Université de Montréal

Analyse de la corrélation conditionnelle dérivée de la coévolution d'un système de trois gènes par un modèle du maximum de vraisemblance

par
Louis Philip Benoit Bouvrette

Département de biochimie
Faculté de médecine

Mémoire présenté à la Faculté des études supérieures
en vue de l'obtention du grade de Maître ès sciences (M.Sc.)
en bio-informatique

Août, 2010

Université de Montréal
Faculté des études supérieures

Ce mémoire intitulé :

Analyse de la corrélation conditionnelle dérivée de la coévolution d'un système de trois gènes par un modèle du maximum de vraisemblance

présenté par :

Louis Philip Benoit Bouvrette

a été évalué par un jury composé des personnes suivantes :

Nadia El-Mabrouk,	président-rapporteur
Miklós Csűrös,	directeur de recherche
Sylvie Hamel,	codirecteur
Hervé Philippe,	membre du jury

Mémoire accepté le : . . . . . .

RÉSUMÉ

Les gènes codant pour des protéines peuvent souvent être regroupés et intégrés en modules fonctionnels par rapport à un organelle. Ces modules peuvent avoir des composantes qui suivent une évolution corrélée pouvant être conditionnelle à un phénotype donné. Les gènes liés à la motilité possèdent cette caractéristique, car ils se suivent en cascade en réponse à des stimuli extérieurs. L'hyperthermophilie, d'autre part, est interreliée à la reverse gyrase, cependant aucun autre élément qui pourrait y être associé avec certitude n'est connu. Ceci peut être dû à un déplacement de gènes non orthologues encore non résolu. En utilisant une approche bio-informatique, une modélisation mathématique d'évolution conditionnelle corrélée pour trois gènes a été développée et appliquée sur des profils phylétiques d'archaea. Ceci a permis d'établir des théories quant à la fonction potentielle du gène du flagelle FlaD/E ainsi que l'histoire évolutive des gènes lui étant liés et ayant contribué à sa formation. De plus, une histoire évolutive théorique a été établie pour une ligase liée à l'hyperthermophilie.

Mots clés : coévolution, archaea, motilité, reverse gyrase, déplacement de gène non orthologue

ABSTRACT

Protein coding gene may often be grouped and integrated in functional modules with respect to an organelle. These modules may have constituents that follow a conditional correlated evolution to a given phenotype. Genes linked to motility posses this characteristic as they follow a cascade in response to external stimuli. Similarly, hyperthermophily is related to reverse gyrase, however no other element that could be associated with certainty is known. This may be caused by an unresolved case of non-orthologous gene displacement. Using a bioinformatic approach, a mathematical model for conditional correlated evolution for three genes has been developed and applied to the phyletic profiles of archaea. This has helped to develop theories about the potential functions of the flagellar gene FlaD/E and the evolutionary history of the genes that are linked to it and that may have contributed to its formation. In addition, a theoretical evolutionary history has been established for a ligase associated with hyperthermophily.

Keywords: coevolution, archaea, motility, reverse gyrase, non-orthologous gene displacement

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES

ADN	Acide désoxyribonucléique
ADP	Adénosine diphosphate
AIC	« Akaike's information criterion »
AMP	Adénosine monophosphate
ARN	Acide ribonucléique
ATP	Adénosine triphosphate
BLAST	« Basic Local Alignment Search Tool »
COG	« Clusters of orthologous group »
DAC	Dernier ancêtre commun
DGNO	Déplacement de gènes non orthologues
FBA	Fructose-1,6-bisphosphate aldolase
NAD	Nicotinamide adénine dinucléotide
NTP	Nucléoside triphosphate
RLV	Ratio de log-vraisemblance
Symbets	« Symetrical best hits »
SAM	S-Adénosyl méthionine
SAH	S-adénosylhomocystéine
t	Temps d'évolution
μ	Taux de perte du gène X
λ	Taux indépendant de gain des gènes Y et Y'
ν	Taux de perte du gène Y en présence individuelle
δ	Taux de perte des gènes Y et Y' présents en paire

REMERCIEMENTS

Je tiens tout d'abord à remercier mon directeur de maîtrise, Miklós Csűrös, ainsi que ma codirectrice, Sylvie Hamel, pour m'avoir accueilli dans leur laboratoire. Je les remercie aussi pour toute l'aide qu'ils m'ont apportée par nos conversations et débats d'idées ainsi que pour le temps et le soutien qu'ils m'ont octroyés.

Je remercie tout autant Nadia El-Mabrouk et Hervé Philippe d'avoir accepté de faire partie du jury de ce mémoire.

Je veux remercier mes parents qui m'ont appuyé et encouragé dans mes études de façon à ce que celles-ci restent mon seul souci.

Finalement, un grand merci à Valérie pour les nombreuses discussions théoriques lors des périodes où tout allait bien et motivantes pour les instants plus laborieux.

CHAPITRE 1

INTRODUCTION

1.1 Introduction

Les gènes d'une espèce influencent sa capacité à survivre, à croître, à se développer et à se reproduire. Les variations de la composition génique d'une espèce permettent l'adaptation de celle-ci à son environnement. La génomique évolutive étudie les mécanismes responsables de cette dynamique des génomes. Au cours de l'évolution, la composition génique se forme par l'interaction de la dérive génétique (sélection neutre) et des forces de sélection positive, qui augmentent le nombre de traits bénéfiques, ou négatives, qui diminuent le nombre de traits néfastes. Les mêmes pressions de sélections peuvent s'appliquer à un ensemble de gènes relatifs à un trait phénotypique. En conséquence, la composition génique reflète la coévolution de gènes.

La coévolution peut être analysée par la comparaison de profils phylétiques. Un profil phylétique est un vecteur binaire qui représente la présence ou l'absence d'un génotype, ou d'un phénotype, pour différentes espèces données. La corrélation entre des profils phylétiques peut indiquer la similarité de forces de sélection. Lors d'un couplage fonctionnel, la perte d'un gène peut rendre un autre gène superflu et ainsi permettre sa perte sans conséquence désavantageuse. Un couplage fonctionnel peut donc être indiqué par une corrélation positive entre deux profils. Inversement, une corrélation négative, où la présence d'un gène implique l'absence de l'autre gène, peut indiquer un « remplacement » où le gain d'un gène rend un autre gène redondant et permet ainsi sa perte. Cette anticorrélation peut refléter un déplacement de gène non orthologue où un gène substitue la fonction d'un gène perdu. La corrélation positive et négative entre des profils peut donc indiquer la fonctionnalité similaire ou reliée entre des gènes.

Ce travail a pour but d'étudier des modèles probabilistes sur la coévolution à partir de deux ou trois profils phylétiques. Ce nouveau modèle permet d'établir des liaisons conditionnelles entre trois états constitutifs d'une espèce, qu'ils soient génotypiques ou phéno-

typiques. Dans le reste du présent mémoire, ces corrélations sont représentées comme un système de trois gènes et symbolisées par les variables X, Y et Y', où Y et Y' sont des gènes ou famille de gènes connexes conditionnels à un constitutif X. Ce constitutif X peut être un trait génotypique ou phénotypique.

Des modèles suivant les mêmes principes existent déjà pour établir des liaisons entre deux profils phylétiques [68]. La motivation derrière la généralisation de cette modélisation est de tester un modèle d'évolution corrélée entre deux gènes (Y et Y') pouvant être conditionnelle à un phénotype donné (X). Ce modèle est une sorte de premier filtre pointant vers des gènes reliés.

Ce mémoire a la structure suivante : le chapitre 1 élabore les grandes lignes biologiques et mathématiques soutenant la présente étude ; le chapitre 2 discute des concepts biologiques inhérents aux cas particuliers utilisés pour appuyer la démarche mathématique ; le chapitre 3 développe les bases fondamentales des regroupements d'homologues ; le chapitre 4 détaille les principes algorithmiques mis en oeuvre pour la résolution des modèles et le chapitre 5 présente les résultats expérimentaux sur des exemples représentatifs.

1.2 Génomique évolutive

Depuis les travaux fondateurs de Gerardus Johannes Mulder et Jöns Jacob Berzelius, les premiers à avoir décrit et nommé, en 1838, les protéines, les biochimistes et biologistes de tout horizon se penchent sur la caractérisation et l'évolution protéique [94]. James Batcheller Sumner, en 1926, a été le premier à démontrer que l'uréase était en fait une protéine, conférant ainsi un rôle central aux protéines chez un organisme [87]. En 1955, Frederick Sanger fut le premier à séquencer une protéine, l'insuline [76]. En ces temps, les techniques de séquençage des protéines ont permis la production d'un nombre croissant de séquences et il a sitôt été évident que les protéines et les acides nucléiques pourraient être utilisés pour documenter l'histoire des événements évolutifs. En 1966 et 1967, Richard V. Eck et Margaret O. Dayhoff ainsi que Walter M. Fitch et Emanuel Margoliash ont respectivement et distinctement utilisé des séquences moléculaires pour

inférer l'évolution [22, 28]. La bio-informatique, encore à ses prémices, a été de plus en plus impliquée dans l'analyse de l'évolution des protéines. Aujourd'hui, les applications bio-informatiques courantes permettant l'étude de l'évolution des protéines comprennent, entre autres, des algorithmes alignant des séquences similaires afin de détecter des homologues dans les grandes bases de données ou permettant de reconstruire des arbres phylogénétiques à partir d'un ensemble donné de séquences. La conjonction moderne de ces deux processus permet dorénavant de déterminer des fonctions et prédire certaines interactions protéiques. Ceci a donné lieu à l'établissement d'une panoplie d'énigmes génétiques qui sont bien étudiées, mais qui sont comprises seulement de façon superficielle, que ce soit par rapport à leurs fonctions ou leurs caractéristiques. Ces nouvelles interrogations amènent les recherches contemporaines à tenter de comprendre la complexité des organismes en analysant la coopération de leurs composants individuels [31]. L'idée de la génomique évolutive est de développer des fondements et des théories relatifs aux différents processus évolutifs majeurs et de les appliquer à la reconstruction de l'histoire de la vie.

Les liens évolutifs entre les ancêtres et leurs descendants peuvent être représentés par un arbre évolutif, aussi appelé phylogénie. Mathématiquement, une phylogénie est un arbre enraciné où les arcs sont orientés selon les relations parentales et les feuilles correspondent aux taxons terminaux.

La répartition de toutes les espèces vivantes répertoriées à ce jour a permis de les séparer en un système à trois domaines, soit les eucaryotes, les bactéries et les archaea. Les deux premiers domaines, eucaryote et bactérie, sont les plus étudiés et par le fait même les plus connus. Les recherches sur le domaine des archaea sont toutefois en constante évolution et les informations et hypothèses qui en ressortent font que de plus en plus de chercheurs s'intéressent à leur génome et à leur évolution.

Le domaine archaea, dont il est largement question dans les sections suivantes, a été proposé par Carl Woese qui voulait, à l'origine, créer une classe d'appartenance phylogénétique distincte pour les espèces produisant du méthane ainsi que les halophiles et thermoacidophiles [55, 95, 96]. Dans un premier temps, seuls les méthanogènes ont été placés dans ce nouveau domaine. Les premiers archaea méthanogènes recensés avaient

principalement comme habitats des marais anoxiques et des sédiments lacustres. Initialement, la plupart des espèces ont été isolées à partir d'habitats extrêmes en ce qui concerne le pH, la température et le taux de chlorure de sodium environnants [24]. Aujourd'hui, des archaea ont été observés dans des environnements tout aussi distincts que conventionnels tels les océans, les sources géothermiques, les sols, le tractus gastro-intestinal de mammifères ainsi que dans le cytoplasme des protozoaires anaérobies comme endosymbiotes. Il est dorénavant admis qu'ils sont plus importants, en termes de nombre et de rôles spécifiques, pour les écosystèmes qu'originalement estimés [90]. Les scientifiques s'entendent maintenant pour dire que les archaea constituent un groupe important et diversifié d'organismes qui sont largement répartis dans la nature.

La pluralité des espèces se regroupant à l'intérieur du domaine des archaea en fait de bons candidats pour appliquer les algorithmes bio-informatiques complémentant les connaissances biochimiques et biologiques. Notamment, l'évolution des répertoires de gènes du domaine des archaea s'avère intéressante. Puisque les archaea sont adaptés à des environnements très différents, ils représentent un défi pour n'importe quelle méthode d'inférence.

1.3 Groupement d'homologues

L'analyse d'homologues est une méthode fondamentale en génomique évolutive. Des homologues sont définis comme des gènes qui descendent d'un ancêtre commun. Une distinction explicite entre les notions principales d'orthologie, de paralogie et de xénologie est primordiale afin d'obtenir des analyses robustes et fiables de classification et de reconnaître des facteurs fonctionnels dans un parcours évolutif [48]. Cette distinction permet de travailler avec une hypothèse donnée sur les liens entre fonction et évolution du répertoire génique.

Les événements de base que peut prendre l'évolution d'un gène sont classifiés selon leurs origines. Un gène chez un organisme donné peut i) suivre une lignée verticale par la spéciation ; ii) faire suite à une duplication entre deux générations ; iii) subir une perte ; iv) être transféré de façon horizontale ; ou v) être le résultat d'une fusion, fission

ou autre réarrangement génique [48]. Avec les notions d'événements évolutifs élémentaires, toutes les descriptions de l'évolution de gènes, de groupe de gènes, et, ultimement, de répertoire de gènes complet, peuvent être regroupées sous les concepts établis d'homologues, d'orthologues et de paralogues [48].

De façon traditionnelle, les analyses de relation entre gènes sont basées sur les similarités de séquences. Par contre, la notion d'orthologie se démarque de ce type d'analyse puisqu'elle vient de l'étude de regroupements à l'intérieur d'un espace phylogénétique. Les orthologues sont des gènes connectés par un héritage évolutif vertical (cas i), d'une espèce ancestrale vers son descendant, puisqu'il s'agit du « même gène » [31, 32]. Cette définition est donc strictement phylogénétique et exclut tout aspect fonctionnel. Identifier ainsi des orthologues permet d'effectuer des recherches sur la coévolution des protéines et sur l'évolution des protéomes complets [31]. Dans la figure 1.1, l'arbre 1 montre un cas simple où l'évolution de la famille X implique uniquement un héritage vertical à partir du dernier ancêtre commun (DAC). Ces gènes sont donc tous orthologues entre eux.

En opposition aux orthologues se situent les paralogues qui sont des gènes liés par une duplication (cas ii) à l'intérieur d'un génome ancestral [32]. La formation de gènes paralogues par duplication est un processus central de l'émergence de nouvelles fonctionnalités génétiques [48].

Les paralogues se classifient plus spécifiquement en in-paralogues et out-paralogues qui sont définis comme des gènes paralogues ayant évolué respectivement lors d'un événement ultérieur ou antérieur à une spéciation [48]. Dans la figure 1.1, les gènes YA1 et YA2 sont in-paralogues relativement aux espèces A et B si leur duplication est arrivée après spéciation, ou des out-paralogues si leur duplication précède la spéciation (impliquant la perte d'une copie dans la lignée de B). Des out-paralogues ne peuvent jamais être des orthologues, mais les in-paralogues peuvent former un groupe de gènes qui, ensemble, sont orthologues à un autre gène d'une autre espèce [67]. À ces définitions doit s'ajouter le concept de coorthologues qui représentent deux gènes ou plus d'une espèce qui sont collectivement orthologues à un gène ou plus d'une autre espèce. Les membres d'un groupe de gènes coorthologues sont des in-paralogues en relation à

leur événement de spéciation respectif [48]. Dans la figure 1.1, les gènes YA1 et YA2 sont coorthologues des gènes YB et YC. Dans l'arbre 3, les gènes ZA1, ZA2 et ZA3 sont collectivement coorthologues des gènes ZB1 et ZB2.

Il est aussi important de considérer les effets des transferts horizontaux (cas iv) sur les relations observées entre des gènes. Des gènes xénologues sont des gènes homologues qui ne sont pas acquis par descente verticale, telles la duplication ou la spéciation. Des gènes xénologues peuvent apparaître comme étant orthologues lors d'une comparaison de génome deux à deux, mais se révèlent comme xénologues lorsque d'autres espèces sont ajoutées [48].

Le postulat d'un gène singulier ancestral est un aspect clé de l'orthologie. L'ancêtre désigné par ce postulat appelle à la présence d'un gène ancestral se trouvant chez le DAC à deux espèces, et non à un ancêtre arbitrairement plus lointain [48]. De plus, la notion d'orthologie n'est pas transitive quand l'histoire évolutive inclut des duplications suivies par différentes pertes entre lignées descendantes [86].

Régulièrement, les orthologues possèdent les mêmes fonctions, ce qui justifie de les utiliser aux fins d'annotation, de positionnement ou de détermination de fonction de génome [32]. Par contre, ceci peut s'avérer risqué puisque la notion d'orthologie n'est pas nécessairement une relation un pour un, mais peut être une relation un pour plusieurs où un gène dans une espèce peut correspondre à toute une famille de gènes paralogues dans une autre branche. Zhang a montré expérimentalement l'évidence de « néo-fonctionnalité », où une duplication est à l'origine d'une nouvelle fonction, et de « sous-fonctionnalité », ou les duplicats sont maintenus de façon stable lorsqu'ils diffèrent dans certains aspects de leurs fonctions [97]. Dans ces cas, l'annotation par homologie est erronée.

Si les gènes orthologues ont régulièrement des fonctions similaires, l'inverse est rarement vrai. De plus, il existe des situations où des fonctions équivalentes sont créées par des gènes non orthologues, souvent même des non homologues, regroupés sous la notion de déplacement de gènes non orthologues (DGNO) [48]. Ce cas particulier sera approfondi à la section 2.5.

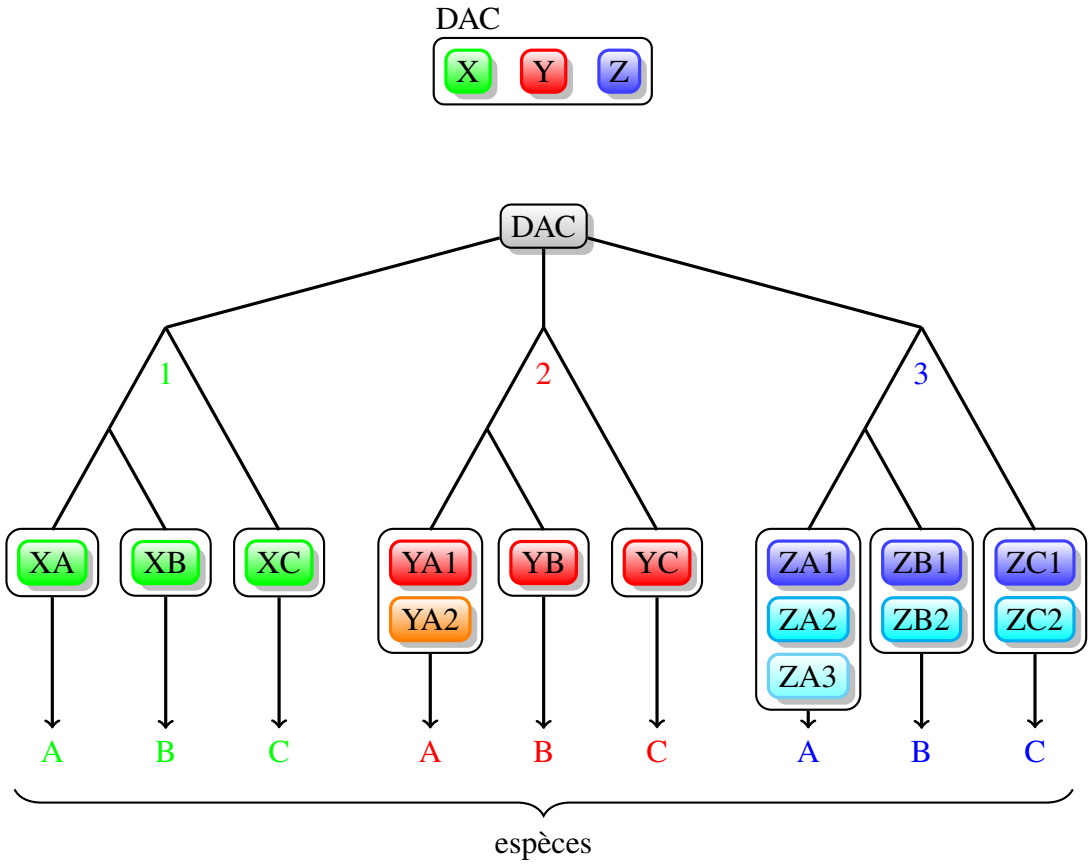


Figure 1.1 – Arbre phylogénétique hypothétique illustrant l'évolution de trois familles dans les génomes de A, B, C. Chaque famille montre un cas distinct de relations entre homologues. Le gène ancestral commun à la famille complète de gènes existait avant le dernier ancêtre commun des trois espèces. DAC : dernier ancêtre commun. Les lettres A, B et C représentent les espèces et X, Y et Z les gènes [48].

1.4 Profil phylétique

Le profilage phylétique décrit par Pelligrini et coll. et par Tatusov et coll. est une méthode d'inférence fonctionnelle basée sur la distribution phylogénétique d'homologues [70, 89]. Un profil phylétique se définit comme un vecteur binaire, présentant une distribution représentée sous forme de 0 pour l'absence et 1 pour la présence d'un gène, d'un groupe de gènes ou d'un phénotype donné chez une panoplie d'espèces.

L'approche par profil phylétique est basée sur l'hypothèse que des protéines qui ont

des fonctions associées à l'intérieur d'une cascade, d'une voie métabolique ou d'une structure plus complexe sont plus disposées à évoluer d'une manière corrélée. Au cours de l'évolution, les pressions sélectives sur des protéines interdépendantes sont similaires puisqu'elles tendent à être conservées ou éliminées de façon concomitante [42, 70]. Dans un exemple illustratif, il est « inutile » de conserver une protéine qui a pour seule fonction un rôle accessoire à une protéine centrale qui aurait été perdue. La sélection contre la perte de cette protéine n'est plus aussi forte à cause de la redondance fonctionnelle.

À partir de séquences d'organismes, qu'elles soient d'acides aminés ou de nucléotides, la décomposition de l'information génétique en profils phylétiques s'effectue en regroupant des gènes entre eux par homologie (figure 1.2 A). La représentation de chaque groupe, ou famille, de gènes dans les génomes étudiés est caractérisée par son profil phylétique (figure 1.2 B). Le couplage fonctionnel est inféré en observant les profils identiques entre familles (figure 1.2 C, 1.2 D) [3, 70].

D'autres relations sont à attendre dans l'inférence de fonction. Les patrons peuvent être soit similaires, mais pas identiques, ou complémentaires. Tel qu'il sera élaboré plus en détail à la section 2.5, le patron de protéines agencé en profil complémentaire peut indiquer un déplacement de gènes non orthologues. Ces protéines peuvent donc avoir des fonctions similaires tout en présentant des profils singuliers [31]. Par contre, la complémentarité ne sera que très rarement parfaite en raison de redondances fonctionnelles. Ces nuances sont importantes puisque le reste du mémoire s'attardera principalement sur deux types de corrélation, soit la similarité et la complémentarité. Les forces complexes d'évolution, les erreurs d'annotation et la redondance fonctionnelle forcent la considération de méthodes de comparaison entre profils plus sophistiquées que la recherche d'identité.

Par exemple, le tableau 1.I propose d'imager ce concept en regardant une analyse hypothétique où les gènes responsables de la synthèse du flagelle seraient à l'étude chez cinq espèces d'archaea, trois motiles et deux non motiles.

En analysant le tableau 1.I, il est valable de conclure que les familles de gènes arCOG1824, arCOG2965 et arCOG5519 pourraient avoir une implication dans le fonc-

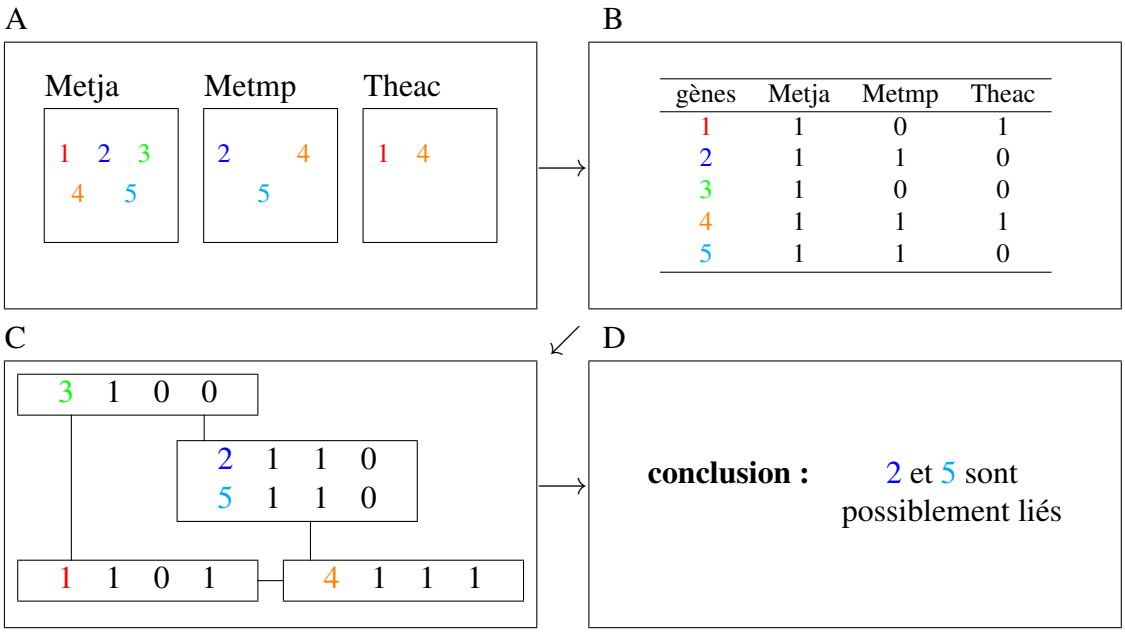


Figure 1.2 – Schématisation de la procédure permettant de décomposer l'information génétique en profil phylétique menant à des conclusions sur les liens entre les gènes. (A) Séquençage génomique et identification de gènes. (B) Mise en tableau, profil phylétique, des gènes selon leur présence et absence dans les espèces à l'étude. (C) Identification de profils identiques. (D) Hypothèse sur les liens possibles entre plusieurs groupes de gènes.

Tableau 1.I – Exemple de profil phylétique pour un phénotype de motilité de cinq archaea avec leur génotype associé.

famille	Metja	Metmp	Theac	Picto	Naneq
Motile	1	1	1	0	0
arCOG1824	1	1	1	0	0
arCOG2965	1	1	1	0	0
arCOG5119	1	1	1	0	0
arCOG1117	0	0	0	1	1
arCOG4402	0	1	0	0	0
arCOG1512	1	0	0	1	0

Les 1 et 0 indiquent respectivement la présence et l'absence d'au moins un gène regroupé sous cette catégorie dans cette espèce.
Metja = *Methanocaldococcus jannaschii* ; Metmp = *Methanococcus maripaludis* ;
Theac = *Thermoplasma acidophilum* ; Picto = *Picrophilus torridus* ;
Naneq = *Nanoarcheota equitans*

tionnement du flagelle chez ces archaea. La description de ces familles est, respectivement, protéine de flagelle présumée F, protéine de flagelle présumée D/E et protéine

de flagelle présumée C. Il est aussi possible de proposer que les familles arCOG4402 et arCOG1512 sont des gènes ayant été soumis à un déplacement de gènes non orthologues. Ces deux familles ont des descriptions respectives de protéine chimiotactique acceptrice de méthyle et N-methylhydantoinase B/acetone carboxylase, soit deux protéines ayant un rôle défini dans le transfert de groupement méthyle. L'approche ici proposée est par contre modeste, les profils phylétiques sont insuffisants par eux-mêmes pour établir une inférence fonctionnelle. Il est nécessaire d'examiner les candidats cas par cas. La famille arCOG1117 est décrite comme un régulateur transcriptionnel, la comparaison des profils présentés au tableau 1.I n'est donc pas suffisante pour lui inférer un rôle dans la motilité.

Ce mémoire propose des méthodes d'analyse bio-informatiques appliquées à des traits phénotypiques et de profils de gènes. Les modèles de corrélation entre profils phylétiques peuvent s'appliquer à n'importe quel trait phénotypique ou génotypique [42, 56]. De façon similaire à lier des gènes entre eux, il est possible de lier des gènes à des phénotypes sans pour autant avoir une connaissance préalable sur l'implication d'autres gènes dans le processus à l'étude [42]. De cette façon, il est possible d'étudier l'interaction génique composant divers phénotypes, tels que la motilité ou l'hyperthermophilie d'un organisme. Il suffit d'encoder le trait phénotypique de façon binaire comme dans le cas d'absence ou présence d'une famille de gènes.

Comme dans l'exemple du tableau 1.I, les preuves apportées par la corrélation de profils sont généralement corroborées par d'autres dans l'inférence de réseaux génétiques. Idéalement, et à plus forte raison dans les cas où la connaissance *a priori* n'est que sommaire, une recherche complète prend en compte une variété de sources telles que l'expression, les profils et les locations chromosomiques. La multiplicité des sources de données permet d'atténuer la faiblesse de l'information liée à une source unique [84].

