovées et sont indiquées en rouge. La surface des diagrammes est proportionnelle à la taille des répertoires.

Les événements de gain multiples sont en règle générale interprétés comme des transferts horizontaux. Si certains auteurs préfèrent négliger ce type d'événement [Blanc *et al.*, 2007; Sakarya *et al.*, 2008], la plupart du temps le transfert horizontal est au cœur des modèles de parcimonie – les pénalités de gain et de perte influencent le nombre de transferts. Il n'est plus à prouver aujourd'hui que le monde procaryote pratique massivement le transfert horizontal [Gogarten et Townsend, 2005], mais la caractérisation de ces événements nécessiterait d'exploiter aussi l'information de séquence afin d'orienter les transferts. Dans notre analyse, les familles pour lesquelles on ne peut pas déterminer une origine précise sont considérées comme transférées horizontalement (tableau 3.2).

Les innovations et les pertes sont des événements que l'on peut plus facilement quantifier

HOGENOM	Total	Procaryotes
Nb familles	194 844	133 477
↪ Nb familles innovées	182 254	122 340
↪ Nb familles transférées	12 590	11 137
Nb familles regagnées	814	776
Nb moyen de transferts par famille	1,8	1,9
Nb moyen de perte par famille	0,34	0,48

TABLEAU 3.2 – **Caractérisation des différents événements pour les familles de protéines.** Ces données ont été calculées sur les scénarios d'évolution bruts obtenus. Chaque famille a été innovée au moins une fois dans son histoire, mais certaines présentent des gains multiples interprétés comme des transferts horizontaux, leur origine n'est donc plus accessible puisque le sens du transfert n'est pas donné par le modèle. Le nombre moyen de transferts a été calculé comme le *nombre de gains* -1 sur les familles ayant subi au moins un transfert au cours de leur histoire. Les familles regagnées sont des familles qui ont été gagnées dans leur histoire (innovation ou transfert) puis perdues et gagnées à nouveau dans la même lignée. Les colonnes procaryotes présentent les résultats pour les jeux de données de familles sans les familles spécifiques des eucaryotes.

le long de l'arbre des espèces. Les événements de perte semblent relativement rares puisqu'en moyenne chaque famille est perdue moins d'une fois au cours de son histoire (tableau 3.2). Si on les rapporte au nombre de familles dans l'espèce parentale (figure 3.6B), les fréquences de perte présentent une distribution asymétrique avec quelques espèces qui ont subi des pertes massives (> 60% du répertoire parental), alors que la majorité des espèces ont perdu moins de 10% des familles parentales.

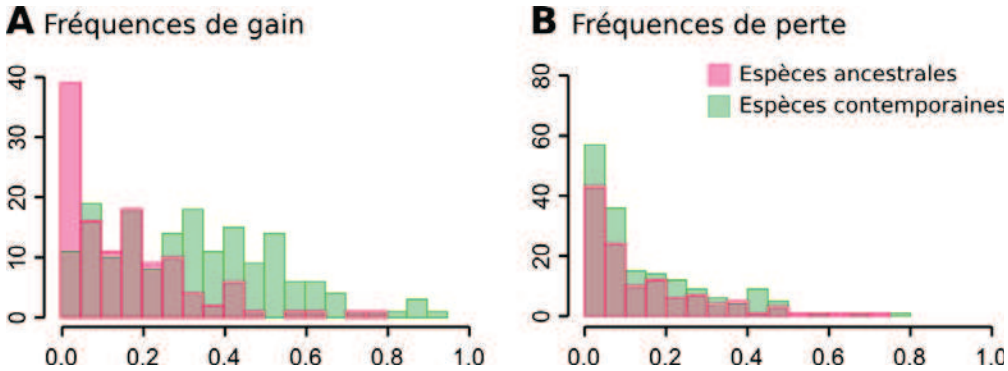


FIGURE 3.6 – **Distributions des fréquences de gain (A) et de perte (B) des familles de protéines.** Les distributions des fréquences dans les espèces contemporaines et ancestrales sont représentées respectivement en vert et en rouge. La fréquence de gain correspond à la proportion des familles de chaque espèce gagnées par innovation, transfert horizontal ou regain. La fréquence de perte correspond à la proportion de familles perdues par rapport au répertoire parental.

La figure 3.7 présente la répartition des familles innovées dans chacun des domaines du vivant. Pour les familles de protéines, plus de 100 000 familles ont été innovées chez les bactéries et sont

spécifiques de ce sous-arbre. Cependant, une minorité de familles remonte à l'ancêtre des bactéries, ce qui signifie que la majorité des protéines a été innovée plus tard au cours de l'évolution. L'innovation est un phénomène continu (confirmé par les diagrammes de répartition des familles dans chacun des répertoires des espèces de l'arbre). Cette caractéristique se retrouve dans l'ensemble du vivant. L'innovation protéique représente en moyenne 21% des répertoires des espèces procaryotes. Cependant, elle apparaît significativement plus grande dans les espèces actuelles (26%) que dans les espèces ancestrales (11%). La figure 3.6A qui présente les distributions des fréquences de gain dans les espèces contemporaines et ancestrales montre une fréquence plus élevée dans les espèces contemporaines.

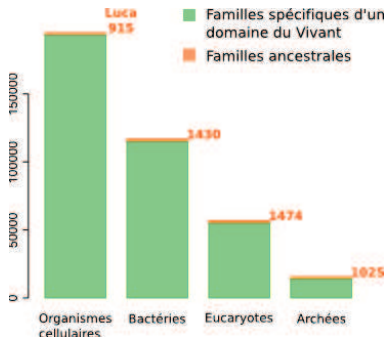


FIGURE 3.7 – **Répartitions des familles de protéines dans les différents domaines du vivant.** La figure représente le nombre de familles de protéines spécifiques de chaque grand domaine du vivant (Eucaryotes, Bactéries et Archées). Parmi elles, les familles retrouvées dans l'ancêtre commun sont indiquées en rouge.

Le sous-arbre des eucaryotes (figure B.3, page 118) présente des taux d'innovation d'un ordre de grandeur plus grand que ceux observés chez les procaryotes. Ce résultat s'explique en partie par des raisons biologiques, notamment par l'augmentation de la complexité des génomes eucaryotes par rapport aux génomes procaryotes : augmentation du nombre de gènes, d'introns et d'éléments mobiles [Lynch et Conery, 2003]. Cependant, une partie significative de l'innovation inférée chez les eucaryotes peut être d'origine artéfactuelle. En effet, la complexité de ces génomes conduit à une difficulté supplémentaire dans la prédiction des gènes, dont certains peuvent avoir été surinterprétés [Clamp *et al.*, 2007]. De plus, l'échantillonnage taxonomique est réduit puisque seules 9 espèces d'eucaryotes sont considérées dans notre étude. En conséquence, les forts taux d'innovation que nous inférons chez les eucaryotes doivent être considérés comme problématiques à ce stade. Les familles de protéines et de modules spécifiques des eucaryotes ont donc été systématiquement retirées de la plupart des analyses présentées dans la suite de ce chapitre pour ne pas biaiser les résultats : cela correspond à 31% des familles de protéines, 34% des familles de Pfam et 42% des familles de ProDom spécifiques des eucaryotes. Pour les autres familles, nous avons pris en compte le sous-arbre eucaryote uniquement comme groupe externe des bactéries et des archées.

3.3 Les modules Pfam : un répertoire ancien privilégiant les réarrangements de modules

3.3.1 Dynamique évolutive du répertoire de modules Pfam

Les répertoires de modules Pfam présentent une dynamique évolutive dominée par les événements de perte et de transferts verticaux (figure 3.8). L'essentiel de la variation des tailles de répertoires au cours de l'évolution est dû aux événements de pertes : chaque famille de modules a été perdue en moyenne 7 fois (tableau 3.3). En revanche, les événements de gains sont de moins en moins fréquents lorsqu'on se rapproche des espèces contemporaines. De plus, les familles sont impliquées dans des événements de gains multiples dans plus de 50% des cas, il n'est donc pas possible de déterminer une origine précise pour ces familles : on dénombre en moyenne 4 événements de transferts horizontaux par famille. Ces transferts constituent la majorité des événements de gains dans les répertoires. En effet, les innovations présentent une répartition inégale où quelques nœuds majoritairement anciens concentrent l'essentiel des événements. Ainsi, un peu moins de 50% des familles Pfam sont innovées à LUCA (Last Universal Common Ancestor) ou dans l'un des trois ancêtres des domaines du vivant. Et aucune innovation n'est inférée dans 128 espèces ancestrales et contemporaines, parmi lesquelles 17 ne présentent aucun gain. La figure 3.9 présente une répartition de l'innovation dans les trois domaines du vivant : 45% des familles bactériennes et 85% des familles archéennes sont présentes dans les ancêtres des domaines. Très peu de familles Pfam sont apparues récemment au cours de l'évolution.

3.3.2 Origine des nouvelles protéines

3.3.2.1 Comparaison des répertoires d'architectures et de modules

Les scénarios d'évolution des familles de protéines et de modules sont indépendants, ils ont été inférés avec des jeux de paramètres ajustés à chacun des jeux de données. Les architectures en modules permettent de faire le lien entre les familles de protéines et les familles de modules Pfam. Avant de mettre en relation directe les scénarios d'évolution des familles de protéines et de modules pour comprendre l'émergence des nouvelles protéines, nous avons analysé les contenus des répertoires en protéines et en modules ancestraux. Deux paradoxes ont été mis en évidence dans ces contenus ancestraux.

Le premier concerne les modules absents chez un ancêtre donné, mais retrouvé dans au moins une architecture de modules inférée présente dans cet ancêtre. Cette situation concerne 949 ($0{,}48\%$) familles de protéines qui sont majoritairement anciennes. Elle s'explique en partie par la différence des jeux de paramètres qui ont été estimés indépendamment pour les familles de protéines et les

FIGURE 3.8 – **Évolution des répertoires des familles de modules Pfam dans les nœuds les plus anciens (A) et chez les Alphaprotéobactéries (B).** Pour la légende, voir figure 3.5, page 59.

familles de modules. En effet, pour certains profils phylogénétiques identiques, le jeu de paramètres des familles de protéines prédit un scénario différent de celui inféré avec le jeu de paramètres des familles de modules. Dans certains cas, l'origine des familles de protéines est plus ancienne que celle des familles de modules Pfam. Dans d'autres cas, les événements de regains des familles de modules, fréquents avec les données de Pfam, semblent également expliquer cette situation. En effet, dans 31% des cas, le module Pfam a été perdu au nœud considéré où dans l'un de ses ascendants et est regagné dans l'un de ses descendants, alors que la famille de protéines est présente dans l'ensemble de la lignée. D'un point de vue biologique, ces situations ne sont pas correctes, cependant, elles n'affectent que très peu le jeu de données protéiques et encore moins l'analyse de l'innovation (moins de 0,22%).

Le second paradoxe correspond aux cas où un module Pfam est présent dans une espèce alors qu'il est absent de l'ensemble des architectures inférées présentes. L'interprétation biologique de tels cas est intéressante puisque cette situation suggère que le domaine était présent dans une protéine ancestrale qui a été perdue dans les espèces contemporaines échantillonnées. Cependant,

Modules Pfam	Total	Procaryotes
Nb familles	6 659	4 389
↪ Nb familles innovées	4 080	2 066
↪ Nb familles transférées	2 579	2 323
Nb familles regagnées	904	863
Nb moyen de transferts par famille	3,9	4,2
Nb moyen de pertes par famille	5,0	7,2

TABLEAU 3.3 – **Caractérisation des différents événements pour les familles de modules Pfam.** Explications supplémentaires dans la légende du tableau 3.2, page 60.

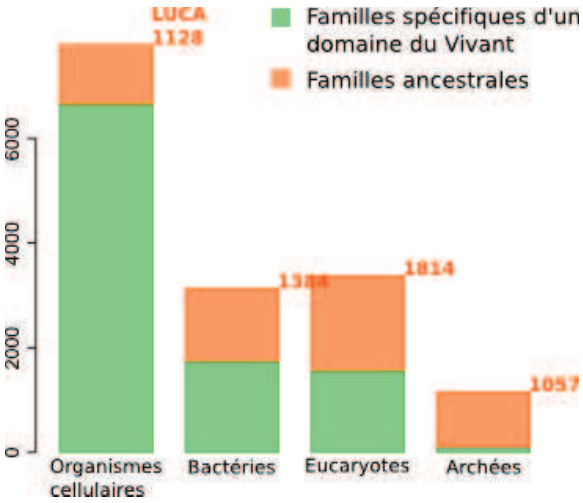


FIGURE 3.9 – **Répartitions des familles de modules Pfam dans les différents domaines du vivant.** La figure représente le nombre de familles Pfam spécifiques de chaque grand domaine du vivant (Eucaryotes, Bactéries et Archées). Parmi elles, les familles retrouvées dans l'ancêtre commun sont indiquées en rouge.

d'autres explications sont possibles comme dans la situation précédente, des différences dans les paramètres des modèles peuvent induire des scénarios différents, localement soutenus ou non par le modèle. Parmi les familles de modules Pfam, 2 685 familles différentes sont impliquées dans une telle situation, soit 40% des familles de modules. Dans une espèce ancestrale donnée, en moyenne 7,8% des modules Pfam présents n'appartiennent à aucune architecture inférée présente. Les modèles probabilistes associés aux scénarios inférés soutiennent localement la présence du module et l'absence de toutes les protéines dans lesquelles le module apparaît dans 35% des cas. La présence de ces modules pourrait donc être la conséquence d'une histoire évolutive dans laquelle la protéine ancestrale n'est pas retrouvée aujourd'hui.

Les modules particuliers ont été analysés en détail à LUCA, la racine de l'arbre (voir tableau 3.4 pour les effectifs). Très peu de modules sont soutenus par les modèles. Nous avons comparé leur versatilité, c'est-à-dire le nombre de contextes protéiques dans lesquels ils sont retrouvés, avec celle des modules présents dans au moins l'une des architectures de LUCA. Les distributions sont présentées dans la figure 3.10. Ces modules particuliers font majoritairement partie des modules les plus versatiles (retrouvés dans plus de 10 architectures différentes). Cela suggère une fréquence de recombinaison plus forte, ce qui étaye l'hypothèse qu'une protéine ancestrale ait existé. À la suite d'un événement de fusion ou de fission par exemple (événements fréquents [Pasek *et al.*, 2006]), celle-ci peut ne pas être retrouvée dans les espèces actuelles.

Nb modules présents à LUCA	1 128
Nb modules présents dans au moins une architecture	850
↪ dont soutenus	569
Nb modules absent des architectures	278
↪ dont soutenus	22

TABLEAU 3.4 – **Répartition des familles de modules Pfam présentes à LUCA.** Les familles sont divisées en deux classes : celles présentes dans au moins l'une des architectures présentes à LUCA et celles absentes de l'ensemble des architectures. L'état inféré d'une architecture ou d'un module est soutenu par le modèle lorsque la probabilité marginale associée est supérieure à 0,95. Le soutien des modules absents des architectures est avéré si la présence du module est soutenue, et que l'absence de l'ensemble des familles de protéines dans lesquelles il apparaît est également soutenue.

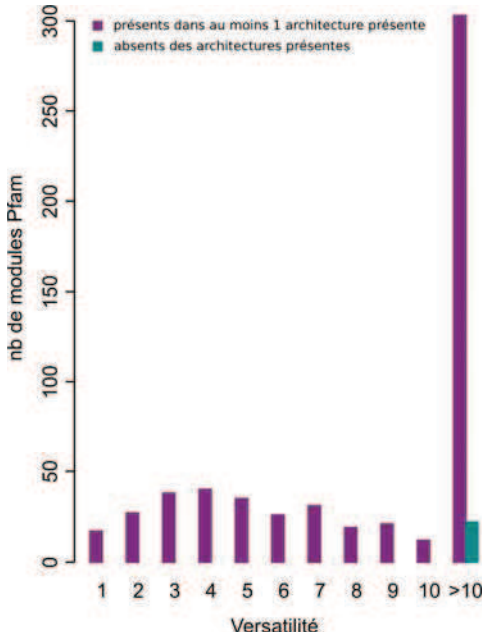


FIGURE 3.10 – **Versatilité des modules Pfam soutenus à LUCA.** La versatilité est comparée pour les modules présents dans au moins une architecture présente à LUCA (barres violettes) et les modules absents des architectures présentes à LUCA (barres cyan).

3.3.2.2 Réconciliation des scénarios de protéines et de modules

L'émergence de nouvelles protéines a été appréhendée à partir des scénarios d'évolution des familles de modules composant leur architecture. Ainsi, nous avons comparé les nouveaux arrangements en modules avec le répertoire parental en modules. Cela nous a permis de définir 3 types de nouvelles protéines. Les *innovations totales* regroupent les architectures composées exclusivement de nouveaux modules (figure 3.11a), c'est-à-dire absents du répertoire en modules parental. Les *réarrangements* regroupent les architectures composées exclusivement de domaines préexistants (figure 3.11b), c'est-à-dire présents dans le répertoire de l'espèce parentale, suggérant un réarrangement de ces modules pour former la nouvelle architecture. Enfin, les *innovations partielles* regroupent les architectures composées à la fois de modules préexistants et de nouveaux modules (figure 3.11c). Dans cette analyse, on considère qu'un module appartenant à la catégorie transférée horizontalement est préexistant puisque l'origine exacte du module n'est pas connue.

La répartition des innovations protéiques en fonction des architectures en modules Pfam présente une majorité de réarrangements, très peu d'innovation totale, les innovations partielles étant

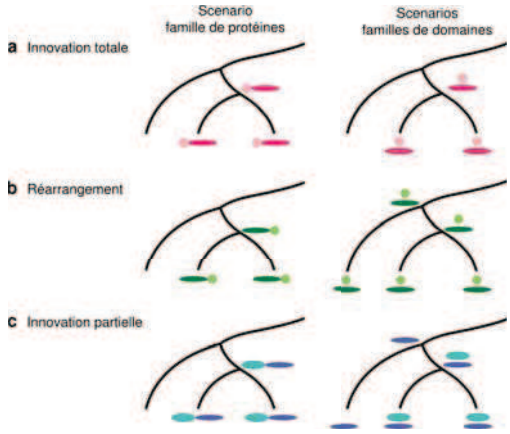


FIGURE 3.11 – **Réconciliation des scénarios d'évolution des familles de modules et des familles de protéines.** Ces exemples montrent les trois types d'innovation protéique possibles à un nœud donné de la phylogénie. (**a**) Innovation totale : la protéine et les modules constituant son architecture apparaissent pour la première fois à ce nœud. (**b**) Réarrangement : la nouvelle protéine provient de modules préexistants seulement. (**c**) Innovation partielle : la nouvelle protéine est une combinaison de nouveaux et d'anciens modules.

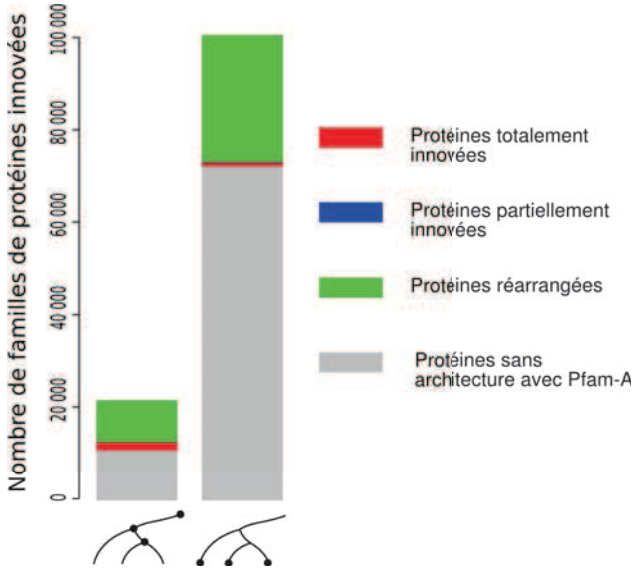


FIGURE 3.12 – **Distribution des différents types d'innovation protéique selon Pfam dans les espèces ancestrales et contemporaines.** Les protéines complètement innovées, partiellement innovées et réarrangées sont respectivement indiquées en rouge, bleu et vert sur la base des modules Pfam. Les familles de protéines ne contenant aucun domaine sont indiquées en gris. Les barres de gauche et de droite correspondent respectivement aux espèces procaryotes ancestrales et contemporaines, comme symbolisées sous chacune d'elles.

quasi inexistantes sur l'ensemble de l'arbre. Ces différents événements sont modélisés sur l'arbre phylogénétique des espèces où les longueurs de branches représentent les effectifs des différents types d'innovation (figures 3.13C-E) [1]. Ces résultats sont en accord avec le paradigme classique de

1. Les distributions des fréquences des différents types d'innovations protéiques par espèces sont présentées dans la figure B.1B, page 109.

l'évolution des protéines modulaires où les nouvelles protéines sont des combinaisons de domaines préexistants. L'arbre des innovations en modules Pfam (figure 3.13B) est extrêmement réduit par rapport à celui des innovations protéiques (figure 3.13A) appuyant l'hypothèse d'un répertoire en module restreint majoritairement ancien pour expliquer l'univers des protéines.

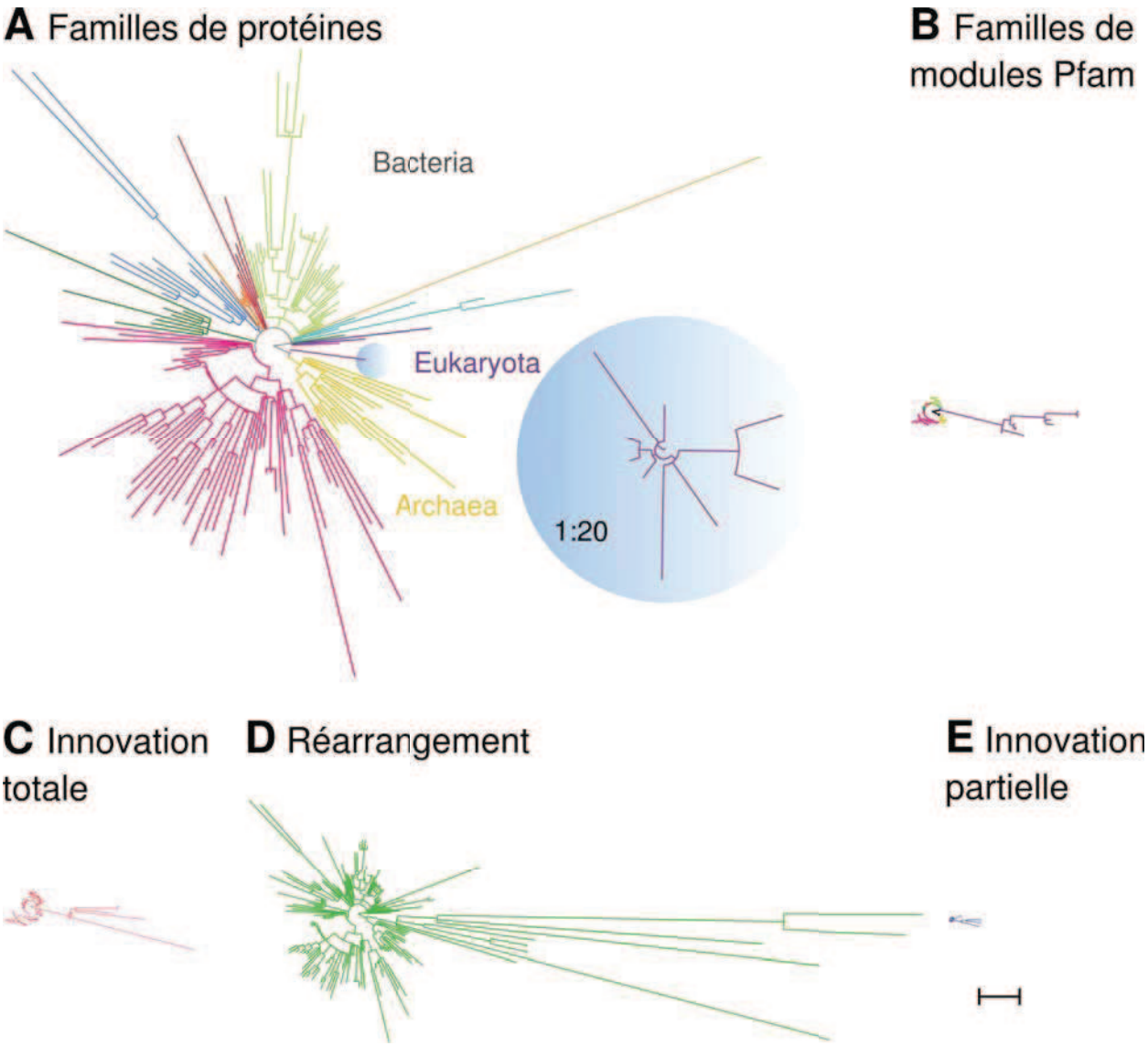


FIGURE 3.13 – **L'innovation protéique et modulaire selon Pfam le long de la phylogénie.** Dans ces arbres phylogénétiques, la longueur des branches représente le nombre de familles innovées de différents types : (A) familles de protéines d'HOGENOM ; (B) familles de modules protéiques de Pfam ; (C) protéines complètement nouvelles ; (D) protéines issues du réarrangement de modules préexistants ; (E) protéines partiellement innovées combinant des modules nouveaux et anciens. Le code couleur utilisé en (A) et (B) correspond aux clades suivants (sens inverse des aiguilles d'une montre) : Archées, jaune ; Eucaryotes, mauve ; Aquifex aeolicus, rose ; Deinococcus, violet ; Thermotoga maritima, vert ; Spirochaetales, bleu clair ; Rhodopirellula baltica, ocre ; Firmicutes, vert clair ; Bacteroidetes, marron ; Fusobacterium nucleatum, rouge ; Chlamydiales, orange ; Actinobacteria, bleu ; Cyanobacteria, vert foncé ; Proteobacteria, magenta. L'échelle du sous-arbre des Eucaryotes a été réduite du facteur indiqué pour améliorer la lisibilité en (A). Dans ces représentations, les arbres ont été racinés au milieu de la branche des bactéries. L'échelle correspond à 500 familles.

Les événements de réarrangement concernent 93% des familles de protéines ayant une architecture en modules Pfam. Cependant, ils ne représentent que 30% de l'ensemble des familles de protéines analysées, puisque 65% d'entre elles ne possèdent pas d'architecture en modules Pfam (figure 3.12). Beaucoup de familles de protéines sont négligées dans les résultats de Pfam : environ 50% des familles de protéines ancestrales et 72% des familles contemporaines n'ont aucune annotation avec les modules Pfam. Les conclusions obtenues avec Pfam négligent donc une part importante de la diversité protéique.

3.4 Les modules ProDom : un répertoire dynamique où l'innovation des modules est au cœur de l'évolution des protéines

3.4.1 Dynamique évolutive du répertoire de modules ProDom

Les répertoires en familles de modules ProDom présentent une dynamique proche de celle des répertoires des familles de protéines. Les innovations représentent la majorité des événements de gains et sont réparties sur l'ensemble de l'arbre où les événements récents sont plus nombreux que les événements anciens (figures 3.14). Seuls 11% des familles présentent des gains multiples, et en moyenne 2 transferts horizontaux sont inférés par famille. La répartition des innovations dans les trois domaines du vivant présente la même caractéristique que celle obtenue pour les familles de protéines. Une minorité des familles spécifiques des trois grands domaines du vivant est présente dans l'ancêtre de ces trois domaines, cela suggère que l'innovation en modules ProDom est un phénomène continu au cours de l'innovation (figure 3.15).

Le grand nombre d'innovations détectées par ProDom au cours de l'évolution est en opposition avec les prédictions du paradigme classique que l'on retrouve avec Pfam. Cela pourrait être une conséquence d'un manque de sensibilité dans la détection de l'homologie ancienne. Ce biais a été pris en compte au départ de ce projet et a conduit à un nouveau regroupement des familles de modules ProDom en augmentant la sensibilité de détection de l'homologie : le seuil de la E-value a été fixé à 10^{-2} contre 10^{-6} dans la construction de ProDom (section 3.1.1.2, page 49). Cependant, le critère utilisé pour la fraction minimale du module chevauchant la séquence protéique (> 80% de la longueur du module) ne permet pas de prendre en compte toutes les homologies locales. Dans le cadre de l'analyse de l'innovation, une homologie locale peut indiquer une probable origine commune plus ancienne. Par exemple, si un petit module est similaire sur toute sa longueur à une fraction d'un plus grand module, alors il est possible que l'un d'eux soit à l'origine de l'autre par des mécanismes de fusion/fission ou d'insertion/délétion par exemple. Dans ce cas, le module pourrait être plus ancien. Pour éviter ce type d'erreur, les modules présentant de la similarité locale

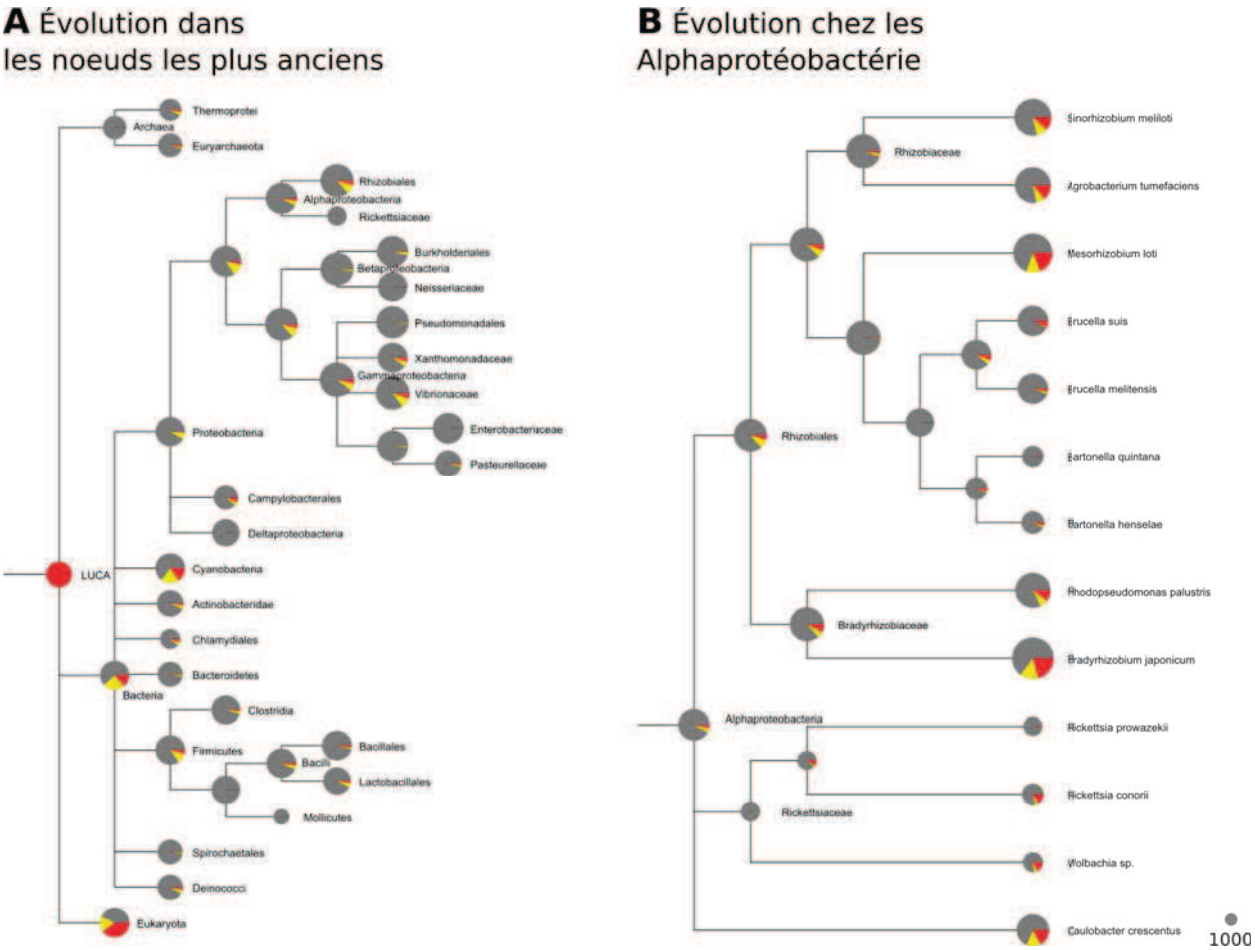


FIGURE 3.14 – **Évolution des répertoires des familles de modules ProDom dans les nœuds les plus anciens (A) et chez les Alphaprotéobactéries (B).** Pour la légende, voir figure 3.5, page 59.

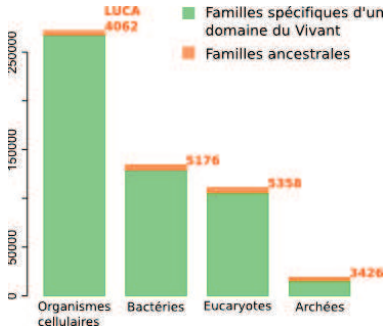


FIGURE 3.15 – **Répartitions des familles de modules ProDom dans les différents domaines du vivant.** La figure représente le nombre de familles ProDom spécifiques de chaque grand domaine du vivant (Eucaryotes, Bactéries et Archées). Parmi elles, les familles retrouvées dans l'ancêtre commun sont indiquées en rouge.

à l'extérieur du sous-arbre dans lequel ils sont prédits innovés sont considérés comme de faux positifs potentiels. La recherche des homologies locales a été effectuée de deux manières différentes. Dans la première, nous avons analysé les relations de similarité détectées par le PSI-BLAST des familles de protéines contre les familles ProDom. Un module ayant une correspondance locale

Modules ProDom	Total	Procaryotes
Nb familles	267 595	154 800
↪ Nb familles innovées	218 446	119 789
↪ Nb familles transférées	49 149	35 011
Nb familles regagnées	3 026	2 779
Nb moyen de transferts par famille	1,9	2,2
Nb moyen de perte par famille	0,7	1,2

TABLEAU 3.5 – **Caractérisation des différents événements pour les familles de modules ProDom.** Ces données ont été calculées sur les scénarios d'évolution obtenus. Les modules innovés présentant de l'homologie locale à l'extérieur du sous-arbre où ils sont innovés sont classés dans la catégorie transfert horizontal. Explications supplémentaires dans la légende du tableau 3.2 page 60.

avec une famille de protéine présente à l'extérieur du sous-arbre dans lequel il est innové n'est plus considéré comme innové. Dans la seconde, nous avons analysé les occurrences des familles de modules dans la base de données ProDom. De nouveau, si l'une des occurrences est détectée à l'extérieur du sous-arbre dans lequel le module est innové, alors il n'est plus considéré comme innové. D'un point de vue opérationnel, ces modules sont traités comme les modules gagnés plusieurs fois et rentrent dans la catégorie des transferts horizontaux. Sur les 232 688 innovations prédites par les scénarios (sans LUCA), seules 18 304 (7,9%) présentent de l'homologie locale à l'extérieur du sous-arbre où ils sont innovés (tableau 3.5).

Nous avons également étudié la complexité des séquences de modules innovés. Les séquences de faible complexité sont des séquences dont la composition en acides aminés est biaisée, pouvant conduire à une forte similarité sans homologie. La procédure de ProDom filtre les séquences de faible complexité qui peuvent être intégrées à un module. Mais les séquences de faible complexité de plus de 20 acides aminés sont décrites comme des modules singletons. Nous avons donc testé l'impact de ces séquences sur l'inférence de l'innovation récente. Nous avons calculé la proportion d'acides aminés filtrés pour chaque famille de modules par le programme SEG [Wootton et Federhen, 1993], qui recherche les régions de faible complexité, et le programme XNU [Claverie et States, 1993], qui recherche les répétitions en tandem d'un nombre quelconque d'acides aminés. Chaque famille de modules est représentée par la séquence consensus de l'alignement des modules de la base ProDom. Les résultats obtenus ont été comparés pour les familles de modules innovés récemment, soit 218 446 familles, et celles présentes à LUCA, soit 4 062 familles.

L'analyse avec XNU a mis en évidence des répétitions dans très peu de modules (386 modules, dont 343 récents). Il n'est cependant pas évident d'interpréter ces répétitions. Prenons l'exemple de PD002264, la séquence de ce module est composée de plusieurs répétitions d'un pentapeptide, dont la séquence est A(D/N)LXX. Ces répétitions en tandem sont retrouvées dans les 3 domaines du vivant (mais essentiellement chez les Cyanobactéries et les plantes). On les retrouve notamment

dans la protéine de résistance aux fluoroquinolones de *Mycobactérium tuberculosis* dont la structure tridimensionnelle a été déterminée par cristallographie aux rayons X et est présentée figure 3.16. Cette structure a la particularité de mimer une molécule d'ADN au niveau de sa taille, de sa forme et de ses propriétés électrostatiques. Cette protéine a la capacité de se fixer aux protéines gyrases à la place de l'ADN et de former un complexe inhibant l'action des fluoroquinolones. Dans cet exemple, PD002264 ne correspond pas à un domaine au sens usuel, c'est-à-dire, un domaine globulaire. Cette structure répétée est retrouvée dans de nombreux contextes protéiques différents, et correspond donc à ce que l'on appelle un module protéique : une unité évolutive. Cependant, les bornes de ce genre de module peuvent être délicates à définir. Ainsi, le module ProDom PD002264 contient 16 répétitions du pentapeptide alors que son homologue dans Pfam (PF00805) n'en contient que 8.

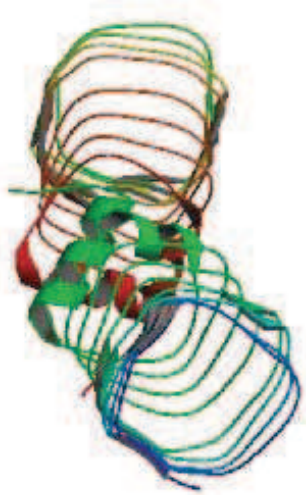


FIGURE 3.16 – **Structure tridimensionnelle de la protéine de résistance aux fluoroquinolones de *Mycobacterium tuberculosis*.** Cette protéine, dont l'identifiant dans la PDB (Protein Data Bank [Hegde *et al.*, 2005]) est 2BM4, est composée d'un motif de 5 acides aminés (pentapeptide) répété. Sa structure et sa taille ont la particularité de mimer celles de l'ADN (la période d'un tour est de 20 acides aminés). De plus, elle présente des particularités électrostatiques similaires à celle de la forme B de l'ADN (la plus commune).

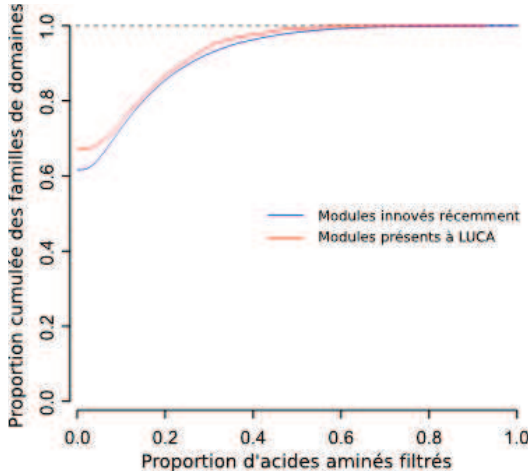


FIGURE 3.17 – **Distributions cumulées des proportions d'acides aminés filtrés par SEG.** Les proportions d'acides aminés filtrés ont été calculées pour les modules innovés récemment (courbe bleue) et ceux présents à LUCA (courbe rouge).

Quant à l'analyse avec SEG, qui recherche les régions de faible complexité, 93 408 modules (∼43%) ont été filtrés sur une portion représentant 7,9% des acides aminés en moyenne. La comparaison de la distribution des proportions d'acides aminés filtrés entre les modules récents et ceux présents à LUCA ne montre pas de différence significative entre les deux (figure 3.17). Cela suggère que les nouveaux modules ne comportent pas plus de régions de faible complexité que les modules anciens, et que ces régions n'expliquent pas la grande proportion d'innovation récente.

3.4.2 Origine des nouvelles protéines

3.4.2.1 Comparaison des répertoires d'architectures et de modules

La comparaison des répertoires en protéines et en modules ProDom a montré le même type de paradoxes que ceux présentés avec les modules Pfam. Le premier paradoxe, qui décrit l'absence du module chez un ancêtre donné, alors qu'il est retrouvé dans au moins une architecture présente, concerne 1 107 ($0{,}57\%$) familles de protéines avec les données de ProDom. Les trois quarts des cas se situent dans les espèces les plus anciennes de l'arbre. Cette situation s'explique majoritairement par la différence des jeux de paramètres entre les familles de protéines et les familles de modules. Dans plus de 99% des cas, la probabilité marginale associée à l'absence du module ou celle associée à la présence de la protéine est inférieure à $0{,}95$. Le second paradoxe correspond aux cas où un module ProDom est présent dans une espèce alors qu'il est absent de l'ensemble des architectures inférées présentes. On trouve 11 000 familles de modules différentes impliquées dans cette situation, soit $4{,}1\%$ du jeu de données total. Dans une espèce ancestrale donnée, en moyenne $8{,}8\%$ des modules ProDom présents sont absents de l'ensemble des architectures inférées présentes. Les modèles probabilistes associés aux scénarios inférés soutiennent localement la présence du module et l'absence de toutes les protéines dans lesquelles le module apparaît dans 28% des cas. L'analyse des modules présents chez LUCA montre que 14% des modules absents des architectures sont soutenus par le modèle (leur probabilité marginale est supérieure à $0{,}95$, tableau 3.6). Ces modules sont biaisés en faveur des modules les plus versatiles (figure 3.18), ils sont majoritairement présents dans plus de 10 contextes protéiques différents. On peut donc poser l'hypothèse qu'une protéine ancestrale, aujourd'hui perdue, ait existé comme dans l'analyse avec Pfam.

3.4.2.2 Réconciliation des scénarios de protéines et de modules

La répartition des innovations protéiques en fonction des architectures en modules ProDom présente 56% d'innovation totale contre seulement 20% de protéines issues d'un réarrangement de modules ProDom anciens. La figure 3.20 présente la répartition des types d'innovation en distinguant les protéines ayant émergé récemment dans une espèce contemporaine et celles plus anciennes. Les deux distributions sont similaires dans l'importance de l'innovation totale, bien qu'elle soit un peu plus importante dans les espèces contemporaines. Le grand nombre d'innovations inférées dans les espèces contemporaines peut s'expliquer de deux manières. La première concerne l'échantillonnage taxonomique restreint à 170 espèces. L'ajout d'autres espèces proches de celles analysées aurait pour conséquence de réduire l'innovation récente sans toutefois la supprimer totalement. En effet, notre échantillon possède certaines espèces très proches, voire différentes

Nb modules présents à LUCA	4 062
Nb modules présents dans au moins une architecture	2 730
↪ dont soutenus	2 422
Nb modules absent des architectures	1 332
↪ dont soutenus	182

TABLEAU 3.6 – **Répartition des familles de modules ProDom présentes à LUCA.** Les familles sont divisées en deux classes : celles présentes dans au moins l'une des architectures présentes à LUCA et celles absentes de l'ensemble des architectures. L'état inféré d'une architecture ou d'un module est soutenu par le modèle lorsque la probabilité marginale associée est supérieure à 0,95. Le soutien des modules absents des architectures est avéré si la présence du module est soutenue et que l'absence de l'ensemble des familles de protéines dans lesquelles il apparaît est également soutenue.

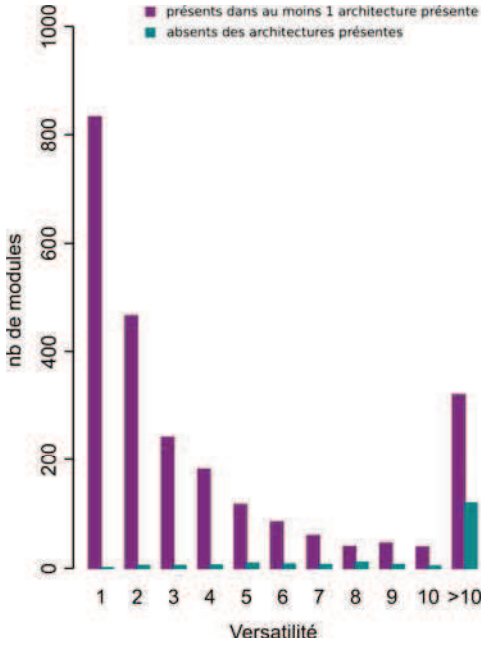


FIGURE 3.18 – **Versatilité des modules ProDom soutenus à LUCA.** La versatilité est comparée pour les modules présents dans au moins une architecture présente à LUCA (barres violettes) et les modules absents des architectures présentes à LUCA (barres cyan).

souches d'une même espèce qui présentent cependant une proportion d'innovation significative [1]. La seconde explication, biologique, postule un turn-over important de modules protéiques nouveaux dont une minorité seulement sera conservée au cours de l'évolution. Ainsi, les innovations ancestrales, moins nombreuses, correspondent aux innovations sélectionnées avec succès.

L'innovation protéique est un phénomène continu comme nous l'avons déjà montré lors de l'analyse de la dynamique des répertoires. Il est intéressant de voir que les trois types d'innovations se retrouvent sur l'ensemble de l'arbre, ils ne sont pas spécifiques de certains groupes d'espèces. La figure 3.19C-E présente les effectifs des différents types d'innovation le long de l'arbre des espèces [2]. La plupart des espèces présentent les trois types d'innovation, mais en proportions variables. L'innovation totale de protéine est majoritaire, en adéquation avec la dynamique d'innovations des modules ProDom (figure 3.19B).

1. Voir par exemple les répertoires des Gammaprotéobactéries page 132 ou des Firmicutes page 126
2. La figure B.1 page 109 présente les distributions des fréquences de ces événements dans chacun des répertoires.

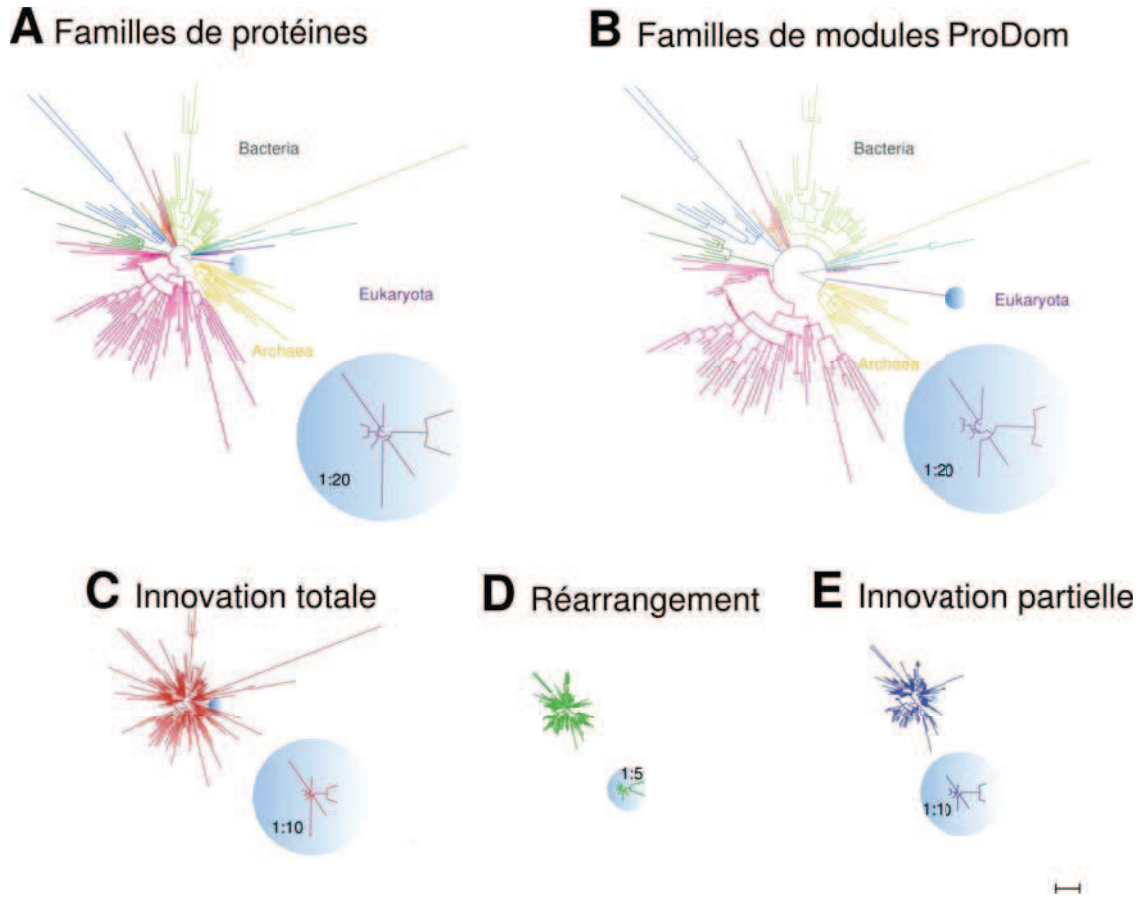


FIGURE 3.19 – **L'innovation protéique et modulaire selon ProDom le long de la phylogénie.** Dans ces arbres phylogénétiques, la longueur des branches représente le nombre de familles innovées de différents types : (A) familles de protéines d'HOGENOM ; (B) familles de modules protéiques de ProDom ; (C) protéines complètement nouvelles ; (D) protéines issues du réarrangement de modules préexistants ; (E) protéines partiellement innovées combinant des modules nouveaux et anciens. Voir la figure 3.13 page 67 pour les éléments de légende complémentaires.

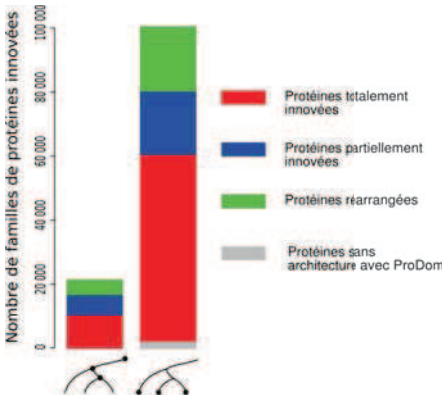


FIGURE 3.20 – **Distribution des différents types d'innovation protéique selon ProDom dans les espèces ancestrales et contemporaines.** Les protéines complètement innovées, partiellement innovées et réarrangées sont respectivement indiquées en rouge, bleu et vert sur la base des modules ProDom. Les familles de protéines ne contenant aucun module ProDom sont indiquées en gris. Dans chaque panneau, les barres de gauche et de droite correspondent respectivement aux espèces procaryotes ancestrales et contemporaines, comme symbolisées sous chaque barre.

3.5 Comparaison des prédictions de ProDom et Pfam

L'analyse des répertoires de modules ProDom et de modules Pfam a mis en évidence une vision différente de l'univers des modules protéiques. D'un côté, les données de ProDom montrent une dynamique évolutive dominée par les innovations ayant lieu tout au long de l'évolution. Cette dynamique est similaire à celle des familles de protéines – les fréquences des différents événements sont similaires sur l'ensemble de l'arbre. On retrouve des fréquences de gains plus élevées dans les espèces contemporaines par rapport aux espèces ancestrales (figure 3.21A et B). De l'autre côté, les données de Pfam décrivent un univers de modules majoritairement anciens où les événements de perte et de transfert sont à l'origine de l'évolution des répertoires. Les fréquences de gain sont globalement plus faibles et légèrement plus élevées dans les espèces ancestrales, ce qui correspond à un univers de modules anciens. Les fréquences de perte sont similaires entre Pfam et ProDom, avec les mêmes variations le long de l'arbre. Les distributions des fréquences de perte (figure 3.21, colonne de droite) présentent la même asymétrie, quels que soient les espèces ou le jeu de données considérés.

La faible couverture de l'univers des protéines par les modules Pfam (39% des familles et 21% des résidus) associée au fait que la majorité des modules soient anciens suggère un biais dans l'échantillonnage des familles de modules Pfam. Ce biais est en faveur des familles anciennes, mais également en faveur des familles dispersées d'un point de vue taxonomique. Ceci est étayé par les nombreux événements de gains multiples et de regain. Les scénarios ne prédisent pas une origine unique pour la plupart des familles Pfam, contrairement à ceux des familles ProDom. La composition moyenne des répertoires actuels (tableau 3.7) précise ce biais. En effet, les modules Pfam récents sont sous-représentés avec moins de 1% des répertoires actuels, alors que les répertoires de ProDom proposent en moyenne 14% de nouveauté et les répertoires de protéines 26% de nouvelles protéines. Les modules très anciens (présents à LUCA) sont surreprésentés dans Pfam.

	Récent	Intermédiaire	Ancien (LUCA)
Familles de protéines	26 %	50 %	24 %
Familles de modules ProDom	14 %	45 %	41 %
Familles de modules Pfam	0,4 %	35 %	64 %

TABLEAU 3.7 – **Ancienneté des familles représentées dans les organismes procaryotes.** Les fréquences sont exprimées en pourcentage de familles de protéines ou de familles de modules trouvées dans les espèces contemporaines qui sont les plus récentes (innovées dans une espèce contemporaine), intermédiaires (innovées dans une espèce ancestrale) ou qui remontent à LUCA. Les fréquences sont une moyenne sur les 161 espèces procaryotes. Les distributions complètes sont disponibles en annexe B.2, page 110.

L'analyse de l'émergence de nouvelles architectures au cours de l'évolution a également montré une grande disparité entre les prédictions de Pfam et celles de ProDom. En effet, d'un côté les

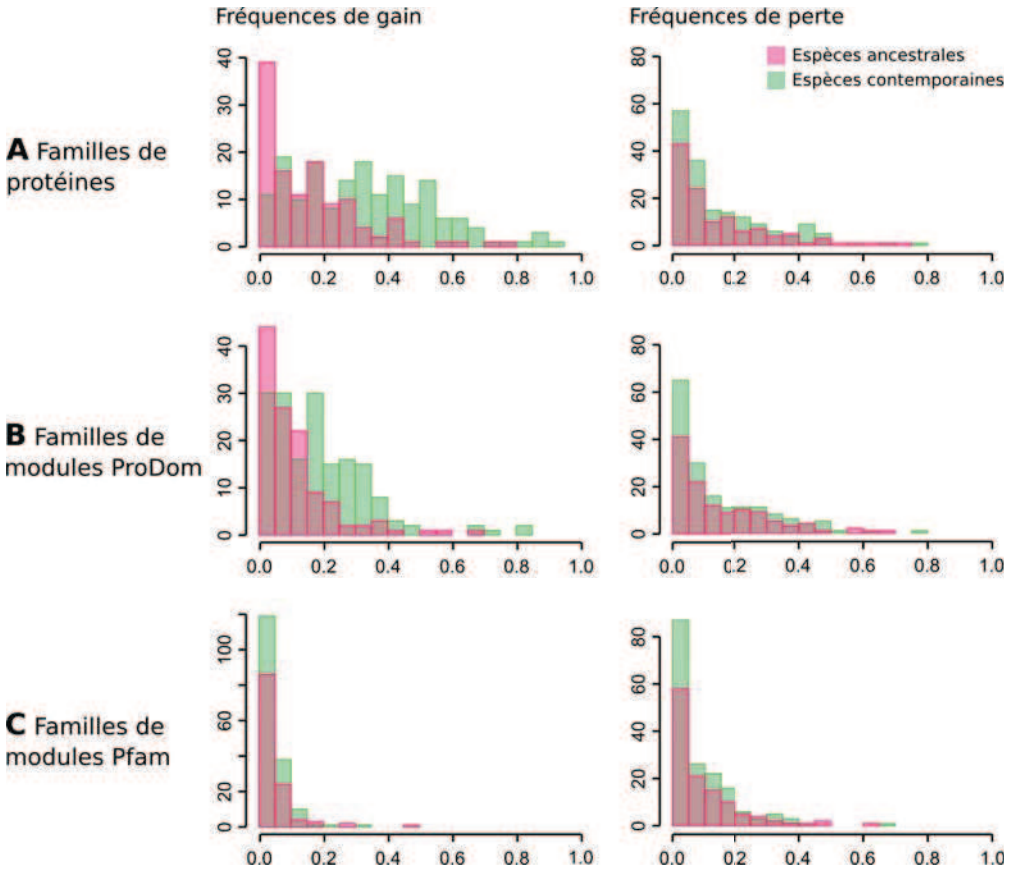


FIGURE 3.21 – **Distributions des fréquences de gain et de perte dans les espèces actuelles et ancestrales.** Les fréquences ont été calculées sur les 170 espèces actuelles et les 150 espèces ancestrales (LUCA n'est pas pris en compte) pour les familles de protéines (A), les familles de modules ProDom (B) et les familles Pfam (C). La fréquence de gain correspond au nombre de familles gagnées (par innovation, transfert horizontal ou regain) dans le répertoire de chaque espèce. La fréquence de perte correspond à la proportion de familles perdues relativement à la taille du répertoire parental.

données de Pfam corroborent le paradigme classique avec une majorité de réarrangements de modules anciens expliquant les nouvelles protéines, mais elles négligent une fraction importante de l'univers des protéines. De l'autre, les données de ProDom prédisent une majorité d'innovations totales. Cependant, la méthodologie de ProDom est moins sensible dans la détection des homologies anciennes et donc certaines innovations pourraient être plus anciennes (ceci est analysé et discuté dans la section suivante). Dans la suite, nous allons comparer plus précisément les prédictions de ProDom et celles de Pfam.

3.5.1 Les nouvelles architectures prédites réarrangées selon Pfam

Nous avons comparé les prédictions des architectures en modules ProDom sur les familles de protéines prédites réarrangées selon Pfam. Ces familles sont majoritairement prédites réarrangées ou innovées partiellement selon ProDom dans 82% des innovations anciennes et 76% des innova-

tions récentes (les effectifs sont présentés dans le tableau 3.8). Celles qui sont prédites innovées totalement sont a priori des prédictions erronées de ProDom puisque Pfam reconnaît une portion de la protéine comme ancienne. Dans le cas des familles prédites innovées partiellement par ProDom, la couverture des séquences est plus grande que celle de Pfam : en moyenne, la couverture de ces familles est de 80% avec ProDom et 56% avec Pfam. La figure 3.22A précise cette couverture en distinguant les acides aminés recouverts à la fois d'un module ProDom et d'un module Pfam, et ceux recouverts seulement par l'un d'eux. En moyenne, 30% de la séquence est recouverte uniquement par des modules ProDom : le découpage de ProDom apporte donc une information supplémentaire non négligeable qui peut permettre de mieux comprendre l'origine d'une protéine. La modularité moyenne des protéines (figure 3.22B), nettement supérieure avec ProDom (2,5 contre 1,4 avec Pfam), met aussi en évidence l'apport d'information par des modules supplémentaires de ProDom.

