Inférence des interactions entre processus évolutifs

Abdelkader Behdenna

THÈSE DE DOCTORAT DE
l'UNIVERSITÉ PIERRE ET MARIE CURIE

Spécialité

Génétique

École doctorale Complexité du Vivant (Paris)

Présentée par

Abdelkader BEHDENNA

Pour obtenir le grade de
DOCTEUR de l'UNIVERSITÉ PIERRE ET MARIE CURIE

Sujet de la thèse :

Inférence des Interactions entre Processus Évolutifs

soutenue le 14 mars 2016

devant le jury composé de :

Mme. Alessandra Carbone	Présidente
M. Guy Perrière	Rapporteur
M. Thomas Bataillon	Rapporteur
Mme. Sophie Schbath	Examinatrice
M. Julien Dutheil	Examinateur
M. Guillaume Achaz	Directeur de thèse
M. Amaury Lambert	Directeur de thèse

Résumé

Au cours de cette thèse, nous avons développé un outil pour détecter la *coévolution*, c'est à dire l'évolution conjointe de différentes entités biologiques (nucléotides, acides aminés, fonctions biologiques), à différentes échelles (moléculaire, organe). Cet outil s'applique sur des arbres phylogénétiques sur lesquels des évènements évolutifs (mutations, gains/pertes de fonctions biologiques) sont placés. Nous nous plaçons dans un cadre abstrait dans le but de travailler sur les processus conduisant à l'apparition d'évènements évolutifs au sens large le long des lignées d'un arbre phylogénétique. Cet outil est constitué de deux parties distinctes, chacune ayant ses propres spécificités.

D'une part, nous avons produit une première méthode simple, très efficace, permettant de détecter parmi un très grand nombre de tels processus, quelles paires d'évènements semblent apparaître de manière conjointe dans l'arbre. Grâce à un formalisme mathématique utilisant les propriétés de l'algèbre bilinéaire, des calculs exacts d'espérance, de variance et même de distributions de probabilités sont possibles et permettent d'associer à ces paires détectées des *p-values* exactes, rendant cette méthode très précise.

D'autre part, nous avons développé un modèle de coévolution entre de tels processus évolutifs. Ce modèle mathématique limite considérablement le nombre de paramètres utilisés et nous a permis de calculer et d'optimiser une fonction de vraisemblance. Cette optimisation revient à rechercher les paramètres du modèles expliquant au mieux les données contemporaines observées, et nous permet ainsi, toujours selon notre modèle, d'établir le scénario le plus probable ayant mené aux données observées.

Cette seconde méthode est plus gourmande en temps de calcul, ce qui invite à associer les deux méthodes dans un *pipeline* nous permettant de traiter efficacement un grand nombre de paires avant d'aller plus loin dans notre étude et tester les paires les plus encourageantes à l'aide de notre modèle mathématique, dans le but de décrire un scénario interprétable dans un contexte biologique. Nous avons testé cet outil à l'aide de simulations, avant de l'appliquer à deux exemples biologiques très différents : le lien entre intracellularité et perte de flagelle chez *Escherichia coli*, et l'étude de toutes les paires de nucléotides dans des séquences d'ARNr 16S de γ-entérobactéries.

Table des matières

Liste des abréviations

Pour des raisons de cohérence avec la littérature, nous avons choisi de conserver certaines abréviations couramment utilisées en langue anglaise pour certains objets et/ou phénomènes cités dans cette thèse. Ces abréviations, ainsi que leurs traductions françaises, sont listées ci-après.

ADN : Acide DésoxyriboNucléique

ARN : Acide RiboNucléique

ARNr : ARN ribosomique

ARNt : ARN de transfert

CPD : Compensated Pathogenic Deviations (Déviations Pathogènes Compensatoires)

MCMC : Markov Chain Monte Carlo

MSA : Multiple Sequence Alignment (Alignement Multiple de Séquences)

ML : Maximum Likelihood (Maximum de Vraisemblance)

WC : (appariement) Watson-Crick

Table des figures

Remerciements

Avant propos :

Le bio-*-ticien, une histoire de modèles

Avant d'entrer dans le vif du sujet, il me semblait important de parler de la façon dont j'ai pu vivre mon travail au cours de ces années de thèse. Venant initialement de l'informatique fondamentale, de la logique mathématique et de la théorie des graphes, j'ai dû apprendre - avant même d'appréhender totalement mon sujet de thèse - que finalement, la Vie, ce n'est pas juste Vrai ou Faux. Qu'il y a bien évidemment tout un monde entre les deux. Et que finalement, la biologie, ce n'est pas répondre à des questions par *oui* ou *non*, mais plutôt de mesurer la part de *oui* et de *non* pour chacune des questions que l'on peut se poser. Ma vision des choses a donc dû nécessairement *évoluer*.

En arrivant à l'Atelier de BioInformatique, donc, plein de mes certitudes (puisque ma formation initiale était justement à propos de certitudes), j'ai appris à relativiser tout cela, à me poser plus de questions, et même ensuite à questionner mes réponses (quand j'en avais). Et je crois qu'il est là, le travail du biologiste, d'être sans cesse en questionnement, et de toujours trouver de nouveaux prétextes pour douter. Douter de ses théories, douter de ses résultats, douter de ses doutes.

En particulier chez les biologistes de l'Évolution. Notre métier est de remonter le temps, ce qui est tout sauf évident, vous en conviendrez. Partir de données actuelles, et remonter le temps, donc, pour comprendre et savoir ce qui a bien pu se passer pendant les millions d'années qui nous ont précédé sur Terre. La Vie ne nous a évidemment pas attendu pour exister, pour changer, pour *évoluer*, et c'est toute cette vertigineuse histoire, qui constitue notre champ d'investigation.

Donc en résumé, la Vie contemporaine est un domaine extrêmement vaste, et nous ajoutons une couche de *temps* par dessus tout cela. Sur des millions d'années, au moins.

Nous en arrivons donc au bio-*-ticien mystérieux du titre de cet avant-propos. Je parle là du bioinformaticien, du biomathématicien, ou du biostatisticien, qui ont un point commun

d'une grande importance : à partir de ce qu'on peut observer, à partir de données brutes, nous voulons raconter une histoire. Mais une histoire cohérente, robuste, et claire. Et pour cela, nous avons des outils théoriques à notre disposition. Ces outils sont à géométrie très variable, mais ils ont un point commun très marqué : ils consistent à étudier des données trop complexes ou trop nombreuses pour être traitées manuellement.

C'est le cas par exemple du *parsing*, qui a finalement été une partie non négligeable de mon travail ces dernières années, et qui consiste à modifier un texte, au sens général, pour en changer la forme, et y réorganiser, voire sélectionner, les informations. En l'occurrence, la majorité du travail de parsing que j'ai fait revenait à sélectionner correctement des nombres parmi d'autres, dans des fichiers potentiellement très vastes. Reformuler, réorganiser, réorienter l'information pour soit pouvoir la lire, l'utiliser pour construire une figure, ou simplement la mettre au bon format pour utiliser d'autres outils.

Mais bien sûr, l'algorithmique et la programmation sont les principaux outils qui ont fait avancer ma recherche. Ils m'ont permis de faire la synthèse d'intuitions et d'hypothèses biologiques d'une part, et d'un formalisme et de résultats mathématiques d'autre part. En ce qui me concerne, c'est là le point vital de mon travail de bioinformaticien : pouvoir faire cette synthèse en construisant des méthodes efficaces pour répondre à ces intuitions, confirmer ou infirmer ces hypothèses, à l'aide des Mathématiques, science exacte par excellence, mais elle-même incapable d'apporter des réponses finales à toutes les questions que le biologiste se pose.

Et finalement arrivent les modèles, de (co)évolution, en ce qui me concerne. Ces modèles nous permettent de formaliser et d'abstraire certains phénomènes qu'il est impossible de quantifier et décrire exhaustivement. À partir de là, il est possible de faire des calculs, d'améliorer notre compréhension de ces phénomènes, voire de prévoir leur évolution future. Malgré tout, il est nécessaire de garder en tête un point très important : l'abstraction ne fait pas tout, elle ne sert qu'à offrir de nouvelles pistes de réflexion, de prévision. Ces modèles doivent être développés de concert avec un *background* biologique robuste et clair. En d'autres termes, pour que les résultats puissent être interprétés et projetés sur un problème biologique concret, le modèle correspondant se doit de coller à ce problème.

Il faut bien entendre que modéliser un système revient tout d'abord à faire des choix. Finalement, nous n'expliquons pas vraiment la Nature, mais nous faisons des hypothèses, nous essayons d'isoler certaines caractéristiques des systèmes étudiés, et, à l'aide d'un formalisme abstrait, dégageons les grandes tendances *mathématiques* définissant le comportement du modèle. Le comportement du vrai système est approché, certes, mais, à partir du moment

où il est impossible de définir exhaustivement la Nature, toute approximation est par essence fausse. Et notre travail est de limiter cette erreur.

Encore une fois, c'est déjà tout un travail intellectuel d'essayer de faire le lien entre des situations concrètes et complexes, et des modèles parfois très abstraits, parfois simplistes. Mais c'est là l'un des grands intérêts de ce processus de formalisation : le contrôle sur les paramètres est total, et on peut ainsi tenter, non pas d'expliquer parfaitement la Nature, mais au moins de dégager certaines grandes lignes. Et enfin, ces grandes lignes, qui constituent autant d'éléments de réponses, serviront de point de départ à de nouvelles questions...

Première partie

Introduction

Remonter le temps, inférer l'histoire de la vie passée à partir d'une "photographie" contemporaine est un des intérêts de l'application de certaines sciences comme les mathématiques, la statistique ou l'informatique à la biologie : pouvoir associer des valeurs numériques, pouvoir quantifier certains phénomènes, et l'incertitude inhérente à ces phénomènes. En l'occurrence, pour ce qui nous intéresse ici, nous pouvons inférer certains phénomènes qui ont conduit à l'existence de la vie telle qu'on la connaît aujourd'hui, sans nécessairement s'appuyer sur des données fossiles.

Nous allons nous concentrer dans cette introduction sur les méthodes et outils existants pour retracer l'histoire évolutive du vivant, définir et quantifier les scénarios hypothétiques qui ont pu mener à l'observable actuel. En particulier, cette thèse décrit une méthode permettant de détecter et étudier les phénomènes d'évolution conjointe de deux agents (ou plus), à une échelle quelconque.

Il est bien sûr important de s'attacher à évaluer dans quelle mesure nous pouvons avoir *confiance* en nos résultats. Quel que soit le phénomène étudié, il est généralement possible d'imaginer de très nombreux modèles et scénarios pouvant les expliquer. C'est pourquoi il est nécessaire d'une part de garder en tête l'importance de la pertinence biologique des outils utilisés, et de pouvoir comparer ces scénarios, de façon non pas à mettre en avant une seule histoire, mais plutôt de décrire l'histoire (ou les histoires) la (les) plus probable(s).

De nombreuses méthodes existent dans la littérature qui attaquent sous différents angles le problème de la détection de la coévolution. Chacune nécessite de faire des hypothèses différentes, c'est pourquoi il est important de lister certaines de ces méthodes, et les mettre en regard, en comparant leurs avantages et inconvénients.

*La coévolution est le personnage principal de ce travail de thèse. Nous la définirons pour l'instant de façon lapidaire comme étant "l'évolution corrélée de deux entités biologiques" (*e.g. *fonctions biologiques, sites sur un génome,* etc*), avant de la détailler dans le deuxième chapitre de cette introduction. Cette première définition, quoique simple dans sa formulation, pose déjà de nombreuses questions que nous nous proposons d'élucider avant de véritablement entrer dans le vif du sujet. Nous nous attacherons principalement à décomposer ce terme, à commencer par parler d'**Évolution**, puis de **Co**évolution. Nous explorerons ensuite différents mécanismes de coévolution, à diverses échelles, qui nous permettront d'éclaircir et illustrer ce terme d'"entités biologiques".*

Chapitre 1

L'Évolution

L'Évolution est un terme très large regroupant les différents mécanismes conduisant à des transformations des êtres vivants au cours du temps par le biais de divers changements, notamment ceux qui s'expriment au niveau phénotypique et jouent un rôle dans l'histoire de vie des individus. Bien entendu, ces changements dépassent largement ces niveaux d'organisation et ne sont évidemment pas toujours observables à l'œil nu. Plus précisément, ils peuvent être situés au niveau génétique, et s'exprimer ou non de façon à être visibles.

Il est difficile de lister tous ces mécanismes de manière exhaustive, qui sont d'ailleurs plus ou moins appuyés par la théorie et les exemples, aussi allons-nous principalement développer les principaux, que l'on retrouve le plus souvent dans la littérature, et généralement sous-entendus au moment d'évoquer l'Évolution. Mais avant toute chose, faisons un rapide historique des différentes théories et propositions qui ont permis d'aboutir à la pensée évolutive actuelle. Encore une fois, il serait bien trop long de faire une étude historique complète, aussi allons-nous nous intéresser à quelques grandes figures qui ont jalonné l'histoire de cette pensée.

1.1 Un bref historique de la théorie de l'évolution

Avant même l'apparition du terme "Évolution", de nombreux scientifiques et philosophes se sont interrogés sur la façon dont la vie que l'on peut observer aujourd'hui a été façonnée. En effet, d'un point de vue purement scientifique, l'observation de la nature nous pousse à nous poser de nombreuses questions, parmi lesquelles :

— Pourquoi existe-t-il autant de diversité parmi les espèces ?

— Dans quelle mesure les espèces sont-elles adaptées à leur environnement ?
— Pourquoi et comment ces espèces sont-elles si adaptées dans leur environnement ?
— Quels sont les mécanismes qui conduisent à cette adaptation ?

Ces questions sont déjà très orientées, tant il est difficile de s'affranchir intellectuellement des théories actuelles qui apportent leurs réponses. Nous entrevoyons même dans la formulation de notre deuxième question une première piste, en l'occurrence, l'importance de l'environnement. Le terme "adaptation", déjà très lourd de sens, rappelle sans équivoque les théories darwiniennes relative à l'évolution des espèces (Darwin, 1859). Charles Darwin n'est pas le premier à avancer des arguments quant à cette *transformation* des êtres vivants au cours de l'histoire. Dès l'antiquité grecque, romaine ou chinoise, plusieurs philosophes proposent même des mécanismes, ou des scénarios, qui sont pour certains aujourd'hui considérés comme bien connus.

Une des premières traces de l'existence de tels scénarios remonte au VI^{e} siècle av. J.-C. où le philosophe grec Anamixandre, élève de Thalès et professeur de Pythagore, entre autres, élabore une théorie, rapportée au III^{e} siècle par Censorin, selon laquelle *"de l'eau et de la terre réchauffées étaient sortis soit des poissons, soit des animaux tout à fait semblables aux poissons. C'est au sein de ces animaux qu'ont été formés les hommes et que les embryons ont été retenus prisonniers jusqu'à l'âge de la puberté ; alors seulement, après que ces animaux eurent éclaté, en sortirent des hommes et des femmes désormais capables de se sustenter eux-mêmes."*(Mangeart, 1843). En d'autres termes, les hommes sont issus des poissons, ou d'animaux semblables, ce qui constitue là un point de vue qui diverge radicalement avec celui des partisans de l'existence éternelle de l'Homme ou de sa création par une ou plusieurs divinités. Deux siècles plus tard, c'est Aristote, le premier naturaliste dont les écrits nous sont parvenus, qui commence à classifier les êtres vivants selon leur niveau de complexité, dans ce qui peut s'apparenter à une "chaîne des êtres".

Faisons un saut dans le temps, et passons maintenant au Moyen-Âge. La plupart des idées issues des grands penseurs antiques sont perdues dans le monde chrétien, au profit d'une interprétation très littérale de la *Genèse*. Les êtres vivants sont classés selon un *ordre divin* et immuable décrit dans les textes religieux, qui placent Dieu et les anges à son sommet. C'est cette théorie *fixiste* qui perdurera jusqu'au siècle des Lumières et aura une influence notable sur de nombreux scientifiques occidentaux, notamment chez les systématiciens. Les premières bribes de la pensée évolutionniste qui ont émergé dans l'Antiquité n'ont malgré tout pas totalement disparu, puisqu'elles reviennent régulièrement dans les écrits de plusieurs savants musulmans. Un des plus marquants est aussi un des plus anciens, l'écrivain Al-Jahiz, qui expose au IX^{e} siècle dans le *Livre des Animaux* la "lutte pour l'existence", et élabore des

ébauches de raisonnements autour de concepts pouvant s'apparenter à l'influence de l'environnement, voire à la sélection naturelle (Zirkle, 1941). Ces idées sont à l'origine d'un courant que l'on pourrait presque qualifier d'évolutionniste chez les savants musulmans pendant tout le X^{e} siècle et au-delà. Le plus notable est Nasir ad-Din at-Tusi, qui dès le XIII^{e} siècle suggère des notions telles que l'adaptation à l'environnement, la variabilité intra-spécifique (même s'il ne la nomme évidemment pas ainsi), ou l'hérédité (Alakbarli, 2001).

Plusieurs siècles passent en Europe, marqués par la domination des théories fixistes, jusqu'à la Renaissance, notamment René Descartes qui au XVII^{e} développe sa philosophie appelée mécanisme, qui compare l'univers à une machine régie par des liens de cause à effet, ouvrant la voie à certaines théories évoquant une *évolution* de la vie, sans nécessaire intervention divine. Un siècle plus tard, Georges-Louis Buffon avance l'idée selon laquelle certaines espèces sont des variétés issues d'un même ancêtre commun, qui aurait évolué du fait de facteurs environnementaux. Nous n'en sommes bien sûr pas encore aux idées évolutionnistes modernes, puisque pour Buffon, c'est par génération spontanée que sont apparues ces espèces originelles. Les idées fixistes restent alors très installées, à l'instar par exemple de Linné qui considère que sa classification binomiale des espèces a avant tout pour but d'expliquer cet ordre divin.

1.2 Les grandes idées modernes

L'avènement de la paléontologie ainsi que la découverte et l'étude des fossiles marquent un tournant majeur dans l'histoire de la pensée évolutionniste. La publication, notamment, en 1796 des résultats de Cuvier sur la comparaison entre des ossements d'éléphants et de fossiles de mammouths prouve la distinction entre ces deux espèces ainsi que l'extinction de la seconde. C'est le premier pas vers un ensemble d'idées et de théories modernes, pour certaines encore d'actualité aujourd'hui. Dans cette section, nous en listerons certaines parmi les plus importantes.

1.2.1 Transformisme

Cette théorie développée par Jean-Baptiste de Lamarck (1744-1829) s'oppose au fixisme en ce sens qu'elle s'inscrit dans la continuité du mécanisme de Descartes. Il décrit dans son ouvrage *Philosophie Zoologique* (Lamarck, 1809) son idée selon laquelle la matière a tendance à se complexifier naturellement. Ainsi les êtres vivants suivent une évolution progressive du simple vers le complexe, à partir de première formes de vies simples apparues par génération

spontanée. Selon Lamarck, l'évolution des êtres vivants peut être décrite par les deux lois suivantes :

— L'usage fréquent d'un organe conduit au développement de celui-ci, et *a contrario*, un organe peu utilisé aura tendance à s'atrophier, puis à disparaître.
— Dans la mesure où ces changements sont acquis par les deux parents, ils sont transmis à la génération suivante.

Ainsi, "la fonction créée l'organe" et ces transformations sont héréditaires. Il étoffe son argumentaire de différents exemples, dont notamment celui de la taupe, qui *"par ses habitudes fait peu usage de la vue"* et est donc devenue par la force des choses quasiment aveugle, mais ne donne pas de mécanisme expliquant l'hérédité qu'il mentionne.

1.2.2 Sélection naturelle

Charles Darwin publie en 1859 l'*Origine des Espèces* après plus de 20 ans de recherches et d'accumulation de preuves, à commencer par ses fameuses observations dans les Îles Galapagos en 1937. Il y propose un mécanisme pour expliquer ces changements et transformations des espèces dans leurs milieux, appelé *sélection naturelle*, et qui repose sur trois principes :

— Principe de **variation** : au sein d'une même espèce, les individus d'une même population possèdent des caractères variables. La taille, la couleur des yeux ou des plumes, la forme du bec, sont autant d'exemples de caractères qui peuvent présenter des variations plus ou moins importantes selon les espèces.
— Principe d'**adaptation** : parmi ces caractères, certaines variations présentent un avantage dans un milieu donné. Elles peuvent apporter un avantage en terme de survie (par exemple en échappant aux prédateurs ou en exploitant plus efficacement les ressources), ou en terme de reproduction (plus grand nombre de descendants). Cet *avantage sélectif* dans un milieu donné implique donc un meilleur *taux de reproduction* par rapport aux congénères ne possédant pas cette variation.
— Principe d'**hérédité** : Ces caractères avantageux doivent pouvoir être transmis à la descendance. Darwin expose un mécanisme d'hérédité, aujourd'hui obsolète (appelé *pangenèse*) que nous n'exposerons pas ici.

Ainsi, selon cette théorie, certains caractères pouvant présenter des variations, peuvent offrir dans un milieu naturel donné, un avantage sélectif. Les individus possédant cette variation ont alors un meilleur taux de reproduction, donc une descendance plus nombreuse, qui possèdera aussi cet avantage, qui aura alors tendance à être de plus en plus présent dans la population,

jusqu'à éventuellement être partagé par tous les individus la composant.

1.2.3 Génétique des populations et Théorie Synthétique de l'Évolution

Dans les années 1900, la conjugaison de la redécouverte des lois de Mendel (1822-1884) et la découverte des chromosomes entraîne la fondation de la génétique formelle au début du XX^{e} siècle. L'information génétique est ainsi le "facteur" (selon l'expression de Mendel) transmis, expliquant l'hérédité.

C'est ainsi entre les années 1920 et 1940 que se développe la génétique des populations, qui a pour but d'étudier les différences de distribution et de fréquences des allèles d'un même gène au sein de populations d'êtres vivants. Ronald Fisher (1890 - 1962) a proposé différents modèles pour formaliser mathématiquement la sélection naturelle, montrant que l'action conjuguée de plusieurs gènes pouvait expliquer les variations continues de certains caractères et que les fréquences alléliques au sein d'une même population pouvaient dépendre de la sélection naturelle (Fisher, 1930). En parallèle, John Haldane (1892 - 1964) étudia différents exemples à l'aide de modèles statistiques, notamment celui très connu de la phalène du bouleau, papillon dont les individus vivant à proximité de zones industrielles (donc polluées) sont noirs, tandis que les autres sont majoritairement blancs. Il démontre ainsi que la vitesse de l'Évolution pouvait être plus importante que ce qui était alors supposé (Haldane, 1949). Enfin, Sewall Wright (1889 - 1988) décrit la théorie de la dérive génétique, qui formalise l'aspect aléatoire des modifications des fréquences alléliques au sein d'une population. Il définit aussi la notion de paysage adaptatif (Wright, 1931) dont nous reparlerons plus loin.

Les travaux de ces trois scientifiques, Fisher, Wright et Haldane permettent ainsi d'unifier dans une même théorie la génétique mendelienne et la sélection naturelle, tout en définissant des formalismes et outils probabilistes pour les étudier. Plus que la génétique des populations qu'ils créèrent, ils ont ouvert la voie à la mise en place d'une unique théorie de l'Évolution. Cette théorie unifiée prend son nom de *Théorie Synthétique de l'Évolution* dans un article de Huxley et al. (1942). Le trait d'union entre la microévolution de la génétique des populations et la macroévolution observée par les paléontologues peut être symbolisé par le livre de Dobzhansky et al. (1941). Avec d'autres ouvrages parus dans la décennie qui suit, cette théorie résume les différents mécanismes de l'Évolution :

— Apparition de nouveaux caractères par mutations, réarrangements chromosomiques, ou recombinaisons. Les migrations constituent aussi une source de nouveaux caractères,

en ce sens qu'elles apportent de la diversité génétique issue d'autres populations de la même espèce.

— Diffusion ou disparition de ces nouveaux caractères au sein des populations par dérive génétique et sélection naturelle. L'hérédité y joue un rôle prédominant comme dans la théorie initiale de Darwin.

1.2.4 Biologie moléculaire et Théorie Neutraliste de l'Évolution

L'avènement de la biologie moléculaire au milieu du XX^{e} siècle a apporté une meilleure compréhension des phénomènes microscopiques qui régissent l'évolution. De la molécule d'ADN à la constitution des protéines, la connaissance du code génétique (notamment par le biais du séquençage) permet en particulier de dater les mutations, notamment par le biais de l'horloge moléculaire (Zuckerkandl et Pauling, 1962, 1965), qui donne un *timing* de l'évolution, et permet d'apporter la notion de temporalité dans l'accumulation des mutations, en supposant que celle-ci se déroule à un rythme constant.

La théorie neutraliste de l'Évolution (Kimura, 1968; King et Jukes, 1969) stipule que la plupart des mutations sont *neutres* en regard de la valeur sélective (mutations synonymes, mutations au niveau des introns), et plus généralement que les mutations ont une influence négligeable sur celle-ci. Elle met ainsi la sélection naturelle en retrait par rapport à la dérive génétique, qui devient le mécanisme prépondérant expliquant la diversité génétique observée à un instant t.

Cette théorie repose sur des modèles mathématiques impliquant plusieurs hypothèses fortes :

— Taux de mutations constants : cette hypothèse est moins forte si l'on restreint les séquences étudiées plutôt que de travailler sur l'ensemble de l'ADN.

— Taille des populations constantes : selon l'échelle de temps étudiée, cette hypothèse n'est pas toujours restrictive, on peut généralement considérer la taille des populations comme étant constante à de petites échelles de temps.

— L'équilibre mutation/dérive doit être respecté : la perte d'allèles due à la dérive génétique doit être égale à la production d'allèles par mutation.

La sélection darwinienne n'est pas remise en question par cette théorie, même si elle est considérée comme plus rare et moins importante que le laissait supposer les théories antérieures.

Chapitre 2

La Coévolution

Nous nous attacherons à conserver dans cette thèse une définition très large de la notion de coévolution. Nous détaillerons ici ce que nous entendons par ce terme, à commencer par la définition rencontrée le plus couramment dans la littérature, c'est à dire la coévolution entre espèces. De manière plus générale, nous appellerons coévolution ***l'évolution corrélée de deux entités biologiques****, quelle que soit l'échelle étudiée. En d'autres termes, cette notion peut aussi concerner la corrélation entre les mutations sur deux sites distincts d'une même séquence, par exemple. Dans un premier temps, nous définirons brièvement la notion de coévolution au niveau de l'individu et de l'espèce, avant de décrire plus en détail la coévolution à d'autres échelles.*

2.1 Interactions à l'échelle de l'individu et de l'espèce

Commençons par évoquer un des aspects de la coévolution que l'on peut le plus facilement observer : celui concernant l'échelle de l'espèce ou de l'individu.

Le concept de coévolution apparaît dans un article de Ehrlich et Raven (1964), décrivant les interactions entre deux grands groupes d'organismes : les plantes et les herbivores, à travers l'exemple des papillons et des plantes dont ils se nourrissent. Il est clair que ces deux groupes partagent d'importantes relations écologiques, aussi les auteurs ont montré que l'on pouvait observer que, les papillons exerçant une évidente pression de sélection sur ces plantes, celles-ci ont développé des mécanismes de défense. Les papillons en question ont alors acquis des caractères permettant de contourner ces mécanismes de défense. Le papillon comme la plante exercent donc l'un sur l'autre une pression de sélection, que l'on

peut qualifier de réciproque, influençant donc l'évolution de chacune de ces espèces. Cet exemple est typiquement représentatif de ce qu'on appelle la "course aux armements" entre les espèces, ou *Théorie de la Reine Rouge*, par allusion à course sans fin entre Alice et la reine du roman *Alice au Pays des Merveilles* de Lewis Carroll. De nouveaux mécanismes auront donc tendance à être sélectionnés chez la proie comme chez le prédateur, l'un pour mieux consommer, l'autre pour mieux échapper à la consommation.

Il en va de même pour les interactions entre les parasites et leurs hôtes, pour lesquels les mécanismes sont très semblables. L'hôte est l'habitat du parasite, qui consomme ses ressources énergétiques, ce qui réduit directement sa valeur reproductive. Ainsi, la présence du parasite constitue une pression de sélection supplémentaire dans une population, où, pour deux individus parasités, celui possédant les meilleurs mécanismes de défense aura une valeur de *fitness* supérieure. Ces mécanismes auront ainsi tendance à envahir la population, favorisés par la sélection naturelle. Cela constitue à son tour une pression pour le parasite, pour lequel des caractères contournant ces défenses seront sélectionnés, et ainsi de suite. Encore une fois, nous observons une "course aux armements" entre les espèces respectivement hôte et parasite (voir Brooks (1979) pour une étude plus détaillée de la coévolution et co-spéciation entre parasites et hôtes).

Nous avons cité ici deux exemples de coévolution *antagoniste*, où les espèces en questions maintiennent une forte pression de sélection les unes sur les autres. D'autres types d'interaction existent, reposant sur les mêmes mécanismes. Nous pouvons par exemple citer celles entre espèces mutualistes, ou même dans un cadre intraspécifique, la coévolution des organes sexuels chez le mâle ou la femelle au sein d'une même espèce.

2.2 L'épistasie

L'épistasie, pour la première fois évoquée par Bateson (1909), est l'influence du contexte génétique sur le changement de fitness dû à une mutation. Une mutation peut en effet être délétère dans certains contextes, tandis qu'elle peut être bénéfique dans d'autres (Phillips, 2008). En d'autres termes, une combinaison de mutations peut avoir un effet différent de la somme des effets individuels de chacune d'entre elles.

2.2.1 L'exemple des Déviation Pathogènes Compensatoires

Un très bon exemple illustrant l'effet du contexte génétique sur les mutations est celui des Déviation Pathogènes Compensatoires. Celles-ci consistent en certaines mutations qui peuvent être délétères pour certaines espèces, entraînant une baisse de *fitness* (maladies, baisse de fécondité, mort...), mais qui ne le sont pas pour d'autres. Ces mutations délétères peuvent donc être *compensées* par une ou plusieurs autres mutations. Ces mutations, dites compensatoires, permettent à la protéine concernée d'assurer convenablement sa fonction (Barešić et Martin, 2011). Par la suite, pour des raisons de commodité et de cohérence avec la littérature, nous garderons l'abréviation anglaise "CPD", pour *Compensated Pathogenic Deviations*.

Plusieurs études ont été menées à propos des CPDs. Kondrashov et al. (2002) ont montré que dans un échantillon de 32 protéines impliquées dans des maladies humaines, environ 10% des substitutions délétères d'acides aminés sont les allèles sauvages chez d'autres vertébrés. Ils détaillent plus avant le cas de la β-hémoglobine chez l'homme et le cheval, pour lesquels les seules différences sont en position 20 et 69 (Val20 et Gly69 chez l'homme et Glu20 et His69 chez le cheval). La substitution 20 $Val \rightarrow Glu$ est délétère chez l'homme, puisqu'elle cause des drépanocytoses, altérant l'hémoglobine et donc le transport de l'oxygène par le sang. Or, cette pathologie n'est pas présente chez le cheval. Les auteurs concluent que la substitution 69 $Gly \rightarrow His$ pourrait compenser cet effet délétère.

Kulathinal et al. (2004) ont étudié dans quelle mesure des mutations pathogènes chez *Drosophila melanogaster* peuvent être fixées (à travers des mécanismes de CPDs, donc d'épistasie) dans des génomes d'autres espèces de diptères. Ils ont montré qu'environ 10% de ces mutations délétères constituaient le génotype sauvage dans ces génomes. Il est intéressant de noter que nous retrouvons le même résultat quantitatif que dans l'échantillon de vertébrés de l'exemple précédent.

2.2.2 Différents types d'épistasie

Il existe de nombreuses définitions de l'épistasie, mais nous resterons ici dans le cadre général de celle citée plus haut. Nous pouvons tout de même caractériser ici l'épistasie d'un point de vue quantitatif, en fonction de son effet.

Caractérisation quantitative

L'effet des interactions génétiques sur la *fitness* est important à appréhender, puisque c'est entre autres lui qui permet de la mesurer et ainsi d'établir les "paysage adaptatifs" que nous définirons dans la section suivante. Nous pouvons donc classer l'effet de l'épistasie dans trois catégories complémentaires. Pour cela, considérons un modèle à deux loci et deux allèles dans $\{0, 1\}$. Le génotype sauvage est $(0, 0)$. Nous avons donc deux simples mutants possibles $(0, 1)$ et $(1, 0)$ et un double mutant $(1, 1)$. Soit $\mathcal{F}(a, b)$ la *fitness* associée au génotype (a, b).

— Absence d'épistasie (**additivité**) : deux mutations sont dites additives si l'effet de la double mutation est le même que la somme des effets des deux mutations séparément (figure 2.1.a) :

$$\mathcal{F}(1,1) - \mathcal{F}(0,0) = (\mathcal{F}(1,0) - \mathcal{F}(0,0)) + (\mathcal{F}(0,1) - \mathcal{F}(0,0))$$

— L'**épistasie de magnitude** décrit les effets en terme d'importance de l'effet d'un double mutant par rapport aux mutants simples. Dans le cas où la *fitness* du double mutant est plus – ou moins – importante qu'attendu en sommant les effets des deux mutations séparées. Si cet effet est plus important, on parle d'*épistasie positive* (figure 2.1.b)

$$\mathcal{F}(1,1) - \mathcal{F}(0,0) > (\mathcal{F}(1,0) - \mathcal{F}(0,0)) + (\mathcal{F}(0,1) - \mathcal{F}(0,0))$$

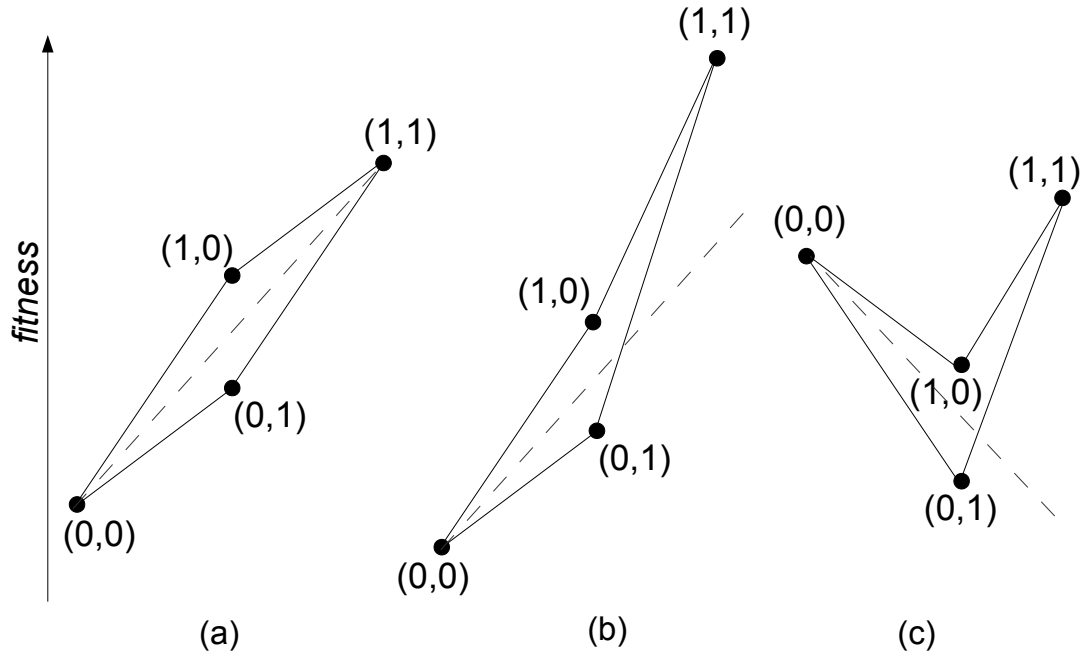
Dans le cas contraire, on parle d'*épistasie négative*. Nous observons donc ici la *fitness* des mutants, dans l'absolu.