À l'origine, ce type d'analyse a été uniquement effectué sur des protéines avec des profils parfaitement identiques. S'en est suivi des études sur des profils entièrement complémentaires. Il est maintenant admis qu'elles sont tout aussi valides si elles sont effectuées sur des profils qui ne sont pas parfaitement identiques ni complémentaires [2, 70]. Dans la pratique, même les gènes dits essentiels et ubiquitaires n'ont pas des profils parfaitement identiques ou complémentaires. Une étude sur des profils d'un noyau de gènes

permet donc d'ajouter ou de soustraire des gènes, qui auraient, au fil de l'évolution, permis un peaufinage de la voie ou du complexe auxquels ils appartiennent.

Les profils phylétiques ne proposent donc pas tous la même qualité d'information. Des gènes peuvent exister uniquement dans une espèce alors que d'autres se retrouvent parmi tous les genres d'un phylum ou d'un domaine. Dans ce dernier cas, il est plus difficile d'inférer une liaison de fonction parce que la corrélation s'explique déjà par la phylogénie. Il est nécessaire d'utiliser des organismes entièrement séquencés et un large éventail d'espèces possédant les divers phénotypes pour obtenir des profils d'une plus grande utilité. En général, il a été démontré que des profils qui contiennent une grande variété d'espèces réparties uniformément parmi différentes niches écologiques donnent de meilleurs résultats [52]. De cette façon, si une paire de gènes est retrouvée dans une fraction d'espèces d'une distribution phylogénique diverse, il est plus justifiable de leur attribuer un lien fonctionnel.

Les modèles étudiant deux familles, ou groupes, de gènes à la fois, se sont rapidement montrés insuffisants pour délier la complexité des réseaux cellulaires et des voies métaboliques. De plus, ne pas connaître la relation phylogénétique entre les organismes étudiés rend l'interprétation de corrélation difficile. Cette corrélation pouvant s'expliquer par la relation phylogénétique entre espèces reliées, ou par similarité de forces de sélection entre diverses espèces. Sur le plan bio-informatique, il est difficile de trouver une pondération sophistiquée entre les entrées d'un profil phylétique. Puisque le but de cette étude est de généraliser l'analyse des réseaux cellulaires en prenant en compte trois états, il est nécessaire d'établir un score de complémentarité entre deux gènes, Y et Y'.

La formule suivante offre une heuristique simple pour quantifier la complémentarité entre profils :

$$C = \frac{n01 + n10 - n11}{n01 + n10 + n11} \tag{1.1}$$

où :

n01 = nombre d'espèces où le gène Y' est présent et Y est absent

n10 = nombre d'espèces où le gène Y est présent et Y' est absent
n11 = nombre d'espèces où les gènes Y et Y' sont présents
n00 = nombre d'espèces où les gènes Y et Y' sont absents. Cette variable n'est pas utilisée puisqu'elle correspond à n-n01-n10-n11 dans la formule.

De cette équation, un nombre entre -1 et 1 sera obtenu. Un $C = -1$ représentera une corrélation positive complète. Un $C = +1$ représentera une complémentarité parfaite, ou anticorrélation. Le score ne considère ni la phylogénie des espèces (il suppose une phylogénie en étoile) ni la variation de pertes entre lignées. Par contre, C permet de sélectionner des paires de profils « prometteur » en prenant arbitrairement, par exemple, des valeurs supérieures ou inférieures à 0,5. Le chapitre 4 introduit un formalisme mathématique pour modéliser l'évolution corrélée entre deux ou trois gènes en un contexte phylogénétique.

CHAPITRE 2

ÉVOLUTION CORRÉLATIONNELLE

2.1 Introduction

Ce chapitre s'intéresse aux connaissances biologiques préexistantes nécessaires afin de bâtir cette présente recherche. Il est défini en quatre parties : la section 2.2 s'intéresse aux généralités des modules de gènes fonctionnels suivant une évolution corrélée ; à la section 2.3, il est question du flagelle des archaea ; la section 2.4 traite de la protéine reverse gyrase, spécifique aux procaryotes hyperthermophiles ; et la section 2.5 porte sur le déplacement de gènes non orthologues. Ces quatre sections proposent des exemples concrets illustrant le phénomène de la modularité qui seront pris en considération lors de l'application des modèles mathématiques définis aux chapitres suivants.

L'évolution corrélationnelle est définie comme le rapport réciproque entre deux, ou plusieurs, gènes, protéines, ou caractères liés à l'évolution d'un individu, qui varient simultanément en fonction l'un de l'autre et qui témoignent d'un lien causal. Cette corrélation, par contre, n'implique pas nécessairement une causalité par lien direct. Par exemple, un système à trois gènes montrant une corrélation dépendante à la reverse gyrase n'attestera pas obligatoirement un lien avec la reverse gyrase, mais pourrait tout aussi bien indiquer deux gènes importants à l'hyperthermophilie.

Si les gènes codant une partie d'une voie ou d'une structure multiprotéique complexe sont perdus, il peut être naturel que les gènes qui composent le reste des protéines impliquées soient inutilisables et perdus. Ceci mène à la modularité de gain et de perte de gènes au cours d'un temps évolutif [5]. L'interaction intermodulaire est mise en exemple avec, comme références, les protéines impliquées dans les modules composant le flagelle ainsi que le module composé uniquement du singleton reverse gyrase. Ces exemples permettent de définir une problématique précise pouvant être analysée par la présente étude.

2.2 Modules de gènes fonctionnels

L'étude des interactions moléculaires est omniprésente au coeur des études biologiques modernes. Ces relations ont été démontrées comme ubiquitaires parmi tous les aspects possibles des fonctions cellulaires. Il a été démontré que les protéines liées par une fonction ont tendance à évoluer ensemble. L'analyse de cette évolution donne lieu à des représentations sous forme de distribution phylogénétique, ou profil phylétique. Des protéines qui ont des distributions phylogénétiques similaires sont souvent des composantes des mêmes voies [10]. Puisque la notion d'orthologie est évolutive, un groupe d'orthologues possède parfois des fonctions différentes. Cette inconstance entre la fonction et l'orthologie est à la base de l'organisation génétique en réseau [83].

Un module est défini comme un ensemble d'unités fonctionnelles. Suivant cette définition, la modularité d'un organisme est contrainte par des processus évolutifs, menant ainsi à la définition de modularité évolutive. Un module évolutif peut être défini comme un groupe de gènes ou de protéines qui ont coévolué ensemble de façon plus accentuée qu'avec des gènes à l'extérieur du module et qui contribuent conjointement à la même fonction cellulaire ou processus biologique [10, 42, 82]. Les modules possèdent un grand nombre de connexions internes entre les protéines fonctionnelles. Les modules correspondent à un complexe physique, ou un complexe fonctionnel, comme celui composant le flagelle [10].

Les modules de gènes fonctionnels ont été observés dans une panoplie de réseaux génétiques tels que les voies enzymatiques métaboliques, le métabolisme d'acides aminés et d'énergie, la motilité cellulaire, le trafic intracellulaire et la sécrétion [10]. En regroupant les gènes en modules, il est possible de réaliser une meilleure analyse de leur évolution.

Les structures modulaires des réseaux biologiques sont dessinées comme étant hiérarchiques et chevauchantes. Des modules différents peuvent donc être reliés entre eux en les combinant en de modules de plus grand ordre par le couplage de leurs protéines (figure 2.1). Un tel lien itératif mène à la connexion de pratiquement toutes les protéines entre elles, de façon directe ou indirecte [42, 83]. Autant des modules de grandes que de

petites tailles peuvent être trouvés chez les modules liés ce qui indique que la taille n'est pas un facteur primordial pour déterminer la cohésion d'un module [10].

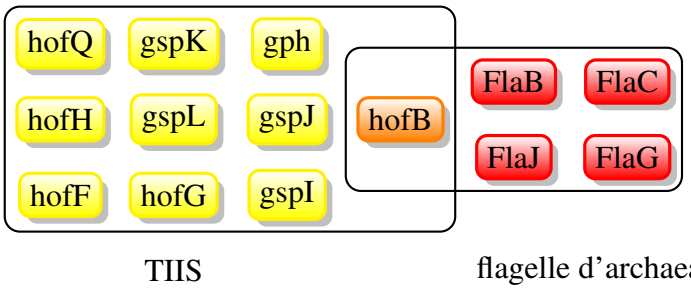


Figure 2.1 – Diagramme de Venn montrant un exemple de deux modules, regroupant les orthologues du flagelle d'archaea et du système de sécrétion de type II (TIIS), se chevauchant [83].

Chacun des petits rectangles arrondis représente un groupement d'orthologues. Les plus grands rectangles arrondis englobent un module. Le groupe d'orthologues hofB fait partie intégrante des deux modules.

Le rapport relatif entre les gènes seuls et les modules de gènes étant affecté à des rôles dans les principales catégories fonctionnelles comme le stockage de l'information, le métabolisme cellulaire et la signalisation, reste à peu près le même. Ceux-ci ne semblent donc pas être biaisés en faveur de certains procédés dans la cellule. Par exemple, il existe environ deux fois plus de modules fonctionnels impliqués dans la métabolisation que dans la traduction, répartis aussi bien en singletons qu'en modules fonctionnels [10].

Les modules ont une évolution de leur cohésion, leurs composantes sont souvent acquises, cédées ou perdues en même temps [10]. La haute connectivité dans un module de gènes liés indique une dépendance entre les gènes d'un module. Quand un des gènes est perturbé, la fonctionnalité des autres gènes peut aussi être compromise. De cette façon, la duplication d'un gène à l'intérieur d'un module est moins probable, surtout si les autres gènes du module ne sont pas dupliqués en même temps. Il est donc attendu d'identifier moins de paralogues dans cette sorte de module [10]. Ce lien entre le dosage et la duplication, par contre, reste hypothétique et est encore matière à controverse, particulièrement dans le cas de module de gènes autorégulés [93].

Cette évolution de la cohésion est reflétée par un profil phylétique tel que décrit à la section 1.4. Dans le tableau 1.I de la page 9, la représentation des arCOG1824,

arCOG2965 et arCOG5119 exprime la connectivité et indique la dépendance entre ces gènes du module lié à la motilité par flagellation.

Il existe une certaine modularité évolutive pour les modules fonctionnels. Les voies métaboliques présentent des variations considérables entre les génomes et la similitude de la répartition des orthologues indique une relation fonctionnelle. Cette flexibilité délimite le potentiel d'utilisation de la cooccurrence de gènes, par l'analyse de profils phylétiques, pour la prédiction de fonctions et de relations, au moins sur de grandes collections de génomes [82].

2.3 Composantes du flagelle des archaea

Dans cette section, il est question d'une brève revue de l'origine et des composantes moléculaires impliquées dans la biogenèse du flagelle, principale organelle responsable de la motilité chez les archaea et les bactéries. Présumément, plusieurs organites sont apparus suite à des modifications de différents types de systèmes de transport. De ceux-ci, les flagelles et les pili sont les principales organelles impliquées, entre autres, au niveau de la motilité. Ils seraient donc tous liés par des systèmes de sécrétion desquels ils auraient évolué de façon indépendante.

Chez les archaea, le flagelle est bien distribué parmi toutes les divisions majeures des crenarchaeota et euryarchaeota, deux principaux phylums de ce domaine. Il existe au moins une espèce flagellée parmi chacun des divers groupes physiologiques et métaboliques, soit les halophiles, haloalkaliphiles, méthanogènes, hyperthermophiles et thermoacidophiles [65, 90].

Les études chez les archaea menées jusqu'à présent se concentrent surtout sur la différenciation entre les deux domaines des procaryotes, soit les bactéries et les archaea. Le principal dispositif de motilité connu chez les archaea est composé d'au moins deux protéines faisant appel à l'énergie résultante d'une hydrolyse d'un nucléoside triphosphate (NTP). Quant au flagelle bactérien, il serait composé d'au moins 50 protéines dans un complexe prenant l'énergie d'une ATPase exploitant un gradient de protons, à première vue d'une plus grande complexité [17]. Chez les archaea, seuls les gènes composant la

flagelline, un élément fondamental du flagelle, sont connus et bien caractérisés [60]. Les archaea posséderaient de multiples flagellines, une douzaine selon les études les plus récentes, alors que les flagelles de bactéries n'en posséderaient qu'une seule [90]. Il a tout d'abord été convenu que les flagelles retrouvés chez les archaea et les bactéries soient similaires, bien qu'ils possèdent quelques différences [11, 12]. Il est maintenant admis qu'il en est tout autrement [25, 65, 90]. Le flagelle des archaea serait plutôt homologue du pilus de type IV des bactéries. Il y a différents types de pili chez les bactéries, dont les pili de type IV qui permettent une motilité par tremblement. Les flagelles d'archaea confèrent tout de même une motilité par propulsion similaire à la motilité par flagelle des bactéries [69].

Il est maintenant généralement admis qu'il y a six types de systèmes de sécrétion de protéines différents chez les bactéries et une très brève introduction de ceux-ci est de mise. Les cinq types bien documentés sont schématisés à la figure 2.2. Le type I est un transporteur ABC (ATP-binding cassette) et forme un pore de trois protéines, une formant le pore, une permettant une fusion membranaire et une protéine ABC, le tout traversant les membranes internes et externes [38]. Le système de type II est dépendant du système Sec, responsable du transport initial de protéines dans le cytoplasme, pour ensuite traverser la membrane. Il est souvent décrit comme étant la branche terminale principale de la voie de sécrétion générale Sec-dépendante [38]. Cette dernière comprend plusieurs embranchements, le système de sécrétion de type II n'est que l'un de ces embranchements. Le système de sécrétion de type II est la voie sécrétoire générale pour le transport des exoprotéines du périplasme vers la membrane externe chez les bactéries [69]. Les bactéries à Gram négatif qui ont un pilus de type IV utilisent une variante de ce type de système pour la biogenèse de leur pilus. Le système de sécrétion de type III est homologue au corps basal des flagelles bactériens et, tout comme le type I, est Sec-indépendant. Il est souvent décrit comme une seringue moléculaire par laquelle une bactérie peut injecter des protéines dans une autre cellule eucaryote [38]. Le système de sécrétion de type IV est un homologue du système de conjugaison des bactéries et des flagelles des archaea [69]. Bien peu est connu et compris sur ce système et plusieurs controverses subsistent encore, y compris lequel du système de motilité ou du système

de conjugaison est apparu le premier sur l'échelle évolutive. Le système de sécrétion de type V est connu sous le nom d'autotransporteur, type Va ou AT-1, ou voie sécrétoire à deux partenaires, type Vb. Un nouveau type Vc s'est aussi récemment ajouté. Les protéines sécrétées par ces voies partagent toutes des similarités autant dans leur structure que dans leur biogenèse. Cet autotransporteur requiert la présence d'un signal peptide N-terminal pour utiliser le système Sec, duquel il est aussi dépendant [38]. Le système de sécrétion de type VI a été découvert seulement en 2006 et n'est que partiellement caractérisé. Les connaissances sur le système de sécrétion de type VI se résument à son implication principale dans la virulence [6].

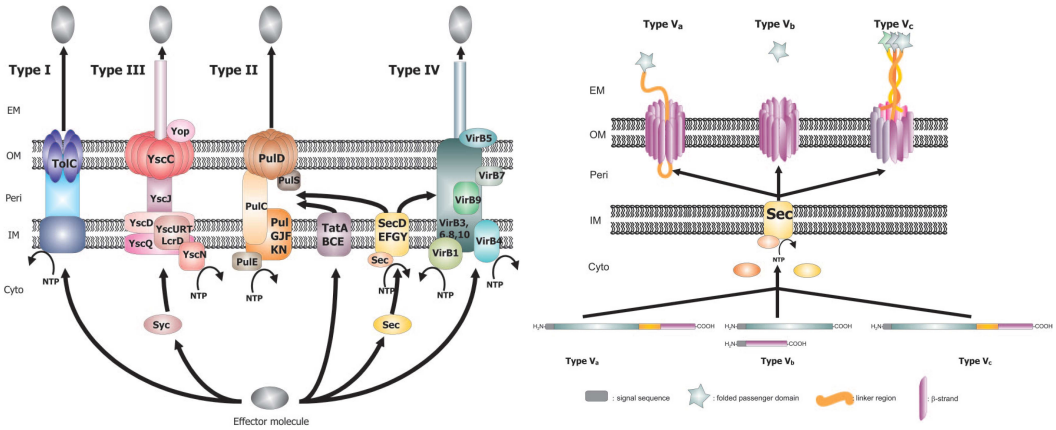


Figure 2.2 – Représentation schématique des systèmes de sécrétion de type I, II, III, IV, et V [38].

Alors que les types I et V sont relativement petits, en termes de nombre et de taille de protéines, les types II à IV sont de larges complexes multiprotéiques. Les systèmes de sécrétion de type III et de type IV ont une, ou possiblement plusieurs, ATPases en commun qu'ils peuvent se partager [6, 69]. Le système de sécrétion de type II possède une ATPase, une protéine transmembranaire, une sécrétine et une prépiline peptidase qui ont chacune un homologue du même nom dans le système de sécrétion de type IV [69]. Par contre, seulement trois ont été démontrées comme homologues du système Fla, codant pour les flagellines d'archaea [69]. Les sécrétines ne se trouvent pas dans le groupe Fla comme il serait attendu. Ceci serait dû à la différence de paroi et à l'absence

de membrane externe chez les archaea [69].

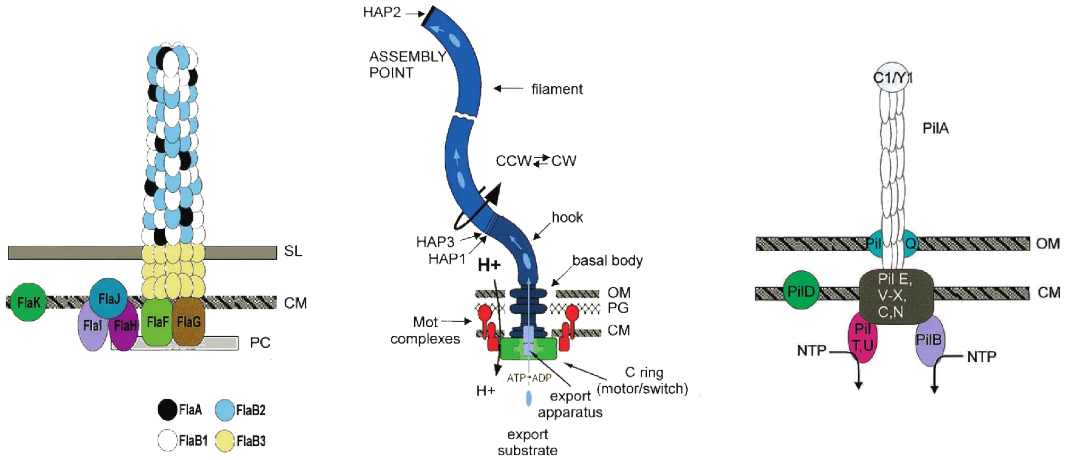
Ces études ont mené à la proposition que les flagelles d'archaea sont plus similaires au pilus de type IV qu'au flagelle bactérien [65]. De plus, les archaea ne possèdent aucun gène homologue chez les bactéries dont le rôle est directement impliqué dans la motilité et la formation du flagelle provenant du même type de système de sécrétion [65]. D'un autre côté, les flagellines ont des séquences d'acides aminés en N-terminal qui sont hautement similaires aux pilines de type IV [65]. Trois types de protéines procaryotes du complexe formant l'enveloppe incluent des protéines prépilines possédant un segment N-terminal hautement similaire d'environ 20 acides aminés. Ces prépilines peuvent s'assembler en structure filamenteuse et composer des parties du système de sécrétion de type II, du système de pilus de type IV ou du système de flagelle d'archaea. De plus, la structure morphologique du flagelle d'archaea, telle qu'observée au microscope, est apparente à celle du pilus de type IV tandis que ces deux organelles sont différentes du flagelle bactérien [4, 65]. De son côté, le flagelle bactérien aurait plutôt émergé d'une évolution de système de transport de type III [4, 65].

Très peu d'information est connue sur les sécrétions protéiques des archaea en tant que telles et encore moins sur leurs liaisons et leurs rôles au niveau de la motilité. Une étude antérieure a par contre trouvé des homologues de Sec connus comme étant des multi-sous-unités liées à la membrane [90]. À ce système de sécrétion s'ajoutent des modules protéiques menant à la formation de l'organelle de motilité. Le locus principal des archaea flagellés est généralement composé des flagellines FlaA ou FlaB auquelles les flagellines de FlaC à FlaJ peuvent s'associer [65]. Les euryarchaeota ont la totalité de ces gènes, alors que les crenarcheota n'en possèdent qu'un sous-groupe qui est différent chez chaque espèce [65]. Parmi les homologues impliqués dans le système de sécrétion de type II, il y en a au moins un possédant un rôle de flagelline, FlaI. FlaI est homologue à TadA, aussi appelée pilT, une ATPase impliquée dans le pilus de types II/IV [4]. Le fait que FlaI soit une parente des flagellines suggère que la biosynthèse s'effectue via un système de sécrétion de type II. FlaJ est quant à elle similaire à TadB, une protéine membranaire [4, 90]. Fla-HIJ sont les seules protéines ayant une présence commune chez tous les archaea motiles, ce qui permet d'avancer l'hypothèse qu'elles forment

le coeur central de cette biogenèse [4]. Les autres flagellines, FlaA-FlaG, ne sont pas aussi communes parmi les archaea motiles. FlaF et FlaG pourraient être équivalentes à pilE et pilV, du pilus type IV, soit des protéines hydrophobes N-terminal similaires aux pilines [90]. Comme les pilines bactériennes, les flagellines d'archaea sont faites de préprotéines avec de courts peptides signaux qui sont coupés par une peptidase spécifique aux archaea, FlaK. Celle-ci montre une similarité de séquences, bien que faible, avec la peptidase maîtresse du pilus bactérien, pilD [4]. Les flagellines bactériennes ne sont pas faites de préprotéines et sont exportées par un système de sécrétion de type III. Les flagellines d'archaea sont faites avec une séquence maîtresse qui est clivée par une peptidase membranaire dans un scénario considéré comme étant similaire au clivage de peptides-chefs du type IV [4, 90].

Sur le plan morphologique, les ressemblances pointent dans la même direction que l'analyse des homologues. Le flagelle est composé de deux parties essentielles : le filament et le crochet. Contrairement aux bactéries, il n'y a pas de corps basal évident chez l'archaea (figure 2.3a). Les bactéries ne possèdent pas de strate de surface (S-layer) par-dessus leurs membranes cytoplasmiques, elles doivent donc avoir un mécanisme d'encrage différent. Leur biogenèse est elle-même différente ; chez les archaea le filament est créé à partir de la base alors que chez les bactéries, les unités supplémentaires sont ajoutées à l'extrémité distale du filament après qu'elles aient traversé le tube que forme celui-ci (figure 2.3b). Chez les bactéries et les archaea, l'organelle responsable de la motilité est une structure rotatoire avec un filament agissant comme propulseur. La rotation peut changer de direction, horaire ou antihoraire, pour se mouvoir en ligne droite ou culbuter, réagissant ainsi à une réponse d'un stimulus chimiotactique, attirant ou repoussant, provenant de l'environnement. Ce système de chimiotactisme est en tout point semblable entre les archaea et les bactéries [4]. Des homologues de la majorité des gènes ont été identifiés, à l'exception de la protéine filM, basculant la rotation du flagelle [4].

Contrairement à la très grande majorité des gènes impliqués dans la flagellation des bactéries, il a été démontré, par analyse sur SDS-PAGE, que les protéines impliquées dans la flagellation des archaea sont glycosilées [90]. Cette modification post-traduction-



(a) Flagelle d'archaea (b) Flagelle de bactérie (c) Pilus de bactérie

Figure 2.3 – Morphologie des différents types de flagelles et de pili d'archaea et de bactéries. (A) Représentation schématique du flagelle d'archaea. (B) Représentation schématique du flagelle de bactéries. (C) Représentation schématique du pilus de bactéries [4].

nelle peut être nécessaire lors des étapes d'assemblage ou d'accrochage du flagelle [90]. Puisque la structure et l'assemblage du flagelle sont complexes et requièrent plus de 40 gènes chez les bactéries, il est envisageable que plusieurs gènes d'archaea de fonction inconnue soient impliqués dans la flagellation et la motilité [90]. En laboratoire, il a été démontré que des mutants ne possédant pas une protéine glycane, ou possédant une protéine tronquée, étaient non flagellés ou possédaient des flagelles défectueux [65]. D'autres gènes comme celui de la glycosyltransférase et de la glycosylase, dont les produits aident à la stabilité protéique dans les environnements extrêmes, ont aussi été identifiés comme ayant un rôle possible dans la biogenèse des flagelles stables [65].

L'absence, chez les archaea, d'homologues de flagellines, de protéines de tige, de protéines de crochet ou associées à celui-ci, d'anneaux interrupteur ou moteur alors qu'il y a une panoplie de gènes de chimiotactisme identifiables entre les archaea et les bactéries est intrigante [90]. Ces deux groupes de protéines sont fortement liés dans leurs interactions chez les bactéries. Il semble donc que les archaea ont adapté un système

de motilité à partir du même système de chimiotactisme avec un système de sécrétion différent.

L'utilisation de profils des gènes du flagelle dans ce mémoire a pour but de répertorier des gènes jusqu'alors connus comme ayant une fonction liée à la membrane cellulaire, aux systèmes de transport de défense, ou chimiotactique, et d'établir des liens avec des gènes du flagelle.

2.4 Reverse gyrase

La reverse gyrase, une topoisomérase, est à ce jour la seule protéine connue comme étant à la fois présente chez les archaea et bactéries hyperthermophiles et absente chez les autres mésophiles et thermophiles [30, 71]. La présence d'une reverse gyrase chez un procaryote donné est déterminée par la température optimale de croissance de cette espèce et non pas par sa position phylogénétique [51, 58]. Ceci lui confère donc un rôle important dans l'adaptation à la vie dans les plus hautes températures des organismes vivant sur Terre, soit des températures supérieures à 80 degrés Celsius [30].

Il existe deux types de topoisomérase, I et II. Elles ont comme fonction principale de contrôler la structure topologique l'ADN [14]. Elles sont une solution aux multiples problèmes associés aux interactions ADN-protéines telles que la réplication, la réparation, la recombinaison, la ségrégation de chromosomes et la régulation de l'expression génique [14, 71].

Alors que les topoisomérases ont plutôt des rôles de relaxation de l'ADN, deux de ces enzymes se distinguent du groupe [14]. La gyrase bactérienne permet un surenroulement négatif et la reverse gyrase est capable d'effectuer un surenroulement positif de l'ADN. Le surenroulement permet de changer l'enlacement de l'ADN. L'enlacement est le nombre de tours contenus dans une molécule d'ADN en conformation de type B relaxé [49]. Le surenroulement se définit comme la structure tertiaire enroulée qui se forme lorsqu'une tension est placée sur une hélice d'ADN par la surtension ou la sous-tension de l'hélice [72]. De l'ADN surtensionné forme un surenroulement positif alors qu'une sous-tension forme un surenroulement négatif [72]. Bien que la nomenclature

des gyrases partage la même origine, il n'y a pas de relation entre leurs structures et leurs mécanismes [71].

La reverse gyrase est une grande protéine composée de 1085 à 1376 acides aminés [30]. Cette enzyme est la seule connue à ce jour à avoir la propriété de permettre un surenroulement positif dans l'ADN circulaire (figure 2.4). Ceci permet de prévenir un désenroulement local de la double hélice à très haute température permettant ainsi de protéger le génome d'une dénaturation [30, 75].

Les différentes reverses gyrases qui ont été caractérisées jusqu'à maintenant montrent toutes des similarités dans leurs comportements et leurs exigences ioniques pour leur activité enzymatique [51]. La reverse gyrase est retrouvée autant chez les archaea que chez les bactéries [37]. Cette répartition commune suggère que la reverse gyrase provient d'un ancêtre commun avant son point de divergence [37]. La reverse gyrase serait formée par la fusion d'un module aminoterminal appartenant à la superfamille des helicases, liant l'ATP, et d'un module carboxyle-terminal d'un ADN topoisomérase IA (type I, classe A), qui relaxe l'ADN [30, 37]. La topoisomérase de type I et l'hélicase formant la reverse gyrase sont membre de deux superfamilles distinctes [37]. Ces deux superfamilles ont deux différentes protéines ancestrales, ce qui implique que les domaines de la reverse gyrase ont évolué séparément l'un de l'autre avant d'être joints [37]. La reverse gyrase et la topoisomérase IA retrouvées chez les bactéries mésophiles possèdent plusieurs particularités communes. Elles requièrent toutes deux du Mg^{2+}, elles sont inhibées par l'ADN simple brin, elles font un lien transitoire covalent avec le groupe ADN brisé phosphorylé 5' et elles clivent préférablement l'ADN sur une séquence 5'-CNNN ↑-3' [51]. La reverse gyrase est donc classée comme étant de la classe A des topoisomérases de type I [51].