		Innovation totale	Innovation partielle	Réarrangement	Sans annotation
Espèces ancestrales	Nb familles	1 534	4 658	2 612	111
	Proportion	17%	52%	30%	1%
Espèces contemporaines	Nb familles	6 000	11 698	8 991	729
	Proportion	22%	43%	33%	2%

TABLEAU 3.8 – **Classification selon ProDom des innovations protéiques prédites réarrangées selon Pfam.** Les données de Pfam prédisent 8 915 (resp. 27 418) familles de protéines réarrangées dans les espèces ancestrales (resp. contemporaines). Le tableau présente la répartition de ces familles en fonction des modules ProDom. Par exemple, parmi les 8 915 familles prédites réarrangées selon Pfam, 4 658 sont partiellement innovées selon les données de ProDom.

La comparaison des prédictions de ProDom sur les familles prédites partiellement innovées selon Pfam montre que plus de 60% des familles sont également partiellement innovées selon ProDom. Mais cette analyse porte sur à peine 300 familles (les résultats sont présentés dans la figure B.2 page 112 en annexe). Quant aux familles prédites innovées totalement selon Pfam, elles le sont également majoritairement avec ProDom (83% des innovations protéiques, voir tableau B.3 page 113).

En conclusion, les prédictions obtenues avec les deux jeux de données de modules semblent complémentaires plutôt que contradictoires. Les différences observées s'expliquent en partie par le fait que le regroupement de ProDom peut manquer de sensibilité dans la détection de l'homologie, mais surtout par la différence de couverture des séquences. ProDom permet ainsi d'analyser l'origine de l'ensemble de la séquence alors que Pfam n'en décrypte en général qu'une partie.

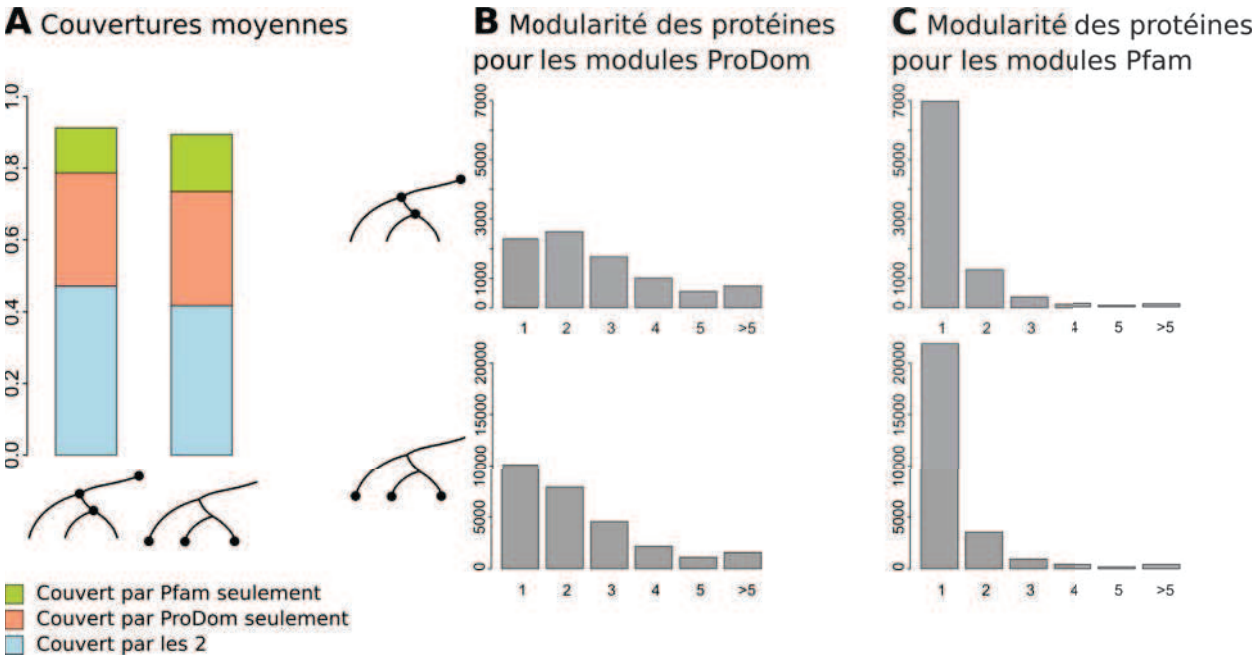


FIGURE 3.22 – **Couverture et modularité des familles de protéines réarrangées selon Pfam.** (**A**) La couverture représente la proportion moyenne d'acides aminés recouverts par seulement un module Pfam (vert), seulement un module ProDom (orange) et par les deux types de modules (bleu). La couverture a été calculée sur l'ensemble des familles de protéines réarrangées selon Pfam dans une espèce ancestrale (barre de gauche) ou contemporaine (barre de droite). (**B,C**) Les distributions de la modularité des protéines réarrangées selon Pfam dans une espèce ancestrale ou contemporaine (symbolisées à gauche des distributions) ont été calculées sur les architectures de modules ProDom (**B**) et les architectures de modules Pfam (**C**).

3.5.2 Comparaison des familles avec et sans architectures de modules Pfam

Une deuxième manière de comparer les prédictions de ProDom et Pfam est de diviser l'univers des familles de protéines en deux groupes : les familles avec et sans architecture en modules Pfam. En effet, la couverture en modules Pfam des familles de protéines procaryotes est de seulement 35%. Ainsi, il reste 65% de l'univers des protéines non pris en compte. Nous avons donc comparé les prédictions de ProDom pour les familles avec et sans annotations Pfam (figure 3.23).

Les distributions obtenues sont nettement différentes. Les innovations totales sont un événement minoritaire pour les familles ayant une annotation Pfam (28% pour les familles de protéines anciennes, figure 3.23B), alors qu'il devient majoritaire pour les familles anciennes sans annotations Pfam (plus de 66%, figure 3.23A). Les distributions mettent également en évidence un biais en faveur des familles de protéines anciennes dans l'annotation Pfam. En effet, Pfam reconnaît 50% des familles de protéines anciennes, mais seulement 28% des familles de protéines contemporaines. Le tableau 3.9 précise ces proportions en distinguant 5 catégories de nœuds, des plus anciens aux plus récents. Ainsi, 99% des familles de protéines à LUCA sont annotées avec Pfam, et plus on se rapproche des espèces contemporaines, plus cette proportion diminue. En revanche avec les données de ProDom, quel que soit l'âge des familles de protéines, plus de 98% d'entre

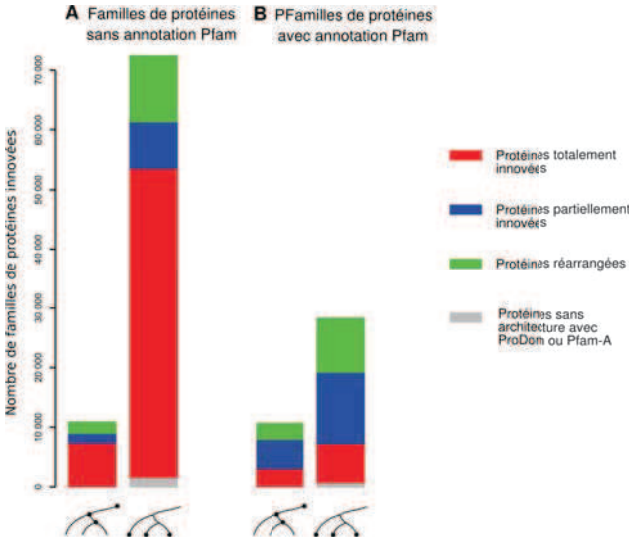


FIGURE 3.23 – **Distribution des différents types d'innovation protéique selon ProDom.** Les protéines complètement innovées, partiellement innovées, et réarrangées sont respectivement indiquées en rouge, bleu et vert sur la base des modules ProDom pour les familles sans (A) ou avec (B) annotations Pfam. Les familles de protéines ne contenant aucun module ProDom sont indiquées en gris. Dans chaque panneau, les barres de gauche et de droite correspondent respectivement aux espèces procaryotes ancestrales et contemporaines, comme symbolisées sous chaque barre.

elles sont annotées. Ainsi, le biais d'échantillonnage des modules Pfam entraîne un biais d'échantillonnage de l'univers des protéines dans lequel les familles de protéines anciennes sont mieux représentées que les familles récentes.

Familles de protéines non annotées	LUCA	Très anciens	Anciens	Récents	Contemporains
ProDom	0,2%	0,6%	1%	1%	2%
Pfam	1%	4%	43%	58%	72%

TABLEAU 3.9 – **Répartition des familles de protéines sans annotations avec ProDom et Pfam le long de l'arbre des espèces.** La proportion des familles de protéines innovées sans architecture en modules ProDom ou Pfam est exprimée en pourcentage des familles innovées dans chacune des 5 catégories de nœuds. Les nœuds de l'arbre ont été répartis en 5 grands groupes d'âges relatifs différents : LUCA, très anciens (les ancêtres des Bactéries et des Archées), anciens (le sous-arbre contient au moins 5 espèces contemporaines, 42 ancêtres), récents (le sous-arbre contient moins de 5 espèces, 69 ancêtres) et contemporains (170 espèces).

3.6 L'innovation de modules : réalité biologique ou artefact ?

L'innovation de modules est un processus qui semble majeur au cours de l'évolution selon ProDom, contrairement à Pfam où la majorité des innovations ont eu lieu au temps de LUCA. La faible couverture de l'univers des protéines et une majorité de modules anciens laissent supposer un biais

d'échantillonnage important des modules Pfam. Cependant, le grand nombre d'innovations détecté par ProDom pourrait aussi être une conséquence d'un manque de sensibilité dans la détection de l'homologie ancienne. Le manque de sensibilité de ProDom a été pris en compte au départ de ce projet et a conduit à un nouveau regroupement des familles de modules ProDom. De plus, les modules innovés pour lesquels de l'homologie locale a été détectée à l'extérieur du sous-arbre dans lequel ils sont innovés ont été systématiquement considérés comme "anciens". Cette section présente les différents tests menés dans le but de tester la sensibilité des regroupements de ProDom. Dans un premier temps, l'impact des vitesses d'évolution des modules ProDom a été analysé. Puis, le regroupement de ProDom a été comparé à celui de Pfam afin de caractériser plus précisément les biais de ces deux bases de données. Enfin, le succès évolutif de ces nouveaux modules a été évalué.

3.6.1 Analyse des taux d'évolution des modules ProDom

Lorsque le taux d'évolution d'une famille de modules est élevé, l'accumulation de mutations peut rendre difficile la détection de l'ensemble de ses homologues. Cela a pour conséquence une mauvaise estimation de l'âge de cette famille (qui sera plus récente) ou de la subdiviser en plusieurs sous-familles, et donc de surestimer le nombre de modules innovés. Lorsqu'une famille de modules est inférée innovée dans l'ancêtre commun aux espèces A et B (voir le schéma de la figure 3.24), deux explications sont possibles : soit le module est véritablement innové à ce nœud, soit il est présent dans l'ancêtre (ABC) et le signal d'homologie a disparu dans les branches menant à l'espèce C et à l'ancêtre (AB) (en rouge dans la figure 3.24) du fait d'un taux d'évolution trop élevé. Dans le cas d'une accélération brutale du taux d'évolution dans la branche menant de l'ancêtre (ABC) à l'ancêtre (AB) suite à un changement des contraintes sur la protéine, on considère que le module est innové dans l'ancêtre (AB).

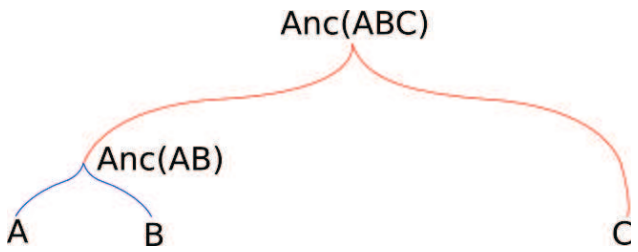


FIGURE 3.24 – Schéma représentant un arbre avec 3 espèces (A, B et C) et deux de leur ancêtres : $Anc(AB)$ et $Anc(ABC)$.

Pour tester les vitesses d'évolution des modules, nous avons comparé pour différents couples d'espèces, les distributions d'un score de similarité des modules innovés dans l'ancêtre de la paire d'espèces considérée et ceux présents à LUCA. En effet, ces derniers ont une vitesse d'évolution suffisamment faible pour que l'homologie ancienne ait été détectée. Les scores de similarité ont

été calculés pour les modules homologues trouvés dans plusieurs paires d'espèces proches : *Methanosarcina mazei* vs. *Methanosarcina acetivorans*, *Mycobacterium avium* vs. *Mycobacterium tuberculosis*, *Bacillus cereus ATCC10987* vs. *Bacillus anthracis* et *Escherichia coli K12* vs. *Salmonella typhi* ATCC700931. Les paires de protéines homologues ont d'abord été détectées à l'aide du programme TreePattern [Dufayard *et al.*, 2005]. Cet outil permet de rechercher un motif dans une banque d'arbres phylogénétiques, ici les arbres des familles de protéines d'HOGENOM. Le motif correspond à une relation de spéciation entre deux séquences du couple d'espèces d'intérêt. Les familles pour lesquelles plusieurs motifs ont été trouvés ou pour lesquelles une duplication d'une des séquences a été détectée après l'événement de spéciation sont retirées des résultats. Pour chaque famille de protéines, le découpage en modules ProDom des séquences protéiques d'intérêt a été effectué à partir de leur position sur l'alignement de la famille. Les gaps communs aux deux séquences, ainsi que les gaps externes sont éliminés de l'alignement. La distance évolutive entre deux modules est celle de Jones-Taylor-Thornton (JTT) [Jones *et al.*, 1992] et a été calculée avec ProtDist [Felsenstein, 1989] sans loi Gamma. Cette distance donne le nombre attendu de substitutions par site. Le nombre k de substitutions par site suit un processus Poissonnien $\mathcal{P}(\lambda)$ de paramètre $\lambda = JTT$:

$$\begin{aligned} P(k|site) &= exp^{-JTT} \times \frac{JTT^k}{k!} \\ P(0|site) &= exp^{-JTT} \end{aligned} \tag{3.2}$$

On déduit de l'équation 3.2 un score de similarité : $l \times exp^{-JTT}$, représentant le nombre attendu de résidus identiques pour un alignement de longueur l. Ce score est en relation directe avec la sensibilité de détection des séquences homologues.

La figure 3.25 présente les distributions des scores de similarité calculés sur les modules présents chez *Escherichia coli* et *Salmonella typhi* et innovés dans leur ancêtre commun le plus récent et ceux remontant à LUCA. Ces derniers évoluent plus lentement, ce qui est attendu, mais les distributions se chevauchent fortement. Plus de 78% des modules récents ont un score supérieur aux 5% des modules les plus rapides détectés à LUCA. Cette proportion augmente avec la profondeur dans l'arbre des ancêtres considérés : plus on remonte dans l'arbre, plus on est confiant dans l'innovation. Des résultats similaires sont obtenus avec chacun des couples d'espèces analysés (les résultats sont présentés dans le tableau B.1 en annexe page 111). En conséquence, bien qu'une fraction des innovations puisse être artefactuelle, la majorité de ces événements ne peut pas être expliquée par des taux d'évolution constitutifs élevés.

3.6.2 Analyse du regroupement de ProDom par rapport à celui de Pfam

Une autre manière d'appréhender le regroupement des familles de ProDom est de le comparer à celui de Pfam. En effet, les modules Pfam sont expertisés manuellement et des HMMs (Modèles de

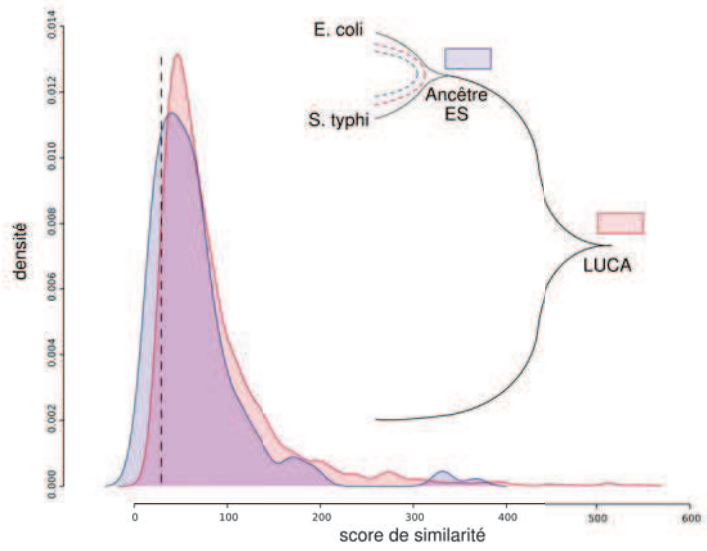


FIGURE 3.25 – **Distributions des scores de similarité de modules ProDom récents et anciens.** Les scores de similarité ont été calculés à partir des alignements de paires de modules orthologues chez *Escherichia coli* et *Salmonella typhi* (voir le tableau B.1 page 111 pour les autres comparaisons et paires d'espèces). La distribution des scores pour les modules relativement récents (inférés innovés chez leur ancêtre commun) est présentée en bleu, alors que celle des modules anciens qui remontent jusqu'à LUCA est présentée en rouge. La ligne en pointillé correspond au quantile à 5% des scores de similarité des modules anciens.

Markov cachés) sont utilisés pour retrouver l'ensemble des modules homologues. Cette méthode est de manière générale plus sensible pour la détection des homologies distantes. En principe, on s'attendrait à ce que les modules Pfam représentent un sous-ensemble des modules ProDom, c'est-à-dire qu'à chaque module Pfam correspond un seul et unique module ProDom. Dans cette section, nous cherchons donc à caractériser les correspondances entre familles de ProDom et de Pfam.

3.6.2.1 Association entre les modules ProDom et Pfam

L'association entre ProDom et Pfam a été établie à partir des architectures des familles de protéines pour lesquelles les positions des modules ProDom et Pfam sont connues. Un module ProDom est associé à un module Pfam s'ils se chevauchent sur au moins 20 acides aminés sur une famille de protéines. L'association entre un module ProDom et Pfam peut être seulement locale (seule une fraction des modules correspond). Nous avons donc également considéré les associations globales, dans lesquelles la longueur du chevauchement est supérieure à 80% de la longueur du module ProDom.

Ces associations permettent de mettre en relation 62 410 modules ProDom (32% du jeu de données total) avec 6 580 familles Pfam (99%). La figure 3.26 présente la distribution du nombre de familles ProDom associées à chaque famille Pfam. Seuls 24% des familles Pfam sont associées à

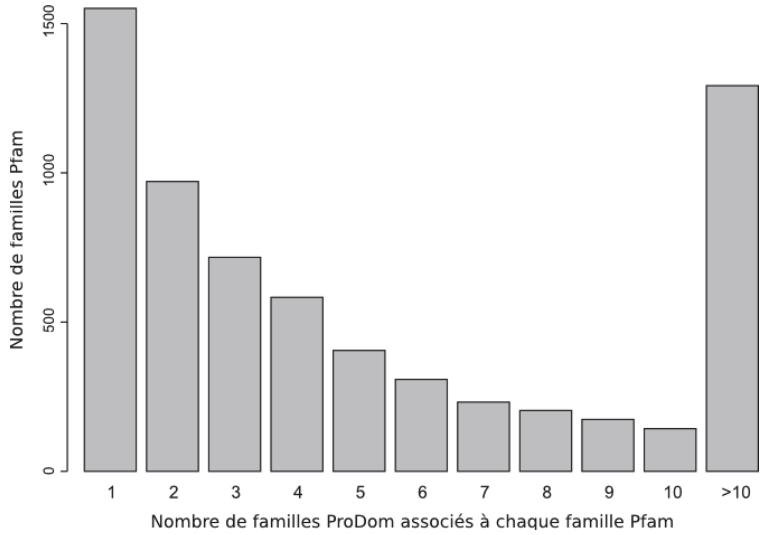


FIGURE 3.26 – **Association entre ProDom et Pfam.** Distribution du nombre de modules ProDom associés à chaque famille de modules Pfam.

un unique module ProDom, et chaque module Pfam est associé en moyenne à 10 familles ProDom (la médiane étant de 4 modules ProDom). Cette distribution présente les mêmes caractéristiques lorsque les jeux de données sont restreints aux modules procaryotes (tableau 3.10) sur lesquels nous allons nous concentrer. Le grand nombre de modules ProDom associés à chaque famille Pfam peut suggérer un manque de sensibilité dans la détection de l'homologie ancienne. Cependant, lorsque l'on considère une correspondance globale, on divise par 2 le nombre moyen de modules ProDom associés à chaque module Pfam. La correspondance entre les modules ProDom et Pfam sur les séquences consensus des familles de protéines est complexe et a été analysée à partir de trois indices décrits ci-dessous.

Notations :

j : indice des familles Pfam
$k^{(j)}$: indice des familles ProDom associées au module Pfam j
$n(k^{(j)})$: nombre de familles ProDom associées au module Pfam j
$n(j,k^{(j)})$: nombre de chevauchements entre j et $k^{(j)}$
$n(j)$: nombre d'occurrences de j
$lg(k^{(j)})$: somme des longueurs du module ProDom $k^{(j)}$ sur les séquences consensus des familles de protéines où il est associé à j
$lg(j,k^{(j)})$: somme des longueurs des chevauchements entre j et $k^{(j)}$ sur l'ensemble des séquences consensus des familles de protéines (seuls les chevauchements $\geq$ 20 acides aminés sont pris en compte)
$lg(j)$: somme des longueurs du module Pfam j sur les séquences consensus des familles de protéines

La couverture d'un module Pfam par ProDom : cet indice reflète la subdivision des familles de modules ProDom lorsque plusieurs familles sont homologues selon Pfam. Dans la figure 3.27, les modules PD1 et PD2 sont associés globalement au même module Pfam (PF1), ce qui indique que le regroupement de ProDom n'est pas suffisant. Cette couverture donne le nombre moyen de

modules ProDom par position du module Pfam :

$$couverture(j) = \frac{n(j)}{lg(j)} \times \sum_{k^{(j)}} \frac{lg(j,k^{(j)})}{n(j,k^{(j)})}$$

FIGURE 3.27 – Superposition des architectures de modules ProDom (PD1 et PD2) et Pfam (PF1) sur les séquences consensus de deux familles de protéines HBG1 et HBG2.

La subdivision d'un module Pfam en modules ProDom : le module Pfam peut être découpé en plusieurs modules différents. Dans la figure 3.28, les modules PD3, PD4 et PD5 n'ont pas à être regroupés en une seule et même famille ProDom puisqu'ils représentent des modules différents. Cette subdivision indique que le point de vue de ProDom est différent de celui de Pfam dans la caractérisation des modules. Cet indice donne le nombre moyen de modules ProDom subdivisant un module Pfam.

$$subdivision(j) = \frac{n(k^{(j)})}{couverture(j)}$$

FIGURE 3.28 – Superposition des architectures de modules ProDom (PD3, PD4 et PD5) et Pfam (PF2) sur la séquence consensus d'une famille de protéines (HBG3).

La couverture d'un module ProDom par un module Pfam : pour chaque couple $(j,k^{(j)})$, la proportion du module ProDom $k^{(j)}$ chevauchant le module Pfam j est calculée grâce à l'indice suivant :

$$couverture(k^{(j)}) = \frac{lg(j,k^{(j)})}{lg(k^{(j)})}$$

Du point de vue des modules Pfam, deux phénomènes se combinent pour expliquer l'association de plusieurs modules ProDom à chaque module Pfam. Le premier décrit la subdivision des familles de ProDom en sous-familles, c'est-à-dire que plusieurs familles ProDom représentent le même module selon Pfam. La figure 3.29A présente la distribution de la couverture des modules Pfam par ProDom. La couverture moyenne de chaque module Pfam est de $4{,}4$ modules ProDom (la médiane étant de $1{,}7$), mais plus de 34% des modules Pfam ont une couverture d'au plus un module ProDom. Parmi elles, 18% des familles Pfam sont associées à plus d'un module ProDom. Cette observation sous-tend que différentes familles ProDom peuvent correspondre à différentes portions des modules Pfam, ce qui correspond au second phénomène. Celui-ci représente le découpage des modules Pfam en plusieurs modules ProDom qui ne modélisent pas le même module, mais des modules différents. La distribution de la subdivision des modules Pfam est présentée dans la figure 3.29B. En moyenne, un module Pfam est découpé en deux modules ProDom distincts et

	Toutes les associations	Procaryotes	Procaryotes et couverture (ProDom) > 0.8
Nb modules ProDom	62 410	38 100	21 876
Nb modules Pfam	6 580	4 513	4 019
Nb moyen de modules ProDom par module Pfam	9,9	8,9	5,6
Couverture moyenne des modules Pfam	5,0	4,4	2,6
Subdivision moyenne des modules Pfam	2,0	1,9	1,9
Couverture moyenne des modules ProDom	73%	75%	96%

TABLEAU 3.10 – **Caractéristiques de l'association entre ProDom et Pfam.** Les résultats sont présentés pour l'ensemble des modules ProDom associés à un module Pfam, les modules ProDom procaryotes, et enfin les modules ProDom procaryotes dont le chevauchement avec le module Pfam est supérieur à 80% de sa longueur. La couverture moyenne des modules Pfam donne le nombre moyen de modules ProDom par position sur le module Pfam. La subdivision moyenne donne le nombre moyen de modules ProDom redécoupant un module Pfam. Enfin, la couverture des modules ProDom donne la proportion du module chevauchant le module Pfam auquel il est associé.

on trouve seulement 14% de modules Pfam non subdivisés. La correspondance entre ProDom et Pfam est globale (plus de 80% du module ProDom associé à Pfam) pour 57% des modules ProDom concernés. De nombreux modules ProDom sont donc associés à un module Pfam localement (figure 3.29C).

L'ensemble de ces résultats suggère qu'une partie des familles de modules ProDom est encore subdivisée. Cette subdivision provient en partie d'un manque de sensibilité dans la détection des modules homologues, mais également en partie d'un découpage des protéines en unités évolutives différentes de celles de Pfam (expliquant les nombreuses homologies locales détectées). Le manque de sensibilité de ProDom peut donc dans certains cas conduire à la surestimation du nombre d'innovations, mais également à un rajeunissement artefactuel de leur innovation. Dans le but de déterminer la marge d'erreur possible pour les innovations inférées par ProDom, nous avons comparé l'âge relatif des nœuds dans lesquels chaque couple de modules ProDom et Pfam associés est apparu.

3.6.2.2 Comparaison des âges relatifs d'apparition des modules ProDom et Pfam associés

Dans cette analyse, seuls les modules ProDom et Pfam innovés sont pris en compte puisque dans le cas de gains multiples, l'origine n'est pas connue. Les modules ProDom spécifiques des eucaryotes ont également été retirés de l'analyse. Le jeu de données est donc restreint à 14 542 modules ProDom associés à 1 871 modules Pfam différents. La comparaison des âges d'apparition revient à comparer la profondeur des nœuds dans l'arbre (voir figure 3.30). Ainsi, les origines sont identiques lorsque les deux modules sont apparus dans le même nœud. L'origine du module

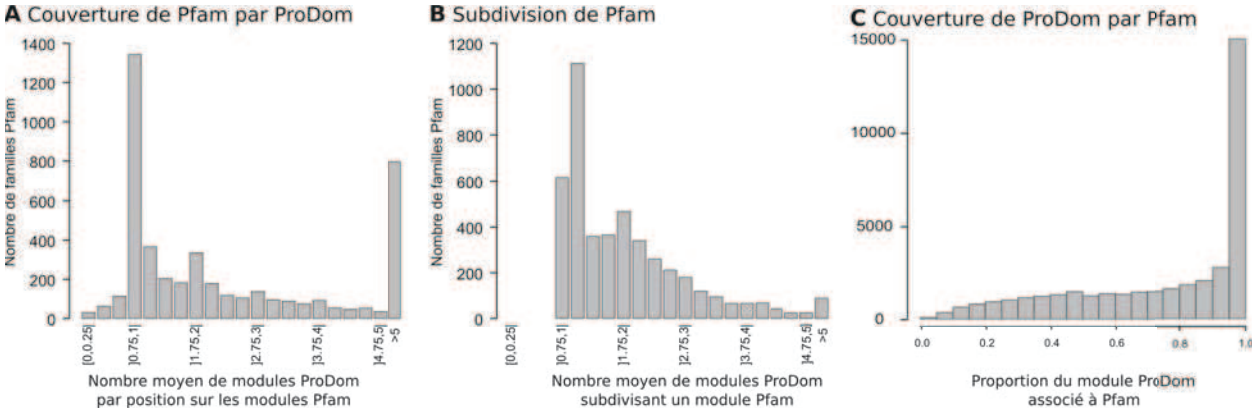


FIGURE 3.29 – **Analyse du chevauchement entre les modules ProDom et Pfam.** Le sous-ensemble des modules ProDom procaryotes a été utilisé. Plusieurs modules ProDom peuvent être associés à un unique module Pfam, ce qui peut être expliqué par une combinaison de deux phénomènes : un manque de regroupement des familles de ProDom représentant le même module (**A**), et le découpage du module Pfam en différents modules ProDom(**B**). Le panneau **A** est une distribution de l'indice $couverture(j)$ (pour tous les modules Pfam j) calculé comme le rapport de la somme des chevauchements entre ProDom et Pfam et la longueur moyenne du module Pfam. Le panneau **B** représente la distribution de l'indice $subdivision(j)$ calculé comme le rapport du nombre de modules ProDom associés à chaque module Pfam et de la couverture de ce dernier. Le panneau **C** donne la distribution de la couverture des modules ProDom ($couverture(k)$) calculée comme la proportion du module associée à Pfam.

ProDom (resp. Pfam) est plus récente lorsque le nœud dans lequel il est apparu appartient au sous-arbre raciné par le nœud dans lequel le module Pfam (resp. ProDom) est apparu. Enfin, si les deux modules sont apparus dans deux sous-arbres distincts, alors la comparaison est impossible et donc les deux origines sont considérées comme différentes. Le tableau 3.11 présente les résultats issus de la comparaison des âges entre les modules ProDom et Pfam. Parmi les modules ProDom, 74% ont une origine plus récente que le module Pfam auquel ils sont associés, et 26% ont la même origine. Nous notons cependant que pour 48% des familles de Pfam, le module ProDom le plus ancien qui lui est associé a la même origine que le module Pfam. Lorsque l'on prend en compte seulement les modules ProDom associés globalement à Pfam (chevauchement sur plus de 80% de la longueur du module ProDom), 57% des modules ProDom anciens ont la même origine que le module Pfam. On trouve donc une meilleure adéquation entre un module ProDom et Pfam lorsque le module ProDom est inclus dans son homologue de Pfam. Dans ce cas, 85% des modules Pfam ont la même origine que l'un au moins des modules ProDom auxquels ils sont associés. Cela suggère que le taux d'évolution de ces familles de modules n'est pas constant sur l'arbre, ce qui peut conduire à la création de sous-familles divergentes dans ProDom.

Comme indiqué plus haut dans ce chapitre, la majorité des modules Pfam sont anciens. De fait 81% des modules ProDom chevauchant Pfam correspondent à un module Pfam remontant jusqu'à LUCA : il s'agit donc de modules très anciens. L'analyse de la distribution des modules ProDom associés à Pfam le long de l'arbre (présentée dans le tableau 3.12) permet d'apprécier l'ampleur de ce biais. En effet si Pfam est représentative on s'attendrait à ce que la fréquence des modules

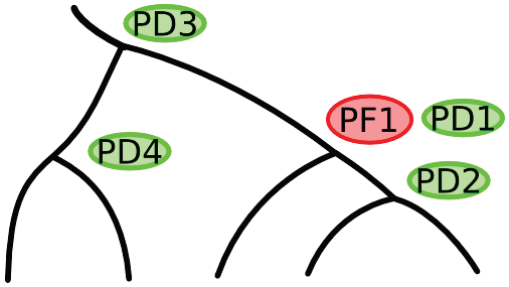


FIGURE 3.30 – **Comparaison des âges relatifs des modules ProDom et Pfam.** Pour chaque module ProDom innové (PD1, PD2, PD3, PD4) et associé à un module Pfam lui-même innové (PF1), l'âge des ancêtres dans lesquels ils sont apparus ont été comparés. On peut distinguer 4 situations : PD1 a la même origine que PF1, PD2 est plus récent que PF1, PD3 est plus ancien, et l'âge de PD4 n'est pas comparable à celui de PF1.

	Espèces contemporaines	Espèces ancestrales	Total
Modules procaryotes innovés	7 521	7 021	14 542
↪ Plus récent	7 123	3 613	10 736
↪ Même origine	398	3 361	3 759
↪ Plus ancien	0	47	47
dont la couverture (ProDom) > 0,8	3 200	4 591	7 791
↪ Plus récent	2 911	1 914	4 825
↪ Même origine	289	2 636	2 925
↪ Plus ancien	0	41	41

TABLEAU 3.11 – **Répartition des modules procaryotes innovés et associés à des modules Pfam en fonction de leur nœud d'innovation.** Un module ProDom peut être innové dans une espèce plus récente que celle où le module Pfam auquel il est associé est lui-même innové, dans la même espèce ou bien dans une espèce plus ancienne. Les modules innovés dans une espèce contemporaine ou ancestrale ont été distingués. Les résultats sont également présentés pour le sous-ensemble de modules ProDom associés à Pfam sur plus de 80% de leur longueur.

ProDom associés aux modules Pfam soit indépendante de l'âge de la famille ProDom. Cela n'est clairement pas le cas : plus les modules ProDom sont récents, moins ils sont associés à un module Pfam.

En conclusion, nous avons analysé les prédictions de différents types d'innovation protéique en fonction de différents jeux de modules : les modules ProDom, les modules Pfam, les modules ProDom associés à Pfam et les modules ProDom associés à Pfam sur toute leur longueur (figure 3.31). Lorsque l'on restreint l'ensemble des modules ProDom au sous-ensemble de modules associés à un module Pfam, les proportions des différents types d'innovation se rapprochent de celles obtenues avec Pfam (figure 3.31C). Les innovations totales et partielles ne sont pas négligeables, mais elles peuvent correspondre à des parties de modules ProDom non représentées dans Pfam. En effet

	LUCA	Très anciens	Anciens	Récents	Contemporains
Nb modules innovés ProDom	4 062	689	7 074	18 292	89 672
↪ proportion associée à Pfam	85%	68%	32%	20%	13%

TABLEAU 3.12 – **Répartition des modules ProDom innovés et associés à Pfam le long de l'arbre des espèces.** Les nœuds de l'arbre ont été répartis en 5 grands groupes d'âges relatifs différents : LUCA, très anciens (les ancêtres des Bactéries et des Archées), anciens (le sous-arbre contient au moins 5 espèces contemporaines, 42 ancêtres), récents (le sous-arbre contient moins de 5 espèces, 69 ancêtres) et contemporains (161 espèces). Les modules ProDom innovés dans chacun de ces groupes de nœuds ont été répertoriés, et la proportion des modules ayant une correspondance avec un module Pfam a été calculée.

lorsque l'on restreint les modules ProDom à ceux complètement associés à Pfam (figure 3.31D), les innovations totales et partielles sont fortement diminuées : on tend vers la distribution obtenue avec les données de Pfam (figure 3.31B). On remarque cependant que l'univers des protéines n'est alors pris en compte que très partiellement.

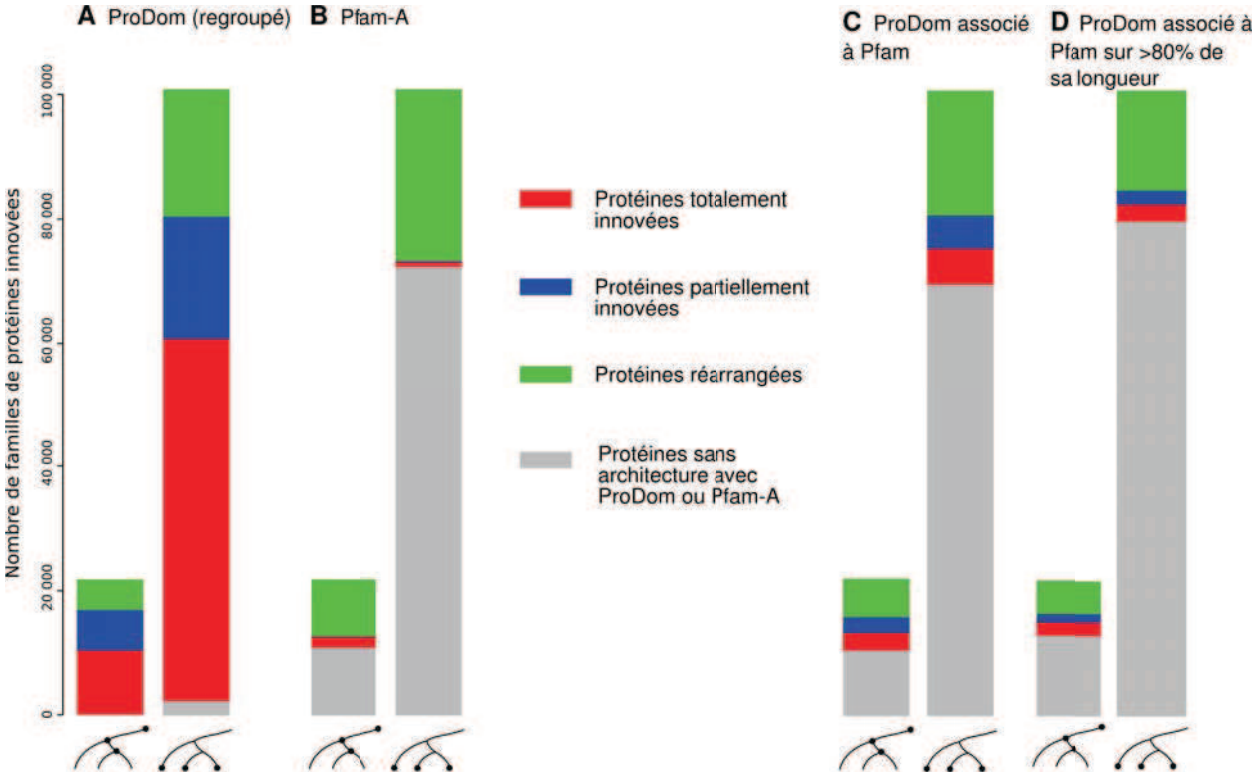


FIGURE 3.31 – **Distribution des différents types d'innovation protéique dans les espèces ancestrales et contemporaines.** Les protéines complètement innovées, partiellement innovées et réarrangées sont respectivement indiquées en rouge, bleu et vert sur la base des modules ProDom (**A**), Pfam-A (**B**), ProDom associés à un module Pfam (**C**) ou des modules ProDom associés globalement à un module Pfam (**D**). Les familles de protéines ne contenant aucun domaine des bases de données correspondantes sont indiquées en gris. Dans chaque panneau, les barres de gauche et de droite correspondent respectivement aux espèces procaryotes ancestrales et contemporaines, comme symbolisées sous chaque barre.

3.6.3 Le succès évolutif des modules innovés

L'analyse de la versatilité des modules innovés, soit le nombre de nouvelles protéines ayant recruté un module particulier avec succès, montre que les modules ProDom ont un certain succès dès leur plus jeune âge. Les distributions de la versatilité des modules ProDom et Pfam en fonction de leur âge sont présentées dans la figure 3.32. On trouve 7% des modules innovés récemment présents dans au moins deux arrangements en modules différents : bien que ces modules soient récents, leur potentiel a déjà été utilisé dans différents contextes protéiques. Le pourcentage de modules versatiles augmente à 35% pour les modules anciens, et jusqu'à 69% pour les modules les plus anciens remontant à LUCA. La versatilité peut donc être vue comme une marque du succès évolutif des modules.

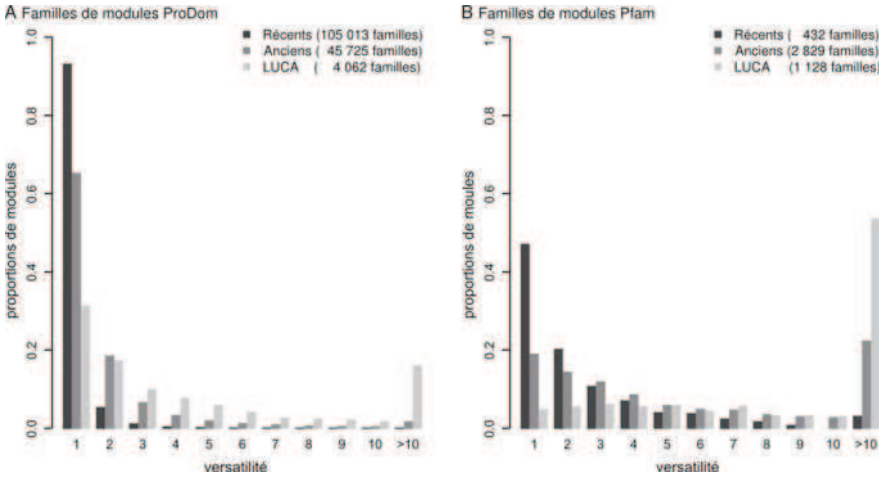


FIGURE 3.32 – **Versatilité des familles de modules anciennes et récentes.** La versatilité est définie comme le nombre d'architectures de modules dans lesquelles le module est impliqué. Les histogrammes des familles de modules de ProDom (**A**) et Pfam (**B**) sont donnés pour les familles de modules dont l'origine est une espèce contemporaine procaryote (gris foncé), un ancêtre intermédiaire (gris) et celles remontant à LUCA (gris clair).

Les modules Pfam sont globalement beaucoup plus versatiles que ceux de ProDom. En effet, la proportion de modules versatiles varie de 53% pour les modules récents à 95% pour les modules remontant à LUCA. Les modules Pfam sont donc biaisés en faveur de modules employés dans de nombreux contextes protéiques différents.

3.7 Conclusion : importance de l'innovation de modules pour l'apparition de protéines nouvelles

La vision classique de l'évolution des protéines modulaires explique l'apparition de nouvelles protéines par le réarrangement de modules protéiques anciens comme observé avec les modules

Pfam. Cependant, seuls 35% des familles de protéines sont prises en compte avec Pfam, avec une couverture en acides aminés de 21%. De plus, nous avons montré que l'échantillonnage de l'univers des protéines par Pfam n'est pas du tout représentatif, avec un biais prononcé en faveur de modules anciens et versatiles. D'où l'importance d'utiliser des méthodes qui puissent identifier les modules protéiques systématiquement : par exemple, ProDom permet d'annoter 98% des familles de protéines avec une couverture en acides aminés de 85%.

Les résultats obtenus présentent une typologie de l'innovation où l'apparition de nouvelles protéines est fortement liée à l'apparition de nouveaux modules protéiques. Cependant, la méthode automatique de regroupement des familles de ProDom présente une faiblesse dans la détection des homologies distantes, ce qui tend à subdiviser les familles de modules en sous-familles. Elle entraîne donc une surestimation de l'innovation des modules, qu'il convient de corriger. En nous basant sur la comparaison entre ProDom et Pfam, nous avons estimé plus haut que l'ancienneté de 57% des familles anciennes de ProDom était en accord avec Pfam et que 42% pourrait être mise en défaut comme étant trop récente (Tableau 3.11). Ce taux de faux positifs étant estimé principalement sur l'analyse des familles les plus anciennes, il constitue une borne supérieure du taux de faux positifs attendu pour des familles plus récentes. En extrapolant de manière conservative ce taux à l'ensemble des modules anciens de ProDom (associés ou non à Pfam), on peut donc corriger les estimations de la figure 3.20 et proposer une borne inférieure de 28% à la fréquence d'apparition de protéines anciennes intégralement innovées, au lieu de 48% dans la figure 3.20. Un calcul analogue nous permet de proposer une borne supérieure de 55% pour la fréquence d'apparition de protéines anciennes à partir de réarrangements de modules préexistants, au lieu de 22% dans la figure 3.20. Ces bornes restent très éloignées des fréquences obtenues avec Pfam, avec 16% seulement des protéines nouvelles intégralement innovées et 83% provenant de réarrangements de modules.

En conclusion, l'innovation en modules protéiques semble être tout au long de l'évolution un événement plus important que ce qui était pensé auparavant. Nos analyses avec Pfam indiquent bien que quelques domaines très anciens ont contribué significativement à la diversification protéique par de nombreux réarrangements, mais ProDom complète cette vision en montrant (1) que ces réarrangements sont fréquemment associés à de nouveaux modules et (2) que des protéines intégralement nouvelles peuvent apparaître qui ne contiennent aucun domaine préexistant. La vision de ProDom permet ainsi d'analyser l'ensemble des protéines prédites pour les espèces échantillonnées et nous éclaire sur l'évolution des protéines plus récentes, non prises en compte par Pfam.

Chapitre 4

Vers une nouvelle vision de l'évolution des protéines modulaires

Dans cette thèse, nous avons cherché à réconcilier deux approches de l'analyse de la diversification des protéines : (1) la combinatoire de domaines dans laquelle les nouvelles architectures sont classiquement des réarrangements de domaines préexistants ; (2) l'inférence de scénarios d'évolution des familles de protéines et de modules dans lesquels les événements évolutifs de gain et de perte sont positionnés sur un arbre des espèces. Ainsi, la combinatoire des domaines a été appréhendée dans un contexte explicitement évolutif dans lequel l'émergence des protéines est décrite à partir de l'histoire évolutive des modules composant leur architecture.

Nous avons développé une méthodologie d'inférence de scénarios d'évolution au maximum de vraisemblance basée sur la structure des réseaux Bayésiens (chapitre 2). En effet, les méthodologies implémentées jusqu'alors étaient essentiellement basées sur des critères de parcimonie non adaptés à l'hétérogénéité des fréquences de gain et de perte observées. Nous avons donc utilisé un modèle d'évolution dans lequel les probabilités de gain et de perte peuvent varier le long de l'arbre. Ce modèle décrit significativement mieux les données que le modèle homogène où les probabilités sont identiques le long de l'arbre. Son association à la structure des réseaux Bayésiens permet d'inférer des scénarios ayant un fort soutien statistique et robustes face à la variabilité de certains paramètres.

Le modèle d'évolution développé avec les réseaux Bayésiens a été utilisé pour inférer l'histoire évolutive des familles de protéines d'HOGENOM et des familles de modules de ProDom et de Pfam. L'analyse des répertoires ancestraux révèle une dynamique évolutive dans laquelle l'in-

novation est omniprésente au cours de l'évolution pour les familles de protéines et de modules ProDom. En revanche, la dynamique des modules Pfam pointe vers une majorité d'innovations dans les espèces les plus anciennes, ce qui a une incidence importante sur l'analyse de l'émergence des protéines. L'expertise manuelle des familles de modules et l'utilisation de HMMs confèrent à Pfam une plus grande sensibilité dans la détection des homologies distantes.Cependant, ces familles présentent une couverture réduite de l'univers des protéines, en privilégiant les familles anciennes et versatiles. Il en résulte une surestimation par Pfam des réarrangements et une sous-estimation de l'innovation des modules. A contrario, l'utilisation de ProDom permet d'obtenir une excellente couverture de l'ensemble des protéines, au prix d'un regroupement moins bon des modules homologues. Il en résulte une sous-estimation par ProDom des réarrangements et une surestimation de l'innovation des modules. Sur la base de la comparaison entre Pfam et ProDom, nous avons pu au chapitre 3 proposer une borne inférieure de 28% pour la fréquence d'innovation intégrale des protéines anciennes et une borne supérieure de 55% pour la fréquence d'innovation par réassortiment pur (p. 89). En conséquence, nous proposons qu'au moins 45% des innovations protéiques anciennes ont incorporé au moins un module protéique nouveau. Dans ce chapitre, nous discutons trois aspects de l'évolution des protéines directement pertinents pour nos résultats : l'univers des modules, la modularité des protéines et l'innovation en modules ou domaines protéiques.

4.1 Points de vue et biais sur l'univers des modules protéiques

Les bases de données de familles de domaines ou modules protéiques se distinguent en général par le compromis entre la sensibilité et la spécificité des familles de domaines représentées. Une bonne sensibilité dans la détection des familles de modules permet d'obtenir une couverture optimale de l'univers des protéines tandis que la spécificité garantit la qualité de la définition des familles de domaines ainsi que le recrutement des domaines appartenant aux différentes familles. Ainsi, il existe les bases de données de familles de modules exhaustives, construites automatiquement comme ProDom ou ADDA. Elles offrent une bonne couverture de l'univers des protéines correspondant, mais les méthodologies employées peuvent manquer de sensibilité pour la détection des homologies anciennes. On trouve également les bases de données construites manuellement ou basées sur les données structurales comme Pfam ou CATH qui ne couvrent que partiellement l'univers des séquences protéiques bien que la définition des familles de domaines soit de bonne qualité et que les outils utilisés permettent la détection d'homologues distants. Ces bases de données offrent des visions assez différentes de l'univers des domaines protéiques. Pour certaines, c'est un univers restreint dont la majorité des familles ont été déterminées [Apic *et al.*, 2001b; Chothia, 1992; Chothia *et al.*, 2003]. Cette vision de l'univers des domaines est basée principalement sur une analyse des structures de protéines disponibles, ce qui ne va pas sans biais. Après avoir présenté ces biais et l'impact qu'ils peuvent avoir sur la vision de la diversification des domaines, nous

discuterons des avancées récentes dans l'exploration de nouveaux domaines protéiques.

4.1.1 Une saturation artéfactuelle de l'univers des domaines protéiques par des domaines majoritairement anciens

Les données structurales permettent de classer les domaines de façon hiérarchique. Ils sont regroupés en familles qui sont elles-mêmes regroupées en superfamilles et enfin les superfamilles sont regroupées en classes de repliements. On s'attend donc par construction à ce que l'univers des repliements ou des superfamilles arrive à saturation avant l'univers des familles de domaines. Le fait que le nombre de nouveaux repliements trouvés diminue avec la détermination de nouvelles structures a fait penser à une proche saturation de l'univers des domaines. De nombreux modèles ont été créés pour extrapoler les effectifs des différentes classes hiérarchiques des domaines à partir des données disponibles. Les différentes estimations sont regroupées dans le tableau 4.1, extrait de Schaeffer et Daggett [2011]. L'ordre de grandeur du nombre estimé de superfamilles varie de 10^3 à 10^5. L'estimation de ces effectifs est assez difficile notamment à cause du fait que tous les repliements (ou les superfamilles) ne sont pas peuplés également en terme de diversité structurale et de séquences. En effet, d'après Coulson et Moult [2002], on peut modéliser l'univers des repliements (ou des superfamilles) en trois zones distinctes : une zone où les repliements sont retrouvés dans une seule famille de domaines (nommés *Unifold*), une zone où les repliements sont trouvés dans un nombre intermédiaire de familles (nommés *Mesofold*) et enfin une zone où les repliements sont retrouvés dans de nombreuses familles (nommés *Superfold*). Ils mettent en évidence que l'usage des repliements suit une loi de puissance et que les repliements uniques sont largement sous-estimés.

La difficulté de l'estimation de la taille de l'univers des domaines, que ce soit au niveau des repliements ou des superfamilles, vient également du fait que l'on se base sur des données structurales globalement biaisées. En effet, les protéines pour lesquelles une structure tridimensionnelle a été déterminée sont de petite taille [Gerstein, 1998] et appartiennent généralement à de grandes familles bien caractérisées [Kunin *et al.*, 2005], les protéines singletons ou les petites familles espèce-spécifique étant sous-représentées [Coulson et Moult, 2002]. La représentation taxonomique des superfamilles (extraites de la base de données SUPERFAMILY [Wilson *et al.*, 2009]) regroupe les trois domaines du vivant pour plus de 63% d'entre elles alors que moins de 1% des familles de domaines de ProDom le sont [Kunin *et al.*, 2005]. De plus, on retrouve également un biais en faveur des protéines ayant un intérêt industriel ou thérapeutique : 10 fois plus de superfamilles sont spécifiques des bactéries que des archées [Kunin *et al.*, 2005].

Ces biais d'échantillonnage se retrouvent dans la plupart des bases de données créées manuellement. En effet, la création manuelle d'une base de données implique la détection préalable des familles sur la base de la structure ou de la séquence. On constate également un biais en faveur de

grandes familles de domaines représentées dans l'ensemble du monde vivant, donc des domaines en général anciens (comparer par exemple les figures 3.9 page 64 et 3.7 page 61). De tels biais sont évités dans les bases de domaines qui regroupent l'ensemble des données de séquence des modules protéiques par des méthodes automatiques, telles que ProDom ou ADDA. On trouve donc dans ces dernières des familles beaucoup plus récentes.

4.1.2 Vers un échantillonnage aléatoire des structures déterminées

L'analyse de la couverture du monde protéique par les domaines a permis d'estimer les données manquantes pour obtenir une couverture optimale. Marsden *et al.* [2006] ont montré que 6 000 familles de domaines extraites de CATH et Pfam permettaient de couvrir 74% des séquences et 56% des résidus. Parmi les familles de Pfam utilisées, il manque environ 1 000 structures pour avoir une couverture structurale de cette grandeur. Ils estiment à 90 000, le nombre de structures nécessaires pour avoir un modèle d'homologie de toutes les séquences analysées et ainsi couvrir les séquences espèces-spécifiques. Ce nombre est nettement supérieur au nombre de structures non redondantes actuellement déterminées. Qu'est-ce que ces nouvelles structures nous diraient sur l'univers des repliements et des superfamilles de domaines ?

Ces dernières années de nombreux projets, comme le projet PSI (Protein Structure Initiative) financé par le NIH (National Institutes of Health), ont été mis en place pour diminuer le biais d'échantillonnage des bases de données structurales. Ces projets visent à choisir les structures à

Années	Repliements	Superfamilles	Référence
1992	< 1 000	1 500	Chothia [1992]
1994	< 7 700	23 100	Orengo *et al.* [1994]
1994	6 727	-	Alexandrov et Go [1994]
1996	455	-	Wang [1996]
1997	< 920	920	Brenner *et al.* [1997]
1997	< 5 200	17 175	Zhang [1997]
1998	650	1 150	Wang [1998]
1998	836	-	Zhang et DeLisi [1998]
1999	3 756	-	Govindarajan *et al.* [1999]
2000	$\sim$ 1 000	4 000-7 000	Wolf *et al.* [2000]
2002	10 000	50 000	Coulson et Moult [2002]
2007	1 613	-	Levitt [2007]
2009	$\sim$ 1 700$\pm$400	$\sim$ 4 000	Sadreyev *et al.* [2009]

TABLEAU 4.1 – **Estimation du nombre de repliements et de superfamilles par année.** Extrait de Schaeffer et Daggett [2011].

déterminer pour augmenter la couverture structurale des séquences et limiter la redondance dans les bases de données [Grabowski *et al.*, 2007; Jaroszewski *et al.*, 2009; Marsden *et al.*, 2006; Todd *et al.*, 2005]. D'autres projets de méta-génomique comme le GOS (Global Ocean Sampling) [Yooseph *et al.*, 2007] ont permis d'obtenir un échantillonnage important de séquences, desquelles de nombreuses nouvelles familles ont été détectées. Ces dernières constituent de bonnes candidates pour augmenter l'échantillonnage de nouvelles structures.

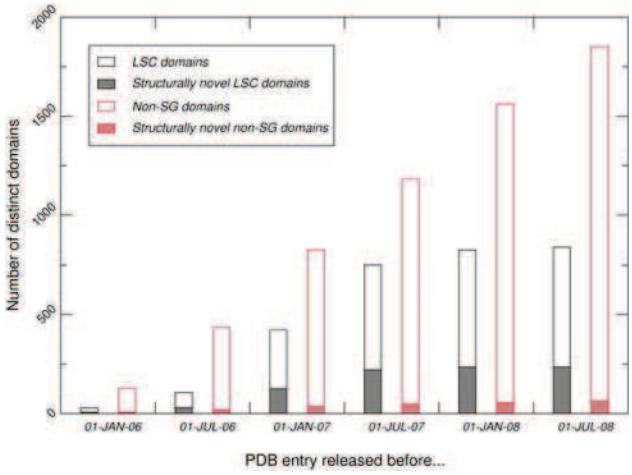


FIGURE 4.1 – **Nouveauté des domaines structuraux récemment déterminés**. Les domaines structuraux ont été déterminés par les centres associés au programme PSI-2 (LSC, Large-Scale Centers), représentés par les barres grises et les centres de biologie structurale traditionnels (à l'exclusion des centres de génomique structurale), représentés par les barres rouges, entre juin 2005 et juin 2008. Seuls les domaines classés dans CATH sont représentés. Extrait de Dessailly *et al.* [2009].

Jaroszewski *et al.* [2009] ont analysé les structures de 250 DUFs (Domain of Unknown Function) dans le cadre du projet PSI. Cette analyse fait suite au constat qu'entre 30% et 40% des protéines sont hypothétiques ou de fonction inconnue, malgré les efforts d'annotation des génomes complets. Ils montrent que 25% de ces familles de domaines peuvent être reliées à d'autres familles existantes par similarité de séquence, la structure venant conforter l'homologie. On trouve également 48% des familles sans similarité de séquence suffisante pour être reliées à des familles caractérisées, mais dont certains repliements caractéristiques permettent de les rattacher à une classe existante. Il reste 27% de nouveaux repliements qui ne peuvent pas être reliés à des familles existantes. L'analyse de 1 502 nouvelles structures issues du programme PSI-2 [Dessailly *et al.*, 2009] a également permis de mettre en évidence que 28% d'entre elles étaient significativement différentes des structures présentes dans CATH (voir figure 4.1). Ainsi, en considérant d'après Marsden *et al.* [2006] qu'il reste environ 90 000 structures à déterminer, on pourrait s'attendre à trouver environ 23 400 nouveaux repliements parmi eux, en faisant l'hypothèse que les nouvelles structures déterminées représentent un échantillon aléatoire des domaines inexplorés.

4.2 Les modules protéiques : un point de vue sur l'évolution des protéines

La diversification de l'univers des protéines peut être expliquée à l'aide de nombreux mécanismes dont les principaux mis en évidence sont les duplications, la divergence par mutations ponctuelles et le réarrangement des domaines composant son architecture. Cette section discute de la pertinence des domaines comme unité évolutive, puis des conséquences possibles sur les relations d'homologie inférées entre protéines et enfin nous reviendrons sur l'interprétation et les limitations de notre modèle d'évolution.

Quelle que soit l'échelle à laquelle on se place pour comprendre l'évolution, comme une espèce, un génome, un gène, une protéine ou encore un nucléotide, on définit toujours une unité qu'on analyse, compare et pour laquelle on recherche des homologues dans le but de comprendre son évolution. La pertinence de ces unités est toujours discutable puisque leurs limites ne sont pas nécessairement durables au cours de l'évolution. Par exemple, dans le cas d'un génome, on peut définir de multiples unités d'évolution différentes : le gène, l'exon, le chromosome ou le génome lui-même.