— L'**épistasie de signe** est relative aux mutations qui ont un effet opposé quand elles sont en présence d'une autre mutation (Weinreich et al., 2005). Ce type d'épistasie est illustré par le cas des CPDs : une mutation peut être délétère pour certaines espèces, mais pour d'autres, pour lesquelles une mutation compensatrice s'est fixée dans les populations, cette mutation devient au minimum neutre, voire positive, puisqu'elle même peut se fixer à son tour.

$$\mathcal{F}(1,1) - \mathcal{F}(1,0) > 0 > \mathcal{F}(0,1) - \mathcal{F}(0,0))$$

Dans le cas particulier de l'*épistasie de signe réciproque* (figure 2.1.c), ce sont cette fois deux mutations délétères qui ont un effet positif quand elles sont en présence l'une de l'autre.

FIGURE 2.1 – **Différents paysages adaptatifs représentant différents types d'épistasie, pour un modèle à deux loci et deux allèles.** (a) absence d'épistasie, (b) épistasie positive, (c) épistasie de signe réciproque. Les lignes en pointillés représentent la fitness attendue en l'absence d'épistasie.



2.2.3 Les paysages adaptatifs

À partir de ces définitions, nous pouvons dès à présent déduire que dans certains cas, toutes les mutations ne sont pas équivalentes (en termes d'avantage évolutif), puisque leur effet peut dépendre du contexte génétique. Nous pouvons même aller plus loin et en déduire que l'*ordre* des mutations est contraint par de telles interactions épistatiques. En effet, en fonction du contexte, certaines mutations pourront engendrer un gain de *fitness* tandis que d'autres seront délétères, et dans un autre contexte, les rôles pourraient être inversés. Or, le contexte lui-même change évidemment à chaque mutation. Ainsi, les *chemins évolutifs* ne sont pas tous équivalents. Ils sont au contraire contraints et dans certains cas, des chemins sont favorisés par rapport à d'autres.

Une façon d'illustrer ce phénomène est de placer sur un plan les divers contextes génétiques, organisés par exemple selon le nombre de mutations qui les caractérise, et d'ajouter dans une troisième dimension leur valeur sélective. C'est une façon intuitive d'envisager ces pay-

sages, puisque le nombre de dimensions étant égal au nombre de sites considérés, dès que les séquences ont une longueur supérieure à 2, on ne peut plus vraiment dessiner ce plan. Ce n'est pas toujours possible, tant la notion de *fitness* peut être vague pour des êtres vivants ou des environnements complexes, mais dans de nombreux cas, il est tout à fait possible de réaliser une approximation et d'assimiler la fitness à une ou plusieurs grandeurs mesurables, comme la résistance à un antibiotique dans le cas d'une bactérie, ou le nombre de descendants moyen. Nous obtenons alors une surface en trois dimensions, que l'on appelle un *paysage adaptatif* (Wright, 1931), et dont la pente la plus élevée en chaque point caractérise les chemins évolutifs les plus bénéfiques, en terme d'adaptation. C'est une façon de conceptualiser et de visualiser l'évolution d'une entité biologique, qui peut aussi permettre de mieux comprendre certains processus sous-jacents.

En particulier, la forme de ce paysage est très importante puisqu'elle peut être un précieux indice quant à l'appréhension de certaines trajectoires évolutives, ce qui nous ramène à étudier les variations de la pente du paysage (Weinreich et Chao, 2005). Prenons le cas simple de deux loci et deux allèles par locus. Ce cas nous permettra de construire des paysages adaptatifs simples et exhaustifs (figure 2.1). *Les effets des mutations sur la fitness sont multiplicatifs, nous raisonnerons donc ici en terme de log-fitness pour illustrer ces effets de façon additive.*

En l'absence d'épistasie, les mutations ont toutes des effets additifs, et ainsi le paysage adaptatif se trouve être un plan (*i.e.* de pente constante) : chaque nouvelle mutation n'ajoutera que son effet absolu. En termes plus mathématiques, nous pouvons établir que dans ce cas :

$$\log(fitness_{genome}) = c + \sum_{loci} \log(fitness_{locus})$$

Où c est une constante correspondant au logarithme de la *fitness* du génome ne portant aucune mutation.

Les choses se compliquent considérablement lorsque l'on introduit de l'épistasie entre les loci. En effet, dans un premier temps, l'épistasie par paire de sites impose d'ajouter à cette formule les termes croisés pour chaque paire de sites, mais on a vu que l'on pouvait aller encore plus loin, puisque c'est finalement tout le contexte génétique qui peut avoir un effet sur la *fitness* de tel ou tel génome. Ainsi, les termes correspondant aux triplets, aux quadruplets, etc, jusqu'à la longueur totale de la séquence étudiée peuvent intervenir dans cette formule, rendant à la fois cette formule bien plus complexe, mais aussi la représentation du comportement de la *fitness* plus difficile, pour ne pas dire impossible (Weinreich et al., 2013).

La question de la dimension des paysage adaptatif est considérée par le modèle NK de Kauffman et Levin (1987), qui permet de construire des paysages adaptatifs pour lesquels

on peut contrôler la dimensionnalité à l'aide des deux paramètres N et K. N représente la longueur des séquences, et K le nombre de sites avec lesquels chaque site interagit. Pour $K = 0$, le paysage est parfaitement plat, puisqu'il n'y a pas d'épistasie, et le cas le plus complexe est représenté par $K = N - 1$, où chaque site interagit avec chaque autre site. Ce modèle permet ainsi l'étude des ces effets dus à la dimension sur des paysages adaptatifs synthétiques complets, puisque générés par celui-ci.

Un paysage en particulier a été étudié par Weinreich et al. (2006) sur des données biologiques. Cinq mutations ponctuelles sur le gène codant pour la β-lactamase déterminent la résistance à la cefotaxime (antibiotique) chez certaines bactéries, le *wild-type* correspondant à l'absence de mutation (et est le moins résistant), et la résistance maximum étant atteinte quand les 5 mutations sont présentes. Ainsi, $5! = 120$ chemins mutationnels sont possibles, à partir de la résistance moindre jusqu'à la plus importante. Les auteurs ont démontré que parmi ces chemins, 102 sont inaccessibles, en raison notamment de l'épistasie de signe qui rend certaines de ces mutations délétères en fonction du contexte génétique (*i.e.* des autres mutations déjà fixées parmi les 4 autres). Par ailleurs, ils montrent que seuls deux chemins – presque identiques, ils n'ont qu'une étape différente – sont très avantageux, cette fois à cause de l'épistasie de signe positive, qui rend certaines mutations plus bénéfiques que d'autres selon le chemin emprunté. Ainsi, l'exploration du paysage adaptatif (relativement simple) dans ce cas permet de mieux comprendre les contraintes importantes qui peuvent peser sur les chemins évolutifs, qui ne sont pas tous équivalents. Une revue sur différents paysages adaptatifs calculés empiriquement peut être trouvée dans l'article de Szendro et al. (2013).

2.3 À l'échelle de l'ARN

L'ARN (Acide RiboNucléique) est une molécule biologique présente chez tous les êtres vivants. Cette molécule, polymère constitué d'une séquence de nucléotides, a de nombreuses formes et fonctions, parmi lesquelles nous pouvons citer les ARN messagers, supports temporaires de l'information génétique, les ARN ribosomiques, "têtes de lecture" de l'information génétique transcrite par l'ARN messager, ou les ARN de transferts, traduisant l'information des codons de l'ARN messager en acides aminés.

En particulier, certains ARN peuvent adopter une structure secondaire et tertiaire stable, et accomplir des fonctions catalytiques, comme c'est le cas pour les ARN ribosomiques (ARNr) ou les ARN de transfert (ARNt). Cette structure a été déterminée expérimentalement (Doty et al., 1959; Leontis et Westhof, 2001), et nous donne de précieuses informations sur les replie-

ments et les interactions entre nucléotides au sein de cette structure. En effet, les positions relatives des différents nucléotides composant l'ARN et leurs propriétés physico-chimiques impliquent différentes contraintes sur ces structures tridimensionnelles.

Nous pouvons par exemple citer le cas de l'ARNr, pour lequel la structure secondaire et tertiaire est largement déterminée par de fortes liaisons hydrogène entre l'adénine et l'uracile, et entre la guanine et la cytosine (Watson et al., 1953). Ces appariements, dits Watson-Crick (WC), ont une telle influence sur la structure des ARNr, qu'une mutation observée au niveau d'un nucléotide d'une paire WC est souvent délétère, et sera généralement observée accompagnée d'une mutation compensatoire sur le deuxième nucléotide de la paire. Les appariements WC ont été très étudiés en tant que cas classiques de coévolution moléculaire (Watson et al., 1953; Leontis et Westhof, 1998; Moore, 1999). Notons que des mécanismes très semblables sont à l'œuvre dans le cas d'ARNt, et ont par exemple été étudiés par Kern et Kondrashov (2004), dans le contexte explicite des CPDs.

2.4 À l'échelle de la protéine

Au même titre que pour l'ARN, les outils de détection de la coévolution sont particulièrement bien adaptés à l'étude des séquences d'acides aminés qui constituent la structure primaire des protéines. Ils permettent ainsi de comparer les paires de sites et de déterminer celles qui partagent significativement une histoire évolutive. Par ailleurs, discriminer ces paires qui coévoluent permet aussi d'obtenir des informations, parfois très importantes, sur la structure tridimensionnelle des protéines concernées, puisque la position relative des acides aminés, et donc leur éventuelle proximité spatiale ou leur localisation sur des résidus en interaction sont autant de critères impliquant de la coévolution (Shindyalov et al., 1994).

L'encombrement stérique (au niveau de la structure tertiaire de la protéine) ou les propriétés physico-chimiques des acides aminés permettent d'identifier de nombreuses contraintes au niveau de la structure en trois dimension de la protéine. Les contacts entre acides aminés sont donc autant d'informations précieuses quant à la connaissance de cette structure. Mais, comme pour l'ARN, de telles contraintes peuvent aussi impliquer l'existence de coévolution au niveau moléculaire entre ces acides aminés. Ainsi, certains auteurs s'attachent à analyser à travers différentes méthodes les mutations corrélées au niveau de la protéine pour prédire les contacts entre les acides aminés concernés dans la structure tridimensionnelle de celle-ci (Pollock et al., 1999; Engelen et al., 2009; Dib et Carbone, 2012).

2.5 Régulation génétique et facteurs de transcriptions

Un autre mécanisme que nous nous devons de mentionner ici est celui de la régulation génétique. Celui-ci fait en effet intervenir tout un ensemble d'entités différentes et pourtant liées par leurs fonctions pour modifier l'expression des gènes. Chez les eucaryotes, typiquement, cela se fait par l'intermédiaire de séquences ADN, dites *séquences cis* ou par des protéines (facteurs de transcription), initiant ou régulant la transcription de ces séquences cis, appelées *facteurs trans*. Plus généralement, de très nombreuses protéines appelées facteurs de transcription sont connues (Ravasi et al., 2010). Elles initient et régulent l'expression des gènes, et ont ainsi un rôle clef à tous les niveaux de la biologie de la cellule.

Il est donc légitime de s'intéresser aux phénomènes de coévolution qui peuvent exister entre ces différents éléments régulant l'expression des gènes. Fraser et al. (2004) ont par exemple montré que la coévolution pouvait être observée non seulement au niveau des protéines, comme nous l'avons évoqué précédemment, mais aussi au niveau de leur degré d'expression. En sus de l'étude les corrélations entre mutations le long d'une séquence d'acides aminés, les auteurs ont analysé les changements coordonnés d'expression de gènes et ont démontré qu'ils portent un signal de coévolution.

Chapitre 3

Comment repérer et étudier la coévolution ?

Nombreuses sont les méthodes qui s'attachent à détecter et expliquer la coévolution. Chaque méthode a ses avantages et ses inconvénients, mais une tendance générale se dégage : il existe un important compromis entre efficacité des méthodes et précision des résultats. Certaines méthodes détectent très efficacement la coévolution entre les sites d'une séquence de nucléotides ou d'acides aminés, par exemple. Elles permettent ainsi l'étude de grands jeux de données, tandis que d'autres sont capables de décrire des scénarios de coévolution probables, mais au prix de temps de calcul parfois très élevés.

Nous distinguons ici deux grandes classes de méthodes : celles indépendantes des phylogénies, et celles qui nécessitent de prendre en compte cette information.

3.1 Méthodes indépendantes de la reconstruction phylogénétique

Les méthodes qui ne font pas usage de la phylogénie se basent uniquement sur l'étude d'alignements multiples (MSA, pour *Multiple Sequence Alignment*), dans le but de déterminer quelles colonnes de l'alignement concerné sont associées statistiquement. Un alignement multiple correspond aux états actuels observés, organisés sous forme de matrice : les lignes correspondent aux séquences utilisées, les colonnes aux positions. On notera donc ici un des principaux avantages de ce type de méthodes : elles ne dépendent nullement d'éventuelles reconstructions (phylogénie, états ancestraux, historiques mutationnels) que nous verrons plus

loin.

3.1.1 Tests statistiques

Une première méthode consiste à appliquer de simples tests statistiques sur les alignements, dans le but de tester s'il existe une corrélation entre les composantes de deux colonnes de l'alignement (acides nucléiques ou acides aminés, la principale différence - en termes d'implémentation - étant la taille de l'alphabet utilisé). Typiquement, on considère une colonne d'un alignement comme un ensemble de réalisations d'une variable aléatoire, ce qui ouvre la porte à différents tests applicables, l'hypothèse nulle étant l'indépendance de ces colonnes. Par exemple, il est possible d'appliquer un test du χ^2 de Pearson, qui permet de tester si un ensemble d'observations sur deux colonnes d'un alignement (*i.e.* deux variables aléatoires) sont indépendants l'un de l'autre, par l'intermédiaire d'une table de contingence (voir Brouillet et al. (2010) pour un exemple d'utilisation d'un tel test pour repérer les colonnes d'un alignement ayant tendance à *ne pas muter* ensemble).

Ce test très simple permet donc de vérifier très rapidement la corrélation entre deux colonnes du MSA, mais implique un effectif minimum (ici, un nombre minimum de lignes dans l'alignement), et une représentation minimum de chaque classe (ici, les nucléotides ou acides aminés). En l'occurrence, la règle la plus utilisée, établie par Cochran (1954), stipule dans notre cas que chaque classe doit être représentée au moins 5 fois, ce qui peut constituer une contrainte importante lors de l'établissement du jeu de données étudié.

3.1.2 Théorie de l'information et entropie

Une autre manière d'envisager et étudier les MSA est de se placer dans le cadre de la théorie de l'information mutuelle. D'un point de vue conceptuel, cela revient à considérer chaque colonne de l'alignement comme une source d'information, et d'estimer la quantité d'information qu'elle émet. Par ailleurs, il est aussi possible de calculer l'information partagée entre deux colonnes. À partir de ces grandeurs, il est possible d'établir la valeur de ce que l'on appelle *information mutuelle*, qui est une mesure de la dépendance statistique entre ces deux colonnes.

La plupart des techniques faisant appel à l'information mutuelle sont basées sur la notion d'*entropie*. La plus couramment utilisée est l'entropie de Shannon, qui est une mesure de l'incertitude pour une variable aléatoire. Considérons une variable aléatoire discrète X tirée dans un alphabet $\{x_1, x_2, ..., x_N\}$, où $N \in \mathbb{N}$, avec la distribution de probabilités associée

$\{p(x_1), p(x_2), ..., p(x_n)\}$. On définit l'entropie $H(X)$ pour cette variable aléatoire par la formule suivante :

$$H(X) = -\sum_{i=1}^{N} p(x_i) \log_b p(x_i)$$

À noter que b représente ici la base du logarithme choisi, et sert à normaliser l'entropie. En particulier, on choisira généralement $b = N$ pour garantir que $0 \leq H(X) \leq 1$ et par convention, on admettra que $0 \log(0) = 0$ (en effet, $\lim_{x \to 0} x \log(x) = 0$).

Prenons une deuxième variable aléatoire Y, tirée dans un alphabet $\{y_1, y_2, ..., y_M\}$, avec $M \in \mathbb{N}$. Nous avons alors une distribution de probabilité jointe pour les variables X et Y correspondant à chaque paire possible. Nous pouvons ainsi définir de manière analogue *l'entropie conditionnelle* $H(X|Y)$, qui représente la quantité d'information nécessaire pour définir X, connaissant Y. Nous pouvons aussi définir *l'entropie jointe* $H(X, Y)$, qui elle représente l'incertitude pour notre couple de variables (X, Y). Elles sont données respectivement, pour ces deux variables aléatoires, par :

$$H(X|Y) = -\sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \log_b \frac{p(x_i, y_j)}{p(x_i)}$$

$$H(X, Y) = -\sum_{i=1}^{N} \sum_{j=1}^{M} p(x_i, y_j) \log_b p(x_i, y_j)$$

On définira alors l'*information mutuelle*, qui est une mesure de la réduction de l'incertitude (elle même mesurée par l'entropie) d'une variable aléatoire, étant donnée l'information concernant une autre variable aléatoire. En d'autres termes, elle estime l'incertitude *partagée* par les variables X et Y et s'exprime de la façon suivante :

$$MI(X, Y) = H(X) - H(X|Y) = H(Y) - H(Y|X) = H(X) + H(Y) - H(X, Y)$$

L'information mutuelle est ainsi comprise entre 0 et 1, et est d'autant plus grande que les variables aléatoires sont corrélées (Cover et Thomas, 2012).

En pratique, concernant les MSAs, nos variables aléatoires sont représentées par des colonnes de l'alignement, l'alphabet étant l'ensemble des acides aminés ou des acides nucléiques, et la distribution de probabilité associée étant calculée à partir des fréquences observées dans la colonne. Ainsi, une entropie nulle correspondra à une colonne complètement conservée, et une entropie de 1, à une colonne où tous les éléments de l'alphabet sont également représentés dans la colonne. De même l'entropie jointe est calculée à partir de paires de colonnes et des fréquences observées pour toutes les paires possibles.

Ainsi, l'information mutuelle entre deux colonnes de l'alignement estime dans quelle mesure les motifs observés dans chaque colonne sont corrélés. Elle est comprise entre 0 et 1, les plus grandes valeurs correspondant à des degrés d'interdépendance plus élevés entre les positions

concernées.

Il est important de noter que plusieurs auteurs soulignent que l'information mutuelle, en tant que mesure directe de la dépendance d'une position à une autre, est malgré une mesure de la coévolution limitée, pour différentes raisons (Carbone et Dib, 2009) :

— L'entropie est d'autant plus importante pour des colonnes très variables, même si la dépendance ne l'est pas nécessairement. Ainsi, pour de telles colonnes, l'information mutuelle peut surestimer la corrélation. On peut corriger, dans une certaine mesure, cet effet en normalisant l'information mutuelle par l'entropie jointe $H(X, Y)$ (Martin et al., 2005).
— La taille de l'échantillon doit être assez importante pour assurer la fiabilité des fréquences observées, ainsi que pour limiter le bruit statistique.
— L'information partagée provient aussi de l'histoire phylogénétique, qui n'est pas prise en compte ici. Il est possible de retirer certaines séquences trop peu divergentes de l'alignement pour limiter cet effet, qu'il est néanmoins impossible de supprimer complètement.

De manière générale, une des approches qui semble parmi les plus pertinentes serait de mesurer les effets d'échantillonnage et ceux correspondant à l'histoire partagée sur des données obtenues par simulation, ce qui permet dans une certaine mesure d'évaluer le bruit statistique dû à ces deux phénomènes (Martin et al., 2005).

3.2 Vraisemblance [1]

3.2.1 Éléments de définitions

Nous appelons "*vraisemblance*" d'un paramètre θ (qui peut être multidimensionnel), sachant une réalisation x d'une loi statistique, la probabilité d'observer cette réalisation sachant la valeur θ de ce paramètre. Elle est notée :

$$\mathcal{L}(\theta|x) = P(x|\theta).$$

On appellera donc "fonction de vraisemblance" la fonction de θ $f_\theta(x) = \mathcal{L}(\theta|x)$, à x fixé. En d'autres termes, cette fonction détermine, sous un modèle, dans quelle mesure le paramètre

1. Les définitions données dans cette section sont basées sur le livre *Probabilités, analyse des données et statistique* de Gilbert Saporta (2006).

θ fixé pour ce modèle explique les données x.

Ainsi, à x fixé, la fonction $\theta \mapsto f_\theta(x)$ décrit une surface, appelée "surface de vraisemblance". Bien entendu, il n'est pas toujours possible de représenter cette surface : tout dépend de la dimension de l'espace dans lequel θ est défini, *i.e.* le nombre de paramètres du modèle. Nous pouvons faire l'analogie avec les paysages adaptatifs vus plus haut : nous appellerons par la suite une telle surface un "paysage de vraisemblance".

Le problème du maximum de vraisemblance, ou ML (*Maximum Likelihood*) est un problème d'optimisation : il consiste à rechercher le maximum de la fonction de vraisemblance, *i.e.* la probabilité maximum de la réalisation observée x, ce qui revient à rechercher le paramètre θ optimal, expliquant le mieux les données. Dans certains cas, il est possible de résoudre analytiquement ce problème. Notons que puisque la vraisemblance est une probabilité conditionnelle, elle est supérieure à 0. Il est donc équivalent de maximiser la vraisemblance et la log-vraisemblance. Si ce n'est pas possible, il faut estimer ce maximum de vraisemblance, en explorant le paysage de vraisemblance défini plus haut et en recherchant ses maxima.

Jusqu'ici, nous considérions que tout l'information sur θ provient des données x. Or, si nous avons une des connaissances *a priori* sur θ sous la forme d'une densité *a priori* $\pi(\theta)$ appelée *prior*, il est possible de "mettre à jour" ce *prior* grâce à notre connaissance des données x et obtenir ce qu'on appelle la densité *a posteriori* $\pi(\theta|x)$. Le théorème de Bayes, appliqué à notre cas précis de vraisemblance, donne en effet :

$$P(\theta|x) = \frac{P(x|\theta)P(\theta)}{P(x)}$$

où $P(x) = \int_\theta P(x,\theta)\,\mathrm{d}\theta = \int_\theta P(x|\theta)P(\theta)\,\mathrm{d}\theta$ est une constante de normalisation.

Nous reconnaissons dans cette formule la vraisemblance $P(x|\theta)$ et pouvons ainsi, à partir d'une connaissance *a priori* sur les paramètres, obtenir une connaissance *a posteriori* sur ces paramètres, en regard de l'information que nous avons par les données. En pratique, pour un jeu de données x fixé, la connaissance de la vraisemblance du paramètre θ et la celle *a priori* sur θ nous permettent de mettre à jour notre connaissance sur ces paramètres, à la lumière des données observées. Par exemple, si nous n'avons aucune connaissance sur θ, un prior uniforme est approprié, puisqu'il n'apporte aucune information sur le paramètre : celui-ci peut prendre n'importe quelle valeur, avec une probabilité égale. Ainsi, à l'aide d'une réalisation donnée x, la vraisemblance de θ sachant x permet à l'aide du théorème de Bayes d'apporter de l'information supplémentaire sur θ et donc de mettre à jour sa distribution sous la forme d'une loi *a posteriori*. Dans le cas où nous avons une meilleure appréhension

de θ, ce *prior* peut par exemple prendre la forme d'une loi normale, dont la variance sera d'autant plus faible que notre connaissance est précise.

Enfin, si θ est de dimension n, en intégrant sur $n-1$ de ses paramètres, nous pouvons obtenir ce que nous appelons *distribution marginale* pour le dernier paramètre libre. Par exemple, pour une fonction à trois paramètres $f(x, y, z)$, la distribution marginale correspondant à x est :

$$Marg(x) = \int_y \int_z f(x, y, z) \mathrm{d}y \mathrm{d}z$$

Dans le cas où cette loi marginale n'est pas calculable analytiquement, il est possible de l'estimer numériquement en projetant ces $n - 1$ paramètres sur l'axe correspondant à la variable concernée.

3.2.2 Estimer numériquement le maximum

Il existe plusieurs méthodes permettant d'explorer le paysage de vraisemblance dans le but de repérer ce maximum. Une première façon "naïve" consiste à choisir un point de départ, puis d'explorer le paysage de proche en proche, en étudiant tous les voisins de ce point et en choisissant celui qui renvoie la meilleure vraisemblance. À partir de ce point, on réitère le procédé, et ainsi de suite, jusqu'à atteindre un point meilleur que tous ses voisins, qui sera considéré comme le maximum (mais qui peut être un maximum local). Encore une fois, cette méthode est très coûteuse en temps de calcul, puisque le nombre de voisins à explorer explose avec l'augmentation du nombre de paramètres.

Il existe d'autres méthodes plus efficaces que cette recherche par voisin, qui permet de limiter l'effet de l'augmentation du nombre de paramètres. L'une d'entre elles consiste, à partir d'un point de départ, à calculer le gradient en ce point, c'est à dire la direction de la pente la plus élevée autour de celui-ci, de la façon suivante, pour une fonction $f(x_1, x_2, x_3, ..., x_n)$, où $n \in \mathbb{N}$ est le nombre de paramètres :

$$\nabla f(x_1, x_2, x_3, ..., x_n) =$$
$$\left(\frac{f(x_1+dx_1,x_2,x_3,...,x_n)-f(x_1-dx_1,x_2,x_3,...,x_n)}{2dx_1}, ... \frac{f(x_1,x_2,x_3,...,x_n+dx_n)-f(x_1,x_2,x_3,...,x_n-dx_n)}{2dx_n}\right)$$

Ce gradient nous donne la direction optimale à suivre dans notre recherche du maximum de la surface. Après avoir trouvé le pic de vraisemblance suivant cette direction, nous réitérons alors le procédé en calculant un nouveau gradient à partir de ce point, et ainsi de suite, jusqu'à obtenir un maximum de vraisemblance, *i.e.* un point autour duquel la pente est toujours négative.

Notons que dans le cas où la surface n'est pas uni-modale, c'est à dire qu'elle possède plusieurs pics, il faut prendre plus de précautions dans le but d'éviter d'atteindre un maximum local, et ainsi manquer le véritable pic de vraisemblance. Il existe d'autres méthodes ou heuristiques permettant d'éviter de genre de problèmes, que nous n'avons pas étudiés au cours de cette thèse.

3.2.3 Échantillonner numériquement le paysage

La recherche du maximum de vraisemblance n'est pas toujours la meilleure solution. En effet, elle dépend très largement de la forme du paysage de vraisemblance considéré. Si ce paysage ne possède qu'un seul pic très marqué, alors, en effet, son maximum est très informatif, puisqu'il correspond à un point "meilleur" que tous les autres, sans ambiguïté. Imaginons un paysage pour lequel plusieurs points sont à une altitude proche, formant un plateau. Son maximum sera alors noyé parmi ses voisins, n'étant meilleur que de façon négligeable. Il n'est donc que très peu informatif, et les paramètres correspondant à deux extrémités opposées de ce plateau sont quasiment équivalents, même s'ils sont très différents. Évidemment, plus ce plateau est vaste, moins ce maximum est informatif. Il en va de même dans le cas où la surface n'admet non pas un unique pic, mais plusieurs. Dans cette situation, comme nous l'avons mentionné plus haut, nous courons le risque de capturer un maximum local, sans possibilité de s'en extraire pour atteindre le maximum global.

Dans ces cas, l'échantillonnage du paysage est une méthode qui nous permet d'obtenir bien plus d'informations pertinentes quant au comportement de la fonction de vraisemblance pour l'ensemble des paramètres possibles. Il consiste à établir un ensemble de points recouvrant le paysage, au *pro rata* de la vraisemblance : le nombre de points échantillonné est plus important dans les régions du paysages où l'altitude (la vraisemblance) est la plus élevée. Ainsi, les régions où l'échantillonnage est le plus dense sont les zones du paysage où la vraisemblance est généralement plus élevée, ce qui permet de pallier efficacement les problèmes mentionnés plus haut. En particulier, concernant les plateaux, nous aurons une densité élevée et homogène sur ceux-ci, et les différentes zones les plus denses correspondront aux différents maxima locaux.

Nous avons utilisé, pour ce travail de thèse, une méthode de Monte-Carlo par chaînes de Markov (MCMC, pour *Markov Chain Monte-Carlo*) assez classique (Metropolis et Ulam, 1949), mettant en œuvre une chaîne de Markov par l'algorithme de Metropolis-Hastings (Metropolis et al., 1953). Encore une fois, nous partons d'un point de départ choisi soit arbitrairement, soit par le biais d'une heuristique permettant de se placer dès le départ au

niveau d'une région "intéressante" du paysage, par exemple en appliquant tout d'abord une recherche de maximum de vraisemblance, et considérant ce maximum comme notre point de départ. La marche aléatoire est ensuite déterminée par l'algorithme suivant : considérons un point $X = (x_1, x_2, ..., x_n)$

- Soit un voisin $Y = (y_1, y_2, ...y_n)$ aléatoire de X
- Si $f(Y) > f(X)$ alors le nouveau point courant choisi est Y, sinon
- Soit r une variable aléatoire tirée uniformément dans l'intervalle $[0; 1]$, si $\frac{f(Y)}{f(X)} < r$ alors Y est le nouveau point courant, sinon nous conservons X
- Réitérer à partir du nouveau point courant

En pratique, nous laissons tourner le MCMC sans prendre en compte les premiers points (*burn in*), dans le but d'éviter de biaiser notre échantillonnage par le choix du point de départ. Ce *burn in* est de longueur variable, qui peut aller d'une centaine de points à plusieurs centaines de milliers de points éliminés. Similairement, nous ne prenons en compte qu'un point tous les k points renvoyés par l'algorithme, où encore une fois, k peut être très variable. Ainsi, l'échantillonnage est robuste, mais cette robustesse est aussi (et surtout) déterminée par la longueur de la chaîne : plus nous échantillonnons de points, meilleur sera notre aperçu du paysage.

Cet algorithme nous assure ainsi d'échantillonner le paysage au *pro rata* de la vraisemblance, ce qui nous donne un bon aperçu de cette surface, sans se contenter de repérer son maximum, parfois peu informatif. En fixant tous les paramètres sauf un, cette méthode permet aussi de déterminer la distribution marginale pour ce paramètre. Il est important de noter que cette méthode est beaucoup plus coûteuse en temps de calcul que la recherche du maximum de ce paysage, il convient donc de considérer ce point avant de choisir la méthode à appliquer.

3.3 Les phylogénies

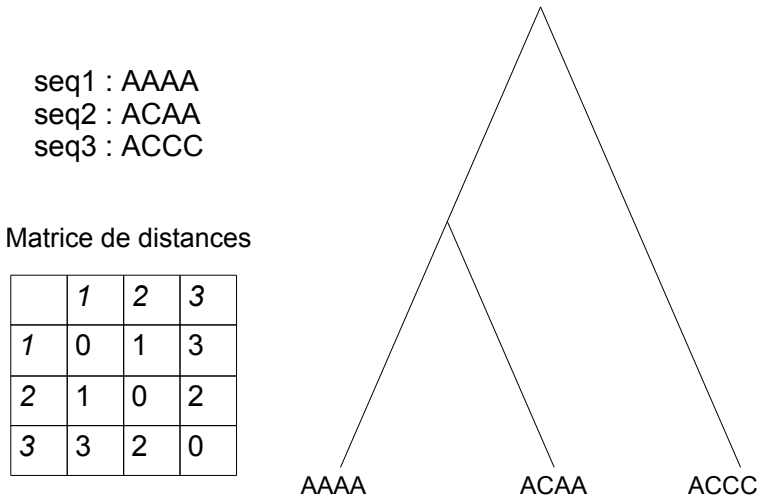
Nous présentons dans cette section les méthodes prenant en compte les phylogénies. Nous appelons phylogénie une topologie (une forme d'arbre) associée à des longueurs de branches. Cette phylogénie n'est pas observable, donc une fois de plus, la matière première reste généralement le MSA, c'est à dire l'alignement des séquences génétiques actuelles. C'est à partir de ces MSAs que nous pouvons reconstruire tout d'abord une phylogénie, mais aussi les historiques mutationnels, par exemple en inférant les états ancestraux correspondant aux séquences, ce qui permet, en comparant les nœuds consécutifs de l'arbre, de déterminer la position des mutations.

Nous considérons pour des raisons de clarté uniquement des arbres binaires, c'est à dire des arbres dont chaque nœud possède exactement deux descendants (nœuds internes) ou aucun (feuilles, ou nœuds terminaux).

3.3.1 Parcimonie

Dans des cas relativement simples, il est possible de reconstruire un arbre à partir d'observations simples, comme certains traits biologiques ou caractères phénotypiques. Cette façon de reconstruire une phylogénie est particulièrement appropriée dans le cas de l'étude d'un échantillon d'espèces suffisamment éloignées et présentant assez de caractères discriminants. C'est aussi possible lors de l'étude de séquences relativement courtes. Pour prendre un exemple simple, considérons les séquences nucléotidiques suivantes : `seq1 : AAAA, seq2 : ACAA, seq3 : ACCC`. Cet exemple nous permet d'illustrer facilement le fait qu'en considérant les différences entre les séquences citées, on déduit facilement que la première et la deuxième sont plus proches l'une de l'autre qu'elles ne le sont de la troisième. Ainsi, cette paire constitue deux nœuds "frères" dans l'arbre phylogénétique inféré, la troisième séquence complétant l'arbre, dont nous pouvons voir un exemple de reconstruction dans la figure 3.1.

FIGURE 3.1 – **Exemple de reconstruction d'un arbre phylogénétique par parcimonie, matrice de distance entre les séquences associées**.



	1	*2*	*3*
1	0	1	3
2	1	0	2
3	3	2	0

Quelle que soit la façon dont l'arbre a été reconstruit, il est aussi possible de reconstruire

les états ancestraux en considérant l'information phylogénétique. De la même manière, la parcimonie - permettant de minimiser le nombre de mutations dans l'arbre - permet cette reconstruction et ainsi l'inférence des historiques mutationnels, pour chaque branche terminale de l'arbre.

3.3.2 Clustering hiérarchique

Les méthodes de clustering hiérarchique constituent une classe de méthode qui, à partir d'une matrice de distance (telle celle de la figure 3.1), permet de reconstruire un arbre phylogénétique en regroupant les séquences considérées de proche en proche, en commençant par les paires minimisant cette distance.

Nous pouvons par exemple citer la méthode UPGMA (Unweighted Pair Group Method with Arithmetic mean) qui fonctionne selon l'algorithme suivant :

- Choisir les deux séquences les plus proches (selon la matrice de distance) et les associer dans l'arbre. Les supprimer de la matrice de distances
- Ajouter un nouvel élément "consensus" dans la matrice, qui sera la moyenne arithmétique des distances aux deux séquences préalablement supprimées
- Itérer le processus jusqu'à ce que la matrice de distance soit vide

Cette méthode suppose que la vitesse d'évolution est constante dans toutes les branches de l'arbre, ce qui est rarement vérifié. D'autres méthodes comme la parcimonie (vue plus haut) ou le Neighbor Joining (Saitou et Nei, 1987) sont plus couramment utilisées. Cette dernière est proche de l'UPGMA, à ceci près qu'elle affine la notion de distance en prenant en compte les vitesse d'évolution le long des différentes branches de l'arbre reconstruit.