Des études ont montré la présence d'un pseudogène de la reverse gyrase chez la bactérie thermophile *Thermus thermophilus*. Ceci suggère que cette enzyme est originaire des espèces thermophiles pour ensuite être intégrée chez les hyperthermophiles [37, 58]. Il semblerait, par l'analyse de la répartition de la reverse gyrase dans divers clades, qu'elle ait évolué séparément entre les bactéries et les archaea hyperthermophiles. Il est possible d'observer, avec un degré de certitude élevé, un clade représentant les bac-

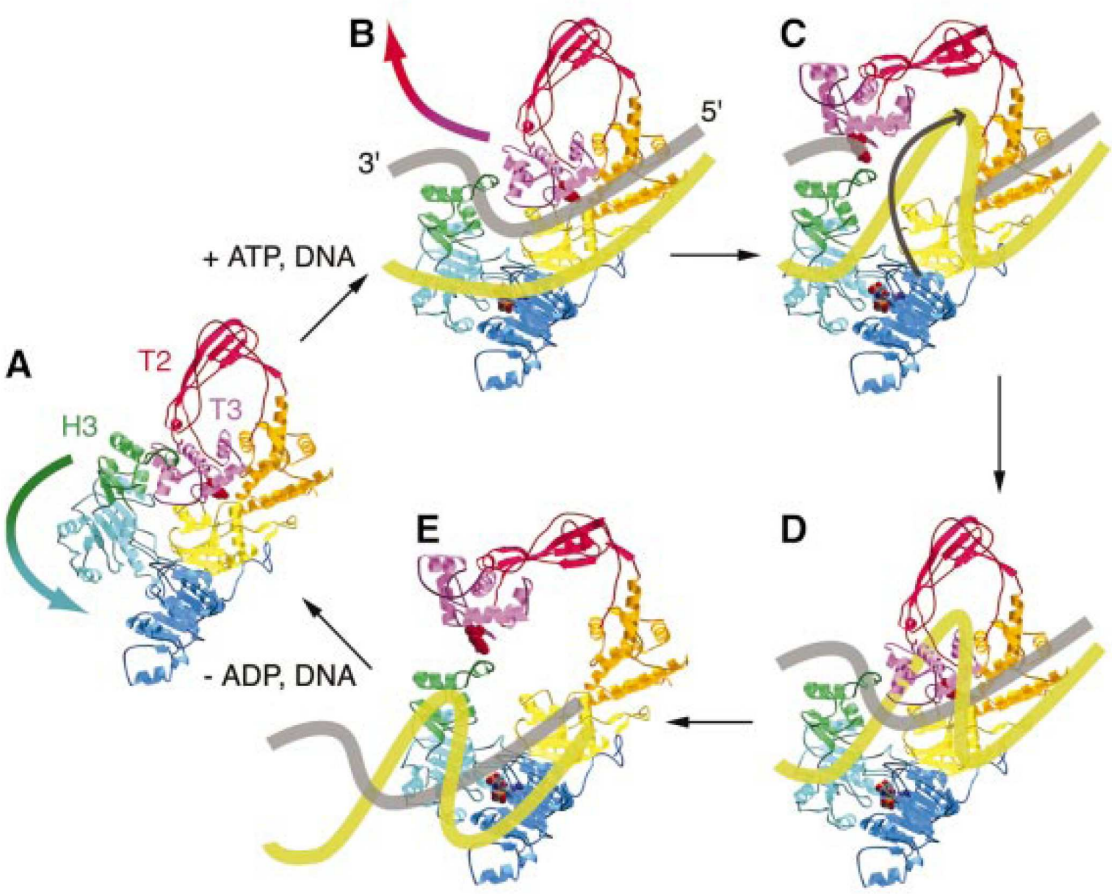


Figure 2.4 – Modèle hypothétique représentant le surenroulement positif de l'ADN par la reverse gyrase. (A) Déclanchement de la fermeture du clivage entre les sous-domaines H1 et H2, tirant H3 loin du sillon T2/T3. (B) Après le clivage de l'ADN, le sillon s'ouvre et le brin non clivé passe au travers la cavité centrale. (C) Le sillon de la topoisomérase se ferme et le bris du brin clivé est lié. (D) Finalement, le sillon s'ouvre une deuxième fois afin de relâcher l'ADN produit. (E) L'ADN contient maintenant un tour supplémentaire. Les différentes couleurs représentent les différents sous-domaines de la protéine. Le jaune et le gris représentent les deux brins de l'ADN [75].

téries et deux autres clades distincts pour les archaea [37]. Selon Heine et coll., la reverse gyrase n'aurait donc pas uniquement évolué de façon indépendante entre les bactéries et les archaea, mais aussi à l'intérieur même des archaea [37]. Malgré tout, la séquence protéique de la reverse gyrase est hautement conservée entre hyperthermophiles [37].

La présence de la reverse gyrase est nécessaire, mais elle n'est pas suffisante à elle

seule pour avoir permis une adaptation aux grandes chaleurs [30]. D'autres traits spécifiques à l'adaptation thermique ont été détectés [30]. Le nombre d'acides aminés chargés formant les protéines et la présence de lipides tétraéther ont sûrement eu une influence importante sur la survie d'une espèce dans un milieu hyperthermique [30]. Par contre, ces autres traits sont aussi présents chez d'autres espèces que les hyperthermophiles, ce qui diminue l'intérêt porté à leur égard dans une recherche sur les hyperthermophiles puisqu'elles y sont moins spécifiques. De plus, le besoin indispensable de la reverse gyrase pour les hyperthermophiles est encore un sujet de controverse puisqu'il a été démontré que l'inactivation de cette enzyme chez l'archaea *Thermococcus kodakaraensis* n'est pas létale, même si sa croissance est grandement réduite [63].

L'utilisation du profil de la reverse gyrase a pour but de trouver d'autres protéines interagissant avec celle-ci ainsi que de rechercher parmi des paires de gènes dans des organismes de phénotype hyperthermophile, sans qu'ils aient pour autant une interaction directe avec la reverse gyrase.

2.5 Déplacement de gènes non orthologues

La distribution de gènes, qui peut être observée dans un profil phylétique, est grandement affectée par des événements évolutionnaires tels que la redondance de gènes, la perte de gènes, les transferts horizontaux ou les déplacements de gènes non orthologues (DGNO) [32].

Koonin et coll. ont été les premiers à définir les déplacements de gènes non orthologues, tels que le remplacement fonctionnel d'une protéine donnée chez une espèce par une protéine paralogue ou *a priori* non homologue, mais possédant une activité enzymatique équivalente [32, 46].

Le cas le plus caractéristique de la détection d'un événement de DGNO est celui de deux classes de tRNA-lysyl-synthétases non homologues. Dans ce cas, leurs profils phylétiques (tableau 2.I) sont quasi parfaitement complémentaires avec un score de complémentarité selon l'équation 1.1 de 0,88. Leur découverte a été faite avant l'arrivée de la comparaison génomique à grande échelle telle que connue aujourd'hui. Toutefois,

en analysant de façon rétrospective leur profil, il est possible de leur conférer un lien fonctionnel.

La fructose-1,6-bisphosphate aldolase (FBA) catalyse une étape essentielle de la glycolyse et est présente dans la plupart des espèces de bactéries, mais est absente chez les archaea et la bactérie *Chlamydia* sp. (tableau 2.I). Les archaea et la bactérie *Chlamydia* sp. ont une enzyme FBA de type DhnA. Ces deux aldolases distinctes ont un profil complémentaire avec un score de 0,76. Il est donc possible de dire que l'enzyme de type DhnA fonctionne comme unique fructose-1,6-bisphosphate aldolase chez les archaea et *Chlamydia* sp.. Dans ce cas-ci, cette hypothèse est soutenue par la démonstration que la protéine DhnA chez *E. coli* a une activité enzymatique de fructose-1,6-bisphosphate aldolase [32].

En dernière analyse du tableau 2.I, l'étude de la complémentarité d'une protéine de la thymidylate synthase, une enzyme précurseure de la biosynthèse d'ADN et d'une protéine à l'activité enzymatique jusque-là inconnue possédant un score de complémentarité de 0,88, permet la prédiction de la fonctionnalité enzymatique de cette nouvelle protéine. Des études subséquentes par analyse d'alignements multiples montrent que ces protéines ont des régions conservées compatibles avec ce type d'activité. De plus, il a été observé que l'homologue de cette protéine, au rôle nouvellement défini, chez *Dictyostelium* sp. permet de contrecarrer une déficience de la thymidylate synthase [21, 32].

Dans le cadre d'études sur des profils phylétiques recherchant des DGNO, l'hypothèse qu'il est peu probable qu'un tel déplacement augmente le nombre de gènes, ou composantes, dans un système biologique peut être émise. Par contre, le remplacement de plusieurs composantes par une seule est facile à concevoir, si celle-ci peut effectuer la même tâche de façon semblable, voire plus efficacement. Ceci est visible dans le cas de l'ARN polymérase des mitochondries et des chloroplastes dans lequel les quatre sous-unités de l'ARN polymérases ont été déplacées et remplacées par une seule unité monomérique [29, 34]. À l'exception du noyau de la machinerie traductionnelle, de quelques ARN polymérases et de quelques chaperones, il n'y a pas de systèmes composés de gènes ubiquitaires. De plus, la plupart de ceux-ci ne sont pas caractérisés par un profil phylétique singulier. Même à l'intérieur d'une voie métabolique d'importance

capitale au niveau biochimique, telle que la glycolyse ou le cycle de Krebs, de nombreuses variations sont observées. Ces variations sont le résultat soit d'une modification de la voie métabolique elle-même soit d'un DGNO [32].

Ce type de déplacement est tout de même estimé comme étant un événement rare pour les protéines impliquées dans les interactions multiprotéiques. Cependant, l'occurrence de ce genre de déplacement a bien été documentée par des études antérieures sur les ARN polymérases protéobactériennes originales remplacées par une ARN polymérase de bactériophage dans l'évolution de la mitochondrie et du chloroplaste [13, 29, 34]. Les DGNO pourraient être bien plus présents dans l'évolution que décrits par les théories originales. Ce phénomène pourrait être une explication raisonnable, entre autres, de la cause des différences chez les facteurs d'initiation de transcription interagissant avec des ARN polymérases homologues chez les bactéries et les archaea [29].

Une étude sur des organismes de très petite taille a décrit qu'il y a au moins 11 gènes non orthologues possédant la même fonction, chez *H. inflenzae* et *M. genitalium*, et ce, sur une reconstruction des 256 gènes orthologues constituant leur ensemble absolu minimal [45, 47, 62]. Au sein d'espèces possédant un génome plus large, il est possible de croire que le nombre de gènes non orthologues, possédant la même fonction, peut être considérablement plus important et pourrait même se chiffrer à quelques centaines d'exemplaires [46]. Les études faites sur ces organismes de plus grande taille génomique ont montré que les DGNO avaient influencé les gènes souvent les plus essentiels, tels que ceux responsables de la traduction, la transcription et, de façon encore plus importante, la réplication [47].

Tableau 2.I – Profil phylétique de trois catégories de gènes pour une panoplie d'espèces montrant des patrons complémentaires suggérant un déplacement de gène non orthologue et permettant une prédiction de certaines fonctions protéiques [32].

	espèces																	
	Arch				Eury				Bact									
	Af	Mj	Mth	Ph	Sc	Aa	Tm	Ssp	Ec	Bs	Mt	Hi	Hp	Mgp	Bb	Tp	Ctp	Rp
voie/ Enzyme																		
Transduction																		
Class II lysyl-tRNA synthetase (COG1190)	-	-	-	-	+	+	+	+	+	+	+	+	+	+	-	+	+	-
Class I lysyl-tRNA synthetase (COG1384)	+	+	+	+	-	-	-	-	-	-	-	-	-	-	+	+	-	+
Glycolyse																		
FBA (COG0191)	-	-	-	-	+	+	+	+	+	+	+	+	+	+	+	+	-	-
DhnA-type FBA (COG1830)	+	+	+	+	-	+	-	-	+	-	-	-	-	-	-	-	+	-
Biosynthèse Thymidylate																		
Thymidylate synthase (COG0207)	+	+	+	-	+	-	-	-	+	+	+	+	-	+	-	-	-	-
novel Thymidylate synthase (COG1531)	-	-	-	+	-	+	+	+	-	-	+	-	+	-	+	+	+	+

Arch=Archaea ; Eury=Eucaryote ; Bact=Bactérie
Aa, *Aquifex aeolicus*, Af, *Archaeoglobus fulgidus*, Bb, *Borrelia burgdorferi*, Bs, *Bacillus subtilis*, Ctp, *Chlamydia trachomatis & Chlamydia pneumoniae*, Ec, *Escherichia coli*, Hi, *Haemophilus influenzae*, Hp, *Helicobacter pylori*, Mgp, *Mycoplasma genitalium & Mycoplasma pneumoniae*, Mj, *Methanococcus jannaschii*, Mt, *Mycobacterium tuberculosis*, Mth,*Methanobacterium thermoautotrophicum*, Ph, *Pyrococcus horikoshii*, Rp, *Rickettsia prowazekii*, Sc, *Saccharomyces cerevisiae*, Ssp, *Synechocystis* sp., Tm, *Thermotoga maritima*, Tp, *Treponema pallidum*. +=présence ;-=absence.

CHAPITRE 3

INFÉRENCE DE GROUPE D'HOMOLOGUES

3.1 Introduction

Afin de pouvoir espérer être en mesure de statuer sur les relations possibles de coévolution conditionnelle, il faut, dans un premier temps, posséder des données qui soient les plus complètes possibles. Puisqu'il existe plusieurs tactiques permettant de regrouper les gènes en fonction de la relation d'orthologie, une attention particulière doit être faite pour procéder par une méthode appropriée au contexte de l'étude.

La base de la recherche sur le regroupement d'homologues et la description de la méthodologie ayant mené à la base de données utilisée, ainsi que les raisons qui ont poussé à l'utilisation de celle-ci en particulier, sont décrites à la section 3.2.

La section 3.3 propose une structure de données efficace afin de regrouper des homologues entre eux selon l'approche naïve par similarité de séquence réciproque. Elle se conclut par une brève discussion émanant de la comparaison entre l'approche implémentée et celle qui fut finalement utilisée.

3.2 Distinction d'homologues

Les études sur les comparaisons de génomes sont rapidement devenues pratique courante. Émanant de ces recherches, une quantité d'information d'une taille considérable a été produite. À l'instant où la comparaison de génomes est devenue une tâche importante, plusieurs questions ont été soulevées quant à la façon de définir l'ensemble des gènes comparables d'une espèce à l'autre. La caractérisation individuelle de chaque gène dans chaque génome se transforme rapidement en une tâche utopique, même pour une analyse informatisée, et devient irréalisable de façon expérimentale au fur et à mesure que de nouveaux génomes sont séquencés et additionnés à ceux qui sont existants. La génomique comparative n'est possible et valable que si le nombre d'entités distinctes à analyser est réduit par l'introduction d'une classification rationnelle des

gènes. Une façon naturelle d'effectuer cette systématisation est de délimiter les gènes en ensembles d'orthologues, en y incluant les coorthologues [48, 89].

La mesure avec laquelle une telle classification peut être appliquée à l'analyse génomique dépend essentiellement de la nature des relations entre les génomes d'espèces différentes. En conséquence, en prenant un cas extrême, si tous les gènes de génomes comparés formaient des groupes parfaits d'orthologues en relation un pour un, le nombre d'entités à l'étude serait égal au nombre de gènes dans chaque génome et resterait constant après l'addition de génomes nouvellement séquencés. Si tel était le cas, toute l'entreprise comparative de la génomique évolutive serait d'une telle simplicité qu'elle serait futile. Par contraste, si le nombre de regroupements d'orthologues est nettement inférieur par rapport au nombre de gènes composant les génomes comparés, la génomique comparative représenterait une tâche ardue [48].

Les études actuelles travaillant à l'échelle génomique usent de simplifications et de raccourcis. Ceux-ci font souvent intervenir l'hypothèse élémentaire, mais essentielle, que les séquences de gènes orthologues doivent être plus similaires entre membres d'un même groupe qu'ils ne le sont à tout autre gène inclus dans les génomes des espèces à l'étude [48, 89]. Ce type de liaison forme ce qui est convenu d'appeler des meilleures ressemblances symétriques, ou symbets pour « symetrical best hits » [48]. Les notions de ressemblance symétrique sont schématisées à la figure 3.1. Il peut être supposé que les symbets sont plus susceptibles d'être formés par des gènes orthologues, ce qui suggère un moyen simple permettant l'identification d'orthologues. En s'appuyant sur la définition formelle d'un gène orthologue et sur la notion que ces orthologues possèdent généralement les mêmes fonctions, l'hypothèse que les gènes partageant les meilleures ressemblances symétriques représentent des gènes orthologues est d'autant plus plausible, à tout le moins statistiquement [48]. Le regroupement d'orthologues selon les principes des symbets est sans doute la méthode la plus simple pour les études sur les génomes d'espèces apparentées. Par contre, cette méthode peut aussi bien servir à une étude sur de plus grandes distances évolutives dans un but de détection de gènes orthologues ayant une correspondance d'un pour un. Il faut cependant noter que cette méthode n'est pas parfaite et ne donne pas toujours le meilleur résultat [66]. L'utilisation exclusive

de symbets résultera certainement dans l'omission d'orthologues [66]. Les inparalogues et les xénologues compliquent la distinction d'orthologues par les symbets [66]. Cette méthode de ressemblance symétrique a été implémentée, pour le présent document, dans le but de regrouper les orthologues. Cette méthode a permis de mettre en place une base de données personnelle. Cette base de données a permis la composition d'arbres phylogénétiques de gènes pouvant être utilisés pour les analyses d'évolution conditionnelle corrélée. L'ajout de cette méthode à celles existantes, telles que les COGs, définis ci-après, et les symbets, a permis de faire une brève analyse comparative des méthodes de regroupement des orthologues.

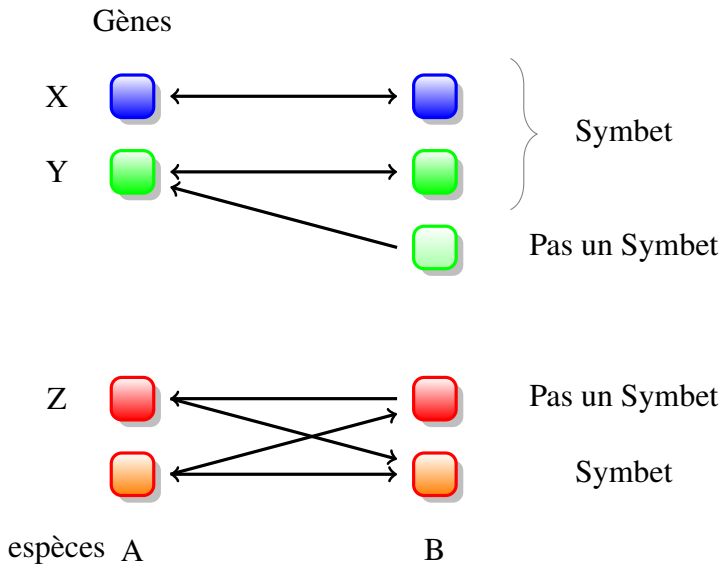


Figure 3.1 – Schématisation de la relation entre orthologues démontrant la notion de meilleure ressemblance symétrique [48]. Les flèches représentent les meilleures ressemblances et les rectangles arrondis de ton similaire représentent les paralogues. X, Y, Z représentent les trois cas possibles de relation entre orthologues ; X : un pour un ; Y : un pour plusieurs ; Z : plusieurs pour plusieurs.

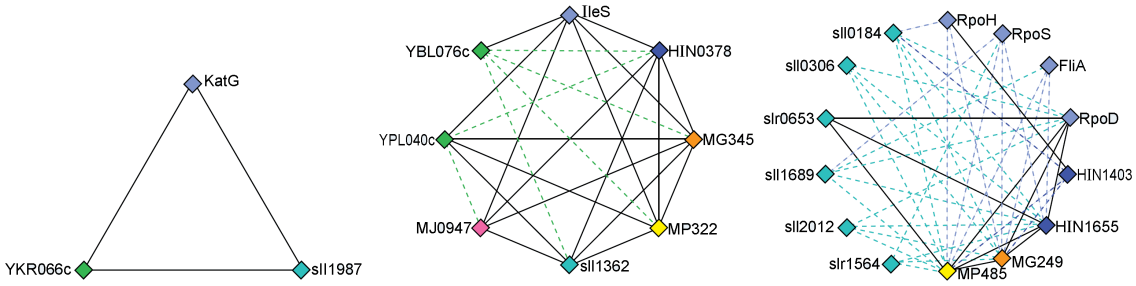
Par contre, pour la création des profils phylétiques indispensables à cette recherche, les dénombrements d'orthologues n'ont pas été effectués par l'approche naïve, mais sont plutôt venus des regroupements issus des « clusters of orthologous group », COGs.

Le concept inhérent aux COGs est de généraliser et d'étendre la notion d'une meilleure

ressemblance à l'intérieur d'un génome [48, 88]. La première étape fut d'éliminer l'exigence de la réciprocité des meilleures ressemblances, comme pour les symbets. Deuxièmement, la notion d'une meilleure ressemblance à l'intérieur d'un génome a été étendue aux génomes multiples de façon à ce que l'algorithme cherche à établir des grappes cohérentes de meilleures ressemblances. Plus précisément, la méthode appliquée pour définir les COGs est basée sur l'hypothèse que trois gènes, ou plus, répertoriés dans des génomes d'espèces relativement éloignées sur le plan évolutif sont plus semblables entre eux qu'ils ne le sont par rapport aux autres gènes de ces mêmes génomes. Ceux-ci sont donc plus susceptibles de faire partie d'un même groupe d'orthologues. Cette prédiction est applicable même dans les cas où les similitudes de séquences comparées sont relativement faibles. En conséquence, les gènes soumis à des pressions sélectives différentes, leur permettant une évolution plus rapide, ou plus lente, peuvent être intégrés dans leur COG approprié [48, 88].

La procédure pour la construction de COG peut se résumer en quatre étapes simples. Premièrement, il faut effectuer une comparaison de type « toutes contre toutes » des séquences protéiques répertoriées des différents génomes. Ceci est couramment fait par le programme BLAST, acronyme qui signifie « Basic Local Alignment Search Tool » [1]. Cet algorithme compare des séquences biologiques primaires afin de détecter des similarités entre une séquence requête et des séquences répertoriées dans une base de données [1, 48]. Ensuite, il faut détecter et regrouper les in-paralogues évidents, ceux-ci sont visibles du fait que ces gènes montrent plus de similarité entre eux tout en provenant d'un même génome qu'en comparaison aux gènes d'autres espèces. La troisième étape consiste à identifier des triangles (figure 3.2a), puisque la méthode repose sur des groupes d'au moins trois membres de meilleure ressemblance. Les groupes d'in-paralogues détectés à l'étape précédente sont traités comme des entités uniques. Finalement, il reste à fusionner les triangles ayant des éléments communs (figures 3.2b et 3.2c) [48, 88, 89].

Les regroupements de gènes en COGs ne sont pas sans erreur [18]. Par contre, des travaux approfondis avec ce type de base de données pour l'annotation de génomes ainsi que pour les études sur l'évolution suggèrent qu'ils sont suffisamment robustes pour inférer des motifs évolutifs et fonctionnels significatifs [48, 64].



(a) Triangle de meilleure ressemblance

(b) COG simple

(c) COG complexe

Figure 3.2 – Étapes pour la création de COGs. Les lignes solides représentent les meilleurs ressemblances symétriques. Les lignes pointillées représentent les meilleurs ressemblances asymétriques, où les couleurs des lignes correspondent à l'espèce à partir de laquelle la ressemblance est observée. Les gènes d'une même espèce sont adjacents et de même couleur, sinon, ils sont positionnés de façon arbitraire. A : Gènes montrant une meilleure ressemblance formant un triangle, le minimum pour former un COG. B : Un COG simple avec deux paralogues de levure. C : Un COG complexe avec plusieurs paralogues [89].

Origine des protéines : KatG, *E. coli* ; sll1987, *Synechocystis* sp. ; YKR066c, *S. cerevisiae*, IleS, *E. coli* ; HIN0378, *H. influenzae* ; MG345, *M. genitalium* ; MP322, *M. pneumoniae* ; sll1362, *Synechocystis* sp ; MJ0947, *M. jannaschii* ; YBL076c et YPL040c, *S. cerevisiae*. RpoH, RpoS, RpoD, et FliA, *E. coli* ; HIN1403 et HIN1655, *H. influenzae* ; MG249, *M. genitalium* ; MP485, *M. pneumoniae* ; sll0184, sll0306, slr0653, sll1689, sll2012, et slr1564, *Synechocystis* sp.

3.3 Association d'orthologues

Dans le but de créer une base de données personnelle d'orthologues, un script permettant de regrouper des gènes entre eux selon des principes de similarité de séquences réciproques a été implémenté. Celui-ci prend comme données un fichier d'alignements produit par BLAST et partage ensuite les séquences en groupe d'orthologues. Cet algorithme prend deux tableaux, un pour le BLAST recto (requête de l'espèce A et base de données de l'espèce B) et un pour le verso (requête de l'espèce B et base de données de l'espèce A) selon des valeurs de seuil arbitraires déterminées par l'usager. Ensuite, il les regroupe selon une structure de données d'appartenance-union, ou « union-find ».

L'algorithme du script regroupant des gènes considère un ensemble d'éléments et le partitionne en un certain nombre de classes disjointes, de telle sorte que chaque élément est dans un seul sous-groupe, ou arbre. Ensemble, ces arbres représentent une re-

lation d'équivalence possédant un membre commun comme étiquette. Cette méthode permet de gérer efficacement une table de hachage pour pallier au problème de connexité. Une telle structure de données pour le problème des classes disjointes maintient la répartition de l'ensemble des éléments en utilisant trois opérations soit *faire* (make-set), *appartenance* (find) et *unir* (union). *Faire* construit en un premier temps une classe d'équivalence contenant uniquement le singleton de l'élément en question. Ceci est fait en pointant l'élément vers lui-même. *Appartenance* détermine la classe d'équivalence d'un élément et si deux éléments appartiennent à une même classe. Étant donné un objet, l'ensemble auquel il appartient est trouvé et l'étiquette de cet ensemble est retournée. *Unir* réunit deux classes d'équivalence en une seule. Étant donné deux objets, les deux ensembles correspondants sont unifiés et une étiquette pour l'ensemble combiné est choisie. La méthode utilisée afin de regrouper entre eux des gènes est celle de l'union rapide équilibrée [7]. Cette méthode se base sur la méthode union rapide. Celle-ci est toujours fondée sur une table indexée par des noms d'objets. Chaque objet a un lien vers un autre objet dans le même ensemble, dans une structure sans cycle. Pour déterminer si deux objets sont dans le même ensemble, l'algorithme suit leurs liens respectifs jusqu'à ce qu'ils atteignent un objet ayant un lien avec lui-même. Ces objets sont considérés comme appartenant au même ensemble s'ils sont connectés à un seul et même objet. Ces objets sont considérés comme n'appartenant pas au même ensemble s'ils sont connectés à des objets différents ayant des liens avec eux-mêmes. Pour réaliser l'union, il suffit de lier un objet à un autre, d'où le nom union rapide. Pour pallier au cas défavorable où l'arbre le plus gros est lié au plus petit, une simple modification de l'algorithme permet de s'assurer que le plus petit est toujours lu en conservant un tableau supplémentaire pour la taille des sous-arbres. Ceci empêche l'extension de longs parcours dans l'arbre et assure que la hauteur de l'arbre reste $O(log\ n)$. C'est pour cela qu'il est appelé union rapide équilibrée. Si chaque consultation ou modification d'un élément du tableau compte comme une opération élémentaire, le temps nécessaire pour exécuter une opération est $O(log\ n)$ quand il y a n objets [7, 80].

L'implémentation d'un script permettant de regrouper en tableau des séquences montrant une réciprocité bidirectionnelle pour ensuite les regrouper en ensemble d'ortho-

logues a été effectuée. La figure 3.3 montre le pseudo-code de l'algorithme appartenance-union. La figure 3.4 montre le pseudo-code prenant deux séquences retenues par BLAST selon des critères de score maximal et de e-valeur pour chaque séquence requête et leur regroupement par appartenance-union. Le score maximal est attribué entre la requête et la meilleure séquence associée selon une matrice de comparaison. Plusieurs séquences peuvent être associées à une même requête et ont donc différents scores inférieurs à la meilleure paire. Un pourcentage arbitraire de ce meilleur score est donc le premier seuil considéré. La e-valeur, ou valeur attendue (e pour « expected ») est un paramètre décrivant le nombre de succès attendus pouvant être obtenus par chance. Plus cette e-valeur est proche de 0, plus significative est l'identification. Ceci est donc le deuxième seuil pris en compte.