L'analyse des protéines modulaires met en avant la pertinence du domaine comme une unité d'évolution des protéines, dans la mesure où ils sont préservés par la sélection. On utilise ainsi une échelle plus fine pour expliquer l'évolution à l'échelle des protéines. Cependant, certaines protéines, comme les protéines fibreuses, échappent à cette définition puisqu'elles ne peuvent pas être découpées en domaines globulaires. C'est la raison pour laquelle dans cette thèse, nous avons majoritairement parlé de modules protéiques qui décrivent seulement des fragments de protéines conservés sans nécessairement une notion de structure ou de fonction.

En considérant l'architecture en domaines des protéines, la définition des relations d'homologie entre protéines est modifiée. En effet, en considérant les protéines entières : soit deux protéines sont suffisamment similaires pour être considérées homologues soit elles ne le sont pas. Avec la prise en compte des domaines, certaines protéines non homologues peuvent partager des domaines homologues. C'est à partir de cette observation que le terme d'homologie partielle a été introduit pour les protéines [Hillis, 1994].

Cette constatation soulève la question de la différenciation entre homologie ou convergence entre des protéines possédant la même architecture en domaines. Est-ce qu'un même arrangement en domaines peut apparaître plusieurs fois indépendamment au cours de l'évolution ? L'apparition d'un même arrangement en modules dans deux clades différents arrive relativement fréquemment. Lorsqu'on analyse les scénarios d'évolution d'une famille de protéines, cet événement correspond aux familles présentant des gains multiples. Dans l'ensemble de l'analyse présentée et de manière générale, ces gains multiples sont interprétés comme des transferts horizontaux. Cependant, on

peut distinguer trois mécanismes différents pour expliquer un tel scénario (figure 4.2) : soit la famille est ancienne et a été perdue dans la majorité des clades, soit il s'agit d'un ou plusieurs cas de transferts horizontaux, soit il s'agit d'un phénomène de convergence. Dans ce dernier cas, les protéines ne sont pas homologues. Les méthodes d'analyse sont alors abusées par une même composition en modules homologues puisque la similarité entre les séquences devient suffisante pour que l'hypothèse d'homologie soit acceptée.

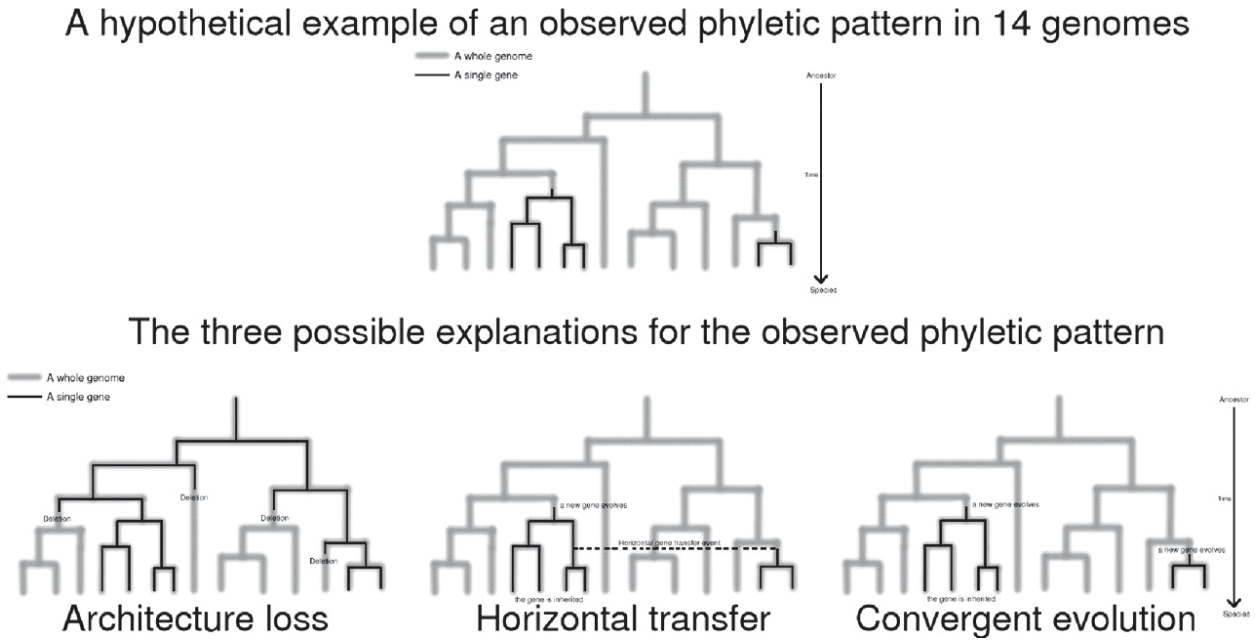


FIGURE 4.2 – **Les trois interprétations possibles d'un gain multiple.** Description d'un exemple de scénario d'évolution d'une architecture de domaines donnée ainsi que des trois explications évolutives possibles. Extrait de Gough [2005].

La convergence architecturale signifie qu'un même arrangement en domaines a été créé indépendamment plusieurs fois. La prévalence de ce phénomène est jugée plutôt rare par Gough [2005] qui prédit entre $0{,}4$ et 4% des séquences impliquées dans une convergence architecturale. Cet auteur est parti de l'hypothèse que la convergence correspond à une contrainte fonctionnelle forte. La même fonctionnalité serait donc créée indépendamment dans différents génomes. Les architectures candidates ont été détectées par des analyses phylogénétiques où l'architecture apparaît dans au moins deux clades distincts. Les architectures retenues correspondent à celles pour lesquelles une relation d'homologie n'a pas pu être inférée à partir d'analyses de similarité sur toute la longueur de la séquence et de taux d'évolution.

Une autre analyse arrive à la conclusion que la convergence architecturale est plus fréquente avec une prévalence variant de $5{,}6$ à $12{,}4\%$ sur les architectures de 96 génomes [Forslund *et al.*, 2008]. Cette analyse reconstruit l'histoire évolutive des architectures non pas sur un arbre des espèces, mais sur les arbres phylogénétiques des séquences des domaines composant les architectures. Les gains multiples des architectures sur ces arbres représentent les cas possibles de conver-

gence. Cette analyse plus fine, réalisée à partir des séquences des domaines, permet de s'affranchir des cas des transferts horizontaux directement pris en compte dans l'analyse.

Bien que le phénomène de convergence ne soit pas un événement très fréquent, il met en évidence l'un des aspects simplificateurs de notre modèle d'évolution des familles de protéines et de modules. En effet, les scénarios d'évolution infèrent la présence ou l'absence d'une famille à un nœud donné de la phylogénie des espèces. Cette méthode permet d'obtenir une image qualitative des contenus des génomes ancestraux. Cependant, cela reste un modèle simplificateur de l'évolution des protéines dans lequel les événements de gains multiples sont interprétés comme des transferts horizontaux. De manière générale, l'analyse des transferts horizontaux nécessiterait également la reconstruction d'un arbre phylogénétique des séquences, à comparer avec un arbre des espèces.

4.3 Qu'est-ce que l'innovation en modules/domaines protéiques ?

Notre analyse de l'émergence de nouvelles protéines indique l'importance de l'innovation en modules. Si cette conclusion apporte un éclairage nouveau sur l'apparition de protéines au cours de l'évolution, elle pose également la question de l'origine des nouveaux modules et plus généralement la question de l'innovation. La notion même d'innovation n'est pas évidente et a fait l'objet de nombreux débats [Moczek, 2008]. L'objectif n'est pas ici de reprendre ce débat, mais plutôt de définir l'innovation sous-entendue dans cette étude à l'échelle des domaines protéiques et de discuter des différents mécanismes possibles.

Ernst Mayr définissait l'innovation en 1960 comme n'importe quelle nouvelle structure ou propriété acquise pour laquelle on peut faire l'hypothèse d'une nouvelle fonction [Mayr, 1960]. Il met en avant l'aspect fonctionnel des innovations, aspect repris aujourd'hui par Tokuriki et Tawfik [2009] pour qui l'innovation est l'apparition d'une nouvelle fonction qui peut initialement n'avoir qu'un petit effet sur la fitness qui peut être renforcé par la sélection au cours du temps. L'innovation suppose donc l'apparition d'un nouveau caractère et sa conservation dans le temps.

L'apparition d'un nouveau domaine suggère l'acquisition d'une nouvelle fonction que ce soit par l'apparition d'une nouvelle séquence codante ou d'une nouvelle structure. On peut ainsi dégager deux types d'origines possibles pour les domaines. Le premier correspond au recrutement d'un nouveau fragment de séquence codante à partir d'une séquence non-codante : apparition de codon initiateur, disparition de codon stop, décalage de la phase de lecture, perte de site d'épissage, insertion, délétion, etc. Le deuxième implique des modifications d'un domaine préexistant. Il peut s'agir d'une accélération brutale du taux d'évolution en raison d'un changement de contrainte

sélective, d'une modification structurale par accrétion d'éléments additionnels ou encore d'une délétion d'éléments de la structure.

4.3.1 L'apparition d'une nouvelle séquence codante

La création *ab initio* de nouveaux domaines est l'une des hypothèses utilisées pour expliquer l'apparition des premiers domaines à partir desquels les autres domaines auraient divergé [Chothia et Gough, 2009]. Ce modèle permet l'acquisition d'un jeu de domaines suffisamment variés pour soutenir une forme de vie basique. Et lorsque la diversité a été suffisante, les mécanismes utilisant la duplication et la divergence, ainsi que la recombinaison des domaines sont supposés plus efficaces pour créer de la nouveauté. L'apparition de nouvelles séquences à partir de matériel non-codant est donc généralement considérée comme peu probable. Cependant, plusieurs exemples d'apparition de nouvelles séquences codantes à partir de séquences non-codantes ont été mis en évidence dans les génomes de la Drosophile et de l'homme, comme nous l'avons présenté en introduction. La figure 4.3 présente la succession des événements permettant une telle innovation mise en évidence pour trois gènes humains [Knowles et McLysaght, 2009]. La comparaison des espèces proches de l'homme a permis de mettre en évidence les mutations à l'origine de la réorganisation du cadre ouvert de lecture ancestral, après la divergence de la lignée humaine.

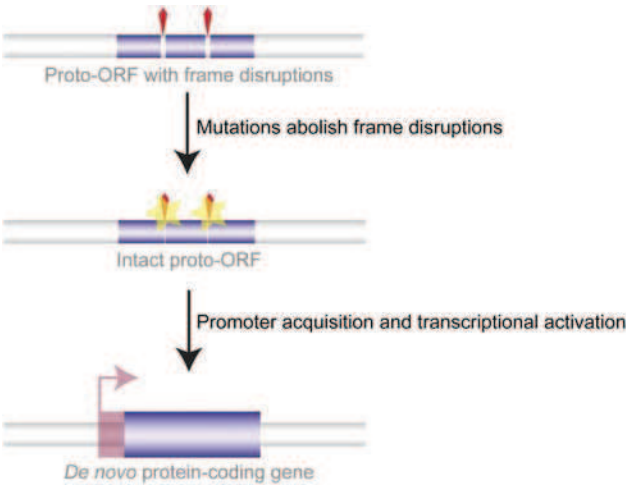


FIGURE 4.3 – **Origine d'une nouvelle séquence codante.** De nouvelles régions codantes peuvent apparaître de novo à partir de séquences génomiques non codantes. Des cadres ouverts de lecture primitifs (proto-ORF : représentées par les petites boîtes bleues) subissent des mutations (substitutions ponctuelles,insertions et délétions ; représentées par les étoiles jaunes) qui retirent morceau par morceau les nucléotides désorganisant le cadre de lecture (représentés par les pointes rouges). L'activation transcriptionnelle des ORFs (par l'acquisition de promoteurs localisés dans la région 5'), codant pour des protéines dont la fonction est potentiellement utile, peut permettre l'évolution vers un nouveau gène codant. Les larges boîtes bleues représentent des exons fonctionnels, la flèche rose le TSS (Transcription Starting Site), et la boîte transparente rose les séquences non transcrites en 5'. A noter, l'étape de l'activation transcriptionnelle peut également précéder la formation d'une ORF utile et complètement fonctionnelle. Extrait de Kaessmann [2010].

L'ensemble de ces analyses suggère que l'émergence de novo est plus fréquente que ce que l'on pensait. Cependant, la prévalence de ce mécanisme dans l'ensemble du monde vivant est toujours inconnue. L'origine des nouveaux modules mis en évidence avec les données de ProDom a été analysée par Louis-Marie Bobay au laboratoire. L'analyse a porté sur 52 génomes bactériens. Les hypothèses concernant le recrutement de séquences non-codantes et le décalage de la phase de lecture ont été testées sur des modules ayant une origine récente. La recherche de similarité des nouveaux modules dans les génomes complets n'a pas été concluante : des séquences similaires ont été détectées dans les génomes voisins pour seulement 25% des nouveaux modules et parmi eux, à peine 9% étaient associés à une séquence intergénique et 6% étaient associés à une séquence codante avec un décalage de la phase de lecture. Cependant, dans le cas des modules associés à des séquences intergéniques, il n'a pas été possible de déterminer d'une manière inambigüe si la séquence ancestrale était non-codante et donc d'orienter les événements. Ainsi le nombre de données exploitables n'a pas été suffisant pour mettre en évidence l'impact du recrutement de séquences intergéniques dans l'apparition de nouveaux modules.

Une autre hypothèse pourrait expliquer une partie des nouveaux domaines observés dans les génomes actuels : l'importation d'éléments de séquence (gènes, fragments de gènes ou séquences non-codantes) à partir de génomes de clades non échantillonnés ou de génomes viraux. En effet, même si l'échantillonnage taxonomique de notre travail couvre les grands domaines du vivant, il reste faible par rapport aux quelque 10^7 espèces existantes. On a donc aujourd'hui accès à un échantillon restreint de la diversité des espèces et des protéines, très probablement biaisé. Cependant, les projets de métagénomique, de plus en plus nombreux (par exemple Yooseph *et al.* [2007]), devraient fournir rapidement du matériel génétique divers et permettre d'obtenir un meilleur échantillonnage de l'univers des espèces et des séquences. Ces projets permettent déjà de détecter des séquences homologues à des séquences considérées comme orphelines jusqu'à présent.

4.3.2 Évolution d'un nouveau domaine à partir d'éléments de domaines anciens

De nombreuses analyses de la diversification des structures tridimensionnelles des domaines ont été réalisées ces dernières années. Elles montrent des changements graduels où des éléments de structure secondaire sont successivement ajoutés ou enlevés à la structure tridimensionnelle conservée : il s'agit d'un phénomène d'embellissement des structures [Cuff *et al.*, 2009a; Grishin, 2001; Kim et Caetano-Anollés, 2010; Reeves *et al.*, 2006]. Cela peut entraîner une modification structurale majeure altérant la géométrie d'un site actif ou la conformation de surface de la protéine, modifiant sa fonction.

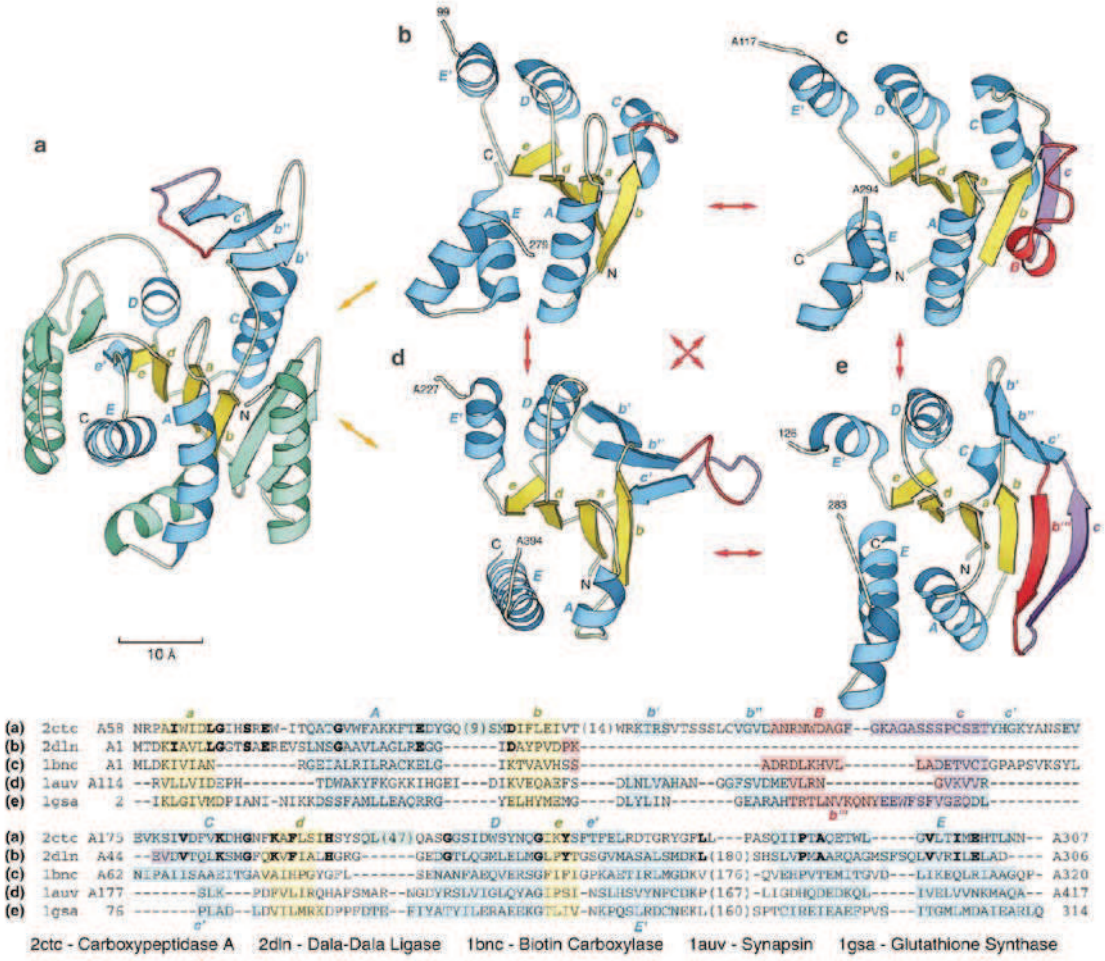


FIGURE 4.4 – **Insertions, délétions et substitutions dans l'évolution des structures des domaines *Rossman Fold* présents dans les protéines ATP-grasp et carboxypeptidase à zinc.** (a) Carboxypeptidase A (2ctc) [Rees *et al.*, 1983] ; (b) D-Ala-D-Ala ligase (2dln) [Fan *et al.*, 1994] ; (c) biotin carboxylase (1bnc) [Waldrop *et al.*, 1994] ; (d) synapsin (1auv) [Esser *et al.*, 1998] ; (e) glutathione synthase (1gsa) [Yamaguchi *et al.*, 1993]. Les représentations en ruban des protéines ont été réalisées avec Bobscript [Esnouf, 1997], une version modifiée de Molscript [KRAULIS, 1991]. Les structures ont d'abord été superposées puis séparées pour une meilleure lisibilité. Les extrémités N et C-terminal sont étiquetées. Les éléments de structure équivalents sont colorés de la même façon dans l'ensemble des structures. Les insertions/délétions sont représentées en vert et violet. Le rouge est utilisé pour souligner certains changements structuraux. Les hélices α et les brins β sont étiquetés en caractères gras et en italique, respectivement en lettres majuscules et minuscules. La couleur de la lettre correspond à la couleur de la structure. Les flèches rouges et oranges entre les structures symbolisent respectivement une relation évolutive avérée et une relation évolutive possible. Les alignements de séquences basés sur les structures sont également représentés. Les numéros des premiers et derniers acides aminés représentés sont donnés pour chaque protéine. La coloration et l'indexation de l'alignement sont reportées sur les éléments de structure correspondants. Les acides aminés conservés sont en caractères gras. L'identifiant PDB et le nom de la protéine sont donnés en dessous de l'alignement. Extrait de Grishin [2001].

Par exemple les membres de la superfamille des ATP-grasp catalysent la ligation ATP-dépendante d'un substrat carboxylate à un groupe thiol ou amine d'un second substrat [Todd *et al.*, 2001]. Quasiment tous les membres de cette superfamille partagent trois domaines, dont deux domaines de liaison de l'ATP qui présentent de nombreux embellissements dans les différents représentants de la superfamille (figure 4.4b-e). La divergence structurale se répercute essentiellement sur la périphérie de la structure notamment par l'ajout d'une unité $\beta\alpha$ (figures 4.4b et 4.4c) ou par la substitution d'une hélice par un brin : le brin β *b"'* de la protéine glutathione synthase est topologiquement similaire à l'hélice α *B* de la biotin carboxylase (figures 4.4c et 4.4e) [Grishin, 2001]. La fonction des domaines embellis varie considérablement : elle inclut des carboxylases, des synthétases et des ligases [Reeves *et al.*, 2006].

L'analyse plus systématique des superfamilles définies dans CATH a permis de préciser les caractéristiques de ces embellissements. Le nombre de structures secondaires peut facilement varier du simple au double entre les différents membres d'une même superfamille. De plus, ces changements sont en général co-localisés dans la structure tridimensionnelle, résultant en une modification structurale finale plus importante susceptible d'altérer la fonction [Reeves *et al.*, 2006] : on parle d'un phénomène d'accrétion de structures secondaires. Il apparaît également qu'un quart des superfamilles présentent des chevauchements structuraux avec d'autres superfamilles, ce qui suggère que l'espace des repliements ressemble plutôt à un espace continu dans lequel les structures peuvent être modifiées graduellement [Cuff *et al.*, 2009a]. Ces modifications peuvent aller jusqu'à un changement de classe de repliement pour les structures, quand bien même les séquences sont homologues. Grishin [2001] avait déjà présenté ces changements de classe à partir de différents mécanismes modifiant la structure comme les mutations ponctuelles et le phénomène d'accrétion de structures secondaires, allant même jusqu'à présenter le modèle d'un chemin possible entre une structure tout-β vers une structure tout-α. Toutes les structures intermédiaires ne sont pas retrouvées dans les bases de données actuelles, mais ce modèle a l'avantage de présenter les principes possibles à l'œuvre dans l'évolution des structures.

En introduction, j'ai présenté une vision classique de l'évolution des protéines impliquant des réassortiments combinatoires de modules, souvent des domaines structuraux, dans une sorte de bricolage de modules. Dans cette thèse, nous avons montré que ces réassortiments de modules ne suffisent pas à expliquer l'ensemble de la diversité des protéines, dans la mesure où des modules nouveaux peuvent résulter eux-mêmes d'un bricolage à une échelle plus fine.

> *L'évolution procède comme un bricoleur qui, pendant des millions et des millions d'années, remanierait lentement son œuvre, la retouchant sans cesse, coupant ici, allongeant là, saisissant toutes les occasions d'ajuster, de transformer, de créer.*
>
> François JACOB (1981)

Annexe A

Notions de base sur les réseaux Bayésiens

Les notions présentées ici sont essentiellement extraites du livre de Jensen et Nielson [Jensen et Nielson, 2007].

Un réseaux Bayésien peut être décrit comme un réseau de causalité dont la force des liens est représentée par des probabilités conditionnelles. Ainsi la modélisation graphique représente la connaissance sur les variables de manière qualitative alors que les probabilités conditionnelles donnent une information quantitative. Cet annexe donne des précisions sur ces deux aspects fondamentaux des réseaux Bayésiens.

Rappels sur les probabilités associées aux réseaux Bayésiens

Le lien de cause à effet entre plusieurs variables peut-être décrit d'un point de vue probabiliste à l'aide des probabilités conditionnelles, qui décrivent la force du lien entre les variables. Les définitions et théorèmes suivants présentent les notions de base associées aux probabilités conditionnelles.

Notations :

- Les noms de variables sont notés en lettres majuscules (ex. A)
- Les états que peut prendre une variable sont notés en minuscule $sp(A) = \{a_1,a_2,\ldots,a_n\}$, avec $sp(A)$ l'espace des états possibles de A.
- On simplifie $P(A = a_i)$ par $P(a_i)$ pour donner la probabilité de la variable A d'être dans l'état a_i.

On ne considère ici que les variables aléatoires discrètes, c'est-à-dire ne pouvant prendre qu'un nombre fini d'états.

Définition A.1. *Pour une variable A pouvant prendre les états $a_1,a_2,\ldots,a_n$, on exprime le caractère incertain de son état à travers une* ***distribution de probabilités*** *$P(A)$ sur ses états :*

$P(A) = (x_1,x_2,\ldots,x_n)$

avec x_i la probabilité de A d'être dans l'état a_i

et $x_i \geq 0$ et $\sum_{i=1}^{n} x_i = 1$

Une distribution est dite ***uniforme*** *si toutes les probabilités sont égales.*

Définition A.2. *On considère deux variables aléatoires discrètes A et B dont les états possibles sont respectivement $sp(A) = \{a_1,a_2,\ldots,a_n\}$ et $sp(B) = \{b_1,b_2,\ldots,b_m\}$. La* ***probabilité conditionnelle de*** *A* ***sachant*** *B, $P(A \mid B)$, est décrite par $n \cdot m$ probabilités conditionnelles $P(a_i \mid b_j)$ qui spécifient la probabilité d'observer l'événement a_i sachant b_j. La propriété suivante est vérifiée :*

$$\forall b_j, \sum_{i=1}^{n} P(a_i \mid b_j) = 1$$

Théorème A.1 (Règle fondamentale)**.**

$$P(A,B) = P(A \mid B)P(B) \tag{A.1}$$
$$= P(B \mid A)P(A)$$
$$P(A,B|C) = P(A \mid B,C)P(B \mid C) \tag{A.2}$$

$P(A,B)$ est la distribution jointe de A et B qui dans le cas de variables discrètes peut être écrite sous la forme d'une table (tableau A.1).

$P(A,B)$	b_1	b_2
a_1	$P(a_1 \mid b_1)P(b_1)$	$P(a_1 \mid b_2)P(b_2)$
a_2	$P(a_2 \mid b_1)P(b_1)$	$P(a_2 \mid b_2)P(b_2)$

TABLEAU A.1 – **Distribution de la probabilité jointe de A et B :** $P(A,B)$. La variable A peut prendre les états a_1 et a_2 et la variable B peut prendre les états b_1 et b_2.

A partir de cette table, on peut calculer la distribution de probabilité de A, $P(A)$ en **marginalisant sur** A, c'est-à-dire en excluant la variable B de la distribution jointe :

$$\forall a_i, P(a_i) = \sum_{j=1}^{m} P(a_i,b_j)$$

Le théorème A.1 permet de définir la **distribution de probabilités jointes** d'un ensemble $\mathcal{U} = A_1,A_2,A_3,...,A_n$ de variables aléatoires. Le théorème suivant se démontre en appliquant récursivement l'équation A.1.

Théorème A.2 (Règle de chaîne générale)**.** *Soit $\mathcal{U} = A_1,A_2,A_3,...,A_n$ un ensemble de variables aléatoires. La distribution de probabilité $P(\mathcal{U})$ s'écrit :*

$$P(\mathcal{U}) = P(A_1,A_2,A_3,...,A_n)$$
$$P(\mathcal{U}) = P(A_n \mid A_1,A_2,...A_{n-1})P(A_{n-1} \mid A_1,A_2,...A_{n-2})...P(A_2 \mid A_1)P(A_1) \tag{A.3}$$

Le théorème A.1 permet également d'écrire le théorème de Bayes :

Théorème A.3 (Règle de Bayes)**.**

$$P(A \mid B) = P(B \mid A)\frac{P(A)}{P(B)} \tag{A.4}$$
$$P(A \mid B, C) = P(B \mid A, C)\frac{P(A \mid C)}{P(B \mid C)} \tag{A.5}$$

Ce théorème donne une méthode pour mettre à jour nos croyances sur A sachant que l'on connaît l'état de B. On appelle $P(A)$ la probabilité **à priori** de A alors que $P(A \mid B)$ est appelée la probabilité **à postériori** de A sachant B ; la probabilité $P(B \mid A)$ est appelé la **vraisemblance** de A. La réécriture du théorème précédent met en évidence la proportionnalité suivante :

$$\textit{Loi a posteriori} \propto \textit{Vraisemblance} \times \textit{Loi a priori}$$

Je termine ces rappels avec les définitions d'indépendance et d'indépendance conditionnelle entre deux variables aléatoires qui sont essentielles dans le cadre des réseaux Bayésiens.

Définition A.3. *La variable A est **indépendante** de la variable B si*

$$P(A \mid B) = P(A)$$

L'indépendance est symétrique : si A est indépendante de B alors B est indépendante de A et $P(B \mid A) = P(B)$

Définition A.4. *Les variables A et B sont **indépendantes conditionnellement** à C si $\forall a_i \in sp(A)$, $b_j \in sp(B)$ et $c_k \in sp(C)$*

$$P(a_i \mid b_j, c_k) = P(a_i \mid c_k)$$

Si A et B sont indépendantes conditionnellement à C alors l'équation A.2 (théorème A.1) est simplifiée à l'aide de la définition A.4, elle devient :

$$P(A,B \mid C) = P(A \mid C)P(B \mid C)$$

Des réseaux de causalité aux réseaux Bayésiens

Les réseaux de causalité sont des graphes dirigés et acycliques dont les nœuds représentent les variables d'intérêt et les arcs les relations de dépendance, ou de cause à effet, entre ces variables. Lorsqu'il existe un lien de A vers B alors on dit que B est le fils de A et A est le parent de B. Ces relations peuvent être utilisées pour suivre la manière dont l'observation d'une variable change la connaissance sur les autres variables. Dans un graphe simple avec seulement deux nœuds A et B, et une relation causale de A vers B, toute information sur A modifie la connaissance sur B et réciproquement, toute information sur B peut modifier la connaissance sur A. L'information ne circule donc pas seulement dans le sens des flèches. Lorsque le graphe est plus complexe, la circulation de l'information dépend du type de connexion entre les nœuds et des informations disponibles sur ceux-ci (voir figure A.1).

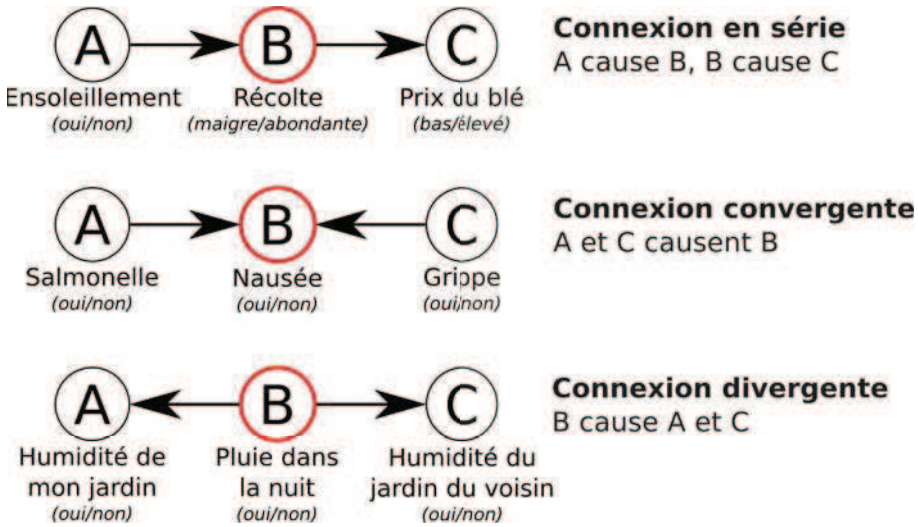


FIGURE A.1 – **Les différents modes de connexion entre les variables d'un graphe causal.** Chaque type de connexion est présenté sur trois variables discrètes A, B et C pour lesquelles des exemples sont donnés (les états possibles des variables sont indiqués entre parenthèses).

On considère ici trois variables A, B et C qui peuvent être connectées selon 3 schémas distincts. Chacun d'eux est décrit dans la figure A.1 où la variable B a une place centrale dans ces connexions : sa connaissance modifie la circulation de l'information entre les variables A et C.

Dans le but de simplifier l'écriture, dire que A influence B signifie que la connaissance de l'état de la variable A modifie les croyances sur l'état de la variable B.

- **Connexion en série**
  Dans cette configuration, A influence B qui elle-même influence C. Il est assez évident que la connaissance de l'état de A (resp. C) influence la connaissance sur B qui influence donc les croyances sur C (resp. A). Cependant, si l'état de B est connu alors la chaîne est bloquée : A et C deviennent indépendants. Par exemple, si la saison a été ensoleillée alors la récolte sera abondante, si la récolte a été abondante alors le prix du blé sera bas. Si l'on sait déjà que la récolte a été abondante, alors connaître l'ensoleillement n'apprend rien sur le prix du blé.
  Conclusion : l'information est transmise à travers une connexion en série lorsque l'état de la variable à l'intérieur de la connexion (B dans cet exemple) est inconnu.
- **Connexion convergente**
  Dans cette configuration, A et C causent B. Si aucune information n'est disponible sur B alors A et C sont indépendantes. Par contre, dès lors que l'état de B est connu, connaître l'état de l'un de ses parents influence la connaissance sur les autres causes possibles. Par exemple, si nous ne savons rien sur l'état nauséeux du patient alors la présence d'une infection aux salmonelles ne nous apprend rien sur l'état grippal du patient. Cependant, s'il a la nausée alors l'information d'une infection aux salmonelles rend moins possible l'hypothèse de la grippe.
  Conclusion : l'information est transmise à travers une connexion convergente seulement si la conséquence est connue.
- **Connexion divergente**
  Dans cette configuration, B cause A et C. Lorsque B est connue alors A et C sont indépendants. Par exemple, si mon jardin est humide alors j'ai tendance à croire qu'il a plu cette nuit et donc que la pelouse de mon voisin est également humide. En revanche, si je sais qu'il a plu cette nuit, je peux affirmer que la pelouse du voisin est humide et l'information relative à mon jardin n'y changera rien.
  Conclusion : l'information est transmise à travers une connexion divergente tant que l'on ne connaît pas la cause.

On remarque que dans chacun des cas, l'information peut être bloquée lorsque certaines informations sont connues. La conséquence de ce blocage est de rendre indépendantes certaines variables. C'est ce qu'on appelle la d-séparation.

Définition A.5. *Deux variables A et B dans un réseau causal sont **d-séparées** (d pour "graphe dirigé") si pour tous les chemins entre A et B, il existe une variable intermédiaire V (différente de A et B) telle que l'une des deux conditions suivantes est réalisée :*

- *la connexion est en série ou divergente et l'état de V est connu*
- *la connexion est convergente et ni V, ni aucun des descendants de V ne sont connus*

Ainsi, si A et B sont d-séparées, alors un changement dans la certitude de A n'a aucun effet sur la connaissance de B.

Dans le cadre des réseaux Bayésiens, la notion de croyance sur une variable A en fonction des autres variables du graphe est traduite en probabilités conditionnelles. Si B est le parent de A, $P(A \mid B)$ représente la force du lien entre A et B. Si C est également un parent de A alors il faut spécifier $P(A \mid B,C)$ pour définir l'interaction entre les variables A, B et C.

La notion de d-séparation présentée ci-dessus est en lien direct avec la définition d'indépendance conditionnelle (définition A.4). En effet, si deux variables A et B sont d-séparées par une variable C alors A et B sont indépendantes conditionnellement à C. Ainsi, la règle de chaîne générale (théorème A.2) est simplifiée dans le cadre des réseaux Bayésiens :

$$P(A_n \mid A_1,A_2,...A_{n-1}) = P(A_n \mid Pa(A_n))$$

La loi jointe globale est décomposée en un produit de lois locales. Cette décomposition de la loi jointe fait des réseaux Bayésien un modèle économique pour représenter les distributions de probabilités.

Annexe B

Données supplémentaires

Analyse de l'innovation

La figure B.1 présente les distributions des fréquences des différents types d'innovation protéique par espèce, en complément des figures 3.13 page 67 et 3.19 page 74.

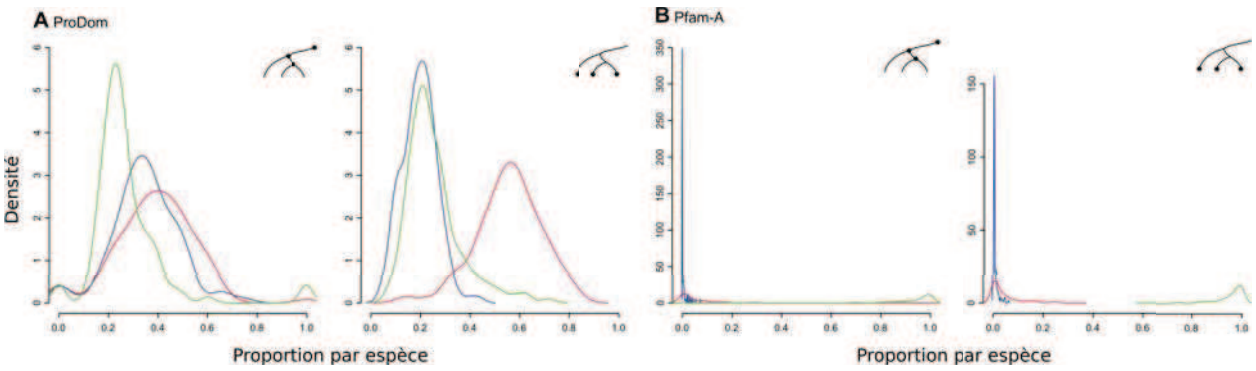


FIGURE B.1 – **Distributions des fréquences des différents types d'innovation protéique par espèce.** Les distributions des fréquences relatives sont reportées pour les protéines complètement innovées, partiellement innovées et réarrangées, respectivement en rouge, bleu et vert sur la base des arrangements en modules de ProDom (**A**) ou de Pfam (**B**). Les distributions ont été calculées sur les 114 espèces ancestrales et les 161 espèces contemporaines procaryotes (comme schématisé sur chaque panneau).

La figure B.2 présente les distributions associées aux valeurs moyennes données dans le tableau 3.7 page 75.

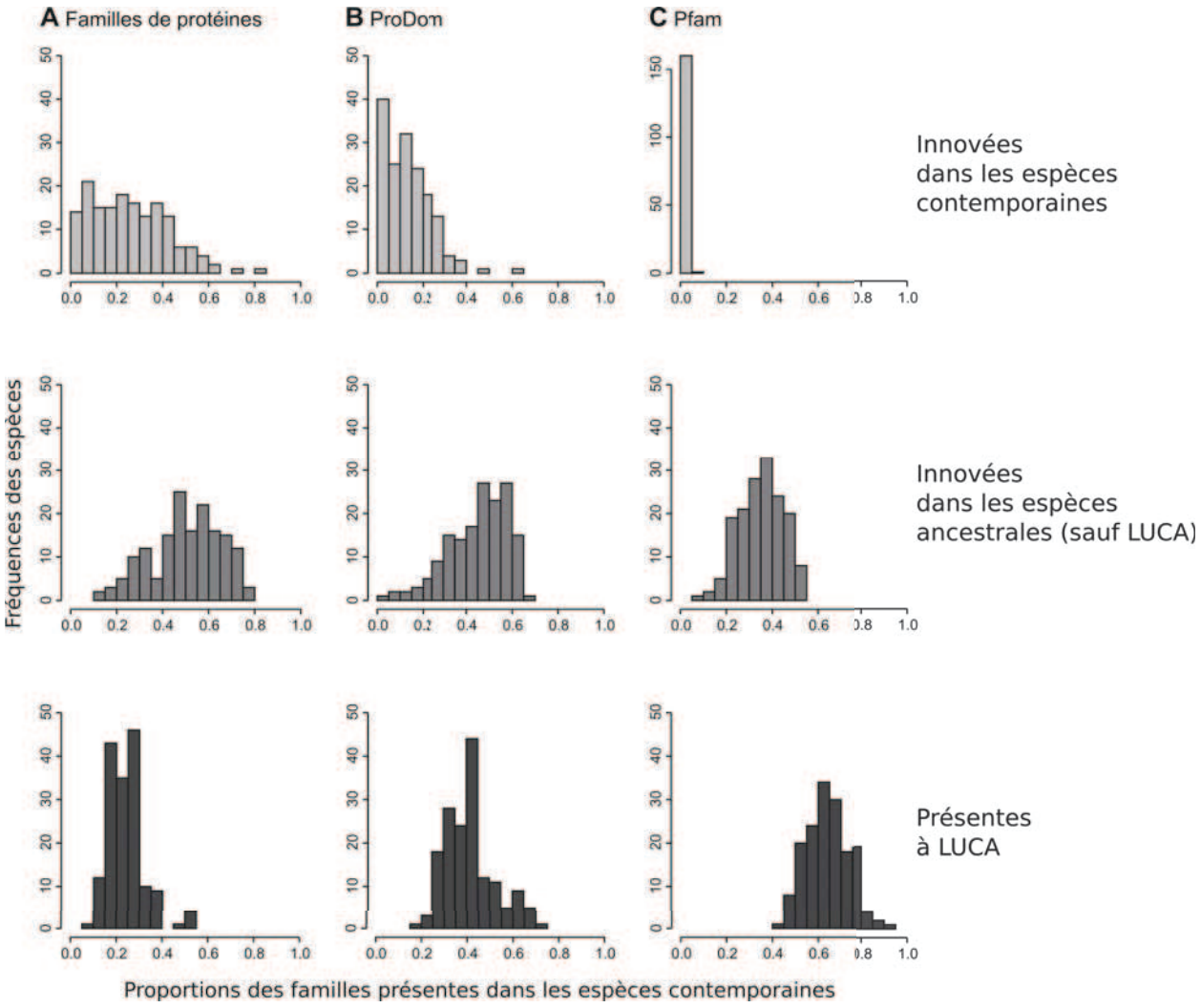


FIGURE B.2 – **Distributions de l'ancienneté des familles représentées dans les organismes procaryotes.** Les histogrammes présentent les proportions des familles de protéines (A), de modules ProDom (B) et de modules Pfam (C) trouvées dans les espèces actuelles qui sont les plus récentes (gris clair, panneaux du haut), qui sont intermédiaires (gris, panneaux au centre) et qui remontent à LUCA (gris foncé, panneaux du bas). Les distributions sont calculées pour les 161 espèces procaryotes.

La table B.1 présente les résultats obtenus avec les 4 paires d'espèces utilisées pour analyser le taux d'évolution des modules ProDom nouveaux (complément de la figure 3.25, page 82).

Paires d'espèces	Nœuds où sont apparus les modules	Chevauchement
Methanosarcina mazai	Methanosarcina	85%
Methanosarcina acetivorans	+ l'ancêtre d'Halobacterium sp.	99%
	+ l'ancêtre d'Archaeoglobus fulgidus	93%
	Euryarchaeota	100%
Mycobacterium avium	Mycobacterium	84%
Mycobacterium tuberculosis	Corynebacterium	85%
	Actinomycetales	91%
	Actinobacteridae	96%
	Bacteria	95%
Bacillus cereus ATCC10987	Bacillus cereus	93%
Bacillus anthracis	Bacillus	88%
	Bacillaceae	97%
	Bacillale	95%
	Bacilli-Firmicutes	96%
	Bacteria	96%
Escherichia coli K12	Ancêtre d'Escherichia coli + Salmonella typhi	79%
Salmonella typhi ATCC700931	+ ancêtre de Photorabdus, Yersinia et Erwinia	86%
	Gammaproteobacteria	87%
	+ ancêtre des Alpha et Betaproteobacteria	91%
	Proteobacteria	90%
	Bacteria	95%

TABLEAU B.1 – **Chevauchement entre les distributions des scores de similarité de modules ProDom récents et anciens.** Les scores de similarité ont été calculés à partir des alignements de paires de modules orthologues dans différentes paires d'espèces proches. Les modules évoluant lentement sont définis comme ayant un score de similarité supérieur au quantile à 5% de la distribution des scores de similarité des modules présents à LUCA. Le tableau donne la fraction des modules évoluant lentement ayant émergé dans des nœuds de plus en plus anciens (indiqués par des points rouges sur la figure C.2, page 141). Ces proportions ont tendance à augmenter avec la profondeur de l'arbre, rendant les estimations de l'innovation en domaines plus conservatives pour les nœuds anciens.

Répartition des innovations protéiques

Les prédictions de Pfam sur l'innovation protéique ont été comparées aux prédictions de ProDom. Les tableaux B.2 et B.3 présentent la répartition selon ProDom des protéines respectivement prédites innovées partiellement et totalement selon Pfam. Les tableaux B.4 et B.5 présentent la répartition selon ProDom des protéines respectivement non annotées et annotées avec les modules Pfam. La modularité et la couverture en acides aminés de ces familles ont été mesurées. La couverture correspond à la proportion moyenne d'acides aminés couverts par un module sur la séquence consensus de la famille de protéines et la modularité est le nombre moyen de modules par architecture.

(a) Classification selon ProDom des familles de protéines prédites partiellement innovées selon Pfam.

		Innovation totale	Innovation partielle	Réarrangement	Sans annotation
Espèces ancestrales	Nb familles	25	62	17	0
	Proportion	24%	60%	16%	0%
Espèces contemporaines	Nb familles	50	105	12	2
	Proportion	30%	62%	7%	1%

(b) Couverture et modularité selon ProDom et Pfam des différents sous-ensembles de familles de protéines.

		Innovation totale		Innovation partielle		Réarrangement	
		Pfam	ProDom	Pfam	ProDom	Pfam	ProDom
Espèces ancestrales	Couverture	63%	79%	64%	87%	77%	82%
	Modularité	2,3	4,5	4,0	5,3	2,6	3,7
Espèces contemporaines	Couverture	49%	72%	54%	75%	65%	67%
	Modularité	3,6	4,4	4,0	6,2	4,1	3,6

TABLEAU B.2 – **Classification et caractéristiques des protéines prédites partiellement innovées selon Pfam.** Les données de Pfam prédisent 104 (resp. 169) familles de protéines innovées partiellement dans les espèces ancestrales (resp. contemporaines). Le tableau (a) présente la répartition de ces familles en fonction des modules ProDom. Le tableau (b) présente la couverture et la modularité moyennes dans chacun des sous-groupes de familles décrits dans le tableau (a).

(a) Classification selon ProDom des familles de protéines prédites totalement innovées selon Pfam.

		Innovation totale	Innovation partielle	Réar-rangement	Sans annotation
Espèces ancestrales	Nb familles	1 398	150	113	11
	Proportion	83%	9%	7%	1%
Espèces contemporaines	Nb familles	447	155	122	11
	Proportion	61%	21%	17%	1%

(b) Couverture et modularité selon ProDom et Pfam.

		Innovation totale		Innovation partielle		Réarrangement	
		Pfam	ProDom	Pfam	ProDom	Pfam	ProDom
Espèces ancestrales	Couverture	81%	90%	68%	89%	83%	89%
	Modularité	1,3	2,9	1,3	3,4	1,0	1,6
Espèces contemporaines	Couverture	67%	84%	58%	80%	71%	81%
	Modularité	1,3	2,0	1,3	3,5	1,1	1,5

TABLEAU B.3 – **Classification et caractéristiques des protéines prédites totalement innovées selon Pfam.** Les données de Pfam prédisent 1 672 (resp. 735) familles de protéines innovées totalement dans les espèces ancestrales (resp. contemporaines). Le tableau (a) présente la répartition de ces familles en fonction des modules ProDom. Le tableau (b) présente la couverture et la modularité moyennes dans chacun des sous-groupes de familles décrits dans le tableau (a).

		Innovation totale	Innovation partielle	Réar-rangement	Sans annotation
Espèces ancestrales	Nb familles	7 270	1 627	1 937	93
	Proportion	66%	15%	18%	1%
	Couverture	93%	90%	90%	
	Modularité	1,2	2,7	1,2	
Espèces contemporaines	Nb familles	51 936	7 708	11 150	1 606
	Proportion	72%	11%	15%	2%
	Couverture	96%	90%	89%	
	Modularité	1,1	2,6	1,2	

TABLEAU B.4 – **Classification et caractéristiques des protéines sans annotation avec Pfam.** Il y a 10 927 (resp. 72 400) familles de protéines innovées dans une espèce ancestrale (resp. contemporaine) qui n'ont aucune annotation avec Pfam. Le tableau présente la répartition de ces familles en fonction des modules ProDom ainsi que la couverture et la modularité moyennes dans chacun des sous-groupes de familles décrits dans le tableau.

		Innovation totale	Innovation partielle	Réar-rangement	Sans annotation
Espèces ancestrales	Nb familles	2 957	4 870	2 742	122
	Proportion	28%	45%	26%	1%
	Couverture	79%	83%	80%	
	Modularité	2,3	3,8	1,9	
Espèces contemporaines	Nb familles	6 497	11 958	9 125	742
	Proportion	23%	42%	32%	3%
	Couverture	71%	80%	74%	
	Modularité	1,5	3,7	1,8	

TABLEAU B.5 – **Classification et caractéristiques des protéines annotées avec Pfam.** Il y a 10 691 (resp. 28 322) familles de protéines innovées dans une espèce ancestrale (resp. contemporaine) qui ont une annotation avec Pfam. Le tableau présente la répartition de ces familles en fonction des modules ProDom ainsi que la couverture et la modularité moyennes dans chacun des sous-groupes de familles décrits dans le tableau.

Évolution des familles de protéines et de domaines

FIGURE B.3 – **Évolution des répertoires des modules protéiques (A, C) et des familles de protéines (B, D).** Les diagrammes circulaires reflètent le nombre de familles inférées présentes à chaque nœud. Les familles héritées de leur parent sont représentées par un secteur gris. Les familles gagnées pouvant être retrouvées dans d'autres clades sont inférées transférées horizontalement et indiquées par le secteur jaune. Toutes les autres familles sont prédites innovées. Elles sont indiquées en rouge pour les familles de domaines (A, C) et en rouge, bleu et vert pour les familles de protéines qui sont respectivement complètement innovées, partiellement innovées et réarrangées (B, D). Les familles de protéines sans annotation avec les modules ProDom (B) et Pfam (D) sont représentées par un secteur noir. Les jeux de données de domaines utilisés sont extraits de ProDom (A, B) et Pfam (C, D). Pour une meilleure lisibilité, l'arbre des espèces a été divisé en plusieurs sous-arbres.