3.3.3 Vraisemblance

Pour des cas plus complexes, en particulier lorsque l'échantillonnage est plus important (par exemple, reconstruction d'un arbre phylogénétique pour de nombreuses espèces, sur des séquences longues), la parcimonie n'est pas la meilleure solution. Du fait que l'histoire des différentes parties de la séquence n'est pas homogène (typiquement du fait d'évènements de recombinaison), il peut être impossible, en minimisant le nombre de mutations, de trouver un arbre consensus satisfaisant.

Dans ce cas, une stratégie couramment adoptée est l'utilisation de méthodes utilisant des calculs de vraisemblance sous des modèles d'évolution moléculaire (voir section 3.2). En

effet, à partir du moment où il est possible de calculer la probabilité d'un arbre conditionné aux états ancestraux, il est aussi possible d'estimer le "meilleur" arbre, expliquant le mieux les données ou d'échantillonner dans l'espace des arbres au *pro rata* de leurs vraisemblances (pour le MCMC). Ainsi, les méthodes de maximum de vraisemblance (Felsenstein, 1973, 1981) et de MCMC (Lartillot, 2006) s'appliquent aussi à la reconstruction phylogénétique (voir le livre de Felsenstein (2004) pour une description complète de ces méthodes). La véritable difficulté ici consiste à explorer correctement l'espace de recherche. Cet espace est constitué de toutes les topologies d'arbres possibles, ainsi que des longueurs de branches, pour chacune des branches de ces arbres. Nous avons donc un nombre très grand de paramètres à optimiser, nous pouvons donc facilement imaginer à quel point cet espace est vaste, aussi est-il nécessaire d'envisager des approches astucieuses permettant cette exploration.

Cette méthode peut impliquer de longs calculs, selon la taille du jeu de données, mais a l'avantage d'être très étudiée (Lartillot, 2006; Mateiu et Rannala, 2006). Ainsi, la littérature autour de ces méthodes est très fournie, et établit de nombreuses façons d'optimiser l'exploration de l'espace des topologies d'arbres possibles associées aux longueurs de branches. Par ailleurs, pour un jeu de données précis, il suffit de reconstruire l'arbre correspondant en amont, aussi est-il parfois judicieux d'appliquer ces méthodes, puisque le coût en temps de calcul n'est à "payer" qu'une seule fois.

3.3.4 Replacer les mutations sur la phylogénie

L'étape suivante consiste à replacer les mutations sur l'arbre phylogénétique inféré à partir des données. Encore une fois, plusieurs méthodes existent, nous revenons ici sur les deux méthodes que nous avons appliquées au cours de cette thèse, qui sont par ailleurs les plus couramment utilisées. En pratique, reconstruire les historiques mutationnels revient à reconstruire les états ancestraux au niveau des nœuds non-terminaux de l'arbre, les feuilles correspondant aux états observés dans les données.

Reconstruction par parcimonie

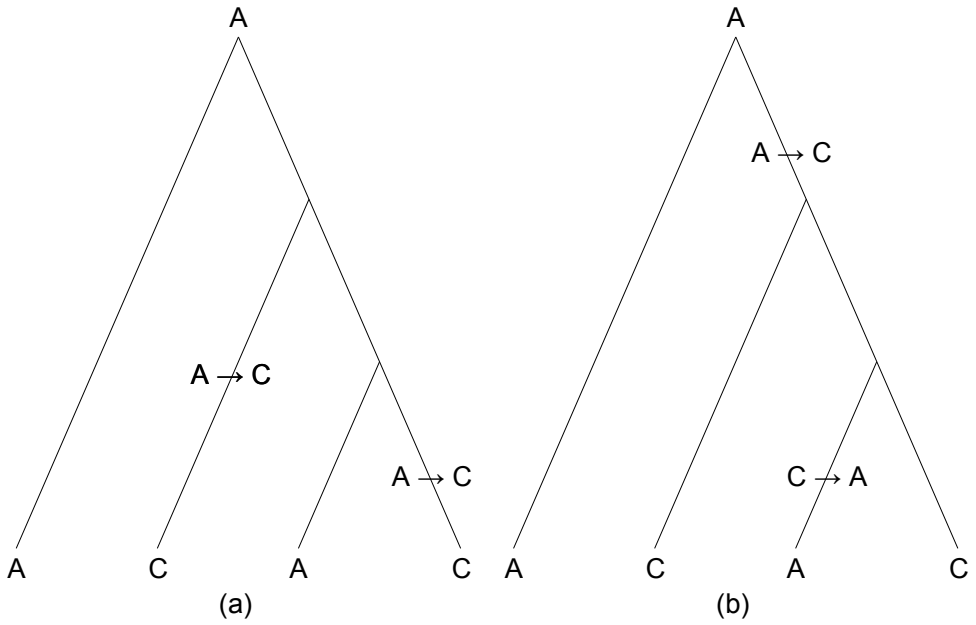
Comme vu précédemment, la connaissance des états aux feuilles (les séquences observées) ne permet pas de connaître de façon exacte et déterministe les mutations ayant conduit aux états actuels. Pour autant, plusieurs techniques permettent de les inférer. La première, vue plus haut, consiste donc à minimiser le nombre de mutations inférées conduisant à l'alignement initial, par *parcimonie*. Dans bien des cas, il faut faire des choix, puisque différentes recons-

tructions impliquant le même nombre de mutations sont possibles. Il existe donc différentes optimisations, dont les deux principales sont les suivantes :

— L'optimisation DELTRAN, pour *DELayed TRANsformation*, favorisant les convergences, et repoussant les mutations vers les feuilles de l'arbre.
— L'optimisation ACCTRAN, pour *ACCelerated TRANsformation*, favorisant les réversions, remontant les mutations vers la racine de l'arbre.

Ces deux optimisations sont illustrées dans la figure 3.2.

FIGURE 3.2 – **Exemple de reconstruction des historiques mutationnels** selon (a) l'optimisation DELTRAN et (b) l'optimisation ACCTRAN.



Le choix de l'optimisation est très important, et donc doit être matière à réflexion en regard du contexte biologique inhérent à la question étudiée. Par exemple, dans le cas de l'étude d'une fonction biologique, le nombre de gènes codant pour cette fonction est à prendre en compte, puisqu'il nous informe à propos de la possibilité de convergence et/ou de réversion : si de nombreux gènes sont nécessaires, il est plus probable de perdre cette fonction que de l'obtenir, aussi l'hypothèse DELTRAN qui multiplie les pertes tardives (plutôt qu'une perte ancienne suivi de plusieurs réversions) serait à privilégier dans cette situation (Agnarsson et Miller, 2008).

Reconstruction par calcul de vraisemblance

Comme pour la reconstruction phylogénétique, celle des états ancestraux est aussi possible à travers des méthodes impliquant le calcul d'une fonction de vraisemblance et l'étude du paysage associé à cette fonction (Pagel, 1999). Considérons ici que la phylogénie associée aux données est connue (par exemple reconstruite selon une des méthodes exposées aux sections 3.3.1 et 3.3.3). Dans ce cas, à partir d'un modèle mutationnel, nous pouvons calculer la probabilité d'observer les états aux feuilles de l'arbre concerné.

À topologie fixée, ces méthodes de reconstruction par calcul de vraisemblance permettent d'estimer :

- L'état ancestral à la racine de l'arbre
- La matrice des taux de transition entre les états dans un alphabet donné (par exemple $\{A, C, G, T\}$ dans le cas de séquences de nucléotides)
- Les longueurs de branches calibrées en temps

Il est bien sûr possible de fixer un ou plusieurs de ces paramètres et d'estimer les autres. C'est le cas par exemple de la méthode Pagel et al. (2004) implémentée dans le logiciel *BayesTraits*, pour laquelle nous fixons la topologie de l'arbre ainsi que les longueurs de branches. Le modèle qu'il utilise admet comme paramètres les taux de substitution pour chaque paire d'états (typiquement, dans le cas de l'ADN, les taux de transitions $A \to C$, $A \to G$, etc, pour toutes les paires possibles) ainsi que l'état ancestral à la racine de l'arbre. Dans ce modèle, tous les sites ont le même comportement, c'est à dire la même matrice de taux. Il est ainsi possible de reconstruire les états à chaque nœud interne, à partir de la racine de l'arbre et de son état, en propageant les mutations selon les taux définis dans le modèle. Il est donc possible, en fonction des paramètres du modèle, de donner la probabilité de chaque état à chaque nœud de l'arbre. Bien sûr, il y a toujours au moins deux façons d'explorer l'espace des paramètres : par ML, pour estimer les états ancestraux les plus probables relativement aux données observées, et par MCMC, permettant l'échantillonnage pour une étude plus fine et plus exhaustive.

Notons que cette méthode renvoie une distribution pour les états de chaque site, à chaque nœud. Les états ancestraux ne sont donc pas nécessairement connus avec exactitude, et ainsi, la reconstruction des historiques mutationnels (*i.e.* les positions des mutations sur les branches de l'arbre) est sous forme de distribution. Nous ne pouvons donc pas systématiquement déterminer la présence d'une mutation sur une branche, mais plutôt donner la probabilité de présence de cette mutation. En conséquence, le nombre de mutations inféré n'est pas toujours entier en espérance.

Cette méthode estime donc la matrice de taux optimale en regard de la topologie de l'arbre et des états aux feuilles. Dans le cas de l'ADN, la matrice est de taille 4×4, soit 16 paramètres à estimer en plus de l'état à la racine de l'arbre. Ajouter des états possibles augmente considérablement le nombre de paramètres à estimer, qui monte par exemple à 20×20, soit 400 paramètres dès qu'il s'agit de séquences d'acides aminés. Il faut donc prendre en compte les temps de calculs potentiellement très longs avant de choisir d'appliquer une telle méthode.

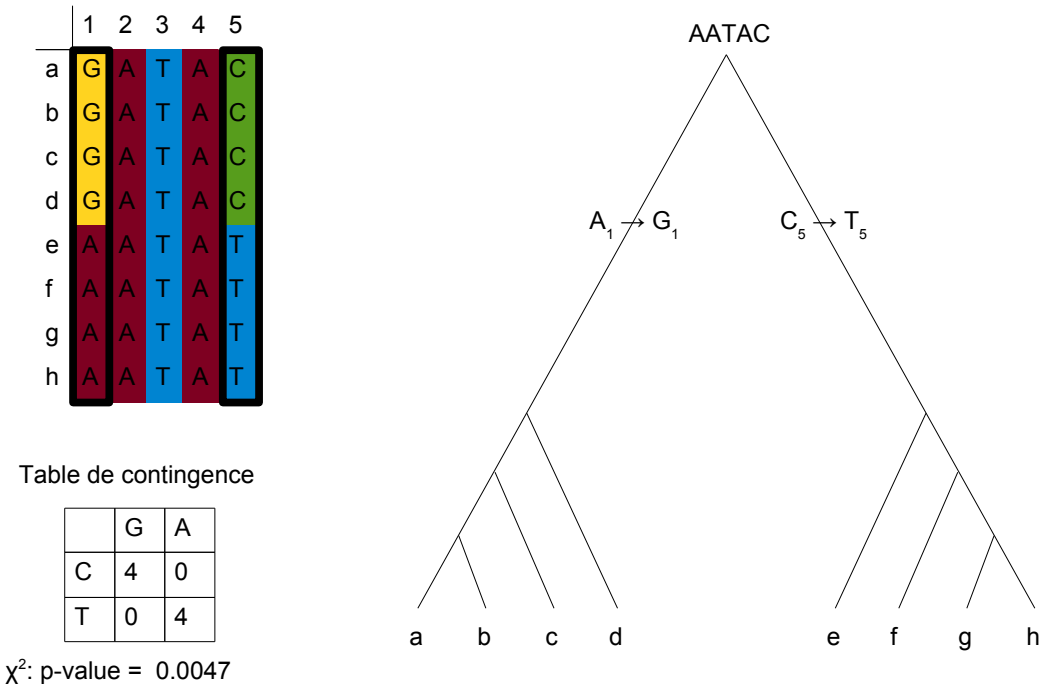
3.4 Méthodes de détection de la coévolution basées sur la phylogénie

Revenons à notre test statistique introduisant cette section (voir 3.1.1). Nous avons vu que de simples tests permettaient d'avoir une première idée de la corrélation qui peut exister entre deux sites, au sein d'un MSA. La figure 3.3 montre l'application d'un tel test, et nous éclaire sur l'influence que peut avoir l'histoire phylogénétique sur l'histoire évolutive commune (Felsenstein, 1983). Dans ce cas particulier, nous avons un alignement pour 8 individus (de a à h), pour une séquence de taille 5. Un test de χ^2 sur la table de contingence associée aux colonnes (donc aux sites) 1 et 5 nous renvoie une *p-value* significative. Or, en observant l'arbre phylogénétique associé, nous remarquons que ce signal de coévolution est principalement porté par la phylogénie.

En particulier, calculer un coefficient de corrélation sur une table de contingence construite à partir d'un MSA ne nous renseigne pas sur la part de la phylogénie dans la corrélation constatée. Il peut donc s'avérer important de prendre en compte celle-ci pour détecter la coévolution et éliminer certains faux-positifs, comme cela a été montré plus en détail par Pagel (2000) et Dutheil (2012).

La seconde classe de méthodes que nous exposons donc dans cette introduction est celle des techniques utilisant l'information phylogénétique explicitement pour inférer la coévolution. Nous détaillerons ici une méthode en particulier (Dutheil et al., 2005), qui a inspiré une partie du travail présenté dans ce document, ainsi qu'une catégorie plus générale de méthodes, très utilisée, y compris par nous-même dans la deuxième partie de cette thèse.

FIGURE 3.3 – **Exemple d'arbre phylogénétique associé à un alignement**, la table de contingence associée pour les sites 1 et 5, *p-value* associée à un test du χ^2 d'indépendance.



3.4.1 Une méthode non-paramétrique (Dutheil et al., 2005)

Cette méthode est très importante pour la suite de cette thèse. En effet, toute la première partie de notre travail peut être vue comme une extension de cette méthode, puisque nous partons des mêmes bases (conversion des données dans un formalisme d'algèbre bilinéaire, calcul de statistiques à partir de ce formalisme), en ajoutant certains aspects clefs de notre étude, en l'occurrence la temporalité et le calcul analytique, donc exact, de *p-values*.

Cette méthode, décrite dans Dutheil et al. (2005) consiste tout d'abord à traduire dans un vecteur les positions des mutations dans l'arbre. Considérons alors l'arbre phylogénétique et les longueurs de branches associées. Nous appelons "branches" les arêtes de l'arbre. Cette méthode situe dans un premier temps les positions des mutations le long des branches de l'arbre ("à la branche près", la position exacte sur chaque branche n'est pas nécessaire). Numérotons les branches de l'arbre entre 1 et n, où n est le nombre de branches. Nous pouvons construire, pour chaque site, le vecteur mutationnel e_i correspondant, pour lequel

la $k^{ème}$ composante contient le nombre de mutations pour le site i, sur la branche k. Ce vecteur représente donc les positions des substitutions sur l'arbre phylogénétique, pour chaque site.

La corrélation entre chaque paire de sites est donnée par le produit scalaire normalisé entre les deux vecteurs de substitution d'intérêt. Cela revient exactement à calculer le cosinus de l'angle entre ces deux vecteurs. Ainsi, un angle petit correspond à deux vecteurs pour lesquels les mutations tendent à se concentrer dans le mêmes composantes, donc dans les mêmes branches de l'arbre. *A contrario*, un grand angle correspond au cas où les mutations ont tendance à se situer dans des branches différentes pour l'un et l'autre des sites de la paire étudiée. Ainsi, les paires sélectionnées comme coévoluant sont celles minimisant cette angle. Cette méthode a été étendue par Dutheil et Galtier (2007) en rassemblant les paires de vecteurs dans des *clusters* concentrés dans de faibles intervalles d'angles, c'est à dire en des ensembles de sites coévoluant en groupe.

Cette méthode est donc très efficace en termes de temps de calculs, ainsi qu'en termes de précision, mais elle détecte principalement des paires de sites en très forte coévolution, puisque le signal qu'elle capture est intégralement constitué de paires dont les mutations sont situées sur les mêmes branches.

3.4.2 Méthodes utilisant la vraisemblance

Cette classe de techniques nous intéresse tout particulièrement, puisqu'elle fait appel à des outils que nous avons utilisés pour développer la seconde partie de notre méthode, qui sera exposée ultérieurement. Comme pour la reconstruction phylogénétique et la reconstruction des états ancestraux, ces méthodes reposent tout d'abord sur des modèles, en l'occurrence ici sur des modèles de coévolution. Une fonction de vraisemblance est associée à ce modèle, permettant l'optimisation des paramètres de ce modèle, de façon à ce qu'ils expliquent au mieux les données. Une fois le modèle mathématique correctement spécifié, cette méthode a pour avantage d'être puissante et de pouvoir caractériser des scénarios de coévolution entre sites, par le biais de paramètres optimaux estimés. Par ailleurs, la méthode de MCMC permet à nouveau d'échantillonner le paysage de vraisemblance, donc de donner un aperçu plus ou moins exhaustif (selon la longueur de la chaîne) des différents scénarios possibles.

Comme vu précédemment, ce type de méthode peut s'avérer très coûteuse, selon le modèle utilisé, et bien sûr, plus le modèle est complexe (*i.e.* plus le nombre de paramètres est grand), plus ce coût est important. C'est par exemple le cas du modèle décrit par Pagel (1994). Celui-ci décrit en effet l'évolution conjointe pour une paire de sites en déterminant une matrice de

transition, pour chaque paire d'états possibles. Par exemple, dans un exemple à deux loci et deux états 0 et 1, les paires possibles sont $(0,0)$, $(1,0)$, $(0,1)$ et $(1,1)$. Ainsi, la matrice de transition correspondant à ce système dépend de 8 paramètres, puisque qu'une transition est définie comme un changement d'état sur un seul locus (la transition $(0,0) \rightarrow (1,1)$ n'est donc pas autorisée).

Estimer 8 paramètres revient donc à explorer un espace à 8 dimensions défini par la fonction de vraisemblance associée à ce modèle. Le temps de calcul nécessaire à cette optimisation est déjà important, donc si l'on multiplie le nombre de sites et le nombre d'états (disons par exemple que nous avons besoin d'au moins 4 états A, C, G et T pour l'étude de séquences d'ADN), alors ce temps de calcul augmente drastiquement, puisque cette fois, ce sont 48 paramètres qu'il faut optimiser pour seulement 2 sites. Cette méthode est donc principalement à utiliser pour des cas très limités, est et difficile à adapter à des jeux de données plus fournis (Pagel et Meade, 2006).

Chapitre 4

Motivation à développer une nouvelle méthode

Cette section aurait aussi pu être titrée “Un bref historique de mon doctorat”, tant la construction des outils et méthodes présentés les parties suivante de cette thèse ont évolué et été construits progressivement, en même temps que sa problématique s'affinait. Bien sûr, je ne vais pas exposer ici de façon exhaustive les divers tâtonnements qui ont conduit à la conclusion de ces années de travail, que vous lisez actuellement, mais il me semblait important d'en parler, au moins pour les grandes lignes.

4.1 La première question

Notre question initiale, au début de cette thèse, était motivée principalement par la considération de deux des exemples cités au cours de cette introduction, les Compensated Pathogenous Deviations et le travail de Dan Weinreich sur les paysages adaptatifs. Au vu de ces exemples, il paraît clair qu'il existe des mécanismes qui contraignent suffisamment les mutations pour favoriser, voire interdire, certains ordres. Les contraintes stériques au sein d'une molécule en sont un exemple, mais on peut imaginer que des contraintes telles que la perturbation d'une chaîne métabolique conduit à privilégier certains chemins évolutifs plutôt que d'autres.

La question était la suivante : *quelles contraintes pour l'ordre des mutations ?* L'idée était donc de développer un outil à même de déterminer quelles mutations, dans une séquence, sont liées dans le temps, voire nécessairement ordonnées. Par ailleurs, en sus de la *détection* de la coévolution, il était aussi intéressant de pouvoir la *quantifier*. Nous nous étions fixé

pour but de construire une méthode efficace en termes de temps de calcul, capable de traiter d'importants jeux de données, de façon à pouvoir estimer, parmi une grande masse de sites, quelle proportion de paires étaient associées de façon privilégiée.

4.2 La temporalité, chronologie des évènements

Cet outil devait donc être à même de discriminer les paires de mutations qui se suivent dans le temps, en autorisant des délais plus ou moins longs, et de façon efficace. Par ailleurs, nous avons vu plus haut qu'il existait déjà plusieurs techniques permettant de répondre dans une certaine mesure à notre question. Mais deux points doivent retenir ici notre attention :

— Plus la technique est précise en termes de statistiques, notamment, plus elle risque d'être coûteuse en temps de calcul. Ce point vaut en particulier pour les méthodes reposant sur des calculs de vraisemblance. On en déduit de là que si nous désirons utiliser une telle méthode, alors le nombre de paramètres du modèle associé doit être petit.
— La plupart des méthodes existantes infèrent une coévolution que l'on pourrait qualifier de "forte", c'est à dire qui repère les mutations ayant fréquemment lieu sur les mêmes branches de l'arbre phylogénétique, donc très proches dans le temps. De là vient une question : Comment intégrer la notion de temporalité sans trop "alourdir" les calculs nécessaires ?

Nous avons discrétisé notre méthode selon les branches de l'arbre pour se contenter de la résolution "à la branche près". Ainsi, pour chaque mutation, la position est définie parmi les branches de l'arbre, en laissant de côté la position exacte sur la branche (ou arête), ce qui laisse autant de choix que l'arbre a de branches, et non plus une infinité. On voit bien ici le gain en terme de calculs, de même que l'analogie avec le travail de Julien Dutheil détaillé plus haut.

Nous avons aussi pu introduire la notion de temporalité, en distinguant les paires d'évènements ayant lieu sur les mêmes branches que nous appelons *coocurrences*, et celles ayant lieu sur des branches distinctes d'une même lignée que nous nommons *chronologies*. Cette distinction sur les paires d'évènements est certes largement dépendante de la résolution de l'arbre, et en particulier des longueurs de branches, mais elle permet de facilement abstraire le problème qui nous intéresse et de développer des statistiques.

4.3 Allons plus loin dans l'abstraction

Nous avons donc décidé de distinguer les cooccurrences des chronologies, qui dépendent de la topologie de l'arbre. L'enjeu est donc de coder la topologie de l'arbre en question - ainsi que ses longueurs de branches - et la position des mutations dans cet arbre, pour pouvoir, à partir de ce formalisme "bas niveau", développer des statistiques pour estimer la dépendance entre les processus évolutifs sous-jacents à notre problème.

Une conséquence directe de ce formalisme est qu'il nous ouvre la porte à des problèmes bien plus larges que ceux abordés initialement. En effet, puisque les mutations placées sur l'arbre phylogénétique sont formalisées comme des entiers dans un vecteur, pourquoi se cantonner seulement à ce type d'évènement évolutif ? En pratique, il est strictement équivalent de traiter ainsi des mutations sur une séquence ou le gain ou la perte d'un caractère ou d'une fonction biologique. Plus généralement, ce formalisme nous permet de traiter n'importe quel type d'évènement évolutif, dans la mesure où il est discret et où ses diverses occurrences peuvent être replacées sur les branches d'un arbre phylogénétique. Nous dirons donc que nous travaillons non plus sur l'ordre des mutations, mais plus généralement sur des *évènements évolutifs*.

4.4 De la construction de statistiques

Évidemment la construction d'un tel formalisme et des statistiques associées n'est pas immédiate, nous avons dû tâtonner avant de pouvoir conclure sur une théorie rigoureuse. La base sur laquelle nous nous sommes appuyés au départ est somme toute assez classique, il s'agit d'étudier des expressions de la forme $\frac{attendu-observé}{N}$. Cette grandeur nous permet typiquement d'estimer l'écart entre les données que l'on observe directement avec ce qu'on attendrait sous une hypothèse H_0 où ces processus seraient indépendants, où N est un facteur de normalisation, qui nous autorise à comparer les paires entre elles. Nous avons donc dans un premier temps développé un ensemble de statistiques variées, plus ou moins sous cette forme, et plus ou moins empiriques. Malgré tout, ce n'était pas complètement satisfaisant, puisque nous cherchions à unifier ces statistiques dans un formalisme élégant, et surtout formel.

Nous avons développé un formalisme matriciel, qui traduit dans une matrice la topologie de l'arbre, selon diverses contraintes développées plus loin. Dans le même esprit, les longueurs de branches ainsi que les positions des occurrences de chaque évènement sont traduits dans des vecteurs. À partir de là, nous pouvons calculer efficacement, et surtout analytiquement,

des $z-scores$ ainsi que des $p-values$ sous l'hypothèse d'indépendance H_0, nous permettant d'une part de classer les paires selon leur intérêt relatif à notre question (quelle dépendance ?) et d'autre part d'estimer la confiance que nous pouvions avoir en nos résultats.

Enfin, nous avons bien entendu testé cette méthode, sur des données biologiques, mais tout d'abord par simulation, pour vérifier la puissance de celle-ci, et sa robustesse relativement à différents paramètres (taux d'occurrence, forme de l'arbre, *etc*). Nous avons donc développé en parallèle un modèle d'interaction entre deux processus évolutifs, ou nous contrôlons l'amplitude de la coévolution. Les vérifications d'usages étaient probantes et la méthode efficace et robuste, mais malgré tout, elle ne restait pas complètement satisfaisantes. Il est intéressant de *détecter* la coévolution, mais est-il possible de la *mesurer* ? Nous nous sommes rendus compte qu'à partir de ce modèle de coévolution, il est tout à fait possible de calculer une fonction de vraisemblance, qui nous permet, par le biais de son optimisation, d'estimer les paramètres du modèle expliquant le mieux les données, et ainsi d'estimer le scénario de coévolution le plus probable (relativement à notre modèle).

Deuxième partie

Résultats

Chapitre 5

Méthode non-paramétrique

5.1 Introduction et résumé des résultats

Dans ce premier article, nous exposons une méthode non-paramétrique de détection de la coévolution. Les données traitées par cette méthode sont un arbre phylogénétique ainsi que les positions des occurrences de deux évènements évolutifs sur celui-ci. Les évènements en question peuvent donc être de natures très variées, tant qu'ils sont ponctuels et que l'on peut replacer leurs occurrences sur une phylogénie. Ainsi, nous pouvons appliquer cette méthode à des mutations sur des sites particuliers d'un génome, mais aussi au gain/perte d'un gène ou même d'une fonction biologique.

Pour une paire d'occurrences de chacun des deux évènements considérés, on dira qu'elle forme une *cooccurrence* si ces elles sont sur une même branche de l'arbre, et une *chronologie* si elles sont situées sur deux branches distinctes d'une même lignée (*i.e.* d'une même suite de branches allant de la racine de l'arbre à une de ses branches terminales). La notion de cooccurrences décrit les interactions fortes, puisque les évènements en questions sont proches dans le temps, mais ne permet pas de les ordonner. *A contrario*, la notion de chronologie décrit des interactions *a priori* plus modérée, mais pour lesquels nous connaissons l'ordre entre les évènements, puisqu'ils sont situés sur des branches distinctes.

D'un point de vue technique, cette méthode traduit la position relative des branches de l'arbre dans différentes matrices S et Id. Nous définissons aussi le vecteur des longueurs de branches et, pour chaque évènement E_i, le vecteur e_i décrivant les positions de ses occurrences sur les branches de l'arbre. Ainsi, tout le système considéré est décrit dans un formalisme matriciel, qui nous permet, pour deux évènements évolutifs, d'une part de compter le nombre

de cooccurrences (resp. de chronologies) qu'ils forment à l'aide de la formule générale $e_1^T M e_2$, qui a pour valeur le nombre de cooccurrences dans lesquels la paire $(E_1; E_2)$ est impliquée si $M = Id$ (la matrice identité), et le nombre de chronologies si $M = S$. Cette même formule permet aussi de compter à la fois les cooccurrences et les chronologies si $M = S + Id$.

Par la suite, ce formalisme permet de calculer les moments *exactes* de ces comptages (espérance, variance) sous une hypothèse d'indépendance H_0. Sous cette même hypothèse, nous pouvons calculer analytiquement une *p-value* associée à ce comptage. Ceci nous permet de rejeter ou non l'hypothèse H_0, avec un seuil de confiance *a priori* (typiquement, 95%).

Nous avons testé avec des résultats satisfaisants cette méthode sur des données simulées grâce à un modèle de coévolution que nous avons développé, et établi des courbes de puissance nous donnant un bon aperçu des forces et des limites de la méthode. Enfin, nous avons testé la méthode sur un exemple biologique, en étudiant le lien qui pouvait exister entre (i) la perte du flagelle chez certaines souches d'*Escherichia coli* et (ii) le passage dans un milieu de vie intracellulaire. Nous avons montré que l'intracellularité précède probablement la perte du flagelle.

Cet article a été accepté par Systematic Biology le 21 Janvier 2016.

5.2 Article 1

Version dated: February 10, 2016

CORRELATED EVOLUTION IN A PHYLOGENETIC TREE

Testing for independence between evolutionary processes

Abdelkader Behdenna[1,2,3], Joël Pothier[5,2], Sophie S Abby[6,7], Amaury Lambert[4,3], and Guillaume Achaz[1,2,3]

[1]*UMR 7138 - Institut de Biologie Paris-Seine , UPMC, Paris*

[2]*Atelier de BioInformatique, UPMC, Paris*

[3]*SMILE - Centre Interdisciplinaire de Recherche en Biologie, Collège de France, Paris*

[4]*Laboratoire de Probabilités et Modèles Aléatoires, UPMC, Paris*

[5]*Enzymologie de l'ARN, UPMC, Paris*

[6]*Institut Pasteur, Microbial Evolutionary Genomics, Paris*

[7]*CNRS, UMR 3525, Paris*

Corresponding author: Abdelkader Behdenna, UMR7138, UPMC, 7 quai St Bernard, Bat. A porte 427, 75252 Paris Cedex 05, France; E-mail: kaderbehdenna@gmail.com

Abstract

Evolutionary events co-occurring along phylogenetic trees usually point to complex adaptive phenomena, sometimes implicating epistasis. While a number of methods have been developed to account for co-occurrence of events on the same internal or external branch of an evolutionary tree, there is a need to account for the larger diversity of possible relative positions of events in a tree. Here we propose a method to quantify to what extent two or more evolutionary events are associated on a phylogenetic tree. The method is applicable to any discrete character, like substitutions within a coding sequence or gains/losses of a biological function. Our method uses a general approach to statistically test for significant associations between events along the tree, which encompasses both events inseparable on the same branch, and events genealogically ordered on different branches. It assumes that the phylogeny and the mapping of branches is known without errors. We address this problem from the statistical viewpoint by a linear algebra representation of the localisation of the evolutionary events on the tree. We compute the full probability distribution of the number of paired events occurring in the same branch or in different branches of the tree, under a null model of independence where each type of event occurs at a constant rate uniformly in the phylogenetic tree. The strengths and weaknesses of the method are assessed via simulations; we then apply the method to explore the loss of cell motility in intracellular pathogens.
(Keywords: Coevolution, phylogeny, statistical test)

Because life implies interactions at all levels of organization (from molecules to species communities), the evolution of any part of a living system depends on how the other parts evolve. The impact of any change that occurs in one part is at least partly influenced by the states of the interacting partners. This, *de facto*, creates a network of dependencies among evolutionary events. At the molecular level, amino acids and nucleic acids interact in their three-dimensional structure. From that physical proximity, constraints emerge from steric effects and physicochemical properties (Lockless and Ranganathan 1999) and, more generally, from any other factors influencing the binding of molecules. Although physical interactions are likely a major source of constraints at the molecular level, proteins involved in metabolism or in regulation networks are also under strong functional constraints (Stelling et al. 2002). These constraints do not necessarily require a physical contact between the partners involved in the same network. At a higher level, ecological interactions between living organisms lead to the establishment of systems where the evolution of one species is intimately linked to the evolution of others (*e.g.*, prey/predator or host/parasite associations). These interactions result in patterns of *coevolution* between the interacting biological entities.

In this study, we describe a generic method to test for the independence of two (or more) evolutionary processes that generate discrete events on a given phylogeny. We will hereafter use a restricted meaning of coevolution, named *correlated evolution*, to refer to the non-independence between these two processes. As the processes considered here must occur in the same phylogenetic tree, we will not address coevolution between entities that do not share the same history. The discrete events generated by the processes are typically changes of state: at the molecular level, they are mutations, but more generally they are a change/gain/loss of a biological function or attribute.

Epistasis is the influence of the genetic background on the fitness change due to a mutation. A mutation can be deleterious in certain backgrounds, while beneficial in others (Phillips 2008). In the illuminating case of the compensated pathogenic

deviations (CPDs), alleles that are highly deleterious and disease associated in an organism can be fixed in closely related populations, because they are presumably compensated by other mutations. In a set of 32 selected proteins involved in human diseases, Kondrashov et al. (2002) reports that around 10% of the deleterious amino acid substitutions are wild-type alleles in related vertebrates. A quantitatively similar pattern was found in insect genomes, where mutations deleterious for *Drosophila melanogaster* were fixed in the genomes of other dipterans (Kulathinal et al. 2004). Both mutations (the deleterious one and the compensatory one) must be in strong epistasis, because without the second one, the first one can hardly survive. More generally, CPDs are instances of Dobzhansky-Muller incompatibilities (Dobzhansky 1934; Welch 2004). CPDs illustrate the notion of temporality of evolutionary events. Two possible chronological scenarios are compatible with CPDs. The first assumes that the deleterious mutation cannot be fixed before the compensatory one. In this scenario, there could be a time lag between both events. The second scenario, on the contrary, assumes that the pathogenic mutation precedes and therefore strongly favors the compensatory one; in that second case, the fixation of the compensatory mutation almost directly follows (or even co-occurs with) the first one (Weinreich and Chao 2005). Both orders are possible but correspond to different biological scenarios.

Several methods based on various approaches were proposed in the literature to quantify correlated evolution. Using continuous Markov models, likelihood methods (Tillier and Collins 1995; Pagel and Meade 2006) allow estimations of the correlation between sites. Bayesian mutational mapping is also used with the same goal, associated with various statistical tests to evaluate the hypothesis of correlated substitutions (see Dimmic et al. (2005) for a comparison of different parametric and non-parametric test statistics; and Dutheil and Galtier (2007) and Kryazhimskiy et al. (2011), detailed below). Information theory defines the entropy as the amount of information produced by a source. Thus the joint entropy is the amount of information shared by two or more sources, and is related to mutual information. Applied

to multiple sequence alignments, mutual information can be used to identify coevolving positions (Martin et al. (2005), and see Carbone and Dib (2009) for a review on different approaches using information theory).

Incorporating phylogenetic information is fundamental to predict *inter alia* the functional links between genes (Barker and Pagel 2005). Because species partially share evolutionary history, they cannot be regarded as independent for statistical purposes (Felsenstein 1985). Taking into account phylogenies (when possible) is essential to investigate evolutionary questions (Pagel 2000). The method we describe here counts the pairs (or n-tuples) of two (or more) types of events on a given phylogenetic tree. Molecular phylogenies can be inferred from sequence data using various inference techniques (*e.g.*, parsimony, distance-based methods and likelihood-based methods) (Felsenstein 2004) and the ancestral trait states can also be inferred using parsimony or a probabilistic framework (Pagel 1999). The method described below assumes that the tree is already inferred and fixed and that the events are correctly placed on the tree. However, it can be easily extended to take as input a tree distribution and/or probabilities of events along the branches.