La base de données d'orthologues obtenue par l'approche naïve telle qu'implémentée ne peut pas être comparée d'égale à égale à la base de données arCOG, une base de données similaire au COG décrit plus tôt, mais spécifique aux archaea, puisque l'approche naïve se base sur des variables subjectives comme seuil d'acceptabilité des similarités obtenues par BLAST [57]. Ce seuil, bien que présent, est inconnu pour arCOG. Des valeurs différentes de seuil pour le pourcentage du score maximum et la e-valeur acceptable donnent des regroupements qui peuvent différer grandement. Par exemple, la meilleure concordance a été observée lorsqu'un pourcentage de 90 % du score maximum et une e-valeur minimale de $1x10^{-12}$ sont utilisés dans l'approche naïve. Dans ce cas, 85 % des regroupements d'orthologues, identifiés par l'approche naïve et l'approche des arCOG, sont identiques. Ce nombre varie grandement, par contre, lorsqu'un pourcentage du score maximal moins élevé est utilisé. Les regroupements diffèrent généralement de quelques séquences seulement. Ces séquences, selon l'analyse de fonction, sont difficiles à classer parmi un des deux groupes de séquences différentes ayant les mêmes fonctions. En fonction des paramètres de e-valeur et de pourcentage utilisés, les séquences traitées peuvent être considérées comme similaires ou différentes et ainsi former un ou plusieurs groupes. Ceci n'est pas uniquement le cas ici, mais des différences similaires entre les bases de données COG et arCOG sont visibles. Les séquences formant les groupes arCOG00001 à 00008, définis comme « régulateur transcriptionnel

Algorithme : Appartenance-Union

```
Procédure : MakeSet
Entrée : Id (ι) d'un gène
Sortie : Rien
M1    id{ι} = {"group" ⇒ ι, "size" ⇒ 1}
```

```
Procédure : Union (a,b)
Entrée : Prend deux gènes
Sortie : Rien
U1    ι ← Find(a)
      ζ ← Find(b)
U2    Si ι != ζ
U3        Si id {ι}{"size"} < id{ζ}{"size"}
U4            id {ι}{"group"} = ζ
              id {ζ}{"size"} = id {ι}{"size"} + id{ζ}{"size"}
U6        Sinon
U7            id {ζ}{"group"} = ι
              id {ι}{"size"} = id{ι}{"size"} + id{ζ}{"size"}
```

```
Procédure : Find
Entrée : Id (ι) d'un gène
Sortie : ι
F1    while(ι != id{ι}{"group"})
F2        ι = id{ι}{"group"}
F3    return ι
```

Figure 3.3 – Pseudo-code de l'algorithme appartenance-union pour la gestion d'une table de hachage.

prédit ; famille PadR », sont regroupées sous le même COG01695 « régulateur transcriptionnel prédit ; famille PadR » [57, 88]. Même dans un cas où l'inférence d'orthologues est parfaite, il peut tout de même y avoir des différences justifiables entre COG, arCOG et l'approche naïve. Une famille COG peut être coupé en plusieurs familles arCOG si l'ancêtre des archaea possède plusieurs membres de la famille COG d'où descendent les familles arCOG. La relation d'orthologies nécessite la spécification d'un ancêtre de référence. Ceci démontre la complexité de regrouper efficacement des séquences par orthologies puisqu'il est difficile d'établir des liens robustes entre les gènes.

Algorithme : Groupement par réciprocité bi-directionnelle
Entrée : Un ensemble, H, d'alignements BLAST (query, hit, pct, e-value)
et une structure de données appartenance-union, G, pour les GIs des gènes.
Query et hit sont des GIs de gènes d'un génome A et d'un autre génome B
(query=A, hit=B ou query=B, hit=A).
Pct est le pourcentage d'identité de l'alignement retrouvé et
e-value est l'e-valeur de l'alignement.
Sortie : Liste d'homologues, G, regroupés par GIs

R1 **Suprimer tout** x $\in$ H avec x.pct < min_pct ou x.e-value > max_e-value et
y $\in$ H avec y.pct < min_pct ou y.e-value > max_e-value
$\backslash*$ borne supérieure sur evalue, inférieure sur pct $*\backslash$
R2 **Pour tout** x.query $\in$ H
R3 **Si** $\exists$ y $\in$ H où y.hit=x.query **et** y.query=x.hit **Alors** $\backslash*$ symbets $*\backslash$
R4 **Si** G.FIND(x.query)=null **Alors** G.MAKE_SET(x.query)
R5 **Si** G.FIND(x.hit)=null **Alors** G.MAKE_SET(x.hit)
R6 G.UNION(x.query, x.hit)

Figure 3.4 – Pseudo-code regroupant naïvement des séquences homologues selon une réciprocité bi-directionnelle par une méthode appartenance-union. *Pct* représente le score minimal selon un pourcentage du score maximum et *e-val* une e-valeur minimale pour retenir deux séquences comme similaires. X et Y représentent respectivement une paire (query, hit) pour le BLAST de A vers B et de B vers A.

CHAPITRE 4

MODÈLE PROBABILISTE DE PROFILS PHYLÉTIQUES

4.1 Introduction

Ce chapitre s'intéresse aux étapes nécessaires, préexistantes et originales, afin de modéliser la corrélation conditionnelle entre deux ou trois profils phylétiques.

La modélisation évolutionnaire probabiliste a déjà été élaborée pour deux caractères discrets. Celle-ci avait permis d'énoncer des hypothèses quant à un rituel d'accouplement chez diverses espèces de singes [68].

Le chapitre détaille les modèles mathématiques en ordre croissant de complexité. La section 4.2 montre la base du modèle en utilisant un seul profil. La section 4.3 présente une méthode mathématique pour l'analyse de la relation entre deux caractères discrets. La section 4.4 instaure les lemmes et équations mathématiques pour les modèles d'évolution conditionnelle corrélée. La section 4.5 montre un algorithme efficace pour résoudre la vraisemblance. La section 4.6 discute de l'approche par maximum de vraisemblance pour établir une estimation de la phylogénie à partir d'un profil. La section 4.7 détaille les bases permettant d'optimiser les taux de mutation, soit ceux de gains et de pertes des familles des gènes X, Y et Y' du profil phylétique assurant une plus grande confiance face aux résultats. Finalement, la section 4.8 discute de la sélection du modèle par le test du χ^2.

4.2 Modèle pour un profil

Considérons une variable pouvant prendre deux états, 0 et 1. L'idée derrière les équations suivantes (4.1 à 4.8) est de caractériser les probabilités dans une branche phylogénique débutant à l'état 0 et où la variable restera la même ou changera à l'état 1, sur un espace de temps arbitraire t, avec les probabilités $P_{00}(t)$ et $P_{01}(t)$. Ces équations décrirons aussi les cas $P_{11}(t)$ et $P_{10}(t)$ représentant respectivement les cas où un embranchement débute et se termine à l'état 1 et change de l'état 1 à l'état 0 sur un temps

t. Si les deux variables changent de façon indépendante l'une par rapport à l'autre, leurs probabilités conjointes de changement sont obtenues par le produit de leurs probabilités distinctes.

Les variables utilisées doivent être déclarées comme évoluant selon un processus de Markov qui sous-tend que la probabilité de changement est indépendante dans chaque branche de l'arbre phylogénétique et que la probabilité de passer d'un état quelconque à un autre ne dépend que de l'état au début de la branche en question, et non pas d'événements antérieurs.

La probabilité que les variables changent de l'état i vers j, définie comme le taux de transition instantanée q_{ij}, sur un intervalle de temps $t + \mathrm{d}t$ est donnée par

$$P_{ij}(t+\mathrm{d}t) = P_{ii}(t)q_{ij}\,\mathrm{d}t + P_{ij}(t)(1-q_{ij})\,\mathrm{d}t \tag{4.1}$$

Le terme $1 - q_{ji}$ représente la probabilité de rester dans l'état j, soit la probabilité de ne pas changer de l'état j vers i. L'équation 4.1 décrit donc qu'un caractère peut changer d'un état i vers j soit en restant inchangé pour une période de temps t suivi d'une transition vers l'état j ou en changeant vers l'état j durant la période de temps t pour conserver l'état j sur une période $\mathrm{d}t$.

Les équations décrivant les quatre probabilités possibles peuvent être représentées dans une matrice sous la forme d'une équation Chapman-Kolmogorov. L'équation Chapman-Kolmogorov est une identité concernant les distributions de probabilité conjointe de différents jeux de coordonnées sur un processus stochastique.

Lemme 1. *Chapman-Kolmogorov*

$$\mathbf{P}(t+\mathrm{d}t) = \mathbf{P}(t)(\mathbf{I}+\mathbf{Q}\,\mathrm{d}t) \tag{4.2}$$

$$\begin{bmatrix} 1-P_{01}(t+\mathrm{d}t) & P_{01}(t+\mathrm{d}t) \\ P_{10}(t+\mathrm{d}t) & 1-P_{10}(t+\mathrm{d}t) \end{bmatrix} = \begin{bmatrix} P_{00}(t) & P_{01}(t) \\ P_{10}(t) & P_{11}(t) \end{bmatrix} \begin{bmatrix} (1-q_{01})\,\mathrm{d}t & q_{01}\mathrm{d}t \\ q_{10}\mathrm{d}t & (1-q_{10})\,\mathrm{d}t \end{bmatrix}$$

Le lemme 1 est la propriété fondamentale d'un processus de Markov à deux états. Pour résoudre $\mathbf{P}(t)$ en termes des paramètres de $\mathbf{Q}$ il faut en premier lieu résoudre

$$\frac{\mathrm{d}\mathbf{P}(t)}{\mathrm{d}t} = \frac{\mathbf{P}(t+\mathrm{d}t) - \mathbf{P}(t)}{\mathrm{d}t} = \mathbf{P}(t)\mathbf{Q} \tag{4.3}$$

d'où provient

$$\mathbf{P}(t) = exp\,[\mathbf{Q}t] + c \tag{4.4}$$

où c est la constante d'intégration.

Puisque $\mathbf{P}(0) = \mathbf{I}$, $c = 0$ et donc :

$$\mathbf{P}(t) = exp\,[\mathbf{Q}t] \tag{4.5}$$

Pour compléter l'équation matricielle $exp[\mathbf{Q}t]$, il est nécessaire de convertir l'équation en considérant le fait que

$$exp[\mathbf{Q}t] = \mathbf{C}exp[\mathbf{D}t]\mathbf{C}^{-1} \tag{4.6}$$

où $\mathbf{C}$ contient les vecteurs propres de $\mathbf{Q}$ et $\mathbf{D}$ est une matrice diagonale contenant les valeurs propres de $\mathbf{Q}$. La valeur de t est constante pour un $P(t)$ donné et est prise de la branche où $P(t)$ s'applique. Donc :

$$\mathbf{P}(t) = \mathbf{C}\begin{bmatrix} e^{\lambda_1 t} & 0 \\ 0 & e^{\lambda_2 t} \end{bmatrix}\mathbf{C}^{-1} \tag{4.7}$$

où les λ_i sont les valeurs propres de $\mathbf{Q}$. Comme $\mathbf{P}$ est une matrice stochastique, la somme des valeurs de chaque rangée est égale à 1. $\mathbf{Q}$ a forcément la forme :

$$\mathbf{Q} = \begin{bmatrix} -\lambda & \lambda \\ \mu & -\mu \end{bmatrix} \tag{4.8}$$

avec λ et μ comme taux de gain et de perte respectivment. Les valeurs propres sont $\lambda_1 = 0, \lambda_2 = -(\lambda + \mu)$. Il en résulte donc :

$$\mathbf{P}(t) = e^{Qt} = exp\left(\begin{pmatrix} -\lambda & \lambda \\ \mu & -\mu \end{pmatrix} t\right) = \begin{pmatrix} \frac{\mu}{\alpha} + \frac{\lambda}{\alpha}e^{-\alpha t} & \frac{\lambda}{\alpha}(1-e^{-\alpha t}) \\ \frac{\mu}{\alpha}(1-e^{-\alpha t}) & \frac{\lambda}{\alpha} + \frac{\mu}{\alpha}e^{-\alpha t} \end{pmatrix} \quad (4.9)$$

où $\alpha = \lambda + \mu$.

L'équation Chapman-Kolmogorov permet donc une solution pour $\mathbf{P}(t)$ uniquement en termes des paramètres de $\mathbf{Q}$ et de la longueur de la branche, t. Par exemple, $P_{01}(t)$ sera une fonction du taux de transition de 0 à 1, donc si q_{01} est grand relativement à q_{10}, plus la probabilité est grande qu'à la fin du temps t, le caractère commençant à l'état 0 sera à l'état 1.

4.3 Modèle d'évolution corrélée entre deux caractères

Pagel a présenté une méthode mathématique destinée à l'analyse de la relation entre deux caractères discrets, X et Y, mesurée parmi des espèces d'après leur arbre phylogénétique [68]. Cette méthode évalue si un patron de profil phylétique établit une évidence pour une évolution corrélée entre deux caractères. Elle tient compte de la longueur des branches de la phylogénie en établissant des estimations pour les taux de variation des caractères. Finalement, elle effectue les tests pour discriminer entre les hypothèses d'évolution corrélée et indépendante sans avoir besoin de reconstruction des états du caractère ancestral [68].

La méthode développée utilise un modèle de Markov pour caractériser les changements évolutifs le long de chacune des branches d'un arbre phylogénétique. Ce modèle est similaire à ceux précédemment employés par Jukes et Cantor (1969), Kimura (1980), Felsenstein (1981), Hasegawa et coll. (1985) et Goldman (1993) dans leurs études de modélisation de substitution d'acides nucléiques [26, 33, 36, 41, 44, 68]. Ce modèle ne n'aborde pas le problème de la construction d'arbres phylogéniques, mais a plutôt comme but d'estimer les taux de transition simultanés de paires de caractères binaires dans une phylogénie donnée. Ensuite, il utilise ces taux pour analyser la corrélation possible dans l'évolution de deux caractères [68].

La méthode prend en considération les valeurs de la longueur des branches. Ainsi, il est possible de détecter l'ordre des changements, même sur une branche montrant un changement simultané. Donc, tous les états possibles des caractères sont utilisés à chaque noeud ancestral. Ceci signifie qu'un test d'hypothèse éventuel est indépendant de toutes méthodes particulières pour assigner les valeurs ancestrales et de l'assignation particulière de ces valeurs [68].

Dans tous les cas, une variable de présence ou d'absence d'un gène peut prendre deux états, 0 ou 1. Deux gènes peuvent donc montrer quatre différents patrons de présence ou d'absence dans chaque espèce. Le diagramme 4.1 montre les liens par des flèches avec leurs paramètres respectifs référant aux taux de transition entre les deux états d'un gène lorsque l'un reste constant.

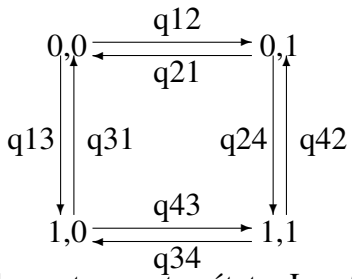


Figure 4.1 – Transitions possibles entre quatre états. Les flèches et leurs paramètres associés représentent les taux de transition entre les deux états alors qu'un d'eux est constant [5].

La méthode teste l'hypothèse d'une évolution corrélée en comparant l'ajustement de deux modèles différents pour les données observées. Premièrement, la méthode ajuste un modèle aux données pour lequel les deux caractères sont indépendants. La qualité de cet ajustement est ensuite comparée à celle d'un autre modèle selon lequel les caractères ont évolué de façon corrélée [68].

4.3.1 Transitions dépendantes avec deux profils

La figure 4.1 montre les taux de transition dans le cas général pour deux profils combinés. Si les changements sont indépendants entre deux caractères, alors les probabilités de transition sont les produits des probabilités des profils individuels. Si le changement

entre deux variables, X et Y, est corrélé, leur changement conjoint ne peut pas être simplement décrit comme le produit de leurs probabilités. Dans cette sous-section, il est question du modèle requis pour calculer ce type de transition où le changement d'une variable est dépendant de l'état d'une autre.

Il existe quatre combinaisons d'états possibles avec deux catégories dichotomiques de variables. Chacun de ces états peut soit rester inchangé, soit se transformer en l'un ou l'autre des trois autres états (figure 4.1). La matrice de probabilités de transition pour un changement corrélé à deux variables est définie selon la matrice suivante, où les lignes et colonnes sont respectivement les états conjoints 00 (X et Y absents), 01 (X absent, Y présent), 10 (X présent, Y absent), 11 (X et Y présents).

$$\mathbf{I}+\mathbf{Q}_D = \begin{pmatrix} 1-(q_{12}+q_{13}) & q_{12} & q_{13} & 0 \\ q_{21} & 1-(q_{21}+q_{24}) & 0 & q_{24} \\ q_{31} & 0 & 1-(q_{31}+q_{34}) & q_{34} \\ 0 & q_{42} & q_{43} & 1-(q_{42}+q_{43}) \end{pmatrix} \tag{4.10}$$

Les valeurs des paramètres décrivant le taux d'une double transition, soit un changement autant en X qu'en Y, sont mises à zéro dans $\mathbf{Q}_D$, la matrice $\mathbf{Q}$ corrélée. Le modèle ainsi conçu peut détecter les relations évolutives entre deux variables de façon à ce que l'état d'une variable affecte la probabilité d'un changement dans l'autre sur une période dt. Si ce n'était pas le cas, le modèle interpréterait les doubles transitions comme un phénomène indépendant et distinct des autres types de transition désirée. Ceci n'empêche pas pour autant la possibilité d'une transition double sur une période de temps t plus longue. Dans ce cas, la transition possible entre 00 et 11 en un temps t pourrait s'effectuer par une transition 00 vers 10, ou 01, puis vers 11. Le modèle fait en sorte qu'il est possible de distinguer ces deux alternatives. Une double transition au cours du temps dt impliquerait que les deux variables différentes aient changé au même instant, soit pendant une durée dt. Une telle implication est incompatible avec la définition que l'état d'une variable repose sur la probabilité d'un changement de l'autre. De plus, le

modèle interpréterait, de façon erronée, comme double transition instantanée tous les cas où il est observé, sur une plus longue période de temps, une transition dans les deux variables.

Comme pour les cas indépendants, les valeurs $P(t)$ sont déterminées à l'aide de l'équation 4.5 : $\mathbf{P}(t) = e^{\mathbf{Q}_D t}$. Pour calculer $\mathbf{P}(t)$, il est nécessaire de déterminer la décomposition matricielle $\mathbf{Q}_D$. Le modèle général de la figure 4.1 a huit paramètres (q_{ab}) pour lesquels la décomposition est difficile. Dans la section suivante, des modèles plus simples avec moins de paramètres sont étudiés. Dans la section 4.4.1, la matrice de taux est déterminée par deux paramètres seulement, soit le taux de perte et de gain commun entre deux gènes. Dans la section 4.4.2, le modèle est décrit par trois paramètres permettant de capturer le changement dans le taux de perte quand deux gènes sont présents en même temps.

4.4 Modèle d'évolution conditionnelle corrélée entre trois caractères

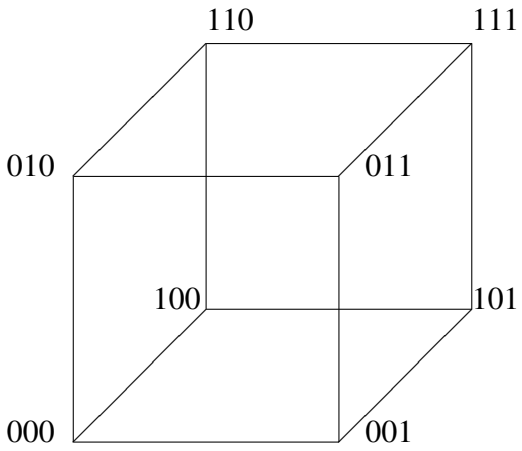


Figure 4.2 – Transitions possibles dans un système à trois états. Les arêtes représentent les changements d'un état lorsque les deux autres restent constants.

Dans tous les cas, une variable peut prendre deux états, 0 ou 1, dénotant leur présence ou leur absence chez une espèce. Trois gènes peuvent donc montrer huit différents patrons de présence ou d'absence dans chaque espèce. La figure 4.2 montre les liens entre

Tableau 4.I – Taux de transition dans un système à trois états suivant les possibilités de la figure 4.2.

Transitions		Taux
X=1	X=0	
$100 \rightarrow 101$	$000 \rightarrow 001$	λ
$100 \rightarrow 110$	$000 \rightarrow 010$	λ
$110 \rightarrow 111$	$010 \rightarrow 011$	λ
$101 \rightarrow 111$	$001 \rightarrow 011$	λ
$110 \rightarrow 100$	$010 \rightarrow 000$	ν
$101 \rightarrow 100$	$001 \rightarrow 000$	ν
$111 \rightarrow 101$	$011 \rightarrow 001$	δ
$111 \rightarrow 110$	$011 \rightarrow 010$	δ
$1ab \rightarrow 0ab$		μ
	$0ab \rightarrow 1ab$	s.o.

les différents états. Les taux de transition entre les états sont établis au tableau 4.I.

Lorsque le gène X est présent, les gènes Y et Y' sont définis comme formant une paire redondante. Ils sont acquis de façon indépendante avec un taux de gain λ. Lorsque seulement un gène, Y ou Y', est présent dans un génome donné, le taux de perte est identifié par ν. Lorsque les gènes Y et Y', sont tous les deux présents dans le génome, le taux de perte est défini par δ. Il est postulé que les taux sont en ordre croissant tel que

$$\lambda \leq \nu \leq \delta \tag{4.11}$$

Par contre, cette condition n'est pas absolument nécessaire pour la réussite de la méthode. Si $\delta \approx \nu$, il n'y a pas de dépendance entre les pertes de Y et Y'. Si $\delta < \nu$ il y a une évidence pour un couplage fonctionnel ou physique entre Y et Y'. Cette évidence partielle doit être combinée avec d'autres sources pour être admise. Si $\delta > \nu$, il peut y avoir une redondance fonctionnelle entre Y et Y'. Cette redondance est une indication d'un DGNO possible. Le cas où $\lambda < \nu$ montre que la perte d'un gène est plus facile que son gain, ce qui est biologiquement admis [61]. Le cas où $\lambda > \nu$, soit que le gain d'un gène est plus facile que sa perte, n'est pas biologiquement raisonnable. Mathématique-

ment, le cas où $\lambda > \nu$ est aussi acceptable dans les équations du modèle. L'imposition d'un ordre de grandeur n'est pas obligatoire pour autant que la variable γ définie au lemme 2, présentée plus loin, reste un nombre réel. Dans le cadre de cette étude, l'ordre défini à l'équation 4.11 est simplement imposé pour concentrer la recherche sur les DGNOs.

Le tableau 4.I correspond à une matrice de taux de 8 par 8 déterminée par les trois paramètres. Pour travailler avec la matrice, les transitions de X sont considérées séparément. La section 4.4.1 s'applique au cas où X reste absent et les deux familles, Y et Y', évoluent séparément. La section 4.4.2 examine le cas où X reste présent et les deux familles, Y et Y', évoluent de façon dépendante, avec une perte de δ lorsque les deux sont présents. La section 4.4.3 montre les probabilités de transition pour la totalité de la matrice $exp(Qt)$ de taille 8 par 8.

4.4.1 Transitions indépendantes avec trois profils

Dans le cadre d'une évolution non corrélée d'un complexe de trois gènes, l'indépendance peut être interprétée comme si le gène X n'est pas présent. Dans ce cas, les gènes Y et Y' forment un système de deux gènes qui peuvent être acquis et perdus de façon indépendante avec les taux de substitution respectifs λ et δ. Les états que peuvent prendre un système Y-Y' sur une branche peuvent être définis comme un processus de Markov tel que :

$$Y(t) \in \{00, 01, 10, 11\} \tag{4.12}$$

Dans le cas d'une évolution indépendante des gènes Y et Y' par rapport à X, étant donné que X est évalué comme absent, la matrice de taux de substitution se présente comme :

$$\mathbf{Q_0} = \begin{pmatrix} -2\lambda & \lambda & \lambda & 0 \\ \delta & -\delta-\lambda & 0 & \lambda \\ \delta & 0 & -\delta-\lambda & \lambda \\ 0 & \delta & \delta & -2\delta \end{pmatrix} \tag{4.13}$$

où les lignes et les colonnes représentent respectivement les états 00, 01, 10, 11 dans cet ordre.

Afin d'être en mesure d'analyser l'évolution indépendante, il est nécessaire d'utiliser une matrice exponentielle, en se basant sur l'équation 4.5, qui est représentée comme

$$S_0(t) = e^{\mathbf{Q_0}t} \tag{4.14}$$

Puisque le gène X est absent, la paire Y-Y' peut être dans un génome sans pour autant former une paire redondante comme ce serait le cas dans une situation où la paire Y-Y' est dépendante de X. Les taux de transition doivent être les mêmes sur les deux gènes Y et Y' puisqu'ils sont indépendants. Dans cette situation d'évolution non corrélée, le taux de transition δ doit être égal à ν. Si le taux de perte δ (Y et Y' présent) était supérieur à ν (Y ou Y' présent), il y aurait le signe d'une influence entre les deux gènes. Cette influence est contraire à une sélection indépendante.

Dans un premier temps, afin d'alléger les équations, il est de mise de définir quelques variables au lemme 2.

Lemme 2. *Définissons*

$$\begin{aligned} \alpha &= \lambda + \nu \\ \gamma &= \sqrt{\alpha^2 + 4(\delta-\nu)(\delta-\lambda)} \\ u &= 3\lambda + 2\delta + \nu + \gamma \\ v &= 3\lambda + 2\delta + \nu - \gamma \\ \omega &= \lambda^2 + 2\lambda\delta + \delta\nu \end{aligned}$$

Lemme 3. *La distibution stationnaire de* $\mathbf{Q_0}$ *peut être écrite comme* $(\pi'_{00}\pi'_{01}\pi'_{10}\pi'_{11})$ *avec*

$$\pi'_{00} = \frac{\delta^2}{\alpha^2} \qquad \pi'_{01} = \frac{\delta\lambda}{\alpha^2}$$
$$\pi'_{10} = \frac{\delta\lambda}{\alpha^2} \qquad \pi'_{11} = \frac{\lambda^2}{\alpha^2}$$

Idée de la preuve. Il peut être vérifié que

$$(\pi'_{00}\pi'_{01}\pi'_{10}\pi'_{11})\mathbf{Q_0} = (0000)$$

□

Les distributions à la racine sont établies par :

$$\pi'_{000} = \pi'_{001} = \pi'_{010} = \pi'_{011} = 0$$
$$\pi'_{100} = \frac{\delta^2}{\alpha^2}$$
$$\pi'_{101} = \pi'_{110} = \frac{\delta\lambda}{\alpha^2}$$
$$\pi'_{111} = \frac{\lambda^2}{\alpha^2}$$

Les distributions à la racine doivent inclure l'état X à cause du modèle de pure perte utilisé où X est présent à la racine. Les probabilités où X est absent sont donc mises à 0.

Lorsque $\delta = \nu$ la matrice exponentielle peut être écrite comme l'équation 4.15 en prenant les formules de 4.16. Ceci est fait en tenant compte des variables définies au lemme 2,

$$\mathbf{S_0}(t) = e^{\mathbf{Q_0}t} = \begin{pmatrix} \cdot & \pi'_{01}+a'(t) & \pi'_{10}+a'(t) & \pi'_{11}+b'(t) \\ \pi'_{00}+c'(t) & \cdot & \pi'_{10}+d'(t) & \pi'_{11}+e'(t) \\ \pi'_{00}+c'(t) & \pi'_{01}+d'(t) & \cdot & \pi'_{11}+e'(t) \\ \pi'_{00}+f'(t) & \pi'_{01}+g'(t) & \pi'_{10}+a'(t) & \cdot \end{pmatrix} \tag{4.15}$$

$$
\begin{aligned}
a'(t) &= \frac{\lambda^2 - \lambda\delta}{\alpha^2} e^{-\alpha t} - \frac{\lambda^2}{\alpha^2} e^{-2\alpha t} \\
b'(t) &= \frac{-2\lambda^2}{\alpha^2} e^{-\alpha t} + \frac{\lambda^2}{\alpha^2} e^{-2\alpha t} \\
c'(t) &= \frac{\lambda\delta - \delta^2}{\alpha^2} e^{-\alpha t} - \frac{\lambda\delta}{\alpha^2} e^{-2\alpha t} \\
d'(t) &= \frac{-2\lambda\delta}{\alpha^2} e^{-\alpha t} + \frac{\lambda\delta}{\alpha^2} e^{-2\alpha t} \\
e'(t) &= \frac{\lambda\delta - \lambda^2}{\alpha^2} e^{-\alpha t} + \frac{\lambda\delta}{\alpha^2} e^{-2\alpha t} \\
f'(t) &= \frac{-2\delta^2}{\alpha^2} e^{-\alpha t} + \frac{\delta^2}{\alpha^2} e^{-2\alpha t} \\
g'(t) &= \frac{2\delta^2 - \lambda\delta}{\alpha^2} e^{-\alpha t} + \frac{\delta^2}{\alpha^2} e^{-2\alpha t}
\end{aligned}
\tag{4.16}
$$

La diagonale représentée par des « · » dans la matrice 4.15 est substituée de façon à ce que les lignes s'additionnent pour égaler 1.

Idée de la preuve. En utilisant le produit Kronecker pour la matrice suivante avec elle-même

$$
exp\left(\begin{pmatrix} -\lambda & \lambda \\ \delta & -\delta \end{pmatrix} t\right) = \begin{pmatrix} \frac{\delta}{\alpha} + \frac{\lambda}{\alpha} e^{-\alpha t} & \frac{\lambda}{\alpha}(1 - e^{-\alpha t}) \\ \frac{\delta}{\alpha}(1 - e^{-\alpha t}) & \frac{\lambda}{\alpha} + \frac{\delta}{\alpha} e^{-\alpha t} \end{pmatrix} \tag{4.17}
$$

□

4.4.2 Transitions dépendantes avec trois profils

Dans le cas d'une évolution corrélée d'un complexe de trois gènes, où Y et Y' montrent une dépendance au gène X présent, la matrice des taux de substitution suivante s'applique.

$$\mathbf{Q_1} = \begin{pmatrix} -2\lambda & \lambda & \lambda & 0 \\ \nu & -\nu-\lambda & 0 & \lambda \\ \nu & 0 & -\nu-\lambda & \lambda \\ 0 & \delta & \delta & -2\delta \end{pmatrix} \tag{4.18}$$

où les lignes et les colonnes représentent respectivement les états 00, 01, 10, 11 dans cet ordre.