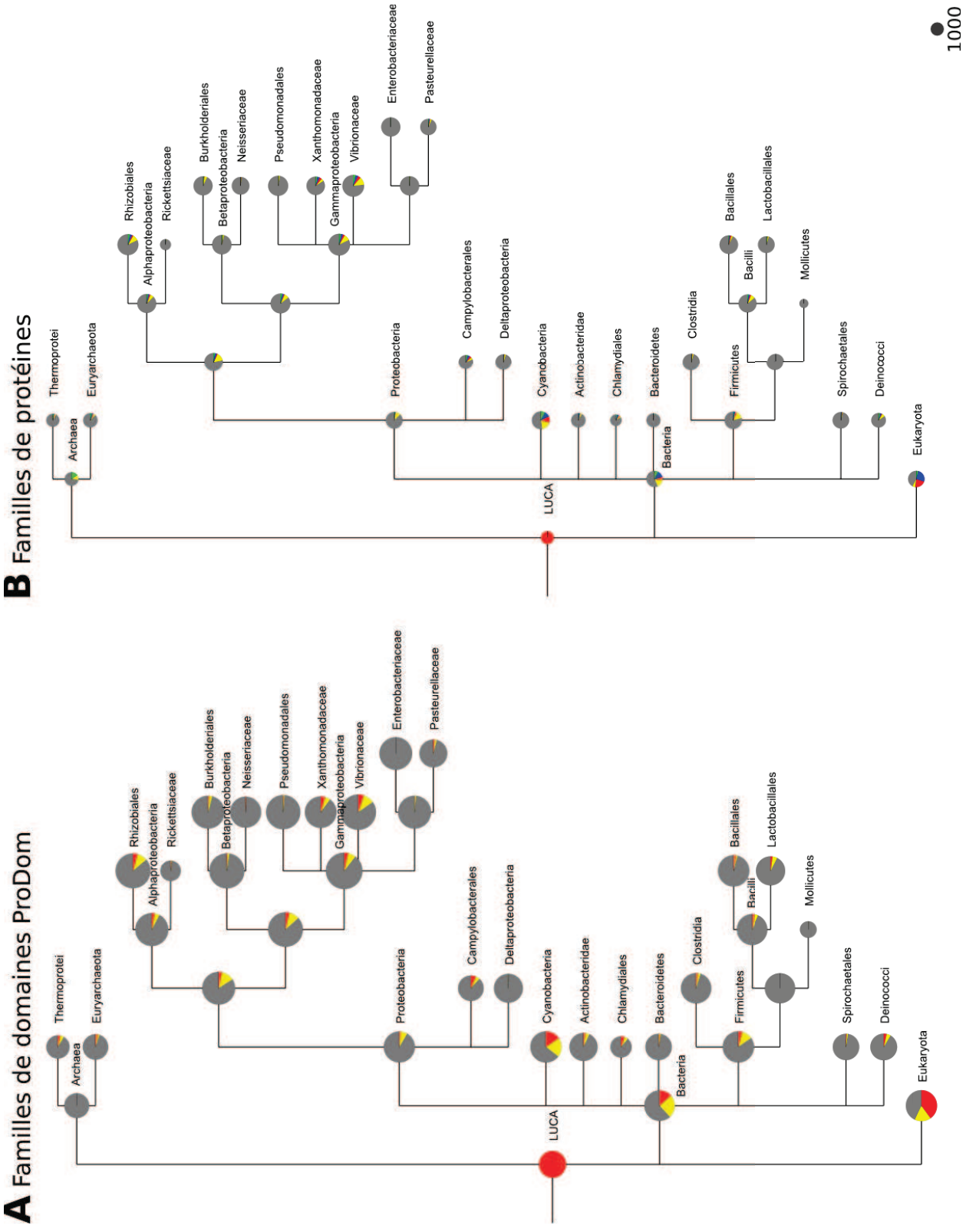


FIGURE B.3 – **Évolution des répertoires des nœuds les plus anciens.**

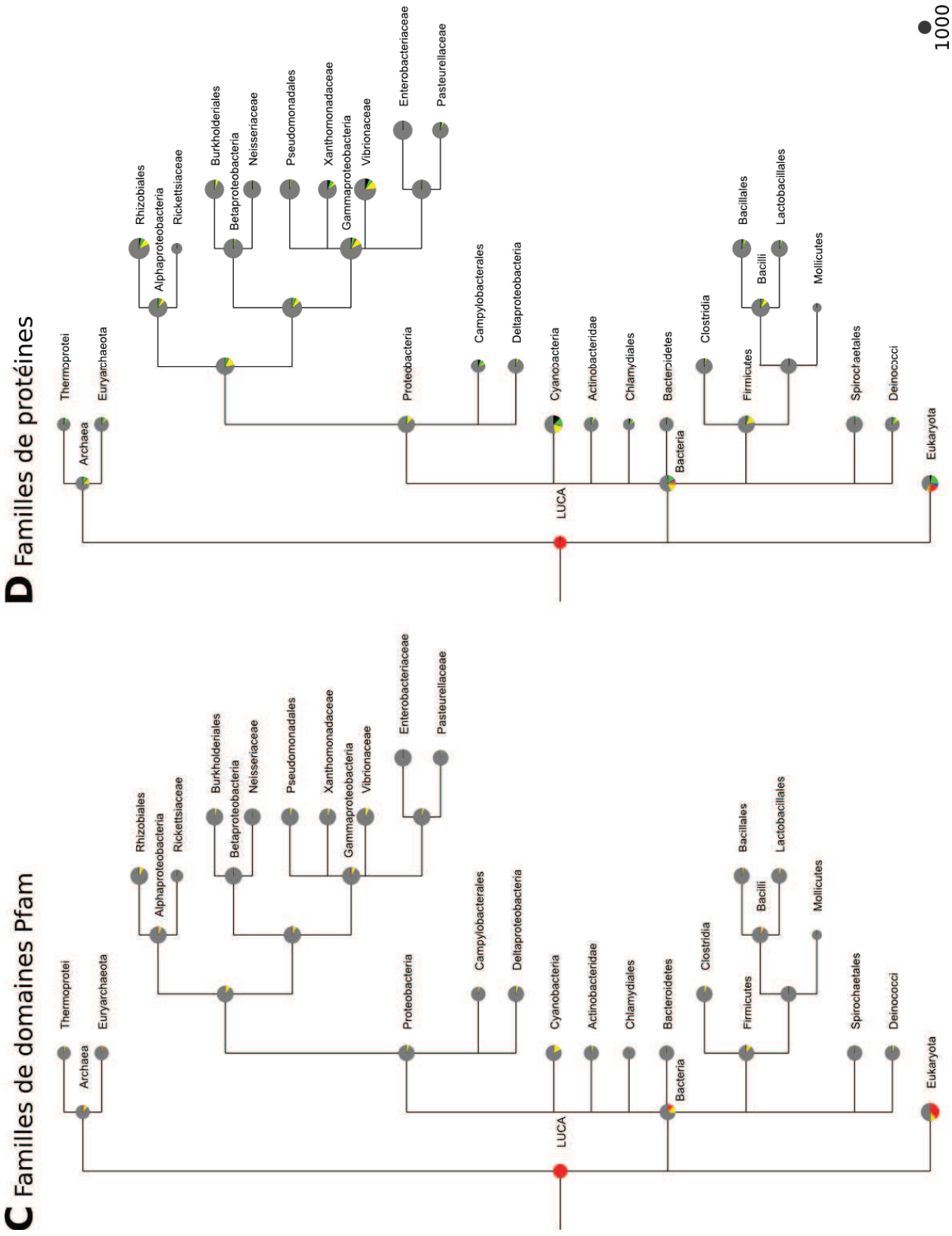


FIGURE B.3 – **Évolution des répertoires des nœuds les plus anciens.**

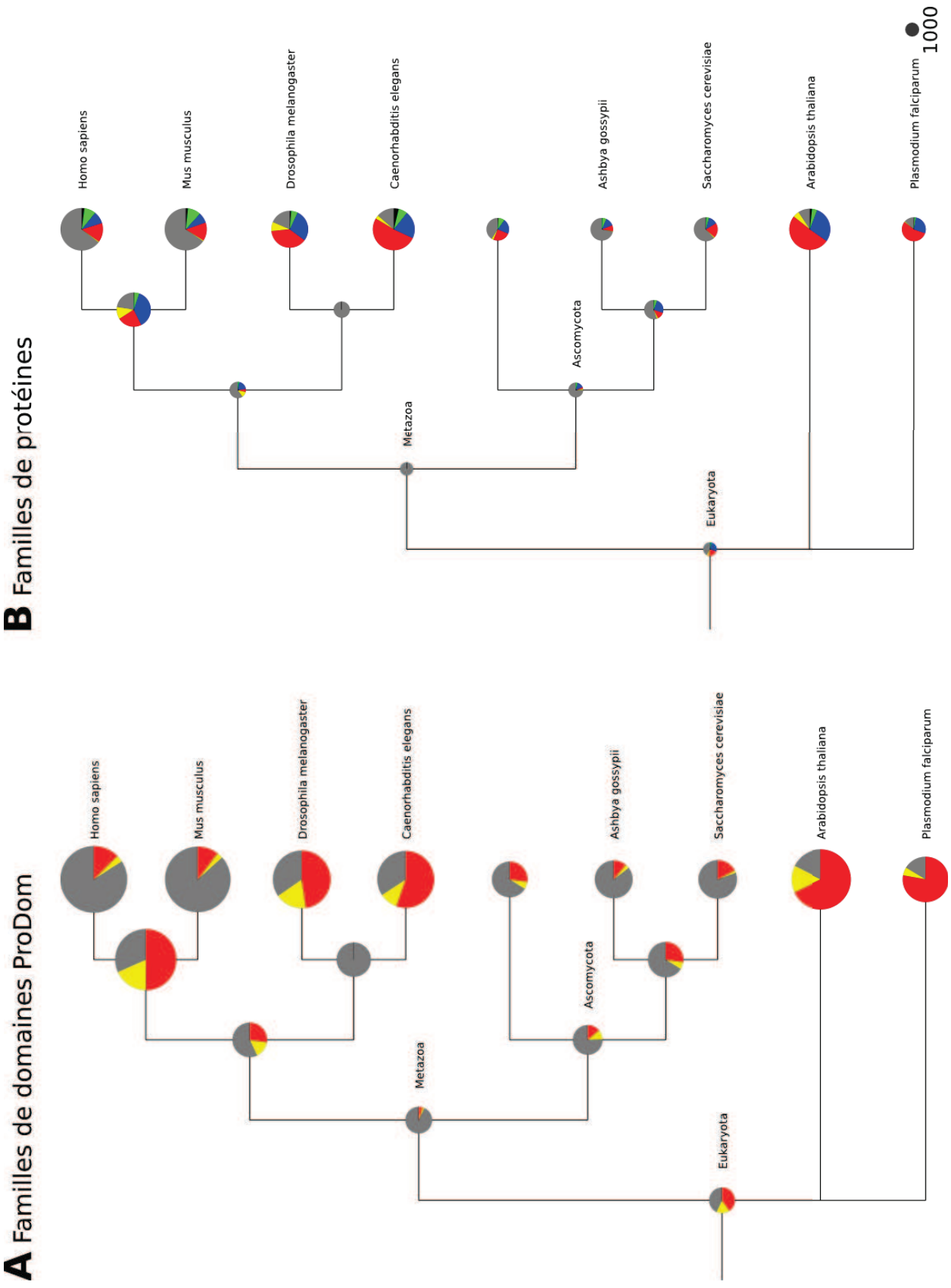


FIGURE B.3 – **Évolution des répertoires des Eucaryotes.**

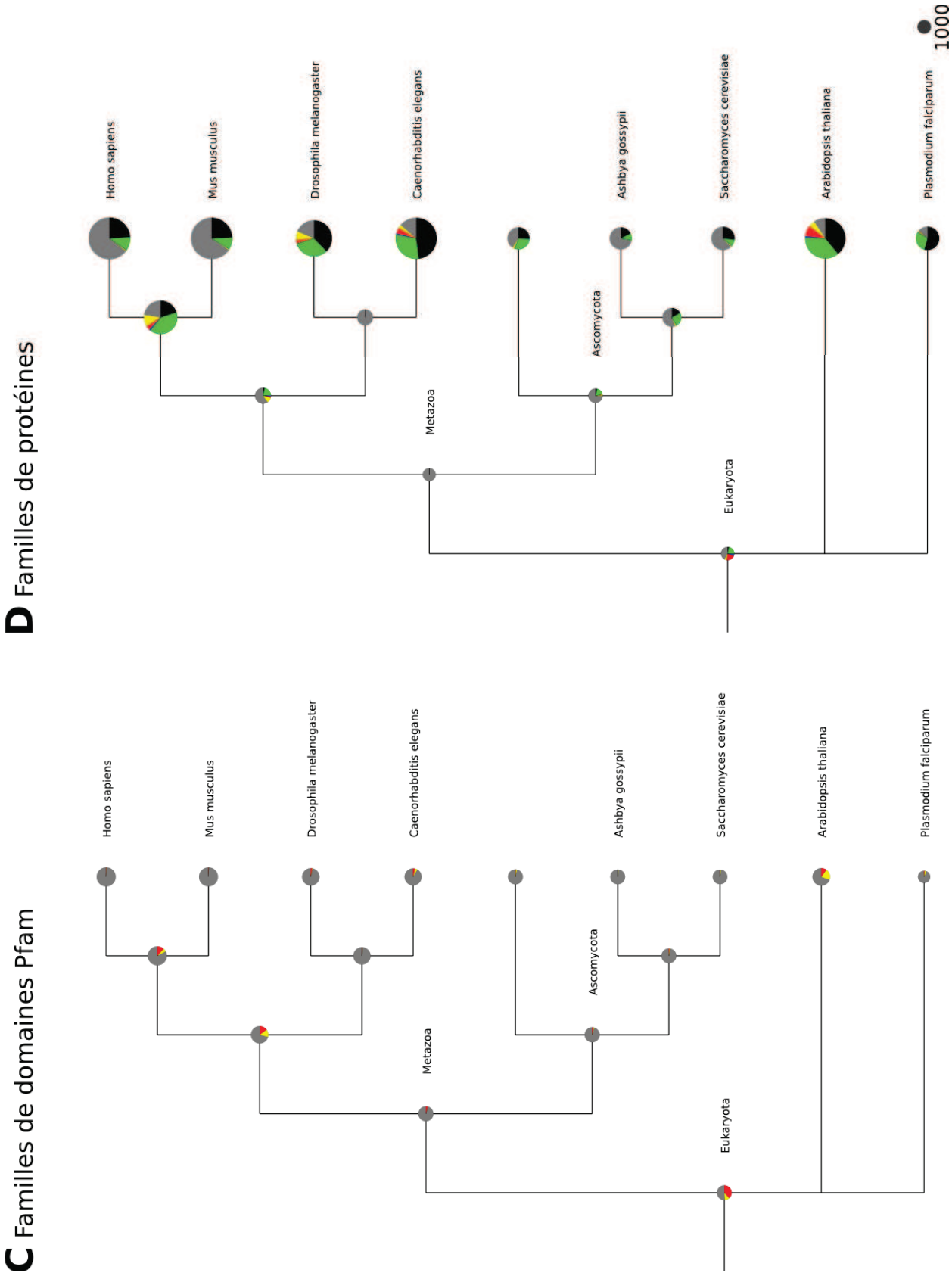


FIGURE B.3 – **Évolution des répertoires des Eucaryotes.**

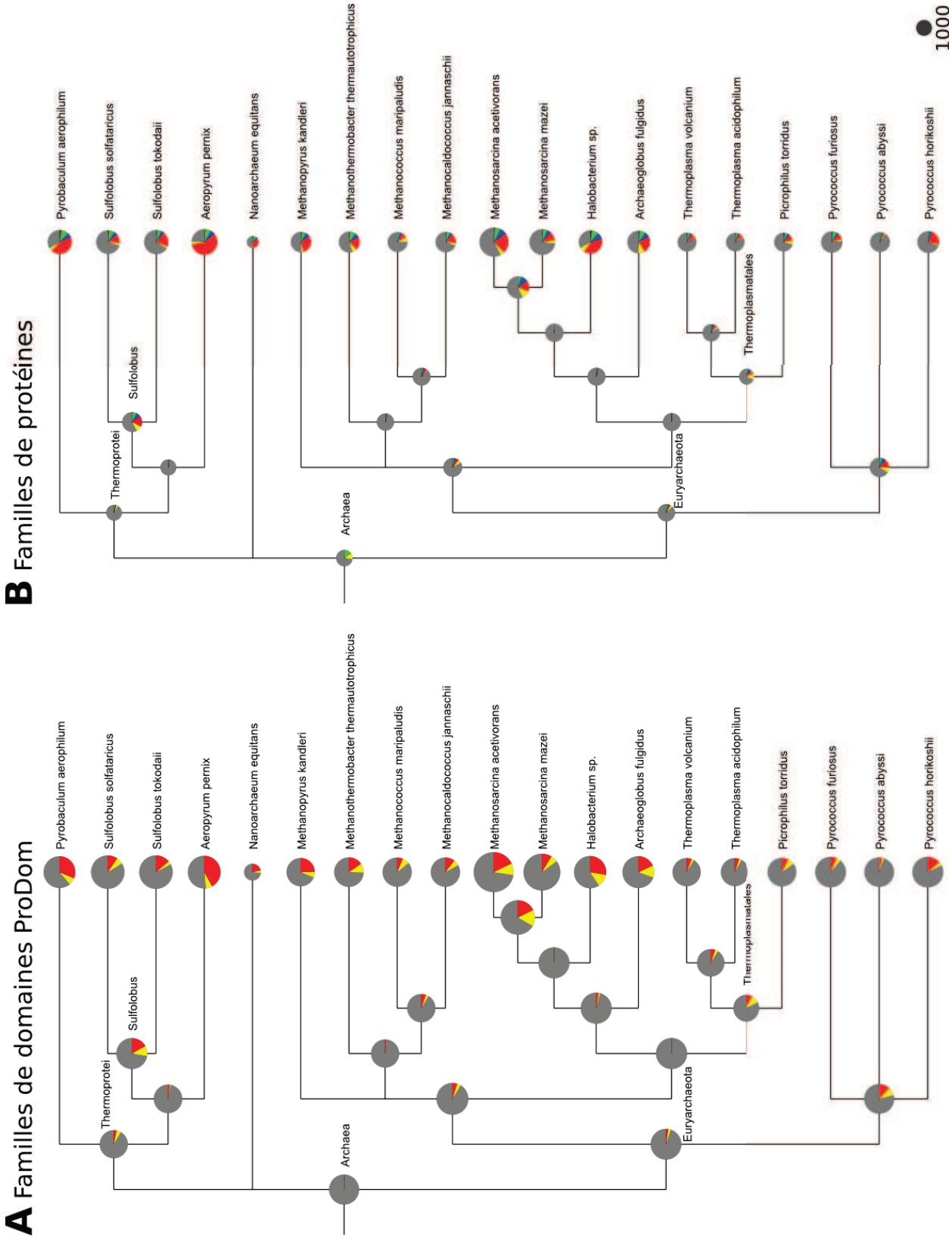


FIGURE B.3 – **Évolution des répertoires des Archées.**

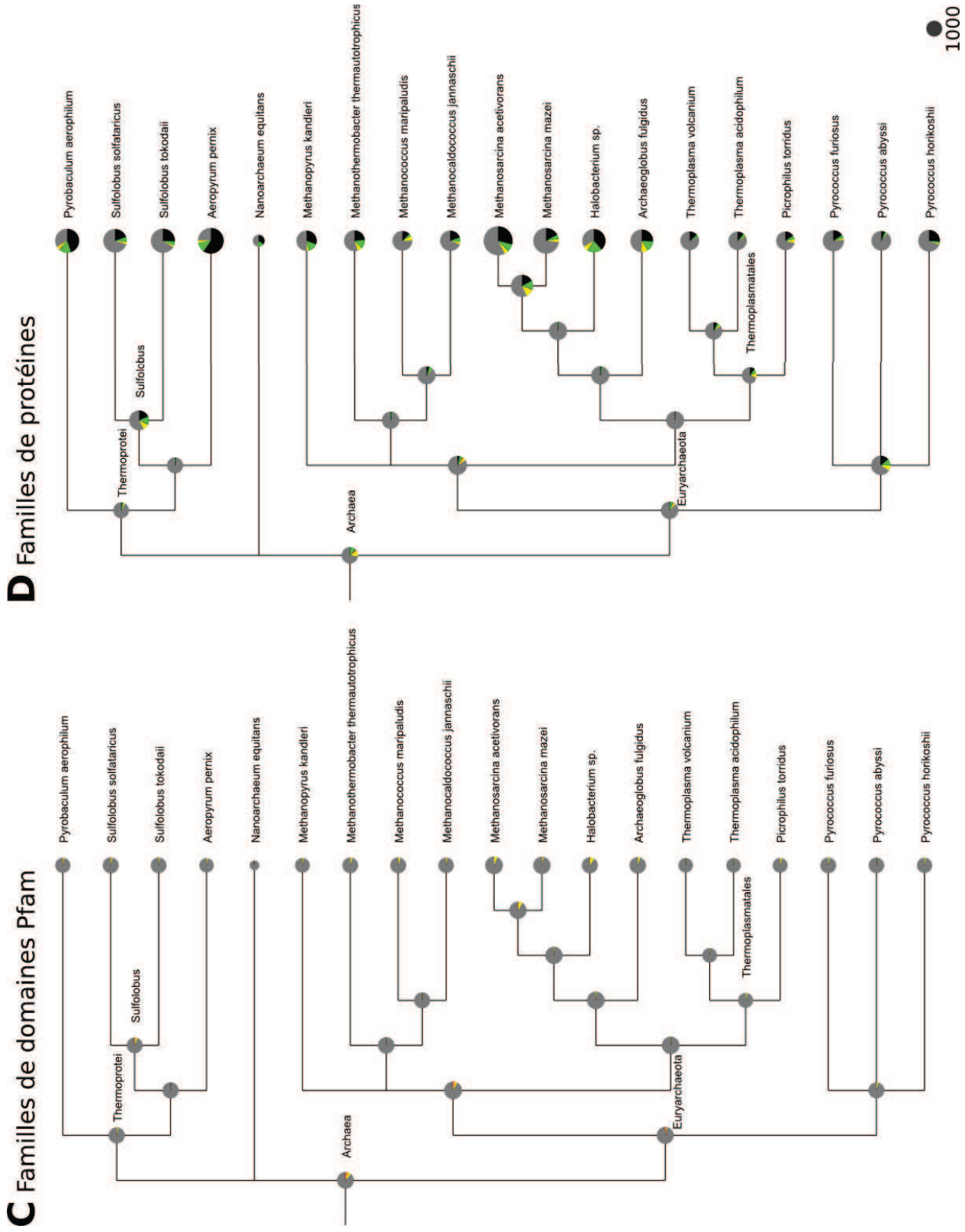


FIGURE B.3 – **Évolution des répertoires des Archées.**

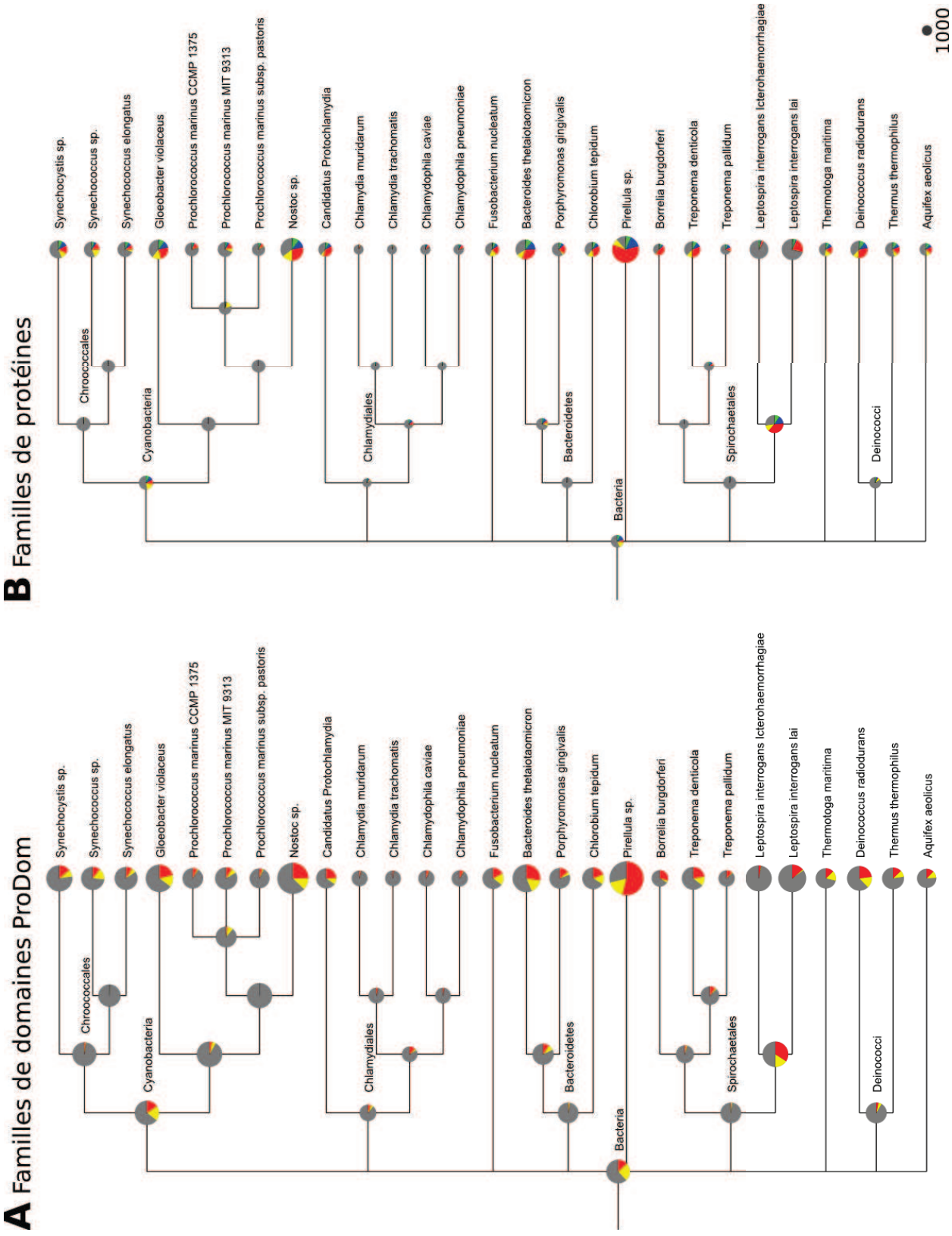


FIGURE B.3 – **Évolution des répertoires des Bactéries à l'exception des sous-arbres des Protéobactéries, des Firmicutes et des Actinobactéries.**

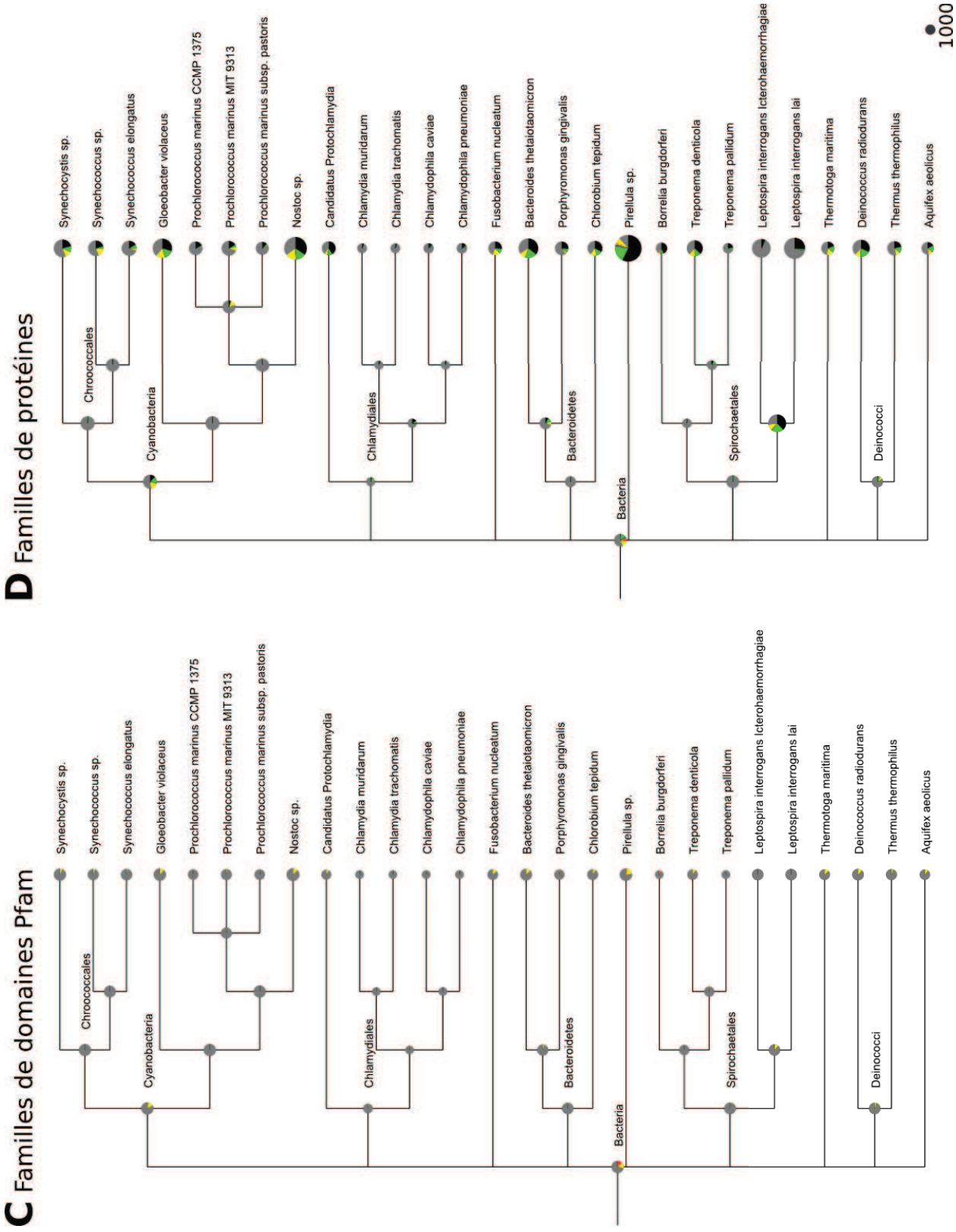


FIGURE B.3 – **Évolution des répertoires des Bactéries à l'exception des sous-arbres des Protéobactéries, des Firmicutes et des Actinobactéries.**

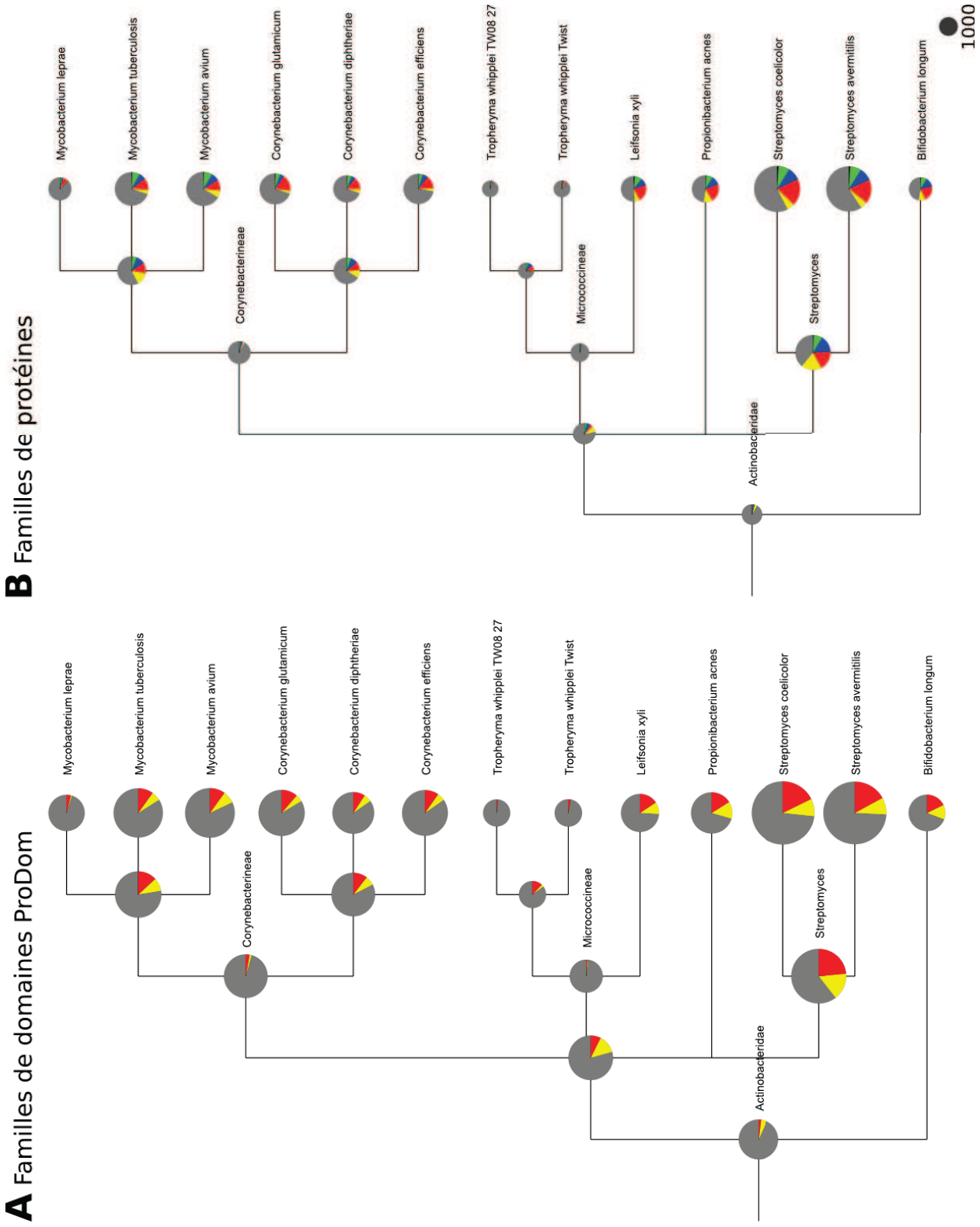


FIGURE B.3 – **Évolution des répertoires des Actinobactéries.**

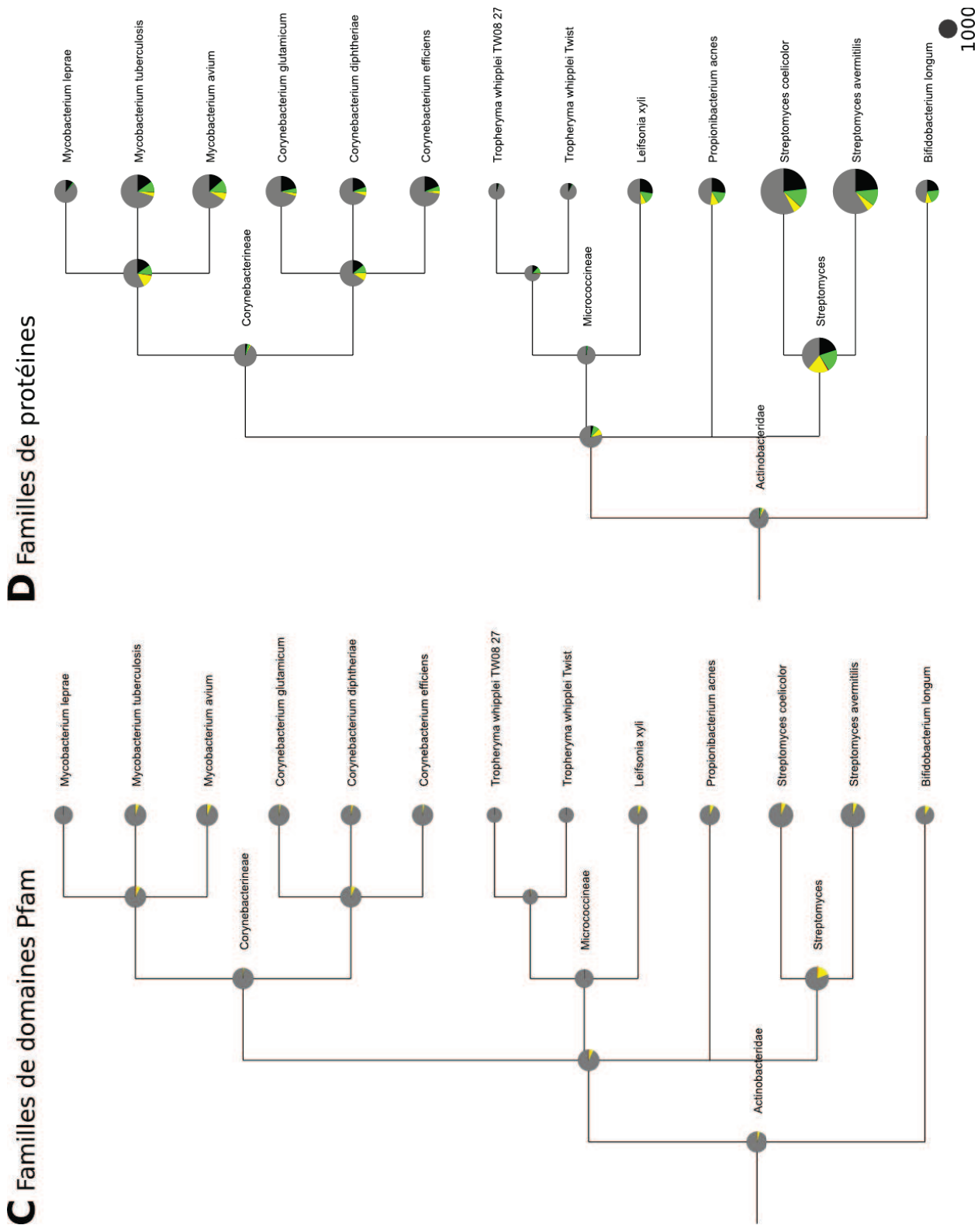


FIGURE B.3 – **Évolution des répertoires des Actinobactéries.**

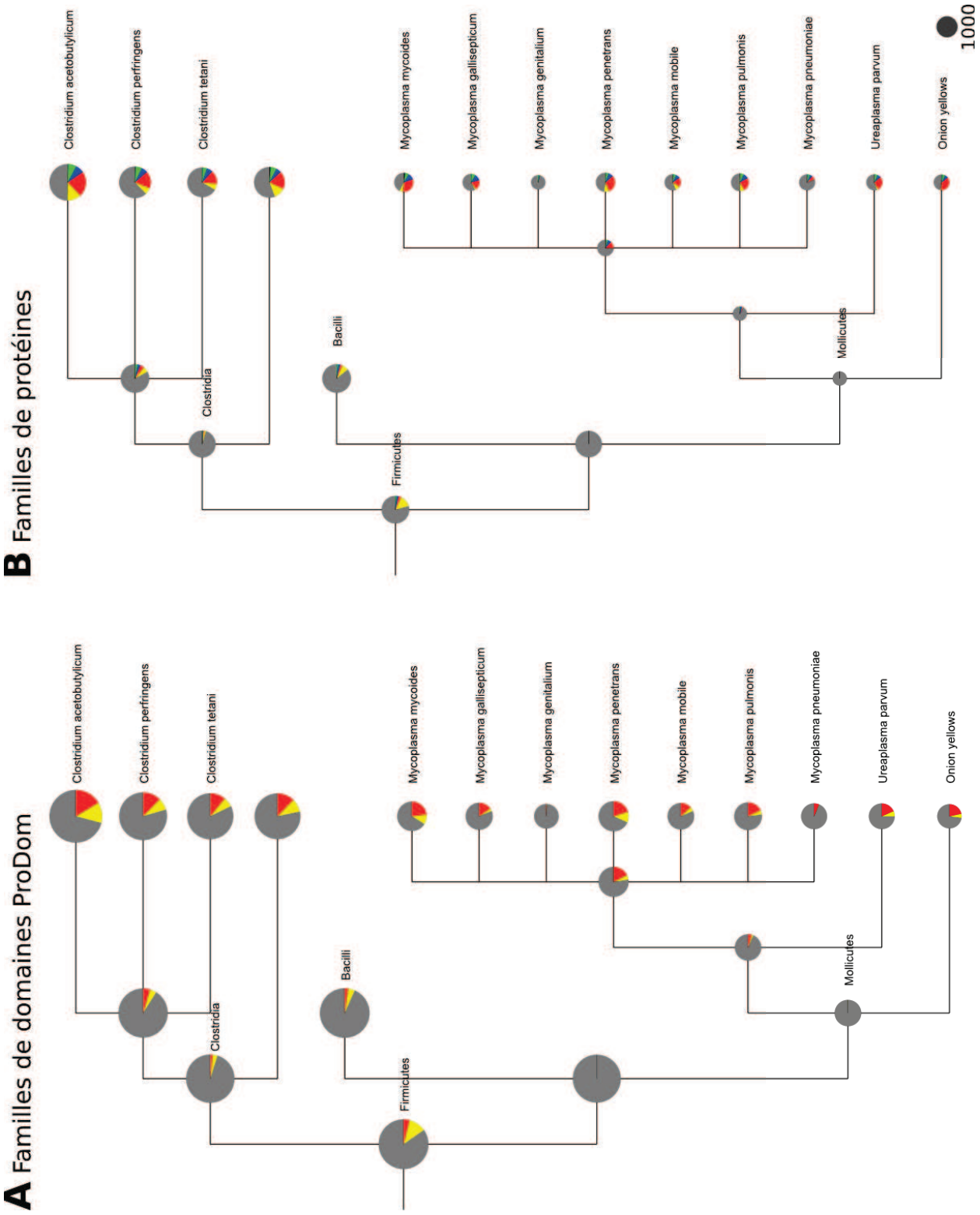


FIGURE B.3 – **Évolution des répertoires des Firmicutes à l'exception du sous-arbre des Bacilli.**

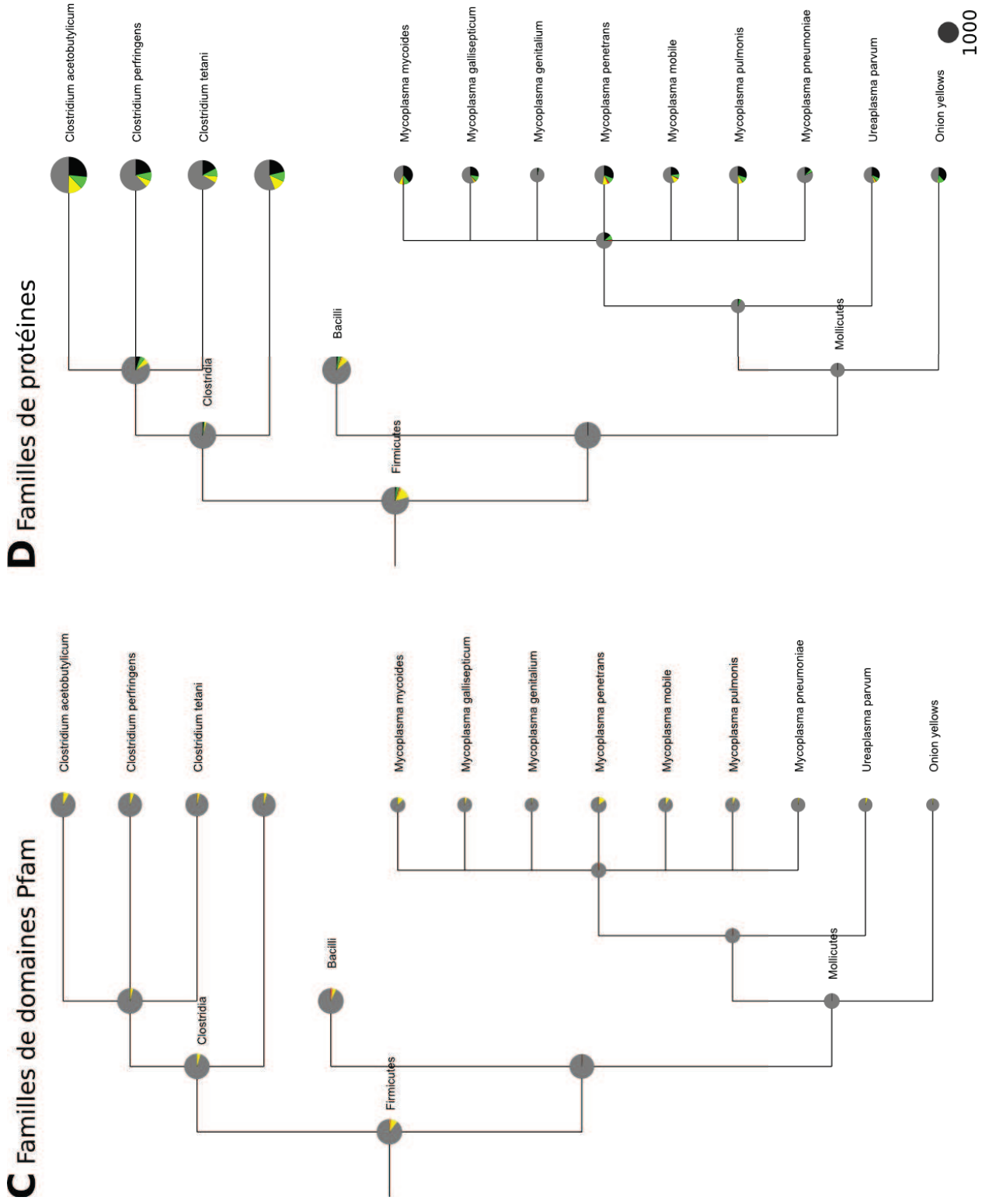


FIGURE B.3 – **Évolution des répertoires des Firmicutes à l'exception du sous-arbre des Bacilli.**

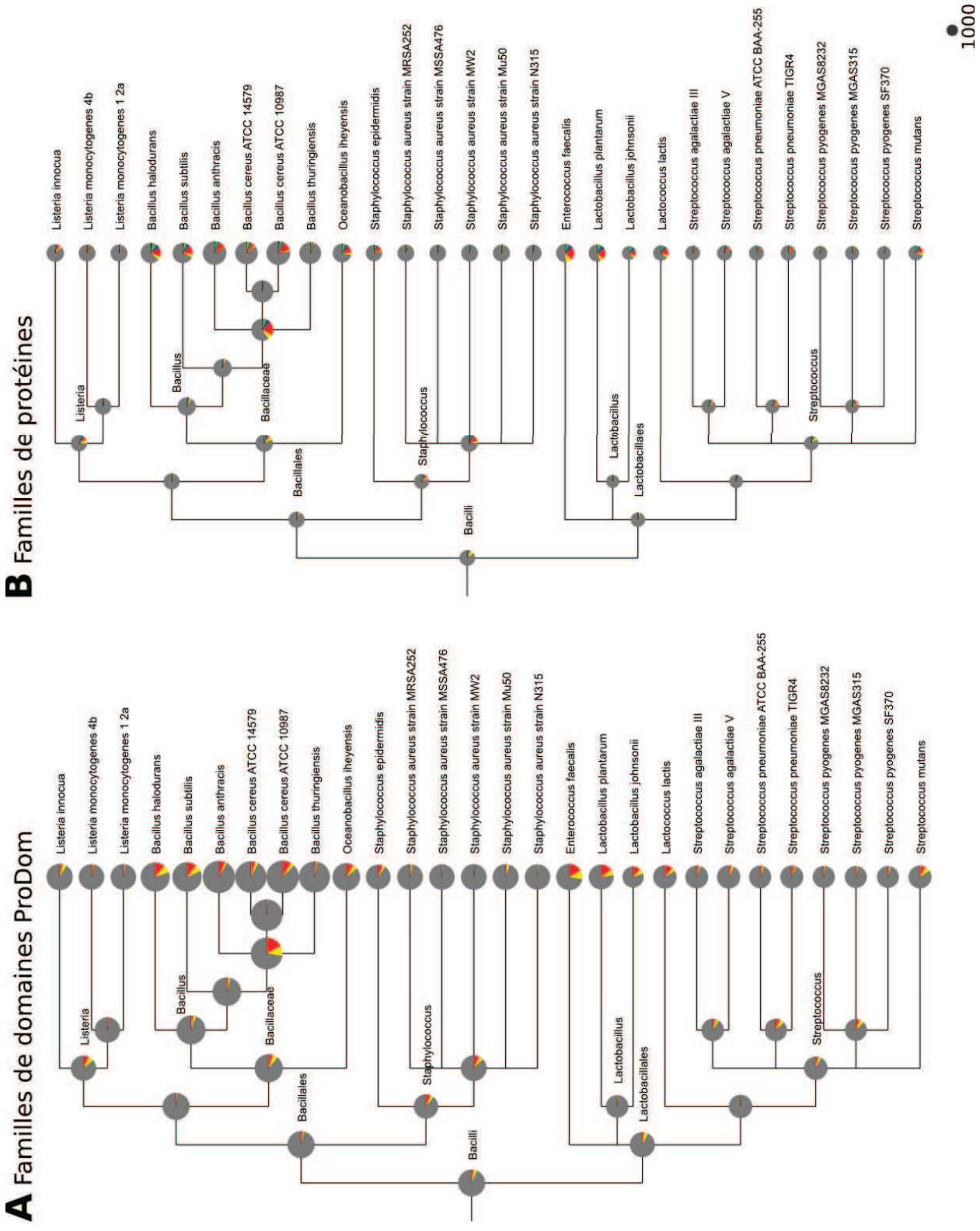


FIGURE B.3 – **Évolution des répertoires des Bacilli.**

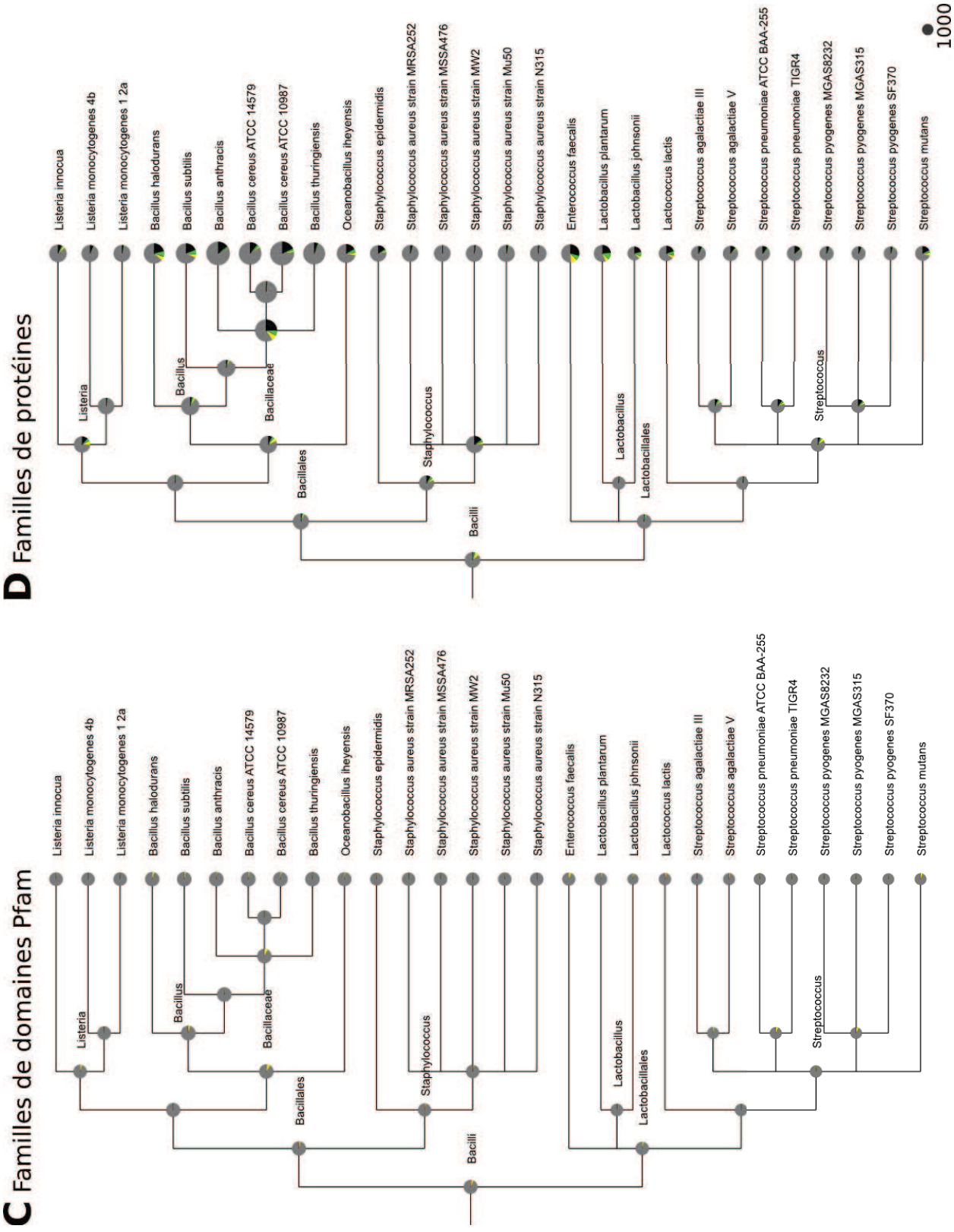


FIGURE B.3 – **Évolution des répertoires des Bacilli.**

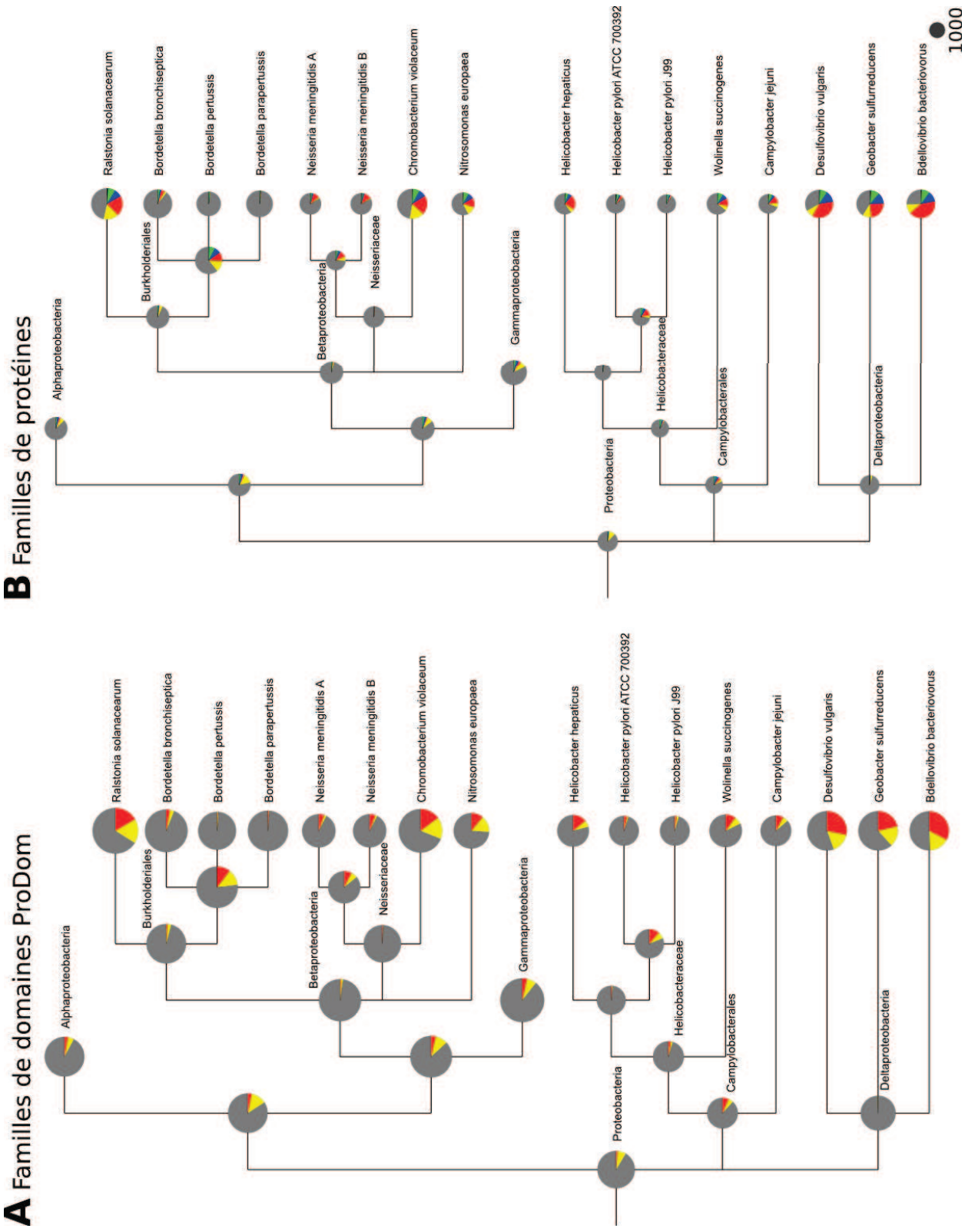


FIGURE B.3 – **Évolution des répertoires des Protéobactéries à l'exception des sous-arbres des Alpha et Gammaprotéobactéries.**

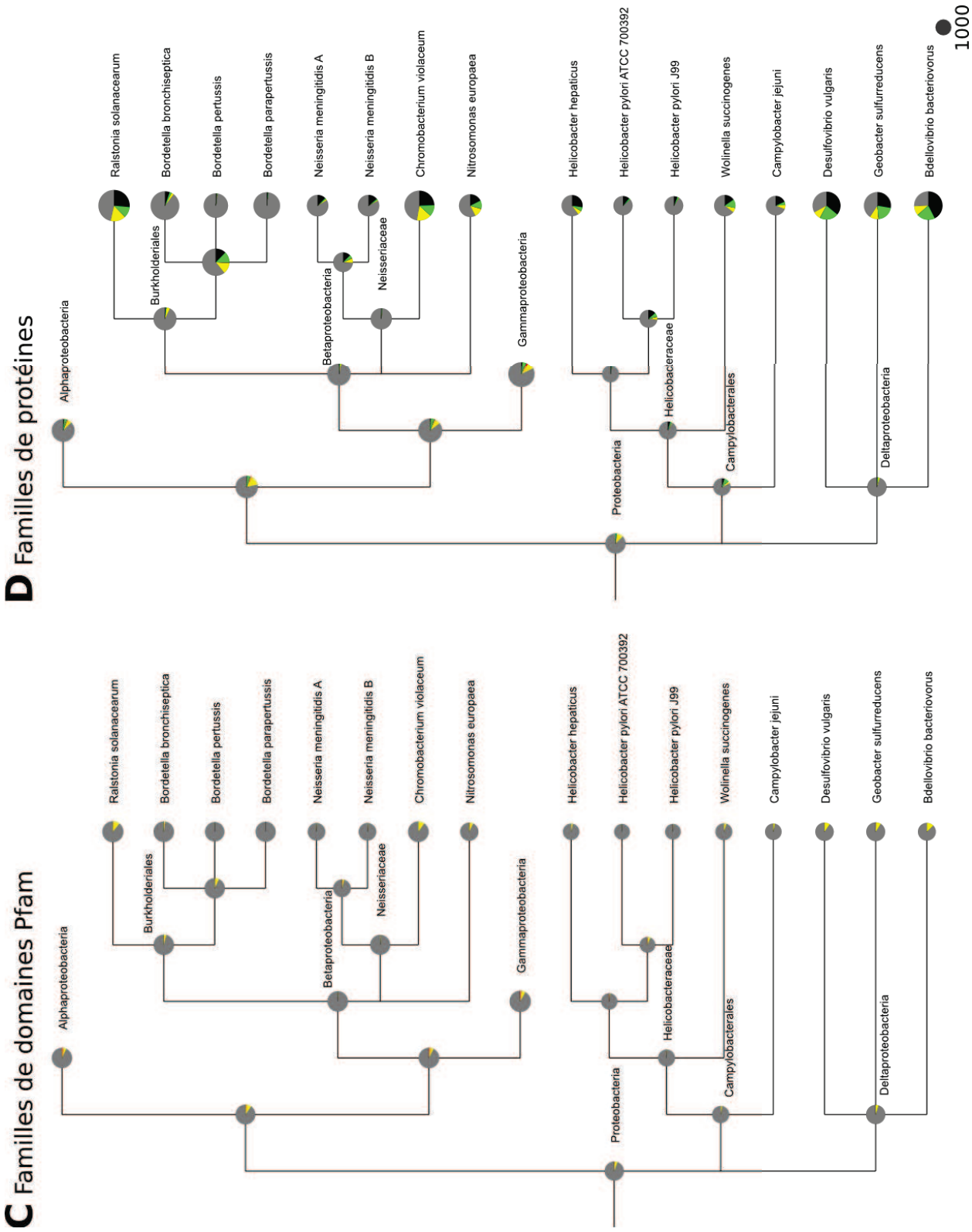


FIGURE B.3 – **Évolution des répertoires des Protéobactéries à l'exception des sous-arbres des Alpha et Gammaprotéobactéries.**