The method extends the one described in Dutheil et al. (2005), that maps the substitutions onto the branches of a phylogenetic tree, then translates their position, for each site and each branch, into a vector of substitution events representing the positions of the substitutions on the phylogenetic tree. The correlation between two sites is estimated by calculating the normalized inner product between the two substitution vectors of interest. Note that this is equivalent to calculating the angle between those vectors (as a normalized inner product): a small angle indicates correlated evolution, since the substitutions for the concerned sites tend to concentrate in the same branches of the tree. This method has been extended by clustering pairs of vectors in groups concentrated in a small range of angles, *i.e.,* that coevolve as a group of sites (Dutheil and Galtier 2007). Nonetheless, this method detects only strong signals of correlated evolution, as the events must be very close in time (*i.e.,* on the same branch of the tree) to contribute to the signal. Kryazhimskiy et al.

(2011) describes a method to measure the acceleration of amino-acid substitution at a given site following the occurrences of amino-acid substitutions at an other site, thus using the notion of temporality. Both methods are based on empirical p-values obtained by simulations. We propose here a method to detect a large range of dependencies between any kind of discrete evolutionary processes. We translate the positions of events as well as the tree itself into a matrix formalism that counts events co-occurring on the same branch of the tree, but also on successive branches of the same lineage. Through this formalism, we can compute the expectation and the variance of the pair counts under a null model where the two substitution processes are independent with constant rates in the whole tree. We therefore build a centered and reduced Z-score statistic to classify the pairs of sites according to the signal of correlated evolution they carry. We further calculate an exact p-value for a given pair count to formally test for the independence between the processes.

1 Model

The model described here translates a tree and the positions of the occurrences of two evolutionary events into a matrix formalism. It then counts the number of the pairs of occurrences of interest and associates each pair count to Z-score and an exact *p-value*.

On a rooted tree T, a *branch* is an edge between two nodes, while a *lineage* is a sequence of consecutive branches on a path between the root and a leaf of the tree. We assume that two types of events, namely E_1 and E_2, occur in T in known branches.

We assume that both types of events are generated by Poisson processes. When the processes associated with each type of events are independent, there is no correlated evolution. Because our method tests if the model of independent processes can account for the observed data, we will refer to it hereafter as H_0. Under H_0, once

the number of events is known, their distributions are independent and uniform on T. Conversely, models where the processes are not independent will generate events that are not uniformly and independently distributed on T.

When an E_1 event occurs on the same branch as an E_2 event, we count it as an *inseparable pair* (Figure 1). When the E_2 event is in the subtree defined by the node under the E_1 event (*i.e.*, the E_1 event precedes an E_2 event), we count it as a *genealogically ordered pair*; in this last case, the order of both events is unambiguously known since at least one node of the tree separates them.

In a scenario where the E_2 events are triggered by the E_1 events, we expect a large number of genealogically ordered pairs and inseparable pairs, their respective numbers depending on the strength of the interaction between both processes. However, the simple counts of genealogically ordered and inseparable pairs are not entirely satisfying for at least two reasons. First, when there are many E_1 or E_2 events, we expect many genealogically ordered and inseparable pairs, even when H_0 holds. Second, when an E_1 event occurs close to the root of T, it caps a large subtree and hence, the probability of observing the genealogically ordered pair (E_1, E_2) is large under H_0.

Let m be the number of branches of T, and $\{b_1, ...b_m\}$ the set of its labelled branches of respective lengths $\{L_1, ...L_m\}$. We further denote the normalized vector of branch length $\mathbf{l} = (l_1, l_2, ..., l_m)$ where for $i = 1, 2, ..., m$, $l_i = \frac{L_i}{\sum_{j=1}^{m} L_j}$.

Adjacency matrices: Two branches of the tree will be said to form an adjacency if they are strictly consecutive in the tree. Let A be the adjacency matrix of the tree (see Figure 1): $A \in M_m$ is a square matrix of order m such that for $i, j = 1, 2, ..., m$,

$$A_{i,j} = \begin{cases} 1 & \text{if } b_j \text{ is consecutive to } b_i \\ 0 & \text{otherwise.} \end{cases}$$

The matrix A contains all the information about the possible adjacencies in the tree and therefore its topology.

We also define S, the matrix corresponding to the possible genealogically ordered pairs in T. We can compute S as follows: for $(i, j) = 1, 2, ..., m$,

$$S_{i,j} = \begin{cases} 1 & \text{if } b_j \text{ is in a subtree of } T \text{ rooted by } b_i \\ 0 & \text{otherwise.} \end{cases}$$

For $n \in \mathbb{N}$, one can show that A^n is the matrix whose generic element $(A^n)_{i,j}$ is 1 if the branches i and j form a n-adjacency (*i.e.,* are separated by exactly $n-1$ branches), 0 otherwise, as A^n represents all the n-length paths in the associated tree. Thus we can use this class of matrices to count in a more precise way the different dependencies in the tree. Also note that the identity matrix of order m is $Id = A^0$ and that $S = \sum_{n=1}^{h} A_n$, where h is the height of the tree, *i.e.,* the length of the longest lineage of the tree.

Hereafter, we will use the generic matrix M to refer to any linear combination of the matrix A^n such as the matrix S.

1.1 Inseparable pairs and genealogically ordered pairs

E_1 and E_2 are the two types of events whose occurrences are respectively characterized by the vectors e_1 and e_2, in which the i_{th} component contains the number of events of type E_1 or E_2 in the i_{th} branch (Figure 1). For any matrix M, the number of times the pair (E_1, E_2) appears in a dependency defined by M can be computed simply by:

$$\text{pairs}_M = e_1^T M e_2 \tag{1}$$

where e_1^T denotes the transpose of the vector e_1.

Note that in this definition of pairs counting, different configurations of variable interest lead to the same result. For instance, one event of type E_1 above 10 events of type E_2 is equivalent to 10 events of type E_1 preceding only one event of type

E_2 (both count as 10 genealogically ordered pairs). In this particular example, only the first case seems relevant. The probabilities of the two configurations are however very different under H_0 (see below).

More than two events: One can expand this formalism to counts of triples among three types of events (E_1, E_2, E_3). Equation 1 simply becomes

$$\text{triplets}_{M_a,M_b} = e_1^T M_a \text{Diag}(e_2) M_b e_3$$

where M_a defines an adjacency of interest for events of types E_1 and E_2, whereas M_b defines the adjacency of interest for events of types E_2 and E_3. Note that M_a and M_b are not necessarily identical. $\text{Diag}(e_2)$ represents the diagonal matrix where the diagonal elements correspond to the elements of e_2. A general formula for k types of events would follow the same structure.

Many combinations are possible and address different questions. For example, considering $M_a = M_b = S$ permits the count of the genealogically ordered triplet $E_1 \rightarrow E_2 \rightarrow E_3$, while $M_a = M_b = Id$ corresponds to the inseparable triplet $E_1 \leftrightarrow E_2 \leftrightarrow E_3$. $M_a = M_b$ is not a necessary condition, so different combinations of matrices are also possible: for instance $M_a = S$ and $M_b = Id$ corresponds to the pattern $E_1 \rightarrow (E_2 \leftrightarrow E_3)$, where a first event triggers a correlated pair of events. Note that when two events are independently triggered by a first event (*e.g.*, $E_1 \rightarrow E_2$ and $E_1 \rightarrow E_3$), they can be studied as two independent pairs.

1.2 Under the model of independence

This matrix formalism allows us to derive by exact calculation the expected pair counts and their associated variances (see supplementary material) by conditioning either a) on the total number of occurrences of both types of events, n_1 and n_2, or b) on the position of the E_1 events in the tree together with n_2, the number of E_2 events. In the first case, all occurrences of both types of events are uniformly placed

on T, whereas in the second case, only the E_2 events are randomly placed on T. In the first case, we take into account the potentially large number of occurrences of both events and, in the second case, we also correct for the position of the E_1 events, that are potentially close to the root of the tree and thus generate a large number of genealogically ordered pairs. It is possible to compute the related Z-scores, the centered and reduced pair counts $(X - \mu_X)/\sigma_X$, associated with both genealogically ordered and inseparable pairs. We respectively note $Z_M^{(n_1,n_2)}$ and $Z_M^{(e_1,n_2)}$ the Z-scores for both conditioning.

1.2.1 The full probability distribution

It is possible to compute numerically the full distribution of pairs_M conditioned on either (e_1, n_2) or on (n_1, n_2). Regardless of the chosen matrix, the number of pair counts concerning the two types of events lies between 0 and $n_1 n_2$. The product corresponds to the case where all events of the first type interact with all events of the second type.

When both processes are independent, pairs_M conditioned on (e_1, n_2) follows a multinomial distribution where the number of trials is n_2. The vector of pair counts that is generated by a single E_1 event occurring on a given branch of T of length l can be computed by $M^T e_1$ and has an associated probability given by l. Once all branches associated with the same number of pair counts are grouped (and their associated probabilities summed), we can compute the probability distribution by iterating recursively over all possible arrangements of the events E_2. To compute a *p-value*, we simply compute the cumulative distribution of realizing *at least* the observed number of pair counts. It is also possible to generalize the probability distributions of pair counts to the case of (n_1, n_2) although the computation time is considerably increased (see Supplementary Material).

1.3 Simulating models of interactions

In all simulations, we assumed a given tree T on which we placed occurrences of E_1 and E_2 events according to different rules. In all cases, for an event E_i, $i \in \{1, 2\}$, we placed its occurrences on the branches of the tree according to a Poisson process with parameter m_i. The parameter m_i can change as a function of the relative positions of the events on the tree. Note that m_i is the intensity of the process governing the distribution of the occurrences of the events of type E_i on the tree.

For each event E_i, its occurrence rate m_i can take two values μ_i and μ_i^*, respectively called the basal and the enhanced occurrence rates. The ratio $\lambda_i = \mu_i^*/\mu_i$ can be assimilated to an induction factor which measures the influence of an event on the other one. Note that there is no order requirement between rates: if $\lambda_i > 1$ the induction is positive whereas it is negative when $\lambda_i < 1$. We assume that the interaction only concerns the next "induced" event, as it models the impact of an E_1 event on the rate of occurrence of the next E_2 event. Framed in terms of compensatory events, the second event can be seen as compensating for the deleterious effect of the first one.

All simulations use the following rules:

1. All rates are basal at the root: $m_i = \mu_i$.
2. When an E_i event occurs, m_i switches to the basal rate μ_i, regardless of its current rate.
3. When an E_i event occurs at its *basal* rate μ_i, m_j $(i \neq j)$ switches to the enhanced rate μ_j^*.

This framework can be used to simulate many different scenarios, four of which are particularly relevant from the biological standpoint and are detailed here.

A scenario of independence. This corresponds to the H_0 model described above. When the processes are independent their intensities are constant, regardless of the position of the occurrences of each type of event on the tree. It therefore corresponds to:

$$\begin{aligned} \mu_1 &= \mu_1^* \\ \mu_2 &= \mu_2^* \end{aligned}$$

The intensities of both processes are constant.

A scenario of prohibition. This scenario models the case where E_2 events are not allowed unless they follow an E_1 event (E_1 triggers the appearance of E_2): for example, a deleterious mutation cannot occur without being compensated upstream. It echoes one of the two possible scenarios that explain Compensatory Pathogenic Deviations (CPDs, highly deleterious alleles compensated by other mutations), as seen for example in Kondrashov et al. (2002).

$$\begin{aligned} \mu_1 &= \mu_1^* \\ \mu_2 &= 0 \\ \mu_2^* &> 0 \end{aligned}$$

A scenario of induction. E_1 events have a constant rate while the rate of E_2 is increased after an E_1 event has occurred. This scenario models, for example, a compensatory mutation that follows the fixation of a deleterious mutation and corresponds to the second hypothesis to explain how CPDs occur in genomes. More generally, it describes the process by which direct compensatory events occur.

$$\begin{aligned} \mu_1 &= \mu_1^* \\ \mu_2 &< \mu_2^* \end{aligned}$$

A scenario of reciprocity. In this last scenario, there is a symmetrical interaction between both processes. E_1 events enhance the occurrence rate of E_2 events and, reciprocally, E_2 events enhance the rate of occurrence of E_1 events. This scenario simulates for example reciprocal sign epistasis, where two mutations are beneficial only in the double mutant (Weinreich et al. 2005; Poelwijk et al. 2007).

$$\begin{aligned} \mu_1 &< \mu_1^* \\ \mu_2 &< \mu_2^* \end{aligned}$$

State-dependent occurrence rates. We finally expand the model, assigning a second occurrence rate (ν_2,ν_2^*) to E_2. This mimics a case where two state transitions $A \to B$ and $B \to A$ are possible for E_2 events with different occurence rates: $A \to B$ occurs at rates (μ_2,μ_2^*) and $B \to A$ at rates (ν_2,ν_2^*). Therefore, the current occurrence rate m_2 switches between μ_2 and ν_2 (or μ_2^* and ν_2^*) at each occurrence of an E_2 event. Importantly the induction factor is kept constant: $\lambda = \mu_2^*/\mu_2 = \nu_2^*/\nu_2$.

Implementation. All the simulations and analyses used programs developed by the authors in the Ocaml language (version 3.09). The source codes are available at http://wwwabi.snv.jussieu.fr/public/EpiCs/. They are also available on Dryad under the DOI doi:10.5061/dryad.847vk.

2 Results

We assess the power of the method on two types of data. We first characterize the theoretical power and limitations of the method on simulated data. Using different controlled scenarios, we evaluate the importance of the chosen matrix, of the shape of the tree, the nature of interactions between both processes and also in the case where E_2 has a state-dependent occurrence rate. We then apply our method to one biological question to exemplify the power and the flexibility of our approach.

2.1 Analysis of simulated data

In each series of simulations, we assess the power of the method by counting the fraction of runs where the p-value computed under H_0, is 5% or less. Unless specified otherwise, we ran simulations with the occurrence rates $\mu_1 = \mu_2 = 5$, a scenario of "induction", the matrix S of genealogically ordered pairs, an *Escherichia coli* tree (see the analysis of *E. coli* biological data) and 10^4 replicates.

The importance of the M matrix The method can be used with any matrix M that will select only pairs of interest. For example, if one sets $M = S$, only genealogically ordered pairs will be counted. When one is only interested in inseparable pairs, the matrix should be $M = Id$. By setting $M = A$, the method will only detect pairs of events (E_1, E_2) that are in consecutive branches. And finally, when $M = S + Id$, the method will count both genealogically ordered and inseparable pairs. Therefore, the choice of M is central to the performance of the method. Mostly, it depends on what the biological question is.

We computed, initially for S and Id, the power curves for the "induction" scenario. Results show that the power (Figure 2a) associated with different matrices indeed depends on the induction factor λ. Id gains power when the induction increases,

reaching a plateau where almost 100% of the cases are detected for very strong induction ($\lambda > 100$).

For moderate inductions ($\lambda \approx 10$), the method detects more significant genealogically ordered than inseparable pairs. Indeed, as mentioned above, in the induction scenario, λ represents the induction of E_1 events on E_2 events; therefore events of type E_2 tend to occur in higher number and more rapidly after events of type E_1 as λ grows. The reduction in time lag between both events explains this phase transition: when the induction is sufficiently strong, inseparable pairs take precedence over genealogically ordered pairs. Note that the power curve for A has a shape similar to the curve for S (see Supplementary Material).

The combination of both genealogically ordered and inseparable pairs, $M = S+Id$, is especially appropriate when $\lambda \lesssim 10$. The method benefits from both genealogically ordered and inseparable pairs at the same time, smoothing the power curve for a gain of power around 10-fold inductions. However, when $\lambda \geq 100$, as genealogically ordered pairs are mostly replaced by inseparable pairs, the combination $M = S + Id$ results in a loss of power.

As a consequence, an adequate choice for M depends on the *a priori* assumptions one can have on the strength of induction between the events (*i.e.,* on λ). With the default simulation parameters, for $\lambda \lesssim 100$, $M = S + Id$ is a good choice, whereas, when $\lambda > 100$, $M = Id$ is a better choice.

Influence of the scenario The nature of the interaction between both processes strongly affects the power of the method. To explore this connection, we test the power to detect genealogically ordered pairs under several representative scenarios: 1) a "prohibition" scenario that forbids E_2 events before the occurrence of E_1 events, 2) an "induction" scenario for which occurrences of E_1 inflate the occurrence rate of E_2, and 3) a "reciprocity" scenario, where an occurrence of any type of event will inflate the occurrence rate of the other type.

The power to detect significant genealogically ordered pairs depends on the simulated scenario (Figure 2b): highest for the "prohibition" scenario and lowest for the "reciprocity" scenario. Indeed genealogically ordered pairs are highly favored in the "prohibition" scenario, which strictly forbids the order (E_2,E_1). As all occurrences of E_2 events follow E_1 events, we observed more genealogically ordered pairs than expected under H_0 (this is true even when $\lambda = 1$).

Because the "reciprocity" scenario is symmetrical by construction, genealogically ordered pairs are not particularly suitable to characterize such interactions.

We observe complementary results when studying Id for the three scenarios (Supplementary Figure a): "reciprocity" is more suited for the characterization of inseparable pairs than "prohibition", while "induction" is intermediate. Furthermore, in the three cases, $S + Id$ presents a smoother power curve while the inseparable pairs alone take precedence for moderate induction, between $\lambda \approx 15$ ("reciprocity") and $\lambda \approx 30$ ("induction", "prohibition").

Influence of the tree topology As genealogically ordered pairs depend largely on the subtrees, while inseparable pairs are principally determined by the branch lengths, we expect that the tree topology has a strong influence on the power. We compute the power curves for a perfectly balanced tree, a caterpillar tree (most unbalanced) and an intermediate case, the phylogenetic tree of a sample of *E. coli* strains.

In all three cases, the power to detect significant genealogically ordered pairs is maximal for induction factors $\lambda \in [10, 20]$, but the height of the power peak varies considerably (Figure 2c). The extreme topologies also have extreme behaviors, with a high power value reached for the balanced tree (45%) while it is barely 10% for the "caterpillar" tree. Since the first tree is perfectly balanced and thus consists of a succession of dense subtrees, it is well suited for genealogically ordered pairs. On the contrary, each node of the unbalanced "caterpillar" tree has one descendant

composed of one single long branch and one descendant composed of a dense subtree, privileging inseparable over genealogically ordered pairs. Note that typical biological trees are intermediate between the two extremes (Figure 2c), where the power to detect genealogically ordered pairs for the *E. coli* tree is indeed intermediate. Nevertheless, this stresses that the balance of the topology and the branch lengths of the tree of interest is a precious clue for predicting the behavior of the statistics.

As expected, the tree topology has almost no effect on the detection of inseparable pairs, since these only depend on the branch lengths, and not on the shape of the tree (Supplementary Figure b). However, the impact of the tree topology on the detection of genealogically ordered pairs influences the shape of the power curve of $S + Id$: the plateau for the *E. coli* topology is lower than for both other topologies, and the perfectly balanced tree is more suited for the use of $S + Id$ for moderate inductions ($\lambda \approx 10$).

State-dependent occurrence rates We explore the cost of ignoring the state of both events by assigning a second rate to E_2. This second rate mimics a state dependent transition, where for E_2, $A \to B$ has a basal rate μ_2, whereas the transition $B \to A$ has a basal rate ν_2. The induction is identical for both states and the rate at the root is μ_2. Although these state-dependent rates have almost no impact on the rate of false positives, they have a dramatic impact on the power of the method. We compute the power curves for the matrix $S + Id$, for three different alternative occurrence rates ν_2: 0.05, 5, and 500 (Figure 2c). Note that $\nu_2 = \mu_2 = 5$ corresponds to the case $M = S + Id$ in Figure 2a.

We observe a substantial gain in power when $\nu_2 = 500$ where the state B has an occurrence rate 100 times larger than the state A. Thus, after an E_1 event, two E_2 events occur ($A \to B$ with rate μ_2^* is immediately followed by $B \to A$ at rate ν_2). On the contrary, when $\nu_2 = 0.05$, only the first event $A \to B$ occurs and then events E_2 are inhibited. Once in state B, E_2 will follow an E_1 event with rate $\lambda\nu_2$ that is

smaller than μ_2 unless λ exceeds 100.

Other factors Other factors have a strong impact on the power of the method. For example increasing the occurrence rate tends to improve the power of the methods independently of M (Supplementary Figure c). Of most importance is how H_0 is tested, since ignoring the position of e_1 results in a complete loss of power to detect genealogically ordered pairs but does not affect the detection of inseparable pairs (Supplementary Figure d).

2.2 Intracellularity and the loss of flagellum in *Escherichia coli*

Although simulations can assess general rules about the performance and, even more importantly, the limits of the method, the analysis of real biological data completes the validation of the method. We apply the method to a non-molecular biological example, testing for the precedence of becoming intracellular over the loss of flagella in *Escherichia coli* strains.

Most species of the *Escherichia* genus are motile and possess at least one flagellum, except for *Shigella*, a non-motile subgenus of *Escherichia* (Yabuuchi 2002). Our dataset encompasses 67 strains of *Escherichia coli* and *Shigella*, plus one strain of *E. fergusonii* that is used as an outgroup (Supplementary Table S1). *Shigella* is a polyphyletic bacterial genus whose members have arisen independently from *E. coli* strains; hence, strictly speaking they belong to the *E. coli* clade (Pupo et al. 2000). We retrieved the complete genomes of these strains from GenBank RefSeq (Benson et al. 2013) and then extracted a core genome of 1,382 genes that are present in all strains. The corresponding proteins were aligned with Muscle v3.6 (Edgar 2004) using default parameters and then back-translated to DNA. The phylogenetic tree of these 68 species (Figure 3) was inferred by maximum likelihood using RAxML

v7.2.8 (Stamatakis 2006) under the model GTR+Γ. As observed before (Touchon et al. 2009), the nodes of this tree are highly supported, suggesting there is little uncertainty in the phylogenetic tree. We used HMM protein profiles from 12 core components of the flagellum along with 9 profiles for the core components of the T3SS (type 3 secretion system), in order to efficiently detect and discriminate the flagella from homologous T3SS (Abby and Rocha 2012) in the analysed genomes using the MacSyFinder framework (Abby et al. 2014). All *Escherichia sensu stricto* present in this phylogeny are motile and possess a flagellum. All *Shigella* have an intracellular living environment. Some strains still encode the essential components of the flagellum (*e.g., S. sonnei* 01) but most of them lack the complete locus. Importantly, *Shigella* strains belonging to different clades have lost different genes that encode for different subsets of the essential components of the flagellum (supporting the assumption of independent losses of the flagellum).

We analyze the correlation between the loss of the flagellum and the intracellular replication in host cells (*i.e.*, the *Shigella* strains). Cell motility through a flagellum is of little use in intracellular bacteria. Furthermore, the flagellum is a costly apparatus that is targeted by the immune system (Hayashi et al. 2001). It is therefore thought that bacteria enduring genetic degradation and niche restriction, with a focus on intracellular replication, rapidly lose the flagellum (Martínez-García et al. 2014). However, one might also consider the alternative hypothesis that loss of the flagellum correlates with niche restriction due to the combination of lower selection for dispersal and selection against appendages targeted by the immune system, and that this occurs before the evolution of an intracellular lifestyle. Hence, we have assessed the chronology of two events and tested the directionality of adaptation. To that purpose, we have first located the events on the phylogenetic tree by parsimony (Figure 3), knowing which strains possess a flagellum or not, and which ones replicate intracellularly (description of strains is given in Supplementary Table).

We first use a DELTRAN (DELayed TRANsformation) optimization, favoring late mutations (Figure 3). This assumes the prevalence of convergent evolution over

possible reversions. In our case, this choice is supported by the fact that it is easier to lose a biological function depending on multiple genes (here, a functional flagellum) than to gain it. The small number of mutations makes this reconstruction feasible by hand. We then used the method consecutively with the matrices S, Id and $S+Id$ to test for significant genealogically ordered pairs, inseparable ones or both between E_1 events, the intracellular transition and E_2 events, the loss of flagellum. For the three types of pairs, we computed the Z-scores and their associated p-values (Table 1). Counting both genealogically ordered and inseparable pairs leads to a higher departure from H_0. This illustrates that when both are observed in a tree, counting them both enhances the power to see a deviation from H_0. Given that there are both genealogically ordered and inseparable pairs, it is tempting to think that the induction of the E_1 events on E_2 events is moderate, likely in the vicinity of a 10-fold induction.

We then perform the same analysis but using an ACCTRAN (ACCelerated TRANsformation) optimization, which instead favors reversions over convergence (Figure 3). The events tend to be mostly drawn to the tree root. In this second reconstruction, we count one less genealogically ordered pair, and one more inseparable pair. Results (Table 1) show that inseparable pairs ($M = Id$) reject H_0 significantly, while counting only the genealogically ordered pairs ($M = S$) does not. This suggests that, in this reconstruction, the induction between E_1 and E_2 may be more important. Nevertheless, using $S+Id$ leads once again to a higher departure from H_0, suggesting that the induction could be, in this case, $\lambda \in [10, 100]$

The conclusion is then basically the same in both cases: the method allows us to reject significantly the independence hypothesis, and gives us an insight of the magnitude of the induction between the events, which varies considerably depending on the reconstruction method. It is noteworthy to mention that the DELTRAN case is reminiscent of the "prohibition" scenario, where E_2 events can only occur after the E_1 events (with moderate induction), whereas the ACCTRAN relates to the induction scenario where E_2 events can only occur after the E_1 with a strong induction.

However, in both cases, we observe ordered pairs ($E_1 \rightarrow E_2$) but never the reverse, suggesting that intracellularity drives the loss of flagellum and not the reverse. We furthermore favor the DELTRAN scenario as it minimizes the number of reversions compared to ACCTRAN.

As the phylogeny used here has entire clades completely devoid of events, we tested if the presence of such clades influences the results. For example, the clade marked with A or B on Figure 3 represent respectively 9.7% and 10.2% of the total tree length and carry no evolutionary events. If both clades are shrunk to a single lineage (we retain the longest), the total tree length is reduced by a factor 1.25 and the resulting p-values (for the DELTRAN reconstruction) are 0.032 for $M = S$, 0.059 for $M = Id$ and 0.0011 for $M = S+Id$. Therefore, the p-values, although slightly higher, are similar to the ones obtained when the whole tree is analyzed. The results for $M = Id$, which slightly rejected H_0 for the whole tree, are not significant anymore, but the use of both matrices ($M = S + Id$) is in this case sufficient to maintain the robustness of the method.

3 Discussion

We have developed a general method to test for the independence of evolutionary processes that generate discrete events on a given phylogenetic tree. The method captures both inseparable pairs (events that are very close in time or on the same long branch) and genealogically ordered pairs (ordered events with longer time lags or in different short branches). The method simply counts pairs (or n-tuples) whose nature is defined by a specific matrix M. The formalism is based on products of vectors that describe the positions of the events on the phylogenetic tree, as it was proposed by Dutheil and Galtier (2007), and adds the notions of temporal ordering between the events that is also used, in a different context, in Kryazhimskiy et al. (2011). It computes efficiently the expectation, variance of the counts as well as the full distribution of the counts under a model of independence. Once discrete events are mapped on a phylogenetic tree, our method can be used to test efficiently for the independence between the processes that have generated them. Assessing the independence between processes along a tree is a question of general interest that can be applied to a wide range of evolutionary questions, ranging from molecular evolution to ecology.

Using simulated controlled scenarios, we have shown that the power of our method depends on the strength of interaction. When both processes are in strong interaction, they generate inseparable pairs because the events appear close in the tree. Milder interactions lead to genealogically ordered pairs for which both events are separated by a time lag. Therefore, the choice of detecting either genealogically ordered or inseparable pairs is crucial but depends on the unknown strength interaction. One way to palliate this problem is to combine the matrices ($M = S + Id$) and to look for both genealogically ordered and inseparable pairs at the same time. However, when interactions are very strong (when the induction factor is larger than 100), there is a cost in power when $M = S + Id$ is used instead of $M = Id$.

We further showed that the performance of the method depends on several param-

eters. The topology of the phylogenetic tree is an important feature as different topologies will have different proportions of internal branches. Balanced trees have more subtrees and can give rise to more genealogically ordered pairs than unbalanced trees which are mainly independent lineages that are associated with inseparable pairs. Another key parameter is the rate of occurrence that should be ideally neither too small (producing too little signal) nor too large (generating too much noise).

Therefore, the optimal choice of the M matrix could be dictated by *a priori* knowledge of the different parameters. For instance, using the matrix S and therefore looking for genealogically ordered pairs is particularly appropriate in the case of a well-balanced tree with a weak interaction between both processes. On the other hand, when the tree is unbalanced and when the interaction is strong, looking for inseparable pairs seems more appropriate. Again, in absence of preliminary knowledge, it seems more appropriate to use $M = S + Id$. However, it is tempting to also compare the results obtained by S, Id and $S + Id$ to build a posterior intuition on the strength of the interaction between the events. This is illustrated by the case of loss of motility that follows intracellularity, where $S + Id$ gives the most significant results, suggesting that the induction factor could be around $\lambda = 10$. In some other cases (*e.g.,* Watson-Crick pairing in RNA), we have observed almost no significant genealogically ordered pairs, suggesting a much stronger induction (data not shown). As a consequence, the choice of the adequate matrix will likely depend on the biological context.

The method can handle a wide range of data as we saw previously, but it does not consider a possible change of state implying different occurrence rates. We have nevertheless shown that in this case the power of the method is ultimately related to the absolute values of the occurrence rates even for a state-dependent model.

The method is deeply grounded in phylogenetic inference. In that regard, the blending of genetic material caused by recombination or horizontal transfer challenges the

phylogenetic framework and therefore impacts our method. However, when horizontal gene transfers occur at a small rate, only few transfers are expected and the phylogeny should still be recovered with good confidence. Our method is meant to detect covariation between processes running on the tree, whatever their underlying generating mechanism. If events of gene gains and losses are studied, two genes that are typically co-transferred between lineages, would lead to significant inseparable pairs, because the gains are not independent. Furthermore, the method depends on the inference of ancestral states and of the mapping of the events on the tree. Therefore, the power of the method (or even worse the fraction of false positives) will be sensitive to errors in the mapping of events. For example, placing the events using parsimony can minimize the number of events and locates them closer to the root. We therefore expect a loss of power to detect significant genealogically ordered pairs in favour of inseparable pairs when parsimony is used.

More generally, the method measures covariations and therefore can only capture correlation between processes that generate evolutionary events in the tree. In absence of any evolutionary events, the method has no power. Obviously this depends on the number of mutations in the sequence, that is, the total tree length in evolutionary distance. For shallow trees, there will be fewer mutations and therefore more monomorphic sites. Reciprocally, for deep trees, we expect fewer monomorphic sites but more divergent regions, for which homology will be uncertain.

We have also explored the power of the method on biological data, for which we have *a priori* knowledge on the mode of evolution. We have observed some limitations and statistical biases related to some features of the data (clades devoid of events, events known to separate clades). We showed that the reconstruction of ancestral states should be based on a robust biological interpretation. In the case of the gain or loss of biological function, for example, if the events are rare, a maximum parsimony reconstruction is possible, but the assumptions for this reconstruction must be biologically sound. If many genes are required for a function, then discriminating against reversions seems to be a sensible hypothesis.

Similarly, it is important to consider the richness of the phylogenetic tree studied as clades devoid of events may represent relevant biological information. It is therefore essential to build the dataset with care. In the example of *E. coli*, we chose to limit the scope of the study to *E. coli* and *Shigella*, taking advantage of the polyphyly of *Shigella*. Reducing the number of taxa leads to similar but larger p-values, but removing all branches except the ones carrying events would result in a complete loss of power. In this line of thought, it is crucial to appreciate that the H_0 hypothesis we consider here implies that E_2 events can occur uniformly over the tree (assuming that p-values are computed for (e_1, n_2) given). The choice of the tree has to be irrespective of the choice of the evolutionary events and *vice versa*. For example, one may consider two reciprocally monophyletic groups separated by two morphological changes (one event E_1 and one event E_2 both giving rise to synapomorphic characters). If the events are located on different branches, then no pair is counted. If, on the contrary, both events occur in the same branch, there would be one correlated pair in this branch. The pair will be significant if the relative length (normalized by the total tree length) of this branch is smaller than the risk (typically 5%). In conclusion, choosing two events that are known to correspond to synapomorphic characters for the tree already violates H_0.

The choices of the sampled species and the events have to be made with care, since H0 assumes that each substitution process occurs at a constant rate everywhere on the tree. H0 assumes that each branch of the tree may potentially carry one - or more - events, without any prior information about the occurrence rates. Even a single occurrence of each event can potentially significantly reject H0 (although the power of the test to detect such events is very low). Furthermore, the test could be distorted by biased sampling, such as intentionally adding more species that are known to concentrate (or to be devoid of) the events.

Importantly, the method can be applied to any discrete evolutionary event that can be placed on a phylogenetic tree, both molecular (*e.g.,* substitution) or non-molecular (*e.g.,* gain or loss of a character). As it has been done in other methods (Lartillot

and Poujol 2011), combining molecular and non-molecular events is an exciting perspective. For example, one could study the emergence of biological functions in parallel to the gains, losses or mutations of genes to collect evidence on the complex phenotype-genotype map.

4 Acknowledgments

The authors would like to thank Marie Touchon for providing the orthologous genes of the *E. coli* dataset, Eduardo Rocha for providing insightful comments about the analysis of the *E. coli* dataset and the manuscript, Julien Y Dutheil for critical comments at early stage of the project. The authors thank the reviewers and editors for their constructive feedback. AB and GA thank the ANR for funding (ANR grant TempoMut ANR-12-JSV7-0007). AB, AL and GA thank the *Center for Interdisciplinary Research in Biology* for funding. SSA was funded by an European Research Council GRANT (EVOMOBILOME 281605).

References

Abby, S. S., B. Néron, H. Ménager, M. Touchon, and E. P. Rocha. 2014. Macsyfinder: a program to mine genomes for molecular systems with an application to CRISPR-Cas systems. PloS One 9:e110726.

Abby, S. S. and E. P. Rocha. 2012. The non-flagellar type III secretion system evolved from the bacterial flagellum and diversified into host-cell adapted systems. PLoS Genetics 8:e1002983.

Barker, D. and M. Pagel. 2005. Predicting functional gene links from phylogenetic-statistical analyses of whole genomes. PLoS Comput Biol 1:e3.

Benson, D. A., M. Cavanaugh, K. Clark, I. Karsch-Mizrachi, D. J. Lipman, J. Ostell, and E. W. Sayers. 2013. Genbank. Nucleic Acids Res 41:D36–42.

Carbone, A. and L. Dib. 2009. Theory and Applications of Models of Computation chap. Co-Evolution and Information Signals in Biological Sequences. Springer.

Dimmic, M. W., M. J. Hubisz, C. D. Bustamante, and R. Nielsen. 2005. Detecting coevolving amino acid sites using Bayesian mutational mapping. Bioinformatics 21 Suppl 1:i126–35.

Dobzhansky, T. 1934. Studies on hybrid sterility. Zeitschrift für Zellforschung und Mikroskopische Anatomie 21:169–223.