Afin d'être en mesure d'analyser l'évolution corrélée de Y et Y', il est nécessaire d'utiliser une matrice exponentielle, en se basant sur l'équation 4.5, qui est représentée comme

$$\mathbf{S_1}(t) = e^{\mathbf{Q_1}t} \tag{4.19}$$

qui donne les probabilités de transition sur une branche de longueur t étant donné que X est constamment présent. Il est possible de calculer la décomposition matricielle de $\mathbf{Q_1}$ symboliquement (la décomposition matricielle de $\mathbf{Q_1}$ étant faite en utilisant Mathématica [39]).

Lemme 4. *La distribution stationnaire de* $\mathbf{Q_1}$ *peut être écrite comme* $(\pi_{00}\pi_{01}\pi_{10}\pi_{11})$ *avec*

$$\begin{aligned} \pi_{00} &= \frac{\delta\nu}{\lambda^2+2\lambda\delta+\delta\nu} \\ \pi_{01} &= \frac{\delta\lambda}{\lambda^2+2\lambda\delta+\delta\nu} \\ \pi_{10} &= \frac{\delta\lambda}{\lambda^2+2\lambda\delta+\delta\nu} \\ \pi_{11} &= \frac{\lambda^2}{\lambda^2+2\lambda\delta+\delta\nu} \end{aligned}$$

Idée de la preuve. Il peut être vérifié que

$$(\pi_{00}\pi_{01}\pi_{10}\pi_{11})\mathbf{Q_1} = (0000)$$

□

Les distributions à la racine sont établies par :

$$\begin{aligned}
\pi_{000} &= \pi_{001} = \pi_{010} = \pi_{011} = 0 \\
\pi_{100} &= \frac{\nu\delta}{\lambda^2 + 2\lambda\delta + \nu\delta} \\
\pi_{101} &= \pi_{110} = \frac{\delta\lambda}{\lambda^2 + 2\lambda\delta + \nu\delta} \\
\pi_{111} &= \frac{\lambda^2}{\lambda^2 + 2\lambda\delta + \nu\delta}
\end{aligned}$$

La matrice exponentielle dans un cadre d'évolution dépendante peut être écrite, en tenant compte des mêmes variables définies au lemme 2, comme l'équation 4.20 en prenant les formules de 4.21.

$$\mathbf{S_1}(t) = e^{\mathbf{Q_1}t} = \begin{pmatrix} \cdot & \pi_{01} + a(t) & \pi_{10} + a(t) & \pi_{11} + b(t) \\ \pi_{00} + c(t) & \cdot & \pi_{10} + d(t) & \pi_{11} + e(t) \\ \pi_{00} + c(t) & \pi_{01} + d(t) & \cdot & \pi_{11} + e(t) \\ \pi_{00} + f(t) & \pi_{01} + g(t) & \pi_{10} + a(t) & \cdot \end{pmatrix} \tag{4.20}$$

$$
\begin{aligned}
a(t) &= \frac{\lambda}{8\gamma\omega}\Big((-3\lambda+2\delta-\nu-\gamma)ve^{-tu/2}+(3\lambda-2\delta+\nu-\gamma)ue^{-tv/2}\Big) \qquad (4.21)\\
b(t) &= \frac{\lambda^2}{2\gamma\omega}\Big(ve^{-tu/2}-ue^{-tv/2}\Big)\\
c(t) &= -\frac{\nu}{8\gamma\omega}\Big((3\lambda-2\delta+\nu+\gamma)ve^{-tu/2}+(-3\lambda+2\delta-\nu+\gamma)ue^{-tv/2}\Big)\\
d(t) &= \frac{-e^{-t\alpha}}{2}+\frac{1}{8\gamma\omega}\Bigg(-\Big((\delta-\lambda)^2-\Big(\frac{\gamma+\alpha}{2}\Big)^2\Big)ve^{-tu/2}\\
&\qquad\qquad +\Big((\delta-\lambda)^2-\Big(\frac{\gamma-\alpha}{2}\Big)^2\Big)ue^{-tv/2}\Bigg)\\
e(t) &= \frac{\lambda}{8\gamma\omega}\Big((\lambda-2\delta-\nu-\gamma)ve^{-tu/2}+(-\lambda+2\delta+\nu-\gamma)ue^{-tv/2}\Big)\\
f(t) &= \frac{\nu\delta}{2\gamma\omega}\Big(ve^{-tu/2}-ue^{-tv/2}\Big)\\
g(t) &= \frac{\delta}{8\gamma\omega}\Big((\lambda-2\delta-\nu-\gamma)ve^{-tu/2}+(-\lambda+2\delta+\nu-\gamma)ue^{-tv/2}\Big)
\end{aligned}
$$

Idée de la preuve. Il faut utiliser la décomposition spectrale de $\mathbf{Q_1} = \mathbf{U\Lambda V}$ et écrire $e^{\mathbf{Q_1}t} = \mathbf{U}e^{\Lambda t}\mathbf{V}$. Les valeurs propres sont $\Lambda = diag(0,-\alpha,-u/2,-v/2)$ □

4.4.3 Probabilités de transition

Dans un système de trois gènes, ou groupe de gènes, identifiés X-Y-Y', un gène X peut interagir avec Y ou Y'. Dans un tel système, l'évolution du gène X est gouvernée par un modèle de pure perte. Il est présent à la racine avec une probabilité de 1 et est perdu avec un taux μ. Cette contrainte a pour but principal de simplifier le calcul mathématique et la décomposition par exponentiation des équations pour deux états. Si le gain était admis, le modèle aurait un paramètre de plus. En particulier, il n'est pas clair si des modèles de corrélation plus compliqués que celui considéré ici permettent une décomposition analytique de la matrice de taux $\mathbf{Q}$ pour l'exponentiation. Cette contrainte est valide biologiquement quand la présence du phénotype, ou du gène X, à la racine est connue et l'évolution comprend seulement des pertes. Le taux μ est estimé en parallèle

des autres taux.

Le modèle d'évolution conditionnelle corrélée ici proposé se veut la complexification la plus simple pouvant être ajoutée au modèle existant pour deux états. Il est voulu que ce modèle ait peu de paramètres de façon à pouvoir interpréter les résultats immédiatement. La puissance statistique du test et des données est faible. Dans les expériences effectuées, la meilleure p-valeur était autour de 0,01. Il est donc nécessaire de se concentrer uniquement sur les plus petites valeurs statistiques. Ce modèle est une étape vers un cas plus général.

Afin de pouvoir étudier un système de trois gènes, il est nécessaire d'avoir les probabilités de transition pour les états joints de X-Y-Y'. Lorsque X est absent, les probabilités de transition sont données par l'équation 4.15, compte tenu du modèle de pure perte utilisé. Lorsque X reste présent, les probabilités de transition établies par l'équation 4.20 sont ajustées par la probabilité de non-perte de X. Si X est perdu sur une branche donnée de l'arbre, les probabilités de transition sont calculées en conditionnant par rapport au moment de cet événement de perte et sur les états d'Y et Y' à ce point. Les probabilités de transition $\mathbf{S_1}$ s'appliquent jusqu'à l'événement de perte pour être alors remplacées par $\mathbf{S_0}$.

Si l'évolution des trois gènes n'est pas corrélée, les probabilités de transition suivent les équations suivantes :

Soit $X[w]$ l'état du gène X au noeud w, et $Y[w]$ l'état des gènes Y-Y' à ce même noeud. Sur une branche ww' de longueur t, les probabilités sont

$$\begin{aligned}
\mathbb{P}\{X[w] = 1 | X[w] = 0\} &= 0 \\
\mathbb{P}\{X[w'] = 0; Y[w'] = y' | X[w] = 0; Y[w] = y\} &= \mathbf{S_0}(t)[y, y'] \\
\mathbb{P}\{X[w'] = 1; Y[w'] = y' | X[w] = 1; Y[w] = y\} &= e^{-\mu t}\mathbf{S_0}(t)[y, y'] \\
\mathbb{P}\{X[w'] = 0; Y[w'] = y' | X[w] = 1; Y[w] = y\} &= (1 - e^{-\mu t})\mathbf{S_0}(t)[y, y']
\end{aligned}$$

Si l'évolution des trois gènes suit une corrélation conditionelle, les probabilités de

transition suivent les équations suivantes :

$$\mathbb{P}\{X[w]=1|X[w]=0\}=0 \tag{4.22}$$
$$\mathbb{P}\{X[w']=0;Y[w']=y'|X[w]=0;Y[w]=y\}=\mathbf{S_0}(t)[y,y']$$
$$\mathbb{P}\{X[w']=1;Y[w']=y'|X[w]=1;Y[w]=y\}=e^{-\mu t}\mathbf{S_1}(t)[y,y']$$
$$\mathbb{P}\{X[w']=0;Y[w']=y'|X[w]=1;Y[w]=y\}$$
$$=\int_{s=0}^{t}\mu e^{-\mu s}\Big(\sum_{z\in\{00,01,10,11\}}\mathbf{S_1}(s)[y,z]\cdot\mathbf{S_0}(t-s)[z,y']\Big)\,\mathrm{d}s$$

L'intégrale est calculée symboliquement à partir des équations 4.15 et 4.20. En particulier, les fonctions peuvent être écrites comme $f(t)=\sum_k \alpha_k e^{(-t\beta_k)}$. Une telle fonction est représentée par les vecteurs α_k et β_k. La multiplication, l'intégration et la convolution se font symboliquement dans cette classe de fonction parce que le résultat reste de la même forme générale. En d'autres mots, les entrées dans la matrice de transition $P()$ sont des fonctions de cette forme. L'implémentation devient plus simple et comme la même représentation symbolique s'applique à toutes les branches, les paramètres α et β doivent être calculés une seule fois selon le modèle λ, ν, μ.

4.5 Calcul de la vraisemblance

Le modèle de Markov permettant de calculer la vraisemblance se définit ici sur un arbre enraciné avec des longueurs de branches v_i et des noeuds $i=0,1,\ldots,m$. Il possède un ensemble d'états possibles s déterminé selon le nombre de profils tel que :

- Un profil $\Rightarrow s=\{0,1\}$
- Deux profils $\Rightarrow s=\{00,01,10,11\}$
- Trois profils $\Rightarrow s=\{000,001,010,011,100,101,110,111\}$

Ceci permet de construire un modèle probabiliste de changement d'états pendant un temps $P_{ss'}(v)$, soit un processus de Markov avec les états s sur chaque branche et avec un état à la racine π_s.

L'expression générale de la vraisemblance d'un arbre a été décrite par Felsenstein [27]. Celle-ci est présentée ici de façon plus concrète à l'aide d'un cas particulier. Les équations suivantes font référence à l'arbre de la figure 4.3. Un format général émergera de cette expression. Les longueurs des branches sont données par les valeurs v_i. Si les états aux noeuds particuliers 0, 6, 7, ou 8 sur l'arbre sont connus et que ceux-ci sont s_0, s_6, s_7 et s_8, la vraisemblance de l'arbre serait le produit des probabilités des changements de chaque segment multiplié par la probabilité *a priori* π_{s0} de l'état s_0 à la racine. De cette façon, avec les s_i définis comme état au noeud i sur l'arbre, il est possible d'obtenir l'équation 4.23 par la propriété de Markov.

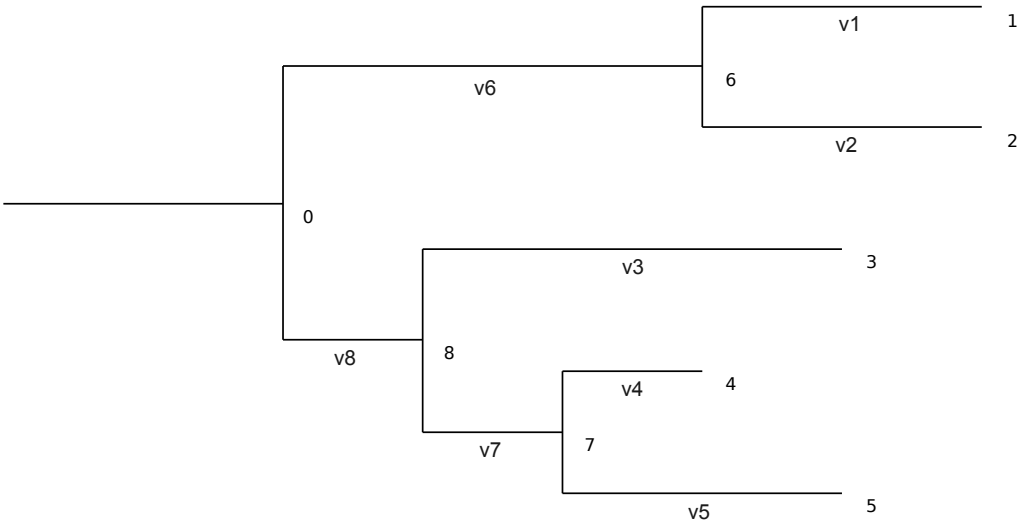


Figure 4.3 – Arbre phylogénétique utilisé pour la discussion de la vraisemblance. Les $v_i, 1 \leq i \leq 8$, représentent la longueur des branches.

$$\begin{aligned} L =& Pr(s_0, s_1, \ldots, s_8) \\ =& \pi_{s0} P_{s_0 s_6}(v6) P_{s_6 s_1}(v1) P_{s_6 s_2}(v2) P_{s_0 s_8}(v8) \\ & P_{s_8 s_3}(v3) P_{s_8 s_7}(v7) P_{s_7 s_4}(v4) P_{s_7 s_5}(v5) \end{aligned} \tag{4.23}$$

Dans les faits, les états ancestraux s_0, s_6, s_7 et s_8 sont inconnus, la vraisemblance est donc la somme de toutes les attributions possibles aux embranchements de l'arbre. Il en résulte l'équation 4.24.

$$\begin{aligned} L =& Pr(s_1, s_2, s_3, s_4, s_5) \\ =& \sum_{s_0}\sum_{s_6}\sum_{s_7}\sum_{s_8} \pi_{s0} P_{s_0 s_6}(v6) P_{s_6 s_1}(v1) P_{s_6 s_2}(v2) P_{s_0 s_8}(v8) \\ & P_{s_8 s_3}(v3) P_{s_8 s_7}(v7) P_{s_7 s_4}(v4) P_{s_7 s_5}(v5) \end{aligned} \tag{4.24}$$

En général, l'expression pour n espèces aura 2^{2n-2} termes dans le cas d'un arbre binaire avec $2n-2$ branches. Il est par contre facile d'économiser des termes en utilisant un analogue de la règle de Horner permettant de calculer les valeurs polynomiales rapidement. Ceci est fait en déplaçant les symboles de sommation vers la droite pour obtenir l'équation 4.25.

$$\begin{aligned} L = & \sum_{s_0} \pi_{s0} \Big\{ \sum_{s_6} P_{s_0 s_6}(v6) \left[P_{s_6 s_1}(v1) \right] \left[P_{s_6 s_2}(v2) \right] \Big\} \\ & \Big\{ \sum_{s_8} P_{s_0 s_8}(v8) \left[P_{s_8 s_3}(v3) \right] \left[\sum_{s_7} P_{s_8 s_7}(v7) (P_{s_7 s_4}(v4)) (P_{s_7 s_5}(v5)) \right] \Big\} \end{aligned} \tag{4.25}$$

Il est important de noter que le motif des parenthèses dans l'expression 4.25 montre la même topologie que l'arbre, soit $\Big\{ [] [] \Big\} \Big\{ [] () () \Big\}$. Il y a une probabilité pour chaque segment de l'arbre. En procédant de l'intérieur vers l'extérieur, il en vient à commencer aux feuilles et parcourir l'arbre jusqu'à la racine. Ceci revient à traverser l'arbre en postordre. Il est possible de redéfinir ceci en termes de probabilité conditionnelle en définissant $L_s^{(k)}$ comme la vraisemblance des données au noeud k ou dans son sous-arbre, en admettant que le noeud k a un état s. Si le noeud k est une feuille, $L_s^{(k)}$ est égale à zéro pour tous les états s sauf ceux observés, pour lesquels $L_{s_k}^{(k)}$ est égale à un. Ceci permet de débuter les calculs en évaluant pour chaque feuille k le nombre d'états possibles pour $L_s^{(k)}$. Dans le cas d'évaluation sur des profils de m gènes, le nombre d'états est 2^m.

Le noeud k, ayant i et j comme descendants immédiats $L_{s_k}^{(k)}$ se définit comme suit :

$$L_{s_k}^{(k)} = (\sum_{s_i} P_{s_k s_i}(v_i) L_{s_i}^{(i)})(\sum_{s_j} P_{s_k s_j}(v_j) L_{s_j}^{(j)}) \tag{4.26}$$

Si ceci ce poursuit jusqu'à ce que la racine de l'arbre soit atteinte, tous les termes de 4.25 sont calculés. Pour la racine, le point 0 sur la figure 4.3, il faudra ensuite calculer les vraisemblances conditionnelles $L_{s_0}^{(0)}$. La vraisemblance complète de l'arbre est donc

$$L = \sum_{s_0} \pi_{s0} L_{s_0}^{(0)} \tag{4.27}$$

Cet algorithme a été nommé « pruning » par Felsenstein, puisqu'il enlève deux feuilles de l'arbre à chaque étape. Cet algorithme est défini par la figure 4.4.

Procédure : Felsenstein pruning
Entrée : Profil de feuilles F, structure de l'arbre T, probabilité $P_{ss'}(v)$
Sortie : Tableau du maximum de vraisemblance pour les transitions possibles $L_a^{(i)}$
Pour tout noeud i de l'arbre T
SI i est une feuille
SET $L_a^{(i)} = \begin{cases} 0 & \text{si } a \neq F[i] \\ 1 & \text{si } a = F[i] \end{cases}$
SINON
j,k ← enfants de i
$L_a^{(i)} = \left(\sum_b \mathbb{P}_{ab}(v_j) \cdot L_b^{(j)}\right) \times \left(\sum_b \mathbb{P}_{ab}(v_k) \cdot L_b^{(k)}\right)$

Figure 4.4 – Pseudo-code décrivant l'algorithme de « pruning ». $L_a^{(i)}$ est un tableau des huit probabilités d'un noeud et L_b est un tableau des huit probabilitiés d'un fils d'un noeud.

4.6 Maximum de vraisemblance

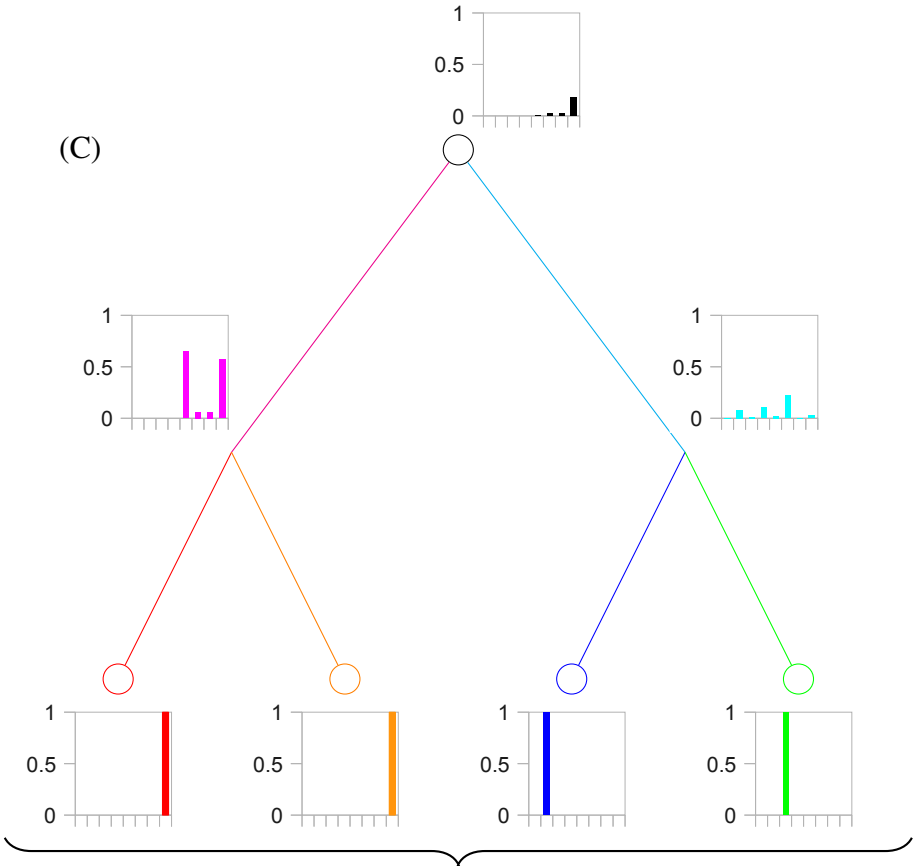
Il existe une panoplie d'approches algorithmiques ayant chacune leurs avantages et leurs inconvénients et permettant de résoudre, par une reconstruction ancestrale sur un arbre donné, les valeurs aux noeuds d'un arbre phylogénique. Une méthode traditionnelle implique généralement, sans être la seule, un calcul du maximum de vraisemblance sur un arbre fixe.

Pour débuter, les valeurs aux feuilles représentent le profil phylétique. Ce profil, sous forme de vecteur binaire, associe à 1 la présence et à 0 l'absence d'un état quelconque. De ce fait, la valeur observée est donc mise comme ayant une probabilité de 100 %. Ensuite, le calcul suit un modèle général de Markov phylogénétique qui permet de calculer la vraisemblance d'un arbre à chaque site selon des taux de substitution différents [27]. Ceci est défini de façon générale comme les pertes et les gains probabilistes sur les arcs donnant lieu à des étiquettes de noeuds d'un arbre binaire [27]. Finalement, la vraisemblance se définit comme la probabilité d'un profil, soit l'étiquetage des feuilles de l'arbre binaire, selon le modèle utilisé [27].

L'approche par maximum de vraisemblance prend en considération la longueur des branches. Ceci fait en sorte qu'une divergence génotypique ou phénotypique est interprétée comme plus probable sur une branche plus longue, en termes de temps évolutif, que sur une branche plus courte [42]. Donc, un changement non corrélé sur une branche de courte distance est jugé comme une probabilité de plus faible intensité qu'un changement non corrélé sur une branche plus longue.

Les calculs de vraisemblance pour les arbres ne sont pas simples. Généralement, ceux-ci exigent des calculs sur tous les états possibles non observés aux noeuds internes d'un arbre hypothétique. Ces calculs ont cependant été rendus possibles dans la pratique depuis les travaux de Felsenstein et son algorithme de « pruning » décrit à la section 4.5 [26, 53].

Puisqu'une estimation par maximum de vraisemblance est visée, le problème de recherche de corrélation conditionnelle se réduit à calculer la probabilité d'un ensemble particulier de données, soit les huit états possibles, des triplets de 0 et de 1, indiquant



	Sp1	Sp2	Sp3	Sp4
X	1	1	0	0
Y	1	1	0	1
Y’	1	1	1	0

	Sp1	Sp2	Sp3	Sp4
X	1	2	0	0
Y	1	2	0	1
Y’	1	2	3	0

Figure 4.5 – Schématisation de la procédure pour obtenir un arbre du maximum de vraisemblance. (A) À la base se trouve le tableau de la répartition des gènes selon les espèces. (B) De celui-ci est créé un profil phylétique. (C) Un arbre est inféré, en déterminant, à chaque noeud, les probabilités que celui-ci ait un profil donné. Ce patron est défini pour les gènes XYY’ comme étant de 000 à 111 de gauche vers la droite. Les valeurs aux feuilles sont connues et donc mises à 1. Les autres valeurs aux noeuds internes et à la racine sont calculées avec les équations de la section 4.4.

les présences et absences des gènes, X-Y-Y', sur un arbre donné et en maximisant cette probabilité sur l'ensemble de l'arbre phylogénétique. Un exemple fait de valeurs réelles est fourni à la figure 4.5. Sur cet exemple, le profil phylétique indique que les espèces 1 et 2 possèdent les trois gènes, ou états, X-Y-Y'. Leur probabilité est donc de 1. Le gène X doit être présent chez l'ancêtre des espèces 1 et 2 puisque les probabilités sur les hypothèses que son génotype n'ait pas l'état X est de 0. Ceci est dû au modèle de pure perte de l'état X utilisé. Les autres génotypes possibles, soit les états 1-Y-Y', ont des valeurs correspondant aux probabilités que cet ancêtre des espèces 1 et 2 ait ce génotype.

4.7 Optimisation

Les taux de mutation, μ, λ, δ et ν, des gains et des pertes de familles de gènes, X, Y et Y', tels que définis au tableau 4.I, sont grandement variables et, de façon encore plus importante, inconnus. Pour pallier ce problème, des paramètres initiaux respectant l'ordre croissant ont été arbitrairement fixés et sont optimisés pour chaque calcul de probabilité d'un arbre, corrélé, non corrélé ou inconditionnel. Ces taux ont été définis comme $\mu = 0,001$, $\lambda = 0,0001$, $\delta = 0,04$, $\nu = 0,01$ et servent de point de départ dans l'optimisation numérique. Ils sont optimisés en prenant le minimum d'une fonction sur ceux-ci entre 0 et un maximum arbitraire, ici défini comme le taux multiplié par la longueur de la plus grande branche de façon à ce que le taux en question soit d'au plus 3. Cette optimisation est faite par la méthode de minimisation en 1D de Brent telle que décrite par Press et coll. [8, 9, 74]. La précision de la fonction de minimisation est de 0,0001.

4.8 Sélection du modèle

Pour inférer les corrélations entre les profils, le maximum de vraisemblance est comparé selon les modèles introduits. Dans le cas de deux gènes, le modèle nul représente l'évolution indépendante avec les taux $\mathbf{Q_0}$ de l'équation 4.13 et le modèle d'évolution corrélée suit les taux $\mathbf{Q_1}$ de l'équation 4.18. Dans le cas de trois gènes, XYY', les trois modèles, nul, corrélé inconditionnel et corrélé conditionnel, sont comparés.

Le modèle nul correspond à la perte indépendante de X où

$$Q = \begin{pmatrix} 0 & 0 \\ \mu & -\mu \end{pmatrix} \text{ et } \pi_s(1) = 1$$

et à l'évolution indépendante $\mathbf{Q_0}$ pour Y-Y'.

Le modèle corrélé inconditionnel correspond à la perte indépendante de X ainsi qu'à l'évolution corrélée pour Y-Y'.

Le modèle corrélé conditionnel correspond à la probabilité de transition selon l'équation 4.22 de la page 54.

Quand X=1 pour toutes les feuilles, le modèle nul correspond au modèle nul de deux gènes Y-Y' et les deux autres modèles corrélés se réduisent au modèle corrélé Y-Y'.

La sélection du modèle revient à choisir un modèle statistique à partir d'un ensemble de modèles possibles. Il existe plusieurs approches pour la comparaison de modèles probabilistes. Ces modèles peuvent être imbriqués ou non et diffèrent en complexité. Parmi les modèles non imbriqués, il y a la pénalisation par « Akaike's information criterion » (AIC) [27]. Celui-ci calcule l'espérance d'un log-vraisemblance pour un nouvel ensemble de données de la même taille que le log-vraisemblance courant [27].

L'approche AIC correspond à prendre la négation du double du log-vraisemblance de chaque hypothèse et de le pénaliser en y additionnant le double du nombre de paramètres. Pour l'hypothèse i de p_i paramètres, le cas général, permettant d'obtenir des quantités pouvant être comparées entre les hypothèses, est présenté à l'équation 4.28.

$$AIC = 2p_i - 2\ln L_i \tag{4.28}$$

L'hypothèse ayant la plus petite valeur est celle retenue.

L'approche AIC est préférable lorsque les hypothèses ont un nombre toujours plus élevé de paramètres. Dans ces cas, un test de ratio de log-vraisemblances (RLV) peut être utilisé. Par contre, pour la sélection entre un modèle de corrélation qui peut être conditionnelle ou inconditionnelle, l'utilisation d'un test RLV, utilisé pour comparer l'ajustement de deux modèles imbriqués, est justifiée. Dans ce cas, l'hypothèse inconditionnelle

est imbriquée dans l'hypothèse conditionnelle. L'imbrication signifie que le modèle le plus complexe peut être transformé en modèle plus simple en imposant certaines contraintes linéaires sur les paramètres [27].

Pour effectuer ce test, les deux hypothèses sont ajustées à leurs données et leurs log-vraisemblances sont enregistrés. Le test statistique est le double de la différence entre ces log-vraisemblances tels que présentés par l'équation 4.29.

$$\begin{aligned} RLV &= -2(\ln(\text{vraisemblance du modèle nul}) - \ln(\text{vraisemblance du modèle alternatif})) \\ &= -2\ln\left(\frac{\text{vraisemblance du modèle nul}}{\text{vraisemblance du modèle alternatif}}\right) \\ &= -2\ln\left(\frac{\text{vraisemblance du modèle indépendant}}{\text{vraisemblance du modèle dépendant}}\right) \end{aligned} \tag{4.29}$$

L'hypothèse ayant le plus de paramètres sera au moins toujours aussi bien ajustée. Elle aura donc une valeur de log-vraisemblance plus grande. Afin de déterminer si elle est préférable, il suffit de dériver la probabilité obtenue par le RLV. Dans la plupart des cas, la distribution de probabilités du test statistique peut être approximée par une distribution χ^2, avec $(dl1 - dl2)$ degrés de liberté, où $dl1$ et $dl2$ sont les degrés de liberté des modèles inconditionnels et conditionnels [68]. Le degré de liberté est le nombre de paramètres libres dans le modèle. Le modèle non corrélé en a trois (μ, λ, ν) et les autres en ont quatre (μ, λ, ν, δ). Les deux modèles corrélés contiennent le modèle non corrélé en tant que cas spécial ($\delta = \nu$). Le test est donc approprié lorsque le χ^2 est utilisé avec un degré de liberté. Par contre, le χ^2 ne s'applique pas à la comparaison des modèles corrélés entre eux parce que la dépendance, ou l'indépendance, de X ne change pas le degré de liberté.