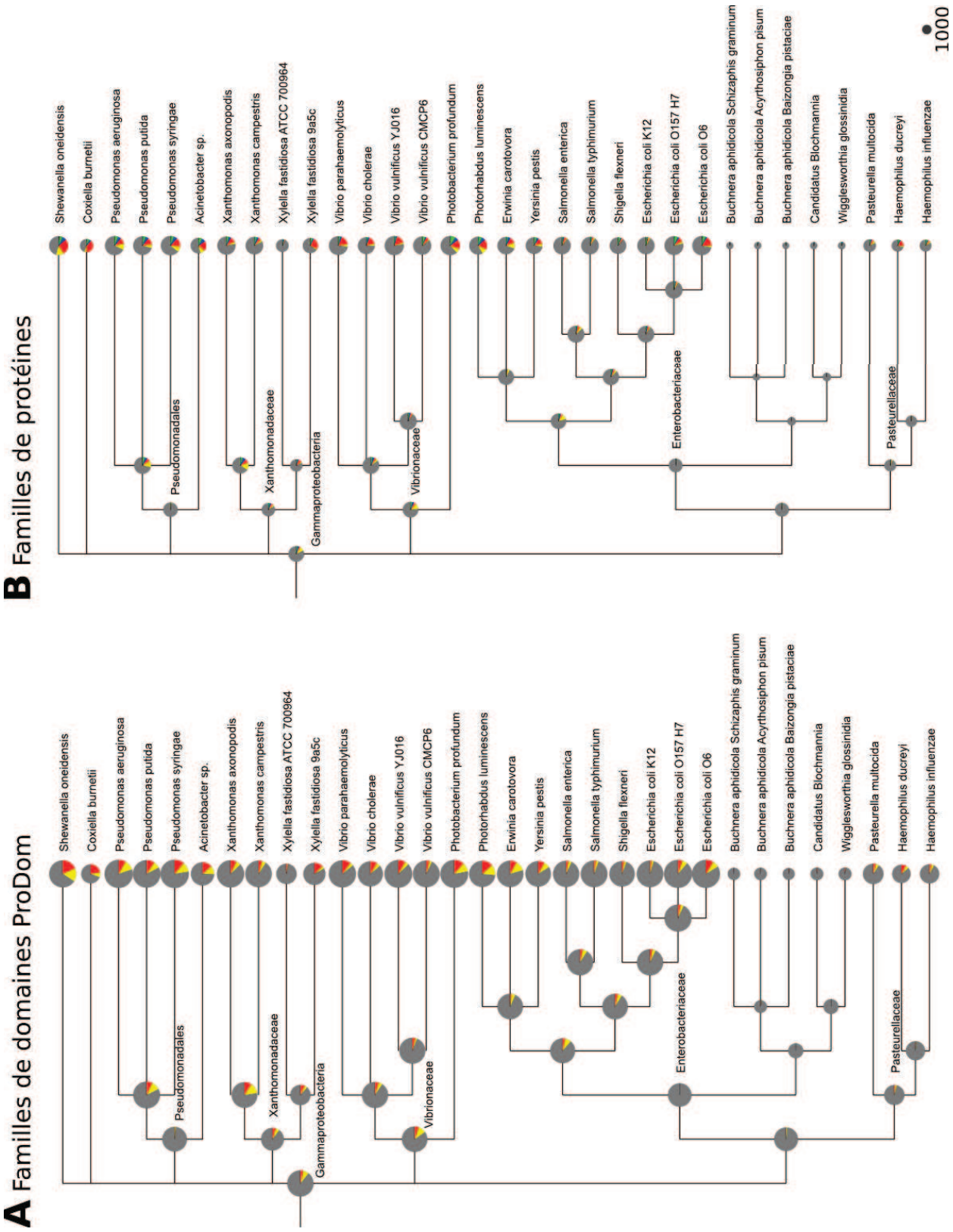


FIGURE B.3 – **Évolution des répertoires des Gammaprotéobactéries.**

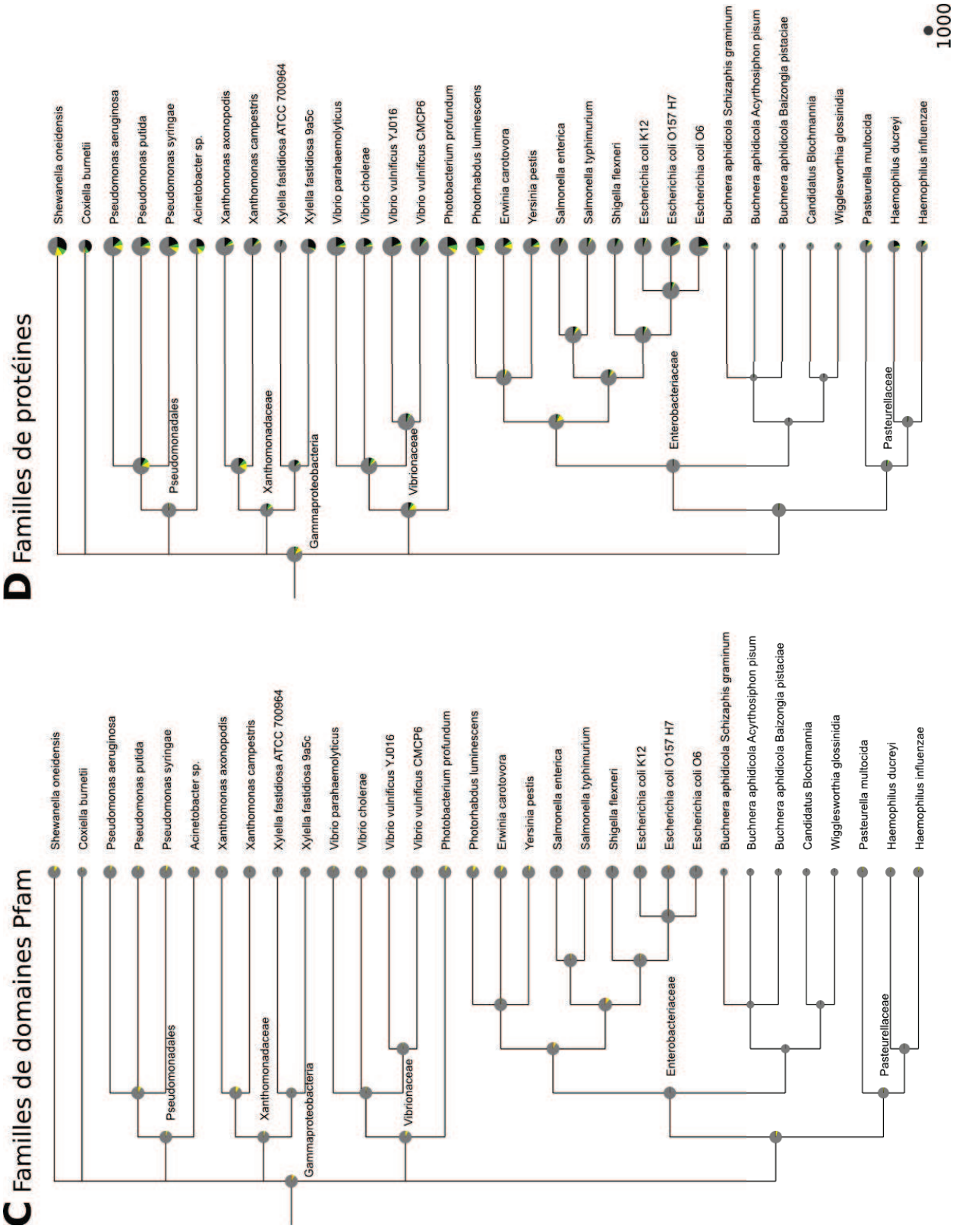


FIGURE B.3 – **Évolution des répertoires des Gammaprotéobactéries.**

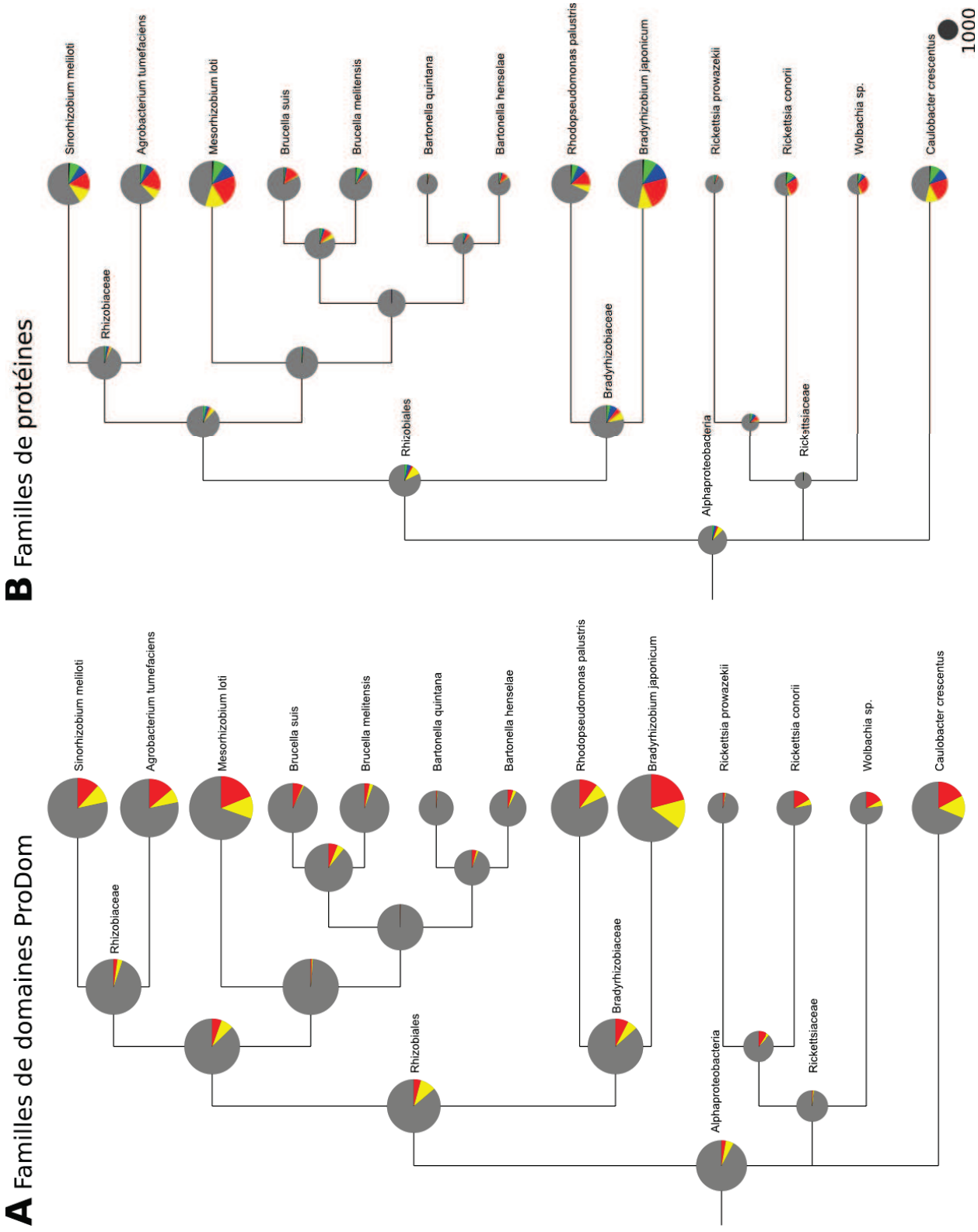


FIGURE B.3 – **Évolution des répertoires des Alphaprotéobactéries.**

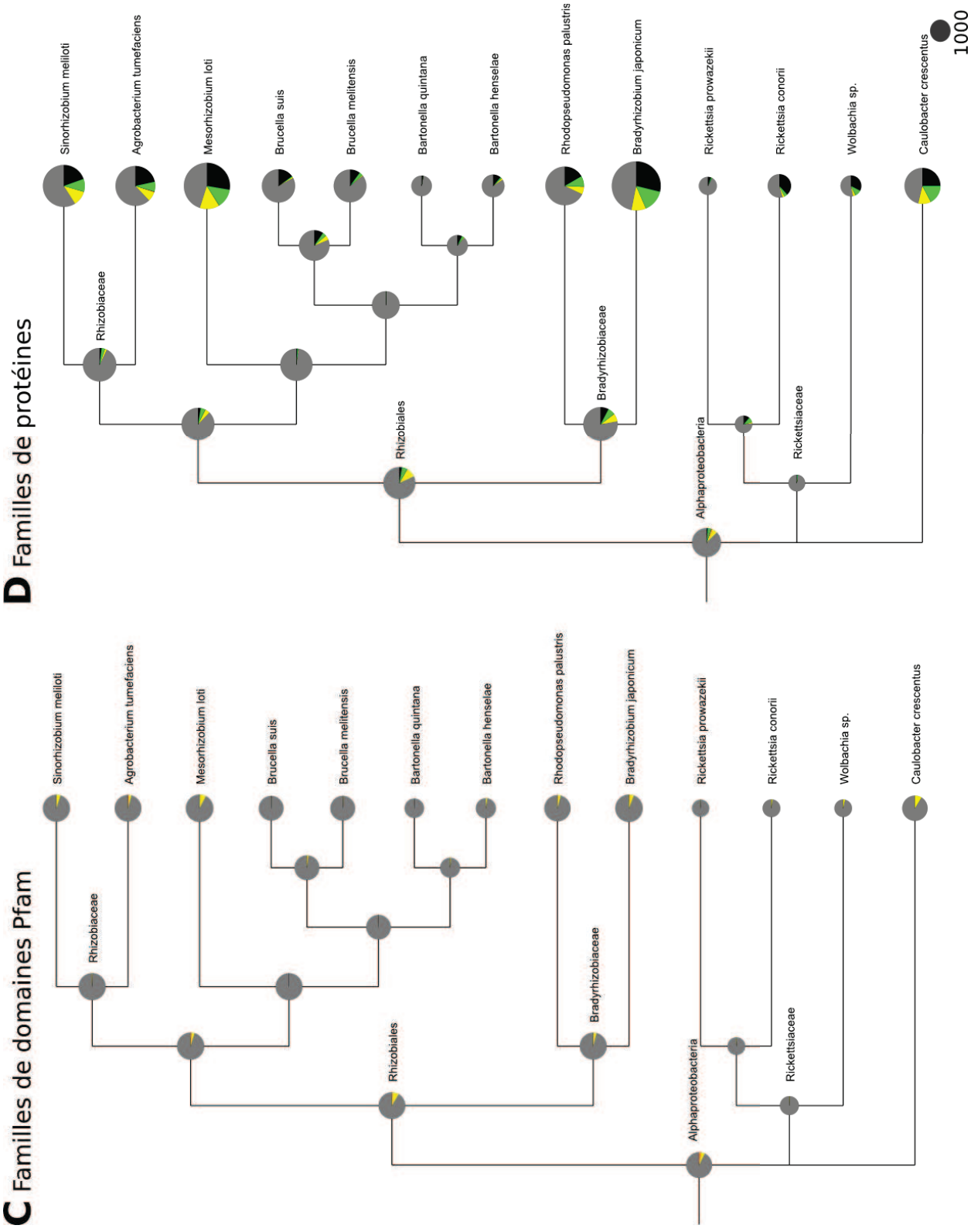


FIGURE B.3 – **Évolution des répertoires des Alphaprotéobactéries.**

Annexe C

Liste des espèces et arbres phylogénétiques utilisés

TABLEAU C.1 – **Liste des 170 espèces avec leur identifiant taxonomique.** La coloration des différents clades correspond à celle utilisée dans les arbres des espèces présentés dans les figures C.1 et C.2

TaxID	Nom
Eucaryotes	
3702	Arabidopsis thaliana
284811	Ashbya gossypii ATCC10895
6239	Caenorhabditis elegans
7227	Drosophila melanogaster
9606	Homo sapiens
10090	Mus musculus
36329	Plasmodium falciparum 3D7
4932	Saccharomyces cerevisiae
284812	Schizosaccharomyces pombe 972h-
Archeae	
272557	Aeropyrum pernix K1
224325	Archaeoglobus fulgidus DSM4304
64091	Halobacterium sp. NRC-1
243232	Methanocaldococcus jannaschii

TaxID	Nom
39152	Methanococcus maripaludis
190192	Methanopyrus kandleri AV19
188937	Methanosarcina acetivorans C2A
192952	Methanosarcina mazei Go1
187420	Methanothermobacter thermautotrophicus
228908	Nanoarchaeum equitans Kin4-M
263820	Picrophilus torridus DSM9790
13773	Pyrobaculum aerophilum
272844	Pyrococcus abyssi GE5
186497	Pyrococcus furiosus DSM3638
70601	Pyrococcus horikoshii OT3
2287	Sulfolobus solfataricus
273063	Sulfolobus tokodaii str.7
273075	Thermoplasma acidophilum DSM1728
273116	Thermoplasma volcanium GSS1

C. Liste des espèces et arbres phylogénétiques utilisés

TaxID	Nom
Bactérie	
224324	Aquifex aeolicus VF5
190304	Fusobacterium nucleatum ATCC 25586
117	Pirellula sp.
243274	Thermotoga maritima MSB8
	Deinococci
1299	Deinococcus radiodurans
262724	Thermus thermophilus HB27
	Spirochaetales
224326	Borrelia burgdorferi B31
267671	Leptospira interrogans icterohaemorrhagiae
189518	Leptospira interrogans lai
243275	Treponema denticola ATCC35405
243276	Treponema pallidum Nichols
	Chlamydiales
243161	Chlamydia muridarum Nigg
272561	Chlamydia trachomatis D-UW-3-CX
227941	Chlamydophila caviae GPIC
115713	Chlamydophila pneumoniae CWL029
264201	Protochlamydia amoebophila UWE25
	Bacteroidetes
226186	Bacteroides thetaiotaomicron VPI-5482
194439	Chlorobium tepidum TLS
242619	Porphyromonas gingivalis W83
	Cyanobacteries
251221	Gloeobacter violaceus PCC7421
103690	Nostoc sp. PCC7120
167539	Prochlorococcus marinus CCMP 1375
74547	Prochlorococcus marinus MIT 9313
59919	Prochlorococcus marinus subsp. pastoris
32046	Synechococcus elongatus
84588	Synechococcus sp. WH8102
1148	Synechocystis sp. PCC6803
	Firmicutes
272562	Clostridium acetobutylicum ATCC824
195102	Clostridium perfringens 13
212717	Clostridium tetani E88
233150	Mycoplasma gallisepticum R
243273	Mycoplasma genitalium G37
267748	Mycoplasma mobile 163K
272632	Mycoplasma mycoides PG1
272633	Mycoplasma penetrans HF-2
272634	Mycoplasma pneumoniae M129
272635	Mycoplasma pulmonis UAB CTIP
262768	Onion yellows phytoplasma OY-M
273068	Thermoanaerobacter tengcongensis MB4
273119	Ureaplasma parvum ATCC700970
	Bacilli
261594	Bacillus anthracis Ames Ancestor
222523	Bacillus cereus ATCC 10987
226900	Bacillus cereus ATCC 14579
272558	Bacillus halodurans C-125
224308	Bacillus subtilis 168
281309	Bacillus thuringiensis ser. konkukian 97-27
226185	Enterococcus faecalis V583
257314	Lactobacillus johnsonii NCC 533
220668	Lactobacillus plantarum WCFS1
272623	Lactococcus lactis Il1403
272626	Listeria innocua Clip11262
169963	Listeria monocytogenes 1/2a
265669	Listeria monocytogenes 4b
221109	Oceanobacillus iheyensis HTE831
282458	Staphylococcus aureus strain MRSA252
282459	Staphylococcus aureus strain MSSA476
158878	Staphylococcus aureus strain Mu50
196620	Staphylococcus aureus strain MW2
158879	Staphylococcus aureus strain N315
176280	Staphylococcus epidermidis ATCC12228
211110	Streptococcus agalactiae III
208435	Streptococcus agalactiae V
210007	Streptococcus mutans UA159
171101	Streptococcus pneumoniae ATCC BAA-255
170187	Streptococcus pneumoniae TIGR4
198466	Streptococcus pyogenes MGAS315
186103	Streptococcus pyogenes MGAS8232
160490	Streptococcus pyogenes SF370

TaxID	Nom
	Actinobacteridae
206672	Bifidobacterium longum NCC2705
257309	Corynebacterium diphtheriae NCTC13129
196164	Corynebacterium efficiens YS-314
196627	Corynebacterium glutamicum ATCC13032
281090	Leifsonia xyli CTCB07
262316	Mycobacterium avium K-10
272631	Mycobacterium leprae
83332	Mycobacterium tuberculosis
267747	Propionibacterium acnes KPA171202
227882	Streptomyces avermitilis MA-4680
100226	Streptomyces coelicolor A3-2
203267	Tropheryma whipplei Twist
218496	Tropheryma whipplei TW08-27
	Protéobactéries
959	Bdellovibrio bacteriovorus
518	Bordetella bronchiseptica
257311	Bordetella parapertussis 12822
520	Bordetella pertussis
192222	Campylobacter jejuni NCTC11168
243365	Chromobacterium violaceum ATCC 12472
882	Desulfovibrio vulgaris
243231	Geobacter sulfurreducens PCA
235279	Helicobacter hepaticus ATCC51449
85962	Helicobacter pylori ATCC 700392
85963	Helicobacter pylori J99
122587	Neisseria meningitidis A
122586	Neisseria meningitidis B
228410	Nitrosomonas europaea ATCC 19718
267608	Ralstonia solanacearum GMI1000
273121	Wolinella succinogenes DSM1740
	Gammaprotéobactéries
62977	Acinetobacter sp. ADP1
203907	Candidatus Blochmannia floridanus
227377	Coxiella burnetii
107806	Buchnera aphidicola Acyrthosiphon pisum
224915	Buchnera aphidicola Baizongia pistaciae
198804	Buchnera aphidicola Schizaphis graminum
218491	Erwinia carotovora

TaxID	Nom
83333	Escherichia coli K12
83334	Escherichia coli O157 :H7
217992	Escherichia coli O6
233412	Haemophilus ducreyi 35000HP
71421	Haemophilus influenzae RdKW20
272843	Pasteurella multocida str.Pm70
74109	Photobacterium profundum
243265	Photorhabdus luminescens
208964	Pseudomonas aeruginosa PAO1
160488	Pseudomonas putida KT2440
223283	Pseudomonas syringae DC3000
209261	Salmonella enterica Typhi Ty2
99287	Salmonella typhimurium LT2
211586	Shewanella oneidensis MR-1
198214	Shigella flexneri 2astr.301
243277	Vibrio cholerae N16961
223926	Vibrio parahaemolyticus RIMD2210633
216895	Vibrio vulnificus CMCP6
196600	Vibrio vulnificus YJ016
36870	Wigglesworthia glossinidia
183190	Xylella fastidiosa ATCC 700964
160492	Xylella fastidiosa 9a5c
190486	Xanthomonas axonopodis
190485	Xanthomonas campestris
214092	Yersinia pestis CO92
	Alphaprotéobactéries
176299	Agrobacterium tumefaciens str. C58
38323	Bartonella henselae
283165	Bartonella quintana str. Toulouse
224911	Bradyrhizobium japonicum USDA 110
224914	Brucella melitensis 16M
204722	Brucella suis 1330
190650	Caulobacter crescentus CB15
266835	Mesorhizobium loti MAFF303099
258594	Rhodopseudomonas palustris CGA009
272944	Rickettsia conorii
272947	Rickettsia prowazekii
266834	Sinorhizobium meliloti 1021
66077	Wolbachia sp.

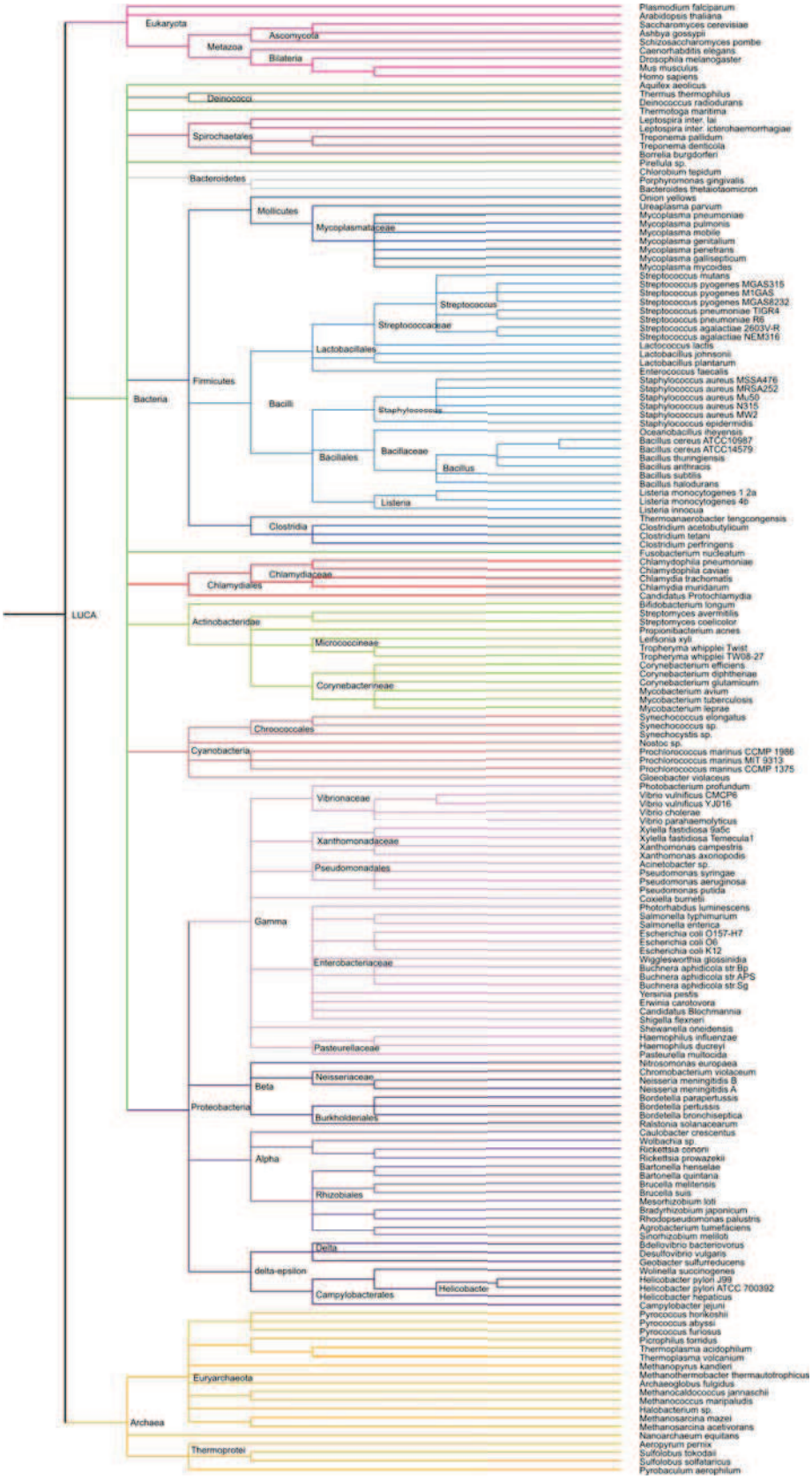


FIGURE C.1 – **Arbre phylogénétique des 170 espèces extrait du NCBI.** Cet arbre a été utilisé dans l'analyse de la méthode des réseaux Bayésiens du chapitre 2.

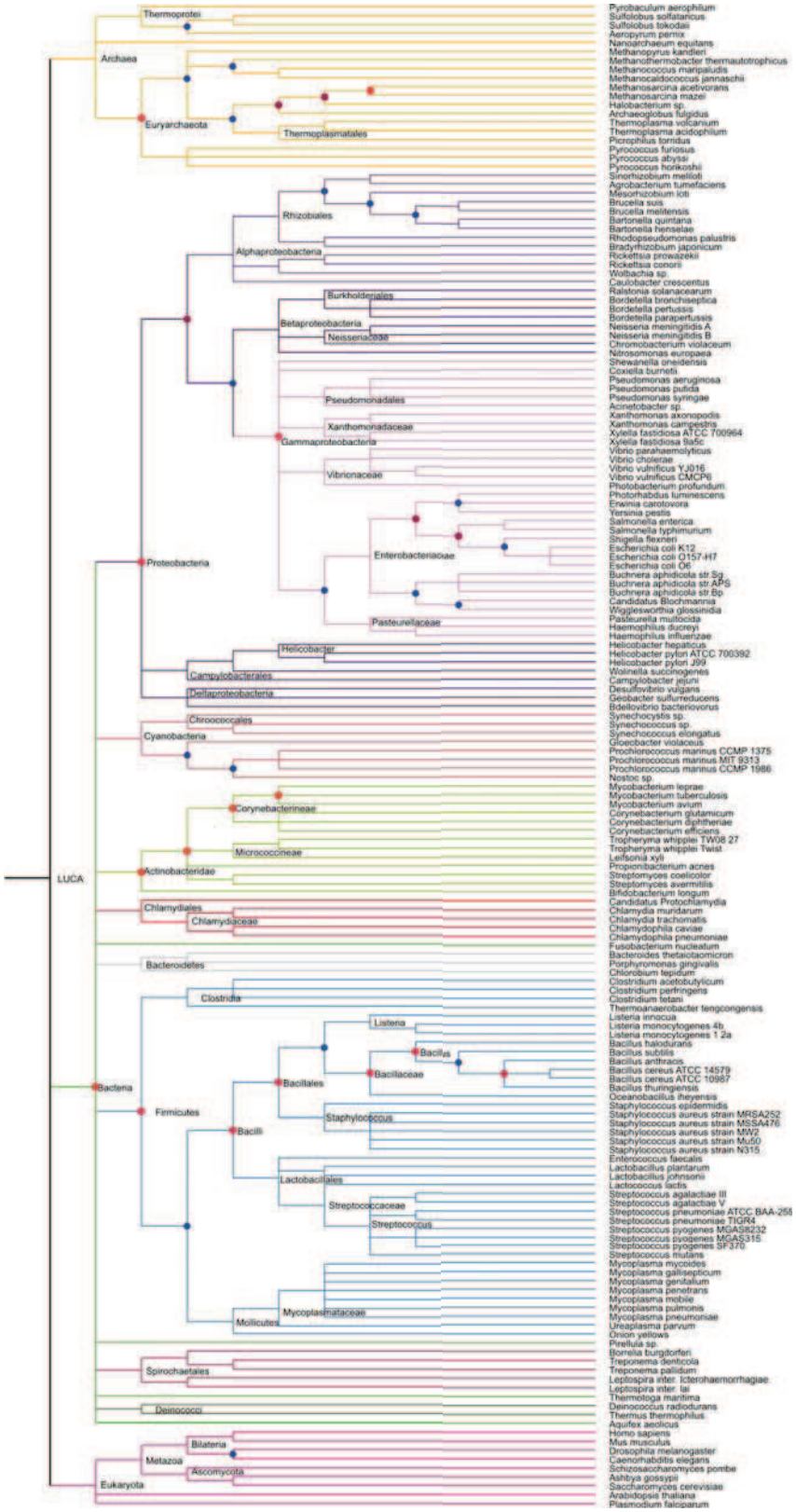


FIGURE C.2 – **Arbre phylogénétique modifié des 170 espèces.** Les nœuds modifiés sont cerclés en bleu. Cet arbre a été utilisé dans l'analyse de l'évolution des protéines modulaires (chapitre 3). Les nœuds cerclés de rouge correspondent à ceux utilisés dans l'analyse des taux d'évolution de nouveaux modules ProDom (section 3.6.1, page 80).

Modification de l'arbre des espèces

L'arbre de la taxonomie du NCBI présente de nombreuses polytomies pour lesquelles des résolutions partielles ou totales ont été proposées dans la littérature et sont aujourd'hui globalement admises. Les sous-arbres suivants ont été modifiés :

- Bilateria : choix de l'hypothèse Ecdysozoa [Holton et Pisani, 2010; Philippe *et al.*, 2005] (regroupant *Drosophila melanogaster* et *Caenorhabditis elegans*) plutôt que l'hypothèse Coelomata [Wolf *et al.*, 2004] (où *Drosophila melanogaster* est le groupe frère des Euarchontoglires).
- Archées : les regroupements effectués dans les sous-arbres des *Thermoprotéi* et des *Euryarchaeota* se basent sur la phylogénie de Brochier *et al.* [2005].
- Les modifications apportées dans les sous-arbres bactériens des Protéobactéries et des Firmicutes viennent de Ciccarelli *et al.* [2006], les α-Protéobactéries ont été modifiées à l'aide de la phylogénie de Boussau *et al.* [2004], quant aux Actinobactéries et aux Cyanobactéries, les modifications proviennent de communications personnelles non publiées de Sophie Abby.

Bibliographie

ALEXANDROV N. N. et GO N. 1994. Biological meaning, statistical significance, and classification of local spatial similarities in nonhomologous proteins. *Protein Sci*, **3**(6):866–875.

ALTSCHUL S. F., MADDEN T. L., SCHÄFFER A. A., ZHANG J., ZHANG Z., MILLER W., et LIPMAN D. J. 1997. Gapped blast and psi-blast : a new generation of protein database search programs. *Nucleic Acids Res*, **25**(17):3389–3402.

ANDREEVA A., HOWORTH D., CHANDONIA J.-M., BRENNER S. E., HUBBARD T. J. P., CHOTHIA C., et MURZIN A. G. 2008. Data growth and its impact on the scop database : new developments. *Nucleic Acids Res*, **36**(Database issue):D419–D425.

APIC G., GOUGH J., et TEICHMANN S. A. 2001a. Domain combinations in archaeal, eubacterial and eukaryotic proteomes. *J Mol Biol*, **310**(2):311–325.

APIC G., GOUGH J., et TEICHMANN S. A. 2001b. An insight into domain combinations. *Bioinformatics*, **17 Suppl 1**:S83–S89.

ARGUELLO J. R., CHEN Y., YANG S., WANG W., et LONG M. 2006. Origination of an x-linked testes chimeric gene by illegitimate recombination in drosophila. *PLoS Genet*, **2**(5):e77.

AROUL-SELVAM R., HUBBARD T., et SASIDHARAN R. 2004. Domain insertions in protein structures. *J Mol Biol*, **338**(4):633–641.

ATTWOOD T. K., BRADLEY P., FLOWER D. R., GAULTON A., MAUDLING N., MITCHELL A. L., MOULTON G., NORDLE A., PAINE K., TAYLOR P., UDDIN A., et ZYGOURI C. 2003. Prints and its automatic supplement, preprints. *Nucleic Acids Res*, **31**(1):400–402.

BABUSHOK D. V., OSTERTAG E. M., et KAZAZIAN H. H. 2007. Current topics in genome evolution : molecular mechanisms of new gene formation. *Cell Mol Life Sci*, **64**(5):542–554.

BARTLETT G. J., BORKAKOTI N., et THORNTON J. M. 2003. Catalysing new reactions during evolution : economy of residues and mechanism. *J Mol Biol*, **331**(4):829–860.

BASHTON M. et CHOTHIA C. 2002. The geometry of domain combination in proteins. *J Mol Biol*, **315**(4):927–939.

BJÖRKLUND A. K., EKMAN D., LIGHT S., FREY-SKÖTT J., et ELOFSSON A. 2005. Domain

rearrangements in protein evolution. *J Mol Biol*, **353**(4):911–923.

BLANC G., OGATA H., ROBERT C., AUDIC S., SUHRE K., VESTRIS G., CLAVERIE J.-M., et RAOULT D. 2007. Reductive genome evolution from the mother of rickettsia. *PLoS Genet*, **3** (1):e14.

BORK P. 1991. Shuffled domains in extracellular proteins. *FEBS Lett*, **286**(1-2):47–54.

BORNBERG-BAUER E., BEAUSSART F., KUMMERFELD S. K., TEICHMANN S. A., et WEINER J. 2005. The evolution of domain arrangements in proteins and interaction networks. *Cell Mol Life Sci*, **62**(4):435–445.

BOUSSAU B., KARLBERG E. O., FRANK A. C., LEGAULT B.-A., et ANDERSSON S. G. E. 2004. Computational inference of scenarios for alpha-proteobacterial genome evolution. *Proc Natl Acad Sci U S A*, **101**(26):9722–9727.

BRENNER S. E., CHOTHIA C., et HUBBARD T. J. 1997. Population statistics of protein structures : lessons from structural classifications. *Curr Opin Struct Biol*, **7**(3):369–376.

BROCHIER C., FORTERRE P., et GRIBALDO S. 2005. An emerging phylogenetic core of archaea : phylogenies of transcription and translation machineries converge following addition of new genome sequences. *BMC Evol Biol*, **5**(1):36.

BRU C., COURCELLE E., CARRÈRE S., BEAUSSE Y., DALMAR S., et KAHN D. 2005. The prodom database of protein domain families : more emphasis on 3d. *Nucleic Acids Res*, **33**(Database issue):D212–D215.

BUJNICKI J. M. 2002. Sequence permutations in the molecular evolution of dna methyltransferases. *BMC Evol Biol*, **2**:3.

BUNTINE W. 1991. Theory refinement on bayesian networks. *Dans Proceedings of the seventh conference (1991) on Uncertainty in artificial intelligence*, pages 52–60, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

CATTANEO R. 1990. Messenger rna editing and the genetic code. *Experientia*, **46**(11-12):1142–1148.

CHOI I.-G. et KIM S.-H. 2006. Evolution of protein structural classes and protein sequence families. *Proc Natl Acad Sci U S A*, **103**(38):14056–14061.

CHOTHIA C. 1992. Proteins. one thousand families for the molecular biologist. *Nature*, **357**(6379): 543–544.

CHOTHIA C. et LESK A. M. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J*, **5**(4):823–826.

CHOTHIA C. et GOUGH J. 2009. Genomic and structural aspects of protein evolution. *Biochem J*, **419**(1):15–28.

CHOTHIA C., GOUGH J., VOGEL C., et TEICHMANN S. A. 2003. Evolution of the protein repertoire. *Science*, **300**(5626):1701–1703.

CICCARELLI F. D., DOERKS T., von MERING C., CREEVEY C. J., SNEL B., et BORK P. 2006.

Toward automatic reconstruction of a highly resolved tree of life. *Science*, **311**(5765):1283–1287.

CLAMP M., FRY B., KAMAL M., XIE X., CUFF J., LIN M. F., KELLIS M., LINDBLAD-TOH K., et LANDER E. S. 2007. Distinguishing protein-coding and noncoding genes in the human genome. *Proc Natl Acad Sci U S A*, **104**(49):19428–19433.

CLAVERIE J.-M. et STATES D. J. 1993. Information enhancement methods for large scale sequence analysis. *Computers & Chemistry*, **17**(2):191 – 201.

COHEN O. et PUPKO T. 2010. Inference and characterization of horizontally transferred gene families using stochastic mapping. *Mol Biol Evol*, **27**(3):703–713.

COHEN O., RUBINSTEIN N. D., STERN A., GOPHNA U., et PUPKO T. 2008. A likelihood framework to analyse phyletic patterns. *Philos Trans R Soc Lond B Biol Sci*, **363**(1512):3903–3911.

CONSORTIUM U. 2009. The universal protein resource (uniprot) 2009. *Nucleic Acids Res*, **37** (Database issue):D169–D174.

CORDES M. H., BURTON R. E., WALSH N. P., MCKNIGHT C. J., et SAUER R. T. 2000. An evolutionary bridge to a new protein fold. *Nat Struct Biol*, **7**(12):1129–1132.

COULSON A. F. W. et MOULT J. 2002. A unifold, mesofold, and superfold model of protein fold use. *Proteins*, **46**(1):61–71.

CSURÖS M. et MIKLÓS I. 2009. Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. *Mol Biol Evol*, **26**(9):2087–2095.

CUFF A., REDFERN O. C., GREENE L., SILLITOE I., LEWIS T., DIBLEY M., REID A., PEARL F., DALLMAN T., TODD A., GARRATT R., THORNTON J., et ORENGO C. 2009a. The cath hierarchy revisited-structural divergence in domain superfamilies and the continuity of fold space. *Structure*, **17**(8):1051–1062.

CUFF A. L., SILLITOE I., LEWIS T., REDFERN O. C., GARRATT R., THORNTON J., et ORENGO C. A. 2009b. The cath classification revisited–architectures reviewed and new ways to characterize structural divergence in superfamilies. *Nucleic Acids Res*, **37**(Database issue):D310–D314.

DAGAN T. et MARTIN W. 2007. Ancestral genome sizes specify the minimum rate of lateral gene transfer during prokaryote evolution. *Proc Natl Acad Sci U S A*, **104**(3):870–875.

DEMPSTER A. P., LAIRD N. M., et RUBIN D. B. 1977. Maximum likelihood from incomplete data via the em algorithm. *J Roy Stat Soc*, **39**:1–38.

DESSAILLY B. H., NAIR R., JAROSZEWSKI L., FAJARDO J. E., KOURANOV A., LEE D., FISER A., GODZIK A., ROST B., et ORENGO C. 2009. Psi-2 : structural genomics to cover protein domain family space. *Structure*, **17**(6):869–881.

DIDELOT X., DARLING A., et FALUSH D. 2009. Inferring genomic flux in bacteria. *Genome Res*, **19**(2):306–317.

DOOLITTLE R. F. 1986. *Of URFS and ORFS – a Primer on How to Analyze Derived Amino Acid*

Sequences. University Science Books, Mill Valley California.

DUFAYARD J.-F., DURET L., PENEL S., GOUY M., RECHENMANN F., et PERRIÈRE G. 2005. Tree pattern matching in phylogenetic trees : automatic search for orthologs or paralogs in homologous gene sequence databases. *Bioinformatics*, **21**(11):2596–2603.

EDGAR R. C. 2004. Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic Acids Res*, **32**(5):1792–1797.

EKMAN D., BJÖRKLUND A. K., et ELOFSSON A. 2007. Quantification of the elevated rate of domain rearrangements in metazoa. *J Mol Biol*, **372**(5):1337–1348.

EMBRACE . 2008. Service registry. http://www.embraceregistry.net/.

ENRIGHT A. J., DONGEN S. V., et OUZOUNIS C. A. 2002. An efficient algorithm for large-scale detection of protein families. *Nucleic Acids Res*, **30**(7):1575–1584.

ENRIGHT A. J. et OUZOUNIS C. A. 2001. Functional associations of proteins in entire genomes by means of exhaustive detection of gene fusions. *Genome Biol*, **2**(9):RESEARCH0034.

ESNOUF R. M. 1997. An extensively modified version of molscript that includes greatly enhanced coloring capabilities. *J Mol Graph Model*, **15**(2):132–4, 112–3.

ESSER L., WANG C. R., HOSAKA M., SMAGULA C. S., SÜDHOF T. C., et DEISENHOFER J. 1998. Synapsin i is structurally similar to atp-utilizing enzymes. *EMBO J*, **17**(4):977–984.

FAN C., MOEWS P. C., WALSH C. T., et KNOX J. R. 1994. Vancomycin resistance : structure of d-alanine :d-alanine ligase at 2.3 a resolution. *Science*, **266**(5184):439–443.

FELSENSTEIN J. 1989. Phylip - phylogeny inference package (version 3.2). *Cladistics*, **5**:164–166.

FINN R. D., MISTRY J., TATE J., COGGILL P., HEGER A., POLLINGTON J. E., GAVIN O. L., GUNASEKARAN P., CERIC G., FORSLUND K., HOLM L., SONNHAMMER E. L. L., EDDY S. R., et BATEMAN A. 2010. The pfam protein families database. *Nucleic Acids Res*, **38** (Database issue):D211–D222.

FLEISCHMANN R. D., ADAMS M. D., WHITE O., CLAYTON R. A., KIRKNESS E. F., KERLAVAGE A. R., BULT C. J., TOMB J. F., DOUGHERTY B. A., et MERRICK J. M. 1995. Whole-genome random sequencing and assembly of haemophilus influenzae rd. *Science*, **269**(5223):496–512.

FONG J. H., GEER L. Y., PANCHENKO A. R., et BRYANT S. H. 2007. Modeling the evolution of protein domain architectures using maximum parsimony. *J Mol Biol*, **366**(1):307–315.

FORSLUND K., HENRICSON A., HOLLICH V., et SONNHAMMER E. L. L. 2008. Domain tree-based analysis of protein architecture evolution. *Mol Biol Evol*, **25**(2):254–264.

FRIEDMAN N., LINIAL M., NACHMAN I., et PE'ER D. 2000. Using bayesian networks to analyze expression data. *J Comput Biol*, **7**(3-4):601–620.

FRIEDMAN N. 2004. Inferring cellular networks using probabilistic graphical models. *Science*, **303**(5659):799–805.

GEMAN S. et GEMAN D. 1984. Stochastic relaxation, gibbs distributions and the bayesian resto-

ration of images. *Journal of Applied Statistics*, **20**(5):721–741.

GERSTEIN M. 1998. How representative are the known structures of the proteins in a complete genome ? a comprehensive structural census. *Fold Des*, **3**(6):497–512.

GEVAERT O., SMET F. D., TIMMERMAN D., MOREAU Y., et MOOR B. D. 2006. Predicting the prognosis of breast cancer by integrating clinical and microarray data with bayesian networks. *Bioinformatics*, **22**(14):e184–e190.

GILBERT W. 1978. Why genes in pieces ? *Nature*, **271**(5645):501.

GOGARTEN J. P. et TOWNSEND J. P. 2005. Horizontal gene transfer, genome innovation and evolution. *Nat Rev Microbiol*, **3**(9):679–687.

GOLDSTEIN R. A. 2008. The structure of protein evolution and the evolution of protein structure. *Curr Opin Struct Biol*, **18**(2):170–177.

GOODSTADT L. et PONTING C. P. 2006. Phylogenetic reconstruction of orthology, paralogy, and conserved synteny for dog and human. *PLoS Comput Biol*, **2**(9):e133.

GOUGH J. 2005. Convergent evolution of domain architectures (is rare). *Bioinformatics*, **21**(8): 1464–1471.

GOUZY J., CORPET F., et KAHN D. 1999. Whole genome protein domain analysis using a new method for domain clustering. *Comput Chem*, **23**(3-4):333–340, y.

GOVINDARAJAN S., RECABARREN R., et GOLDSTEIN R. A. 1999. Estimating the total number of protein folds. *Proteins*, **35**(4):408–414.

GRABOWSKI M., JOACHIMIAK A., OTWINOWSKI Z., et MINOR W. 2007. Structural genomics : keeping up with expanding knowledge of the protein universe. *Curr Opin Struct Biol*, **17** (3):347–353.

GRISHIN N. V. 2001. Fold change in evolution of protein structures. *J Struct Biol*, **134**(2-3):167–185.

HAFT D. H., SELENGUT J. D., et WHITE O. 2003. The tigrfams database of protein families. *Nucleic Acids Res*, **31**(1):371–373.

HAHN M. W., BIE T. D., STAJICH J. E., NGUYEN C., et CRISTIANINI N. 2005. Estimating the tempo and mode of gene family evolution from comparative genomic data. *Genome Res*, **15** (8):1153–1160.

HAO W. et GOLDING G. B. 2006. The fate of laterally transferred genes : life in the fast lane to adaptation or death. *Genome Res*, **16**(5):636–643.

HAO W. et GOLDING G. B. 2008. Uncovering rate variation of lateral gene transfer during bacterial genome evolution. *BMC Genomics*, **9**:235.

HECKERMAN D. 1999. A tutorial on learning with bayesian networks. pages 301–354.

HECKERMAN D., GEIGER D., et CHICKERING D. M. 1995. Learning bayesian networks : The combination of knowledge and statistical data. *Dans Machine Learning*, volume 20, pages 197–243.

HEGDE S. S., VETTING M. W., RODERICK S. L., MITCHENALL L. A., MAXWELL A., TAKIFF H. E., et BLANCHARD J. S. 2005. A fluoroquinolone resistance protein from mycobacterium tuberculosis that mimics dna. *Science*, **308**(5727):1480–1483.

HEGER A. et HOLM L. 2003. Exhaustive enumeration of protein domain families. *J Mol Biol*, **328** (3):749–767.

HEGER A., WILTON C. A., SIVAKUMAR A., et HOLM L. 2005. Adda : a domain database with global coverage of the protein universe. *Nucleic Acids Res*, **33**(Database issue):D188–D191.

HILLIS D. M. 1994. Homology in molecular biology. *Homology : The Hierarchical Basis of Comparative Biology.* Academic Press, New York, page 339–368.

HOLM L. et ROSENSTRÖM P. 2010. Dali server : conservation mapping in 3d. *Nucleic Acids Res*, **38 Suppl**:W545–W549.

HOLTON T. A. et PISANI D. 2010. Deep genomic-scale analyses of the metazoa reject coelomata : evidence from single- and multigene families analyzed under a supertree and supermatrix paradigm. *Genome Biol Evol*, **2**:310–324.

HUGHES A. L. 1994. The evolution of functionally novel proteins after gene duplication. *Proc Biol Sci*, **256**(1346):119–124.

ITOH M., NACHER J. C., ichi KUMA K., GOTO S., et KANEHISA M. 2007. Evolutionary history and functional implications of protein domains and their combinations in eukaryotes. *Genome Biol*, **8**(6):R121.

IWASAKI W. et TAKAGI T. 2007. Reconstruction of highly heterogeneous gene-content evolution across the three domains of life. *Bioinformatics*, **23**(13):i230–i239.

JACOB F. 1977. Evolution and tinkering. *Science*, **196**(4295):1161–1166.

JAIN R., RIVERA M. C., et LAKE J. A. 1999. Horizontal gene transfer among genomes : the complexity hypothesis. *Proc Natl Acad Sci U S A*, **96**(7):3801–3806.

JANSEN R., YU H., GREENBAUM D., KLUGER Y., KROGAN N. J., CHUNG S., EMILI A., SNYDER M., GREENBLATT J. F., et GERSTEIN M. 2003. A bayesian networks approach for predicting protein-protein interactions from genomic data. *Science*, **302**(5644):449–453.

JAROSZEWSKI L., LI Z., KRISHNA S. S., BAKOLITSA C., WOOLEY J., DEACON A. M., WILSON I. A., et GODZIK A. 2009. Exploration of uncharted regions of the protein universe. *PLoS Biol*, **7**(9):e1000205.

JELTSCH A. 1999. Circular permutations in the molecular evolution of dna methyltransferases. *J Mol Evol*, **49**(1):161–164.

JENSEN F. V., LAURITZEN S. L., et OLESEN K. G. 1990. Bayesian updating in causal probabilistic networks by local computations. *Computational Statistics Quaterly*, **4**:269–282.

JENSEN F. V. et NIELSON T. D. 2007. *Bayesian network and decision graphs*. Springer, 2nd édition.

JIANG H. et BLOUIN C. 2007. Insertions and the emergence of novel protein structure : a structure-

based phylogenetic study of insertions. *BMC Bioinformatics*, **8**:444.

JONES D. T., TAYLOR W. R., et THORNTON J. M. 1992. The rapid generation of mutation data matrices from protein sequences. *Comput Appl Biosci*, **8**(3):275–282.

KAESSMANN H. 2009. Genetics. more than just a copy. *Science*, **325**(5943):958–959.

KAESSMANN H. 2010. Origins, evolution, and phenotypic impact of new genes. *Genome Res*, **20** (10):1313–1326.

KAESSMANN H., ZÖLLNER S., NEKRUTENKO A., et LI W.-H. 2002. Signatures of domain shuffling in the human genome. *Genome Res*, **12**(11):1642–1650.

KETTLER G. C., MARTINY A. C., HUANG K., ZUCKER J., COLEMAN M. L., RODRIGUE S., CHEN F., LAPIDUS A., FERRIERA S., JOHNSON J., STEGLICH C., CHURCH G. M., RICHARDSON P., et CHISHOLM S. W. 2007. Patterns and implications of gene gain and loss in the evolution of prochlorococcus. *PLoS Genet*, **3**(12):e231.

KIM H. J., CHOI M. Y., KIM H. J., et LLINÁS M. 2010. Conformational dynamics and ligand binding in the multi-domain protein pdc109. *PLoS One*, **5**(2):e9180.

KIM K. M. et CAETANO-ANOLLÉS G. 2010. Emergence and evolution of modern molecular functions inferred from phylogenomic analysis of ontological data. *Mol Biol Evol*, **27**(7):1710–1733.

KINCH L. N. et GRISHIN N. V. 2002. Evolution of protein structures and functions. *Curr Opin Struct Biol*, **12**(3):400–408.

KIRCHMAIR J., MARKT P., DISTINTO S., SCHUSTER D., SPITZER G. M., LIEDL K. R., LANGER T., et WOLBER G. 2008. The protein data bank (pdb), its related services and software tools as key components for in silico guided drug discovery. *J Med Chem*, **51**(22):7021–7040.

KNOWLES D. G. et MCLYSAGHT A. 2009. Recent de novo origin of human protein-coding genes. *Genome research*, **19**(10):1752–1759.

KOONIN E. V. 2003. Comparative genomics, minimal gene-sets and the last universal common ancestor. *Nat Rev Microbiol*, **1**(2):127–136.

KOONIN E. V., FEDOROVA N. D., JACKSON J. D., JACOBS A. R., KRYLOV D. M., MAKAROVA K. S., MAZUMDER R., MEKHEDOV S. L., NIKOLSKAYA A. N., RAO B. S., ROGOZIN I. B., SMIRNOV S., SOROKIN A. V., SVERDLOV A. V., VASUDEVAN S., WOLF Y. I., YIN J. J., et NATALE D. A. 2004. A comprehensive evolutionary classification of proteins encoded in complete eukaryotic genomes. *Genome Biol*, **5**(2):R7.

KRAULIS P. 1991. Molscript - a program to produce both detailed and schematic plots of protein structures. *JOURNAL OF APPLIED CRYSTALLOGRAPHY*, **24**(Part 5):946–950.

KUMMERFELD S. K. et TEICHMANN S. A. 2005. Relative rates of gene fusion and fission in multi-domain proteins. *Trends Genet*, **21**(1):25–30.

KUMMERFELD S. K. et TEICHMANN S. A. 2009. Protein domain organisation : adding order. *BMC Bioinformatics*, **10**:39.

KUNIN V., CASES I., ENRIGHT A. J., de LORENZO V., et OUZOUNIS C. A. 2003. Myriads of protein families, and still counting. *Genome Biol*, **4**(2):401.

KUNIN V. et OUZOUNIS C. A. 2003a. The balance of driving forces during genome evolution in prokaryotes. *Genome Res*, **13**(7):1589–1594.

KUNIN V. et OUZOUNIS C. A. 2003b. Genetrace-reconstruction of gene content of ancestral species. *Bioinformatics*, **19**(11):1412–1416.

KUNIN V., TEICHMANN S. A., HUYNEN M. A., et OUZOUNIS C. A. 2005. The properties of protein family space depend on experimental design. *Bioinformatics*, **21**(11):2618–2622.

LAGOMARSINO M. C., SELLERIO A. L., HEIJNING P. D., et BASSETTI B. 2009. Universal features in the genome-level evolution of protein domains. *Genome Biol*, **10**(1):R12.

LANDER E. S., LINTON L. M., BIRREN B., NUSBAUM C., ZODY M. C., et INTERNATIONAL HUMAN GENOME SEQUENCING CONSORTIUM . 2001. Initial sequencing and analysis of the human genome. *Nature*, **409**(6822):860–921.

LAURITZEN S. L. et SPIEGELHALTER D. J. 1988. Local computations with probabilities on graphical structures and their application to expert systems. *Journal of the Royal Statistical Society*, **50**(2):157–224.

LEE D., GRANT A., BUCHAN D., et ORENGO C. 2003. A structural perspective on genome evolution. *Curr Opin Struct Biol*, **13**(3):359–369.

LEES J., YEATS C., REDFERN O., CLEGG A., et ORENGO C. 2010. Gene3d : merging structure and function for a thousand genomes. *Nucleic Acids Res*, **38**(Database issue):D296–D300.

LETUNIC I., DOERKS T., et BORK P. 2009. Smart 6 : recent updates and new developments. *Nucleic Acids Res*, **37**(Database issue):D229–D232.

LEVINE M. T., JONES C. D., KERN A. D., LINDFORS H. A., et BEGUN D. J. 2006. Novel genes derived from noncoding dna in drosophila melanogaster are frequently x-linked and exhibit testis-biased expression. *Proc Natl Acad Sci U S A*, **103**(26):9935–9939.

LEVITT M. 2007. Growth of novel protein structural data. *Proc Natl Acad Sci U S A*, **104**(9):3183–3188.

LIU J. et ROST B. 2004. Chop proteins into structural domain-like fragments. *Proteins*, **55**(3):678–688.

LIU M. et GRIGORIEV A. 2004. Protein domains correlate strongly with exons in multiple eukaryotic genomes–evidence of exon shuffling ? *Trends Genet*, **20**(9):399–403.

LUKJANCENKO O., WASSENAAR T. M., et USSERY D. W. 2010. Comparison of 61 sequenced escherichia coli genomes. *Microb Ecol*, **60**:708–720.

LYNCH M. et FORCE A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics*, **154**(1):459–473.

LYNCH M. et CONERY J. S. 2003. The origins of genome complexity. *Science*, **302**(5649):1401–1404.

MADEJ T., GIBRAT J. F., et BRYANT S. H. 1995. Threading a database of protein cores. *Proteins*, **23**(3):356–369.

MAKAROVA K. S., SOROKIN A. V., NOVICHKOV P. S., WOLF Y. I., et KOONIN E. V. 2007. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. *Biol Direct*, **2**:33.

MARCOTTE E. M., PELLEGRINI M., NG H. L., RICE D. W., YEATES T. O., et EISENBERG D. 1999. Detecting protein function and protein-protein interactions from genome sequences. *Science*, **285**(5428):751–753.

MARRI P. R., HAO W., et GOLDING G. B. 2007. The role of laterally transferred genes in adaptive evolution. *BMC Evol Biol*, **7 Suppl 1**:S8.

MARSDEN R. L., LEE D., MAIBAUM M., YEATS C., et ORENGO C. A. 2006. Comprehensive genome analysis of 203 genomes provides structural genomics with new insights into protein family space. *Nucleic Acids Res*, **34**(3):1066–1080.

MARSDEN R. L. et ORENGO C. A. 2008. The classification of protein domains. *Methods Mol Biol*, **453**:123–146.

MARUYAMA K. 1994. Connectin, an elastic protein of striated muscle. *Biophys Chem*, **50**(1-2):73–85.

MAYR E. 1960. The emergence of evolutionary novelties. *The evolution of life*. Chicago : University of Chicago, pages 349–380.

MEDINI D., DONATI C., TETTELIN H., MASIGNANI V., et RAPPUOLI R. 2005. The microbial pan-genome. *Curr Opin Genet Dev*, **15**(6):589–594.

MEIER S., JENSEN P. R., DAVID C. N., CHAPMAN J., HOLSTEIN T. W., GRZESIEK S., et OZBEK S. 2007. Continuous molecular evolution of protein-domain structures by single amino acid changes. *Curr Biol*, **17**(2):173–178.

MEINEL T., KRAUSE A., LUZ H., VINGRON M., et STAUB E. 2005. The systers protein family database in 2005. *Nucleic Acids Res*, **33**(Database issue):D226–D229.

MIRKIN B. G., FENNER T. I., GALPERIN M. Y., et KOONIN E. V. 2003. Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC Evol Biol*, **3**:2.

MOCZEK A. P. 2008. On the origins of novelty in development and evolution. *Bioessays*, **30**(5):432–447.

MOLINA N. et van NIMWEGEN E. 2008. The evolution of domain-content in bacterial genomes. *Biol Direct*, **3**:51.

MOORE A. D., BJÖRKLUND A. K., EKMAN D., BORNBERG-BAUER E., et ELOFSSON A. 2008. Arrangements in the modular evolution of proteins. *Trends Biochem Sci*, **33**(9):444–451.

MORAN N. A., MCCUTCHEON J. P., et NAKABACHI A. 2008. Genomics and evolution of heritable bacterial symbionts. *Annu Rev Genet*, **42**:165–190.

MURPHY K. P. 2001. The bayes net toolbox for matlab. *Comput. Sci. Stat.*, **33**:2001.

MURPHY K. P. 2005. Software packages for graphical models/bayesian networks. http://www.cs.ubc.ca/~murphyk/Software/BNT/bnsoft.html.

NEYMAN J. et PEARSON E. 1928. On the use and interpretation of certain test criteria for purposes of statistical inference, part i. *Biometrika*, **20A**:175–240.

OCHMAN H. et MORAN N. A. 2001. Genes lost and genes found : evolution of bacterial pathogenesis and symbiosis. *Science*, **292**(5519):1096–1099.

OGURA A., IKEO K., et GOJOBORI T. 2005. Estimation of ancestral gene set of bilaterian animals and its implication to dynamic change of gene content in bilaterian evolution. *Gene*, **345** (1):65–71.

OHNO S. 1970. *Evolution by gene duplication*. Springer Verlag.

ORENGO C. A. 1999. Cora–topological fingerprints for protein structural families. *Protein Sci*, **8** (4):699–715.

ORENGO C. A., JONES D. T., et THORNTON J. M. 1994. Protein superfamilies and domain superfolds. *Nature*, **372**(6507):631–634.

ORENGO C. A. et TAYLOR W. R. 1996. Ssap : sequential structure alignment program for protein structure comparison. *Methods Enzymol*, **266**:617–635.

ORGEL L. E. et CRICK F. H. 1980. Selfish dna : the ultimate parasite. *Nature*, **284**(5757):604–607.

OUZOUNIS C. A., KUNIN V., DARZENTAS N., et GOLDOVSKY L. 2006. A minimal estimate for the gene content of the last universal common ancestor–exobiology from a terrestrial perspective. *Res Microbiol*, **157**(1):57–68.

PAGEL M. 1994. Detecting correlated evolution on phylogenies : a general method for the comparative analysis of discrete characters. *Proceedings of the Royal Society London Series B*, **255**:37–45.

PAGEL M., MEADE A., et BARKER D. 2004. Bayesian estimation of ancestral character states on phylogenies. *Systematic Biology*, **53**(5):673–684(12).

PASEK S., BERGERON A., RISLER J.-L., LOUIS A., OLLIVIER E., et RAFFINOT M. 2005. Identification of genomic features using microsyntenies of domains : domain teams. *Genome Res*, **15**(6):867–874.

PASEK S., RISLER J.-L., et BRÉZELLEC P. 2006. Gene fusion/fission is a major contributor to evolution of multi-domain bacterial proteins. *Bioinformatics*, **22**(12):1418–1423.

PATTHY L. 1999. Genome evolution and the evolution of exon-shuffling–a review. *Gene*, **238** (1):103–114.

PATTHY L. 2001. Exons and protein modules. *Encyclopedia of Life Sciences*. John Wiley & Sons, Ltd.

PENEL S., ARIGON A.-M., DUFAYARD J.-F., SERTIER A.-S., DAUBIN V., DURET L., GOUY M., et PERRIÈRE G. 2009. Databases of homologous gene families for comparative genomics.

BMC Bioinformatics, **10 Suppl 6**:S3.

PHILIPPE H., LARTILLOT N., et BRINKMANN H. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol Biol Evol*, **22**(5):1246–1253.

PIROVANO W. et HERINGA J. 2010. Protein secondary structure prediction. *Methods Mol Biol*, **609**:327–348.

POLLASTRI G. et MCLYSAGHT A. 2005. Porter : a new, accurate server for protein secondary structure prediction. *Bioinformatics*, **21**(8):1719–1720.

PORTUGALY E., LINIAL N., et LINIAL M. 2007. Everest : a collection of evolutionary conserved protein domains. *Nucleic Acids Res*, **35**(Database issue):D241–D246.

PUTNAM N. H., SRIVASTAVA M., HELLSTEN U., DIRKS B., CHAPMAN J., SALAMOV A., TERRY A., SHAPIRO H., LINDQUIST E., KAPITONOV V. V., JURKA J., GENIKHOVICH G., GRIGORIEV I. V., LUCAS S. M., STEELE R. E., FINNERTY J. R., TECHNAU U., MARTINDALE M. Q., et ROKHSAR D. S. 2007. Sea anemone genome reveals ancestral eumetazoan gene repertoire and genomic organization. *Science*, **317**(5834):86–94.

QIAN J., LUSCOMBE N. M., et GERSTEIN M. 2001. Protein family and fold occurrence in genomes : power-law behaviour and evolutionary model. *J Mol Biol*, **313**(4):673–681.

REES D. C., LEWIS M., et LIPSCOMB W. N. 1983. Refined crystal structure of carboxypeptidase a at 1.54 a resolution. *J Mol Biol*, **168**(2):367–387.

REEVES G. A., DALLMAN T. J., REDFERN O. C., AKPOR A., et ORENGO C. A. 2006. Structural diversity of domain superfamilies in the cath database. *J Mol Biol*, **360**(3):725–741, y.

ROST B. 1998. Marrying structure and genomics. *Structure*, **6**(3):259–263.

ROTH C., RASTOGI S., ARVESTAD L., DITTMAR K., LIGHT S., EKMAN D., et LIBERLES D. A. 2007. Evolution after gene duplication : models, mechanisms, sequences, systems, and organisms. *J Exp Zool B Mol Dev Evol*, **308**(1):58–73.

RUAN J., LI H., CHEN Z., COGHLAN A., COIN L. J. M., GUO Y., HÉRICHÉ J.-K., HU Y., KRISTIANSEN K., LI R., LIU T., MOSES A., QIN J., VANG S., VILELLA A. J., URETA-VIDAL A., BOLUND L., WANG J., et DURBIN R. 2008. Treefam : 2008 update. *Nucleic Acids Res*, **36**(Database issue):D735–D740.

RUBIN D. B. 1976. Inference and missing data. *Biometrika*, **63**(3):581–592.