Dutheil, J. and N. Galtier. 2007. Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol 7:242.

Dutheil, J., T. Pupko, A. Jean-Marie, and N. Galtier. 2005. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22:1919–28.

Edgar, R. C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7.

Felsenstein, J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1–15.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Inc.

Hayashi, F., K. D. Smith, A. Ozinsky, T. R. Hawn, E. C. Yi, D. R. Goodlett, J. K. Eng, S. Akira, D. M. Underhill, and A. Aderem. 2001. The innate immune response to bacterial flagellin is mediated by toll-like receptor 5. Nature 410:1099–1103.

Kondrashov, A. S., S. Sunyaev, and F. A. Kondrashov. 2002. Dobzhansky-muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 99:14878–83.

Kryazhimskiy, S., J. Dushoff, G. A. Bazykin, and J. B. Plotkin. 2011. Prevalence of epistasis in the evolution of influenza a surface proteins. PloS Genetics 7(2):e1001301.

Kulathinal, R. J., B. R. Bettencourt, and D. L. Hartl. 2004. Compensated deleterious mutations in insect genomes. Science 306:1553–4.

Lartillot, N. and R. Poujol. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol. Biol. Evol. 28:729–44.

Lockless, S. W. and R. Ranganathan. 1999. Evolutionarily conserved pathways of energetic connectivity in protein families. Science 286:295–9.

Martin, L. C., G. B. Gloor, S. D. Dunn, and L. M. Wahl. 2005. Using information theory to search for co-evolving residues in proteins. Bioinformatics 21:4116–24.

Martínez-García, E., P. I. Nikel, M. Chavarría, and V. Lorenzo. 2014. The metabolic cost of flagellar motion in pseudomonas putida kt2440. Environmental microbiology 16:291–303.

Pagel, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Systematic Biology 48:612–622.

Pagel, M. 2000. Phylogenetic–evolutionary approaches to bioinformatics. Brief Bioinform 1:117–30.

Pagel, M. and A. Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. Am Nat 167:808–25.

Phillips, P. C. 2008. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855–67.

Poelwijk, F. J., D. J. Kiviet, D. M. Weinreich, and S. J. Tans. 2007. Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445:383–6.

Pupo, G. M., R. Lan, and P. R. Reeves. 2000. Multiple independent origins of *Shigella* clones of *Escherichia coli* and convergent evolution of many of their characteristics. Proceedings of the National Academy of Sciences 97:10567–10572.

Stamatakis, A. 2006. Raxml-vi-hpc: maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. Bioinformatics 22:2688–90.

Stelling, J., S. Klamt, K. Bettenbrock, S. Schuster, and E. D. Gilles. 2002. Metabolic network structure determines key aspects of functionality and regulation. Nature 420:190–3.

Tillier, E. R. M. and R. A. Collins. 1995. Neighbor joining and maximum likelihood with RNA sequences: Adressing the interdependence of sites. Mol Biol Evol 12:7–15.

Touchon, M., C. Hoede, O. Tenaillon, V. Barbe, S. Baeriswyl, P. Bidet, E. Bingen, S. Bonacorsi, C. Bouchier, O. Bouvet, et al. 2009. Organised genome dynamics in the *Escherichia coli* species results in highly diverse adaptive paths. PLoS Genetics 5:e1000344.

Weinreich, D. M. and L. Chao. 2005. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. Evolution 59:1175–82.

Weinreich, D. M., R. A. Watson, and L. Chao. 2005. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59:1165–74.

Welch, J. J. 2004. Accumulating Dobzhansky-Muller incompatibilities: reconciling theory and data. Evolution 58:1145–56.

Yabuuchi, E. 2002. Bacillus dysentericus (sic) 1897 was the first taxonomic rather than bacillus dysenteriae 1898. Int J Syst Evol Microbiol 52:1041.

	DELTRAN			ACCTRAN		
Type of pairs	Count	$Z_M^{(e_1,n_2)}$	p-value	Count	$Z_M^{(e_1,n_2)}$	p-value
Genealogically ordered pairs ($M = S$)	2	3.66	0.021	1	1.57	0.223
Inseparable pairs ($M = Id$)	2	2.96	0.039	3	4.74	0.0024
Both ($M = S + Id$)	4	4.82	0.00047	4	4.81	0.00047

Table 1: Testing for the independence between the loss of motility and the loss of flagella in *E. coli* for both parsimony reconstruction DELTRAN (no reversion for the loss of flagellum) and ACCTRAN

.

Legends of figures

Figure 1 A tree to illustrate the labeling of the branches, the genealogically ordered and inseparable pairs between events E_1 (plain circles) and events E_2 (plain squares). We successively label the branches of T through a so called "depth-first traversal" of the tree (*i.e.,* a recursive root-first traversal, exploring as far as possible along each lineage before backtracking). This numbering ensures that each branch has a smaller label than all subsequent branches in the tree. We also represent the matrix A of adjacencies between the branches and the vector for both events.

Figure 2 Power to reject the independence model (H_0), with a risk of 5%, as a function of the induction factor. The power is assessed with varying parameters. (a) sensitivity to the choice of the M matrix (b) impact of the nature of interaction between both processes (c) influence of topology from a complete balanced tree (binary) to a completely unbalanced tree (caterpillar) and finally (d) a second asymmetric occurrence rate for E_2.

Figure 3 Phylogenetic maximum likehihood tree for 67 strains labeled *E. coli* and *Shigellas*. The tree was rooted using one strain of *E. fergusonii* as an outgroup (not represented). Two events are placed on the tree by parsimony (either using DELTRAN in black or ACCTRAN in grey): circles represent the intracellular transition and squares represent the loss of flagellum (or gain for ACCTRAN). A and B denote the clades that were shrunk into a single lineage to assess for the impact of empty clades.

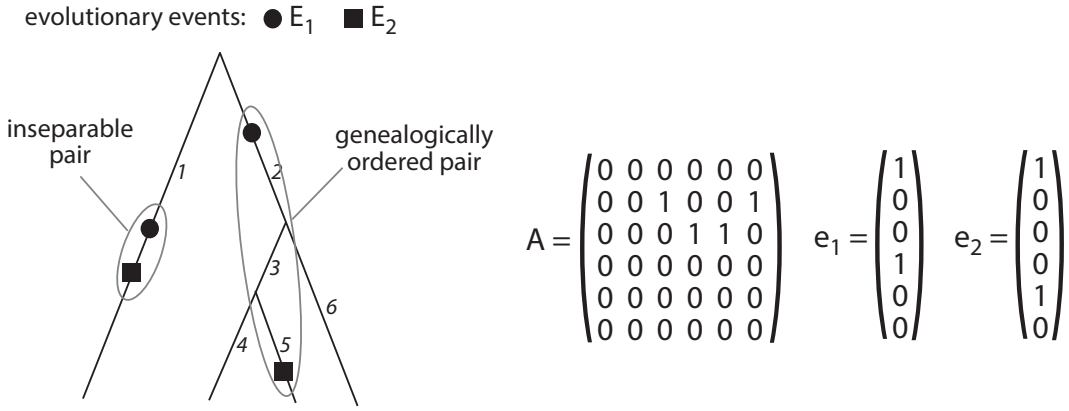


Figure 1: Tree numbering of its branches, genealogically ordered and inseparable pairs, the adjacency matrix A and the two vectors of events

.

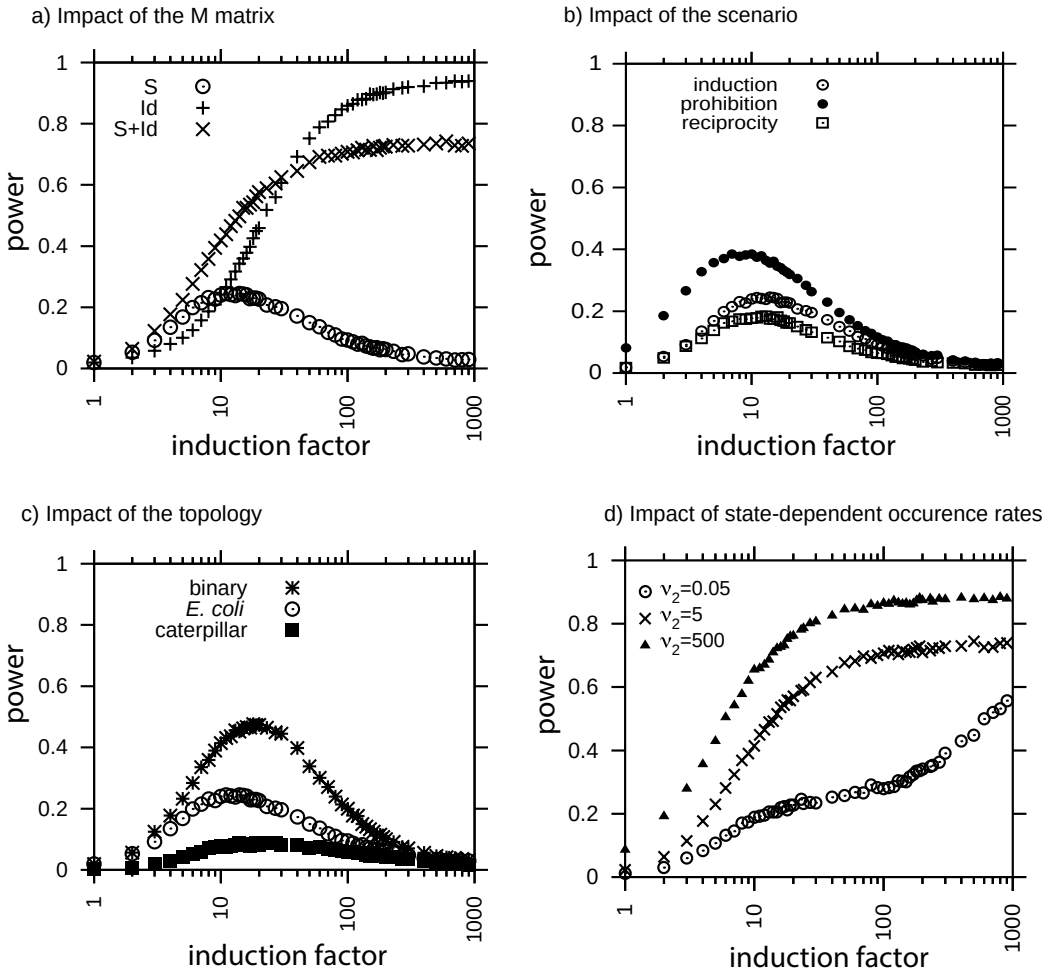


Figure 2: Power curves obtained by simulation

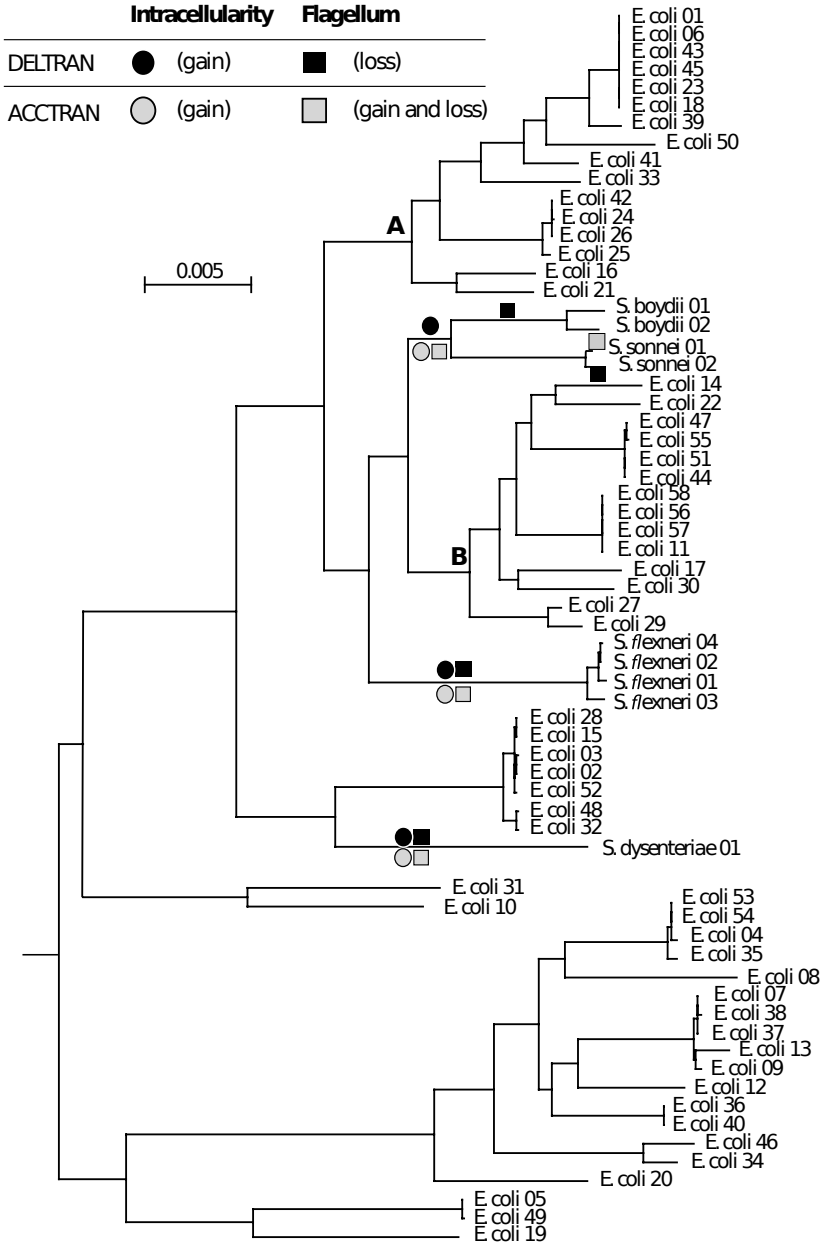


Figure 3: Phylogenetic tree for 67 strains labelled *Escherichia* and *Shigella*.

Chapitre 6

Méthode paramétrique

6.1 Introduction et résumé des résultats

Ce deuxième article est dans la continuité du premier, puisqu'il reprend le modèle de coévolution utilisé dans les simulations que nous y avons conduites. Nous nous plaçons donc dans un même contexte : un arbre phylogénétique associé à deux évènements évolutifs, dont nous connaissons les positions des occurrences sur cet arbre.

Précisément, le modèle de coévolution que nous avons développé décrit pour chaque évènement évolutif un processus de Poisson régissant la répartition de ses occurrences le long des lignées de la phylogénie concernée. Ce processus admet deux taux que nous appelons respectivement *taux basal* et *taux induit*. Deux règles régissent le basculement d'un taux à l'autre le long des lignées : (i) un processus bascule sur son taux induit après une occurrence de l'autre évènement et (ii) un processus bascule sur son taux basal après une occurrence de l'évènement qui le concerne. Nous appelons le rapport entre le taux induit et le taux basal l'*induction*, c'est à dire l'effet positif ou négatif qu'a l'autre évènement sur sa vitesse d'apparition. Ainsi, pour une paire d'évènements, nous avons 4 taux qui modélisent leur coévolution : les deux taux basaux et les deux taux induits.

Nous avons calculé la fonction de vraisemblance associée à ce modèle, pour un arbre et les positions sur cet arbre des occurrences de deux évènements donnés. Cette fonction renvoie donc, pour quatre paramètres fixés, la probabilité des données, c'est à dire la probabilité de la co-répartition observée des évènements sur l'arbre en question, conditionnée à la valeur de ces paramètres. En maximisant cette vraisemblance, nous calculons les quatre paramètres optimaux, donc les paramètres expliquant au mieux notre jeu de données. Ils nous permettent

ainsi de décrire le scénario le plus vraisemblance, *i.e.* le scénario optimal expliquant la co-répartition observée dans les données traitées. Le maximum de vraisemblance n'était pas toujours informatif, nous avons développé une méthode de MCMC permettant d'échantillonner cette fonction de vraisemblance dans l'espace des valeurs que peuvent prendre les 4 paramètres.

Nous avons montré sur des données simulées que notre méthode est efficace pour inférer des taux de mutation, ainsi que pour inférer le scénario selon lequel les données ont été générées. Nous avons par ailleurs testé notre méthode sur des données biologiques, en retrouvant avec succès les paires en interaction Watson-Crick dans l'ARNr 16S d'un échantillon de γ-entérobactéries. Par ailleurs, nous avons commencé à construire l'ébauche d'un *benchmark*, en d'autres terme, d'un jeu de donnée neutre nous permettant de comparer différentes méthodes de détection de la coévolution (qu'elles s'appuient sur des phylogénies ou non). Nous avons testé quelques méthodes ainsi que celle décrite dans cet article dans le but de les comparer.

Cet article est une version préliminaire. Il sera complété et soumis à Systematic Biology *avant la date de la soutenance. Nous avons prévu d'ici là d'étudier plus avant certains aspect de la méthode décrite dans cet article, en particulier l'effet des taux de mutation sur ses performances. En effet, nous avons remarqué lors de nos tests sur les données simulées que la puissance de la méthode dépend des taux des processus de Poisson impliqués.*

Les courbes de puissance que nous présentons dans cet article montrent que la puissance de la méthode LTR, pour un scénario d'induction, atteint un plateau à environ 65%. *Or, cette courbe correspond aux taux basaux* $\mu_1 = \mu_2 = 5$ *qui semblent amputer la puissance de la méthode. Nous avons observé par exemple que pour des taux très asymétriques, la méthode gagne largement en puissance et plafonne à* 95% *des cas détectés. Nous allons donc étudier ce phénomène avant de compléter cet article, qui y gagnera en exhaustivité.*

Par ailleurs, la fonction de vraisemblance explicitée en annexe est présentée sous sa forme intégrale. La version simplifiée de cette fonction, telle qu'elle a été implémentée sera présente dans la version finale de l'article. Nous y intégrerons en outre les résultats liés à la méthode de MCMC.

6.2 Article 2

Assessing correlated evolution by a likelihood approach

Abdelkader Behdenna[1,2,3], Amaury Lambert[4,3], Guillaume Achaz[1,2,3]

[1]UMR 7138 - Institut de Biologie Paris-Seine , UPMC, Paris
[2]Atelier de BioInformatique, UPMC, Paris
[3]SMILE - Centre Interdisciplinaire de Recherche en Biologie, Collège de France, Paris
[4]Laboratoire de Probabilités et Modèles Aléatoires, UPMC, Paris
Email corresponding author: kaderbehdenna@gmail.com

Abstract

Evolutionary history is not distinct for each living being, but rather the product of shared histories. Those shared mechanisms, which are part of what we call here *correlated evolution* are key to a better understanding of evolution. A wide range of methods have already been proposed to detect correlated evolution. Those methods often require significant computing time, since they depend on the estimation of many parameters. Here we propose a method to describe the association of two evolutionary processes on a phylogenetic tree. The method is applicable to any discrete character, like substitutions within sequence or gains/losses of a biological function. Our method uses a very simple 4-parameter model of correlated evolution between two evolutionary processes, allowing us to calculate the likelihood of a co-distribution of two types of evolutionary events on a known tree, conditioned on the choice of parameters. Maximum likelihood estimates of these parameters correspond to the most likely scenario leading to the observed co-distribution. The strengths and weaknesses of the method, as well as its relative power toward other existing methods, are assessed via simulation. The method is also applied to explore interactions among rRNA 16S nucleotides in a sample of γ-enterobacteria.

1 Introduction

Observing a current photograph of Nature is not sufficient to understand evolutionary processes, since they turn out to be greatly intricated at all level of organization. One change in any part of a living system may have consequences on the evolution of other parts. The three-dimensional structure and composition of amino acids and nucleic acids imply for example steric effects or physico-chemical properties which create constraints at the molecular level (Shindyalov et al., 1994). Functional constraints are also widely present within metabolism or regulation networks (Fraser et al., 2004), without any necessary physical contact. At an event greater scale, ecological interactions can also interconnect the evolution of living organisms, or species.

In this article, we use the term *evolutionary process* in the sense of "process leading to the distribution of discrete evolutionary events on a phylogenetic tree". Those *discrete events* can be mutations, for example, at the molecular level, but at other scales they may be a change, a gain, or a loss of a biological function or attribute. We derive from this definition the notion of *correlated evolution* referring to the non-independence of two such processes. We describe here a likelihood-based method to (i) test for independence between evolutionary processes, and (ii) describe the most likely scenario leading to the correlated evolution of different agents, at different scales. As we study the co-distribution of two types of evolutionary events on one fixed phylogenetic tree, we only address correlated evolution between entities sharing the same history.

Epistasis is the influence of genetic background on fitness change due to mutation. A mutation may in fact be deleterious in certain contexts, while it may be beneficial in others (Phillips, 2008). In other words, a combination of mutations may have a different effect than the sum of the individual effects of each. This is well illustrated by the case of the Compensated Pathogenic Deviations (CPDs), that are instances of Dobzhansky-Muller incompatibilities (Dobzhansky, 1934; Welch, 2004). They correspond to alleles that are highly deleterious in some species but can be fixed in others. In these species, this deleterious effect seems to be compensated by one or ore other mutations, called compensatory. Depending on the importance of the deleterious effect, we can assume that these mutations (the disease-associated one and the compensatory one) are under more or less strong epistasis. This phenomenon is not so rare, since in a set of 32 selected proteins involved in human diseases, Kondrashov et al. (2002) report that around 10% of the deleterious amino acid substitutions are wild-type alleles in related vertebrates.

In this example, we can imagine different mutational histories that may have led to this situation. Indeed, the mere fact of trying to determine an order between mutations leads us to make strong assumptions. If the first fixed mutation is the deleterious one, then the fixation of the compensatory one must follow very quickly. In another possible scenario where the compensatory mutation is fixed first, then the deleterious one is "authorized" after a possibly long time lag. Thus we can define

the concept of *induction*, which describes the intensity with which a mutation (and more generally an evolutionary event) may favor another one. In the first scenario, we may observe a stronger induction than in the second case.

Based on different approaches, several methods to detect correlated evolution have been proposed in the literature. A first approach is based on the direct study of multiple sequence alignments (MSAs). Considering each column of the MSA (*i.e.* each site of the alignment) as a source of information, it is possible to estimate the amount of information produced by this source using entropy, the most commonly used being Shannon entropy. This principle can be extended by studying the information shared by two columns of the MSA through the joint entropy of these columns. Finally, the entropy of each column and the joint entropy is used to establish the value of the so-called mutual information, which is a measure of the statistical dependency between two columns. Mutual information can be used to identify coevolving positions (Martin et al. (2005), and see Carbone and Dib (2009) for a review on different approaches using information theory). Nevertheless, this approach has several limitations reported in the literature. In particular, the mutual information can be overestimated for highly variable columns of the MSA and the size of the sample must be sufficiently large for the estimation of the entropy to be robust and to minimize the statistical noise. The third major limitation is that the shared information also comes from the shared phylogenetic history, which can create false positives if this shared history is not taken into account.

Thus, because species partially share evolutionary history, they cannot be regarded as independent for statistical purposes (Felsenstein, 1985). Taking into account phylogenies (when possible) is essential to investigate evolutionary questions (Pagel, 2000). For example, Bayesian mutational mapping is also used, locating substitutions on a phylogenetic tree, and associating various statistical tests to evaluate the hypothesis of correlated substitutions (see Dimmic et al. (2005) for a comparison of different parametric and non-parametric test statistics). For instance, the method described by Dutheil and Galtier (2007), implemented in CoMap, maps the substitutions onto the branches of a phylogenetic tree, then translates their position, for each site and each branch, into a vector of substitution events representing the positions of the substitutions on the phylogenetic tree. The correlation between two sites is estimated by calculating the normalized inner product between the two substitution vectors of interest.

Another class of methods we describe here takes this phylogenetic information into account through continuous Markov models, estimating correlation between sites (Tillier and Collins, 1995; Pagel and Meade, 2006) using likelihood methods. Different likelihood-based approaches are described in the literature, and pinpoint the fact that these approaches are very efficient to detect and describe correlated evolution (according to a predefined model), but those are generally extremely consuming in computation time. Indeed, these methods involve exploring a likelihood landscape defined by as

many dimensions as the number of parameters of the underlying model. Considering for example a minimal model with 2 loci and 2 alleles per loci, the corresponding 4-by-4 transition matrix implies the estimation of 16 parameters. A more complex model must depend on more parameters, thus the required computation time may be greatly increased.

The method described below completes and enhances the method we described in (Behdenna et al., 2016), that efficiently detects coevolving pairs of events by translating the positions of events as well as the tree itself into a matrix formalism. Through this formalism, we computed the expectation and the variance of the pair counts under a null model. We therefore calculated an exact p-value for a given pair count to formally test for the independence between the processes. We use two different matrices Id and S that permit through this formalism to respectively count the number of *unseparable pairs* (*i.e.* pairs of events located on the same branch of the tree) and *genealogically ordered pairs* (*i.e.* pairs of events located on different branches of the same path from the root of the tree to a terminal branch). The use of the matrix $S + Id$ permits the count (and the computation of the statistics) of both unseparable pairs and genealogically ordered pairs at the same time.

We propose here a method based on a simple model of correlated evolution and the corresponding likelihood function. The method assumes that the tree is already inferred and fixed and that the events are correctly located on the tree. We seek the maximum likelihood but also sample the likelihood landscape in order to describe the scenario of correlated evolution and estimate the induction between the processes, while rejecting or not a null scenario of independence.

2 Model and scenarios

2.1 A model of interactions

The model detailed here describes the co-distribution of the occurrences of two evolutionary events E_1 and E_2 on a tree T.

We assume given the tree T on which occurrences of E_1 and E_2 events are placed according to different rules. In all cases, for an event E_i, $i \in \{1, 2\}$, its occurrences are placed on the branches of the tree according to a Poisson process with parameter m_i. The parameter m_i can change as a function of the relative positions of the events on the tree. Note that m_i is the intensity of the process governing the distribution of the occurrences of the events of type E_i on the tree.

For each event E_i, its occurrence rate m_i can take two values μ_i and μ_i^*, respectively called the basal and the enhanced occurrence rates. The basal rate can be seen as a "natural" rate of occurrence while the enhanced one represents the influence of E_2 on the occurrence rate of E_1. The ratio $\lambda_\mathrm{i} = \mu_i^*/\mu_i$ can be thought of as an induction factor which measures the influence of an event on another. Note that there is no order requirement between rates: if $\lambda_\mathrm{i} > 1$ the induction is positive whereas it is negative when $\lambda_\mathrm{i} < 1$. We assume that the interaction only concerns the next "induced" event, as it models the impact of an E_1 event on the rate of occurrence of the next E_2 event. Framed in terms of compensatory events, the second event can be seen as compensating for the need created by the first one.

All simulations use the following rules:

1. When an E_i event occurs, m_i switches to the basal rate μ_i, regardless of its current rate.
2. When an E_i event occurs at its *basal* rate μ_i, m_j $(i \neq j)$ switches to the enhanced rate μ_j^*.
3. The rates at the root are drawn within $\{(\mu_1, \mu_2), (\mu_1, \mu_2^*), (\mu_1^*, \mu_2)\}$ with respective probabilities $(\frac{\mu_1^*\mu_2^*}{N}, \frac{\mu_1\mu_1^*}{N}, \frac{\mu_2\mu_2^*}{N})$, where $N = \mu_1\mu_1^* + \mu_2\mu_2^* + \mu_1^*\mu_2^*$ is a normalizing coefficient. It corresponds to the equilibrium for the following rate matrix Q corresponding to our model.

$$\begin{pmatrix} -(\mu_1 + \mu_2) & \mu_1 & \mu_2 \\ \mu_2^* & -\mu_2^* & 0 \\ \mu_1^* & 0 & -\mu_1^* \end{pmatrix}$$

This framework can be used to simulate many different scenarios, four of which are particularly relevant from the biological standpoint and are detailed hereafter.

2.2 Simulating different scenarios

A scenario of independence. This corresponds to the H_0 model. When the processes are independent their intensities are constant, regardless of the position of the occurrences of each type of event on the tree. It therefore corresponds to:

$$\begin{aligned} \mu_1 &= \mu_1^* \\ \mu_2 &= \mu_2^* \end{aligned}$$

The intensities of both processes are constant.

A scenario of prohibition. This scenario models the case where E_2 events are not allowed unless they follow an E_1 event (E_1 triggers the appearance of E_2): for example, a deleterious mutation cannot occur without being compensated upstream. It echoes one of the two possible scenarios that explain Compensatory Pathogenic Deviations (CPDs, highly deleterious alleles compensated by other mutations).

$$\begin{aligned} \mu_1 &= \mu_1^* \\ \mu_2 &= 0 \\ \mu_2^* &> 0 \end{aligned}$$

A scenario of induction. E_1 events have a constant rate while the rate of E_2 is increased after an E_1 event has occurred. This scenario models, for example, a compensatory mutation that follows the fixation of a deleterious mutation and corresponds to the second hypothesis to explain how CPDs occur in genomes. More generally, it describes the process by which direct compensatory events occur.

$$\begin{aligned} \mu_1 &= \mu_1^* \\ \mu_2 &< \mu_2^* \end{aligned}$$

A scenario of reciprocity. In this last scenario, there is a symmetrical interaction between both processes. E_1 events enhance the occurrence rate of E_2 events and, reciprocally, E_2 events enhance the rate of occurrence of E_1 events. This scenario mimics for example reciprocal sign epistasis, where two mutations are beneficial only in the double mutant (Weinreich et al., 2005; Poelwijk

et al., 2007).

$$\begin{aligned} \mu_1 &< \mu_1^* \\ \mu_2 &< \mu_2^* \end{aligned}$$

3 Estimating the occurrence rates: a likelihood framework

Let E_1 and E_2 be two types of events whose occurrences are distributed on a tree T, whose topology and branch lengths are known. We compute the likelihood function as the probability of the positions of the occurrences of both events on T, under the model previously described, conditioned on a set of parameters $(\mu_1, \mu_1^*, \mu_2, \mu_2^*)$.

Calculating the likelihood of one branch

The likelihood of a branch b of the tree depends on (i) the branch length, (ii) the current substitution rates at the beginning of the branch and (iii) which event(s) occur on the branch. Considering an initial state and the possible occurrences of E_1 and/or E_2 on b, we can determine the final state of the branch, which is the initial state for its successors, if this branch is non-terminal.

Concerning the substitution rates at the beginning of the branch, we distinguish three cases:

1. Both events are at their basal rates (state 00)
2. E_1 is at its basal rate and E_2 is at its enhanced rate (state 01)
3. E_1 is at its enhanced rate and E_2 is at its basal rate (state 10)

Note that the state "E_1 and E_2 both are at their enhanced rates" is strictly forbidden under our model.

Let l be the length of b. Let $\mathcal{L}^p_{\mu_1,\mu_1^*,\mu_2,\mu_2^*}(b)$ be the likelihood for a branch b, for the set of parameters $(\mu_1, \mu_1^*, \mu_2, \mu_2^*)$ and the initial state $\Gamma \in \{00, 01, 10\}$.

The likelihood of the branch b is calculated through the four following cases : (the exact formulas are detailed in Annex A):

1. No event on the branch
2. One occurrence of E_1 on the branch
3. One occurrence of E_2 on the branch

4. One occurrence of each event on the branch. This case is separated into two sub-cases :
    - 4a. The occurrence of E_1 precedes the occurrence of E_2
    - 4b. The occurrence of E_2 precedes the occurrence of E_1
5. More than one occurrence of at least one event on the branch. This case is neglected in our work.

In the case 4 where E_1 and E_2 are represented on b, we cannot assume any order between them, *i.e.* both orders $(E_1 \to E_2)$ and $(E_2 \to E_1)$ are possible, and the final state of the branch depends on this order. In this particular case, we compute separately the likelihood of the branch for both cases $4a$ and $4b$ as well as the corresponding final cases. This way, we can separate both orders $(E_1 \to E_2)$ and $(E_2 \to E_1)$ on this branch, allowing us to transmit the final state to the immediately descending branch.

Calculating the likelihood of the whole tree

For a branch b, f_b is the likelihood for a single branch, associated with the terminal states of the occurrence rates, thus we can calculate the likelihood for a whole tree through a recursive algorithm.

Once all the possible deterministic configurations are listed, we apply on each resulting tree the following algorithm:

- the initial state on the root is either 00, 01 or 10. We calculate the likelihoods considering separately each of those states. The likelihood of the tree is the sum of those three values, weighted by the probability of each initial state at the equilibrium, respectively $\frac{\mu_1^*\mu_2^*}{N}$, $\frac{\mu_1\mu_1^*}{N}$ and $\frac{\mu_2\mu_2^*}{N}$, where $N = \mu_1\mu_1^* + \mu_2\mu_2^* + \mu_1^*\mu_2^*$ is a normalizing coefficient.

- by recursion hypothesis, we assume we know the initial state on a branch

- determine the presence/absence of the events on the concerned branch (one in 5 cases) and calculate its likelihood and the final state on the branch. If the branch is in the case 4, we then calculate separately the likelihood for the sub-cases $4a$ and $4b$ and the respective final states

- if the branch is not terminal, apply this algorithm to both its descendants, transmitting the final state. In the case 4, we transmit both final states separately to two different branches of recursion

The likelihood of the whole tree is the product of the likelihoods of its branches. Note that all the possible trees are not explicitly generated. The only non-deterministic case is case 4, since the order of two events on the same branch is not known, instead of considering a non-deterministic tree,

we separate the sub-cases $4a$ and $4b$. The recursion is sufficient to implicitly list all deterministic cases. It is equivalent to consider a forest of all the deterministic trees representing all the orders $(E_1 \to E_2)$ and $(E_2 \to E_1)$ for all the concerned branches. Note that if no branch of the tree carries both types of events, then only one tree is produced.

3.1 Estimating the maximum likelihood (ML)

The set of parameters $(\mu_1, \mu_1^*, \mu_2, \mu_2^*)$ maximizing the likelihood describes the most likely scenario leading to the observed co-distribution of the occurrences of E_1 and E_2 on the tree T under our model of correlated evolution. The quadruplet maximizing the likelihood is searched by a gradient descent (Debye, 1909) together with golden section search (Avriel and Wilde, 1966).

In practice, we start the exploration from an arbitrary point (*i.e.* a quadruplet of parameters), which is here the quadruplet $(\theta_1, \theta_1, \theta_2, \theta_2,)$ where θ_i $(i \in \{1, 2\})$ is the observed occurrence rate n_i/L with n_i the observed number of occurrences of the event of type E_i and L the total length of the tree. We then calculate the gradient from this point and follow its direction as the likelihood increases. Once a likelihood peak is passed (*i.e.* at the first decreasing step), we situate the peak by a binary search, weighted by the golden ratio, ensuring a constant-speed convergence. We then repeat the process from this new point until the gain in likelihood is negligible (*e.g.* under an arbitrarily small threshold).

Scenario selection

Maximizing the likelihood is also possible while constraining some of the parameters: for example, setting $\mu_1 = \mu_1^*$ is equivalent to assuming that the occurrence rate of E_1 is not influenced by the occurrences of E_2. By constraining one or more parameters, we can define nested models, where one of them is a particular case of the other. Comparing those models amounts to comparing a reference model to a reduced model.

We compare nested models through a likelihood-ratio test (Neyman and Pearson, 1933). Considering two nested models, twice the difference of their log-likelihood follows approximately a χ^2 distribution with degrees of freedom equal to the difference of the degrees of freedom of each model. We call the model with the most degrees of freedom the $reference$ model and the other one the *null* model. Then by comparing the difference of the log-likelihoods for each model to the corresponding value in a χ^2 distribution table, we can reject or not the null model.