CHAPITRE 5

APPLICATIONS SUR DES MODÈLES BIOLOGIQUES

5.1 Introduction

Ce chapitre présente la validation de l'approche mathématique et algorithmique développée ainsi que différents éléments de réponses obtenues suite à la mise en application de cette approche avec des données biologiques concrètes.

La section 5.2 montre la validation statistique de l'approche mathématique par l'analyse de regroupement de paires de gènes de même famille. De plus, cette section montre, par un exemple sur des profils phylétiques représentant des DGNOs connus, comment analyser les tableaux de résultats obtenus.

L'arbre et les longueurs de branches utilisées lors de toutes les expérimentations suivantes sont établis à la figure 5.1. Cet arbre phylogénétique a été établi par Csűrös et Miklós [16]. La topologie de l'arbre et les longueurs de branches sont nécessaires pour le modèle, mais l'influence de celles-ci n'a pas été évaluée pour cette étude. Cette tâche était trop laborieuse dans le cadre d'une maîtrise.

Le but de l'expérience est de trouver des paires de gènes (Y et Y') montrant une évolution corrélée qui est conditionnelle à un état constitutif X. Dans le cadre de cette étude, cet état constitutif X est soit la motilié soit l'hyperthermophilie. Inévitablement, considérant la masse d'information disponible dans la base de données utilisée, arCOG, plusieurs éléments de réponses, soit des paires de gènes Y et Y' montrant une corrélation conditionnelle, ont été obtenus. Seul un élément de réponse par constitutif X a été retenu pour une analyse plus complète. Les sections 5.3 et 5.4 établissent un regroupement de gènes ayant été décelé comme étant possiblement corrélé et conditionnel à la motilité et à l'hyperthermophilie. Une interprétation de l'histoire évolutive, considérant la fonction de chaque gène, est par la suite donnée pour expliquer les profils observés.

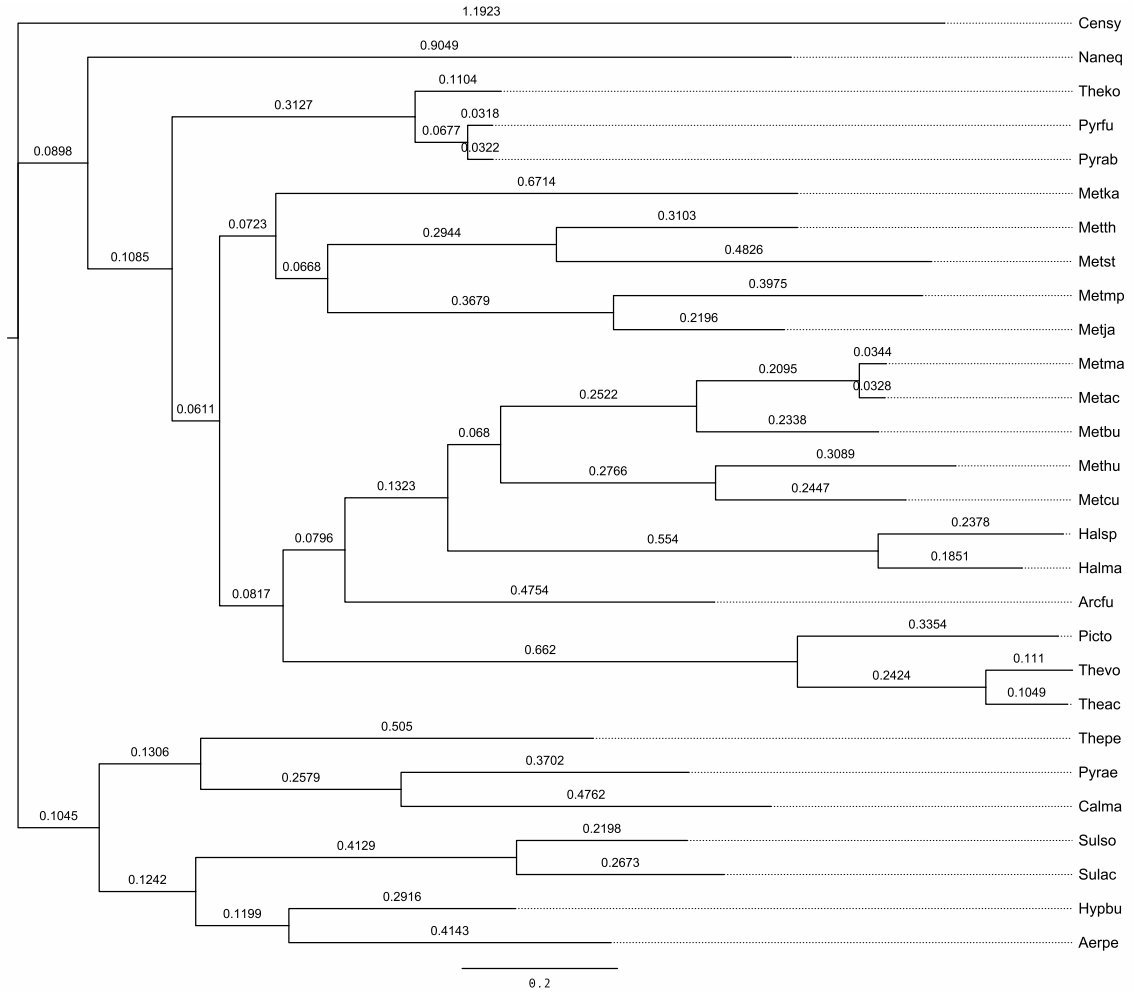


Figure 5.1 – Arbre phylogénétique avec longueurs de branches pour des espèces d'archaea. Cet arbre est utilisé par les modèles mathématiques définis à la section 4.4 pour la recherche sur les DGNOs

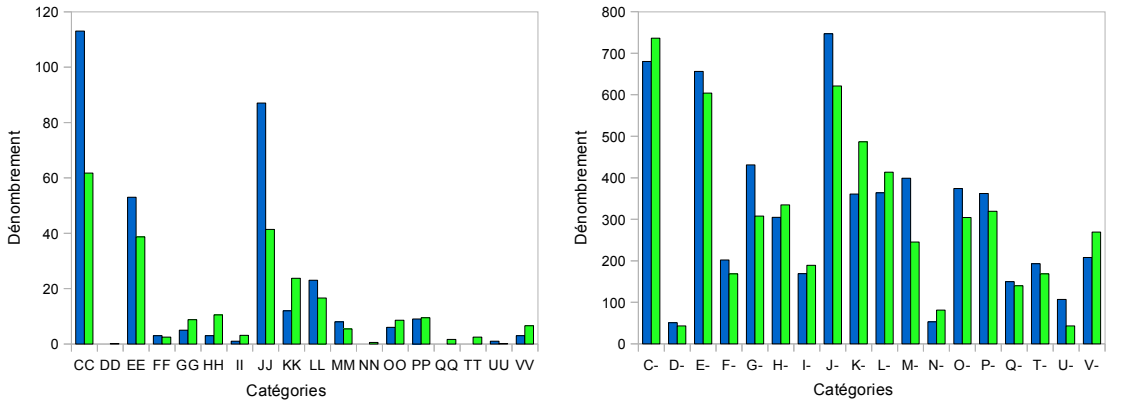
Aerpe, *Aeropyrum pernix* ; Arcfu, *Archaeoglobus fulgidus* ; Calma, *Caldivirga maquilingensis* IC-167 ; Censy, *Cenarchaeum symbiosum* ; Halma, *Haloarcula marismortui* ATCC 43049 ; Halsp, *Halobacterium* sp. ; Hypbu, *Hyperthermus butylicus* ; Metac, *Methanosarcina acetivorans* ; Metbu, *Methanococcoides burtonii* DSM 6242 ; Metcu, *Methanoculleus marisnigri* JR1 ; Methu, *Methanospirillum hungatei* JF-1 ; Metja, *Methanococcus jannaschii* ; Metka, *Methanopyrus kandleri* ; Metma, *Methanosarcina mazei* ; Metmp, *Methanococcus maripaludis* S2 ; Metst, *Methanosphaera stadtmanae* ; Metth, *Methanobacterium thermoautotrophicum* ; Naneq, *Nanoarchaeum equitans* ; Picto, *Picrophilus torridus* DSM 9790 ; Pyrab, *Pyrococcus abyssi* ; Pyrae, *Pyrobaculum aerophilum* ; Pyrfu, *Pyrococcus furiosus* ; Sulac, *Sulfolobus acidocaldarius* DSM 639 ; Sulso, *Sulfolobus solfataricus* ; Theac, *Thermoplasma acidophilum* ; Theko, *Thermococcus kodakaraensis* KOD1 ; Thepe, *Thermofilum pendens* Hrk 5 ; Thevo, *Thermoplasma volcanium*.

5.2 Validation

Dans un premier temps, afin de déterminer si le modèle mathématique décrit à la section 4.4 est adéquat pour établir des liens de corrélation, un test de regroupement aléatoire a été effectué. Pour ce test, le modèle de corrélation le plus simple a été utilisé. Le test vérifie l'hypothèse que le modèle de corrélation entre deux gènes est utile dans l'inférence fonctionnelle. Ce test vérifie si les corrélations entre les familles de mêmes catégories fonctionnelles sont plus fréquentes qu'attendu avec un modèle nul d'étiquetage aléatoire. Ce test comparatif a été fait sur des paires de gènes de même famille puisqu'il est légitime de croire que ceux-ci aient des liens corrélés. Ce test a été réalisé sur l'ensemble des valeurs obtenues dans le cas où les protéines recensées dans la base de données arCOG ont été prises comme des Y et Y' conditionnelles à un X toujours égale à 1. Ceci a permis de faire une recherche exhaustive des paires corrélées puisque la variable X est constante, éliminant le principe de conditionnalité. Afin de procéder à ce test, le nombre de chaque catégorie définissant un arCOG a été calculé pour chaque paire, Y et Y'. Par exemple, le nombre de fois que Y est un arCOG de catégorie C et Y' est de catégorie C, ou Y est de catégorie C et Y' de catégorie D, et ainsi de suite. Afin de réduire l'interférence que pourrait introduire une annotation érronée, deux conditions ont été posées. La première est que pour être additionné, un arCOG doit être présent dans au moins 3 espèces, soit 10 % des espèces dans le cas présent où la phylogénie a 28 espèces. La deuxième est que cet arCOG ne doit pas être de la catégorie R, fonction générale prédite seulement, ou S, fonction inconnue. Dans un premier temps, ce dénombrement a été fait sur la liste des paires obtenue par le modèle d'évolution corrélée. Ensuite, un étiquetage aléatoire des gènes a été effectué. Pour ce faire, une catégorie aléatoire a été attribuée à chaque gène pris sur la liste de paires de façon à ce qu'un même gène soit de la même catégorie dans toutes les paires. Une permutation aléatoire des paires a été effectuée. Cette permutation a été faite 100 000 fois et à chaque fois le nombre de paires de chaque catégorie a été calculé.

La figure 5.2 montre les variations entre les valeurs obtenues par les modèles de la section 4.4 et celles de la moyenne obtenue aléatoirement. Lors de ce test compara-

tif, dans le cas où les deux catégories sont les mêmes (figure 5.2a), l'observation d'un nombre plus important de corrélations trouvées par les modèles de la section 4.4 que par le regroupement aléatoire est espéré. Dans le cas des autres paires possibles, d'une catégorie avec une autre qui lui est différente (figure 5.2b), il est souhaité que ce dénombrement soit semblable entre les deux méthodes. C'est ce qui est observé dans la majorité des familles comme le démontre la figure 5.2.



(a) Fréquence observée entre les mêmes catégories (b) Fréquence totale additionnée observée entre des catégories différentes

Figure 5.2 – Tableaux comparatifs entre le nombre de paires observées pour différentes catégories par le modèle mathématique développé (bleu) et une fonction aléatoire (vert). (a) Entre les mêmes catégories (b) Sommation d'une catégorie et de toutes les catégories différentes. Les définitions des familles sont décrites au tableau 5.II.

Un taux XX/X-, où X est une catégorie donnée et « - » toutes les autres, par exemple CC/C- montre le rapport de la paire CC sur la somme de toutes les paires avec un C (en Y ou Y') possible, a été calculé à chaque itération et pour toutes les catégories. Les rangs percentiles des taux entre paires obtenues par le modèle ont ensuite été calculés pour chaque catégorie par rapport aux taux des paires aléatoires. Les rangs percentiles sont décrits au tableau 5.I.

Cette validation démontre que les résultats calculés par les modèles mathématiques peuvent être généralement considérés significatifs. Dans les cas où les percentiles sont bas, le nombre d'occurrences rend les résultats un peu moins significatifs. Ces cas ne de-

Tableau 5.I – Rangs percentiles du taux entre les catégories identiques et les catégories différentes (XX/X-) du modèle mathématique de recherche de corrélation parmi les taux du modèle aléatoire.

Catégorie	CC	DD	EE	FF	GG	HH	II	JJ	KK
Percentile	0,99	0,00	0.90	0,58	0,07	0,03	0,16	0,99	0,13
Catégorie	LL	MM	NN	OO	PP	QQ	TT	UU	VV
Percentile	0,97	0,48	0,00	0,16	0,38	0,00	0,00	0,89	0,21

Tableau 5.II – Définition des fonctions de chaque famille d'arCOG

famille	Fonction
	Traitement et entreposage de l'information
J	Traduction ; biogenèse et structure ribosomique
K	Transcription
L	Réplication ; recombinaison et réparation
	Processus cellulaire et signalisation
D	Contrôle du cycle cellulaire ; division ; partition du chromosome
V	Mécanismes de défense
T	Mécanismes de transduction du signal
M	Biogenèse de la paroi cellulaire/membrane/enveloppe
N	Motilité cellulaire
U	Trafic intracellulaire ; sécrétion ; transport
O	Modification post-traductionnelle ; renouvellement protéique ; chaperones
	Métabolisme
C	Production et conversion d'énergie
G	Transport et métabolisme de carbohydrate
E	Transport et métabolisme d'acide aminé
F	Transport et métabolisme de nucléotide
H	Transport et métabolisme de coenzyme
I	Transport et métabolisme de lipide
P	Transport et métabolisme d'ion inorganique
Q	Biosynthèse de métabolites secondaires ; transport et catabolisme
	Caractérisation insuffisante
R	Fonction générale prédite seulement
S	Fonction inconnue

vraient donc pas être pris en considération pour évaluer la force du modèle mathématique développé. Ceci montre simplement que le modèle permet de trouver des corrélations

entre deux gènes.

Le modèle mathématique a ensuite été appliqué à des profils phylétiques représentant des cas de DGNO connus et définis à la section 2.5. Dans ce cas-ci, la protéine conditionnelle X a la valeur constante de 1, puisque les espèces possédant une protéine (Y) ou l'autre (Y') possèdent toujours le phénotype X. L'article de Galperin et Koonin duquel les profils phylétiques ont été pris ne propose pas d'arbre phylogénétique des espèces qu'ils utilisent [32]. Un arbre a donc été établi pour les espèces composant les profils phylétiques cités dans l'article. Cet arbre a été fait (i) en prenant les séquences génomiques des espèces de la base de données du NCBI (ii) en regroupant par orthologies les gènes avec l'algorithme tel que décrit à la section 3.3 (iii) en alignant les gènes orthologues retrouvés dans les groupes formés d'au moins un exemplaire de toutes les espèces avec muscle (iv) en inférant un arbre consensus avec phyml [23, 35].

Les valeurs calculées se trouvent dans le tableau 5.III. Le type de corrélation, ou non-corrélation, attribué à un triplet X-Y-Y', est analysé en prenant la valeur absolue minimale parmi les valeurs de log-vraisemblance. Celle-ci, corroborée par une p-valeur suffisamment petite, permet d'inférer le lien de corrélation et de conditionnalité entre les trois gènes.

Tableau 5.III – Valeurs de log-vraisemblance et taux de transition des pertes des gènes Y et Y' en présence individuelle (ν) et en paire (δ) entre différentes paires de protéines établissant des liens de corrélation. Les noms et fonctions des gènes sont donnés au tableau 5.IV

Y	Y'	Log-vraisemblance	p-valeur	δ	ν
		Valeurs non-corrélées inconditionnelles			
cog1190	cog1384	-20,63	s.o.	0,473	s.o.
cog0191	cog1830	-22,25	s.o.	0,622	s.o.
cog0207	cog1531	-22,38	s.o.	0,564	s.o.
		Valeurs corrélées inconditionnelles			
cog1190	cog1384	-17,8	0,017	2,909	0,473
cog0191	cog1830	-20,58	0,068	2,335	0,622
cog0207	cog1531	-18,95	0,009	3,849	0,564

Dans tous les cas, la valeur absolue minimale permet de définir le duo de protéines

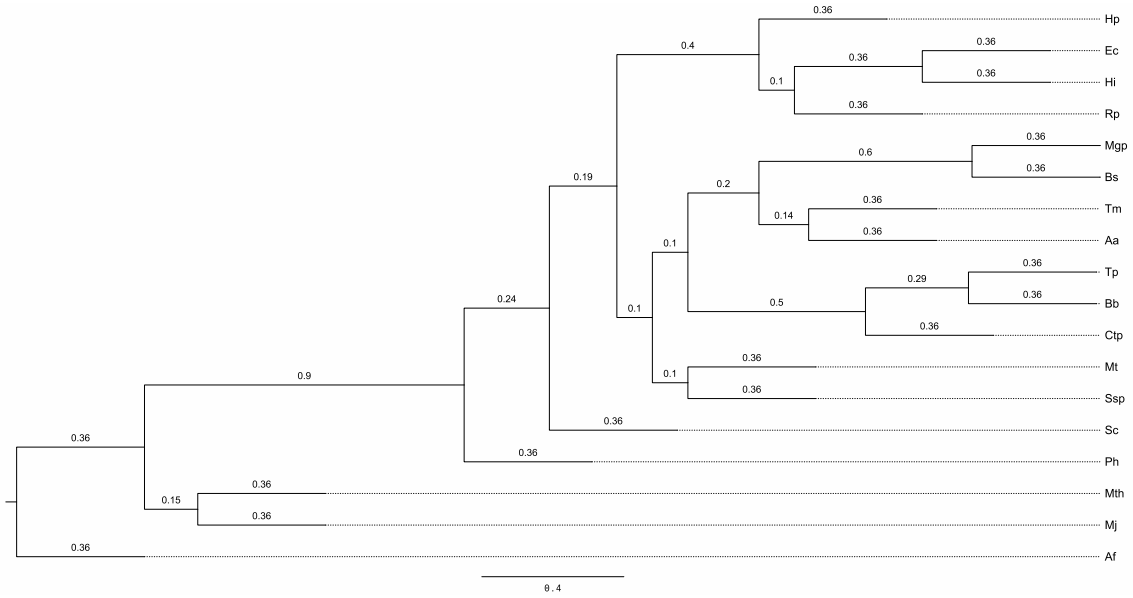


Figure 5.3 – Arbre phylogénétique pour les espèces utilisées conjointement avec le tableau 2.I pour la validation du modèle mathématique dans le cas de DGNO connus.

Aa, *Aquifex aeolicus*, Af, *Archaeoglobus fulgidus*, Bb, *Borrelia burgdorferi*, Bs, *Bacillus subtilis*, Ctp, *Chlamydia trachomatis* & *Chlamydia pneumoniae*, Ec, *Escherichia coli*, Hi, *Haemophilus influenzae*, Hp, *Helicobacter pylori*, Mgp, *Mycoplasma genitalium* & *Mycoplasma pneumoniae*, Mj, *Methanococcus jannaschii*, Mt, *Mycobacterium tuberculosis*, Mth, *Methanobacterium thermoautotrophicum*, Ph, *Pyrococcus horikoshii*, Rp, *Rickettsia prowazekii*, Sc, *Saccharomyces cerevisiae*, Ssp, *Synechocystis* sp., Tm, *Thermotoga maritima*, Tp, *Treponema pallidum*.

Tableau 5.IV – Catégories et fonctions des COGs corrélés

COG	Catégorie	Fonction
cog1190	J	Synthétase lysyl-ARNt (classe II)
cog1384	J	Synthétase lysyl-ARNt (classe I)
cog0191	G	FBA
cog1830	G	FBA de type DhnA
cog0207	F	Thymidylate synthase
cog1531	S	Protéine d'archaea à la fonction inconnue

J : Traduction ; biogenèse et structure ribosomique ; G : Transport et métabolisme de carbohydrate ; F : Transport et métabolisme de nucléotide ; S : Fonction inconnue.

comme corrélé. Cette corrélation est en accord avec la théorie pour ces protéines [32]. De plus, la nature de la corrélation peut être inférée en examinant les valeurs des taux de perte ν (Y ou Y' présent) et δ (Y et Y' présent). Lorsque $\delta > \nu$, une corrélation négative

est observée. Il peut alors être supposé que Y et Y' possèdent une force sélective les restreignant à ne pas être ensemble. Une redondance de fonctionnalité entre les deux peut donc être supposée. C'est le cas pour les DGNO et c'est ce qui est observé au tableau 5.III.

Tel qu'il peut déjà être observé ici, la puissance statistique de la méthode mathématique d'inférence de corrélation conditionnelle est faible. Par contre, cette méthode est tout de même valide puisque le tableau 5.I montre des cas où le modèle mathématique est meillleur que le modèle aléatoire. Ceci montre que pour des catégories fonctionnelles, le modèle trouve des corrélations à partir de profils phylétiques.

5.3 Motilité

La recherche de corrélation liée à la motilité a été effectuée selon le profil où X est égal à la présence ou à l'absence de motilité par flagellation, Y est égal à chacun des arCOGs de la famille N, soit celle liée à la motilité, et Y' est égal à chacun des arCOGs d'une autre famille.

Afin de faciliter l'analyse parmi l'ensemble de résultats obtenus, un seuil de précision subjectif a été établi. Pour être conservé, un triplet se doit d'avoir un taux de complémentarité, établi par l'équation 1.1, supérieur à 0,75 et une p-valeur associée à un log-vraisemblance indiquant une corrélation conditionnelle possible inférieure à 0,01.

De ce groupe restreint, une corrélation entre la protéine de flagelle FlaD/E et certaines protéines membranaires a été retenue pour une analyse.

5.3.1 Protéine du flagelle FlaD/E - protéines membranaires

Plusieurs molécules sont impliquées directement et indirectement dans le déplacement physique des espèces motiles. La section 2.3 résume les différents mécanismes connus et les liens marquants entre les systèmes de chimiotactisme et de motilité entre les espèces d'archaea et les bactéries. Elle conclut en émettant l'hypothèse que les archaea ont probablement adapté un système de motilité à partir d'un système de chimiotactisme plus général et commun avec les bactéries, tout en possédant un système de

sécrétion différent.

Parmi les protéines d'archaea dont la ou les fonctions sont encore mal comprises, se trouve la flagelline FlaD/E. Elle se nomme ainsi puisque les deux flagellines FlaD et FlaE sont très semblables. FlaD est une polyprotéine avec une région C-terminale montrant une similarité de séquence avec la région C-terminale de FlaE et les deux sont des protéines membranaires [91]. La protéine FlaD/E n'est pas obligatoire pour la motilité, mais lorsque le gène l'encodant est présent dans le génome, elle est nécessaire pour former un flagelle complet et fonctionnel [78]. Des études récentes concluent que la fonction de FlaD/E est soit impliquée dans la sécrétion et l'assemblage du flagelle soit impliqué dans le moteur du flagelle, ou d'une structure reliée, permettant l'inversion du mouvement [17, 78, 90].

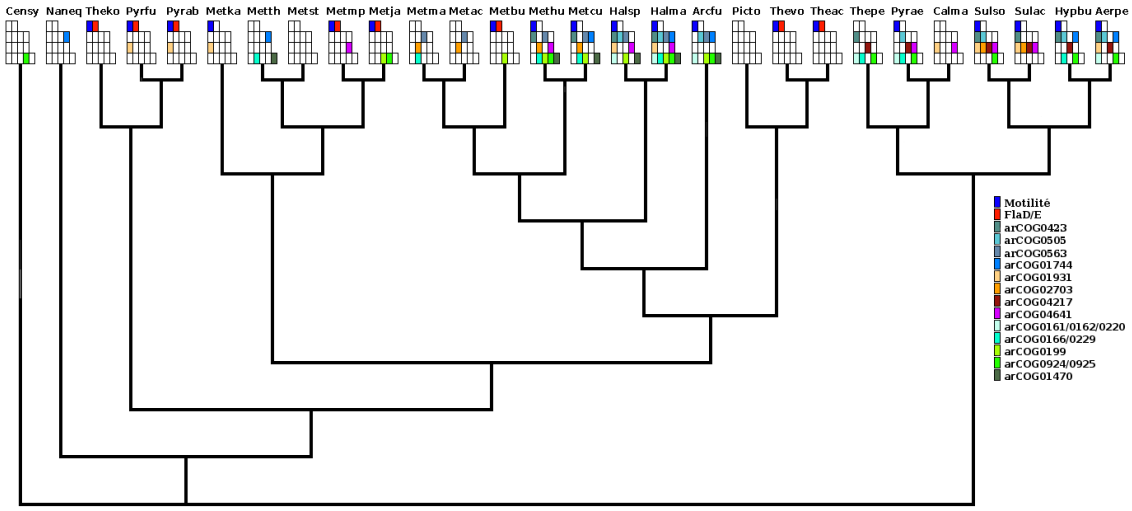


Figure 5.4 – Arbres phylogénétiques présentant le profil phylétique aux feuilles de protéines corrélées avec FlaD/E chez différentes espèces d'archaea. Un carré blanc représente l'absence d'un gène et un carré de couleur représente la présence d'un (des) gène(s) selon la légende. De gauche à droite, Ligne du haut : Motilité, FlaD/E ; 2^{e} ligne : arCOG0423, arCOG0505, arCOG0563, arCOG01744 ; 3^{e} ligne : arCOG01931, arCOG02703, arCOG04217, arCOG04641 ; ligne du bas : arCOG0161/0162/0220, arCOG0166/0229, arCOG0199, arCOG0924/0925, arCOG01470.

L'analyse de la figure 5.4 montre qu'en présence d'un phénotype de motilité la protéine FlaD/E peut être substituée par une panoplie de protéines. Les valeurs de log-

vraisemblances et les taux de transition décrits au tableau 5.V montrent une corrélation conditionnelle entre la protéine FlaD/E et toutes les protéines de ce groupe. Selon le tableau 5.VI, ce groupe de protéines remplaçant FlaD/E peut être divisé en deux catégories distinctes, soit celle des transporteurs de types ABC et celle des hydrolases. Il s'agit donc dans un cas, d'une protéine membranaire ayant un rôle important dans la sécrétion en utilisant l'énergie de l'hydrolyse de l'ATP et, dans l'autre, d'une enzyme qui catalyse une réaction chimique d'hydrolase, séparant une molécule d'eau en cation d'hydrogène (H+) et anion hydroxide (OH-). De plus, donnant un indice plus informatif quant au rôle spécifique, FlaD/E montre une corrélation avec une méthyltransférase S-Adénosyl méthionine-dépendante (SAM). La méthyltransférase SAM-dépendante est un co-substrat, souvent cyclique, communément impliqué dans le transfert d'un groupement méthyle [20, 50, 81]. Il est supposé que la méthylation est associée au chimiotactisme [20, 81].

Les transporteurs ABC répertoriés ont plusieurs caractéristiques en commun qui peuvent être regroupés en deux fonctions chevauchantes, soit un lien avec les méthyltransférases SAM-dépendantes, soit un lien avec des transporteurs métal-dépendants. Les méthyltransférases peuvent être classées en deux catégories, dont une nécessite un métal aidant à sa stabilité [15].

Le transporteur de spermidine/putrescine joint à la protéine périplasmique liant la spermidine/putrescine se classe, dans le cas présent, comme étant lié à la méthyltransférase SAM-dépendante. Un rôle important de la SAM est son implication dans la biosynthèse de polyamines telle que la spermidine/putrescine. La figure 5.5 montre un diagramme simplifié du cycle de la méthionine et du rôle de la SAM dans la voie polyamine. Dans cette voie polyamine, suite à une succession de transformations biochimiques, la putrescine attaque une SAM décarboxylée et la convertit en spermidine. Cette transformation biochimique produit et régénère la SAM de façon cyclique. La première étape de ce cycle est la méthylase SAM-dépendante. Par la suite, des hydrolases catalysent la réaction complète de dégradation jusqu'à la formation d'une homocystéine qui est recyclée en méthionine par le transfèrt d'un groupement méthyle [50]. Cette étape est intéressante puisque c'est précisément ce que peut faire une méthionine cobalamine-

dépendante [79]. Le fait que l'autre transporteur ABC soit celui d'une cobalamine vient ajouter un argument supplémentaire favorisant l'association de ces deux transporteurs pour l'étape de méthylation du chimiotactisme.