RUSSELL R. B., SAQI M. A., SAYLE R. A., BATES P. A., et STERNBERG M. J. 1997. Recognition of analogous and homologous protein folds : analysis of sequence and structure conservation. *J Mol Biol*, **269**(3):423–439.

SADREYEV R. I., KIM B.-H., et GRISHIN N. V. 2009. Discrete-continuous duality of protein structure space. *Curr Opin Struct Biol*, **19**(3):321–328.

SAKARYA O., KOSIK K. S., et OAKLEY T. H. 2008. Reconstructing ancestral genome content based on symmetrical best alignments and dollo parsimony. *Bioinformatics*, **24**(5):606–612.

SALI A. et BLUNDELL T. L. 1990. Definition of general topological equivalence in protein structures. a procedure involving comparison of properties and relationships through simulated annealing and dynamic programming. *J Mol Biol*, **212**(2):403–428.

SCHAEFFER R. D. et DAGGETT V. 2011. Protein folds and protein folding. *Protein Eng Des Sel*, **24**:11–19.

SIGRIST C. J. A., CERUTTI L., HULO N., GATTIKER A., FALQUET L., PAGNI M., BAIROCH A., et BUCHER P. 2002. Prosite : a documented database using patterns and profiles as motif descriptors. *Brief Bioinform*, **3**(3):265–274.

SNEL B., BORK P., et HUYNEN M. 2000. Genome evolution. gene fusion versus gene fission. *Trends Genet*, **16**(1):9–11.

SNEL B., BORK P., et HUYNEN M. A. 2002. Genomes in flux : the evolution of archaeal and proteobacterial gene content. *Genome Res*, **12**(1):17–25.

SPENCER M., SUSKO E., et ROGER A. J. 2006. Modelling prokaryote gene content. *Evol Bioinform Online*, **2**:157–178.

SPIEGELHALTER D. J. et LAURITZEN S. L. 1990. Techniques for bayesian analysis in expert systems. *Annals of Mathematics and Artificial Intelligence*, **2**:353–366.

TATUSOV R. L., FEDOROVA N. D., JACKSON J. D., JACOBS A. R., KIRYUTIN B., KOONIN E. V., KRYLOV D. M., MAZUMDER R., MEKHEDOV S. L., NIKOLSKAYA A. N., RAO B. S., SMIRNOV S., SVERDLOV A. V., VASUDEVAN S., WOLF Y. I., YIN J. J., et NATALE D. A. 2003. The cog database : an updated version includes eukaryotes. *BMC Bioinformatics*, **4**:41.

TODD A. E., ORENGO C. A., et THORNTON J. M. 2001. Evolution of function in protein superfamilies, from a structural perspective. *J Mol Biol*, **307**(4):1113–1143.

TODD A. E., MARSDEN R. L., THORNTON J. M., et ORENGO C. A. 2005. Progress of structural genomics initiatives : an analysis of solved target structures. *J Mol Biol*, **348**(5):1235–1260.

TOKURIKI N. et TAWFIK D. S. 2009. Protein dynamism and evolvability. *Science*, **324**(5924):203–207.

TULLER T., BIRIN H., GOPHNA U., KUPIEC M., et RUPPIN E. 2010. Reconstructing ancestral gene content by coevolution. *Genome Res*, **20**:122–132.

van DONGEN S., 2000. *Graph clustering by flow simulation.* Thèse de doctorat, University of Utrecht, The Nederlands.

VOGEL C., BASHTON M., KERRISON N. D., CHOTHIA C., et TEICHMANN S. A. 2004a. Structure, function and evolution of multidomain proteins. *Curr Opin Struct Biol*, **14**(2):208–216.

VOGEL C., BERZUINI C., BASHTON M., GOUGH J., et TEICHMANN S. A. 2004b. Supra-domains : evolutionary units larger than single protein domains. *J Mol Biol*, **336**(3):809–823.

VOGEL C., TEICHMANN S. A., et PEREIRA-LEAL J. 2005. The relationship between domain duplication and recombination. *J Mol Biol*, **346**(1):355–365.

WALDROP G. L., RAYMENT I., et HOLDEN H. M. 1994. Three-dimensional structure of the biotin

carboxylase subunit of acetyl-coa carboxylase. *Biochemistry*, **33**(34):10249–10256.

WANG Z. X. 1996. How many fold types of protein are there in nature ? *Proteins*, **26**(2):186–191.

WANG Z. X. 1998. A re-estimation for the total numbers of protein folds and superfamilies. *Protein Eng*, **11**(8):621–626.

WEINER J., BEAUSSART F., et BORNBERG-BAUER E. 2006. Domain deletions and substitutions in the modular protein evolution. *FEBS J*, **273**(9):2037–2047.

WEINER J. et BORNBERG-BAUER E. 2006. Evolution of circular permutations in multidomain proteins. *Mol Biol Evol*, **23**(4):734–743.

WEINER J., THOMAS G., et BORNBERG-BAUER E. 2005. Rapid motif-based prediction of circular permutations in multi-domain proteins. *Bioinformatics*, **21**(7):932–937.

WETLAUFER D. B. 1973. Nucleation, rapid folding, and globular intrachain regions in proteins. *Proc Natl Acad Sci U S A*, **70**(3):697–701.

WILSON D., PETHICA R., ZHOU Y., TALBOT C., VOGEL C., MADERA M., CHOTHIA C., et GOUGH J. 2009. Superfamily–sophisticated comparative genomics, data mining, visualization and phylogeny. *Nucleic Acids Res*, **37**(Database issue):D380–D386.

WOLF Y. I., GRISHIN N. V., et KOONIN E. V. 2000. Estimating the number of protein folds and families from complete genome data. *J Mol Biol*, **299**(4):897–905.

WOLF Y. I., ROGOZIN I. B., et KOONIN E. V. 2004. Coelomata and not ecdysozoa : evidence from genome-wide phylogenetic analysis. *Genome Res*, **14**(1):29–36.

WOOD V., RUTHERFORD K. M., IVENS A., RAJANDREAM M. A., et BARRELL B. 2001. A re-annotation of the saccharomyces cerevisiae genome. *Comp Funct Genomics*, **2**(3):143–154.

WOOTTON J. C. et FEDERHEN S. 1993. Statistics of local complexity in amino acid sequences and sequence databases. *Computers & Chemistry*, **17**(2):149 – 163.

WORTH C. L., GONG S., et BLUNDELL T. L. 2009. Structural and functional constraints in the evolution of protein families. *Nat Rev Mol Cell Biol*, **10**(10):709–720.

YAMAGUCHI H., KATO H., HATA Y., NISHIOKA T., KIMURA A., ODA J., et KATSUBE Y. 1993. Three-dimensional structure of the glutathione synthetase from escherichia coli b at 2.0 a resolution. *J Mol Biol*, **229**(4):1083–1100.

YANG S. et BOURNE P. E. 2009. The evolutionary history of protein domains viewed by species phylogeny. *PLoS One*, **4**(12):e8378.

YOOSEPH S., SUTTON G., RUSCH D. B., HALPERN A. L., WILLIAMSON S. J., REMINGTON K., EISEN J. A., HEIDELBERG K. B., MANNING G., LI W., JAROSZEWSKI L., CIEPLAK P., MILLER C. S., LI H., MASHIYAMA S. T., JOACHIMIAK M. P., van BELLE C., CHANDONIA J.-M., SOERGEL D. A., ZHAI Y., NATARAJAN K., LEE S., RAPHAEL B. J., BAFNA V., FRIEDMAN R., BRENNER S. E., GODZIK A., EISENBERG D., DIXON J. E., TAYLOR S. S., STRAUSBERG R. L., FRAZIER M., et VENTER J. C. 2007. The sorcerer ii global ocean sampling expedition : expanding the universe of protein families. *PLoS Biol*, **5**(3):e16.

ZHANG C. et DELISI C. 1998. Estimating the number of protein folds. *J Mol Biol*, **284**(5):1301–1305.

ZHANG C. T. 1997. Relations of the numbers of protein sequences, families and folds. *Protein Eng*, **10**(7):757–761.

ZHOU Q. et WANG W. 2008. On the origin and evolution of new genes–a genomic and experimental perspective. *J Genet Genomics*, **35**(11):639–648.

ZHOU Q., ZHANG G., ZHANG Y., XU S., ZHAO R., ZHAN Z., LI X., DING Y., YANG S., et WANG W. 2008. On the origin of new genes in drosophila. *Genome Res*, **18**(9):1446–1455.

Liste des publications en cours

- Article publié
  S. PENEL, A.-M. ARIGON, J.-F. DUFAYARD, A.-S. SERTIER, V. DAUBIN, L. DURET, M. GOUY ET G. PERRIÈRE. 2009. Databases of homologous gene families for comparative genomics. *BMC Bioinformatics*. **10 Suppl 6** :S3
- Article en cours de révision
  G. MARAIS, C. DUFAURE DE CITRES, A.-S. SERTIER ET V. DAUBIN. Unexpected genomic degeneration in the highly abundant free-living marine cyanobacteria Prochlorococcus. *Genome Biology and Evoluion*
- Article soumis
  A.-S. SERTIER, V. DAUBIN ET D. KAHN. A phylogenetic view of the expanding protein universe. *Plos Biology*.
- Article en cours de préparation
  A.-S. SERTIER, V. DAUBIN ET D. KAHN. Bayesian network models for evolutionary scenarios inference.. *Bioinformatics*.

BMC Bioinformatics

BioMed Central

Proceedings

Open Access

Databases of homologous gene families for comparative genomics

Simon Penel[1], Anne-Muriel Arigon[2], Jean-François Dufayard[2], Anne-Sophie Sertier[1], Vincent Daubin[1], Laurent Duret[1], Manolo Gouy[1] and Guy Perrière*[1]

Address: [1]Laboratoire de Biométrie et Biologie Évolutive, CNRS, Université Claude Bernard – Lyon 1, 43 bd. du 11 Novembre 1918, 69622 Villeurbanne Cedex, France and [2]Laboratoire d'Informatique, de Robotique et de Microélectronique de Montpellier, 161 rue Ada, 34392 Montpellier, France

Email: Simon Penel - penel@biomserv.univ-lyon1.fr; Anne-Muriel Arigon - Anne-muriel.Arigon@lirmm.fr; Jean-François Dufayard - jeanfrancois.dufayard@gmail.com; Anne-Sophie Sertier - sertier@biomserv.univ-lyon1.fr; Vincent Daubin - daubin@biomserv.univ-lyon1.fr; Laurent Duret - duret@biomserv.univ-lyon1.fr; Manolo Gouy - mgouy@biomserv.univ-lyon1.fr; Guy Perrière* - perriere@biomserv.univ-lyon1.fr

* Corresponding author

from European Molecular Biology Network (EMBnet) Conference 2008: 20th Anniversary Celebration
Martina Franca, Italy. 18–20 September 2008

Published: 16 June 2009

BMC Bioinformatics 2009, **10**(Suppl 6):S3 doi:10.1186/1471-2105-10-S6-S3

This article is available from: http://www.biomedcentral.com/1471-2105/10/S6/S3

Abstract

Background: Comparative genomics is a central step in many sequence analysis studies, from gene annotation and the identification of new functional regions in genomes, to the study of evolutionary processes at the molecular level (speciation, single gene or whole genome duplications, etc.) and phylogenetics. In that context, databases providing users high quality homologous families and sequence alignments as well as phylogenetic trees based on state of the art algorithms are becoming indispensable.

Methods: We developed an automated procedure allowing massive all-against-all similarity searches, gene clustering, multiple alignments computation, and phylogenetic trees construction and reconciliation. The application of this procedure to a very large set of sequences is possible through parallel computing on a large computer cluster.

Results: Three databases were developed using this procedure: HOVERGEN, HOGENOM and HOMOLENS. These databases share the same architecture but differ in their content. HOVERGEN contains sequences from vertebrates, HOGENOM is mainly devoted to completely sequenced microbial organisms, and HOMOLENS is devoted to metazoan genomes from Ensembl. Access to the databases is provided through Web query forms, a general retrieval system and a client-server graphical interface. The later can be used to perform tree-pattern based searches allowing, among other uses, to retrieve sets of orthologous genes. The three databases, as well as the software required to build and query them, can be used or downloaded from the PBIL (Pôle Bioinformatique Lyonnais) site at http://pbil.univ-lyon1.fr/.

Background

HOVERGEN, a database devoted to homologous gene families in vertebrates [1,2] has been first released in 1994. The motivation to develop this database was to build a system allowing to do large-scale comparative genomic studies on vertebrates. HOVERGEN allows to retrieve sets of orthologous genes in order to do evolutionary studies on gene families [3-12].

Two other systems based on the same architecture: HOGENOM and HOMOLENS are presented here. HOGENOM contains homologous gene families from all available complete genomes from bacteria, archaea and unicellular eukaryotes, plus some representative plants and animals. HOMOLENS contains gene families from complete animal genomes found in Ensembl [13]. In the three databases, after family assembly, protein sequences are aligned and the alignments produced are used to build phylogenetic trees. Those two steps are realized through an automated procedure.

These databases are structured under the ACNUC sequence database management system [14]. Access to these databases is possible through different implementations of the ACNUC libraries. The first one is the Web server available at PBIL [15]. The second one is the program Query, a retrieval system allowing to query local or remote ACNUC databases [16]. Lastly, a graphical interface named FamFetch allows to retrieve families and display associated data [17,18]. This program allows to perform pattern searches on the phylogenetic trees through a pattern-matching algorithm. This feature is especially helpful to retrieve sets of orthologous sequences, but also for any kind of studies involving the detection of phylogenetic profiles.

Materials and methods

Data harvesting and pre-processing

For the three systems, two ACNUC databases are built, one for the protein sequences and one for the corresponding nucleotide sequences. Protein sequences are stored in UniProtKB format [19] while nucleotide sequences are stored in EMBL format [20]. To build those databases, the sequences are gathered from different sources. In the case of HOVERGEN, protein sequences represent the primary source of information, and they are taken from UniProt. Nucleotide sequences are taken from EMBL, using the cross-references provided in UniProt. For HOMOLENS, nucleotide annotated sequences come from Ensembl and protein sequences are generated from the corresponding Coding DNA Sequences (CDS) described in Ensembl annotations. In the case of HOGENOM, data sources are represented by various nucleotide sequence collections that are used in a hierarchical manner. Sequences from Genome Reviews [21] are used first and then supplemented with various systems such as Ensembl, the NCBI microbial data repository and complete genomes collection, the European Bioinformatics Institute (EBI) complete genome data, sequences from the Joint Genome Institute (JGI), the Sanger Institute and the John Craig Venter Institute (JCVI). The CDS from these collections are translated, using the adequate genetic code and reading frame, to generate the corresponding protein sequences except when alternative splicing occurs. In this case only the longest CDS is translated. Annotations of the CDS are analysed to get information related to protein annotations. When cross-references to UniProt are found, UniProt entries are scanned to get information on function, product and bibliography to improve the annotations. The UniProt identifier is inserted into the annotations as a keyword and the UniProt accession number is inserted as a secondary accession number.

Inconsistencies or lack of precision in the taxonomic information present in some source databases are corrected, mostly in HOGENOM. In HOVERGEN, UniProt and EMBL sequence names and accession numbers are used. In HOGENOM and HOMOLENS, devoted to complete genomes, entries are renamed to directly provide information about the species identity and the location of genes in chromosomes. For nucleotide sequences, the first two letters of the genus, the first three letters of the species, a number identifying the strain, and another identifying the replicon, the chromosome, or the organelle make up sequence names. For each individual CDS, a suffix containing the two letters "PE" (for peptide) followed by its rank number in the replicon is added to the containing sequence's name. For example, ESCOL2_1.PE76 and ESCOL2_2.PE3371 correspond respectively to the sequence of the *traL* gene on plasmid *F* and the sequence of the *glgX* gene on the chromosome of *Escherichia coli* K12. For protein sequences, the same naming is employed, except that the CDS rank number is integrated in the sequence name (*e.g.*, ESCOL2_1_PE76 for the above mentioned *traL* gene). Note that original accession numbers are conserved and added to sequence annotations so that the coherence with original data source is conserved.

Clustering algorithm

To build families, a similarity search of all proteins against themselves, after filtering low complexity regions with SEG [22], is performed with the BLASTP2 program [23], the BLOSUM62 amino-acid similarity matrix [24], and a threshold of 10^{-4} for BLAST *E*-values. The Build_Fam program is used to cluster protein sequences into families. This program filters BLAST output in order to remove Homologous Segment Pairs (HSPs) that are incompatible with a global alignment (Figure 1). For complete protein sequences, two sequences in a pair are included in the

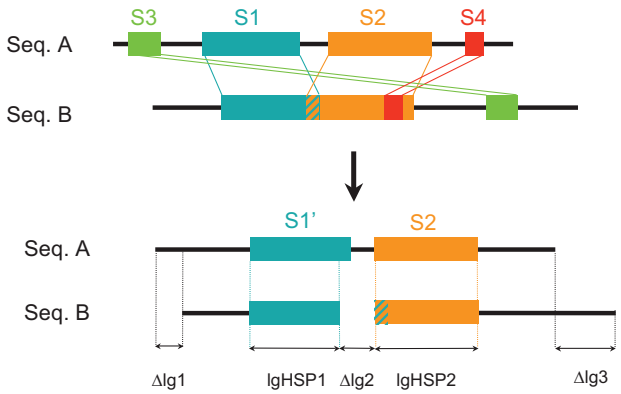


Figure 1
Removal of incompatible HSPs. For each couple of homologous sequences found by BLASTP, HSPs that are incompatible with a global alignment are removed. In this example, segments *S1* and *S2* are compatible, but segments *S3* and *S4* are not. They are therefore ignored by further computations on similarity measures which allow one to classify (or not) these two sequences in the same family.

same family if remaining HSPs cover at least 80% of the protein length and if their similarity is over 50% (two amino-acids are considered similar if their BLOSUM62 similarity score is positive). This couple of parameters will be denoted by 50/80 below.

Build_Fam uses a simple transitive link to build families. It means that if the pair of sequences {A, B} matches the conditions to be integrated in the same family and if the pair {A, C} also matches them, then sequences A, B and C will be clustered together, even if the pair {B, C} does not match the conditions. Once families of complete protein sequences are built, partial sequences are included in the classification. A partial sequence having similarity with a complete protein is included in a family if it fulfils the two conditions required for a complete sequence and if its length is ≥100 amino-acids or ≥50% of the length of the complete protein. When several families can be associated with a partial sequence, the sequence is included in the family that presents the complete sequence with the highest similarity.

Extensions of sequence annotations

Further sequence annotations are created after the clustering step. For protein sequences, a family identifier is added in the "CC" field. In the case of nucleotide sequences, this information is added in a "/gene_family" qualifier associated to each CDS. In both cases, this identifier is incorporated in the keywords associated to the corresponding entries in the ACNUC structure. It is thus possible to retrieve all the sequences in a family with this number when using any of the retrieval systems developed for our three databases. Some supplementary features corresponding to descriptions of non-coding regions are also introduced in the nucleotide sequences: "INT_INT" for internals introns (*i.e.*, within CDS), "5'NCR" for 5' non-coding regions, and "3'NCR" for 3' non-coding regions (*i.e.*, regions respectively upstream and downstream of annotated CDS, including UTRs and intergenic regions). Those supplementary features define what we call sub-sequences [14] which can be selected and extracted from the databases in the same way as CDS or structural RNAs.

Alignments and phylogenetic trees

Once the families are built, multiple alignments are computed on protein sequences using MUSCLE [25] with all default parameters. Alignments are filtered with Gblocks [26] in order to keep only their reliable parts. Based on our experience, Gblocks is used with parameters corresponding to relaxed conditions, in agreement with Talavera and Castresana [27]. Phylogenetic trees are computed with the fast maximum-likelihood algorithm implemented in PhyML [28], the JTT amino acid substitution model [29], and across-site rate variation modelled by a gamma distribution with four rate classes. Estimation of the α parameter for gamma distributions is carried out by PhyML. Internal branch support is estimated using the approximate Likelihood Ratio Test (aLRT) available in PhyML [30]. Due to the amount of time and memory required by computations on large families, alignments and tree computations were limited to families up to 1,000 sequences in HOVERGEN and up to 2,000 sequences in HOGENOM and HOMOLENS.

Tree reconciliation

All individual phylogenetic trees are reconciled with a species tree using the program RAP [18]. The reconciliation consists in the comparison of a gene tree with a species tree. When inconsistencies are detected between the two, they are explained by the presence of duplication events followed by selective losses in different lineages (Figure 2). The reference species tree used is the one provided by the NCBI taxonomic database http://www.ncbi.nlm.nih.gov/sites/entrez?db=taxonomy. During this process, some annotations are added to the reconciled trees. Those annotations consist in taxonomic data (*i.e.*, species names) and cross-references to the CDS corresponding to the protein sequences used to build the trees. Trees are rooted using the same reconciliation procedure. The root is placed to maximize the similarity between the gene tree and the species tree. All possible positions of the root in the gene tree are explored, and the one that requires the minimal number of gene duplications is retained. Tree reconciliation is used for HOVERGEN and HOMOLENS but not for HOGENOM because RAP does not model Horizontal Gene Transfers (HGTs),

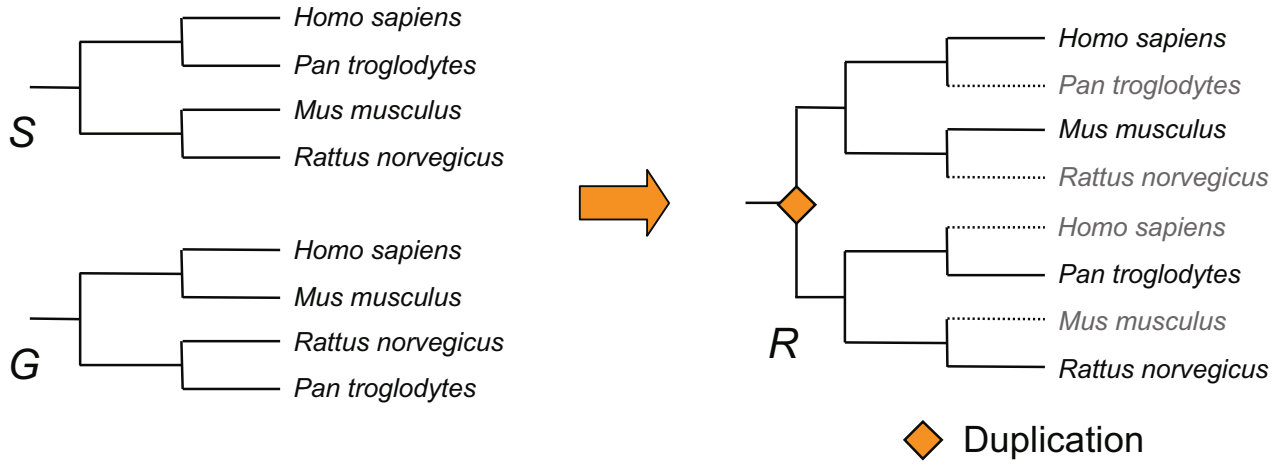


Figure 2
Tree reconciliation between a gene tree *G* and a species tree *S* showing different topologies. The result is the reconciled tree *R*. *R* is a variation of *S*, in which duplication nodes have been inserted in order to explain incongruence with *G*.

which are thought to be an important source of phylogenetic inconsistencies in prokaryotes [31-33].

Evaluation of clustering criteria

The efficiency and reliability of our clustering algorithm was assessed through a comparison with alternative approaches. We selected all the 219,951 protein sequences from 50 complete genomes including a panel of bacterial, archaeal and eukaryotic species in HOGENOM. For Build_Fam, three similarity/length-percentage combinations were experimented: the above-mentioned 50/80 and also the 40/80 and 40/90 combinations. We also applied the OrthoMCL and TribeMCL clustering programs on the same dataset. OrthoMCL is used to build the OrthoMCL-DB database [34]. This approach attempts to use evolutionary concepts such as orthology (*i.e.*, divergence after speciation events) and paralogy (*i.e.*, divergence after duplication events) to enforce a lower weight to paralogous relationships during the MCL clustering procedure [35]. This algorithm uses an inflation parameter (I) which regulates the cluster tightness. The default value for OrthoMCL is I = 1.5, but we also examined its behaviour with I = 1.1 and 4.0. TribeMCL is the algorithm used to build Tribes [36], and it is based on a similarity criterion provided by the user. Two different similarity criteria for TribeMCL were used: i) the simple BLAST *E*-value; and ii) our own score, Tribe(HSP), defined as:

$$\text{Tribe}(\text{HSP}_{xy}) = \sum_{\text{all HSP}} \frac{s\left(HSP_{xy}\right)}{\max\left(s_{xx}, s_{yy}\right)}$$

where x and y are two homologous protein sequences, $s(HSP_{xy})$ is the BLAST bit score for an HSP in an ordered list of HSPs found between x and y, and s_{zz} is the BLAST bit score between sequence z and itself. The value given to the inflation parameter for MCL in this case was the default one (I = 2).

The desired properties of a clustering algorithm for phylogenetic database reconstruction are twofold: first, the algorithm should be able to cluster homologous sequences from divergent organisms; second, the resulting alignments should nevertheless remain of high quality. After clustering, families based on each algorithm were aligned using MUSCLE with default parameters. To estimate the quality of alignments, six subsets of families were considered for each clustering algorithm: three containing all families with 10, 25 and 50 sequences, and three containing all families of 10, 25 and 50 species. The quality of alignments was assessed using two approaches: the NorMD index [37] which computes a similarity score over the entire alignment based on amino acid similarity (measured with PAM250 in this study); and Gblocks filtering [26] which we used as a measure of the number of gaps introduced in the alignment. When NorMD ≥ 0.5, the alignment is considered to be of good quality [37]. For Gblocks, the higher the percentage of sites conserved after filtering, the better the alignment. We used the default parameters for Gblocks (all gaps are removed), and we considered empirically that the alignments were of good quality if the percentage of conserved sites was ≥50%.

Databases access

As of October 2008, HOGENOM and HOMOLENS gather the information of complete genomes from respectively 513 and 41 species, while HOVERGEN contains 415,383 vertebrate proteins, and these three databases are regularly updated. They all provide high quality alignments and phylogenetic trees that can be queried and downloaded using a wide variety of tools, allowing to perform from very simple text searches to complex queries. Contents in terms of sequences and families for the present releases of the three databases are given in Table 1.

Web services

Sequences and families can be selected and retrieved via the PBIL server http://pbil.univ-lyon1.fr/. This server provides convenient and flexible web forms for selecting sequences and families by many different criteria in several databases [38], including the general repository collections such as Ensembl, UniProt, GenBank [39] or EMBL. The core of the service is represented by the WWW-Query application [15]. The corresponding form allows the combination of up to four criteria to retrieve sequences or gene families. Among the allowed criteria are: sequence names, accession numbers, keywords, taxonomic data, organelle, molecule type (CDS, RNA, or the supplementary features described in the **Extensions to sequence annotations** section), bibliographical references, date of insertion in the repository collections. Each time a query is performed, the list of matching sequences is stored on the server, and it is possible to re-use previously created lists to refine queries. The Quick Search form represents a simpler version of this application. With this form, the user enters only a string corresponding to a sequence name, an accession number, a keyword or a species name, and all the sequences or families associated to a criterion matching the string will be sorted. Note that the use of wildcard for fuzzy searches is allowed with both WWW-Query and Quick Search.

The Cross Taxa application gives access to a family retrieval system based on taxonomic criteria. It allows to retrieve gene families that are shared by a first set of taxa and (optionally) that are not present in a second set of taxa. Any taxonomic level can be used and mixed to compose the query (*e.g.*, *Homo sapiens*, Mammalia, Metazoa). For example it is possible to retrieve all gene families specific to a toxic bacterial strain, all gene families present in human but not in rodents, or all metazoan-specific gene families.

Alignments can be displayed on static HTML pages with several colouring options and they can be edited in order to visualize only a subset of sequences (Figure 3). Alternatively they can be visualised with the JalView applet [40] or downloaded on local disk. Phylogenetic trees are displayed as a clickable Portable Network Graphics (PNG) picture generated with Perl modules [41] and coloured according to taxonomy. Several displaying options are available, allowing to visualize species names, sequence name. Alternatively, trees can be visualised with the ATV applet [42] or downloaded.

Standard BLAST similarity searches can be performed on the three databases, but it is also possible to use a specific tool named HoSeqI [43]. With HoSeqI, instead of simply identifying the sequences in a database that are the most similar to a query sequence, the application identifies the most similar family. Then the query sequence is integrated into this family and the corresponding alignment and tree are recomputed on the fly. For that purpose, a panel of different multiple alignment and tree building programs is proposed to the user. Especially, it is possible to use profile alignments algorithms instead of performing *de novo* complete alignments. Therefore, the complete identification process can be very fast. Again, alignments and trees can be visualized on static HTML pages or through the use of JalView and ATV applets.

Lastly, note that HOVERGEN and HOGENOM families and phylogenetic trees can be directly accessed from the UniProt Web site http://www.uniprot.org/uniprot, through cross-references of the "Phylogenomic databases" field.

ACNUC remote connection

The ACNUC database system handles any sequence collection structured with the GenBank, EMBL or UniProt flat file formats. Recently, network access to ACNUC data-

Table 1: Databases content for HOVERGEN, HOGENOM and HOMOLENS.

	HOVERGEN	HOGENOM	HOMOLENS
Proteins	415,383	2,142,639	672,064
CDS	613,473	2,128,552	892,572
Genomic sequences	541,405	135,105	178,069
Families	16,673	147,586	23,155
Orphans	24,234 (5%)	397,545 (18%)	90,953 (13%)
Proteins associated to a family	311,647 (75%)	1,742,390 (81%)	579,620 (86%)
Unclassified partial sequences	79,502 (19%)	-	-

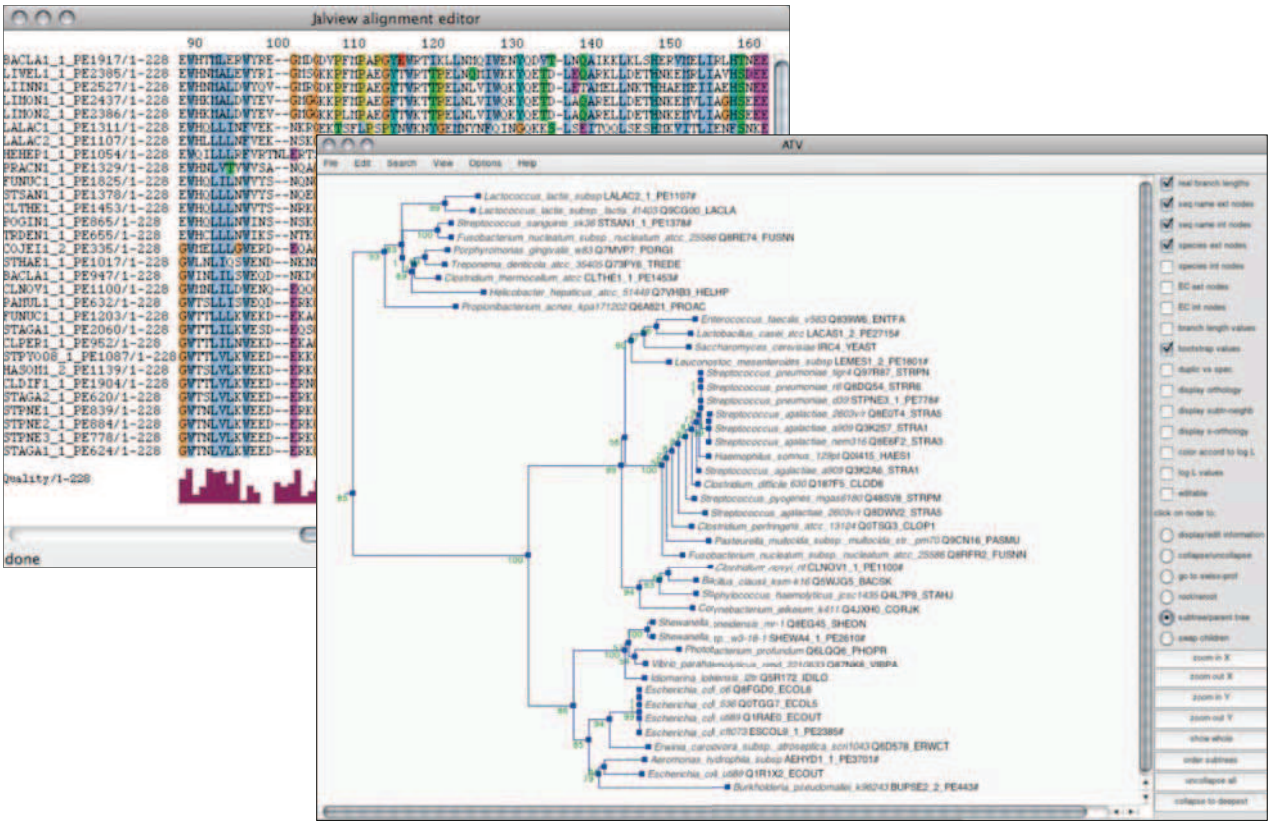


Figure 3
Multiple alignments and phylogenetic trees visualization through the PBIL Web interface. In this exemple, the alignment is displayed with the JalView applet and the phylogenetic tree is displayed with the ATV applet.

bases has been achieved by the definition and implementation of a remote ACNUC access protocol that governs information exchanges between the PBIL and remote clients [16]. This protocol uses a TCP/IP socket connection to a dedicated server and makes retrieval operations to remote ACNUC databases nearly as fast as to local databases with usual academic Internet connections.

HOVERGEN, HOGENOM and HOMOLENS can be accessed with two client programs: Query_win with a graphical user interface, and raa_query with a command-line interface. The latter is useful in a scripting context, possibly to repeatedly execute fixed retrieval operations. Both of them allow to compose complex queries involving multiple criteria, extraction of sequences and subsequences into local files, and access to keywords and taxonomic data browsers. Query_win executables are available for major computing platforms, therefore most Internet-connected computers can run an ACNUC client and access the PBIL databases.

The remote ACNUC access protocol has also been interfaced with two programming languages, C and Python, and the widely used statistical computing environment R [44]. Therefore, it is possible for users to write their own programs in any of these languages in order to access ACNUC databases. Furthermore, the R binding is included into an official R package called seqInR [45]. This package provides various tools for statistical and evolutionary analyses of biological sequences and access to the very large set of libraries available in the R environment.

FamFetch interface

FamFetch is a Java client allowing to access sequence data, as well as the alignments and trees present in HOVERGEN, HOGENOM and HOMOLENS [17]. Starting from the main window of the interface it is possible to access the whole list or a personal subset of families and to make queries to retrieve those matching specific criteria (Figure 4). An equivalent of the Cross Taxa application is also implemented. After selection of a family, the correspond-

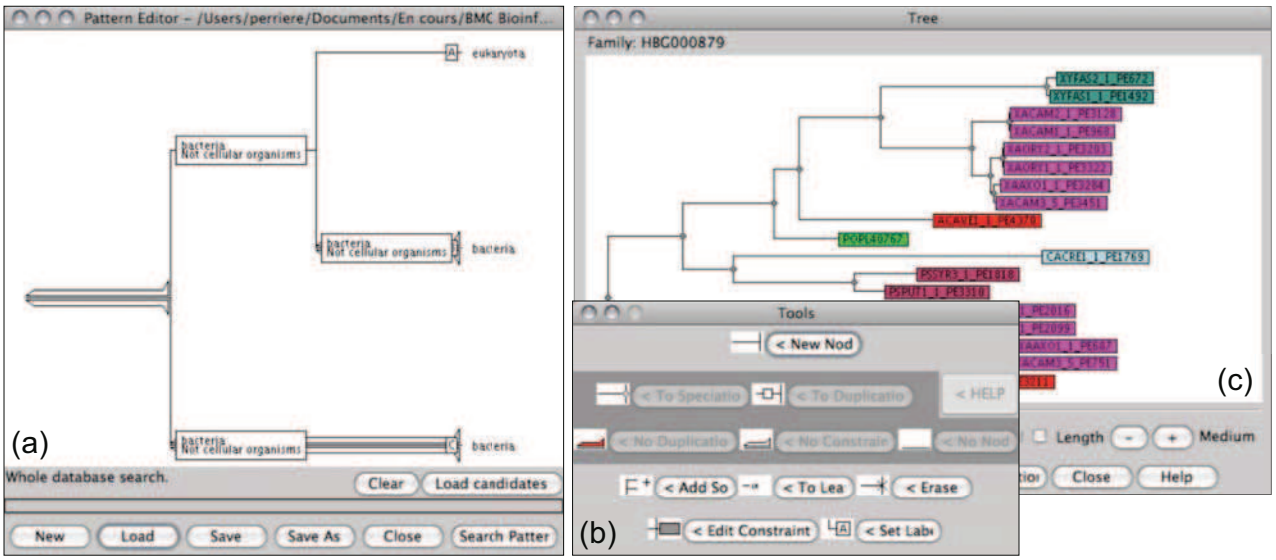


Figure 4
Three different frames of the FamFetch interface. Frame (a) is an interactive editor that allows users to build any pattern, node by node and leaf by leaf. Here the pattern entered allows to detect families in which an eukaryotic species is placed within a clade of bacterial species. Frame (b) allows to choose between tools to use in the editor. Tools surrounded by dark grey are those that use the gene duplication predictions, and can be avoided if the user does not want to trust this information. Frame (c) is the tree display. In this frame, sequence are displayed using a colour code corresponding to the taxonomy.

ing phylogenetic tree is displayed in the tree window. In this tree, sequences are coloured using a code reflecting the taxonomic position of the corresponding species. A choice of four different editable colouring schemes is proposed to the user. The tree display is active, with options of re-rooting, node swapping, subtree selection or zooming. Clicking on leaves allows users to visualize the entries from UniProt and EMBL or the alignment of the selected sequences.

A major feature of FamFetch is the possibility to retrieve families showing specific tree patterns [18]. The interface integrates a tree pattern editor allowing to define a pattern that will be searched in the set of phylogenetic trees. After the pattern matching operation, the main frame of FamFetch displays the list of matching families. The results can be saved in a file, each pattern being numbered and described with its gene list. Thanks to the possibility to introduce duplications and/or taxonomic data constraints in search patterns, it is possible to easily detect ancient gene duplications or to select orthologous genes. For that purpose, the user only needs to build a pattern in which duplications are forbidden. The whole tree pattern search operation really makes sense with the tree reconciliation performed with RAP. Indeed, with reconciled trees, even hidden paralogies due to duplications followed by gene losses in some lineages are taken into account in the pattern search process.

The use of the tree pattern matching algorithm to retrieve sets of orthologous genes has been previously described [18]. The approach to orthology inference implemented by the RAP tree pattern matching algorithm is very different from that used by most other systems such as COGs [46], OrthoMCL-DB [47] or Inparanoid [48], and is the only one based on phylogenetic analysis. But this tool can be also used for other purposes, and in the case of HOGENOM, it is possible to search for genes that may have been obtained by HGT in some species. HGTs are known to be an important driving force in prokaryotes evolution [31-33], and the question of their detection has raised a lot of methodological problems [49-51]. It is generally admitted that the phylogenetic methods (*i.e.*, the methods based on the use of phylogenetic trees) are the most efficient ones to identify HGTs [50-52]. In order to detect transfers with a database like HOGENOM, the simplest thing to do is to search for anomalous patterns in trees, for instance patterns that are violating the monophyly of a well-established group of species.

A possible example of search of this kind is summarized in Figures 4 and 5. In this search, the pattern entered allows to detect families in which an eukaryotic species is placed within a clade of bacterial species (Figure 4). When performed on the release 4 of HOGENOM (February 2008), this search returns 1,304 trees, two of which are shown in Figure 5. Many of these patterns represent prob-

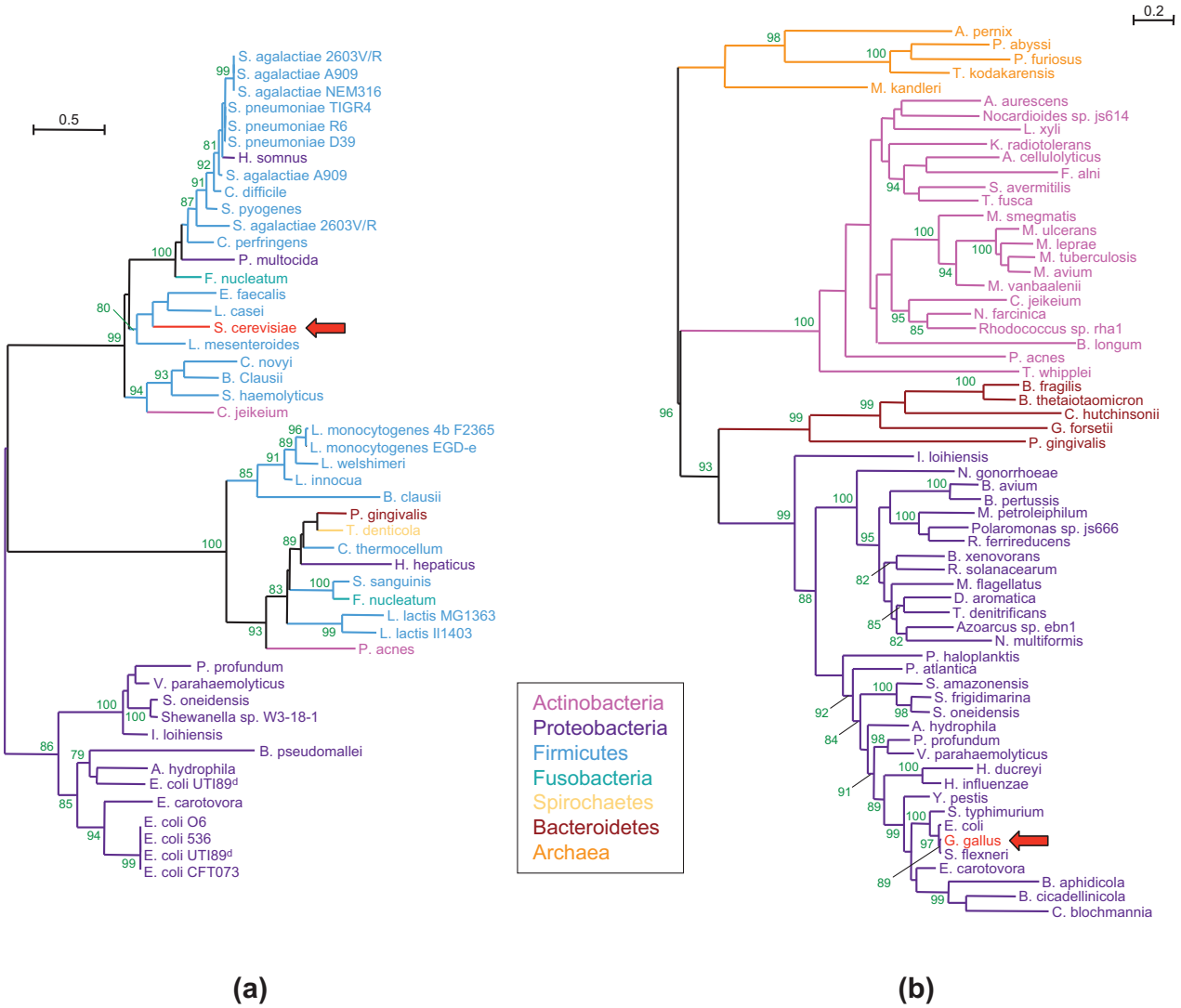


Figure 5

Exemple of trees containing anomalous patterns involving eukaryotes and bacteria. A search on the pattern shown in Figure 4 has been performed on HOGENOM release 4, and this search returned a total of 1,304 families. Two trees taken among the 1,304 are shown in this figure. Family HBG082165 (a) corresponds to a conserved hypothetical protein, and it shows a *S. cerevisiae* sequence among Lactobacillales species. Family HBG459980 (b) corresponds to the 3-phosphoshikimate 1-carboxyvinyltransferase enzyme, and it shows a *G. gallus* sequence among Proteobacteria species. Values of the aLRT test are given for the internal branches, and only values with a $P > 80\%$ are shown.

able contaminations rather than real HGTs, an example of this being the presence of *Gallus gallus* among Proteobacteria sequences in HBG459980 family. More plausible is the case of family HBG082165 that shows a possible HGT of a gene encoding an hypothetical protein from a Lactobacillales species to the yeast *Saccharomyces cerevisiae*.

Programs and data availability

All software, and databases can be freely used and/or downloaded from the PBIL server at http://pbil.univ-lyon1.fr. Executable files for Windows, MacOSX, Linux X86 and Solaris of the graphical interface version of Query are distributed, as well as standard C sources for the command-line version. For the FamFetch and RAP programs, Java sources as well as their compiled classes are provided. For the databases, ACNUC index tables, sequence files in

Table 2: Clustering results for Build_Fam, OrthoMCL and TribeMCL

	Build_Fam				Ortho_MCL			Tribe_MCL
Parameters	50/80	40/80	*E*-value	HSP	1.5	4.0	*E*-value	HSP
Nb. clustered seq.	119222	144956	157993	186779	171129	169507	157993	186779
% clustered seq.	54%	66%	72%	85%	78%	77%	72%	85%
Nb. families	20706	17043	19608	19344	23966	31343	19608	19344
Avg. seq./family	5.76	8.51	8.06	9.66	7.14	5.41	8.06	9.66
Families ≥ 1000	1	6	1	1	0	0	1	1
Largest family	1580	2642	1121	1185	479	281	1121	1185
Families sp. = 1	10359 (50%)	8050 (47%)	8379 (43%)	6735 (35%)	7828 (33%)	10134 (32%)	8379 (43%)	6735 (35%)
Families sp. = 50	13 (0.6‰)	34 (2‰)	19 (1‰)	30 (1.6‰)	27 (1.1‰)	5 (0.2‰)	19 (1‰)	30 (1.6‰)
Famílles sp. ≥ 25	504 (2.4%)	620 (3.6%)	630 (3.2%)	744 (3.9%)	734 (3.1%)	554 (1.8%)	630 (3.2%)	744 (3.9%)

The parameters used for the algorithms correspond to the similarity/length combination in the case of Build_Fam, to the inflation parameter in the case of OrthoMCL, and to the two scores used in the case of TribeMCL. The three last lines give the number and percentage of families containing only one species (sp. = 1), 50 different species (sp. = 50), and at least 25 different species (sp. ≥ 25).

EMBL and UniProt format, alignments in Clustal format [53], and trees in Newick format [54] are provided. The seqInR package is available from any Comprehensive R Archive Network (CRAN) mirror. All data used to estimate the reliability of Build_Fam and its comparison with other clustering algorithms can be downloaded at ftp://pbil.univ-lyon1.fr/pub/datasets/BMC2009/ .

Results and discussion

Tree reconciliation

The main originality of our system is the possibility to make queries using tree patterns, as this allows users not only to search for orthologs but also for HGTs, gene duplications or any phylogenetic profile of interest. Also, it is possible to perform tree pattern searches on reconciled or non-reconciled databases, the only difference being that duplications need to be described explicitly by the user in a non-reconciled database.

Clustering algorithm

The comparison of clustering methods revealed that different approaches have different desirable properties. An ideal algorithm for building phylogenetic tree databases would be fast, producing high quality alignments while maximizing species representation in protein families. In terms of speed, Build_Fam indisputably outperformed both TribeMCL and OrthoMCL (respectively less than an hour, 3 hours and 41 hours to cluster 219,951 sequences on a Sun Fire 880, UltraSparc-III, 8 × 900 MHz CPUs, 28 Gbytes RAM). In the clustering procedure, OrthoMCL and TribeMCL always cluster a significantly larger fraction of sequences than Build_Fam with respectively 77–78% and 72–85% against 54–58%, depending on the parameters used for each program (Table 2). As expected, when the Build_Fam similarity threshold is made less stringent, the number of families generated decreases while the average number of sequences per family increases. This average number of sequences per family is usually low because many families have a small number of sequences. An important difference is the fact that Build_Fam and TribeMCL have a tendency to generate a small number of very large families (containing >1,000 sequences), in contrast with OrthoMCL. Overall the clustering criteria appear more stringent in Build_Fam, and therefore the proportion of families that include representatives from more than one kingdom is lower (Table 3). Furthermore when the number of species represented in a family grows, Build_Fam tends to have more sequences per species, and thus to have more redundancy than OrthoMCL (excepted for I = 1.1). The tendency of reducing redundancy in families is a build-in characteristic of the OrthoMCL algorithm and is therefore not surprising. It may not, however, be a desirable property for the present databases.

Although it detected less universal families, Build_Fam almost consistently produced better alignments than other methods, either for the NorMD index or the number of gaps as detected by Gblocks (Table 4). When the number of sequences or species is low, Build_Fam 50/80

Table 3: Proportion of families integrating sequences from one, two or the three kingdoms of life (Bacteria, Archaea and Eukaryota).

	Build_Fam			OrthoMCL			TribeMCL	
Parameters	50/80	40/80	40/90	1.1	1.5	4.0	*E*-value	HSP
1 kingdom	91%	88%	89%	86%	84%	85%	87%	83%
2 kingdoms	7%	9%	8%	10%	13%	13%	10%	13%
3 kingdoms	2%	3%	3%	4%	4%	2%	3%	4%

Table 4: Alignment quality results for the Build_Fam, OrthoMCL and TribeMCL algorithms

Algo.	Families	Nb. families	Mean nb. seq.	Mean nb. sp.	Mean %Gbl.	Nb. fam. %Gbl. >50%	Mean NorMD	Nb. fam. NorMD >0.5
BF 50/80		213	10	5.99	**63%**	**172 (80.8%)**	**0.73**	**207 (97.2%)**
BF 40/80		190	10	5.44	51%	96 (50.5%)	0.67	151 (79.5%)
BF 40/90		179	10	6.01	53%	104 (58.1%)	0.65	144 (80.4%)
Ortho 1.1	Seq. = 10	270	10	5.09	36%	76 (28.1%)	0.34	149 (55.2%)
Ortho 1.5		447	10	6.02	38%	136 (30.4%)	0.44	246 (55.0%)
Ortho 4.0		450	10	6.3	45%	186 (41.3%)	0.59	300 (66.7%)
Tribe *E*-value		290	10	5.09	43%	111 (38.3%)	0.59	199 (68.6%)
Tribe HSP		373	10	5.59	31%	77 (20.6%)	0.13	149 (39.9%)
BF 50/80		35	25	16.6	**51%**	**18 (51.4%)**	**0.61**	**26 (74.3%)**
BF 40/80		37	25	16.03	34%	9 (24.3%)	0.46	14 (37.8%)
BF 40/90		45	25	17.27	41%	15 (33.3%)	0.50	25 (55.6%)
Ortho 1.1	Seq. = 25	49	25	15.47	22%	5 (10.2%)	0.05	14 (28.6%)
Ortho 1.5		70	25	16.96	27%	8 (11.4%)	0.33	31 (44.3%)
Ortho 4.0		51	25	18.22	35%	12 (23.5%)	0.45	25 (49.0%)
Tribe *E*-value		38	25	13.683	27%	4 (10.5%)	0.38	11 (28.9%)
Tribe HSP		55	25	14.75	23%	5 (9.1%)	0.13	12 (21.8%)
BF 50/80		7	50	23.29	**35%**	**2 (28.6%)**	**0.49**	2 (28.6%)
BF 40/80		9	50	29.33	28%	0 (0.0%)	0.43	**5 (55.6%)**
BF 40/90		15	50	25.8	22%	1 (6.7%)	0.39	8 (53.3%)
Ortho 1.1	Seq. = 50	23	50	29.91	11%	1 (4.3%)	-0.30	4 (17.4%)
Ortho 1.5		18	50	28.28	17%	1 (5.6%)	0.14	4 (22.2%)
Ortho 4.0		4	50	37	25%	0 (0.0%)	0.48	1 (25.0%)
Tribe *E*-value		11	50	29.64	13%	0 (0.0%)	0.14	0 (0.0%)
Tribe HSP		17	50	30.88	16%	0 (0.0%)	0.00	3 (17.6%)
BF 50/80		107	12.1	10	**60%**	**82 (76.6%)**	**0.66**	**101 (94.4%)**
BF 40/80		102	15.06	10	44%	46 (45.1%)	0.53	63 (61.8%)
BF 40/90		113	13.65	10	48%	51 (45.1%)	0.59	80 (70.8%)
Ortho 1.1	Sp. = 10	121	14.58	10	34%	34 (28.1%)	0.22	59 (48.8%)
Ortho 1.5		224	12.3	10	40%	73 (32.6%)	0.43	119 (53.1%)
Ortho 4.0		221	11.59	10	46%	100 (45.2%)	0.43	142 (64.3%)
Tribe *E*-value		128	14.52	10	47%	56 (43.8%)	0.53	85 (66.4%)
Tribe HSP		172	15.37	10	33%	45 (26.2%)	0.31	74 (43.0%)
BF 50/80		32	30.91	25	**40%**	8 (25.0%)	0.51	19 (59.4%)
BF 40/80		23	42.13	25	28%	7 (30.4%)	0.34	8 (34.8%)
BF 40/90		31	41.74	25	37%	**11 (35.5%)**	0.39	13 (41.9%)
Ortho 1.1	Sp. = 25	36	51	25	16%	2 (5.6%)	0.14	13 (36.1%)
Ortho 1.5		33	37.64	25	28%	5 (15.2%)	0.43	18 (54.5%)
Ortho 4.0		30	27.97	25	35%	7 (23.3%)	**0.52**	**20 (66.7%)**
Tribe *E*-value		26	45.19	25	22%	1 (3.8%)	0.30	10 (38.5%)
Tribe HSP		42	46.29	25	18%	2 4.8%)	0.24	11 (26.2%)
BF 50/80		13	61.38	50	30%	1 (7.7%)	**0.54**	**8 (61.5%)**
BF 40/80		34	181.15	50	23%	4 (11.8%)	0.16	15 (44.1%)
BF 40/90		23	206.3	50	26%	**4 (17.4%)**	0.16	10 (43.5%)
Ortho 1.1	Sp. = 50	55	70.35	50	20%	2 (3.6%)	0.27	18 (32.7%)
Ortho 1.5		27	57.04	50	27%	2 (7.4%)	0.50	14 (51.9%)
Ortho 4.0		5	53.4	50	**32%**	0 (0.0%)	0.51	3 (60.0%)
Tribe *E*-value		19	87.42	50	19%	2 (10.5%)	0.21	5 (26.3%)
Tribe HSP		30	87.67	50	18%	1 (3.3%)	0.16	6 (20.0%)

The different parameters used for Build_Fam (BF), OrthoMCL (Ortho) and TribeMCL (Tribe) are given in the first column. The best scores in four last columns are shown in bold.

generates alignments that are much better than those obtained with OrthoMCL or TribeMCL. On the other hand, for large and very large families, the quality of the alignments considerably decreases. Considering the largest family generated by Build_Fam 50/80 (1,580 sequences) it happens that it is split by OrthoMCL 1.5 into 104 different families (corresponding to 92% of the total of sequences). The alignments of those 104 families are good as their average NorMD index is >0.5 for 96 families and their average site selection by Gblocks is >60%. There is therefore a tendency of Build_Fam to integrate divergent sequences on very large families.

On average, the better alignments obtained with Build_Fam for families up to 50 sequences or species can be explained by the double constraint put on the similarity and length of the pair of proteins. This increase in quality is partly counter-balanced by the use of a simple transitive link to incorporate sequences in a family. The use of a complete link would probably ensure an even better alignment quality, but at the cost of many families split. As this phenomenon of families splitting is already important with Build_Fam in its present state, it is probably not worth considering this model of sequence integration for the moment. Remarkably, the 50/80 parameter combination – which was chosen empirically for the first HOBACGEN release [17] – gives better results than the other combinations tested (40/80 and 40/90). This parameter choice thus appears as a good compromise between family size (and therefore family exhaustivity) and alignment quality. As the quality of a phylogenetic tree is the direct consequence of the quality of the corresponding sequence alignment, it is of special importance to have good alignments in our databases. In that context, the lower exhaustivity – materialized by the fact that Build_Fam tends to include only sequences from one kingdom in a family – is acceptable.

Parallel computing

The sizeable computational volume required by the construction of HOGENOM, HOVERGEN and HOMOLENS has been performed using the computing facilities of the Institut National de Physique Nucléaire et de Physique des Particules (IN2P3). This computing centre provides access to a 2,300 CPU cluster that can efficiently parallelize tasks such as BLAST searches or the construction of thousands of alignments and trees. The use of parallel computing brought a major improvement since computation time has been reduced by a factor of 50 to 100.

Conclusion

The different databases descibed in this paper are useful tools that has been used in many published biological studies but it might be desirable to create a general gene family database, combining sequence data from all available taxa. One important difficulty is that this would considerably increase the size of many gene families, and hence this would make the phylogenetic trees much more difficult to browse and interpret. Moreover, the global quality of the trees themselves would be drastically lowered because of the difficulty to compute reliable multiple alignments with very large families. Given that users are generally interested only in a particular clade, we decided to maintain three different databases (HOVERGEN, HOGENOM and HOMOLENS), whose content is partly overlapping, but that focus on different clades and different kinds of data (complete genome sequences *vs.* all data available for one clade). Also, we plan to develop a strategy including an incremental all-against-all BLAST search performed on a whole general protein sequence repository collection (such as UniProt). We will provide procedures allowing users to: i) extract a subset from the exhaustive set of protein similarities detected; ii) use this subset to create a specific database. Moreover, we wish to develop tools that would allow the user to automatically edit phylogenetic trees to display only a subset of sequences representative of the taxa of interest.

Competing interests

The authors declare that they have no competing interests.

Authors' contributions

SP is in charge of the database maintenance, the development of HOGENOM and HOMOLENS, and the present developments on the Web site. AMA developed the HoSeqI system. JFD developed the RAP program and the tree pattern search algorithm implemented in FamFetch. ASS and VD did the comparisons of the different clustering algorithms. LD conceived the database structure and wrote the Build_Fam program. MG developed the ACNUC system and the Query program, as well as its C API. GP developed the FamFetch interface, the core of the Web interface and wrote the manuscript.

Acknowledgements

This work has been supported by the ANR grant ANR-08-EMER-011-03 "Phylariane". We thank the IN2P3 (Villeurbanne) for the computing resources and Pascal Calvat for his technical help.

This article has been published as part of *BMC Bioinformatics* Volume 10 Supplement 6, 2009: European Molecular Biology Network (EMBnet) Conference 2008: 20th Anniversary Celebration. Leading applications and technologies in bioinformatics. The full contents of the supplement are available online at http://www.biomedcentral.com/1471-2105/10?issue=S6.

References

1. Duret L, Mouchiroud D, Gouy M: **HOVERGEN: a database of homologous vertebrate genes.** *Nucleic Acids Res* 1994, **22:**2360-2365.
2. Duret L, Perrière G, Gouy M: **HOVERGEN: database and software for comparative analysis of homologous vertebrate genes.** In *Bioinformatics Databases and Systems* Edited by: *Letovsky S. Boston: Kluwer Academic Publishers*; 1999:13-29.