Within the specifications of our model, we can define different scenarios, some of them being nested. We built a nesting tree (Figure 1) that we can explore to infer the best fitting scenario.

Testing for independence In particular, the independence between E_1 and E_2, which is here the null model, is represented by the constraints $\mu_1^* = \mu_1$ and $\mu_2^* = \mu_2$. This model has 2 degrees of freedom, while the general model has 4 degrees of freedom. Thus $D = 2(\ln(\mathcal{L}_{\mu_1,\mu_1^*,\mu_2,\mu_2^*}(T)) - \ln(\mathcal{L}_{\mu_1,\mu_1,\mu_2,\mu_2}(T)))$ follows approximately a χ^2 distribution with 2 degrees of freedom. By comparing D to a selected value in a χ^2 distribution table, we can reject or not the null model, *i.e.* statistically reject or not the independence hypothesis between E_1 and E_2.

3.2 Markov Chain Monte Carlo (MCMC)

In some cases, the likelihood landscape may be flat around its maximum. In this case, the ML loses some of its informational interest, since different sets of parameters may return likelihoods only very slightly inferior to the maximum. For example, under our model of correlated evolution, if $\mu_1 = 0$, then E_2 never switches to its enhanced rate, thus the parameter μ_2^* has no influence on the computed likelihood. In this case, the likelihood landscape will be constant when μ_2^* varies, which is not observable when only studying the ML.

To complete the study of the ML, we implemented a Markov Chain Monte Carlo method (Metropolis and Ulam, 1949), allowing us to sample the likelihood landscape through a random walk generated by a Metropolis-Hastings algorithm (Metropolis et al., 1953). The equilibrium distribution of this sampling is the same as the likelihood function distribution, giving us an overview of the likelihood landscape. Of course, the longer the random walk, the better the estimation of the distribution.

Implementation. All the simulations and analyses used programs developed by the authors in the Ocaml language (version 3.09). The source codes are available at www.-.com.

4 Results on simulated data

We first assessed the power of the method on simulated data. Those simulations give us an overview of the theoretical power and the limitations of the method. Through different controlled scenarios described previously, we evaluate to what extent the method correctly estimates the occurrence rates as well as its power depending on the tree topology and the nature of the interactions between the evolutionary processes. Finally, we test the ability of the method to correctly identify the scenario selected in the simulations.

Note that we frequently use the term "specialized model" in the following section. These models correspond to the "$E_1 \rightarrow E_2$ induction model" and the "both inductions are equal" model, when we respectively simulate data under an induction scenario, or a reciprocal one. Both have 3 *degrees of liberty and are generally tested against the independence model by a likelihood-ratio test (LRT).*

4.1 Estimating the occurrence rates

We first tested the capacity of the method to retrieve the occurrence rates that have been arbitrary selected for a series of simulation. We simulated $10,000$ scenarios where $\mu_1 = 20$, $\mu_2 = 4$, $\mu_1^* = 40$ and $\mu_2^* = 80$ through the model we described previously. Each simulation generated a codistribution of the occurrences of the events E_1 and E_2, on which we used the maximum likelihood method to estimate the 4 optimal rates.

Figure 2 shows the distribution of the rates estimated by this method. We observe that μ_1 and μ_1^* are correctly estimated: their corresponding distributions are centered on the theoretical values and the dispersion around these values is low. The ML method is able to effectively estimate these two rates. The dispersion we observe may be due to the uncertainty of the conducted simulations: since the events are distributed according to Poisson processes, in some cases we can get more or less occurrences than the average corresponding to the theoretical rates.

The distribution corresponding to μ_2 is located on the right order of magnitude, but there is no well-marked peak and the dispersion is greater. This may be due to the fact that $\mu_2 = 4$ is relatively low, thus the method is not able to infer its value. μ_2^* is under-estimated, though the order of magnitude is also correct, while the corresponding distribution is as well associated to a low dispersion (*i.e.* a marked peak). This may be a consequence of the poor estimate of μ_2 which affects the accuracy to which μ_2^* is itself estimated.

4.2 Power

For a series of simulations, we assess the power of the method by counting the fraction of runs where the p-value computed under H_0, is 5% or less. Unless specified otherwise, we ran simulations with the occurrence rates $\mu_1 = \mu_2 = 5$, the tree we reconstructed for the analysis of the γ-enterobacteria biological data (see section 5 below) and 500 replicates. We tested the capacity of the method to reject a null model of independence ($\mu_1 = \mu_1^*$ and $\mu_2 = \mu_2^*$), for different reference models :

- A model of induction:

$$\begin{aligned} \mu_1 &= 5 \\ \mu_2 &= 5 \\ \mu_1 &= \mu_1^* \end{aligned}$$

- A model of symmetrical induction:

$$\begin{aligned} \mu_1 &= 5 \\ \mu_2 &= 5 \\ \lambda_1 &= \lambda_2 \end{aligned}$$

To do that, we calculate the maximum likelihood for the corresponding models, and compare it to the independence model through a likelihood-ratio test. The power is then the proportion of the replicates that reject significantly the null model of independence.

In both cases, we compare the results with a likelihood-ratio test between the same null model and the reference model where all rates are free. We also compare the results with the power of the non-parametric method, with the matrix $S + Id$ that allows the count (and the computing of the *p-values* under a model of independence) of both unseparable pairs and genealogically ordered pairs at the same time.

Induction scenario The power to detect an induction scenario is broadly similar for all three methods tested (Figure 3a). The more induction, the more power, and the methods reach a plateau (corresponding to a power of 70% for the specialized model to 80% for the non-parametric method) for inductions greater than 100. Although we may anticipate that the specialized model (Induction) has more power, the generalist one (All rates are free) has surprisingly the same overall power, as well as the non-parametric method. For moderate induction ($\lambda < 11$) we can observe a slight advantage for the specialized model, while the non-parametric method becomes better than the

other two for greater induction ($\lambda > 11$).

Symmetrical scenario The comparison between the methods for a scenario of reciprocity is more discriminating. Again, as anticipated, the more induction, the more power. All three method reach a plateau of 100% of the cases detected, but the main difference is the speed at which this plateau is reached. The power of the specialized method increases much more rapidly than the other ones and the plateau is reached for moderate induction around $\lambda \approx 10$. The generalized ML method is also very effective, and reaches the plateau for $\lambda \approx 12$. Finally, the non-parametric method is slower to reach its maximum power, for $\lambda \approx 120$. This method requires a very strong signal of coevolution to detect efficiently reciprocal interactions while the ML method can discriminate the cases of reciprocity, even for low inductions.

4.3 Inference of the scenario

To complete the previous point, we finally tested the capacity of the method to correctly retrieve the simulated scenario, for different intensities of induction. To do this, we explore the scenarios tree described in Figure 1, representing the nesting of different models, from the most constraint one ($\mu_1 = \mu_2 = \mu_1^* = \mu_2^*$) to the one with 4 free parameters. We advance in the tree following its branches as long as the gain in likelihood from a scenario to another is significant. When this is not the case, the scenario where we stopped is selected. Since some models are not nested into each other (*e.g.* induction $E_1 \to E_2$ and induction $E_2 \to E_1$), several scenarios can be selected at once.

We tested the inference of the scenario on simulated data on the tree we reconstructed for the analysis of the γ-enterobacteria biological data (see section 5 below), for different values of induction (500 replicates for each), for both following scenarios:

- A scenario of induction:

$$\begin{aligned} \mu_1 &= 20 \\ \mu_2 &= 4 \\ \mu_1 &= \mu_1^* \end{aligned}$$

- A scenario of symmetrical induction:

$$\begin{aligned} \mu_1 &= 20 \\ \mu_2 &= 4 \\ \lambda_1 &= \lambda_2 \end{aligned}$$

The results are presented in Figure 4:

- On the one hand we calculated the proportion of cases detected for each model. Since more than one scenario can be selected at once, we count 1 if the scenario is the only one selected, 1/2 if another one is selected, 1/3 if two other scenarios are selected, and so on. In other words, for each replicate, we weight each scenario selection by the total number of selected scenarios (Figure 4 a1 and a1).
- On the other hand we calculated, for each scenario, the power of the corresponding specialized method to accurately detect it (Figure 4, a2 and b2).

Induction scenario (Figure 4a) We observe that, for a given induction, several scenarios coexist. For low inductions ($\lambda < 8$), the 1-parameter model and the independence model are the most recurrent, which is expected due to the lack of statistical signal to capture (low inductions correspond to the independence between the processes). For moderate induction between $\lambda = 10$ and $\lambda = 100$, the induction scenario $E_1 \to E_2$ is the most present: the method seems to detect the accurate scenario in more than half of the cases. Finally, when the induction increases ($\lambda > 100$), the majority of interactions are unseparable pairs (events located on the same branch), thus it becomes impossible to chronologically order the interactions between the events. Both reciprocal models ("events are equivalent" and "inductions are equal") as well as both induction models coexist and the scenario selection is not able to discriminate between those scenarios anymore. When studying the corresponding power curve, we observe the same pattern: the power of the specialized method increases with the induction until a peak is reached: 90% of the cases are detected when $\lambda \approx 20$. Then the unseparable pairs take precedence to the genealogically ordered pairs, and a chronological order between the events becomes difficult, or in certain cases even impossible, to discriminate, which explains the loss of power as the induction continues to increase.

Symmetrical scenario (Figure 4b) We observe a similar pattern for the reciprocal scenario. Here, the $E_1 \to E_2$ scenario and the "inductions are equal" scenario coexist for almost all values of induction (except for the lowest values). We expect an advantage for the reciprocal scenario, but the induction model is also selected very often. This may be due to the asymmetry of the

rates, $\mu_1 = 20$ being far greater than $\mu_2 = 4$. Once again, the power curve is very informative: the power of the specialized method ("inductions are equal" against "independence") increases with the induction and reaches a peak at $\lambda \approx 20$ (almost 90% of the cases detected). For greater induction, we observe a slight decrease in power. The method is more accurate for moderate inductions.

5 Strong interactions among rRNA 16S nucleotides

The ribosomal ribonucleic acid (rRNA) is the principal component of the ribosome. rRNA sequences were present in LUCA (Last Universal Common Ancestor), are well conserved and therefore were used to show the diversity of living organisms, in particular the existence of the three kingdoms of life (Fox and Woese, 1975; Woese and Fox, 1977). The structure of RNA has been determined experimentally (Doty et al., 1959; Leontis and Westhof, 2001), and gives information in particular on its folds and on interactions between nucleotides. In particular, the single-stranded structure of rRNA is in a large part determined by strong hydrogen bonding patterns, the Watson-Crick pairs (hereafter WC pairs). As the RNA structure is strongly stabilized by stems of WC pairs, the mutation of one nucleotide involved in a WC pair is very often deleterious and is typically observed along with a second compensatory mutation on the paired nucleotide. WC pairs constitute a classical case of correlated evolution at the molecular level (Watson et al., 1953; Leontis and Westhof, 1998; Moore, 1999). As the interactions between both nucleotides of a pair are very strong, we expect a very short time lag, if any, between the two substitutions at paired bases (*i.e.* unseparable pairs).

5.1 Using the non-parametric method

Here, we aim to detect all pairs of nucleotides that show a significant correlated evolution in the RNA 16S gene in a sample of 54 gamma-enterobacteria (plus one beta-enterobacterium as an outgroup). We retrieved the sequences of 16S rRNAs from the SILVA rRNA database (Pruesse et al., 2007; Quast et al., 2013), aligned them using Muscle v3.6 (Edgar, 2004), and trimmed the poorly aligned positions (*i.e* too divergent regions) of the alignment using Gblocks v0.91b with default parameter (Castresana, 2000; Talavera and Castresana, 2007). From the resulting 1,233 nucleotides alignment, we inferred a phylogenetic tree, using Phyml (Guindon et al., 2010) under the model GTR; finally, we inferred the ancestral states by maximum likelihood with Bayestraits Multistate (Pagel et al., 2004).

At each site, we assimilated all substitutions to events regardless of the exact nature of the substitution. This implies that we will detect pairs that coevolve regardless of their state. For each pair

(i, j) of nucleotides, we calculated ${Z_{\mathrm{Id}}}^{(\mathrm{e_i},\mathrm{n_j})}$ (*i.e.* the score for unseparable pairs) and its associated p-value.

To gain insight into the performance of the method, we used the WC pairs as a positive control. The comparison between the pairs that are significantly coevolving and the documented WC pairs show that there is an overall good overlap between them (Figure 5, "Id matrix"). However, some WC pairs were not detected. At least two reasons can be invoked. First, we have trimmed the alignment with Gblocks. Out of the complete set of 477 WC pairs (0.0004 of all 1,169,685 pairs), only 349 WC pairs (0.0004 of the 780,625 unmasked pairs) were left after Gblocks. Furthermore, all sites that are monomorphic on the alignment can neither be found significantly coevolving as we look for correlated substitutions. This again reduces the dataset.

Figure 5 shows unsurprisingly that there is an important trade-off between specificity (the fraction of WC pairs among the ones we detect) and sensitivity (*i.e.* the fraction of WC pairs that are detected): we detect more interacting pairs if the threshold is less stringent (*i.e.* for greater ranks), but at the cost of accuracy. Consequently, for a stringent (rank < 10) threshold, we show that all 10 pairs are WC pairs.

5.2 Comparison to other methods

We applied four different methods (non-parametric (Id matrix), likelihood-ratio test (LRT), CoMap and Mutual Information corrected for the phylogeny (MIp)) to the same dataset in order to compare their performance. We selected the 75 most significantly coevolving pairs for each method and used the crystal structure of the 16S ribosome (Korostelev et al., 2006) to compute the distance (in Å) between each nucleotide, taking the barycenter of the base as its position in space. We then sorted them between WC pairs, non-WC pairs separated by less than 10Å and non-WC and separated by more than 10Å (Figure 5b). The LRT method is the most efficient to detect the WC pairs. As a matter of fact, almost all the best ranked pairs are WC pairs, whereas the other method detect close and distant non-WC pairs at lower ranks. MIp and Id are the most efficient methods to detect pairs that are close in space, and their rate of false positive is the lowest for those 75 selected pairs.

To further characterize the coevolving pairs, we remapped them on the secondary structure (WC stems) and on the tertiary structure (computing the average distance in Å). Results (Figure 5a) confirm that the pairs with the strongest signal of correlated evolution are the WC pairs, which distance is 6.8Å on average, for all the four methods. There are however other pairs that show a significant pattern of correlated evolution for small distances. Indeed, once all pairs of nucleotides are ranked according to their distance and assigned to bins of 2 Å, only the $4-6$Å, $6-8$Å (that

includes the WC pairs) and the $8-10$Å bins show more significantly co-evolving pairs (for a given risk of 1%) than expected (that is 1% of all pairs in the bin) for the parametric, LTR and MIp methods. This again confirms that pairs of nucleotides that are not WC pairs show a significant signal of correlated evolution. Note that CoMap is a special case since it also shows an over-representation of the significative pairs for distance lower than 10Å but is not discriminating for distance greater than 20Å, for which it shows the same pattern of over-representation.

5.3 Estimation of the occurrence rates with the likelihood method

Although most of the pairs with a strong correlated evolution signal are WC pairs, other types of pairs are also observed. We selected the 75 most significantly coevolving pairs for each method and computed the distance (in Å) between each nucleotide, taking the barycenter of the base as its position in space. Interestingly, some of the selected pairs are less than 10 Å away (between 5 and 18 pairs depending on the method). Given that the average distance is 83.8Å and that only 0.6% of all pairs have a distance smaller than 10Å, there is a very strong overrepresentation of close pairs among the coevolving pairs. We further investigated the positions of both nucleotides of the close pairs and show that they belong to WC stems. The counts of different types of pairs (Figure 6a) show that they are mostly either consecutive (stacked nucleotides) or that the interacting nucleotide is consecutive to the WC pair. This suggests that alternative interactions also constrain the evolution of the rRNA 16S. While all the method present similar results for the detection of WC pairs, MIp and Id have better performances for the detection of pairs that are close in space.

Finally, for the LTR method, we have established the distribution of the estimated induction for the 75 most significantly coevolving pairs (Figure 6b). Results show a general trend: the lowest inductions correspond to non-WC pairs and the greater ones, to WC pairs. The intermediate inductions are shared between the two types of pairs. Interestingly, some induction bins (in log, $1.6-1.8$ and $1.8-2.0$) correspond almost all to non-WC pairs. This confirms the specialization of our LTR method that is very efficient to detect strong and symmetrical signals of coevolution, typical to Watson-Crick interactions but is less accurate when the interactions are potentially less strong or asymmetrical.

6 Discussion and conclusion

We have developed a method based on a simple model of correlated evolution between two evolutionary processes that generate discrete events on a given phylogenetic tree. This model depends on four parameters - two for each process - allowing the description, in a basic way, of the *induction* of one process on an other. For a given tree and the positions of the events on it, we calculate the likelihood, *i.e.* the probability of this set of data under our model, for a given set of parameters. Maximizing this probability is equivalent to searching for an optimal set of parameters describing the processes that led to the observed distribution of the events on the tree. In certain cases, maximizing the likelihood is not informative enough, *e.g.* in the case of a plateau in the likelihood landscape. Thus we also explore this landscape through a Monte Carlo Markov Chain method, sampling it in proportion to the likelihood. Both techniques give (i) the exact maximum, corresponding to the best scenario and (ii) a general overview of the space of possible scenarios, if many of them equally describe the data. Nested models also permit the direct comparison of scenarios, in particular to assess the independence between the concerned evolutionary processes.

Using simulated controlled scenarios, we have shown that the accuracy of our method depends on the occurrence rates of the studied processes. Our method accurately infers those rates if they are large enough. In the case of lower rates, the variance becomes greater and the rates may be under-estimated. However, it correctly estimates the orders of magnitude, allowing a biological interpretation.

We also shown that the power of the method depends on the scenario of coevolution that has been chosen to generate the simulated data. In the case of an induction scenario ($\mu_1 = \mu_1^*$, $\mu_2^* > \mu_2$), the method is about as powerfull as the non-parametric method we developed in a previous article (Behdenna et al., 2016). While it has more power for moderate inductions, it becomes less effective when the induction becomes greater. In also depends on the chosen occurrence rates: if $\mu_1 = \mu_2$, *i.e.* if the basal rates are equal, it becomes more difficult to discriminate against other symmetrical scenarios. Thus the method has more power to detect asymmetrical inductions if the basal rates of the processes are initially sufficiently different. The power of our method is better for a reciprocity scenario, where both events have the same induction on each other. Then the specialized method (the "inductions are equal" model, especially designed to detect those kinds of symmetrical scenarios) has been proved to have more power, for low inductions as for greater ones.

We built a tree of scenarios, from the nesting of the different models contained in the general model we presented in this article. We then used likelihood-ratio tests to explore this tree conditionally to simulated data, testing the ability of our method to select the right model. Our method is capable of selecting the correct scenario in most of the cases, for inductions around $\lambda \approx 20$, and remains

effective for greater inductions. The statistical signal seems to low to get good performances for inductions lesser that 20. Furthermore, due to the shape of the scenarios tree, multiple model can be selected at once. For example, for two events E_1 and E_2, if they only perform unseparable pairs (they always appear in pairs on the same branches), then three scenarios may explain this co-repartition: (i) a scenario of reciprocal induction, (ii) a scenario of strong induction $E_1 \rightarrow E_2$ and (iii) a scenario of strong induction $E_2 \rightarrow E_1$. Consequently one must take great caution when interpreting the results of such a method.

Since the method is deeply grounded in the field of phylogenetic inference, some possible biases and limitations can appear, but also be avoided by being very careful when constructing the dataset. Our working hypotheses implies that the chosen events may occur wherever on the chosen phylogenetic tree, without any *a priori* on the concentration of the events on the different parts of the tree. Typically, the presence of empty clades in the tree is an information itself, to the extent that these clades are devoid of any event by chance, and not due to an *a priori* on the occurrence rates on this part of the tree. Furthermore, the method takes a known phylogenetic tree as well as the positions of the occurrences of the different types of events as an input. Consequently, the uncertainty due to the reconstruction of the phylogeny and the ancestral states has to be taken into account. Furthermore, the likelihood function the method calculates and optimizes requires the assumption of the presence of a maximum of only one occurrence of each event on each branch.

We have demonstrated the robustness of the method on biological data, the study of the strong interactions among rRNA 16S nucleotides within a sample of γ-enterobacteria. We compared it to different methods that aim to detect coevolution patterns. Each method has its pro and cons, and we have shown that our method is the most efficient for the detection of Watson-Crick pairs, that correspond to a very strong and symmetrical signal of coevolution. However, it has less power for the detection of weaker interactions, for example pairs that are spatially close without being in Watson-Crick interaction.

We can efficiently use the method in different ways, on different types of data. It can be used prospectively, to handle a large amount of data, or as a specific test, to determine if two processes distributing events on a tree significantly reject an independence model. Due to the computing time needed to apply such a method, one must keep in mind that a dataset of reasonable size is more convenient. Moreover, regardless of its nature, any discrete evolutionary event that can be placed on a phylogenetic tree can be treated, whether molecular (*e.g.* substitution) or non-molecular (*e.g.* gain or loss of a character). Combining those two approaches is another perspective and mixing molecular and non-molecular events may be useful to study for example the apparition of functions or organs in parallel with mutations on a sequence and leading to clues to determining coding regions (as in Lartillot and Poujol (2011)).

References

Avriel, M. and D. J. Wilde. 1966. Optimality proof for the symmetric fibonacci search technique. Fibonacci Quarterly 4:265–269.

Behdenna, A., J. Pothier, S. S. Abby, A. Lambert, and G. Achaz. 2016. Testing dependencies between evolutionary processes. Systematic Biology (in press) .

Carbone, A. and L. Dib. 2009. Co-evolution and Information Signals in Biological Sequences chap. Co-Evolution and Information Signals in Biological Sequences. Springer.

Castresana, J. 2000. Selection of conserved blocks from multiple alignments for their use in phylogenetic analysis. Mol Biol Evol 17:540–52.

Debye, P. 1909. Näherungsformeln für die zylinderfunktionen für große werte des arguments und unbeschränkt veränderliche werte des index. Mathematische Annalen 67:535–558.

Dimmic, M. W., M. J. Hubisz, C. D. Bustamante, and R. Nielsen. 2005. Detecting coevolving amino acid sites using bayesian mutational mapping. Bioinformatics 21 Suppl 1:i126–35.

Dobzhansky, T. 1934. Studies on hybrid sterility. Zeitschrift für Zellforschung und Mikroskopische Anatomie 21:169–223.

Doty, P., H. Boedtker, J. R. Fresco, R. Haselkorn, and M. Litt. 1959. Secondary structure in ribonucleic acids. Proc Natl Acad Sci U S A 45:482–99.

Dutheil, J. and N. Galtier. 2007. Detecting groups of coevolving positions in a molecule: a clustering approach. BMC Evol Biol 7:242.

Edgar, R. C. 2004. Muscle: multiple sequence alignment with high accuracy and high throughput. Nucleic Acids Res 32:1792–7.

Felsenstein, J. 1985. Phylogenies and the comparative method. The American Naturalist 125:1–15.

Fox, G. E. and C. R. Woese. 1975. 5s rna secondary structure. Nature 256:505–7.

Fraser, H. B., A. E. Hirsh, D. P. Wall, and M. B. Eisen. 2004. Coevolution of gene expression among interacting proteins. Proceedings of the National Academy of Sciences of the United States of America 101:9033–9038.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, and O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies: assessing the performance of phyml 3.0. Syst Biol 59:307–21.

Kondrashov, A. S., S. Sunyaev, and F. A. Kondrashov. 2002. Dobzhansky-muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 99:14878–83.

Korostelev, A., S. Trakhanov, M. Laurberg, and H. F. Noller. 2006. Crystal structure of a 70s ribosome-trna complex reveals functional interactions and rearrangements. Cell 126:1065–1077.

Lartillot, N. and R. Poujol. 2011. A phylogenetic model for investigating correlated evolution of substitution rates and continuous phenotypic characters. Mol Biol Evol 28:729–44.

Leontis, N. B. and E. Westhof. 1998. A common motif organizes the structure of multi-helix loops in 16 s and 23 s ribosomal rnas. Journal of molecular biology 283:571–583.

Leontis, N. B. and E. Westhof. 2001. Geometric nomenclature and classification of rna base pairs. RNA 7:499–512.

Martin, L. C., G. B. Gloor, S. D. Dunn, and L. M. Wahl. 2005. Using information theory to search for co-evolving residues in proteins. Bioinformatics 21:4116–24.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller. 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21:1087–1092.

Metropolis, N. and S. Ulam. 1949. The monte carlo method. Journal of the American statistical association 44:335–341.

Moore, P. B. 1999. Structural motifs in rna. Annual review of biochemistry 68:287–300.

Neyman, J. and E. Pearson. 1933. On the problems of the most efficient tests of statistical hypotheses. Philosophical Transactions of the Royal Society of London .

Pagel, M. 2000. Phylogenetic–evolutionary approaches to bioinformatics. Brief Bioinform 1:117–30.

Pagel, M. and A. Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. Am Nat 167:808–25.

Pagel, M., A. Meade, and D. Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53:673–84.

Phillips, P. C. 2008. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9:855–67.

Poelwijk, F. J., D. J. Kiviet, D. M. Weinreich, and S. J. Tans. 2007. Empirical fitness landscapes reveal accessible evolutionary paths. Nature 445:383–6.

Pruesse, E., C. Quast, K. Knittel, B. M. Fuchs, W. Ludwig, J. Peplies, and F. O. Glöckner. 2007. Silva: a comprehensive online resource for quality checked and aligned ribosomal rna sequence data compatible with arb. Nucleic Acids Res 35:7188–96.

Quast, C., E. Pruesse, P. Yilmaz, J. Gerken, T. Schweer, P. Yarza, J. Peplies, and F. O. Glöckner.

2013. The silva ribosomal rna gene database project: improved data processing and web-based tools. Nucleic Acids Res 41:D590–6.

Shindyalov, I. N., N. A. Kolchanov, and C. Sander. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations? Protein Eng 7:349–58.

Talavera, G. and J. Castresana. 2007. Improvement of phylogenies after removing divergent and ambiguously aligned blocks from protein sequence alignments. Syst Biol 56:564–77.

Tillier, E. R. M. and R. A. Collins. 1995. Neighbor joining and maximum likelihood with rna sequences: Adressing the interdependence of sites. Mol Biol Evol 12:7–15.

Watson, J. D., F. H. Crick, et al. 1953. Molecular structure of nucleic acids. Nature 171:737–738.

Weinreich, D. M., R. A. Watson, and L. Chao. 2005. Perspective: Sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59:1165–74.

Welch, J. J. 2004. Accumulating dobzhansky-muller incompatibilities: reconciling theory and data. Evolution 58:1145–56.

Woese, C. R. and G. E. Fox. 1977. Phylogenetic structure of the prokaryotic domain: the primary kingdoms. Proc Natl Acad Sci U S A 74:5088–90.

Legends of figures

Figure 1 Tree representing the nesting between different models. The most constraint scenario is on the left ($\mu_1 = \mu_2 = \mu_1^* = \mu_2^*$) and the less constraint on the right. Each arrow corresponds to the releasing of on parameter (*i.e.* the adding of one degree of freedom). From left to right, the number of degrees of freedom (*i.e.* required number of parameters to describe the corresponding model) increases.

Figure 2 Empirical distribution of the rates estimated by maximum likelihood for a scenario where $\mu_1 = 20$, $\mu_2 = 4$, $\mu_1^* = 80$ and $\mu_2^* = 40$. The theoretical values are reported on the horizontal axis.

Figure 3 Power to reject the independence model, with a risk of 5%, as a function of the induction factor. The power is assessed for two scenarios. (a) an induction scenario (b) a scenario of reciprocity. We tested three different methods: LRT with a null model against a specialized reference model (ML 1dof), LRT with a null model against a generalist reference model (ML 2dof) and the non-parametric method with the matrix $S + Id$.

Figure 4 Results of scenario selection for an induction scenario (a1 and a2) and for a scenario of reciprocity (b1 and b2), depending on the induction factor. (a1) and (b1) correspond to the distribution of the selected scenarios, each color corresponding to a different scenario, (a2) and (b2) correspond to the power to detect the accurate scenario, as a function of the induction factor.

Figure 5 Coevolving pairs of nucleotides in the 16S of γ-enterobacteria. (a) Number of significant pairs (filled box) for a risk of 1% with the expected number of false positive –1% of the pairs in each bin– (line) once we correct for the local abundance of pairs at a given distance, for four different methods : non-parametric (Id matrix), likelihood-ratio test (LRT), CoMap and Mutual Information corrected for the phylogeny (MIp). All pairs were assigned to bins of 2Å. Pairs showing significant correlated evolution are mostly located at distance $4-6$Å, $6-8$Å (this includes the WC pairs) and $8-10$Å. (b) The 75 pairs associated to the lowest p-values, ranked according to their p-values, for the four methods. The vertical axis and colors correspond to their sorting according to the nature of the pairs : Watson-Crick (black), non-WC and separated by less than 10Å (grey) and non-WC and separated by more than 10Å (light grey).

Figure 6 Coevolving pairs of nucleotides in the 16S of γ-enterobacteria. (a) Sketches to represent the 75 pairs that show the highest signal of correlated evolution. (b) distribution of the estimated inductions for the Watson-Crick pairs, all pairs were assigned to bins of size 2. We reported in each bin the significant pairs for a risk of 1% (black) and the non-significant pairs (white), for the likelihood-ratio test (LRT) method.