Tableau 5.V – Valeurs de log-vraisemblance et taux de transition des pertes des gènes Y et Y' en présence individuelle (ν) et en paire (δ) entre FlaD/E et des protéines membranaires établissant des liens de corrélation.

Y1	Y2	Log-vraisemblance	p-valeur	δ	ν
		Valeurs non-corrélées inconditionnelles			
arCOG02965	arCOG00161	-41.68	s.o.	1.78	s.o.
arCOG02965	arCOG00162	-41.68	s.o.	1.78	s.o.
arCOG02965	arCOG00166	-46.83	s.o.	2.61	s.o.
arCOG02965	arCOG00199	-41.48	s.o.	1.21	s.o.
arCOG02965	arCOG00220	-41.68	s.o.	1.78	s.o.
arCOG02965	arCOG00229	-46.83	s.o.	2.61	s.o.
arCOG02965	arCOG00423	-42.10	s.o.	0.87	s.o.
arCOG02965	arCOG00505	-42.62	s.o.	1.64	s.o.
arCOG02965	arCOG00563	-39.18	s.o.	0.85	s.o.
arCOG02965	arCOG00924	-46.96	s.o.	1.84	s.o.
arCOG02965	arCOG00925	-46.96	s.o.	1.84	s.o.
arCOG02965	arCOG01470	-40.42	s.o.	1.37	s.o.
arCOG02965	arCOG01744	-44.05	s.o.	2.26	s.o.
arCOG02965	arCOG01931	-44.72	s.o.	1.81	s.o.
arCOG02965	arCOG02703	-39.79	s.o.	1.50	s.o.
arCOG02965	arCOG04217	-38.21	s.o.	0.76	s.o.
arCOG02965	arCOG04641	-43.51	s.o.	1.45	s.o.
		Valeurs corrélées inconditionnelles			
arCOG02965	arCOG00161	-39.96	0.06392	94.27	1.55
arCOG02965	arCOG00162	-39.96	0.06392	94.27	1.55
arCOG02965	arCOG00166	-44.30	0.02452	94.27	2.19
arCOG02965	arCOG00199	-41.41	0.70225	1.58	1.09
arCOG02965	arCOG00220	-39.96	0.06392	94.27	1.55
arCOG02965	arCOG00229	-44.30	0.02452	94.27	2.19
arCOG02965	arCOG00423	-39.83	0.03313	94.27	0.94
arCOG02965	arCOG00505	-40.63	0.04597	94.26	1.45
arCOG02965	arCOG00563	-36.86	0.03125	94.27	0.44
arCOG02965	arCOG00924	-45.93	0.15102	7.55	1.59
arCOG02965	arCOG00925	-45.93	0.15102	7.55	1.59
arCOG02965	arCOG01470	-38.77	0.06910	94.27	1.04
arCOG02965	arCOG01744	-41.85	0.03598	94.27	1.90
arCOG02965	arCOG01931	-44.19	0.30414	3.74	1.50
arCOG02965	arCOG02703	-38.24	0.07861	94.27	1.19
arCOG02965	arCOG04217	-37.05	0.12753	94.27	0.85
arCOG02965	arCOG04641	-43.24	0.46598	3.41	1.42
		Valeurs corrélées conditionnelles			
arCOG02965	arCOG00161	-36.35	0.00110	26.23	0.79
arCOG02965	arCOG00162	-36.35	0.00110	26.23	0.79
arCOG02965	arCOG00166	-42.89	0.00498	13.00	1.24
arCOG02965	arCOG00199	-38.00	0.00831	8.70	0.46
arCOG02965	arCOG00220	-36.35	0.00110	26.23	0.79
arCOG02965	arCOG00229	-42.89	0.00498	13.00	1.24
arCOG02965	arCOG00423	-35.38	0.00025	28.17	0.87
arCOG02965	arCOG00505	-33.65	0.00002	94.26	1.53
arCOG02965	arCOG00563	-34.45	0.00210	9.10	0.33
arCOG02965	arCOG00924	-43.31	0.00684	12.06	1.47
arCOG02965	arCOG00925	-43.31	0.00684	12.06	1.47
arCOG02965	arCOG01470	-35.21	0.00125	20.99	0.52
arCOG02965	arCOG01744	-40.70	0.00957	16.60	1.36
arCOG02965	arCOG01931	-41.05	0.00676	8.15	0.82
arCOG02965	arCOG02703	-36.16	0.00707	10.90	0.81
arCOG02965	arCOG04217	-32.97	0.00121	15.53	0.35
arCOG02965	arCOG04641	-39.79	0.00643	14.03	0.98

Tableau 5.VI – Catégories et fonctions des arCOGs corrélés à la protéine FlaD/E

arCOG	Catégorie	Fonction
arCOG02965	N	Protéine du flagelle d'archaea D/E
arCOG00161	E	Transporteur ABC de spermidine/putrescine ; composante permease I
arCOG00162	E	Transporteur ABC de spermidine/putrescine ; composante permease II
arCOG00166	H	Transporteur ABC de tungstate ; composante periplasmique
arCOG00199	P	Transporteur ABC de cobalamin/Fe3+- ; composante ATPase
arCOG00220	E	Protéine périplasmique liant la spermidine/putrescine
arCOG00229	H	Transporteur ABC de tungstate ; composante permease
arCOG00423	R	Phosphohydrolase Predite
arCOG00505	R	Hydrolase Zinc-dépendante
arCOG00563	S	Protéine membranaire indéfinie
arCOG00924	E	Transporteur ABC d'acide aminé ; composante ATPase
arCOG00925	E	Transporteur ABC d'acide aminé ; composante ATPase
arCOG01470	C	Transporteur ABC de Na+ ; composante permease
arCOG01744	R	Hydrolase métalo-dépendante lié à la membrane
arCOG01931	R	Hydrolase métalo-dépendante lié à la membrane prédite
arCOG02703	Q	Methyltransferase SAM-dépendante
arCOG04217	R	Metallopeptidase prédite
arCOG04641	S	Protéine membranaire prédite

E : Transport et métabolisme d'acide aminé ; H : Transport et métabolisme de coenzyme ; R : Prédiction de fonctions générales seulement ; N : Motilité ; P : Transport et métabolisme d'ions inorganiques ; S : Fonction inconnue ; C : Production et conservation d'énergie ; Q : Biosynthèse de métabolite secondaire, transport et catabolisme.

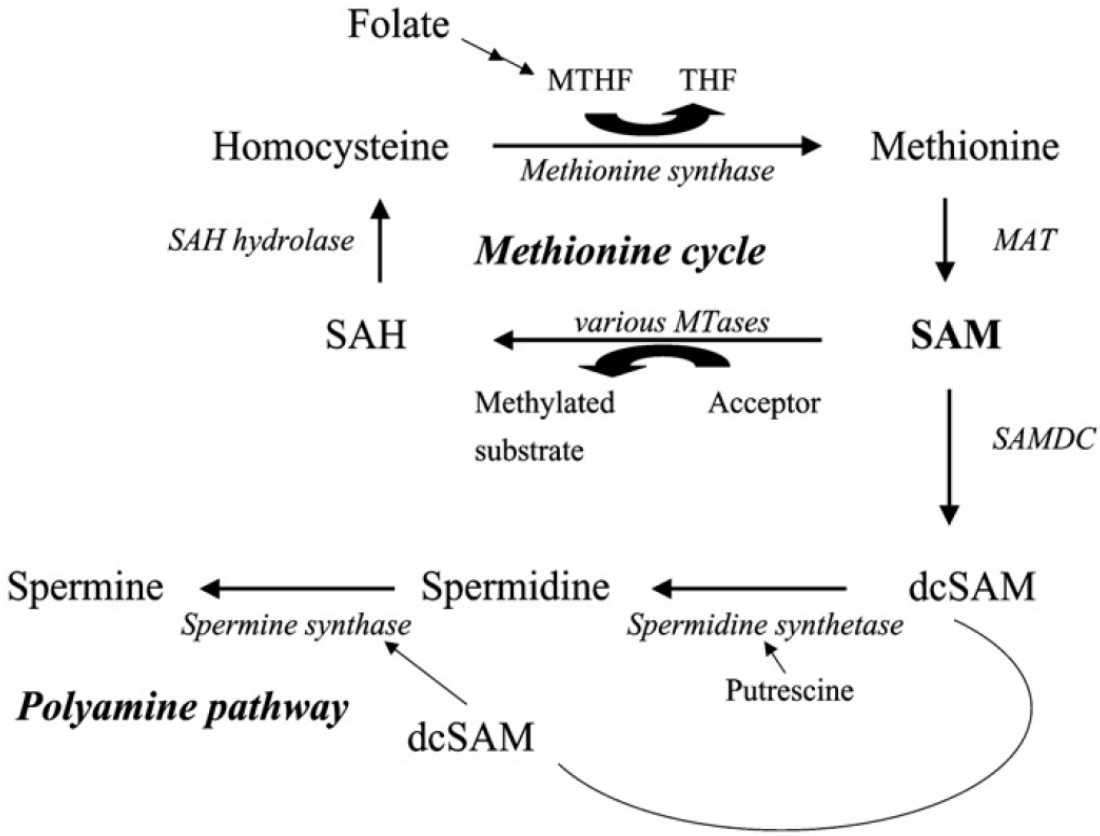


Figure 5.5 – Diagramme simplifié du métabolisme de la SAM [54]

Les trois autres transporteurs, soit ceux de tungstate, de sodium et d'acides aminés, peuvent jouer un rôle dans la stabilisation ou dans l'apport d'atomes aux réactions biochimiques impliquées dans le transfert du méthyl. Le métal utilisé pour la stabilité dans le cas du transfert du méthyl par l'O-méthyltransférase est le magnésium [92]. Par contre, des simulations suggèrent que ce magnésium agit principalement pour organiser le site de liaison aux substrats et n'agit pas comme base générale [79, 92]. Certaines O-méthyltransférases, qui transfèrent spécifiquement un méthyle sur un oxygène, comme celles qui peuvent attaquer le glutamate qui offre deux oxygènes en cible, ne requièrent pas d'ion métallique, mais nécessitent plutôt l'acide aminé histidine, qui est chargé positivement [79]. L'apport du transporteur d'acide aminé pourrait donc intervenir en cet endroit. De plus, le zinc et le calcium sont connus comme pouvant remplacer le mag-

nésium [92]. L'une des hydrolases répertoriées comme étant possiblement corrélée est dépendante du zinc. D'autres études sur la possibilité d'utiliser certains métaux ou sur l'implication que peuvent avoir ces derniers transporteurs sont par contre nécessaires pour valider ou rejeter les hypothèses avancées ici.

Les hydrolases, formant l'autre groupe principal, peuvent avoir un rôle dans la méthylation/déméthylation. La S-adénosylhomocystéine (SAH), soit le produit formé lorsqu'un groupement méthyle d'un donneur SAM est transféré vers une molécule réceptrice, est une inhibitrice forte de toute réaction de transméthylation [81]. Dans les cellules eucaryotes, cette inhibition est tempérée par l'hydrolyse de la SAH en adénosine et en homocystéine catalysée par une hydrolase [81]. La phosphohydrolase aussi identifiée peut avoir un rôle dans le chimiotactisme. La phosphorylation est très importante dans les principes chimiotactiques connus car, chez les bactéries, les protéines CheA et CheY doivent être phosphorylées pour être actives [19, 20]. La métallopeptidase peut aussi contribuer au chimiotactisme bien qu'elle est généralement impliquée dans le développement, la reproduction et le remodelage de tissus. La métallopeptidase a une activité catalytique d'hydrolase de lien peptidique dans lequel l'eau agit comme nucléophile. Ensuite, un ou deux ions métalliques tiennent cette molécule d'eau en place et un acide aminé chargé se lie à cet ion métallique. Le metal est généralement du zinc, mais il peut aussi être du cobalt, du manganèse ou du cuivre [73]. D'autres protéines membranaires aux rôles inconnus, dont l'identification serait sûrement intéressante, complètent le tableau.

Le fait que les deux transporteurs, spermidine/putrescine et cobalamine, renforcés par l'utilisation d'hydrolases, puissent avoir été adaptés en méthyltransférase pour le chimiotactisme n'est pas inimaginable. Les protéines connues de chimiotactisme, principalement CheABR, agissent d'ailleurs selon un principe similaire. Des liens fonctionnels entre elles peuvent donc être envisagés.

Chez les bactéries, le processus de chimiotactisme connu agit en deux étapes, soit la régulation du flagelle et la régulation du récepteur. Pour ce dernier, CheB, lorsqu'activée par CheA, agit comme une méthyltransférase et enlève le méthyle d'un résidu glutamate. Elle agit de façon antagoniste à CheR, aussi une méthyltransférase, qui ajoute un

méthyle à ce même glutamate de façon SAM-dépendante. De plus, CheB appartient à la classe des sérines hydrolases. Les méthyltransférases CheB et CheR modulent le signal de sortie du récepteur chimiotactique en contrôlant le niveau de méthylation. Plus il y a de groupements méthyles d'attachés au récepteur, plus il se désensibilise. De cette façon, cette cascade permet un cycle de méthylation et de déméthylation possédant la caractéristique de mémoire à court terme de la dernière concentration externe de résidu attirant ou repoussant. Les enzymes de modification du récepteur, CheB et CheR, catalysent la méthylation et la déméthylation au même site sur le chimiorécepteur, amenant à des questions encore non résolues sur l'évolution de ces deux enzymes collaboratrices [19, 20]. L'inclusion de FlaD/E, des transporteurs ABC et des hydrolases acquis de par cette étude pourrait permettre une meilleure compréhension des mécanismes inhérents à la motilité et de leurs évolutions.

L'hypothèse peut donc être proposée que les liens chimiotactiques encore inobservés chez les archaea motiles proviendraient d'une adaptation du système de transport ABC jumelé à des hydrolases. Cette hypothèse ne contredit pas les informations préalablement publiées sur la protéine FlaD/E et sa situation membranaire ainsi que son rôle avec le moteur du flagelle. Il s'agit, par contre, de suppositions tirées de modèles probabilistes appuyés par la litérature. Une investigation biologique par expérience en laboratoire directement sur ces organismes est de mise avant d'établir les hypothèses ici avancées.

5.4 Reverse gyrase

La recherche de corrélation liée à l'hyperthermophilie a été effectuée selon le profil où X est égal à la présence ou à l'absence de la reverse gyrase, tandis qu'Y et Y' sont égaux à chacun des arCOGs.

Afin de faciliter l'analyse parmi l'ensemble de résultats obtenus, un seuil de précision subjectif a été établi tel que, pour être conservé, un triplet se doit d'avoir un taux de complémentarité, établi par l'équation 1.1, supérieur à 0,50 et une p-valeur associée à un log-vraisemblance indiquant une corrélation conditionnelle possible inférieure à 0,01.

De ce groupe restreint, une corrélation entre deux ligases a été retenue pour analyse.

5.4.1 Ligases

Parmi les résultats intéressants conditionnels à l'hyperthermophilie, un triplet de résultats a montré une corrélation conditionnelle possible entre deux protéines impliquées dans un rôle de ligase. Il s'agit d'une ligase ADN ATP-dépendante, arCOG04218, et d'une nucléotidyltransférase prédite, arCOG01204. Leur profil de présence et d'absence est indiqué à la figure 5.6. Leur complémentarité établie par l'équation 1.1 est de 0,8. Le tableau 5.VII indique clairement une valeur de log-vraisemblance minimale penchant vers une relation corrélée conditionnelle et cette valeur est supportée par une p-valeur de 0,01.

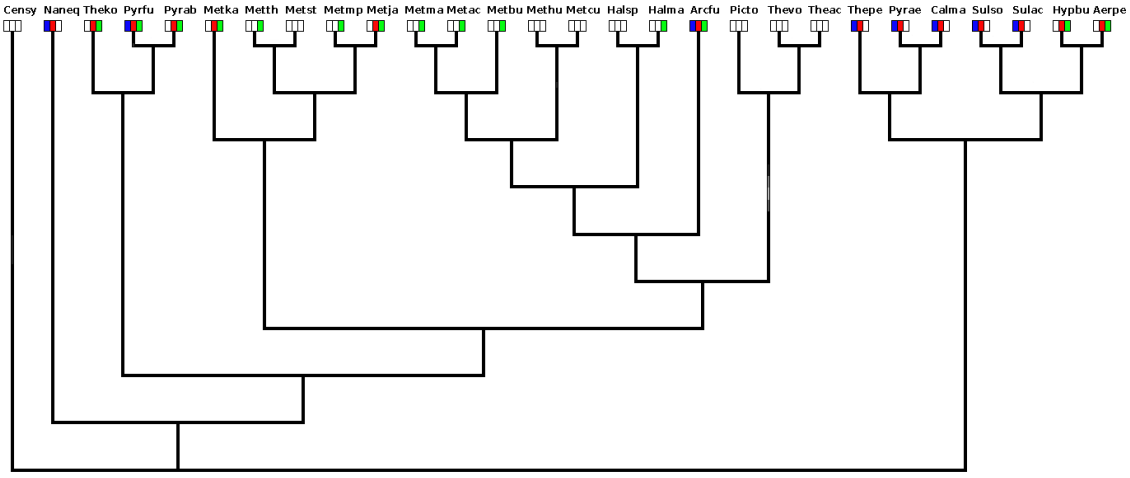


Figure 5.6 – Arbre phylogénétique présentant le profil phylétique aux feuilles de deux ligases et de la reverse gyrase chez différentes espèces d'archaea. Un carré blanc représente l'absence d'un gène et un carré de couleur représente la présence d'un gène. En bleu : nucléotidyltransférase, en rouge : reverse gyrase, en vert : ligase ADN ATP-dépendante.

La ligase ADN ATP-dépendante, surtout présente chez les euryarchaeota, arCOG-04218, est définie comme étant membre de la famille L, soit celle englobant les protéines de réplication, de recombinaison et de réparation. Elle est aussi décrite comme un homologue de la ligase III eucaryote. De plus, elle est la descendante présumée de la nucléotidyltransférase [59].

Tableau 5.VII – Valeurs de log-vraisemblance et taux de transition des pertes des gènes Y et Y' en présence individuelle (ν) et en paire (δ) entre deux ligases établissant des liens de corrélation.

Y1	Y2	Log-vraisemblance	p-valeur	δ	ν
arCOG04218	arCOG01204	Valeurs non-corrélées inconditionnelles			
		-44,33	s.o.	0,99	s.o.
		Valeurs corrélées inconditionnelles			
		-42,17	0,04	3,86	0,64
		Valeurs corrélées conditionnelles			
		-40,65	0,01	3,11	0,8

arCOG04218 : ligase ADN ATP-dépendante ; arCOG01204 : nucléotidyltransférase prédite

La ligase ADN III eucaryote forme un complexe avec la protéine de réparation de l'ADN XRCC1, permettant le scellage soit de l'ADN suite à l'excision d'un nucléotide, soit de fragments recombinants [59]. Par contre, un homologue de XRCC1 n'a pas été identifié chez les espèces d'archaea possédant cette ligase spécifique suite à une recherche par BLAST d'une séquence consensus des XRCC1 eucaryotes et procaryotes connus. Ceci peut être en accord avec la théorie puisque seule la partie alpha de cette enzyme interagit avec XRCC1 et la nucléotidyltransférase correspondrait, selon les bases des données courantes (entre autres, pfam, interpro et uniprot), à la partie bêta, impliquant que d'autres acteurs peuvent interagir avec cette ligase [59]. De plus, tous les isoformes connus de cette ligase ATP-dépendante possèdent un motif de doigt de zinc aidant à la stabilisation du repliement protéique ou à la liaison d'éléments structurels secondaires tels que les sites endommagés de l'ADN [59].

La nucléotidyltransférase prédite, arCOG01204, quant à elle, est définie comme étant membre de la famille R, soit celle englobant des fonctions prédites générales seulement. Elle est identifiée comme étant principalement présente chez les crenarchaeota. Tout comme la ligase ADN ATP-dépendante, la nucléotidyltransférase est une polymérase qui est une composante de la voie de réparation de l'ADN suite à l'excision d'un nucléotide.

Le tableau 5.VIII propose des détails visant à départager les espèces d'archaea selon leur température optimale. La moyenne de température optimale de croissance des es-

Tableau 5.VIII – Température et pH minimal, maximal et optimal, présence de la reverse gyrase et domaine d'appartenance de différents archaea [40].

Surnom	Temp. min	Temp. max	Temp. opt	RG	pH min	pH max	Domaine
Censy	8	18	10	0	s.o.	s.o.	Cren
Naneq	75	98	90	1	5.5	6	Nano
Theko	60	100	86	1	5	9	Eury
Pyrfu	70	103	100	1	5	9	Eury
Pyrab	70	103	102	1	7	7	Eury
Metka	84	110	98	1	5.5	7	Eury
Metth	35	70	65	0	6	8.5	Eury
Metst	36	40	36	0	s.o.	s.o.	Eury
Metmp	20	40	35	0	6.5	8	Eury
Metja	48	94	94	1	5.2	7	Eury
Metma	s.o.	s.o.	37	0	5.5	8	Eury
Metac	35	40	35	0	6.5	7	Eury
Metbu	0	28	23	0	s.o.	s.o.	Eury
Methu	s.o.	s.o.	37	0	6.6	7.4	Eury
Metcu	15	45	15	0	s.o.	s.o.	Eury
Halsp	s.o.	s.o.	37	0	s.o.	s.o.	Eury
Halma	s.o.	s.o.	30	0	2	4.5	Eury
Arcfu	60	95	83	1	s.o.	s.o.	Eury
Picto	55	90	65	0	4	-0.2	Eury
Thevo	40	70	60	0	1	4	Eury
Theac	55	60	55	0	0.5	4	Eury
Thepe	67	93	85	1	5	6	Cren
Pyrae	75	104	100	1	5.5	9	Cren
Calma	60	92	85	1	3.7	4.2	Cren
Sulso	55	90	75	1	0.9	5.8	Cren
Sulac	55	85	80	1	0.9	5.9	Cren
Hypbu	95	107	95	1	7	7	Cren
Aerpe	90	95	90	1	5	7	Cren

RG : Reverse Gyrase, 1 : présence, 0 : absence.

pèces possédant la ligase ADN ATP-dépendante est de $96,5\,°C$ alors que celles possédant la nucléotidyltransférase ont une température moyenne de $85,83\,°C$. Cette variation de $10\,°C$ permet d'envisager que les espèces ont une ligase utilisant l'ATP acquise au cours d'une évolution vers des températures plus élevées. Hypothétiquement, cette li-

gase est plus thermostable qu'une nucléotidyltransférase. Il est d'autant plus intéressant de noter que, pour l'échantillon d'espèces utilisé, les euryarchaeota hyperthermophiles ont une température de croissance optimale moyenne supérieure de 10 °C par rapport aux crenarchaeota qui, eux, sont strictement hyperthermophiles.

Une étude faite sur la ligase de *Methanobacterium thermoautotrophicum*, possédant la ligase ADN ATP-dépendante, montre que cette ligase est thermophilique [85]. Cette ligase est très sensible à la température et est active à une température optimale de $60 - 70$ °C. Cette étude montre aussi que cette ligase atteint son activité optimale à un pH se situant entre 7,5 et 8.8 [85]. Une étude similaire faite sur *Pyrococcus horikoshii*, possédant la ligase ADN ATP-dépendante, montre que l'activité adénylyltransférase se fait à une température optimale de 90 °C et que l'assemblage des bouts (« nick-joining ») se fait entre 70 et 90 °C [43]. Une étude sur *Pyrobaculum aerophilum*, possédant la nucléotidyltransférase, a montré que son activité de réparation était optimale à une température de 60 °C [77]. Ceci met de l'emphase sur l'hypothèse que la ligase ATP-dépendante a une plus haute température optimale possiblement reliée à sa stabilité ou à son activité enzymatique que sa contre-partie nucléotidyltransférase.

De plus, la ligase ATP-dépendante est incapable d'utiliser l'adénosine diphosphate (ADP) ou le nicotinamide adénine dinucléotide (NAD) pour initier la liaison, apportant une preuve supplémentaire pour l'hypothèse de la thermostabilité de cette enzyme à de plus hautes températures [43, 85]. D'un point de vue évolutif, ceci suggère que les ligases ATP et NAD-dépendantes peuvent avoir évoluées d'un ancêtre commun par l'acquisition d'éléments protéiques structuraux qui interagissent avec le phosphate gamma de l'ATP ou du nicotinamide ribonucléoside du NAD. Dans ce cas, la ligase originale pourrait avoir utilisé l'ADP comme substrat intermédiaire. Dans cette optique, les travaux démontrant que la ligase de *Aeropyrum pernix* catalysant la liaison, soit en présence d'ATP ou d'ADP, mais pas en présence de NAD ou de l'adénosine monophosphate (AMP), peut être importante dans l'histoire évolutive reliant les ligases ATP-dépendantes et la nucléotidyltransférase [43].

L'analyse du pH montre aussi une histoire évolutive débutant avec l'utilisation d'une nucléotidyltransférase pour éventuellement utiliser une ligase ADN ATP-dépendante

plus complexe. L'ATP est stable entre un pH de 6.8 et 7.4. Le pH moyen de la nucléotidyltransférase se situe entre 3.58 et 6.15 tandis que celle de l'ATP-dépendante est entre 5.78 et 7.33. Il faut noter que ces valeurs de pH dénotent le pH nécessaire au milieu de croissance et non le pH cytoplasmique qui peut être régulé et donc différent. En se basant sur ces données concernant la température et le pH, il est possible de croire qu'il y a eu une évolution de la nucléotidyltransférase vers une ligase ADN ATP-dépendante.

En conclusion, au sujet de l'évolution possible de la ligase chez les espèces hyperthermophiles, il est possible d'émettre l'hypothèse qu'en s'adaptant à des températures plus élevées, la nucléotidyltransférase aurait acquis l'usage de l'ATP comme source de phosphate dans des environnements qui en plus possèdent un pH plus propice à sa stabilité. De plus, puisque le taux de NaCl influence aussi grandement la stabilité de l'ATP et des ligases, des études supplémentaires en ce sens seraient sûrement hautement intéressantes.

CHAPITRE 6

CONCLUSION

Ce mémoire s'est intéressé à développer une généralisation de l'analyse de caractères discrets d'un profil phylétique à deux états pour trois états. À l'origine de ce travail, l'un des objectifs était d'appliquer la recherche de corrélation de deux gènes à un phénotype duquel ils peuvent être conditionnels. Afin d'être en mesure de déterminer cette conditionnalité il faut, dans un premier temps, mettre au point une méthode afin de regrouper les gènes par homologie. Un algorithme de recherche d'orthologues par ressemblance symétrique a donc été mis au point. Cependant, l'utilisation de celui-ci, bien qu'adéquate, a révélé quelques lacunes dans le cadre d'une expérimentation reproductible. En conséquence, une base de données publique de gènes orthologues a été préférée. De plus, il est nécessaire de s'intéresser et comprendre divers phénotypes et les cascades de gènes impliqués dans ceux-ci. C'est pourquoi la motilité et l'hyperthermophilie ont été étudiées comme modèles typiques des archaea. Dans le but de déceler des déplacements de gènes non orthologues, ces études se sont concentrées sur des profils montrant un niveau donné de complémentarité. En ce sens, les buts ont bel et bien été atteints. Une histoire évolutive a été imputée à des gènes impliqués dans le transport membranaire et à des hydrolases leur conférant des liens avec un gène de la motilité. Le cheminement de deux ligases ayant un lien avec l'hyperthermophile a aussi été proposé. Ceci a permis de corroborer des études antérieures sur ces ligases.

Ce mémoire est donc divisé en deux parties. L'une est dédiée à généraliser des algorithmes et des théorèmes mathématiques et l'autre est consacrée à appliquer l'information acquise à des concepts biologiques. Un résumé de l'apport de ces recherches ainsi que les perspectives qu'elles engendrent sont présentés aux sections suivantes.

6.1 Contributions

L'algorithme de regroupement de gènes homologues défini au chapitre 3 a montré que le regroupement par meilleures ressemblances symétriques était une bonne base à une classification fiable des gènes. La comparaison avec des méthodes plus raffinées a fait ressortir des problèmes inhérents à la complexité des liens entre les gènes, principalement entre les orthologues et les paralogues. Par contre, l'utilisation de l'algorithme d'appartenance-union s'est avérée être très efficace pour effectuer les regroupements suite au couplage par BLAST. Ceci a permis de prendre l'information de la génomique comparative sur le groupement d'homologues pour obtenir des profils phylétiques tels que décrits au chapitre 1.

Le chapitre 4 s'est consacré à la généralisation de la recherche de corrélation pour trois caractères discrets à partir d'un profil phylétique acquis par le regroupement et la classification d'homologues. Le chapitre 2 a montré l'importance de cette généralisation pour l'analyse plus sensible de la classification de gènes en modules fonctionnels par profil phylétique sur des exemples concrets de mécanismes menant à des phénotypes que peuvent avoir les archaea.