3. Graur D, Duret L, Gouy M: **Phylogenetic position of the order Lagomorpha (rabbits, hares and allies).** *Nature* 1996, **379:**333-335.
4. Hedges SB, Parker PH, Sibley CG, Kumar S: **Continental breakup and the ordinal diversification of birds and mammals.** *Nature* 1996, **381:**226-229.
5. Makalowski W, Boguski MS: **Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences.** *Proc Natl Acad Sci USA* 1998, **95:**9407-9412.
6. Eyre-Walker A, Keightley PD: **High genomic deleterious mutation rates in hominids.** *Nature* 1999, **397:**344-347.
7. Duret L, Mouchiroud D: **Determinants of substitution rates in mammalian genes: expression pattern affects selection intensity but not mutation rate.** *Mol Biol Evol* 2000, **17:**68-74.
8. Chen FC, Li WH: **Genomic divergences between humans and other hominoids and the effective population size of the common ancestor of humans and chimpanzees.** *Am J Hum Genet* 2001, **68:**444-456.
9. Nei M, Xu P, Glazko G: **Estimation of divergence times from multiprotein sequences for a few mammalian species and several distantly related organisms.** *Proc Natl Acad Sci USA* 2001, **98:**2497-2502.
10. Lercher MJ, Urrutia AO, Hurst LD: **Clustering of housekeeping genes provides a unified model of gene order in the human genome.** *Nat Genet* 2002, **31:**180-183.
11. Kim SH, Elango N, Warden C, Vigoda E, Yi SV: **Heterogeneous genomic molecular clocks in primates.** *PLoS Genet* 2006, **2:**e163.
12. Studer RA, Penel S, Duret L, Robinson-Rechavi M: **Pervasive positive selection on duplicated and nonduplicated vertebrate protein coding genes.** *Genome Res* 2008, **18:**1393-1402.
13. Flicek P, Aken BL, Beal K, Ballester B, Caccamo M, Chen Y, Clarke L, Coates G, Cunningham F, Cutts T, Down T, Dyer SC, Eyre T, Fitzgerald S, Fernandez-Banet J, Graf S, Haider S, Hammond M, Holland R, Howe KL, Howe K, Johnson N, Jenkinson A, Kahari A, Keefe D, Kokocinski F, Kulesha E, Lawson D, Longden I, Megy K, Meidl P, Overduin B, Parker A, Pritchard B, Prlic A, Rice S, Rios D, Schuster M, Sealy I, Slater G, Smedley D, Spudich G, Trevanion S, Vilella AJ, Vogel J, White S, Wood M, Birney E, Cox T, Curwen V, Durbin R, Fernandez-Suarez XM, Herrero J, Hubbard TJ, Kasprzyk A, Proctor G, Smith J, Ureta-Vidal A, Searle S: **Ensembl 2008.** *Nucleic Acids Res* 2008, **36:**D707-714.
14. Gouy M, Gautier C, Attimonelli M, Lanave C, di Paola G: **ACNUC – a portable retrieval system for nucleic acid sequence databases: logical and physical designs and usage.** *Comput Applic Biosci* 1985, **1:**167-172.
15. Perrière G, Gouy M: **WWW-Query: an on-line retrieval system for biological sequence banks.** *Biochimie* 1996, **78:**364-369.
16. Gouy M, Delmotte S: **Remote access to ACNUC nucleotide and protein sequence databases at PBIL.** *Biochimie* 2008, **90:**555-562.
17. Perrière G, Duret L, Gouy M: **HOBACGEN: database system for comparative genomics in bacteria.** *Genome Res* 2000, **10:**379-385.
18. Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G: **Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases.** *Bioinformatics* 2005, **21:**2596-2603.
19. The UniProt Consortium: **The Universal Protein Resource (UniProt) 2009.** *Nucleic Acids Res* 2009, **37:**D169-174.
20. Cochrane G, Akhtar R, Bonfield J, Bower L, Demiralp F, Faruque N, Gibson R, Hoad G, Hubbard T, Hunter C, Jang M, Juhos S, Leinonen R, Leonard S, Lin Q, Lopez R, Lorenc D, McWilliam H, Mukherjee G, Plaister S, Radhakrishnan R, Robinson S, Sobhany S, Hoopen PT, Vaughan R, Zalunin V, Birney E: **Petabyte-scale innovations at the European Nucleotide Archive.** *Nucleic Acids Res* 2009, **37:**D19-25.
21. Sterk P, Kulikova T, Kersey P, Apweiler R: **The EMBL nucleotide sequence and Genome Reviews databases.** *Methods Mol Biol* 2007, **406:**1-22.
22. Wootton JC, Federhen S: **Analysis of compositionally biased regions in sequence databases.** *Methods Enzymol* 1996, **266:**554-571.
23. Altschul SF, Madden TL, Schaffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ: **Gapped BLAST and PSI-BLAST: A new generation of protein database search programs.** *Nucleic Acids Res* 1997, **25:**3389-3402.
24. Henikoff S, Henikoff JG: **Amino acid substitution matrices from protein blocks.** *Proc Natl Acad Sci USA* 1992, **89:**10915-10919.
25. Edgar RC: **MUSCLE: a multiple sequence alignment method with reduced time and space complexity.** *BMC Bioinformatics* 2004, **5:**113.
26. Castresana J: **Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis.** *Mol Biol Evol* 2000, **17:**540-552.
27. Talavera G, Castresana J: **Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments.** *Syst Biol* 2007, **56:**564-577.
28. Guindon S, Gascuel O: **A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood.** *Syst Biol* 2003, **52:**696-704.
29. Jones DT, Taylor WR, Thornton JM: **The rapid generation of mutation data matrices from protein sequences.** *Comput Appl Biosci* 1992, **8:**275-282.
30. Anisimova M, Gascuel O: **Approximate likelihood-ratio test for branches: A fast, accurate, and powerful alternative.** *Syst Biol* 2006, **55:**539-552.
31. Ochman H, Lawrence JG, Groisman EA: **Lateral gene transfer and the nature of bacterial innovation.** *Nature* 2000, **405:**299-304.
32. Gogarten JP, Townsend JP: **Horizontal gene transfer, genome innovation and evolution.** *Nat Rev Microbiol* 2005, **3:**679-687.
33. Ochman H, Lerat E, Daubin V: **Examining bacterial species under the specter of gene transfer and exchange.** *Proc Natl Acad Sci USA* 2005, **102(Suppl 1):**6595-6599.
34. Li L, Stoeckert CJ Jr, Roos DS: **OrthoMCL: identification of ortholog groups for eukaryotic genomes.** *Genome Res* 2003, **13:**2178-2189.
35. Van Dongen S: **Graph clustering by flow simulation.** In *PhD thesis Centre for Mathematics and Computer Science, Amsterdam*; 2000.
36. Enright AJ, Kunin V, Ouzounis CA: **Protein families and TRIBES in genome sequence space.** *Nucleic Acids Res* 2003, **31:**4632-4638.
37. Thompson JD, Plewniak F, Ripp R, Thierry JC, Poch O: **Towards a reliable objective function for multiple sequence alignments.** *J Mol Biol* 2001, **314:**937-951.
38. Perrière G, Combet C, Penel S, Blanchet C, Thioulouse J, Geourjon C, Grassot J, Charavay C, Gouy M, Duret L, Deléage G: **Integrated databanks access and sequence/structure analysis services at the PBIL.** *Nucleic Acids Res* 2003, **31:**3393-3399.
39. Benson DA, Karsch-Mizrachi I, Lipman DJ, Ostell J, Sayers EW: **GenBank.** *Nucleic Acids Res* 2009, **37:**D26-31.
40. Clamp M, Cuff J, Searle SM, Barton GJ: **The Jalview Java alignment editor.** *Bioinformatics* 2004, **20:**426-427.
41. Ruan J, Li H, Chen Z, Coghlan A, Coin LJ, Guo Y, Hériché JK, Hu Y, Kristiansen K, Li R, Liu T, Moses A, Qin J, Vang S, Vilella AJ, Ureta-Vidal A, Bolund L, Wang J, Durbin R: **TreeFam: 2008 Update.** *Nucleic Acid Res* 2008, **36:**D735-740.
42. Zmasek CM, Eddy SR: **ATV: display and manipulation of annotated phylogenetic trees.** *Bioinformatics* 2001, **17:**383-384.
43. Arigon AM, Perrière G, Gouy M: **HoSeqI: automated homologous sequence identification in gene family databases.** *Bioinformatics* 2006, **22:**1786-1787.
44. Ihaka R, Gentleman R: **R: A language for data analysis and graphics.** *J Comp Graph Stat* 1996, **5:**299-314.
45. Charif D, Lobry JR: **SeqinR 1.0–2: a contributed package to the R project for statistical computing devoted to biological sequences retrieval and analysis.** In *Structural Approaches to Sequence Evolution: Molecules, Networks, Populations* Edited by: *Bastolla U, Porto M, Roman HE, Vendruscolo M. New York: Springer Verlag*; 2007:207-232.
46. Tatusov RL, Fedorova ND, Jackson JD, Jacobs AR, Kiryutin B, Koonin EV, Krylov DM, Mazumder R, Mekhedov SL, Nikolskaya AN, Rao BS, Smirnov S, Sverdlov AV, Vasudevan S, Wolf YI, Yin JJ, Natale DA: **The COG database: an updated version includes eukaryotes.** *BMC Bioinformatics* 2003, **4:**41.
47. Chen F, Mackey AJ, Stoeckert CJ Jr, Roos DS: **OrthoMCL-DB: querying a comprehensive multi-species collection of ortholog groups.** *Nucleic Acids Res* 2006, **34:**D363-368.
48. Berglund AC, Sjolund E, Ostlund G, Sonnhammer EL: **InParanoid 6: eukaryotic ortholog clusters with inparalogs.** *Nucleic Acids Res* 2008, **36:**D263-266.

49. Koski LB, Morton RA, Golding GB: **Codon bias and base composition are poor indicators of horizontally transferred genes.** *Mol Biol Evol* 2001, **18:**404-412.
50. Beiko RG, Hamilton N: **Phylogenetic identification of lateral genetic transfer events.** *BMC Evol Biol* 2006, **6:**15.
51. Galtier N: **A model of horizontal gene transfer and the bacterial phylogeny problem.** *Syst Biol* 2007, **56:**633-642.
52. Beiko RG, Ragan MA: **Detecting lateral genetic transfer: a phylogenetic approach.** *Methods Mol Biol* 2008, **452:**457-469.
53. 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* 1994, **22:**4673-4680.
54. Felsenstein J: **PHYLIP – Phylogeny inference package (Version 3.2).** *Cladistics* 1989, **5:**164-166.

Unexpected genomic degeneration in the highly abundant free-living marine cyanobacteria *Prochlorococcus*

Gabriel A. B. Marais*, Caroline Dufaure de Citres, Anne-Sophie Sertier, Vincent Daubin

Université Lyon 1; CNRS; UMR5558; Laboratoire de Biométrie et Biologie évolutive; Villeurbanne, F-69622 cedex, France

* *Author for Correspondence*: Gabriel Marais, Laboratoire de Biométrie et Biologie évolutive, Université Lyon 1, Villeurbanne, France, tel: (+33) (0) 4 72 43 29 09, fax: (+33) (0) 4 72 43 13 88, e-mail address: Gabriel.Marais@univ-lyon1.fr

Abstract

The genomes of bacterial endosymbionts such as *Buchnera aphidicola* living in Aphids are smaller than their free-living close relatives. This genome reduction is mainly explained by Muller's ratchet, a degenerative process affecting small asexual populations. Genome reduction has been recently reported in the marine cyanobacteria *Prochlorococcus* but here it is thought to be an adaptation to life in nutrient-poor oceanic surface waters in which a small genome could be advantageous. Using nine *Prochlorococcus* genomes (some of them reduced) and two *Synecochoccus* genomes (as outgroups), we show that the reduced *Prochlorococcus* genomes show signs of degeneration not adaptation: inefficient selection on codon usage, reduced selection at protein level and preferential loss of genes under low selective pressure. This mirrors what is found in bacterial endosymbionts and raises the possibility of Muller's ratchet operating in *Prochlorococcus*, a highly abundant free-living bacteria. This and other hypotheses to explain our very unexpected findings are discussed. This work has implications for understanding the long-standing question of why genome size varies among organisms.

Introduction

In bacteria, small, AT-rich, fast-evolving genomes were considered the hallmark of the endosymbiotic lifestyle (reviewed in Moran 2002, Wernegreen 2002, Moya *et al*. 2008, Moran *et al*. 2008) until it was unexpectedly found in two free-living oceanic bacteria: first in *Prochlorococcus marinus* (Dufresne *et al*. 2003; Rocap *et al*. 2003; Dufresne *et al*. 2005) and then in *Pelagibacter ubique* (Giovannoni *et al*. 2005). In the cyanobacteria *P. marinus*, some strains (hereafter called "reduced") have undergone a 30% genome reduction, a 40% increase in A+T content and a two to four-fold acceleration in DNA evolution compared to the other strains (hereafter called "normal") and using *Synechococcus sp*. as outgroup (Dufresne *et al*. 2005, Kettler *et al*. 2007, and see Figure 1). This observation does call for an explanation since these strains are in fact highly differentiated species (>30% sequence divergence) that diverged about 80 millions years ago (Dufresne *et al*. 2005) and the situation here is very different from closely related bacterial strains such as *E. coli* strains (<3% sequence divergence) that show polymorphism in genome size (Welch *et al*. 2002, Touchon *et al*. 2009).

In bacterial endosymbionts such as *Buchnera aphidicola*, many observations and theoretical work point towards Muller's ratchet as a strong driver of genome evolution (Wernegreen & Moran 1999, Rispe & Moran 2000, Abbot & Moran 2002, Van Ham *et al*. 2003, Delmotte *et al*. 2006, Pettersson & Berg 2007). Muller's ratchet is a well-known degenerative process that affects small asexual populations such as that of *Buchnera*. Reduced polymorphism and inefficient selection at different levels (codon usage, proteins) are expected under Muller's ratchet and have been found in *Buchnera* (Wernegreen & Moran 1999, Abbot & Moran 2002, Van Ham *et al*. 2003). Massive gene loss is expected under Muller's ratchet especially genes under weak selective pressure, which has been shown again in *Buchnera* (Delmotte *et al*. 2006). These include some DNA repair genes and their loss explains why the genome tends to accumulate GC to AT mutations and deletions, the most common mutations in bacteria (Mira *et al*. 2001, Rocha & Danchin 2002, Moran *et al*. 2009, Hershberg & Petrov 2010, Hildebrand *et al*. 2010). Host-specialization also contributes to gene loss since some genes become useless in a host (Moya *et al*. 2008).

In *P. marinus,* genome shrinkage is considered an adaption to nutrient-poor surface waters where the first strains with reduced genomes were found (Dufresne *et al*. 2005). The

idea is that small genomes are advantageous in such environments because DNA replication is less costly (especially in nitrogen and phosphorus, two limiting elements in the surface waters). However, when more *P. marinus* strains became sequenced it appeared that the relationship between genome size and ocean depth was fairly loose since some reduced strains (NATL1A, NATL2A, SS120) can be found in as deep waters as normal strains (Kettler *et al*. 2007, Partensky & Garczarek 2010 and see Figure 1). Although recent data suggest the NATL strains can live both in deep and surface waters, SS120 has been shown to be restricted to deep waters (Partensky & Garczarek 2010). The hypothesis of "adaptation by streamlining" also predicts that proteins should be shorter because of selective pressure to reduce energy and nutrient consumption related to DNA, RNA and protein synthesis and this prediction is not supported by the data (Marais *et al*. 2008). Moreover, there is no compelling evidence that reduced strains got rid of all their junk DNA, which is again not in agreement with the "adaptation by streamlining" hypothesis (see Figure 1). Finally, the idea itself of reducing genome size to grow better is not supported by the data since there is no correlation between genome size and growth rate in bacteria (Viera-Silva *et al*. 2009).

An important question is whether reduced *P. marinus* show genomic degeneration as the bacterial endosymbionts do? To address this question, the classical approach is to study the efficiency of purifying selection, using dN/dS ratio and look for evidence of reduced selection at protein level. A recent work has shown that dN/dS is lower in *P. marinus* than in *Synechococcus sp*., which suggests Muller's ratchet is not operating in *P. marinus* (Hu & Blanchard 2009). However, we lack a direct comparison between the reduced and normal *P. marinus* strains. Here we studied eleven cyanobacterial genomes including nine *P. marinus* strains and two *Synechococcus sp*. and looked for evidence of genomic degeneration in three different ways. First, we identified the optimal codons for translation in *Synechococcus sp*. and *P. marinus* genomes, and found compelling evidence that selection on codon usage is no longer efficient in the reduced strains. Second, we show that the differences in dN/dS observed between reduced and normal strains are strongly impacted by codon usage bias. When we consider only genes with low codon usage bias to get reliable estimates of dN/dS, we found a higher dN/dS in reduced strains than that in normal strains. We show that this higher dN/dS ratio is due to reduced purifying selection at protein level using appropriate tests. Last, we studied the genes lost in the reduced strains and show that these genes have a higher dN/dS than the "core" genes shared by all *P. marinus* strains and *Synechococcus sp*., which suggests that genes with low selective constraints have been preferentially lost.

Very surprisingly, this pattern mirrors what has been observed in bacterial

endosymbionts and is consistent with Muller's ratchet (Wernegreen & Moran 1999, Van Ham *et al*. 2003, Delmotte *et al*. 2006). A combination of reduced sex and high mutation rate (and maybe reduced *Ne*) could explain why Muller's ratchet is affecting *P. mar*inus strains with small genomes. Other Hill-Roberston effects (clonal interference, ruby-in-the-rubbish, genetic draft) might also contribute to the pattern of genomic degeneration that we observe. Genome reduction could be driven by niche-specialization and loss of unnecessary genes. Genome-wide degeneration at synonymous and amino acid sites is more difficult to explain under this scenario though. Mutation pressure due to loss of DNA repair genes could explain degeneration in reduced *P. marinus*. We discuss these possible explanations and the implications for the evolution of genome size in bacteria.

Material and methods

Identification of optimal codons and computation of codon usage bias

All the coding sequences were retrieved from the Hogenom database (see below), classified as ribosomal proteins and others, and analysed using a multivariate method called within-group factorial correspondence analysis (WCA), using a R script and the seqinR package (Charif *et al*. 2005). WCA is used to find the codons preferentially used in the highly expressed genes (supposedly under the strongest selection on codon usage, here ribosomal proteins) compared to other genes. It is robust to biases in amino acid composition and is the best available method to identify optimal codons (Perriere & Thioulouse 2002, Suzuki *et al*. 2008). The optimal codons were those preferentially used by the ribosomal proteins and were identified using the two axes projection of the WCA results. The frequency of optimal codons (*Fop*) was estimated with a C program that we ran on coding sequences.

Identification of orthologous genes

We started with the Hogenom database, a gene family database including most of the prokaryotic and eukaryotic complete genomes (release 04, Feb 21 2008, including 511 fully sequenced genomes, Penel *et al*. 2009). We searched the database for one to one orthologous genes shared by 9 *Prochlorococcus marinus* strains (MIT9312, MIT9313, NATL2A, SS120, MED4, MIT9515, NATL1A, MIT9601, MIT9303) and 2 *Synechococcus sp*. strains (CC9605, CC9311) using a program called TreePattern (Dufayard *et al*. 2005) included in the FamFetch graphical interface (http://pbil.univ-lyon1.fr/software/famfetch.html). TreePattern screens all

the gene families looking for a particular tree motif. The tree motif that we used here was the published phylogeny of *Prochlorococcus marinus* with *Synechococcus sp*. as outgroups (Kettler *et al*. 2007). We excluded all genes with duplication within the tree motif. We retrieved 51 one to one orthologous genes.

Identification of core/lost genes

Core genes: The set of core genes is the same as the 51 orthologous genes described above and for which we conserved only normal *P. marinus* strains (MIT9313, MIT9303) and both *Synechococcus sp*. (CC9605, CC9311) for comparison with lost genes (see below).

Lost genes: We first established a list of genes lost in all reduced *P. marinus* strains using a Bayesian framework. This analysis uses the published phylogenetic tree of *Prochlorococcus marinus* with *Synechococcus sp*. as outgroups (Kettler *et al*. 2007). Each node of the phylogenetic tree was associated with two possible states: either presence or absence of the Hogenom gene families including some of our species. Evolution of each family was modelled using a Bayesian network (Bru 2005, Jensen 2001), with conditional probabilities for gain and loss associated to each edge and non-conditional probabilities associated to the states of the root of the tree. Bayesian tree model was implemented using the Bayesian Network Toolbox for MATLAB (Murphy 2001). Conditional and non-conditional probabilities were estimated by the EM algorithm (Dempster *et al*. 1977) so as to maximize the likelihood of the observed occurrences over the entire set of 5,901 gene families. The most probable evolutionary scenario was inferred for each gene family using the junction tree algorithm (Jensen 2001, Dempster *et al*. 1977). Timing of gene family gains and losses were inferred from these scenarios. We then selected genes with both normal strains and *Synechococcus sp*. plus other cyanobacterial outgroups to get reliable orthologous genes using a TreePattern search. We retrieved a first set of 37 lost genes. We excluded genes with duplication events and got 17 remaining genes. Only results with the 17 genes are presented.

Alignment and phylogeny

In all cases, coding sequences were extracted using a list of accessions provided by FamFetch for all the selected genes that we entered directly in the alignment graphical interface Seaview using the option “import from database” on Hogenom (Gouy *et al*. 2010). Alignment of each set of orthologous coding sequences was performed using the graphical interface Seaview with the Muscle program (Gouy *et al*. 2010). Alignments were clean manually removing non-conserved regions at the 5’ and 3’ ends. Alignments were

concatenated in a super-alignment for further analysis using the concatenation module in Seaview. Phylogenetic trees were obtained using PhyML in Seaview with the following parameters: GTR+γ model, estimated base frequencies, estimated proportion of invariable sites, 4 substitution rate categories, estimated gamma distribution parameter. All the trees fully agreed with the published species tree.

dN/dS analysis

We used the codeml program from the PAML 4.1 package (Yang 2007) to estimate dN/dS, which was run on the alignment and the PhyML tree for each dataset (see previous section). We performed several analyses:

Branch model: codeml with parameter (model = 2, NSsites = 0) was used to estimate global dN/dS ratio in various lineages. The statistical significance of the differences in dN/dS among lineages was assessed with a likelihood ratio test (LRT) using nested models (Yang 2007).

Site models: codeml with parameter (model = 0, NSsites = 2) was used to estimate the frequency of sites in three categories (0<dN/dS< 1, dN/dS=1, dN/dS >1) and the dN/dS value for sites evolving under purifying selection and positive selection (model M2a). We also used FMutSel and FMutSel0 to check for the effect of selection on codon usage on the results (Yang & Nielsen 2008). Statistical significance was assessed comparing models M2a and model M1a (model= 0, NSsites = 1; only 2 site categories) and also FMutSel/FMutSel0, and FMutSel with different number of site categories with a LRT.

Branch-site model: codeml with parameter (model = 2, NSsites = 2) was used to identify fast-evolving codons and estimate their mean dN/dS ratio in a particular lineage. When the dN/dS ratio was > 1, we tested whether it was significantly different from 1 using a LRT (comparing models with estimated dN/dS and dN/dS=1).

Clade model: codeml with parameter (model = 3, NSsites = 2) was used to identify all codons evolving differently in a given lineage compared to other lineages. The dN/dS and the number of sites in this category were estimated and statistical significance was assessed comparing the clade model with M1a (model= 0, NSsites = 1), a model without a category of sites evolving differently among lineages using a LRT as recommended in the PAML manual.

Results

Inefficient selection on codon usage in the "reduced" *P. marinus* strains

In bacteria, selection on codon usage is widespread (Sharp *et al*. 2010). We identified the optimal codons in our set of species using within-group correspondence analysis (WCA) on their complete genomes retrieved from the Hogenom database (see Material and Methods). We found very similar sets of optimal codons in our *Synechococcus sp*. strains and in the two normal *P. marinus* strains (see legend of Figure 2 for the list of optimal codons). In the reduced *P. marinus* strains, the WCA yielded to no optimal codons since the set of highly expressed genes (e.g. the ribosomal proteins) were found to use the same codons as the other genes. All this suggests that selection on codon usage has affected both *Synechococcus* and *Prochlorococcus* lineages and has stopped in the reduced *P. marinus* strains. The optimal codons are all GC-ending codons and are consistent with the GC-content in our set of species: with normal strains being GC-rich and reduced strains being AT-rich (see Figure 1). In Figure 2, we show the comparison of the frequency of optimal codons (Fop) for orthologous genes in one reduced strain (MED4) and one normal strain (MIT9313). This clearly shows that Fop is low and homogeneous in the reduced strain and is fully in agreement with inefficient selection on codon usage in this strain.

Reduced selection on proteins in the "reduced" *P. marinus* strains

To study differences in dN/dS in *P. marinus*, we used high quality orthologous gene sets identified with phylogenetic tools (Dufayard *et al*. 2005, see Material and Methods). Table 1 shows the results of a dN/dS analysis of different lineages in the tree from Figure 1 using codeml (branch-model analysis, see Material and Methods). We found that dN/dS is significantly higher in normal strains and *Synechococcus sp*. than that in reduced strains, which is in agreement with previous results (Hu & Blanchard 2009). However, we showed in the previous section that *P. marinus* strains display major differences in codon usage bias. Selection on codon usage can make the interpretation of dN/dS difficult because dS can hardly be considered neutral and differences between dN/dS among strains can reflect differences in selection on codon usage and not in selection on proteins (reviewed in Yang 2002). Selection on codon usage varies among genes within a genome and in our set of orthologous genes the Fop values are quite heterogeneous. The effect of selection on codon usage on dN/dS estimates is expected to be the strongest for genes under high selection on codon usage in the normal strains and the outgroups. We thus computed the mean of Fop values in the normal strains and the outgroups and found that highly biased genes show a higher dN/dS in normal versus reduced strains, whereas lowly biased genes show the opposite

pattern (see Figure 3).

We then ranked genes according to this value as low codon usage bias (50% of the genes with the lowest *Fop* values, i.e. *Fop* ≤ 0.49), very low codon usage (25 % of the genes with the lowest *Fop* values, i.e. *Fop* ≤ 0.525), and also high codon usage bias (50% of the genes with the highest *Fop* values, i.e. *Fop* ≥ 0.525), very high codon usage (25 % of the genes with the highest *Fop* values, i.e. *Fop* ≥ 0.57). In Table 1, we show that the pattern of dN/dS variations among strains strongly depends on the set of genes considered, which confirms the trend observed in Figure 3. We also found that the very lowly biased genes, the genes that are less likely to be biased by differences in selection on codon usage among lineages, show a significantly higher dN/dS in reduced strains than that in normal ones. This suggests reduced selection at protein level in reduced strains. However, because we estimated global dN/dS ratios, these results could also be explained by a higher proportion of sites under positive selection in reduced strains. To test this hypothesis, we used the branch-site and clade model analyses in codeml (see Material and Methods). The branch-site analysis identifies codons under positive selection in a particular lineage. Table 2 shows that genes with very low codon usage bias (the best set of genes to perform unbiased dN/dS analysis) show no evidence for positive selection in reduced strains, which confirms the hypothesis of reduced selection in these strains. The clade model analysis identifies fast-evolving codons in various lineages and estimates their mean dN/dS. Applied to genes with very low codon usage bias, the clade model analysis revealed a slight increase in dN/dS in reduced strains compared to normal ones (see Table 2). The same analyses conducted on all the orthologous genes, did not detect any positive selection (see Table 2).

In all the above analysis, we considered separately the ancestral branch and other branches in the reduced lineage (see Tables 1 and 2). A preliminary dN/dS analysis showed that the ancestral branch had a very high dS value (x60) compared to the other branches whereas its dN value was roughly conform to molecular clock. This intriguing pattern results in a very low dN/dS value for the ancestral branch compared to other branches in the reduced lineage. Table 1 shows that the dN/dS of the ancestral branch tends to increase when codon usage bias decreases as expected if inefficient selection on codon usage explained the very high dS value for this branch. However, even for genes with very low codon usage bias, dN/dS of the ancestral branch is still a lot lower than in the rest of reduced lineage (see Table 1). This is probably because the switch from efficient to inefficient selection on codon usage occurred on this branch (where other dramatic changes such as massive gene loss has been

observed, Kettler *et al*. 2007, AS Sertier, unpublished data) and implies a "out of equilibrium" evolution at the synonymous sites. Such mode of evolution is expected to increase to number of substitutions before the new equilibrium is reached and to yield to meaningless dN/dS values (i.e. not relevant for diagnosing selection), which probably explains why the ancestral branch has peculiar dN/dS values.

Preferential loss of genes under low selective pressure in the reduced *P. marinus* strains

To study what kinds of genes were lost in the reduced lineage, we prepared two sets of genes (see Material and Methods): a set of genes shared by all the *P. marinus* strains and outgroups (core genes) and another set of genes present in normal strains and outgroups and lost in all the reduced strains (lost genes). Lost genes have been identified using a Bayesian approach inferring events of gene loss and gain in the *Prochlorococcus* phylogeny and excluding gene duplications to focus on strictly orthologous sequences (see Material and Methods). To make a fair comparison between core and lost genes, we included in the analysis the four same species (i.e. *Synechococcus sp*. and the normal *P. marinus* strains). We then studied the intensity and form of selection affecting both gene sets using codeml (see Material and Methods). We first estimated the global dN/dS and found that lost genes have a higher dN/dS than core genes, which suggests purifying selection is stronger on core genes than on lost genes (see Table 3). We also ran a site model analysis -a codeml analysis identifying codons with different selection regimes- and found that the fraction of site under purifying selection is bigger in the core genes than in the lost genes, with no identified sites under positive selection (see Table 3). This suggests that the genes lost have lower selective constraints than the genes that persisted in the reduced strains.

Importantly, these results are not affected by selection on codon usage. Codeml includes a pair of site models FMutSel/FMutSel0 that can be used to test whether dN/dS estimates are affected by selection on codon usage (Yang & Nielsen 2008). Compared to a classical site model FMutSel has extra parameters for selection on codon usage (selection coefficients for each synonymous codon). FMutsel0 is nested in FMutSel (selection coefficients are set to 0) and is used to test whether selection on codon usage is significant with a likelihood ratio test. We found no evidence for positive selection in both core and lost genes by comparing FMutSel with and without a site category under positive selection (see Table 3). We did find that FMutSel is significantly better than FMutSel0, which further confirms selection on codon usage in the data but the results (dN/dS and proportion of sites evolving on purifying selection or neutral evolution) were unchanged using FMutSel or the standard site model (see Table 3).

Discussion

Genomic degeneration in free-living and endosymbiotic bacteria

The patterns of inefficient selection on codon usage, reduced selection at protein level and loss of genes under weak selective pressure in the reduced *P. marinus* are indicative of a global reduction of the efficacy of selection on these strains. This seems to have affected differently codon usage and protein evolution, which is expected. Indeed, selection on codon usage is of weak intensity (Hartl et al. 1994), and a global reduction of the efficacy of selection will affect it strongly. Selection operating on proteins is overall much stronger than selection on codon usage and a global reduction of the efficacy of selection will probably affect a small fraction of the amino acids (under weak selection). Only a slight reduction of selection on proteins is expected, which is fully in agreement with the small differences in *dN/dS* observed between reduced and normal strains. The patterns of genomic degeneration that we found in *Prochlorococcus* are very similar to those that have been found in bacterial endosymbionts (Wernegreen & Moran 1999, Van Ham et al. 2003, Delmotte et al. 2006). However, the extent of genomic degeneration is much larger in endosymbionts. In *Buchnera*, comparisons with *E. coli* showed that about 80% of the genes originally present in the ancestror of *Buchnera* have been lost (Moran & Mira 2001), whereas in *Prochlorococcus*, this figure is probably lower than 30% (Dufresne et al. 2005, Kettler et al. 2007). *Buchnera* and many other obligate endosymbiont bacteria have much smaller genomes (with about 200-600 genes, Moran et al. 2008) than reduced *P. marinus* strains do (with >1500 genes, Dufresne et al. 2005, Kettler et al. 2007). In *Buchnera*, 3' ends of genes have degenerated due to small deletions and AT substitutions, and genes are smaller in *Buchnera* than in free-living relatives (Charles et al. 1999). Such pattern was not found in *Prochlorococcus* where genes have very similar size in both normal and reduced strains (Marais et al. 2008). *dN/dS* values are also much more increased in *Buchnera* (Wernegreen & Moran 1999) compared to *Procholorococcus* where we found only a slight (but significant) increase. Genomic degeneration is probably much weaker in reduced *P. marinus* than in bacterial endosymbionts, but we still need to explain why we observe degeneration in bacteria that have a very different lifestyle from endosymbiosis.

Codon usage bias and unreliable *dN/dS* estimates

dN/dS analysis is one of the most common analysis in molecular evolution and programs estimating *dN/dS* such as codeml are heavily used and cited (codeml has >3000 citations in Pubmed). An important assumption underlying the interpretation of *dN/dS* ratio is that the synonymous sites evolve neutrally and dS is merely affected by mutation rate. There has been a debate on how violation of this assumption might affect the reliability of the *dN/dS* estimates (for reviews, see Yang & Bielawski 2000, Yang 2002). Constant selection on codon usage among lineages is probably not a problem as argued by Yang and colleagues but differences in the level of selection on codon usage among lineages may lead to unreliable estimates as we show here (see Table 1 and Figure 3). Recent developments have introduced selection on codon usage in codon-based models used for *dN/dS* analysis (Nielsen et al. 2007, Yang & Nielsen 2008). However, this makes the models much more complicated and the *dN/dS* estimation much more computationally-demanding. Only site-model analysis (with selection on codon usage) is currently included in codeml (see Tables 2 and 3), popular analyses such as branch-model and branch-site do have this option and caution is needed when working on sequence data from species differing in the level of selection on codon usage.

Causes of degeneration in *Prochlorococcus*: Muller's ratchet or other Hill-Robertson effects?

The pattern of degeneration that we found in *Prochlorococcus* is reminiscent of what has been found in bacterial endosymbionts, and this raises the possibility that Muller's ratchet is also operating in *Prochlorococcus*. It has been shown theoretically that the ratchet affects small asexual populations, the rate of deleterious mutations being a key parameter (Gordo & Charlesworth 2000). *Prochlorococcus* are highly abundant bacteria: chlorophyll from *Prochlorococcus* can be seen from space by ocean-monitoring satellite and cell concentrations as high as 10^5 cells /ml have been observed (Johnson et al. 2006, Partensky & Garczarek 2010). However, very little is known about the population genetics of *P. marinus* strains and we have currently no estimates for the effective population size (*Ne*) - the relevant parameter for Muller's ratchet - for the different *P. marinus* strains. Another issue is whether *P. marinus* strains are asexual, another condition for Muller's ratchet to work. Some cases of horizontal gene transfer have been documented in *Prochlorococcus*; for instance the *hli* genes are found only in the high-light (HL) *P. marinus* strains (all reduced) and are from phage origin (Lindell et al. 2004). Some regions of the genome –called genomic islands – seem to be hot spots for horizontal gene transfer in reduced *P. marinus* strains (Coleman et al. 2006).

Recent studies of horizontal gene transfer in *Prochlorococcus* and *Synechococcus* have found a substantial amount of those within *Prochlorococcus*, between *Prochlorococcus* and *Synechococcus* and even more distantly related group (one transfer from a gamma-proteobacteria has been documented), which suggests *P. marinus* strains are not asexual (Zhaxybayeva et al. 2006, Luque et al. 2008, Dufresne et al. 2008, Zhaxybayeva et al. 2009). However, the horizontal gene transfers are much more frequent in the normal strains than in the reduced strains (Zhaxybayeva et al. 2009). In the reduced strains from the HL ecotype, horizontal gene transfers are much rarer than in the low light (LL) one, and the GC content of the transferred genes suggest that they are ancient events (Zhaxybayeva et al. 2009). There seems to be a correlation between the level of sex and genome size in *Prochlorococcus*. Another key issue is the rate of deleterious mutations, which is currently not known in *Prochlorococcus*. However, DNA sequence analysis suggests that reduced strains have a higher mutation rate than the normal strains, which is consistent with the loss of several DNA repair genes in the reduced strains (Dufresne et al. 2005, Kettler et al. 2007). It is possible that the reduced strains have small *Ne*, low or no recombination and a high deleterious mutation rate, which would make Muller's ratchet operating in these strains. However, estimates for these parameters are clearly needed to give further support to this hypothesis.

Muller's ratchet is one of the degenerative processes known as Hill-Robertson effects (for review, see Gordo & Charlesworth 2001). Other such processes might explain the genomic degeneration in *Procholorococcus*. Clonal interference for instance happens when beneficial mutations on different genetic backgrounds compete with one another in the absence of sex (Gerrish & Lenski 1998). Clonal interference is expected to affect large asexual bacterial populations, which might be the case of the reduced *P. marinus* strains. However, if reduced strains probably have high mutation rate, clonal interference may not be strong since genomes combining all beneficial mutations can be created by mutation and recombination is not needed (Bollback & Huelsenbeck 2007). The ruby-in-the-rubbish model where mild beneficial mutations arise on genetic backgrounds with strong deleterious mutations and are eliminated could also apply to *Prochlorococcus* since this model requires high mutation rate (Orr 2000). *Prochlorococcus* may have huge *Ne* and recurrent episodes of adaption could sweep the polymorphism and reduce the efficacy of selection as in the genetic draft model developed by Gillespie (2000, 2001). Unfortunately, critical parameters (*Ne*, level of recombination, mutation rate) are missing to discuss further the possible contribution of all these processes to genomic degeneration in *Prochlorococcus*.

Causes of degeneration in *Prochlorococcus*: byproduct of niche-specialization?

The adaptation to surface waters by "genome streamlining" was initially suggested as the explanation for small genomes in some *P. marinus* strains (Dufresne et al. 2005). However, the correlation between depth and genome size is weak in *P. marinus* strains as reduced strains can be found close to the surface (HL ecotype) or deeper (LL ecotype). However, the ecology of *P. marinus* strains is poorly known and reduced strains might occupy a specific niche. Specialization can lead to loss of unnecessary genes. Part of gene loss in bacterial endosymbionts is indeed explained by host-specialization (Moya et al. 2008). If the environment of the reduced *P. marinus* strains is very specific and very stable, which is possible (Partensky & Garczarek 2010), then niche-specialization could explain the gene loss and genome reduction that we observe. It is difficult to anticipate which genes are expected to be lost without knowing much about this niche-specialization. However, both high and low selective pressure genes could become unnecessary and lost, and niche-specialization does not necessarily predicts a higher *dN/dS* value for the lost genes than that of the core genes as we found. In addition to gene loss, we also have to explain inefficient selection on codon usage and reduced selection on proteins in reduced strains, which are not straightforward expectations from the niche-specialization hypothesis. In bacteria, selection on codon usage is related to growth rate, with species with low selection on codon usage being slow growers and species with high selection on codon usage being fast growers (Viera-Silva et al. 2010). Reduced *P. marinus* strains may have become slow-grower while adapting to a new niche where slow growth rate may be an advantage. Reduced selection on proteins that we observe is much more difficult to reconcile with the niche-specialization hypothesis. Loss of DNA repair genes and increase of mutation rate may have cause degeneration of proteins (by Muller's ratchet or increase of mutation pressure, see next section). In this case, degeneration would be a byproduct of niche-specialization. The loss of DNA repair genes when adapting to a new niche is well-known in bacteria. It has been shown that loosing such genes and increasing mutation rate can be advantageous for a bacterial population undergoing a change in the environment (and the need to adapt) since it will increase the rate of beneficial mutations (Taddei et al. 1997, Tenaillon et al. 1999, Sniegowski et al. 2007). But once the adaptive event has passed, selection will favour bacteria with reduced rate of deleterious mutations and the mutation is expected to decrease, which is actually observed in nature (Denamur et al. 2000). It is not clear at all why the reduced strains would benefit never re-gaining the DNA repair genes and maintaining a high mutation rate for millions of years of evolution.

Causes of degeneration in *Prochlorococcus*: increase in mutation pressure?

The reduced *P. marinus* strains evolve fast at DNA level, which suggests a high mutation rate (Dufresne et al. 2005). This is clearly not due to UV exposure since some of the reduced strains live at the same depth (and with similar UV exposure) compared to normal strains (SS120 and NATL strains, Ketller et al. 2007, Partensky & Garczarek 2010). Moreover, it has been shown that there is no deficit of pyrimidine dinucleotides (the targets of UVs) in the reduced MED4 strain, a HL strain with high UV exposure (Palmeira et al. 2006). The reduced strains all lack several DNA repair genes and this may explain fast evolution and high mutation rate (Dufresne et al. 2005, Kettler et al. 2007, Marais et al. 2008). Using population genetics and other theoretical frameworks, it has been shown that mutation rate can strongly affect the evolution of genome size (Knibbe et al. 2007, Lynch 2007). Lynch has shown that nonfunctional DNA (e.g introns, transposable elements, duplicate genes) accumulates in species with small *Ne* due to genetic drift (reviewed in Lynch 2007). This nonfunctional DNA not only makes the genome bigger but also increases the mutational targets in the genome (e.g. mutations in introns can disrupt the coding regions of a gene). This makes the mutation rate a critical parameter for the accumulation of such DNA: when the mutation rate is very high, deleterious mutations in nonfunctional DNA will be very frequent and selection to purge nonfunctional DNA out of the population will be efficient. The mitochondria are a clear example of the effect of mutation on genome size. Whereas plant and animal mitochondria are supposed to have similar Ne, they strongly differ in mutation rate the animal mitochondria having a much higher mutation rate than that of the plant ones. Consistent with Lynch's hypothesis, animal mitochondrial genomes are much smaller than plant mitochondrial genomes, which include a large fraction of intergenic DNA (Lynch et al. 2006, Lynch 2006). Using an evolution *in silico* approach, it has been shown that the fraction of non-coding DNA in a genome is correlated with the mutation rate (especially the rate of rearrangements) in a bacteria-like system (Knibbe et al. 2007). We have also shown that an increase in mutation rate could also affect the number of genes in a genome (Marais et al. 2008). In very large populations, the efficacy of selection will depend on the ratio s/u (s=selection coefficient and u=mutation rate) and an increase in u will make selection on genes with small s inefficient. A modest increase in mutation rate (x10) can result in the loss of 30% of the genes (Marais et al. 2008). The effect of mutation rate on other genomic features (protein evolution, codon usage) has not been modeled. But following Marais et al. (2008), an increase in mutation rate should reduce both selection on proteins (especially at

codons with small effect on fitness, i.e. small *s*) and selection on codon usage. However, further work is needed to check this.

The question of the loss of DNA repair genes in the reduced *P. marinus* lineage is a puzzling one. It has been shown that loosing DNA repair genes and increasing mutation rate can be advantageous for a bacterial population undergoing a change in the environment (and the need to adapt) since it will increase the rate of beneficial mutations (Taddei et al. 1997, Tenaillon et al. 1999, Sniegowski et al. 2007). But once the adaptive event has passed, selection will favour bacteria with reduced rate of deleterious mutations and the mutation is expected to decrease, which is actually observed in nature (Denamur et al. 2000). Why the reduced strains have maintained such a high rate of mutation for so long is puzzling. One possibility is that re-gaining DNA repair genes may be difficult for those bacteria. However, it seems that horizontal gene transfer has repeatedly occurred in the evolutionary history of the reduced *P. marinus* strains (Lindell et al. 2004, Coleman et al. 2006). Another possibility is an ever-changing environment with an ever-changing need to adapt and maintain a high mutation rate. Interestingly, it has been shown that such evolutionary dynamics can select very high mutation rates that may be disadvantageous in the long run and may lead to extinction in some circumstances (Gerrish et al. 2007). A possibility for such an ever-changing environment is a very fast evolutionary arm race between phage and bacteria, which we know would favour high mutation rate (Tenaillon et al. 1999). However, we have no indication phage pressure is particularly strong on reduced *P. marinus* strains.

Implications for the evolution of genome size in bacteria

It has been proposed that genetic drift is the main determinant of genome size in bacteria (reviewed in Lynch 2006, 2007). A recent analysis in bacteria has shown that this is supported by the data (Kuo et al. 2009). Within bacteria, genetic drift strongly affects genome size: low *Ne* results in loosing genes, accumulating deletions (the most common indels in bacteria) and getting a small genome. This relationship still holds when endosymbionts are removed from the dataset suggesting this is not explained only by this peculiar lifestyle. In this analysis, *P. marinus* is clearly an outlier (Kuo et al. 2009). Genome reduction in this cyanobacteria does not seem to follow the general trend and is probably not associated to genetic drift only. An unusually high mutation rate may explain genome reduction in *Prochlorococcus*. Further investigation is needed to confirm association between genome reduction and high mutation rate and also to identify the precise mechanism underlying genome reduction in this bacteria (Muller's ratchet, mutation pressure, other).

Acknowledgements

We thank Olivier Tenaillon, Sylvain Glémin, Caroline Knibbe, Frédéric Partensky, Laurence Garczarek and Daniel Kahn for stimulating discussions and helpful comments on an earlier version of this manuscript. GABM is supported by a grant from Agence Nationale de la Recherche (ANR-08-JCJC-0109).

References

Abbot P, Moran NA. Extremely low levels of genetic polymorphism in endosymbionts (Buchnera) of aphids (Pemphigus). Mol Ecol. 2002 Dec;11(12):2649-60.

Bollback JP, Huelsenbeck JP. Clonal interference is alleviated by high mutation rates in large populations. Mol Biol Evol. 2007 Jun;24(6):1397-406.

Bru C. Analyse évolutive des familles de domaines protéiques. PhD thesis, Paul Sabatier University, 2005.

Charif D, Thioulouse J, Lobry JR, Perrière G. Online synonymous codon usage analyses with the ade4 and seqinR packages. Bioinformatics. 2005 Feb 15;21(4):545-7.

Charles H, Mouchiroud D, Lobry J, Gonçalves I, Rahbe Y. Gene size reduction in the bacterial aphid endosymbiont, Buchnera. Mol Biol Evol. 1999 Dec;16(12):1820-2.

Coleman ML, Sullivan MB, Martiny AC, Steglich C, Barry K, Delong EF, Chisholm SW. Genomic islands and the ecology and evolution of *Prochlorococcus*. Science. 2006 Mar 24;311(5768):1768-70.

Delmotte F, Rispe C, Schaber J, Silva FJ, Moya A. Tempo and mode of early gene loss in endosymbiotic bacteria from insects. BMC Evol Biol. 2006 Jul 18;6:56.

Dempster A. P., Laird, N. M. and Rubin D. B.. Maximum likelihood from incomplete data via the em algorithm. J Roy Stat Soc, 39 :1–38, 1977.

Denamur E, Lecointre G, Darlu P, Tenaillon O, Acquaviva C, Sayada C, Sunjevaric I, Rothstein R, Elion J, Taddei F, Radman M, Matic I. Evolutionary implications of the frequent horizontal transfer of mismatch repair genes. Cell. 2000 Nov 22;103(5):711-21.

Dufayard JF, Duret L, Penel S, Gouy M, Rechenmann F, Perrière G. Tree pattern matching in phylogenetic trees: automatic search for orthologs or paralogs in homologous gene sequence databases. Bioinformatics. 2005 Jun 1;21(11):2596-603.

Dufresne A, Garczarek L, Partensky F, *et al*. 2005. Accelerated evolution associ¬ated with genome reduction in a free-living prokaryote. Genome Biol 6:R14

Dufresne A, Salanoubat M, Partensky F, *et al*. 2003. Genome sequence of the cyanobacterium *Prochlorococcus* marinus SS120, a nearly minimal oxyphototrophic genome. Proc Natl Acad Sci USA, 100:10020-10025

Dufresne A, Ostrowski M, Scanlan DJ, Garczarek L, Mazard S, Palenik BP, Paulsen IT, de Marsac NT, Wincker P, Dossat C, Ferriera S, Johnson J, Post AF, Hess WR, Partensky F. Unraveling the genomic mosaic of a ubiquitous genus of marine cyanobacteria. Genome Biol. 2008;9(5):R90.

Gerrish PJ, Lenski RE. The fate of competing beneficial mutations in an asexual population. Genetica. 1998;102-103(1-6):127-44.

Gillespie JH. Genetic drift in an infinite population. The pseudohitchhiking model. Genetics. 2000 Jun;155(2):909-19.

Gillespie JH. Is the population size of a species relevant to its evolution? Evolution. 2001 Nov 11;55(11):2161-9.

Giovannoni SJ, Tripp HJ, Givan S, Podar M, Vergin KL, Baptista D, Bibbs L, Eads J, Richardson TH, Noordewier M, Rappé MS, Short JM, Carrington JC, Mathur EJ. Genome streamlining in a cosmopolitan oceanic bacterium. Science. 2005 Aug 19;309(5738):1242-5.

Gordo I, Charlesworth B. The degeneration of asexual haploid populations and the speed of Muller's ratchet. Genetics. 2000 Mar;154(3):1379-87.

Gordo I, Charlesworth B. Genetic linkage and molecular evolution. Curr Biol. 2001 Sep 4;11(17):R684-6.

Gouy M, Guindon S, Gascuel O. SeaView version 4: A multiplatform graphical user interface for sequence alignment and phylogenetic tree building. Mol Biol Evol. 2010 Feb;27(2):221-4

Hartl DL, Moriyama EN, Sawyer SA. Selection intensity for codon bias. Genetics. 1994 Sep;138(1):227-34.

Hershberg R, Petrov DA. Evidence That Mutation Is Universally Biased towards AT in Bacteria. PLoS Genet. 2010 Sep 9;6(9).

Hu J, Blanchard JL. Environmental sequence data from the Sargasso Sea reveal that the characteristics of genome reduction in *Prochlorococcus* are not a harbinger for an escalation in genetic drift. Mol Biol Evol. 2009 Jan;26(1):5-13.

Jensen FV. Bayesian network and decision graphs. Springer, 1st edition, September 2001.

Johnson ZI, Zinser ER, Coe A, McNulty NP, Woodward EM, Chisholm SW. Niche partitioning among *Prochlorococcus* ecotypes along ocean-scale environmental gradients. Science. 2006 Mar 24;311(5768):1737-40.

Hildebrand F, Meyer A, Eyre-Walker A. Evidence of selection upon genomic GC-content in bacteria. PLoS Genet. 2010 Sep 9;6(9). pii: e1001107.

Kettler G, Martiny A, Chisholm S. *et al*. 2007. Patterns and implications of gene gain and loss in the evolution of *Prochlorococcus*. Plos Genetics volume 3 issue 12 e231

Knibbe C, Coulon A, Mazet O, Fayard JM, Beslon G. A long-term evolutionary pressure on the amount of

noncoding DNA. Mol Biol Evol. 2007 Oct;24(10):2344-53.

Kuo CH, Moran NA, Ochman H. The consequences of genetic drift for bacterial genome complexity. Genome Res. 2009 Aug;19(8):1450-4.

Lindell D, Sullivan MB, Johnson ZI, Tolonen AC, Rohwer F, Chisholm SW. Transfer of photosynthesis genes to and from *Prochlorococcus* viruses. Proc Natl Acad Sci U S A. 2004 Jul 27;101(30):11013-8.

Luque I, Riera-Alberola ML, Andujar A, Ochoa de Alda JAG. 2008. Intraphylum diversity and complex evolution of cyanobacterial aminoacyl-tRNA synthetases. Mol Biol Evol. 25:2369–2389.

Lynch M, Koskella B, Schaack S. Mutation pressure and the evolution of organelle genomic architecture. Science. 2006 Mar 24;311(5768):1727-30

Lynch M. Streamlining and simplification of microbial genome architecture. Annu Rev Microbiol. 2006;60:327-49.

Lynch M. The Origins of Genome Architecture. Editeur : Sinauer Associates Inc.,U.S.; Édition : 1 (1 mars 2007)

Marais GA, Calteau A, Tenaillon O. 2008. Mutation rate and genome reduction in endosymbiotic and free-living bacteria. Genetica 134(2):205-10

Mira A, Ochman H, Moran NA. Deletional bias and the evolution of bacterial genomes. Trends Genet. 2001 Oct;17(10):589-96.

Moran NA, McCutcheon JP, Nakabachi A. Genomics and evolution of heritable bacterial symbionts. Annu Rev Genet. 2008;42:165-90.

Moran NA, McLaughlin HJ, Sorek R. The dynamics and time scale of ongoing genomic erosion in symbiotic bacteria. Science. 2009 Jan 16;323(5912):379-82.

Moran NA, Mira A. The process of genome shrinkage in the obligate symbiont Buchnera aphidicola. Genome Biol. 2001;2(12):RESEARCH0054.

Moran NA. 2002. Microbial minimalism: genome reduction in bacterial pathogens. Cell 108:583-586

Moya A, Peretó J, Gil R, Latorre A. Learning how to live together: genomic insights into prokaryote-animal symbioses. Nat Rev Genet. 2008 Mar;9(3):218-29.

Murphy KP. The Bayes Net Toolbox for Matlab. Comput. Sci. Stat., 33 :2001, s2001.

Orr HA. The rate of adaptation in asexuals. Genetics. 2000 Jun;155(2):961-8.

Palmeira L, Guéguen L, Lobry JR. UV-targeted dinucleotides are not depleted in light-exposed prokaryotic genomes. Mol Biol Evol. 2006 Nov;23(11):2214-9.

Partensky F, Garczarek L (2010) *Prochlorococcus*: Advantages and Limits of Minimalism. Annual Review of Marine Science 2: 305-331.

Penel S, Arigon AM, Dufayard JF, Sertier AS, Daubin V, Duret L, Gouy M, Perrière G. Databases of homologous gene families for comparative genomics. BMC Bioinformatics. 2009 Jun 16;10 Suppl 6:S3.

Perrière G, Thioulouse J. Use and misuse of correspondence analysis in codon usage studies. Nucleic Acids Res.

2002 Oct 15;30(20):4548-55.

Pettersson ME, Berg OG. Muller's ratchet in symbiont populations. Genetica. 2007 Jun;130(2):199-211.

Rispe,C. & Moran,N.A. 2000. Accumulation of deleterious mutations in endosymbionts: Muller's ratchet with two levels of selection. The American Naturalist 156, 425-411.

Rocap G, Larimer FW, Lamerdin J, Malfatti S, Chain P, *et al*. 2003. Ge¬nome divergence in two *Prochlorococcus* ecotypes reflects oceanic niche differentiation. Nature, 424:1042-1047

Rocha EP, Danchin A. Base composition bias might result from competition for metabolic resources. Trends Genet. 2002 Jun;18(6):291-4.

Sharp PM, Emery LR, Zeng K. Forces that influence the evolution of codon bias. Philos Trans R Soc Lond B Biol Sci. 2010 Apr 27;365(1544):1203-12

Sniegowski PD, Gerrish PJ, Lenski RE. 2007. Evolution of high mutation rates in experimental populations of E.coli. Nature 387:703-705

Suzuki H, Brown CJ, Forney LJ, Top EM. Comparison of correspondence analysis methods for synonymous codon usage in bacteria. DNA Res. 2008 Dec;15(6):357-65.

Taddei F, Radman M, Maynard-Smith J, Toupance B, Gouyon PH, Godelle B. Role of mutator alleles in adaptive evolution. Nature. 1997 Jun 12;387(6634):700-2.

Tenaillon O, Toupance B, Le Nagard H *et al*. 1999. Mutators, population size, adaptive landscape the adaptation of asexual populations of bacteria. Ge¬netics 97:485-493

Touchon M, Hoede C, Tenaillon O, Barbe V, Baeriswyl S, Bidet P, Bingen E, Bonacorsi S, Bouchier C, Bouvet O, Calteau A, Chiapello H, Clermont O, Cruveiller S, Danchin A, Diard M, Dossat C, Karoui ME, Frapy E, Garry L, Ghigo JM, Gilles AM, Johnson J, Le Bouguénec C, Lescat M, Mangenot S, Martinez-Jéhanne V, Matic I, Nassif X, Oztas S, Petit MA, Pichon C, Rouy Z, Ruf CS, Schneider D, Tourret J, Vacherie B, Vallenet D, Médigue C, Rocha EP, Denamur E. Organised genome dynamics in the Escherichia coli species results in highly diverse adaptive paths. PLoS Genet. 2009 Jan;5(1):e1000344.

Van Ham RC, Kamerbeek J, Palacios C, Raussel C, *et al*. 2003. Reductive genome evolution in Buchnera aphidicola. Proc Natl Acad Sci USA 100:581-586

Vieira-Silva S, Touchon M, Rocha EP. No evidence for elemental-based streamlining of prokaryotic genomes. Trends Ecol Evol. 2010 Jun;25(6):319-20

Welch RA, Burland V, Plunkett G 3rd, Redford P, Roesch P, Rasko D, Buckles EL, Liou SR, Boutin A, Hackett J, Stroud D, Mayhew GF, Rose DJ, Zhou S, Schwartz DC, Perna NT, Mobley HL, Donnenberg MS, Blattner FR. Extensive mosaic structure revealed by the complete genome sequence of uropathogenic Escherichia coli. Proc Natl Acad Sci U S A. 2002 Dec 24;99(26):17020-4.

Wernegreen JJ, Moran NA. Evidence for genetic drift in endosymbionts (Buchnera): analyses of protein-coding genes. Mol Biol Evol. 1999 Jan;16(1):83-97.

Wernegreen JJ. Genome evolution in bacterial endosymbionts of insects. Nat Rev Genet. 2002 Nov;3(11):850-

61.

Yang Z, Bielawski JP. 2000. Statistical methods for detecting molecular adaptation. Trends Ecol Evol. 15:496–503.

Yang Z, Nielsen R. Mutation-selection models of codon substitution and their use to estimate selective strengths on codon usage. Mol Biol Evol. 2008 Mar;25(3):568-79.

Yang Z. Inference of selection from multiple species alignments. Curr Opin Genet Dev. 2002 Dec;12(6):688-94. Review.

Yang, Z. 2007. PAML 4: a program package for phylogenetic analysis by maximum likelihood. Molecular Biology and Evolution 24: 1586-1591

Zhaxybayeva O, Doolittle WF, Papke RT, Gogarten JP. Intertwined evolutionary histories of marine Synechococcus and *Prochlorococcus* marinus. Genome Biol Evol. 2009 Sep 2;1:325-39.

Zhaxybayeva O, Gogarten JP, Charlebois RL, Doolittle WF, Papke RT. Phylogenetic analyses of cyanobacterial genomes: quantification of horizontal gene transfer events. Genome Res. 2006 Sep;16(9):1099-108.

Tables

Table 1: Comparison of *dN/dS* ratios in normal versus reduced *P. marinus* strains taking codon usage bias into account.