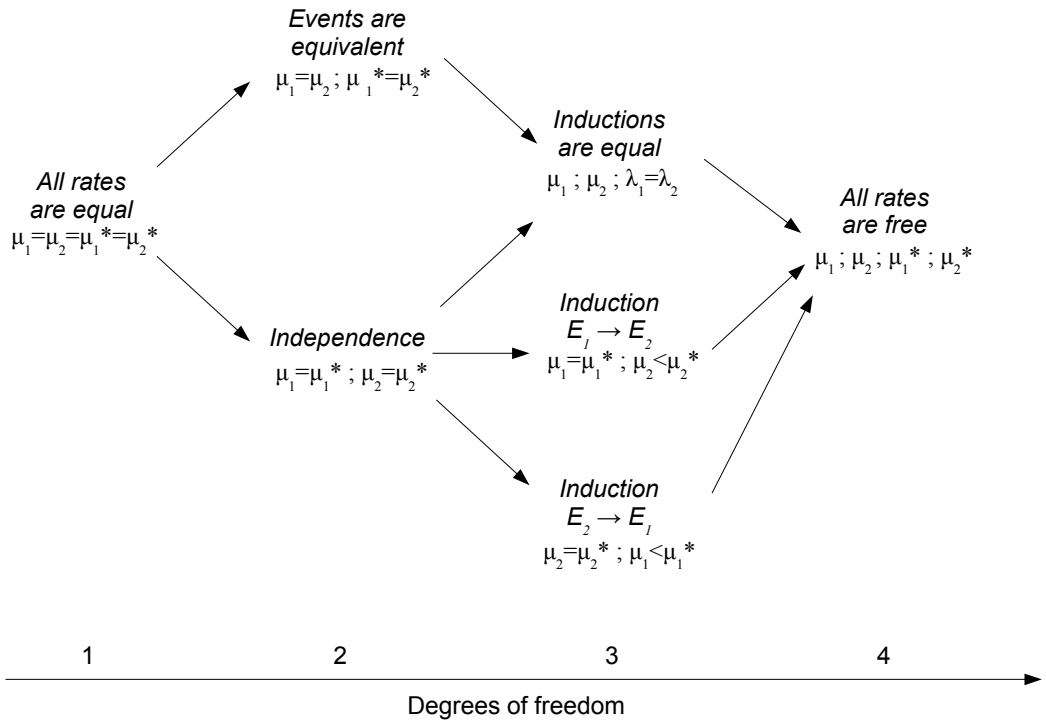


Figure 1: Nesting of different scenarios and corresponding rates

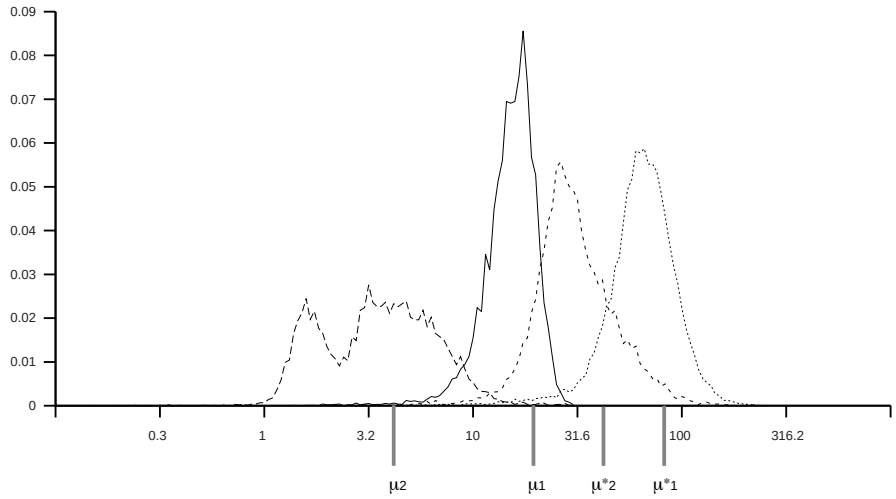


Figure 2: Empirical distribution of the rates estimated by maximum likelihood

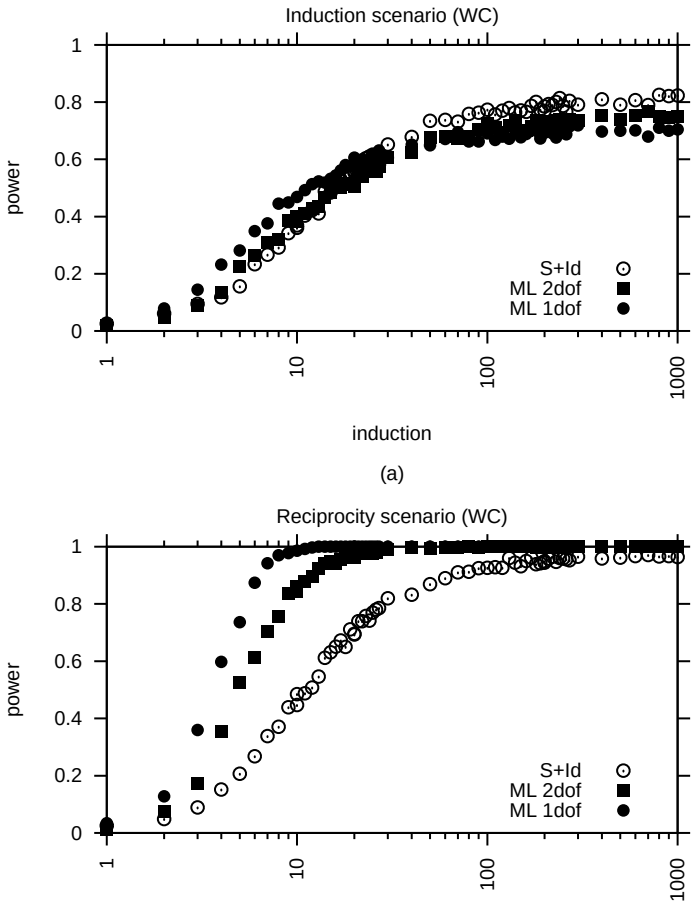


Figure 3: Power curves obtained by simulation

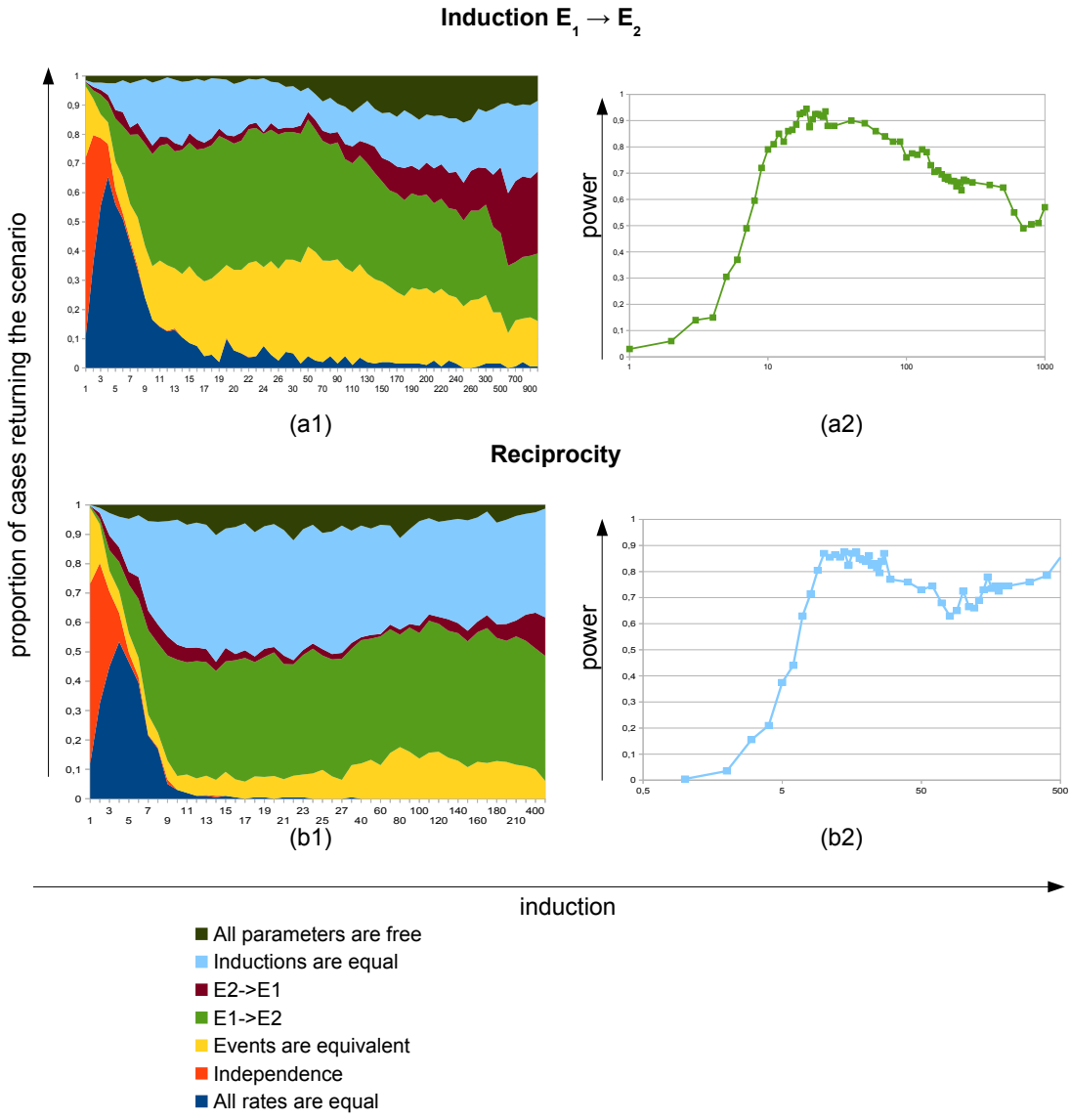


Figure 4: Scenario selection

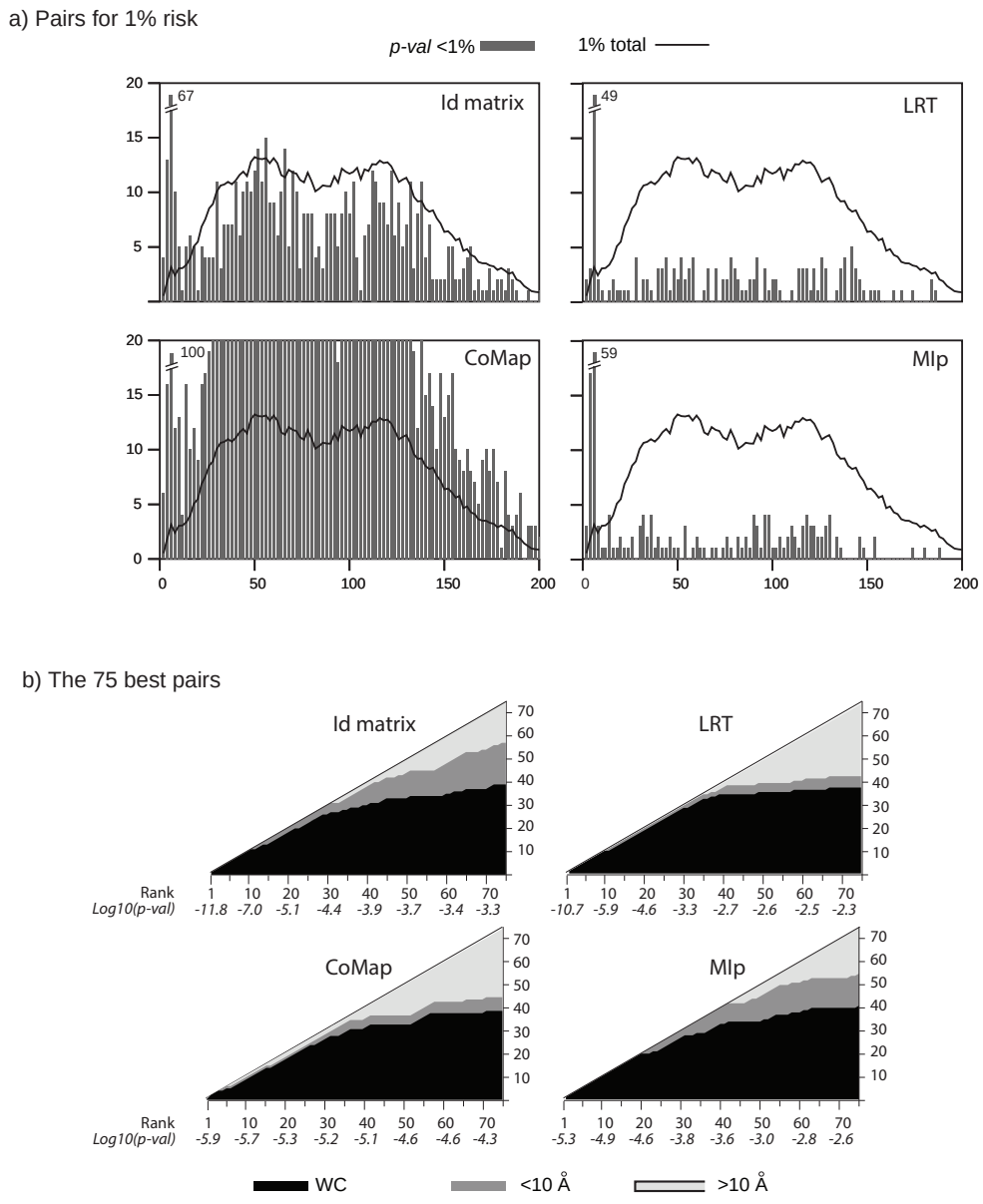
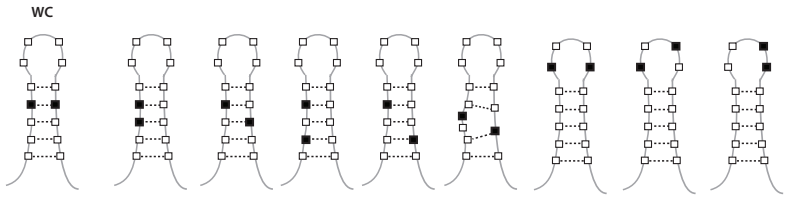


Figure 5: Coevolving pairs of nucleotides in the 16S of gamma-enterobacteria

a) Nature of co-evolving pairs



Id matrix	39	7	6	2	1	1	1	-	-	-	18
LRT	38	1	2	-	1	-	-	-	1	-	32
CoMap	39	1	1	-	1	-	1	1	1	-	30
Mlp	39	5	3	2	2	-	1	1	1	1	20

b) Induction for WC pairs

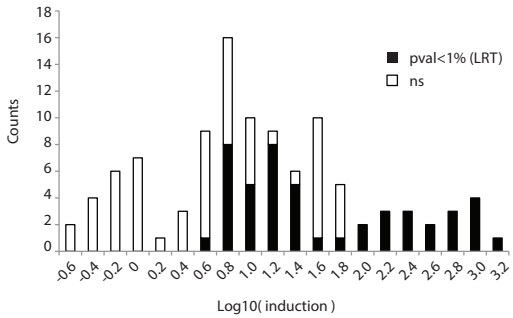


Figure 6: Replacing the coevolving pairs of nucleotides in the 16S of gamma-enterobacteria in a biological context

Annex

Likelihood function for a single branch

Let T be a tree, and b a branch of this tree, of length l. We consider two events E_i, $i \in \{1, 2\}$, of respective occurrence rates (μ_i, μ_i^*). Let $\Gamma \in \{00, 01, 10\}$ be the initial state of this branch b. The last column corresponds to the final states (FS).

We detail here the likelihood function for this single branch, the five possible cases being treated separately.

1. No event on the branch

$$f_b(\mu_1, \mu_1^*, \mu_2, \mu_2^*) = \begin{cases} e^{-l(\mu_1+\mu_2)} & \text{if } \Gamma = 00 \quad FS = 00 \\ e^{-l(\mu_1+\mu_2^*)} & \text{if } \Gamma = 10 \quad FS = 10 \\ e^{-l(\mu_1^*+\mu_2)} & \text{if } \Gamma = 01 \quad FS = 01 \end{cases}$$

2. One occurrence of E_1 on the branch

$$f_b(\mu_1, \mu_1^*, \mu_2, \mu_2^*) = \begin{cases} \mu_1 e^{-\mu_1 l} \int_0^l e^{-\mu_2 t} e^{-\mu_2^*(l-t)} dt & \text{if } \Gamma = 00 \quad FS = 01 \\ \mu_1 e^{-\mu_1 l} \int_0^l e^{-\mu_2^* t} e^{-\mu_2^*(l-t)} dt & \text{if } \Gamma = 01 \quad FS = 01 \\ \int_0^l \mu_1^* e^{-\mu_1^* t} e^{-\mu_2 t} e^{-\mu_1(l-t)} e^{-\mu_2(l-t)} dt & \text{if } \Gamma = 10 \quad FS = 00 \end{cases}$$

3. One occurrence of E_2 on the branch

Same results as in previous section, with 1 and 2 inverted.

4. One occurrence of each event on the branch

4a. E_1 **precedes** E_2

$$f_b(\mu_1,\mu_1^*,\mu_2,\mu_2^*) = \quad \int_0^l \mu_1 e^{-\mu_1 t} e^{-\mu_2 t}(\int_0^{l-t} \mu_2^* e^{-\mu_2^* x} e^{-\mu_1 x} e^{-\mu_2(l-t-x)} e^{-\mu_1(l-t-x)} dx)dt \quad \text{if } \Gamma = 00 \quad FS = 00$$

$$\int_0^l \mu_1^* e^{-\mu_1^* t} e^{-\mu_2 t}(\int_0^{l-t} \mu_2 e^{-\mu_2 x} e^{-\mu_1 x} e^{-\mu_2(l-t-x)} e^{-\mu_1^*(l-t-x)} dx)dt \quad \text{if } \Gamma = 10 \quad FS = 10$$

$$\int_0^l \mu_1 e^{-\mu_1 t} e^{-\mu_2^* t}(\int_0^{l-t} \mu_2^* e^{-\mu_2^* x} e^{-\mu_1 x} e^{-\mu_2(l-t-x)} e^{-\mu_1(l-t-x)} dx)dt \quad \text{if } \Gamma = 01 \quad FS = 00$$

4b. E_2 **precedes** E_1

$$f_b(\mu_1,\mu_1^*,\mu_2,\mu_2^*) = \quad \int_0^l \mu_2 e^{-\mu_2 t} e^{-\mu_1 t}(\int_0^{l-t} \mu_1^* e^{-\mu_1^* x} e^{-\mu_2 x} e^{-\mu_1(l-t-x)} e^{-\mu_2(l-t-x)} dx)dt \quad \text{if } \Gamma = 00 \quad FS = 00$$

$$\int_0^l \mu_2 e^{-\mu_2 t} e^{-\mu_1^* t}(\int_0^{l-t} \mu_1^* e^{-\mu_1^* x} e^{-\mu_2 x} e^{-\mu_1(l-t-x)} e^{-\mu_2(l-t-x)} dx)dt \quad \text{if } \Gamma = 10 \quad FS = 00$$

$$\int_0^l \mu_2^* e^{-\mu_2^* t} e^{-\mu_1 t}(\int_0^{l-t} \mu_1 e^{-\mu_1 x} e^{-\mu_2 x} e^{-\mu_1(l-t-x)} e^{-\mu_2^*(l-t-x)} dx)dt \quad \text{if } \Gamma = 01 \quad FS = 01$$

Chapitre 7

Un *pipeline* pour détecter et expliquer la coévolution

Les méthodes décrites dans les deux articles écrits pendant cette thèse ont donc bien leurs avantages et leurs inconvénients :

— La méthode non-paramétrique est très efficace en termes de temps de calcul, elle peut donc brasser de très grands jeux de données et discriminer les paires qui semblent les plus intéressantes. Par ailleurs, elle fournit des statistiques exactes. Elle ne permet cependant pas d'aller plus loin que rejeter ou non l'hypothèse d'indépendance.

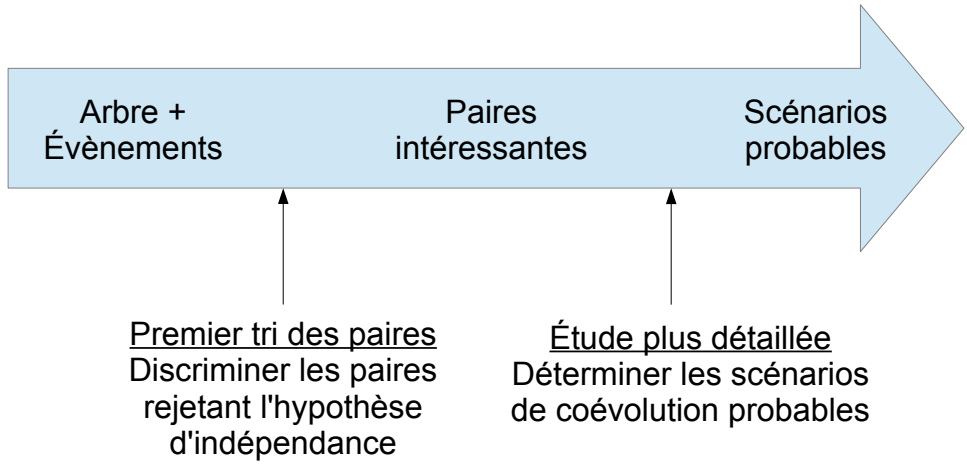
— La méthode paramétrique est bien plus coûteuse en temps de calcul, puisqu'elle revient *a minima* à maximiser une fonction de 4 paramètres qui est longue à évaluer. Elle permet quant à elle d'inférer de probables scénarios de coévolution, en estimant les paramètres, et donc en estimant les inductions, c'est à dire les effets en termes d'accélération ou de décélération des processus sous-jacents, qu'ont les deux évènements l'un sur l'autre.

Les deux méthodes s'applique en outre au même type de données : un arbre phylogénétique et les positions des occurrences de deux évènements évolutifs sur celui-ci. La complémentarité des deux outils que nous avons développés nous poussent donc à imaginer les utiliser conjointement dans le *pipeline* décrit dans la figure 7.1. Dans le cas typique d'un grand jeu de données, par exemple l'étude d'un alignement multiple de séquences, nous pouvons traiter toutes les paires de sites possibles par la première méthode, qui nous permettrait d'efficacement classer les paires, selon leur *p-values* ou leurs *z-score*. Les paires les plus intéressantes pourraient ensuite être traitées plus avant par l'intermédiaire de la deuxième méthode, qui renverrait donc pour chacune d'entre elles les scénarios les plus probables expliquant la co-répartition

observée, selon notre modèle.

Nous pourrions ainsi pallier aux inconvénients inhérents à chacune des méthode, et, à partir de jeux de données conséquents, décrire la dépendance entre les processus évolutifs sous-jacents pour les paires d'évènements les plus pertinentes.

FIGURE 7.1 – ***Pipeline*** **associant les deux méthodes**.



Implémentation

Avant de discuter plus avant les résultats de ce travail et conclure, il est important de discuter certains choix techniques, notamment en matière d'implémentation des deux méthodes que nous décrivons dans cette thèse, et les logiciels utilisés.

Langages de programmation

OCaml

Le choix d'Ocaml pour développer les méthodes s'est fait naturellement, puisque j'ai été formé à la programmation sur ce langage, et que je l'ai pour ainsi dire toujours utilisé. Par ailleurs, les différentes structures de données utilisées, notamment les arbres phylogénétiques et les différentes informations stockées au niveau des nœuds de l'arbre vont dans le sens de son utilisation. En effet, Ocaml a l'avantage d'être très fortement typé, ce qui pose de très nombreuses contraintes quant à la façon de programmer, mais évite de très nombreuses erreurs de typage en forçant le programmeur à plus de rigueur.

C

Malgré tout, quand il a été temps de distribuer EpiCs, le logiciel correspondant à la première méthode décrite dans cette thèse, nous avons basculé vers le langage C. En termes de performance, la différence est saisissante : nous avons gagné un facteur 200 en termes de temps de calcul. Par ailleurs, ce langage est plus répandu, son compilateur existe de manière native sur la plupart des systèmes d'exploitations (notamment basés sur Unix) et le code est ainsi plus facilement portable. Ainsi, l'utilisation du C se fait dans une logique de facilitation de l'usage à un public plus large.

Distribution

EpiCs est donc un logiciel open-source développé en C et compilé avec gcc 4.2.1. Il est uniquement utilisable en ligne de commande. Il est disponible, ainsi que son guide d'utilisation, à l'adresse suivante :

`http://wwwabi.snv.jussieu.fr/public/EpiCs/`.

Il faudrait à terme développer une interface graphique pour rendre ce programme utilisable par le plus grand nombre, notamment les utilisateurs peu ou pas familiers avec l'utilisation de logiciels en ligne de commande.

Logiciels

Nous avons utilisé aussi peu de logiciels externes que possible, préférant coder nous-même la plupart de nos outils. La plupart des logiciels utilisés au cours de cette thèse l'ont donc été dans un but de vérification de calculs ou de visualisation. Certains d'entre eux nous ont été utiles, en particulier pour la construction de nos jeux de données à partir de MSA.

Construction des jeux de données

Nous avons utilisé le logiciel *PhyML* (Guindon et al., 2010) pour reconstruire les différentes phylogénies utilisées au cours de cette thèse, en particulier celle issue de l'échantillon de 54 γ-entérobactéries utilisée pour l'étude des appariements Watson-Crick dans l'ARNr 16S. Pour cette même étude, nous avons par la suite utilisé *BayesTraits Multistate* (Pagel et al., 2004) avec les paramètres par défaut pour reconstruire les états ancestraux.

Calculs et visualisation

Nous avons utilisé *Mathematica* sous license étudiante de l'UPMC pour visualiser diverses fonctions et surfaces, notamment au moment de développer la fonction de vraisemblance. Les arbres phylogénétiques ont quant à eux été visualisés à l'aide des logiciels *NJplot* et *Seaview* (en plus de visualiser les alignements).

Troisième partie

Discussion

Au cours de cette thèse, nous avons développé deux méthodes permettant de détecter et caractériser, à l'aide d'outils mathématiques et informatiques, les interactions entre deux processus évolutifs. Plus précisément, nous avons développé une méthode générale permettant d'étudier la co-répartition de deux catégories de points sur un arbre phylogénétique connu, et de déterminer si cette co-répartition est due au hasard ou non. Les processus évolutifs dont nous parlons sont ceux qui ont conduit à la répartition des points de chaque type, qui pour nous sont des *observées*.

Dans un premier temps, nous avons construit un formalisme mathématique, ancré dans l'algèbre linéaire et le calcul matriciel, permettant de traduire les données (arbre phylogénétique et positions des évènements sur l'arbre), et de calculer, pour deux principaux types d'interactions (cooccurrences et chronologies), des comptages, et des statistiques exactes sur ces comptages. Nous pouvons donc associer un *Z-score* et une *p-value* sous l'hypothèse d'indépendance H_0 – tous deux calculés analytiquement – à une configuration d'évènements sur un arbre. De cette manière, nous sommes capables de rejeter ou non, avec une erreur contrôlée, l'hypothèse selon laquelle les processus ayant conduit à la co-répartition observée sont indépendants. Cette première étape est très rapide et efficace en temps de calcul et permet de traiter de grands jeux de données.

Dans un second temps, nous avons développé un modèle mathématique d'interaction entre deux processus évolutifs, dépendant de 4 paramètres. Nous avons associé une fonction de vraisemblance à ce modèle, nous permettant de calculer la probabilité d'une configuration, connaissant les paramètres du modèle, c'est à dire la probabilité des données conditionnée au modèle. Ainsi, à partir de cette fonction de vraisemblance, nous avons deux possibilités : (i) estimer les paramètres maximisant cette vraisemblance, donc expliquant au mieux les données, ou (ii) échantillonner le paysage au *pro-rata* de la vraisemblance, nous permettant ainsi d'avoir une vue globale de ce paysage. Associer ces deux axes permet d'inférer le ou les scénario(s) expliquant le mieux les données. Bien entendu, cette méthode est plus coûteuse en temps de calcul, mais reste tout à fait raisonnable, puisque ne dépendant que de peu de paramètres, contrairement à d'autres méthodes semblables rencontrées dans la littérature (Pagel et Meade, 2006).

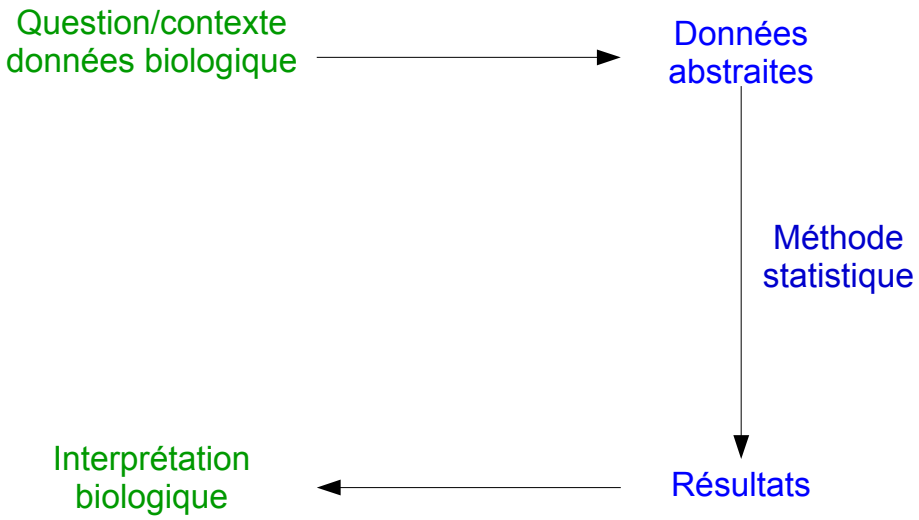
Nous avons finalement testé ces deux outils selon deux axes différents. D'une part sur des données simulées pour lesquels nous contrôlons le scénario de coévolution, nous permettant d'établir diverses courbes de puissance, qui constituent une importante source d'information quant aux biais possibles, ainsi qu'au comportement des méthodes selon les données (topologie de l'arbre étudié, taux de mutation, *etc*). Concernant la méthode faisant intervenir la vraisemblance, les simulations nous ont aussi permis d'avoir une idée précise de la fiabilité de

la méthode quant à sa capacité à retrouver des taux de mutation choisis arbitrairement, ainsi que de retrouver automatiquement le scénario de coévolution suivi par la simulation. Enfin, nous avons testé les méthodes sur deux jeux de données biologiques réels : la perte du flagelle chez *Escherichia coli* pour la méthode non paramétrique, et les appariements Watson-Crick dans un jeu de donnée de γ-entérobactéries pour les deux méthodes, avec des résultats très satisfaisants.

Cette façon de caractériser notre méthode permet de soulever plusieurs points que nous discuterons dans la suite de cette partie. Le premier est le lien entre les données biologiques réelles qui nous servent de base de travail et les données abstraites qui sont traitées par notre méthode. Il est important de poser la question de la pertinence du contexte biologique qui se cache derrière ce travail, et du lien qui existe entre ces hypothèses biologiques et la construction des jeux de données traités. Ensuite, il va sans dire qu'il existe des biais associés à la nature même de la méthode que nous décrivons ici. Ces biais peuvent être liés aux données que la méthode prend en entrée comme aux outils que nous employons. Enfin, nous terminerons par une réflexion sur les questions biologiques auxquelles nous tentons de répondre, ainsi que sur la valeur de nos réponses, en particulier en termes d'interprétation biologique.

La figure 7.2 schématise ce lien entre les données et résultats biologiques, par l'intermédiaire de données abstraites traitées par notre méthode.

FIGURE 7.2 – **Schéma résumant la démarche, depuis une question biologique à l'interprétation des résultats renvoyés par notre méthode.**



Chapitre 8

Des données aux modèles

La base de ce travail de thèse est une question biologique, en l'occurrence, la caractérisation de la coévolution. Comme nous le disions dans l'avant-propos de cette thèse, et comme cela a été répété dans les deux articles présentés précédemment dans ce manuscrit, il est d'une importance primordiale de pousser la réflexion à propos des hypothèses, ne serait-ce que pour questionner les modèles et le rapport entre les données, et les résultats que nous donnent les diverses méthodes mathématiques. Il est nécessaire de rattacher les modèles aux données, par le biais d'une réelle interrogation quant à la nature des données traitées et du contexte biologique sous-jacent.

8.1 De l'intérêt de se poser les bonnes questions

Quelles sont les bonnes questions à se poser ? Voilà une phrase qui a le mérite de bien résumer ce qui doit rester la base de ce travail. Quels types de conclusions voulons nous faire ressortir à partir des données ? À quel genre de questions voulons-nous tenter de répondre à travers les méthodes mathématiques que nous décrivons ici ?

Nous ne pouvons évidemment pas décrire l'histoire passée avec certitude, aussi est-il important de faire preuve de précision quant aux questions biologiques, et aussi mathématiques, que nous nous posons. D'un point de vue très pragmatique, donc, nous avons développé une méthode qui nous permet d'étudier la co-répartition de deux catégories de points sur un arbre. Ces points correspondent à des occurrences d'évènements évolutifs, et la distribution de ces points sur l'arbre phylogénétique est une réalisation de processus sous-jacents. Étudier cette co-répartition peut donc se faire, à partir de cet observé (co-répartition des deux catégories de

points sur l'arbre phylogénétique en question), en estimant si, sous un modèle mathématique donné de coévolution, nous sommes capables de rejeter l'hypothèse d'indépendance.

Le lien entre les données et la méthode est clair et nous permet de mieux situer le cadre dans lequel se place notre méthode, en particulier les questions fondamentales auxquelles nous tentons de répondre. Tout d'abord une première question biologique : à partir de données observées actuelles, du type présence/absence de telle ou telle fonction biologique sur un arbre phylogénétique, les processus évolutifs à l'origine de ces données sont-ils indépendants ? Ajoutons à cette question initiale une deuxième, cette fois plutôt liée au contexte mathématique : à quel niveau de confiance pouvons nous placer la réponse donnée à la première interrogation ? Enfin, répondre à ces deux questions n'est pas complètement satisfaisant, les résultats de la méthode non paramétrique nous le prouvent : une fois que nous avons rejeté l'indépendance entre les processus évolutifs d'intérêt, quels sont les scénarios de coévolution que nous pouvons inférer, qui aurait pu conduire à la "photographie" contemporaine observée ?

8.2 L'hypothèse H0

Ainsi, l'hypothèse dite "nulle", qu'on appelle ici H_0, découle à la fois du modèle mathématique et statistique, des outils développés, mais aussi des arguments biologiques sous-jacents. Il était donc important de consacrer une section à cette hypothèse, qui a été longuement discutée, modifiée et travaillée au cours de ce travail de thèse.

La première façon que nous avons eue d'envisager H_0 a été de simplement l'appeler "hypothèse d'indépendance", sans la questionner plus avant. Il est assez naturel de la considérer ainsi, puisque finalement, les tests statistiques que nous avons développés ont pour but de rejeter ou non l'indépendance entre les processus évolutifs. Mais cette définition purement mathématique n'est pas suffisante à partir du moment où nous mettons en perspective l'aspect biologique des donnes traitées. En effet, les points sur un arbre, qui finalement constituent les données que nous traitons à l'aide de notre méthode, proviennent de données biologiques qu'il convient de mieux préciser.

La distribution des occurrences des évènements sur l'arbre phylogénétique peut dépendre d'un *a priori* sur les différentes branches de l'arbre, ce qui peut poser problème, et nous a amenés à mieux spécifier H_0. Nous avons, au fur de notre avancée, identifié un écueil principal qu'il convient de citer ici. Il est possible d'abaisser artificiellement la *p-value*, pour une paire d'évènements, si l'on a un *a priori* sur les taux d'occurrences de ces évènements sur certaines parties de l'arbre. Ceci peut être caractérisé de deux façons différentes.

La première façon de voir ici les choses est de considérer un arbre pour lequel on connaît des régions (en général des sous-arbres, ou au moins des ensembles de branches) dans lesquels l'un ou l'autre des évènements (ou les deux) sera absent. C'est typiquement le cas lorsque l'on étudie certains clades particuliers qui possèdent une fonction biologique (par exemple la mobilité chez *Escherichia coli*) en sachant à l'avance que cette même fonction est absente dans le reste de l'arbre. Ainsi, l'évènement qui correspond au gain de cette fonction n'est présent que dans certaines parties de l'arbre, ce qui peut aisément générer des cooccurrences ou des chronologies avec d'autres évènements évolutifs qui n'auraient pas lieu d'être significatives si l'on se restreint aux parties de l'arbre où cette fonction est potentiellement présente. Notons aussi que la longueur totale de l'arbre a une influence notable sur la significativité des résultats, puisqu'elle entre en compte lors de la normalisation des longueurs de branches, elles-mêmes prises en compte lors du calcul des *p-values* ou de la vraisemblance des données.

Le deuxième type de connaissances *a priori* que l'on peut avoir à propos des données étudiées – et qui encore une fois fausse la méthode – correspond au cas où certaines lignées sont connues pour porter plus d'évènements que d'autres. Encore une fois, cela peut passer l'enrichissement de l'échantillon d'espèces étudiées, en ajoutant par exemple à l'arbre phylogénétique des lignées privilégiée quant à l'apparition d'occurrences de tel ou tel évènement. L'accumulation de mutations conduit au comptage de cooccurrences et de chronologies qui n'apparaissent pas nécessairement dans le reste de l'arbre, et ainsi à des *p-values* plus faibles.

Nous avons donc défini avec bien plus de précision notre hypothèse de travail qui est alors "Pour deux évènements évolutifs, les processus conduisant à la distribution des points sur l'arbre ont **des intensités constantes dans tout l'arbre**, et sont indépendants l'un de l'autre.". Cette légère modification sous-entend à présent que nous n'avons pas d'*a priori* sur les variations de taux le long des lignées, et donc, concernant les positions des occurrences des évènements sur l'arbre, pas d'*a priori* sur (i) leur présence/absence et (ii) sur leur concentration.

Chapitre 9

Les biais

Dans ce chapitre, nous nous attacherons à décrire et discuter divers biais - principalement statistiques - que nous avons rencontrés au cours de cette thèse. Il va sans dire que certains d'entre eux étaient attendus, puisqu'à partir du moment où nous appliquons des méthodes statistiques et abstraites à des données réelles, la transition $données \rightarrow formalisme$ implique nécessairement la définition d'hypothèses (comme nous avons par exemple pu le voir plus haut avec la description de l'hypothèse d'indépendance H_0), d'approximations, et tout simplement de calculs qui par essence ne nous donnent qu'une version probable de l'histoire passée. Le concept même de *certitude absolue* nous est complètement inaccessible, et ainsi, il est important de comprendre d'où provient cette incertitude.

Bien que nous partions du principe que, pour deux évènements évolutifs donnés, l'*input* pour notre méthode est constitué d'un arbre phylogénétique et des positions des occurrences des évènements sur cet arbre, nous nous devons de détailler un minimum les biais dus à la construction de tels jeux de données.

9.1 Les biais dus aux données brutes

Un premier degré d'incertitude se situe au niveau des données brutes. En effet, les arbres phylogénétiques et les positions des évènements d'intérêt dépendent fondamentalement de la qualité des données. Avant même de reconstruire les historiques mutationnels, il arrive que des erreurs de séquençage ou de manipulation conduisent à fausser les données. Il convient donc de vérifier la qualité des données en amont. Ces erreurs peuvent se manifester de différentes façons, selon le type et la nature des données traitées, et peuvent ainsi générer différentes

erreurs d'analyse.

Les alignements multiples de séquences sont souvent le matériel de départ lors de la reconstruction des arbres phylogénétiques et des états ancestraux. Ainsi, une erreur de séquençage aura pour conséquence de fausser un état au niveau d'une feuille de l'arbre phylogénétique, d'où un biais à ce niveau, qui peut être transposé aux niveaux supérieurs de cet arbre, en remontant les lignées jusqu'à sa racine. Par ailleurs, dans un jeu de donnée biologique où les séquences divergent peu, les mutations sont généralement rares. Une telle erreur peut donc avoir un impact non négligeable. Malgré tout, ce type d'erreur peut être parfois "noyé" dans la masse, si la séquence étudiée est suffisamment longue, par exemple.

Nous pouvons aussi rencontrer des erreurs au niveau de l'observation de telle ou telle fonction biologique chez une espèce étudiée. L'erreur typique peut être un caractère noté absent alors qu'il est présent. Dans ce cas, l'impact est d'autant plus grand que nous ne parlons plus ici d'une erreur ponctuelle le long d'une séquence de parfois plusieurs centaines voire milliers de loci, mais d'une erreur sur un trait, qui sera, nous l'avons vu plus haut et nous y reviendrons en fin de partie, généralement étudié dans le cadre d'une seule paire précise.

9.2 Reconstruction des phylogénies

La première source importante de biais statistique concerne la reconstruction de l'arbre phylogénétique, nécessaire à l'application de la méthode. Dans la mesure où nous avons principalement travaillé avec des arbres déjà existants, nous avons peu observé ces biais dans nos études. Nous développerons dans la section suivante les biais dus aux méthodes de reconstruction des états ancestraux, qui sont largement similaires. Malgré tout, précisons que, quelle que soit la façon dont l'arbre est reconstruit, il existe une incertitude quant à sa topologie, d'une part, et aux longueurs de branches d'autre part. Il convient donc de garder à l'esprit que cet arbre n'est pas définitivement fixé, mais plutôt associé à une probabilité (calculée dans le cas d'une reconstruction par une méthode de vraisemblance).