L'approche proposée au chapitre 4, inspirée des travaux sur une modélisation à deux états, a pris en compte tous les éléments disponibles. Ces éléments étant les profils phylétiques ainsi que l'arbre phylogénétique montrant des taux de mutation optimisés à chacune de ses branches. L'utilisation de modèles de Markov phylogénétiques s'est avérée appropriée pour de multiple raisons. Premièrement, l'héritage que laisse l'évolution n'a pas de souvenir et dépend seulement des parents et non de l'histoire évolutive complète. Deuxièmement, les taux de mutation diffèrent indépendamment à chaque branche de l'arbre. Il a donc été possible de créer un modèle probabiliste d'inférence de corrélation qui en plus incorpore une notion de conditionnalité. Des modèles mathématiques ont de ce fait été établis pour correspondre aux possibilités qu'a pu prendre un triplet donné de gènes. Ces modèles représentent les cas où ce triplet aurait suivi une évolution non conditionnelle et non corrélée, non conditionnelle mais corrélée, ou conditionnelle et corrélée. À chaque fois, le meilleur modèle a été retenu suite à une sélection par un test

de χ^2.

Des comparaisons sur des données connues et sur d'autres données dont la corrélation pouvait être légitimement anticipée ont démontré que ce modèle mathématique probabiliste constitue une première analyse adéquate. Il est donc convenable de l'utiliser pour explorer des profils réels de génotype ou phénotype d'organismes possédant des relations inexpliquées. Faisant suite à des preuves apportées par des travaux antérieurs permettant d'espérer que des relations conditionnelles corrélées soient observées, les phénotypes de motilité et d'hyperthermophilie ont été retenus. Le modèle a donc été appliqué sur la totalité de la base de données arCOG pour ces deux phénotypes. Ceci a permis dans les deux cas de calculer la probabilité de plus d'un million de paires de gènes en un temps acceptable. Une recherche d'une telle ampleur serait absurde si elle était entreprise en utilisant uniquement des méthodes classiques de laboratoire. L'analyse d'une partie de la masse imposante de données ainsi obtenues a permis d'établir des hypothèses quant à l'histoire évolutive qu'aurait pu suivre deux groupes de gènes par rapport au phénotype de motilité et d'hypertermophilie.

Dans le cas de l'étude sur la motilité, des indices sur l'histoire évolutive de la flagelline FlaD/E ont été apportés par le modèle probabiliste. Ceci a permis d'émettre l'hypothèse d'une évolution de différents gènes de transport ABC, d'hydrolases et de méthyltransférases vers une protéine FlaD/E. Elle vient apporter une ligne d'étude possible pour de plus amples recherches.

Dans le cas de l'étude sur l'hyperthermophilie, une histoire évolutive de deux ligases s'est détachée. Le modèle probabiliste ici développé a permis d'effectuer une étude entièrement indépendante arrivant aux mêmes conclusions que des études biologiques classiques. Le dénouement de ces deux études distinctes mène à la conclusion que la ligase ADN ATP-dépendante est la descendante de la nucléotidyltransférase. Le recoupement de ces études vient corroborer cette hypothèse.

Cette méthode apporte donc une quantité d'information qui à elle seule est insuffisante pour établir une quelconque hypothèse solide et autonome. Cette méthode s'est montré statistiquement faible, mais offre tout de même un test qui permet de rechercher des paires de gènes intéressants pour une analyse approfondie sur un phénotype donné.

Lorsqu'elle confère une corrélation à un gène à la fonction encore inconnue, peu d'information immédiatement interprétable est réellement générée. Par contre, lorsqu'elle établit des corrélations pour des gènes ayant des fonctions spécifiques *a priori* non liées à l'état conditionnel, comme dans le cas de la motilité, des hypothèses valides peuvent être émises. Le présent travail établit donc une base probabiliste qui peut s'avérer un guide fiable permettant de restreindre les recherches et les analyses futures que pourraient effectuer des biologistes en laboratoire directement sur les espèces étudiées. Dans une optique plus globale, la collaboration interdisciplinaire devient de plus en plus essentielle. Pour ce type de recherche, en limitant le nombre de gènes à l'étude à quelques dizaines, et non pas à quelques millions, les méthodes classiques en laboratoire demeurent pertinentes.

6.2 Perspectives

Le modèle probabiliste aide à rechercher une corrélation entre les gènes pour des espèces d'une phylogénie donnée. Dans plusieurs cas, la différentiation entre un modèle corrélé ou non n'est pas significative. Dans le cas où les log-vraisemblances des deux modèles corrélés sont supérieurs à celui non corrélé et que les valeurs des deux premiers modèles corrélés sont proches, c'est-à-dire que la différence relative entre leurs p-valeurs est en deçà d'une valeur déterminée, cela peut simplement signifier que le profil de X n'est pas suffisamment accentué sur la phylogénie. Il sera alors nécessaire d'augmenter la diversité en ajoutant d'autres espèces à la phylogénie.

Il est important de réaliser que la méthode ici proposée n'est pas nécessairement optimale. Le modèle mathématique utilise $3\times$ *le nombre de feuilles* bits d'information pour l'inférence. Dans les exemples utilisés dans ce mémoire, ceci représente 3×28 bits, ce qui est peu. Il est donc possible que l'estimation des taux de mutation soit considérablement erronée, puisqu'une seule famille est utilisée pour estimer la perte de X, μ, et seulement deux familles sont utilisées pour les taux de perte et de gain d'Y et Y', ν, λ et δ. Une prochaine étape dans l'avancement de ce projet serait d'utiliser une très grande quantité de triplets comme échantillon. Ceci pourrait être fait en définissant un certain

nombre de classes de dépendance définies par le modèle et les taux associés et en performant un groupage sur des paires de familles, avec X fixe, en estimant les paramètres des catégories simultanément. Malgré l'erreur d'estimation des taux, la question importante est plutôt celle d'inférence de corrélation. Ceci vient à déterminer s'il est possible de déduire l'existence de corrélation en trouvant un meilleur ajustement d'un modèle corrélé à l'ajustement d'un modèle non corrélé. Le but de ces expériences sur de grands ensembles de triplets sera d'évaluer si les p-valeurs obtenues par le test de χ^2 sont de bons indicateurs d'évolution corrélée.

Le chemin qu'a emprunté ce mémoire part de profils phylétiques et de modèles mathématiques probabilistes pour mener à un algorithme plus général d'inférence de corrélation conditionnelle d'un système de trois gènes. Dans le futur, il serait intéressant de voir ce type d'algorithme renforcé et utilisé pour des études sur la coévolution et la corrélation à partir de modules complexes de gènes fonctionnels.

BIBLIOGRAPHIE

[1] S F Altschul, W Gish, W Miller, E W Myers et D J Lipman. Basic local alignment search tool. J Mol Biol, 215(3):403–10, Oct 1990. URL http://view.ncbi.nlm.nih.gov/pubmed/2231712.

[2] Alexey V Antonov et Hans W Mewes. Complex phylogenetic profiling reveals fundamental genotype-phenotype associations. Comput Biol Chem, 32(6):412–6, Dec 2008. URL http://view.ncbi.nlm.nih.gov/pubmed/18753010.

[3] L Aravind. Guilt by association : contextual information in genome analysis. Genome Res, 10(8):1074–7, Aug 2000. URL http://view.ncbi.nlm.nih.gov/pubmed/10958625.

[4] Sonia L Bardy, Sandy Y M Ng et Ken F Jarrell. Prokaryotic motility structures. Microbiology, 149(Pt 2):295–304, Feb 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/12624192.

[5] Daniel Barker, Andrew Meade et Mark Pagel. Constrained models of evolution lead to improved prediction of functional linkage from correlated gain and loss of genes. Bioinformatics, 23(1):14–20, Jan 2007. URL http://view.ncbi.nlm.nih.gov/pubmed/17090580.

[6] Lewis Eh Bingle, Christopher M Bailey et Mark J Pallen. Type vi secretion : a beginner's guide. Curr Opin Microbiol, 11(1):3–8, Feb 2008. URL http://view.ncbi.nlm.nih.gov/pubmed/18289922.

[7] Gilles Brassard et Paul Bratley. Fundamentals of algorithmics. Prentice-Hall, 1996. ISBN 0-13-335068-1.

[8] Richard P Brent. An algorithm with guaranteed convergence for finding a zero of a function. The Computer Journal, 14(4):422, 1971. URL http://comjnl.oxfordjournals.org/cgi/content/abstract/14/4/422.

[9] Richard P Brent. Algorithms for Minimization Without Derivatives. Prentice-Hall, 1973. ISBN 0-486-41998-3.

[10] Monica Campillos, Christian von Mering, Lars Juhl Jensen et Peer Bork. Identification and analysis of evolutionarily cohesive functional modules in protein networks. Genome Res, 16(3):374–82, Mar 2006. URL http://view.ncbi.nlm.nih.gov/pubmed/16449501.

[11] T Cavalier-Smith. The evolutionary origin and phylogeny of eukaryote flagella. Symp Soc Exp Biol, 35:465–93, 1982. URL http://view.ncbi.nlm.nih.gov/pubmed/6764046.

[12] T Cavalier-Smith. The origin of eukaryotic and archaebacterial cells. Ann N Y Acad Sci, 503:17–54, 1987. URL http://view.ncbi.nlm.nih.gov/pubmed/3113314.

[13] N Cermakian, T M Ikeda, P Miramontes, B F Lang, M W Gray et R Cedergren. On the evolution of the single-subunit rna polymerases. J Mol Evol, 45(6):671–81, Dec 1997. URL http://view.ncbi.nlm.nih.gov/pubmed/9419244.

[14] J J Champoux. Dna topoisomerases : structure, function, and mechanism. Annu Rev Biochem, 70:369–413, 2001. URL http://view.ncbi.nlm.nih.gov/pubmed/11395412.

[15] Jang-Hee Cho, Younghee Park, Joong-Hoon Ahn, Yoongho Lim et Sangkee Rhee. Structural and functional insights into o-methyltransferase from bacillus cereus. J Mol Biol, 382(4):987–97, Oct 2008. URL http://view.ncbi.nlm.nih.gov/pubmed/18706426.

[16] Miklos Csuros et Istvan Miklos. Streamlining and large ancestral genomes in archaea inferred with a phylogenetic birth-and-death model. Mol Biol Evol, 26(9):2087–95, Sep 2009. URL http://view.ncbi.nlm.nih.gov/pubmed/19570746.

[17] Elie Desmond, Celine Brochier-Armanet et Simonetta Gribaldo. Phylogenomics of the archaeal flagellum : rare horizontal gene transfer in a unique motility structure. BMC Evol Biol, 7:106, 2007. URL http://view.ncbi.nlm.nih.gov/pubmed/17605801.

[18] Christophe Dessimoz, Brigitte Boeckmann, Alexander C J Roth et Gaston H Gonnet. Detecting non-orthology in the cogs database and other approaches grouping orthologs using genome-specific best hits. Nucleic Acids Res, 34(11):3309–16, 2006. URL http://view.ncbi.nlm.nih.gov/pubmed/16835308.

[19] S Djordjevic, P N Goudreau, Q Xu, A M Stock et A H West. Structural basis for methylesterase cheb regulation by a phosphorylation-activated domain. Proc Natl Acad Sci U S A, 95(4):1381–6, Feb 1998. URL http://view.ncbi.nlm.nih.gov/pubmed/9465023.

[20] S Djordjevic et A M Stock. Crystal structure of the chemotaxis receptor methyltransferase cher suggests a conserved structural motif for binding s-adenosylmethionine. Structure, 5(4):545–58, Apr 1997. URL http://view.ncbi.nlm.nih.gov/pubmed/9115443.

[21] J L Dynes et R A Firtel. Molecular complementation of a genetic marker in dictyostelium using a genomic dna library. Proc Natl Acad Sci U S A, 86(20):7966–70, Oct 1989. URL http://view.ncbi.nlm.nih.gov/pubmed/2813371.

[22] Richard V Eck et Margaret O Dayhoff. Evolution of the structure of ferredoxin based on living relics of primitive amino acid sequences. Science, 152(3720):363–366, Apr 1966. URL http://view.ncbi.nlm.nih.gov/pubmed/17775169.

[23] Robert C Edgar. Muscle : multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res, 32(5):1792–7, 2004. URL http://view.ncbi.nlm.nih.gov/pubmed/15034147.

[24] Thijs J G Ettema, Willem M de Vos et John van der Oost. Discovering novel biology by in silico archaeology. Nat Rev Microbiol, 3(11):859–69, Nov 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/16175172.

[25] D M Faguy, K F Jarrell, J Kuzio et M L Kalmokoff. Molecular analysis of archael flagellins : similarity to the type iv pilin-transport superfamily widespread in bacteria. Can J Microbiol, 40(1):67–71, Jan 1994. URL http://view.ncbi.nlm.nih.gov/pubmed/7908603.

[26] J Felsenstein. Evolutionary trees from dna sequences : a maximum likelihood approach. J Mol Evol, 17(6):368–76, 1981. URL http://view.ncbi.nlm.nih.gov/pubmed/7288891.

[27] Joseph Felsenstein. Inferring phylogenies. Sinauer Associates, 2004. ISBN 0-87893-177-5.

[28] W M Fitch et E Margoliash. Construction of phylogenetic trees. Science, 155(760): 279–84, Jan 1967. URL http://view.ncbi.nlm.nih.gov/pubmed/5334057.

[29] P Forterre et H Philippe. Where is the root of the universal tree of life ? Bioessays, 21(10):871–9, Oct 1999. URL http://view.ncbi.nlm.nih.gov/pubmed/10497338.

[30] Patrick Forterre. A hot story from comparative genomics : reverse gyrase is the only hyperthermophile-specific protein. Trends Genet, 18(5):236–7, May 2002. URL http://view.ncbi.nlm.nih.gov/pubmed/12047940.

[31] Toni Gabaldon. Evolution of proteins and proteomes : a phylogenetics approach. Evol Bioinform Online, 1:51–61, 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/19325853.

[32] M Y Galperin et E V Koonin. Who's your neighbor ? new computational approaches for functional genomics. Nat Biotechnol, 18(6):609–13, Jun 2000. URL http://view.ncbi.nlm.nih.gov/pubmed/10835597.

[33] N Goldman. Statistical tests of models of dna substitution. J Mol Evol, 36(2): 182–98, Feb 1993. URL http://view.ncbi.nlm.nih.gov/pubmed/7679448.

[34] M W Gray et B F Lang. Transcription in chloroplasts and mitochondria : a tale of two polymerases. Trends Microbiol, 6(1):1–3, Jan 1998. URL http://view.ncbi.nlm.nih.gov/pubmed/9481814.

[35] Stephane Guindon et Olivier Gascuel. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Syst Biol, 52(5):696–704, Oct 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/14530136.

[36] M Hasegawa, H Kishino et T Yano. Dating of the human-ape splitting by a molecular clock of mitochondrial dna. J Mol Evol, 22(2):160–74, 1985. URL http://view.ncbi.nlm.nih.gov/pubmed/3934395.

[37] Michelle Heine et Sathees B C Chandra. The linkage between reverse gyrase and hyperthermophiles : a review of their invariable association. J Microbiol, 47 (3):229–34, Jun 2009. URL http://view.ncbi.nlm.nih.gov/pubmed/19557338.

[38] Ian R Henderson, Fernando Navarro-Garcia, Mickael Desvaux, Rachel C Fernandez et Dlawer Ala'Aldeen. Type v protein secretion pathway : the autotransporter story. Microbiol Mol Biol Rev, 68(4):692–744, Dec 2004. URL http://view.ncbi.nlm.nih.gov/pubmed/15590781.

[39] Wolfram Research Inc. Mathematica Edition : Version 7.0. Wolfram Research, Inc., Wolfram Research, Inc., 2008.

[40] DOE Joint Genome Institute. The Regents of the University of California, 2010. URL http://www.jgi.doe.gov/.

[41] T H Jukes et C R Cantor. Evolution of protein molecule in H. N. Munro, ed. Mammalian protein metabolism, volume III. New York : Academic Press, 1969.

[42] Philip R Kensche, Vera van Noort, Bas E Dutilh et Martijn A Huynen. Practical and theoretical advances in predicting the function of a protein by its phylogenetic distribution. J R Soc Interface, 5(19):151–70, Feb 2008. URL http://view.ncbi.nlm.nih.gov/pubmed/17535793.

[43] Niroshika Keppetipola et Stewart Shuman. Characterization of a thermophilic atp-dependent dna ligase from the euryarchaeon pyrococcus horikoshii. J Bacteriol, 187(20):6902–8, Oct 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/16199559.

[44] M Kimura. A simple method for estimating evolutionary rates of base substitutions through comparative studies of nucleotide sequences. J Mol Evol, 16(2):111–20, Dec 1980. URL http://view.ncbi.nlm.nih.gov/pubmed/7463489.

[45] E V Koonin. How many genes can make a cell : the minimal-gene-set concept. Annu Rev Genomics Hum Genet, 1:99–116, 2000. URL http://view.ncbi.nlm.nih.gov/pubmed/11701626.

[46] E V Koonin, A R Mushegian et P Bork. Non-orthologous gene displacement. Trends Genet, 12(9):334–6, Sep 1996. URL http://view.ncbi.nlm.nih.gov/pubmed/8855656.

[47] Eugene V Koonin. Comparative genomics, minimal gene-sets and the last universal common ancestor. Nat Rev Microbiol, 1(2):127–36, Nov 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/15035042.

[48] Eugene V Koonin. Orthologs, paralogs, and evolutionary genomics. Annu Rev Genet, 39:309–38, 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/16285863.

[49] O I Kovalsky, S A Kozyavkin et A I Slesarev. Archaebacterial reverse gyrase cleavage-site specificity is similar to that of eubacterial dna topoisomerases i. Nucleic Acids Res, 18(9):2801–5, May 1990. URL http://view.ncbi.nlm.nih.gov/pubmed/2160070.

[50] Piotr Z Kozbial et Arcady R Mushegian. Natural history of s-adenosylmethionine-binding proteins. BMC Struct Biol, 5:19, 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/16225687.

[51] S A Kozyavkin, R Krah, M Gellert, K O Stetter, J A Lake et A I Slesarev. A reverse gyrase with an unusual structure. a type i dna topoisomerase from the hyperthermophile methanopyrus kandleri is a two-subunit protein. J Biol Chem, 269(15): 11081–9, Apr 1994. URL http://view.ncbi.nlm.nih.gov/pubmed/8157633.

[52] DA Liberles, A Thoren, G von Heijne et A Elofsson. The use of phylogenetic profiles for gene predictions. Current Genomics, 3(3):131, June 2002.

[53] P Lio et N Goldman. Models of molecular evolution and phylogeny. Genome Res, 8(12):1233–44, Dec 1998. URL http://view.ncbi.nlm.nih.gov/pubmed/9872979.

[54] W A M Loenen. S-adenosylmethionine : jack of all trades and master of everything ? Biochem Soc Trans, 34(Pt 2):330–3, Apr 2006. URL http://view.ncbi.nlm.nih.gov/pubmed/16545107.

[55] L J Magrum, K R Luehrsen et C R Woese. Are extreme halophiles actually "bacteria" ? J Mol Evol, 11(1):1–8, May 1978. URL http://view.ncbi.nlm.nih.gov/pubmed/660662.

[56] Mahmood A Mahdavi et Yen-Han Lin. Prediction of protein-protein interactions using protein signature profiling. Genomics Proteomics Bioinformatics, 5(3-4): 177–86, Dec 2007. URL http://view.ncbi.nlm.nih.gov/pubmed/18267299.

[57] Kira S Makarova, Alexander V Sorokin, Pavel S Novichkov, Yuri I Wolf et Eugene V Koonin. Clusters of orthologous genes for 41 archaeal genomes and implications for evolutionary genomics of archaea. Biol Direct, 2:33, 2007. URL http://view.ncbi.nlm.nih.gov/pubmed/18042280.

[58] Kira S Makarova, Yuri I Wolf et Eugene V Koonin. Potential genomic determinants of hyperthermophily. Trends Genet, 19(4):172–6, Apr 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/12683966.

[59] Ina V Martin et Stuart A MacNeill. Atp-dependent dna ligases. Genome Biol, 3(4):REVIEWS3005, 2002. URL http://view.ncbi.nlm.nih.gov/pubmed/11983065.

[60] A L Metlina. Bacterial and archaeal flagella as prokaryotic motility organelles. Biochemistry (Mosc), 69(11):1203–12, Nov 2004. URL http://view.ncbi.nlm.nih.gov/pubmed/15627373.

[61] Boris G Mirkin, Trevor I Fenner, Michael Y Galperin et Eugene V Koonin. 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, Jan 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/12515582.

[62] A R Mushegian et E V Koonin. A minimal gene set for cellular life derived by comparison of complete bacterial genomes. Proc Natl Acad Sci U S A, 93(19):10268–73, Sep 1996. URL http://view.ncbi.nlm.nih.gov/pubmed/8816789.

[63] Alessandra Napoli, Anna Valenti, Vincenzo Salerno, Marc Nadal, Florence Garnier, Mose Rossi et Maria Ciaramella. Functional interaction of reverse gyrase with single-strand binding protein of the archaeon sulfolobus. Nucleic Acids Res, 33(2):564–76, 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/15673717.

[64] D A Natale, U T Shankavaram, M Y Galperin, Y I Wolf, L Aravind et E V Koonin. Towards understanding the first genome sequence of a crenarchaeon by genome annotation using clusters of orthologous groups of proteins (cogs). Genome Biol,

1(5):RESEARCH0009, 2000. URL http://view.ncbi.nlm.nih.gov/pubmed/11178258.

[65] Sandy Y M Ng, Behnam Zolghadr, Arnold J M Driessen, Sonja-Verena Albers et Ken F Jarrell. Cell surface structures of archaea. J Bacteriol, 190(18):6039–47, Sep 2008. URL http://view.ncbi.nlm.nih.gov/pubmed/18621894.

[66] Richard A Notebaart, Martijn A Huynen, Bas Teusink, Roland J Siezen et Berend Snel. Correlation between sequence conservation and the genomic context after gene duplication. Nucleic Acids Res, 33(19):6164–71, 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/16257980.

[67] Kevin P O'Brien, Maido Remm et Erik L L Sonnhammer. Inparanoid : a comprehensive database of eukaryotic orthologs. Nucleic Acids Res, 33(Database issue): D476–80, Jan 2005. URL http://view.ncbi.nlm.nih.gov/pubmed/15608241.

[68] Mark Pagel. Detecting correlated evolution on phylogenies : A general method for the comparative analysis of discrete characters. Biological Sciences, 255(1342): 37, Jan 1994. URL http://www.jstor.org/stable/49836.

[69] Christopher R Peabody, Yong Joon Chung, Ming-Ren Yen, Dominique Vidal-Ingigliardi, Anthony P Pugsley et Milton H Jr Saier. Type ii protein secretion and its relationship to bacterial type iv pili and archaeal flagella. Microbiology, 149(Pt 11):3051–72, Nov 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/14600218.

[70] M Pellegrini, E M Marcotte, M J Thompson, D Eisenberg et T O Yeates. Assigning protein functions by comparative genome analysis : protein phylogenetic profiles. Proc Natl Acad Sci U S A, 96(8):4285–8, Apr 1999. URL http://view.ncbi.nlm.nih.gov/pubmed/10200254.

[71] Giuseppe Perugino, Anna Valenti, Anna D'amaro, Mose Rossi et Maria Ciaramella. Reverse gyrase and genome stability in hyperthermophilic organisms.

Biochem Soc Trans, 37(Pt 1):69–73, Feb 2009. URL http://view.ncbi.nlm.nih.gov/pubmed/19143604.

[72] B Pierce. Genetics : A conceptual approach. W. H. Freeman and Company, 2e édition, 2005.

[73] Julio Polaina et Andrew P MacCabe. Industrial enzymes : Structure, fonction and applications. Springer, 2007. ISBN 978-1-4020-5376-4.

[74] William H Press, Saul A Teukolsky, William T Vetterling et Brian P Flannery. Numerical Recipes in C : The art of scientifique computing. Press syndicate of the University of Cambridge, 1992. ISBN 0-521-43108-5.

[75] A Chapin Rodriguez et Daniela Stock. Crystal structure of reverse gyrase : insights into the positive supercoiling of dna. EMBO J, 21(3):418–26, Feb 2002. URL http://view.ncbi.nlm.nih.gov/pubmed/11823434.

[76] A P Ryle, F Sanger, L F Smith et R Kitai. The disulphide bonds of insulin. Biochem J, 60(4):541–56, Aug 1955. URL http://view.ncbi.nlm.nih.gov/pubmed/13249947.

[77] Alessandro A Sartori et Josef Jiricny. Enzymology of base excision repair in the hyperthermophilic archaeon pyrobaculum aerophilum. J Biol Chem, 278(27):24563–76, Jul 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/12730226.

[78] Matthias Schlesner, Arthur Miller, Stefan Streif, Wilfried F Staudinger, Judith Muller, Beatrix Scheffer, Frank Siedler et Dieter Oesterhelt. Identification of archaea-specific chemotaxis proteins which interact with the flagellar apparatus. BMC Microbiol, 9:56, 2009. URL http://view.ncbi.nlm.nih.gov/pubmed/19291314.

[79] Heidi L Schubert, Robert M Blumenthal et Xiaodong Cheng. Many paths to methyltransfer : a chronicle of convergence. Trends Biochem Sci, 28(6):329–35, Jun 2003. URL http://view.ncbi.nlm.nih.gov/pubmed/12826405.

[80] Robert Sedgewick. Algorithmes en Java. Pearson Education, 3e édition, 2004. ISBN 2-7440-7024-6.

[81] Shi Shu, Dana C Mahadeo, Xiong Liu, Wenli Liu, Carole A Parent et Edward D Korn. S-adenosylhomocysteine hydrolase is localized at the front of chemotaxing cells, suggesting a role for transmethylation during migration. Proc Natl Acad Sci U S A, 103(52):19788–93, Dec 2006. URL http://view.ncbi.nlm.nih.gov/pubmed/17172447.

[82] Berend Snel, Peer Bork et Martijn A Huynen. The identification of functional modules from the genomic association of genes. Proc Natl Acad Sci U S A, 99 (9):5890–5, Apr 2002. URL http://view.ncbi.nlm.nih.gov/pubmed/11983890.

[83] Berend Snel et Martijn A Huynen. Quantifying modularity in the evolution of biomolecular systems. Genome Res, 14(3):391–7, Mar 2004. URL http://view.ncbi.nlm.nih.gov/pubmed/14993205.

[84] Balaji S Srinivasan, Antal F Novak, Jason A Flannick, Serafim Batzoglou et Harley H McAdams. Integrated protein interaction networks for 11 microbes. RECOMB 2006 Proceedings, 2006.

[85] V Sriskanda, Z Kelman, J Hurwitz et S Shuman. Characterization of an atp-dependent dna ligase from the thermophilic archaeon methanobacterium thermoautotrophicum. Nucleic Acids Res, 28(11):2221–8, Jun 2000. URL http://view.ncbi.nlm.nih.gov/pubmed/10871342.

[86] Christian E V Storm et Erik L L Sonnhammer. Automated ortholog inference from phylogenetic trees and calculation of orthology reliability. Bioinformatics, 18(1):92–9, Jan 2002. URL http://view.ncbi.nlm.nih.gov/pubmed/11836216.

[87] James B Sumner. The isolation and crystallization of the enzyme urease. preliminary paper. Journal of Biological Chemistry, 69:435, 1926.

[88] R L Tatusov, M Y Galperin, D A Natale et E V Koonin. The cog database : a tool for genome-scale analysis of protein functions and evolution. Nucleic Acids Res, 28(1):33–6, Jan 2000. URL http://view.ncbi.nlm.nih.gov/pubmed/10592175.

[89] R L Tatusov, E V Koonin et D J Lipman. A genomic perspective on protein families. Science, 278(5338):631–7, Oct 1997. URL http://view.ncbi.nlm.nih.gov/pubmed/9381173.

[90] N A Thomas, S L Bardy et K F Jarrell. The archaeal flagellum : a different kind of prokaryotic motility structure. FEMS Microbiol Rev, 25(2):147–74, Apr 2001. URL http://view.ncbi.nlm.nih.gov/pubmed/11250034.

[91] N A Thomas et K F Jarrell. Characterization of flagellum gene families of methanogenic archaea and localization of novel flagellum accessory proteins. J Bacteriol, 183(24):7154–64, Dec 2001. URL http://view.ncbi.nlm.nih.gov/pubmed/11717274.

[92] J Veser. Kinetics and inhibition studies of catechol o-methyltransferase from the yeast candida tropicalis. J Bacteriol, 169(8):3696–700, Aug 1987. URL http://view.ncbi.nlm.nih.gov/pubmed/3611026.

[93] Tobias Warnecke, Guang-Zhong Wang, Martin J Lercher et Laurence D Hurst. Does negative auto-regulation increase gene duplicability ? BMC Evol Biol, 9:193, 2009. URL http://view.ncbi.nlm.nih.gov/pubmed/19664220.

[94] Jaime Wisniak. Jons jacob berzelius a guide to the perplexed chemist. The Chemical Educator, 5:343, 2000.

[95] C R Woese, O Kandler et M L Wheelis. Towards a natural system of organisms : proposal for the domains archaea, bacteria, and eucarya. Proc Natl Acad Sci U S A, 87(12):4576–9, Jun 1990. URL http://view.ncbi.nlm.nih.gov/pubmed/2112744.

[96] C R Woese, L J Magrum et G E Fox. Archaebacteria. J Mol Evol, 11(3):245–51, Aug 1978. URL http://view.ncbi.nlm.nih.gov/pubmed/691075.

[97] Jianzhi Zhang. Evolution by gene duplication : an update. TRENDS in ecology and evolution, 18(6):292–298, june 2003.