Dataset	**Branch model 4 ratios**	**LRT (statistical test for differences between ω1 and ω2)**
All genes (n=51)	ω0= 0.1051 ω1= 0.1081 ω2= 0.0630 ω3= 0.0009	$p < 10^{-4}$
Low codon usage bias (n=26)	ω0= 0.0866 ω1= 0.0771 ω2= 0.0633 ω3= 0.0009	p=0.0121
Very low codon usage bias (n=13)	ω0= 0.0857 ω1= 0.0640 ω2= 0.0792 ω3= 0.0011	p=0.0483
High codon usage bias (n=26)	ω0= 0.1029 ω1= 0.1281 ω2= 0.0605 ω3= 0.0008	$p < 10^{-4}$
Very high codon usage bias (n=13)	ω0= 0.0961 ω1= 0.1340 ω2= 0.0559 ω3= 0.0009	$p < 10^{-4}$

ω= *dN/dS*. 0: outgroups, 1: normal *P. marinus* strains, 2: reduced *P. marinus* strains, 3: ancestral branch. LRT = likelihood ratio tests. Simpler models with just 1, 2 or 3 ratios are significantly less likely than a model with 4 ratios as presented here. The comparison between a model with 4 ratios (ω1≠ω2) and a model with 3 ratios (ω1=ω2) to test whether reduced and normal strains have different *dN/dS* is shown in the last column. All analyses were done on combined data.

Table 2: *dN*/*dS* analysis to distinguish positive selection and reduced selection in the reduced *P. marinus* strains

Dataset*	**Analysis**	**Results**	**LRT**
All genes	Branch-site (test for positive selection in 1)	P=6% ω= 1.19666	p=0.17, not significant No positive selection
All genes	Clade model	P'= 37% ω0= 0.26186 ω1= 0.25022 ω2= 0.12240 ω3= 0.03385	$p < 10^{-4}$ Significant improvement of purifying selection in 2 versus 1
Very low codon usage bias	Branch-site (test for positive selection in 2)	P=6% ω= 1	Not necessary No positive selection
Very low codon usage bias	Clade model	P'=45% ω0= 0.18109 ω1= 0.12548 ω2= 0.12858 ω3= 0.03891	$p < 10^{-4}$ Significant reduction of purifying selection in 2 versus 1

ω= *dN*/*dS*. 0: outgroups, 1: normal *P. marinus* strains, 2: reduced *P. marinus* strains, 3: ancestral branch. LRT = likelihood ratio tests. All analyses were done on combined data. P=fraction of codons under positive selection in the tested lineage, *dN*/*dS* ratio of these codons is provided. P'=fraction of fast-evolving codons in the tested lineages, *dN*/*dS* ratio of these codons for all tested lineages is provided. For the branch-site, LRT is comparing a model with estimated ω for positively selected sites with a model where ω is set to 1. For clade model, LRT is comparing a model with a category for codons evolving differently among species (with different *dN*/*dS* ratios) and a model with no such codons.

Table 3: Global and site-by-site *dN*/*dS* analysis of core genes versus lost genes

Analysis	**Core genes**	**Lost genes**
Branch Model (1 ratio)	ω=0.0495	ω=0.0825
Site Model (M1a)*	P0=0.84797, ω0=0.02894 P1=0.15203, ω1=1	P0=0.77335, ω0=0.04094 P1=0.22665, ω1=1
Site Model with selection on codon usage (FMutSel)**	P0=0.85024, ω0=0.03125 P1=0.14976, ω1=1	P0=0.77594, ω0=0.04111 P1=0.22406, ω1=1

ω= *dN*/*dS*. P=proportion. 0: sites under purifying selection, 1: sites evolving neutrally. LRT = likelihood ratio tests. All analyses were done on combined data. Very similar results were obtained for a larger set of lost genes (n=37, see Material and Methods).

* a LRT between M1a and M2a (site category with ω > 1) is not significant

** a LRT between FMutSel (2 site categories) and FMutSel (3 site categories including a site category with ω > 1) is not significant; a LRT between FMutSel and FMutSel0 (no selection on codon usage) is significant, $p < 10^{-4}$

Figures and legends

Figure 1: **Phylogenetic tree of *Prochlorococcus marinus* strains and statistics on their genomes**. This tree has been built using the 51 orthologous genes shared by all species identified in this study (see Material and Methods) with Seaview (Gouy *et al*. 2010) using PhyML with GTR+γ. Reduced *P. marinus* strains are in red, other *P. marinus* are in blue and *Synechococcus sp*. (outgroups) are in black. The numbers denote the groups studied in the *dN/dS* analysis. Data on genomes comes from NCBI and ecological data comes from Kettler *et al*. 2007. HL=High light zone (near the surface), LL=Low light zone (deeper waters).

Figure 2: **Comparison of codon usage bias between a reduced strain (MED4) and a normal strains (MIT9313)**. We used 51 orthologous genes in all *P. marinus* and our two *Synechococcus sp*. (see Material and Methods). Codon usage bias is measured by the frequency of optimal codons (*Fop*) in the coding sequences. The list of optimal codons have been obtained using a standard WCA analysis (see Material and Methods) and includes the following codons: AAC, AAG, ACC, ACG, AGC, ATC, CAC, CAG, CCC, CCG, CGC, CGG, CTC, CTG, GAC, GAG, GCC, GCG, GGC, GTG, TAC, TCC, TCG, TGC, TTC. All the optimal codons are GC-ending codons. The spearman non-parametric correlation coefficient is -0.27 and is not significantly different from 0 (p=0.98). The line Y=X is shown.

Figure 3: **Relationship between codon usage bias (*Fop*) and the ratio of *dN/dS* in reduced versus normal *P. marinus* strains (ω_r/ω_n)**. Both parameters have been estimated independently for all the 51 orthologous genes. The *Fop* values correspond to the mean of *Fop* in MIT9313, MIT9303 and both *Synechococcus sp*. strains. *dN/dS* ratio have been estimated by branch model analysis (two different ratios for *P. marinus* strains with reduced genomes and other *P. marinus* strains) with codeml (see Material and Methods). The spearman non-parametric correlation coefficient is -0.48 and is significantly different from 0 ($p < 10^{-3}$).

Figure 1

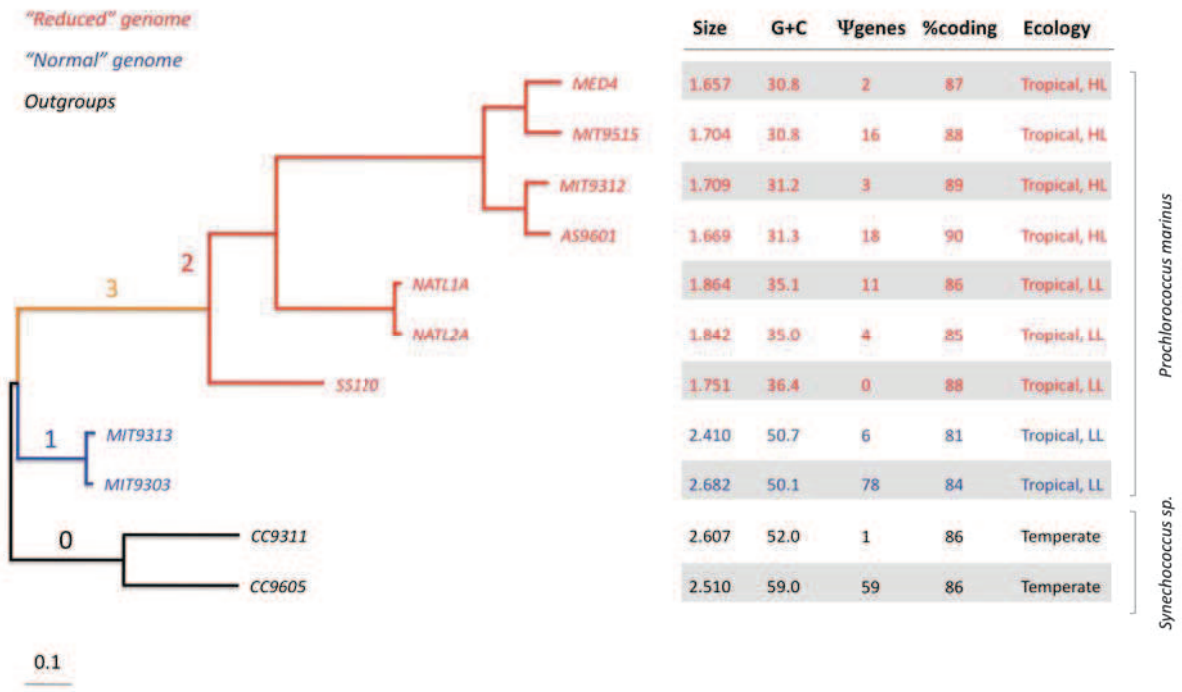

"Reduced" genome
"Normal" genome
Outgroups
MED4
MIT9515
MIT9312
AS9601
NATL1A
NATL2A
SS120
MIT9313
MIT9303
CC9311
CC9605
3
2
1
0
0.1
Size
G+C
Ψgenes
%coding
Ecology
1.657 30.8 2 87 Tropical, HL
1.704 30.8 16 88 Tropical, HL
1.709 31.2 3 89 Tropical, HL
1.669 31.3 18 90 Tropical, HL
1.864 35.1 11 86 Tropical, LL
1.842 35.0 4 85 Tropical, LL
1.751 36.4 0 88 Tropical, LL
2.410 50.7 6 81 Tropical, LL
2.682 50.1 78 84 Tropical, LL
2.607 52.0 1 86 Temperate
2.510 59.0 59 86 Temperate
Prochlorococcus marinus
Synechococcus sp.

Figure 2

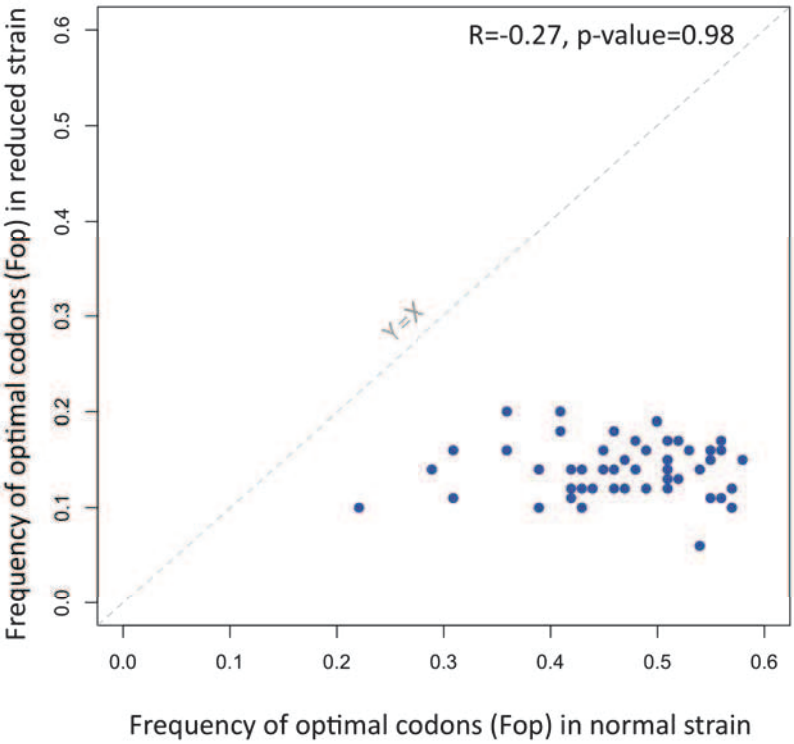
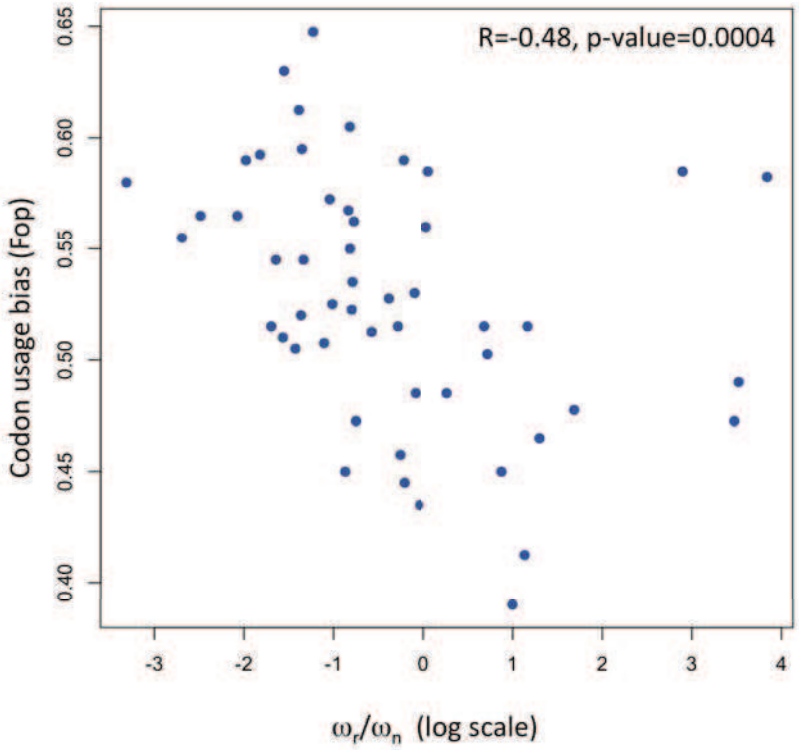

R=-0.27, p-value=0.98
Y=X
Frequency of optimal codons (Fop) in reduced strain
Frequency of optimal codons (Fop) in normal strain
0.0
0.1
0.2
0.3
0.4
0.5
0.6

Figure 3


R=-0.48, p-value=0.0004
Codon usage bias (Fop)
0.40
0.45
0.50
0.55
0.60
0.65
-3
-2
-1
0
1
2
3
4
ω_r/ω_n (log scale)

A phylogenetic view of the expanding protein universe

Anne-Sophie Sertier, Vincent Daubin and Daniel Kahn

Université de Lyon; Université Lyon 1; CNRS; INRIA; UMR5558; Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France.

One-sentence summary

Systematic inference of evolutionary scenarios for proteins and their constituent modules shows that protein module innovation, not domain shuffling, is the major drive for the tremendous diversification of proteins that has been experienced throughout the history of life.

Corresponding author:

Daniel Kahn

Laboratoire de Biométrie et Biologie Evolutive, UMR CNRS 5558

Université Lyon 1

43 boulevard du 11 novembre

69622 Villeurbanne Cedex, France

Phone : +33 (0)472 43 13 44

FAX : +33 (0)472 43 13 88

Email : kahn@biomserv.univ-lyon1.fr

ABSTRACT

The tremendous diversity of proteins is classically seen as resulting primarily from the combination of a few thousand primary domain types through domain shuffling. However most genomes also contain orphan genes with no resemblance to anything in the known protein universe, which points to a completely different source of protein diversity. In order to determine the importance of domain shuffling in protein innovation, we used Bayesian network analysis and derived evolutionary scenarios for both protein and domain families. We show that domain shuffling only explains a minority of protein innovations throughout the history of life. This points to an evolution of protein modules considerably more dynamic than previously envisioned, with a high turn-over of novel protein modules being the major drive of protein innovation.

The development of genomics has initiated the revision of several paradigms, from the definition of the gene *(1)* to the representation of life's genetic diversity *(2,3)*. As a mechanism generating protein novelty, the theory of domain shuffling has received tremendous support from comparative genomics *(4,5)*, and the immense diversity of proteins is classically seen as resulting mostly from combinatorial arrangements of a few thousand primary domain types *(6)*. However, most genomes contain significant numbers of orphan genes that cannot be explained by the shuffling of preexisting domains *(3,7)*. This raises the question whether the modular nature of proteins is the major source of their diversity. Here we use Bayesian network analysis to systematically derive evolutionary scenarios for proteins and their constituent modules in a full phylogenetic framework. This allowed us to characterize the emergence of novel proteins in terms of modular arrangements and novel protein modules. Contrary to common belief, we find that module shuffling is not the major process responsible for protein diversification. We explain why this was not apparent in previous studies that relied on manually curated protein domain family databases. We conclude that the protein domain repertoire is much more dynamic than hitherto envisioned and that protein domain innovation, not domain shuffling, has been the major drive of protein innovation throughout the history of life.

In the absence of structural information, modular proteins are better described as an assemblage of evolutionary modules than of structural domains. While comparative genomics has established the role of duplication, insertion, deletion, fusion and fission in the evolution of proteins *(4,8,9)*, the possibility that the repertoire of elementary modules has been diversifying throughout the evolution of life has received little consideration. Understanding the relative contributions of these processes in the expansion of the protein universe requires a proper phylogenetic framework. Based on a consensus phylogeny of 170 completely sequenced genomes spanning the three domains of life, we were able to independently reconstruct ancestral repertoires of proteins and evolutionary modules for every node of the phylogenetic tree using a Bayesian network model *(10-12)*. In contrast to maximum parsimony *(13,14)*, this method allows each branch of the tree to have its own probability of gene (or module) loss and gain, and therefore adequately accounts for processes of genome reduction and expansion that are known to vary widely *(12,15)*. Protein and module repertoires of extant organisms were derived respectively from gene families of the HOGENOM database *(16)* and domain families of a reclustered version of the ProDom database *(12,17)*. A decisive advantage of using these databases lies in their inherent comprehensive coverage, providing systematic access to all protein families or modules demonstrable from genomic sequences even when they have not yet been characterized. For each of these families, a scenario of gain, loss, and lateral gene transfer (LGT) was reconstructed and a most likely repertoire of genes and modules could be deduced for each node of the tree. The subsequent mapping of protein modules onto gene families established a correspondence between these scenarios, which could be readily interpreted in terms of processes of protein evolution. For each protein family originating at a node of the tree, we asked whether its component modules could be found elsewhere.

Protein modules were considered as innovated at this node if they could not be found in the parental node nor in other clades, which is conservative because it also excludes potentially novel modules that have been subsequently subject to LGT. Innovated proteins were classified as completely innovated when all its modules were innovated: both protein and its modules were new at this node. In contrast, when all modules were ancestrally present or found in other clades, the protein was considered as a rearrangement of preexisting modules, otherwise it was tagged as partly innovated.

The relative importance of these different types of innovation throughout evolution is shown in Fig. 1, where branch lengths represent numbers of events. Not surprisingly, the eukaryotic part of the tree presents rates of innovation typically an order of magnitude higher than the prokaryotic parts. This may be due to a combination of biological reasons *(18)* and over-annotation of potential protein-coding genes *(19)*. However our sample of eukaryotic species was too small to further investigate this question here. Therefore in what follows we specifically focus on prokaryotes that are much better sampled and for which gene prediction methods are robust. Overall, innovated proteins represented 21% of a species repertoire at a given node. Innovation appeared significantly higher in extant species (26%) than in ancestral species (11%), suggesting that a large fraction of novel genes observed in current genomes may have little evolutionary prospect. In extant species 12% to 83% of innovated proteins are orphans, *i.e.* they share no detectable homology with proteins from other genomes at the module level, even though a majority of these orphan genes is thought to be functional *(20)*. Although these general tendencies depend on the overall structure of the dataset, the relative proportions of different innovation mechanisms show a striking pattern: the proportion of innovated proteins composed of only new modules is much higher (56%) than those resulting from module shuffling only (20%). In other words orphan proteins dominate protein innovation, not module shuffling, which is conspicuous both for recent and ancestral protein innovations (Fig. 1C-D and Fig. S4).

A clear definition of what we mean by module innovation is required here to further interpret these results. We consider that appearance of a module should be seen as innovation when it results from a sudden evolutionary change. The formation of a coding sequence from non-coding DNA, as well as the sudden increase of evolutionary rate or the frame-shift of a previously coding gene are, according to this definition, genuine innovations. In contrast constitutively fast evolving sequences, for which homologs are difficult to detect, should not be seen as innovated. To measure the effect of fast evolving genes on our estimates of domain innovation, we compared, for different pairs of species, the evolutionary rates of modules innovated in their common ancestors with those of ancient module families that originated in the Last Universal Common Ancestor (LUCA). An example of the distributions obtained is plotted in Fig. 2. As expected, modules that originated at LUCA evolve more slowly than more recent modules *(21)*. However the distributions strongly overlap: over 78% of the modules originating at recent nodes evolve more slowly than the 5% fastest modules detected in LUCA. This proportion increases for modules originating at deeper nodes (Table S2). Therefore homologs to such modules

would be expected to be detected even after several billion years of evolution: although a fraction of innovated modules may be false positives, the majority of these innovation events cannot be explained by high constitutive evolutionary rates.

Throughout the evolution of life, protein modules appear to have originated continuously (Fig. 1B) and the vast majority of innovated proteins arose through generation of novel protein elements rather than through reassortment of preexisting modules. This view differs from previous studies suggesting that most of the protein universe can be described as combinations of a few thousands of distinct module types *(4,6)*. These analyses were based mainly on curated protein families derived from either sequence comparisons *(22,23)* or protein structures *(24,25)*. When restricting our dataset to domains represented in these databases, our approach also supports a predominant role of module shuffling in the origin of protein novelty (Fig. 3B-C). However, a significant fraction of both protein and domain diversity is then neglected. For instance the Pfam-A database covers 51% of the residues from 73% of all proteins in the UniProt database *(23)*. Therefore most frequently occurring protein modules are well represented in these curated databases, but half of the protein universe remains uncharacterized. This blind spot becomes even wider when considering protein families: 67% of all prokaryotic protein families have no match with Pfam-A (Fig. 3B). Therefore automated protein clustering methods such as implemented in ProDom are essential tools to illuminate the entire protein module repertoire.

A dynamic view of protein module innovation emerges when this extensive repertoire is taken into consideration: protein innovation is an active ongoing process, with an average of 11% of protein modules in extant prokaryotes having emerged very recently (Table 1). A significant percentage of these protein modules (3.3%) is used in at least two different modular arrangements showing that, although recent, they have been successfully recruited by other novel proteins. This percentage increases to 21% for more ancient modules and up to 69% for protein modules tracing back to LUCA, so that this versatility can be seen as an indicator of evolutionary success (see also Fig. S3). The fact that contemporary organisms contain many more novel proteins and protein modules than inferred in ancestral species (Figs. 1 and 3A) indicates a high turn-over of protein innovation. While long-term selection or mutational drift is expected to eliminate a majority of these innovations, the minority that withstands selection has a profound influence in reshaping protein repertoires. Indeed almost half of the protein modules in extant prokaryotes are ancestral, clade-specific innovations that appeared after LUCA (Table 1). We conclude that the protein module repertoire is much more dynamic than previously thought and that the novel protein modules generated by this dynamics is the major drive of protein innovation. We anticipate that these novel protein modules will exhibit a broad range of novel structural features, as is already becoming apparent from structural genomics projects *(26)*. A major challenge will now be to identify the genomic mechanisms responsible for the generation of this diversity.

REFERENCES

1. M. B. Gerstein et al., Genome Res. 17, 669-681 (2007).
2. J. C. Venter et al., Science 304, 66-74 (2004).
3. S. Yooseph et al., PLoS Biol. 5, e16 (2007).
4. A. D. Moore, A. K. Björklund, D. Ekman, E. Bornberg-Bauer, A. Elofsson, Trends Biochem. Sci. 33, 444-451 (2008).
5. M. Bashton, C. Chothia, Structure 15, 85-99 (2007).
6. C. Chothia, J. Gough, C. Vogel, S. A. Teichmann, Science 300, 1701-1703 (2003).
7. V. Daubin, H. Ochman, Curr. Opin. Genet. Dev. 14, 616-619 (2004).
8. S. Pasek, J. Risler, P. Brézellec, Bioinformatics 22, 1418-1423 (2006).
9. J. Weiner, F. Beaussart, E. Bornberg-Bauer, FEBS J. 273, 2037-2047 (2006).
10. C. Bru, Thesis, Toulouse Paul Sabatier University (2005).
11. F. Jensen, Bayesian network and decision graphs (Springer, ed. 1, 2001).
12. Methods are available as supporting material on Science Online.
13. J. H. Fong, L. Y. Geer, A. R. Panchenko, S. H. Bryant, J. Mol. Biol. 366, 307-315 (2007).
14. M. Wang, G. Caetano-Anollés, Structure 17, 66-78 (2009).
15. B. Boussau, E. O. Karlberg, A. C. Frank, B. Legault, S. G. E. Andersson, Proc. Natl. Acad. Sci. U.S.A. 101, 9722-9727 (2004).
16. S. Penel et al., BMC Bioinformatics , in press (2009).
17. C. Bru et al., Nucleic Acids Res. 33, D212-D215 (2005).
18. M. Lynch, J. S. Conery, Science 302, 1401-1404 (2003).
19. M. Clamp et al., Proc. Natl. Acad. Sci. U.S.A. 104, 19428-19433 (2007).
20. V. Daubin, H. Ochman, Genome Res. 14, 1036-1042 (2004).
21. M. M. Albà, J. Castresana, Mol. Biol. Evol. 22, 598-606 (2005).
22. S. Hunter et al., Nucleic Acids Res. 37, D211-215 (2009).
23. R. D. Finn et al., Nucleic Acids Res. 36, D281-288 (2008).
24. A. Andreeva et al., Nucleic Acids Res. 36, D419-425 (2008).
25. A. L. Cuff et al., Nucleic Acids Res. 37, D310-314 (2009).
26. A. E. Todd, R. L. Marsden, J. M. Thornton, C. A. Orengo, J. Mol. Biol. 348, 1235-1260 (2005).

27. We would like to thank Catherine Ngom-Bru for initiating the use of Bayesian networks for ProDom and Laurent Duret and Simon Penel for stimulating discussions. Parallel Bayesian network computations were performed on the IN2P3 computer center in Lyon (http://cc.in2p3.fr). This work was supported in part by the France-Israel program in Bioinformatics, grants LHSG-CT-2004-512092 and INFRA-2007-213037 from the European Union and ANR-08-EMER-011-03 *Phylariane* from the Agence Nationale de la Recherche.

Table 1. Extent of protein innovation in prokaryotic organisms. Frequencies are expressed as percentages of protein families or module families found in extant organisms that are most recent (associated with only one leaf of the species tree), intermediary (innovated at an internal node) or trace back to LUCA. Frequencies are averaged over 161 prokaryotic species. Full distributions are available in Fig. S2.

	Most recent	Intermediary	Ancient (LUCA)
Protein families	26%	50%	24%
Protein modules	11%	48%	41%

Figure legends

Figure 1. Pervasive innovation of proteins and protein modules throughout phylogeny. In these phylogenetic trees, branch lengths reflect numbers of innovated families of different types: (A) protein families from HOGENOM *(16)*; (B) protein module families from reclustered ProDom *(17)*; (C) entirely novel proteins; (D) proteins that arose by shuffling of preexisting protein modules; (E) partly innovated proteins combining novel and preexisting modules. Color codes in (A) and (B) correspond to the following clades (anticlockwise): Archaea, yellow; Eukaryota, mauve; Aquifex aeolicus, pink; Deinococcus, purple; Thermotoga maritima, green; Spirochaetales, light blue; Rhodopirellula baltica, ochre; Firmicutes, light green; Bacteroidetes, brown; Fusobacterium nucleatum, red; Chlamydiales, orange; Actinobacteria, blue; Cyanobacteria, dark green; Proteobacteria, magenta. Eukaryotic parts of these phylogenetic trees were downscaled by the indicated factor for legibility. In this representation the tree has been rooted at the middle of the bacterial branch. Scale bar corresponds to 500 families.

Figure 2. Distribution of similarity scores of recent *vs* ancient protein modules. Similarity scores were calculated from alignments of orthologous protein modules in *Escherichia coli* and *Salmonella typhi* (see Table S2 for other pairs of species). The distribution for relatively recent modules inferred to have originated in their last common ancestor is shown in blue, while that for ancient modules tracing back to LUCA is shown in red. The dashed line corresponds to the first 5% quantile of ancient module similarity scores. The strong overlap between these distributions shows that the majority of module innovations cannot be explained by a constitutive fast rate of sequence divergence. Therefore they correspond to *bona fide* protein module innovations.

Figure 3. Distribution of protein innovation in ancestral and contemporary species. Completely innovated, partly innovated and module shuffled proteins are indicated in red, blue and green respectively on the basis of (A) reclustered ProDom, (B) Pfam-A and (C) SCOP family homology (per species distributions are available in Fig. S4). Protein families from HOGENOM not containing any module from the corresponding database are indicated in gray. In each panel the left and right bars correspond to ancestral and contemporary prokaryotic species, respectively, as symbolized below each bar.

Fig. 1. Pervasive innovation of proteins and protein modules throughout phylogeny

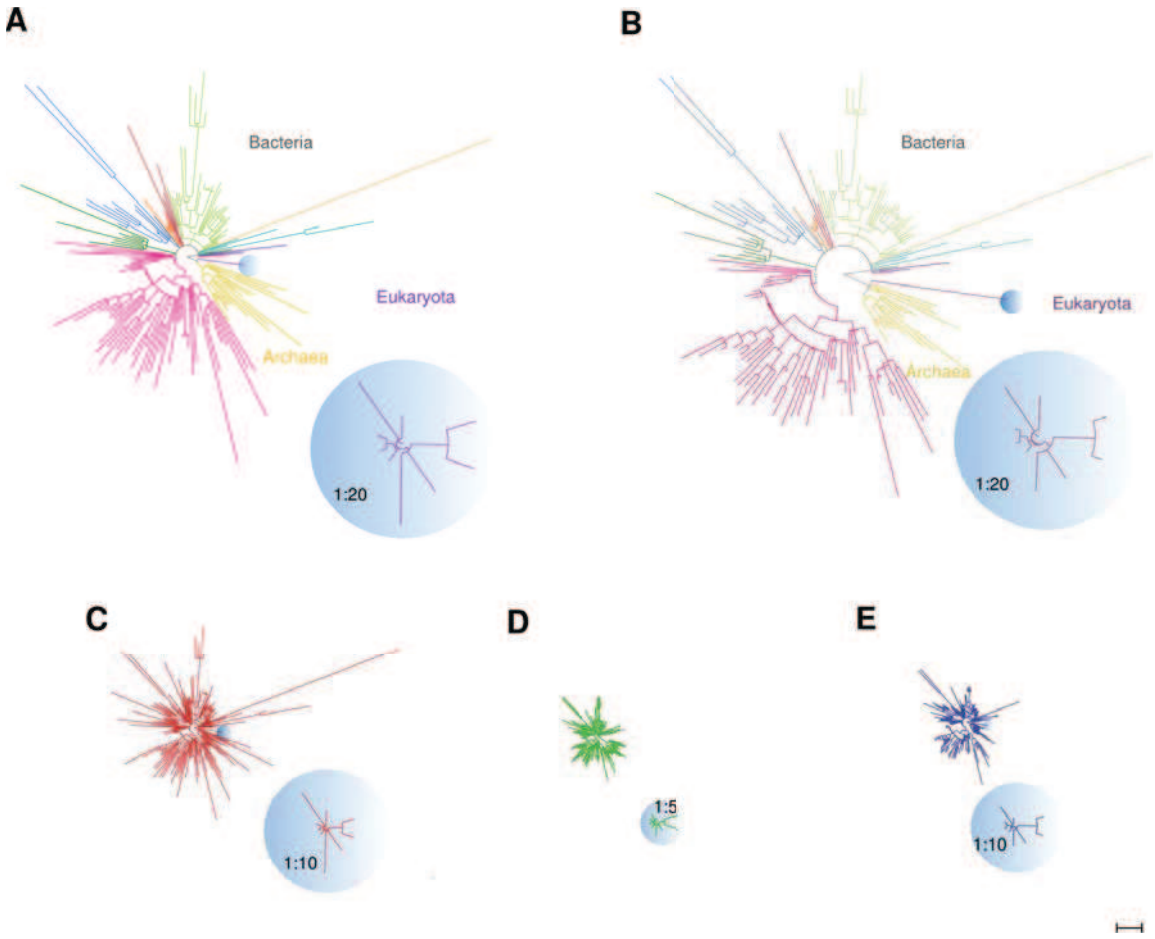


Fig. 2. Distribution of similarity scores of recent *vs* ancient protein modules

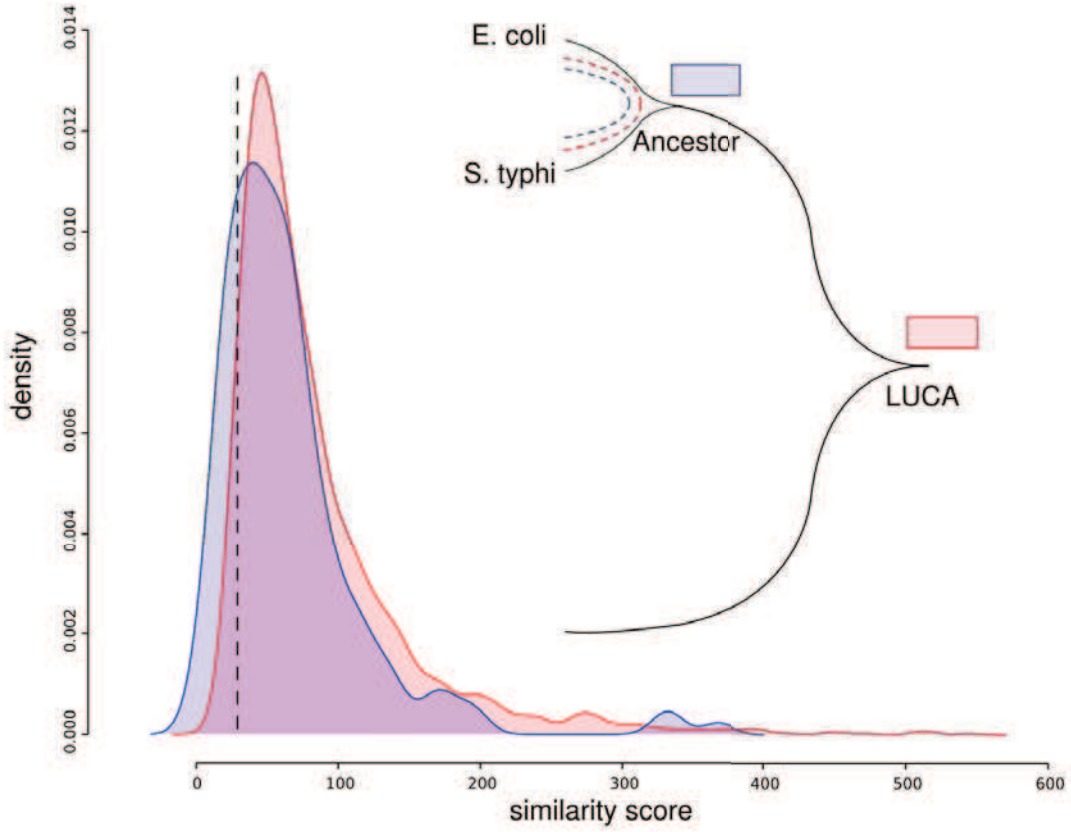
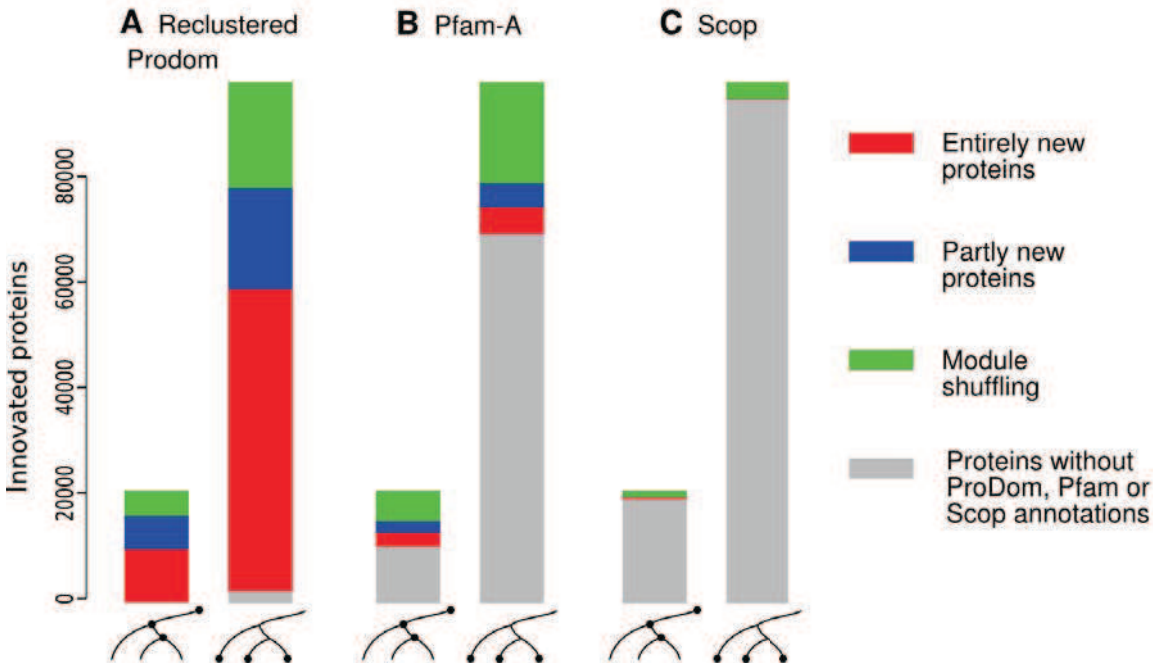


Fig. 3. Distribution of protein innovation in ancestral and contemporary species



Supporting online material for

A phylogenetic view of the expanding protein universe

Anne-Sophie Sertier, Vincent Daubin, Daniel Kahn

Université de Lyon; Université Lyon 1; CNRS; INRIA; UMR5558; Laboratoire de Biométrie et Biologie Evolutive, F-69622 Villeurbanne, France.

Contents

METHODS

Source of protein and domain families

We used HOGENOM *(1)* release 3 and the complete genome section of ProDom *(2)* release 2005.1 to define protein and module families, respectively. The overlap of the two databases defined a set of 170 fully sequenced genomes spanning the three kingdoms of life (9 Eukaryota, 19 Archaea and 142 Bacteria; Table S1). In HOGENOM full-length proteins are grouped by single linkage clustering on the basis of extensive protein similarity covering at least 80% of protein length, so that distinct modular arrangements cluster into different families. In contrast, ProDom comprises elementary protein modules defined through a procedure of local similarity search and clustering of homologous protein regions. The stringent PSI-BLAST 10^{-6} E-value cutoff that is used in the construction of ProDom tends to subdivide families into subfamilies, which was not desirable for the present work. We therefore used a Markov clustering procedure *(3)* to recluster ProDom module families on the basis of a sensitive PSI-BLAST search (E-value=10^{-2}; sequence span above 80%). ProDom clusters were mapped onto HOGENOM families by PSI-BAST at the same E-value threshold. In case two matches overlapped by more than 20 aminoacids, priority was given to the ProDom cluster with the highest number of domains, resulting in a linear arrangement of ProDom clusters characteristic of each protein family from HOGENOM.

Inference of evolutionary scenarios

The phylogenetic tree relating the 170 genomes was taken from the NCBI taxonomy *(4)* with minor modifications (Fig. S1). Phylogenetic profiles of protein families and protein modules were derived from the corresponding species inventories from HOGENOM and ProDom clusters, respectively. Each node of the phylogenetic tree was associated with two possible states – either presence or absence of a protein or protein module in the corresponding ancestral species. Evolution of each family was modeled *(5)* using a Bayesian network *(6)* based on this tree, with conditional probabilities for gain and loss associated to each edge and non-conditional probabilities associated to the states of the root LUCA. Bayesian tree models for protein and protein module families were implemented using the Bayesian Network Toolbox *(7)* for MATLAB (The MathWorks, Inc.). Conditional and non-conditional probabilities were estimated by the EM algorithm *(8)* so as to maximize the likelihood of the observed occurrences over a set of 10,000 randomly drawn families from HOGENOM and ProDom clusters, respectively. The resulting Bayesian tree models allowed us to infer the most probable evolutionary scenario for each protein family and protein module family using the junction tree algorithm *(6,7)*, predicting its pattern of occurrence in ancestral species. Conversely the most probable protein and protein module repertoires were inferred for each ancestral species as the unions of all proteins and protein modules respectively, predicted to be present from the previous analysis (Figs S5-S13).

This Bayesian network methodology appears to be a useful generalization of classical

parsimony. Indeed parsimony analysis uses a uniform value for gain and loss penalties, which corresponds to a special configuration of the Bayesian network with uniform gain and loss probabilities. Note however that applying parsimony results in empirical frequencies of gene gain and loss that vary widely *(9)*, which contradicts the use of uniform probabilities. We used the Likelihood Ratio Test *(10)* to verify whether the much larger number of parameters in the full Bayesian network is justified by the data, a set of 10,000 randomly drawn protein families. Let L_0 and L_1 be the maximum likelihoods obtained with the parsimony model and the full Bayesian network model, respectively, and k the difference between the numbers of parameters in these nested models. Under the null hypothesis of the parsimony model, the test statistic $-2\ln\frac{L_0}{L_1}$ will asymptotically follow a χ^2 distribution with k degrees of freedom ($k = 578$). The null hypothesis was unambiguously rejected by this test when applied to protein family data ($\chi^2 = 19{,}462$; P-value ≈ 0), which justifies the use of the full Bayesian network model.

Detection of recent *vs.* ancient protein modules

Similarity scores were calculated for orthologous protein modules found in pairs of closely related species: *Methanosarcina mazei vs. Methanosarcina acetivorans, Mycobacterium avium vs. Mycobacterium tuberculosis, Bacillus cereus* ATCC10987 *vs. Bacillus anthracis, Escherichia coli* K12 *vs. Salmonella typhi* ATCC700931. Orthologous module pairs were extracted from orthologous proteins as detected with TreePattern *(11)*. The Jones-Taylor-Thornton (*JTT)* distance *(12)* was calculated with ProtDist *(13)*. Similarity scores were taken as the expected numbers of identical residues $l.\exp(-JTT)$ in each pairwise alignment of length l. The distributions of these scores were compared between relatively recent modules and ancient modules tracing back to LUCA and were always found to overlap strongly (Table S2). In all cases of module innovation reported, we verified the absence of detectable homologs in other clades, so that protein modules were conservatively not considered as innovated in case of LGT.

REFERENCES

1. S. Penel *et al.*, *BMC Bioinformatics***,** in press (2009).
2. C. Bru *et al.*, *Nucleic Acids Res.* **33,** D212-D215 (2005).
3. S. Van Dongen, *SIAM J Matrix Anal Appl* **30,** 121-141 (2008).
4. D. L. Wheeler *et al.*, *Nucleic Acids Res.* **32,** D35-40 (2004).
5. C. Bru, Thesis, Paul Sabatier University (2005).
6. F. Jensen, *Bayesian network and decision graphs* (Springer, ed. 1, 2001).
7. K. Murphy, *Comput. Sci. Stat.* **33,** 331-350 (2001).
8. A. P. Dempster, N. M. Laird, D. B. Rubin, *J Roy Stat Soc* **39,** 1-38 (1977).
9. B. Boussau, E. O. Karlberg, A. C. Frank, B. Legault, S. G. E. Andersson, *Proc. Natl. Acad. Sci. U.S.A.* **101,** 9722-9727 (2004).
10. J. Neyman, E. Pearson, *Biometrika* **20A,** 175-240 (1928).

11. J. Dufayard *et al.*, *Bioinformatics* **21,** 2596-2603 (2005).
12. D. T. Jones, W. R. Taylor, J. M. Thornton, *Comput. Appl. Biosci.* **8,** 275-282 (1992).
13. J. Felsenstein, *Cladistics* **5,** 164-166 (1989).

Fig. S1. Phylogenetic tree of the 170 species used in this study. Colors used are the same as in main text Fig. 1: *Eukaryota*, mauve; *Aquifex aeolicus*, pink; *Deinococcus*, purple; *Thermotoga maritima*, green; *Spirochaetales*, light blue: *Rhodopirellula baltica*, ochre; *Firmicutes,* light green; *Bacteroidetes*, brown; *Fusobacterium nucleatum*, red; *Chlamydiales*, orange; *Actinobacteria*, blue; *Cyanobacteria*, dark green; *Proteobacteria,* magenta; *Archaea*, yellow. Blue dots correspond to nodes analyzed in Table S2.

This figure corresponds to the figure C.2 of the manuscript.

Figure S2. Extent of protein innovation in prokaryotic organisms. Histograms of the proportions of protein families (A) or protein module families (B) found in extant organisms that are most recent (dark gray, left panels), intermediary (gray, middle panels), or trace back to LUCA (light gray, right panels). Distributions are computed over 161 prokaryotic species.

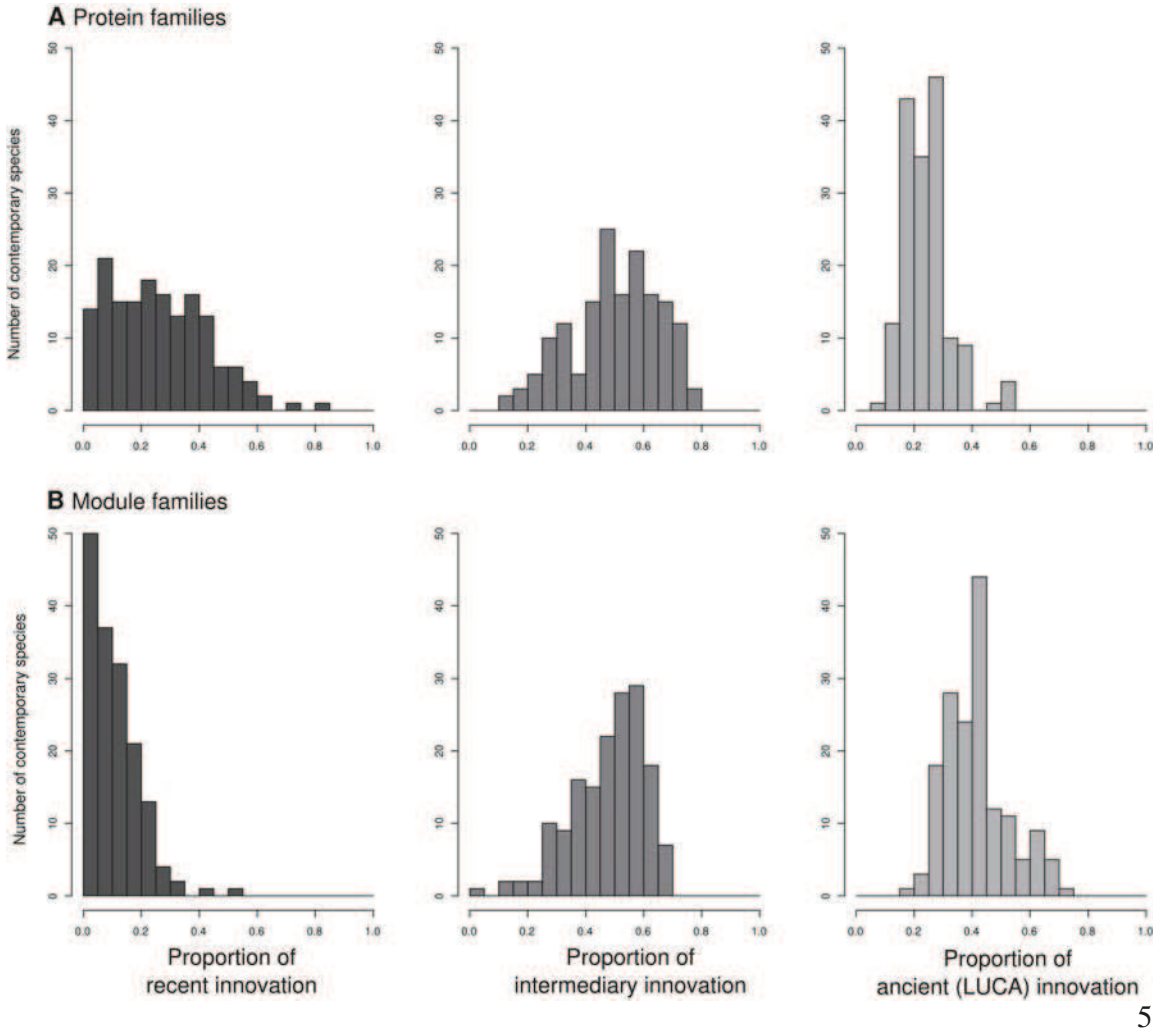


Fig. S3. Versatility of recent *vs.* ancient protein modules. Histograms of protein module versatility, defined as the number of different module arrangements in which a module participates. Histograms are given for module families that are most recent (dark gray), intermediary innovations (gray), or trace back to LUCA (light gray).

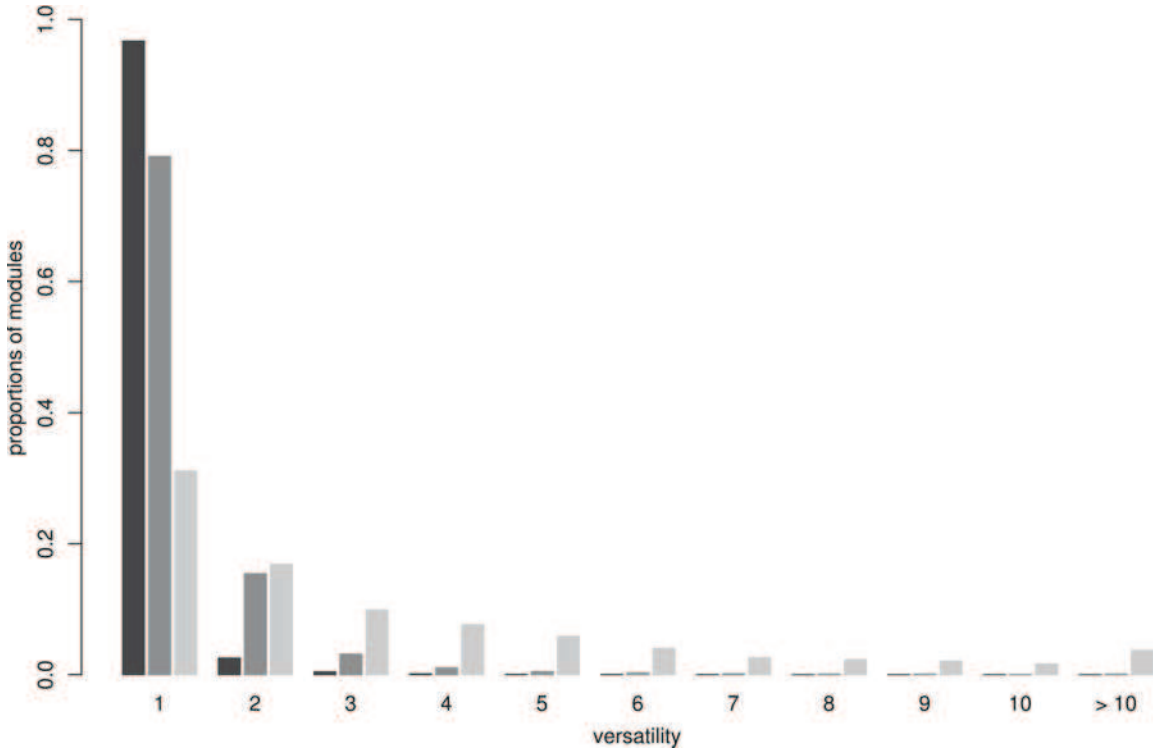
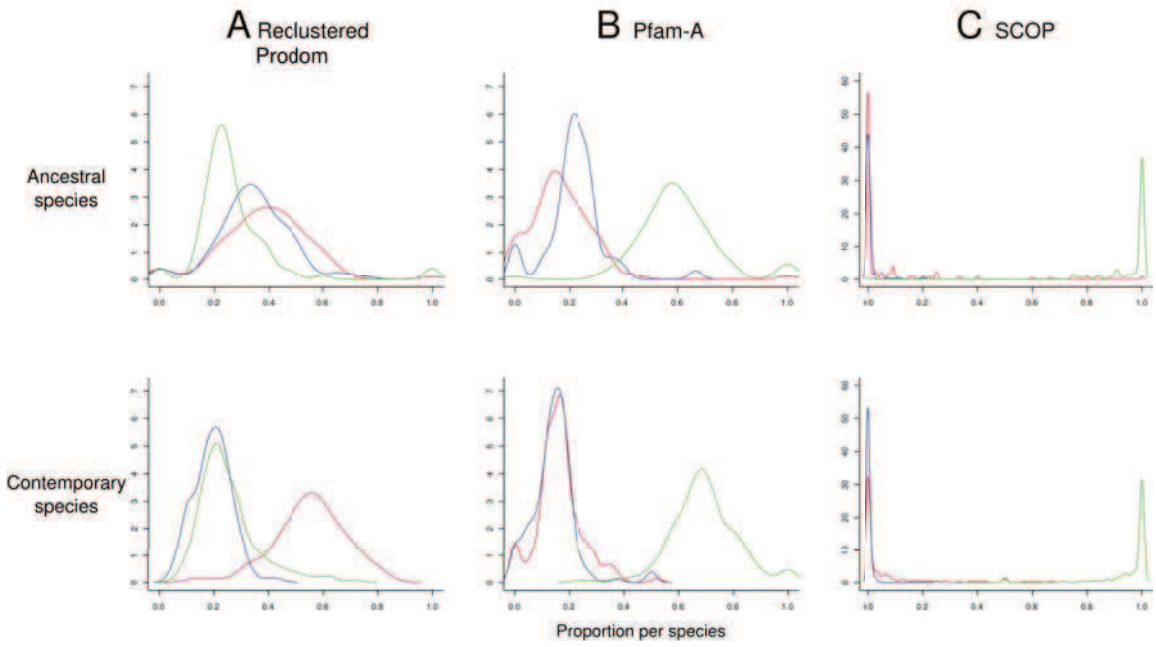


Fig. S4. Distributions of the different types of protein innovation per species. Distributions of relative frequencies are reported for completely innovated, partly innovated and module shuffled proteins in red, blue and green respectively. Distributions were computed over 114 ancestral species (top) or 161 contemporary species (bottom), on the basis of (A) reclustered ProDom, (B) Pfam-A and (C) SCOP family homology.



Figures S5-S13. Evolution of (a) protein module repertoires and (b) protein family repertoires. Pie chart areas reflect numbers of families inferred at each node. Families inherited from the parental node are represented in gray. Gained families that can be found in other clades are inferred as horizontally transferred and indicated in yellow. All other families are predicted to be innovated. They are indicated in red for protein module families (a) and in red, blue and green for protein families that were completely innovated, partly innovated and module shuffled, respectively (b). For clarity the full phylogenetic tree (Fig. S1) was split into several subtrees.

These figures correspond to figures B.3 parts A and B in the manuscript.

Table S1. List of the 170 species used in this study. TaxIDs correspond to taxonomic identifiers from the NCBI taxonomy.

This table corresponds to the table C.1 in the manuscript.

Table S2. Overlap between similarity scores of recent vs. ancient protein modules. Similarity scores were calculated from alignments of orthologous protein modules in different pairs of closely related species. Slow evolving modules are defined here as having a similarity score larger than the 5% quantile of the distribution of similarity scores for modules having originated at LUCA. The table reports the fraction of slow evolving modules having originated at increasingly ancient nodes (indicated by blue dots on Fig. S1). This proportion tends to increase for deeper nodes, making estimates of module innovation more conservative for ancient nodes.

This table corresponds to the table B.2 in the manuscript.

L'ÉVOLUTION MODULAIRE DES PROTÉINES : UN POINT DE VUE PHYLOGÉNÉTIQUE

La diversité du monde vivant repose pour une large part sur la diversité des protéines codées dans les génomes. Comment une telle diversité a-t-elle été générée ? La théorie classique postule que cette diversité résulte à la fois de la divergence de séquence et de la combinatoire des arrangements de protéines en domaines à partir de quelques milliers de domaines anciens, mais elle n'explique pas les nombreuses protéines orphelines.

Dans cette thèse, nous avons étudié l'évolution des protéines du point de vue de leur décomposition en domaines en utilisant trois bases de données : HOGENOM (familles de protéines homologues), Pfam (familles de domaines expertisées) et ProDom (familles de modules protéiques construites automatiquement). Chaque famille d'HOGENOM a ainsi été décomposée en domaines de Pfam ou modules de ProDom.

Nous avons modélisé l'évolution de ces familles par un réseau Bayésien basé sur l'arbre phylogénétique des espèces. Dans le cadre de ce modèle, on peut reconstituer rigoureusement les scénarios d'évolution les plus probables qui reflètent la présence ou l'absence de chaque protéine, domaine ou module dans les espèces ancestrales. La mise en relation de ces scénarios permet d'analyser l'émergence de nouvelles protéines en fonctions de domaines ou modules ancestraux. L'analyse avec Pfam suggère que la majorité de ces événements résulte de réarrangements de domaines anciens, en accord avec la théorie classique. Cependant une part très significative de la diversité des protéines est alors négligée. L'analyse avec ProDom, au contraire, suggère que la majorité des nouvelles protéines ont recruté de nouveaux modules protéiques. Nous discutons les biais de Pfam et de ProDom qui permettent d'expliquer ces points de vue différents. Nous proposons que l'émergence de nouveaux modules protéiques peut résulter d'un turn-over rapide de séquences codantes, et que cette innovation au niveau des modules est essentielle à l'apparition de nombreuses protéines nouvelles tout au long de l'évolution.
Mots-clefs : module protéique, domaine, réseau Bayésien, scénario d'évolution, réarrangement, innovation

---

A PHYLOGENETIC VIEW OF THE MODULAR EVOLUTION OF PROTEINS.

The diversity of life derives mostly from the variety of proteins coded in genomes. How did evolution produce such a tremendous diversity ? The classical theory postulates that this diversity results both from sequence divergence and from the combinatorial arrangements of a few thousand primary protein domain types. However this does not account for the increasing number of entirely unique proteins as found in most genomes.

In this thesis, we study the evolution of proteins from the point of view of their domain decomposition and rely on three databases : HOGENOM (homologous protein families), Pfam (manually curated protein domain families) and ProDom (automatically built protein module families). Each protein family from HOGENOM has thus been decomposed into Pfam domains or ProDom modules.

We have modelled the evolution of these families using a Bayesian network based on the phylogenetic species tree. In the framework of this model, we can rigorously reconstitute the most likely evolutionary scenarios reflecting the presence or absence of each protein, domain or module in ancestral species. The comparison of these scenarios allows us to analyse the emergence of new proteins in terms of ancestral domains or modules. Pfam analysis suggests that the majority of protein innovations results from rearrangements of ancient domains, in agreement with the classical paradigm of modular protein evolution. However a very significant part of protein diversity is then neglected. On the other hand ProDom analysis suggests that the majority of new proteins have recruited novel protein modules. We discuss the respective biases of Pfam and ProDom underlying these contrasting views. We propose that the emergence of new protein modules may result from a fast turnover of coding sequences and that this module innovation is essential to the emergence of numerous novel proteins throughout evolution.
Keywords : protein module, domain, Bayesian network, evolutionary scenario, domain shuffling, innovation

---

DISCIPLINE : Bioinformatique
INTITULÉ ET ADRESSE DU LABORATOIRE
Laboratoire de Biométrie et Biologie Évolutive - UMR 5558 CNRS
Bâtiment Gregor Mendel - Université Claude Bernard Lyon 1
43, bd. du 11 Novembre 1918 - 69622 Villeurbanne CEDEX