Par ailleurs, il n'est pas pertinent de reconstruire l'arbre en question à partir des données brutes, c'est à dire l'alignement multiple, si nous sommes dans un cadre moléculaire, ou la matrice de présence/absence d'un trait biologique. En effet, dans ce cas, l'information qui nous sert à reconstruire l'arbre est redondante avec la position des évènements évolutifs. Pour prendre un exemple simple, nous savons typiquement que dans le cas d'un alignement multiple, les polymorphismes conditionneront grandement les branchements dans cet arbre, les longueurs de branches (directement liées au nombre moyen de mutations portées par ces

branches), mais aussi la forme même de l'arbre. Or, nous savons que l'histoire évolutive, par exemple de deux gènes, peut différer du fait de recombinaisons.

Toutefois, il est possible d'assurer un certain degré de confiance, ou au moins de consensus, quant aux phylogénies manipulées, par exemple en les reconstruisant sur le *core genome*, c'est à dire la partie du génome partagé par tout l'échantillon étudié dans le cas de bactéries. Par exemple, lors du travail sur notre premier exemple biologique, en l'occurrence la perte du flagelle chez *Escherichia coli*, nous avons obtenu un arbre que nous pouvons qualifier de "consensus" pour toutes les souches de l'échantillon, puisque construit sur ce *core genome*, et non pas uniquement inféré à partir des données directement liées aux gènes concernés par le flagelle en question.

9.3 Reconstruction des états ancestraux

Par analogie avec la reconstruction de l'arbre phylogénétique, celle des états ancestraux est aussi sujette à divers niveaux d'incertitude. Nous sommes d'autant plus concernés par cette problématique que, plus encore que pour les phylogénies, nous pouvons être amenés à reconstruire nous-même les états ancestraux.

Nous avons vu plus tôt que, pour éviter la redondance de l'information portée par les données observées (gain/perte d'une fonction, alignement multiple) et l'arbre phylogénétique étudié, il est possible de reconstruire un arbre à partir de données plus "générales", c'est à dire par exemple le *core genome* dans le cas des bactéries. *A contrario*, une fois l'arbre reconstruit, l'inférence des états ancestraux, nécessaire pour replacer les évènements sur les arbres - notre matière première - ne peut évidemment se faire qu'à partir de ce que nous appelons "photographie" contemporaine, c'est à dire l'état des caractères étudiés, qu'ils soient moléculaires (*e.g.* séquences contemporaines) ou non-moléculaires (*e.g.* "présence d'un flagelle fonctionnel").

9.3.1 Parcimonie

Nous avons vu que la reconstruction des états ancestraux par parcimonie était aussi potentiellement génératrice d'incertitude. Il est en effet parfois nécessaire de poser des hypothèses pour appliquer cette méthode de reconstruction. Généralement, cette méthode de reconstruction, qui tend à minimiser le nombre d'évènements sur l'arbre expliquant les données observées, dépend du choix du modèle : plusieurs reconstructions possibles sont équivalentes

en ce sens et ainsi discriminer parmi celles-ci revient à faire des choix, parfois importants. Citons ici les deux grandes hypothèses, auxquelles nous avons été confrontés, notamment lors de l'étude de la perte du flagelle et de la mobilité chez *Escherichia coli* et *Shigella*.

La première hypothèse, dite ACCTRAN (pour *ACCelerated TRANsformation*) favorise les réversions et concentre ainsi les évènements vers la racine de l'arbre. En effet, permettre les réversions revient à autoriser le placement d'évènements anciens, tandis que l'hypothèse DELTRAN favorise la convergence évolutive, donc les évènements tardifs, plus proches des branches terminales de l'arbre phylogénétique. Ces deux hypothèses sont illustrées par la figure 3.2.

Dans tous les cas, citons une fois de plus une précaution importante à prendre : une bonne compréhension biologique des données est vitale pour pouvoir appliquer la méthode tout en limitant l'incertitude inhérente à la construction du jeu de donnée. Prenons l'exemple de la perte du flagelle chez *Escherichia coli*. Nous avons décidé d'appliquer l'hypothèse DELTRAN, qui concentre de manière privilégiée les mutations vers les branches terminales de l'arbre. Ce choix est justifié par le fait que la perte du flagelle est génétiquement moins coûteuse que son gain, du fait de la multitude de gènes nécessaires à l'obtention d'un flagelle fonctionnel. Les réversions sont ainsi moins justifiées et les évènements sont plus concentrés vers les branches terminales. Notons par ailleurs qu'il ne faut pas négliger l'impact de ce choix, puisqu'il tend à générer plus de cooccurrences que si nous avions appliqué l'hypothèse ACCTRAN (une chronologie supplémentaire pour une cooccurrence en moins, voir figure 3.2), ce qui fait directement inférer par la méthode une induction plus importante.

9.3.2 Méthodes utilisant la vraisemblance

La principale critique que nous pouvons formuler pour les méthodes utilisant la vraisemblance pour reconstruire les états ancestraux est qu'elles dépendent du choix d'un modèle évolutif. Ce modèle se doit de reposer sur des bases biologiques solides pour être justifiable. En parallèle, comme nous l'avons vu, ces méthodes peuvent s'avérer très coûteuses en termes de temps de calcul. Ainsi, un modèle sera plus précis si il possède plus de paramètres, mais la méthode nécessitera plus de temps pour être appliquée. Il faut donc être précautionneux quant à ce compromis entre précision et temps de calcul, et éviter les modèles trop simplistes qui ne retranscrivent qu'une part de la réalité. *A contrario*, les modèles plus détaillés ont certes l'avantage d'être plus proches d'une réalité biologique, mais empêchent de travailler sur des jeux de données conséquents.

9.4 Biais dus à la méthode

Enfin, mis à part les données qui servent de matière première à l'application de notre méthode, elle-même est potentiellement affectée par plusieurs biais que nous exposons ici.

9.4.1 Les sites monomorphes

Le premier point que nous abordons ici est plus une limite qu'un biais. Pour appliquer la méthode, nous avons donc besoin d'un arbre phylogénétique, et d'au moins deux types d'évènements pour lesquels nous connaissons la distribution des occurrences dans cet arbre. Bien entendu, dans le cas d'une étude ponctuelle de deux évènements évolutifs précis, nous pouvons partir du principe que nous avons une observée variable, nous permettant d'inférer au moins une occurrence de chaque évènement dans l'arbre, ce qui suffit à l'application de notre méthode. En revanche, si nous étudions par exemple un alignement multiple, certaines de ses colonnes seront potentiellement parfaitement homogènes, correspondant aux sites monomorphes. Dans ce cas, il n'y aura aucune mutation dans l'arbre phylogénétique inférée pour ces sites. Même si ces sites coévoluent, cette information nous est totalement inaccessible.

Toute l'information ne nous est donc pas accessible, et nous n'avons aucun moyen direct d'éliminer ce problème. Nous pouvons toutefois le contourner en enrichissant le jeu de données et en ajoutant des espèces pour lesquels les évènements en question existent. Plus pratiquement, si nous revenons à notre exemple de sites monomorphes pour un alignement multiple, il suffirait d'ajouter à notre échantillon une espèce pour laquelle ce site diverge par rapport aux autres. Bien entendu, il est dangereux de faire un tel raccourci sans précaution : rappelons que l'hypothèse H_0 à laquelle nous confrontons les données pour estimer la dépendance entre les évènements évolutifs impose que les évènements aient lieu à taux constants dans tout l'arbre en question, sans *a priori* sur ces taux. Ainsi, on ne peut pas partir du constat qu'un site est monomorphe pour en déduire qu'il faut enrichir l'échantillon. Nous pouvons en revanche limiter ce genre de problème en considérant *a priori* des échantillonnages plus larges, assurant un degré de divergence assez élevé pour augmenter le nombre de sites polymorphes sur l'alignement.

9.4.2 Un nombre entier de mutations est nécessaire

Nous avons aussi vu que la méthode que nous décrivons ici permet d'établir de nombreux résultats complémentaires relatifs à l'indépendance de paires d'évènements évolutifs. En d'autres termes, nous avons accès à un calcul analytique d'espérances, de variances, de distributions exactes (et donc de *p-values*) par la première partie, et d'un outil utilisant un calcul de vraisemblance pour la deuxième partie. Concernant l'espérance et la variance, la formule que nous avons dérivée de notre formalisme matriciel ne pose aucun problème quant au nombre d'évènements, puisqu'il s'agit là de calculs matriciels facilement applicables pour des matrices à coefficients non entiers.

La question se corse au moment de calculer la distribution exacte qui nous amène à l'établissement des *p-values*. En effet, nous utilisons une méthode associant du calcul matriciel et un raisonnement combinatoire pour accéder à cette distribution. Mathématiquement parlant, nous devons lister toutes les façons de distribuer n points dans m cases d'un vecteur. Ce travail nécessite de manipuler des nombres entiers, or selon les méthodes de reconstruction des états ancestraux (voir plus haut), il est possible d'inférer, du fait de l'incertitude inhérente à ces méthodes, non pas un nombre d'évènements, mais une distribution sur les états possibles et ainsi des mutations, par exemple. En d'autres termes, l'ambiguïté due à certaines méthodes de reconstruction doit être levée, par exemple en faisant des choix arbitraires. Typiquement, dans le cas de séquences d'ADN, les méthodes paramétriques associeront à chaque noeud de l'arbre, pour chaque site, une distribution de probabilité sur $\{A, C, G, T\}$. Cette ambiguïté peut donc être levé en faisant des choix parfois forts, par exemple en considérant que l'état associé à la probabilité la plus élevée est l'état choisi (ce qui peut poser problème lorsque cette distribution est quasiment uniforme).

9.4.3 Les mutations multiples sur une même branche

Un prolongement du point précédent concerne cette fois uniquement la méthode utilisant les vraisemblances. En effet, nous venons de voir que nous avons besoin d'un nombre entier de mutations pour pouvoir l'appliquer, mais en réalité il faut aller plus loin. La fonction de vraisemblance que nous avons calculée est exacte, ce qui est très important pour la robustesse de la méthode et pour l'efficacité en termes de temps de calcul. Néanmoins, puisque cette même fonction prend en compte toutes les configurations possibles pour chaque branche de l'arbre prise séparément, nous avons pris un raccourci nécessaire en imposant le fait que chaque évènement est *au maximum* présent une seule fois par branche (et c'est à ce prix que la vraisemblance est exacte).

En effet, imaginons que nous considérons deux types d'évènements que nous appelons E_1 et E_2. Par hypothèse, nous connaissons "à la branche près" les positions des occurrences de chaque évènement, donc pour chaque branche prise individuellement, nous n'avons pas l'information de la position exacte de ces occurrences sur celle-ci. Prenons le cas de figure où chaque évènement possède deux occurrences sur une branche. Ainsi, sur celle-ci, le calcul de la vraisemblance doit prendre en compte tous les ordres chronologiques possibles, c'est à dire $[E_1E_1E_2E_2]$, $[E_2E_2E_1E_1]$, $[E_1E_2E_1E_2]$ et $[E_2E_1E_2E_1]$, soit quatre termes à calculer. On imagine assez facilement que ce nombre de termes explose quand le nombre d'occurrences augmente pour une branche, d'autant plus que ce nombre est arbitrairement grand, donc le nombre de termes à calculer l'est aussi. La fonction de vraisemblance pour une seule branche étant par ailleurs impossible à calculer récursivement, à partir de termes d'ordres inférieurs, nous avons donc fait le choix de considérer au maximum une seule occurrence pour un évènement et pour une branche.

Notons que cette approximation pose rarement un problème pour des cas biologiques réels, où les mutations sont généralement rares. Nous pouvons principalement le rencontrer dans le cas où l'arbre possède de très longues branches, qui pourraient donc *par hasard* porter de nombreuses mutations. Nous pouvons dans une certaine mesure contourner ce problème en choisissant en amont une résolution phylogénétique telle que les branches ne peuvent que rarement porter plus d'une mutation.

Il est aussi important de discuter ce biais du point de vue biologique. En effet, ces cas de mutations (ou plus généralement, évènements) multiples sur une seule et même branche de l'arbre ne sont généralement pas observables. Prenons par exemple le cas de mutations multiples sur un locus le long d'une branche. Il suffit d'associer une mutation et une réversion sur un même locus pour qu'aucun de ces deux évènements ne soit observable. Plus la branche est longue, plus le nombre de mutations qu'elle portera sera importante, et ainsi, plus le nombre de mutations/réversions sera statistiquement important. Ainsi, sans même considérer l'aspect modèle du problème que nous exposons dans les paragraphes précédents, nous sous-estimons très probablement le nombre d'évènements, en particulier pour les longues branches. Ce problème est par ailleurs difficile à contourner, du fait que nous reconstruisons notre matériel d'étude à partir de données contemporaines observées, l'inférence de ces mutations/réversions se révélant de ce fait très difficile.

9.4.4 Notre modèle de coévolution n'est pas état-dépendant

Notre modèle de coévolution est assez complet, puisqu'il permet de formaliser un très grand nombre de scénarios que nous avons bien décrits dans les parties précédentes. C'est en particulier le cas des scénarios que nous avons nommés "Prohibition", "Réciprocité" ou "Induction". Malgré tout, certains autres scénarios rencontrés *in natura* ne sont pas accessibles grâce à ce formalisme à quatre paramètres.

C'est le cas du scénario de la perte de gène, par exemple. Dans ce cas, nous pouvons imaginer, par analogie avec notre modèle, que le processus évolutif lié à l'évènement "Perte du gène X" est un processus poissonnien le long des lignées de l'arbre phylogénétique, de taux λ. Pourtant, à la première occurrence de cet évènement, le gène est perdu, donc nous ne devrions logiquement pas pouvoir observer d'autres occurrences de cet évènement dans la suite de cette lignée. Notre modèle ne permet pas de formaliser de telles situations, puisqu'il n'est pas état-dépendant. En d'autre termes, chaque évènement peut avoir un effet sur tous les autres évènements, mais pas sur lui-même, ce qui interdit de traiter, et donc évidemment d'inférer certains scénarios pourtant possibles dans la nature.

Chapitre 10

Conclusions biologiques

Avant d'aborder les perspectives, et de ce qui pourrait améliorer ou étendre ce travail de thèse, il est important de parler des conclusions biologiques que nous pouvons tirer de ce travail.

10.1 À quelles questions pouvons nous tenter de répondre ?

Nous avons donc établi au début de cette partie trois grandes questions qui résument notre démarche, à savoir :

- À partir de données observées actuelles, du type présence/absence de telle ou telle fonction biologique, ou à partir d'un alignement multiple de séquences, dans la mesure où nous pouvons reconstruire avec confiance l'histoire qui a mené à cette observation, les processus évolutifs à l'origine de ces données sont-ils indépendants ?
- À quel niveau de confiance pouvons-nous placer la réponse donnée à la première interrogation ?
- quels sont les scénarios de coévolution que nous pouvons inférer, qui auraient pu conduire à la "photographie" contemporaine observée ?

À la première question, nous pouvons répondre qu'il faut prendre de grandes précautions quant à la "mise en forme des données", c'est à dire le choix de l'échantillonnage pour entrer au mieux sous les conditions de l'hypothèse de travail (processus évolutifs poissonniens à taux constants le long des lignées de l'arbre), et qu'il faut garder à l'esprit que la reconstruction de l'histoire évolutive conduit nécessairement à introduire de l'incertitude. Ces précautions prises, il est possible d'estimer si les processus évolutifs dont cette histoire est une réalisation sont indépendants. Concernant la deuxième question, c'est le calcul analytique des *Z-scores*,

mais surtout des *p-values* qui permet d'estimer la significativité de ce résultat, en donnant une estimation numérique de celle-ci. Enfin, nous pouvons estimer, selon notre modèle simple de coévolution, quel est, ou quels sont les scénarios les plus probables, c'est à dire estimer les paramètres du modèle qui décrivent le mieux ces processus évolutifs.

D'un point de vue biologique, c'est la nature des données étudiées qui permet de poser un cadre concret à ces trois questions. En pratique, la méthode que nous avons développée est suffisamment formelle pour traiter des données d'origines et de natures très variées : puisque nous travaillons avec des points sur un arbre, tout processus discret dans le temps peut être ainsi étudié. Nous pouvons travailler avec des données moléculaires, pour étudier la coévolution entre les locus d'un génome, ou les résidus d'une protéine. Comme nous l'avons vu, ce genre d'étude peut aussi inférer des liens plus concrets, par exemple des contraintes physico-chimiques, ou stériques sur des protéines ou de l'ARN. Nous pouvons aussi considérer des données à une échelle physique plus grande, comme l'apparition ou la disparition d'un trait ou d'une fonction biologique. Notons par ailleurs qu'il est possible d'utiliser des données de différentes natures lors de la même étude, par exemple pour rechercher les parties d'un génome codant pour telle ou telle fonction, en confrontant des données moléculaires (mutations) avec l'apparition de ce trait.

10.2 De l'intérêt de bien doser l'information biologique

Pour pouvoir répondre à toutes ces questions, il faut bien entendu que l'on puisse d'une part abstraire les données étudiées, mais aussi que ces données portent un signal, une information que nous pouvons capturer à l'aide de notre méthode, et ainsi pouvoir donner un résultat interprétable. Là encore, plusieurs questions se posent, et autant d'écueils sont à éviter.

Une question que nous n'avons pas précisée plus haut, mais qui reste aussi vitale est celle de l'échelle de temps à laquelle nous voulons placer notre étude. Celle-ci est directement liée à l'échantillonnage des données que l'on souhaite étudier. Une échelle de temps trop grande, *i.e.* un choix d'espèce trop éloignées impliquera de très longues branches au niveau de l'arbre phylogénétique reconstruit, et ainsi de très nombreuses mutations le long de ces branches. Nous risquons alors d'observer un trop-plein de signal, et donc principalement du bruit statistique, du fait de la saturation de l'arbre en évènements. *A contrario*, des espèces trop proches entraîneront l'inférence de trop peu de mutations, d'un arbre phylogénétique ramassé sur lui-même, et ainsi, aucun signal statistique exploitable. D'une manière générale, comme nous l'avons montré notamment à l'aide de nos courbes de puissance, les évènements

doivent être rares, au sens où ils ne doivent pas saturer l'arbre, mais être tout de même assez présents pour offrir un signal qui peut être capturé par la méthode. **Pour cela, il faut prendre en compte l'équilibre qui existe dans les données entre l'échelle de temps considérée, les taux de mutations et l'intensité de l'échantillonnage.** Ce sont ces trois paramètres très liés qui définiront la résolution de l'étude menée et la qualité du signal statistique exploitable.

Il ne faut pas non plus oublier les évènements de recombinaison. Nous en avons déjà parlé au moment d'évoquer les biais statistiques dus à la reconstruction des phylogénies : chez les bactéries, ces évènements rendent quasiment nécessaire cette reconstruction sur les *core-genomes* plutôt que gène par gène. La recombinaison pose donc la question de l'existence même d'une phylogénie consensus pour les différents gènes étudiés, mais de manière encore plus générale, elle pose la question de l'unicité d'une histoire évolutive sous forme d'arbre à inférer. En effet, de forts taux de recombinaison impliquent une divergence complète de l'histoire évolutive que nous voulons inférer à l'aide de notre méthode, et donc plus de signal statistique exploitable par celle-ci puisque les différents gènes ne partagent plus d'histoire commune. C'est en particulier le cas chez les eucaryotes : les taux de recombinaison sont si élevés *au sein d'une même espèce* que nous ne pouvons typiquement pas inférer d'arbre consensus autre qu'en étoile, et ainsi perdons l'information chronologique, qui est pourtant une des plus importantes pour l'application de notre méthode.

10.3 Deux manières complémentaires d'utiliser nos méthodes

Maintenant que nous avons vu les précautions avec lesquelles il est nécessaire de construire son jeu de données, il faut noter que notre méthode permet de les aborder de deux manières différentes, et finalement très complémentaires. C'est principalement la nature des données, les contraintes de temps de calcul, et le degré de précision des réponses apportées qui dictent l'angle à choisir pour traiter les données.

D'une part, il est possible d'utiliser notre méthode non-paramétrique sur de très grands jeux de données, par exemple sur des séquences d'acides nucléiques ou d'acides aminés pour rechercher les paires de sites coévoluant parmi toutes les paires possibles. Ainsi, sur une séquence complète potentiellement longue, nous pouvons "mesurer" la quantité de coévolution, c'est à dire la proportion de paires de sites qui coévoluent. Nous pouvons aussi établir un classement de ces paires, ordonné selon les *Z-scores* ou les *p-values*. Comme nous l'avons vu, cette

première façon d'utiliser la méthode est à appliquer en particulier à des jeux de données moléculaires que nous pouvons ainsi traiter en masse. En contrepartie, le degré de temporalité sous-jacent aux résultats est à choisir en amont de l'étude, puisqu'il est conditionné par le choix de la matrice M, comme nous l'avons vu plus tôt.

D'autre part, nous pouvons aussi appliquer la méthode à seulement deux types fixés d'évènements. Dans ce cas, une étude complète est possible avec la méthode utilisant le calcul de vraisemblance. En plus de se positionner par rapport à H_0, par un calcul de *p-value* (par l'emboîtement des modèles), nous pouvons inférer l'histoire évolutive la plus probable. Par ailleurs, une étude plus poussée du paysage de vraisemblance grâce à la méthode de MCMC permet d'avoir une vision globale de tous les scénarios possibles, sans se contenter de localiser le pic de vraisemblance, qui n'est d'ailleurs pas nécessairement très marqué. Cette méthode est donc particulièrement adaptée à de petits jeux de données, par exemple constitués de traits biologiques peu nombreux, en raison du temps de calcul nécessaire qui est quadratique en le nombre d'évènements candidats, et bien plus long que pour la méthode non-paramétrique. Malgré tout, dans ce cas, nous sommes capables d'obtenir une réponse complète quant à la coévolution des évènements concernés.

Mises bout à bout, ces deux méthodes se révèlent très complémentaires, puisqu'elles peuvent être les deux parties d'un seul et même outil, un *pipeline* constitué d'une part de la première méthode qui permet de traiter rapidement de (très) grands jeux de données, dans le but de rechercher parmi ces données quelles sont les paires d'évènements qui co-évoluent, et d'autre part de la seconde méthode qui permet d'inférer plus précisément les scénarios expliquant le mieux les données observées. En d'autres termes, parmi un nombre très important d'évènements évolutifs, ce pipeline permet de tester toutes les paires possibles (ce qui augmente considérablement le nombre de cas à traiter), et de discriminer celles qui coévoluent de manière significative.

10.4 Correctement interpréter les résultats

La dernière étape d'une étude utilisant la méthode décrite dans cette thèse est bien entendu l'interprétation des résultats. C'est là qu'à partir de l'output abstrait renvoyé par la méthode, nous pouvons déduire une "histoire" biologique cohérente et crédible, appuyée par les résultats. En ce sens, l'un des principaux avantages de la méthode que nous exposons ici est le calcul exact des statistiques (*z-scores* et *p-values* sous l'hypothèse d'indépendance H_0). Nous tirons de ce fait un avantage important : les *p-values* sont exactes et non pas calculées

par simulation. La méthode gagne donc en efficacité en termes de temps de calcul, puisque le calcul analytique nous épargne de fastidieuses simulations qui renvoient par ailleurs un résultat dont la précision dépend du nombre de simulations engagées.

Par ailleurs, nous avons fait le choix d'un modèle simple à 4 paramètres pour le calcul de vraisemblance. Ce modèle certes simpliste, a plusieurs avantages importants par rapport aux modèles utilisant un nombre important de paramètres, comme par exemple celui de Pagel et Meade (2006) :

— Il limite considérablement les temps de calcul, puisqu'il restreint le nombre de paramètres à optimiser (voir la comparaison des différentes méthodes mettant en jeu des calculs de vraisemblance, en introduction).

— Il permet d'obtenir des résultats plus fiables avec des jeux de données restreints. En effet, un modèle qui repose sur plus de paramètres nécessitera plus de données pour que le signal soit suffisant pour tous les optimiser.

— Il permet d'aisément interpréter les résultats, notamment parce qu'il met en jeu deux paramètres rassemblant l'essentiel de l'information portée par le modèle, *i.e.* les inductions λ_1 et λ_2 qui représentent l'intensité de la coévolution.

Devant cette variété de résultats différents accessibles, il ne faut surtout pas négliger l'importance des plus basiques, c'est à dire du comptage en soi. Imaginons par exemple le cas de figure où, lors de l'étude d'une séquence nucléotidique, une paire de sites est retenue, pour laquelle est associée une *p-value* très significative. Estimer la pertinence de cette paire par cette valeur, ou par le *z-score* associé est certes pertinent, mais pas suffisant. En effet, nous pouvons facilement imaginer le cas où cette paire correspond à une unique cooccurrence sur une très petite branche, auquel cas cette association est en effet très significative, mais à relativiser par rapport à d'autres paires qui pourraient être tout aussi significatives mais associées à des comptages plus importants. Dans tous les cas de figures, il faut bien entendu retenir les paires significatives, mais c'est à l'utilisateur d'avoir le dernier mot et d'interpréter les résultats selon le contexte biologique de l'étude en question.

Chapitre 11

Perspectives

Pour conclure cette thèse, parlons des perspectives d'amélioration des outils décrits dans cette thèse, et de correction de certains biais. Plusieurs pistes existent, que nous nous efforçons de lister ici.

11.1 Traiter des graphes qui ne sont pas des arbres

Une première piste consisterait à travailler non pas seulement sur des arbres phylogénétiques, mais aussi sur des graphes plus généraux, ou *réseaux*, de plus en plus souvent rencontrés dans la littérature. Un arbre est par définition un graphe non-orienté, acyclique et connexe. En l'absence de la propriété de connexité, il suffit de travailler séparément sur les composantes connexes de ce graphe, puisque ce sont elles qui, individuellement, représentent une histoire partagée par leurs feuilles. L'aspect non-orienté n'est pas non plus discutable ici, puisque nous considérons, dans le cas des arbres, qu'il existe une racine, qui sert de point de départ pour l'axe temporel, lui aussi nécessaire pour établir la notion de chronologie entre les évènements. En réalité, nous travaillons donc avec des arbres enracinés, donc pour lesquels le sens dans lequel le temps se déroule est fixé.

C'est donc l'hypothèse d'acyclicité qui pourrait être relaxée, de façon à ce que la méthode soit applicable à des réseaux, par exemple des arbres pour lesquels nous ajoutons des relations entre des nœuds autres que la descendance, typiquement les transferts horizontaux ou les recombinaisons. La véritable difficulté ici est de traiter les cycles de façon à ce que la notion de temporalité ne soit pas perdue. Ce n'est pas forcément le cas, puisqu'à partir du moment où nous pouvons par exemple orienter les transferts horizontaux, nous avons la garantie de

connaître le sens du temps sur tout l'arbre. Notons que la matrice d'adjacence A que nous décrivons dans le premier article de cette thèse se construit de la même façon pour un graphe possédant des cycles, la principale différence venant du fait que dans ce cas, cette matrice n'est plus triangulaire, ce qui encore une fois ne change rien aux calculs, tant qu'aucun cycle n'est présent dans le réseau.

11.2 Enrichir le modèle de coévolution

Nous avons aussi vu que plusieurs biais liés directement à notre méthode sont dus au modèle de coévolution que nous avons développé, qui est probablement trop simple pour formaliser certains phénomènes que l'on peut rencontrer dans la nature. Deux axes permettraient de l'enrichir et ainsi couvrir plus de cas de figure.

11.2.1 Traiter plus d'une occurrence par branche

Le premier axe qui va dans ce sens est de modifier le modèle de façon à traiter plus d'une occurrence par évènement et par branche. Nous avons aussi vu qu'il n'est mathématiquement pas possible d'exprimer cette fonction "à la volée", c'est à dire en la construisant récursivement à partir de cas simples. Nous devons donc calculer cette fonction pour tous les cas de figure possibles, ce qui pose problème, puisqu'il existe un nombre infini de combinaisons possibles.

Nous pouvons en revanche enrichir cette fonction en considérant non pas une occurrence par évènement et par branche au maximum, mais par exemple deux ou trois, ce qui permettrait de couvrir plus de situations. Bien sûr, la fonction est dans ce cas beaucoup plus complexe, mais elle reste calculable analytiquement, et le principe reste le même. Encore une fois, si l'on considère les mutations comme étant rares dans la nature, alors cette modification laisserait de coté des cas très rares, probablement assez pour être négligés.

11.2.2 Décrire un modèle de coévolution état-dépendant

Enfin, la perspective qui est probablement la plus importante est de compléter ce modèle de coévolution en associant non pas deux taux (basal et augmenté) à chaque évènement, mais plutôt quatre. En effet, nous ne modélisons pour l'instant que l'effet d'un évènement sur tous les autres, sans considérer qu'une occurrence d'un évènement peut aussi avoir un

effet sur lui-même, comme dans le cas de l'évènement "perte du gène X". Ajouter deux états permettrait de formaliser cette rétroaction et nous donnerait ainsi accès à plus de phénomènes. Il faut malgré tout faire attention au fait que modéliser plus d'états implique nécessairement l'utilisation de plus de paramètres. Ces paramètres doivent être optimisés, donc nous devons aussi raisonner en termes de complexité et de temps de calcul.

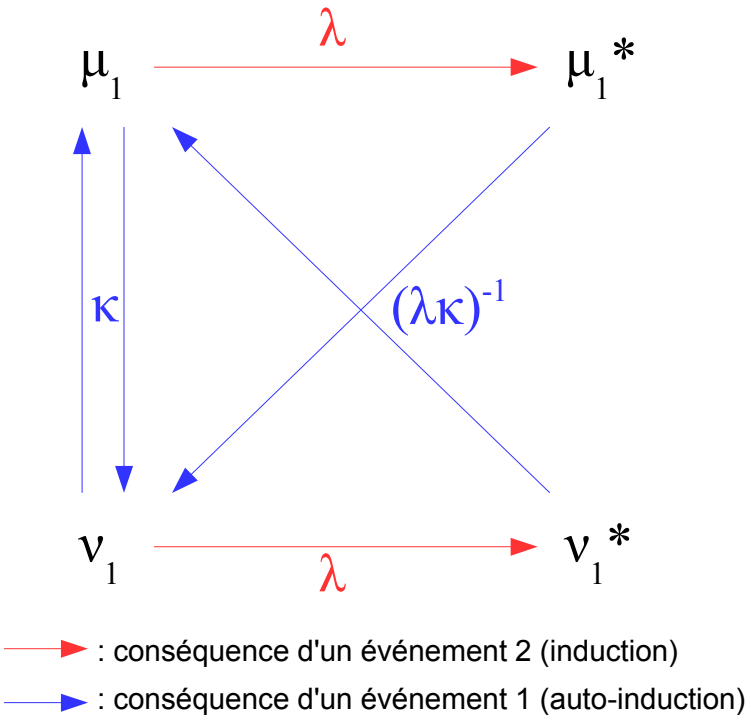
Une idée reviendrait ainsi par exemple à définir, par analogie avec les inductions λ_1 et λ_2 déjà décrits dans le modèle actuellement utilisé, ce que l'on pourrait appeler les "auto-inductions" κ_1 et κ_2 qui modéliseraient la notion d'état-dépendance, les évènements possédant deux états, et basculant d'un état à l'autre à chacune de leurs *propres* occurrences. Les règles du modèle, décrites pour l'évènement E_1 dans la figure 11.1, restent très proches des spécifications que nous avons déjà établies, et ainsi, le calcul de la fonction de vraisemblance, bien que plus complexe, resterait largement accessible. Il faut toutefois garder à l'esprit que manipuler une telle fonction à 6 paramètres au lieu de 4 demande un temps de calcul plus élevé, qu'il faudrait précisément déterminer.

11.3 Mieux comprendre les différentes méthodes

Une dernière perspective intéressante, mais qui s'inscrit dans un cadre plus large que le travail d'une thèse est la comparaison entre les différentes méthodes de détection de la coévolution. Nous avons vu qu'il en existait de nombreuses, pour certaines radicalement différentes les unes des autres. Nous avons commencé au cours de ce travail une ébauche de comparaison entre les méthodes d'information mutuelle, CoMap, notre méthode non-paramétrique et notre méthode mettant en jeu les calculs de vraisemblances, notamment sur des données simulées, mais aussi en les testant sur le même exemple biologique des interactions Watson-Crick dans l'ARN 16S chez les γ-entérobactéries.

Ce n'est bien entendu pas suffisant, en particulier pour les données synthétiques, ne serait-ce que parce que les simulations que nous avons menées se placent dans le cadre de notre modèle de coévolution, ce qui ne nous assure aucunement que toutes les méthodes sont traitées avec équité. Il faudrait en pratique plutôt développer un *benchmark*, ou banc d'essai, neutre vis à vis de toutes les méthodes, permettant de comparer les puissances des différentes méthodes. Par ailleurs, quels que soient les résultats de cette première comparaison, une question subsistera : les différentes méthodes capturent-elles le même signal de coévolution, ou sont-elles plutôt complémentaires, chacune étant adaptée à différents cas de figure ? Par exemple, toutes les méthodes sont-elles sensibles à des signaux de coévolution modérés (que

FIGURE 11.1 – **Modèle de coévolution état-dépendant**, spécification pour les taux de l'évènement E_1, et facteurs d'induction associés. Si l'évènement E_1 est associé à deux états A et B, alors μ_1 est le taux d'occurrence dans l'état A et ν_1 le taux dans l'état B.



nous décrivons ici comme générant des chronologies plutôt que des cooccurrences) ?

Une fois de plus, ce type de projet nécessite une collaboration élargie entre les différents chercheurs ayant construit et développé de telles méthodes, qui travailleraient de concert à l'élaboration d'un ensemble de cas qui constituerait ce *benchmark*.

Bibliographie

Agnarsson, I. et J. A. Miller. 2008. Is acctran better than deltran ? Cladistics 24 :1032–1038.

Alakbarli, F. 2001. A 13th-Century Darwin ? Tusi's Views on Evolution. Azerbaijan International Pages 48–49.

Barešić, A. et A. C. R. Martin. 2011. Compensated pathogenic deviations. Biomol Concepts 2 :281–92.

Bateson, W. 1909. Mendel's Principles of Heredity. Cambridge Univ. Press.

Brooks, D. R. 1979. Testing the context and extent of host-parasite coevolution. Systematic Biology 28 :299–307.

Brouillet, S., T. Valere, E. Ollivier, L. Marsan, et A. Vanet. 2010. Research co-lethality studied as an asset against viral drug escape : the hiv protease case. Biology Direct .

Carbone, A. et L. Dib. 2009. Co-evolution and Information Signals in Biological Sequences chap. Co-Evolution and Information Signals in Biological Sequences. Springer.

Cochran, W. G. 1954. Some methods for strengthening the common χ 2 tests. Biometrics 10 :417–451.

Cover, T. M. et J. A. Thomas. 2012. Elements of information theory. John Wiley & Sons.

Darwin, C. 1859. On the Origin of Species. 1 ed. John Murray.

Dib, L. et A. Carbone. 2012. Protein fragments : functional and structural roles of their coevolution networks. PLoS One 7 :e48124.

Dobzhansky, T. G. et al. 1941. Genetics and the Origin of Species. Columbia University Press.

Doty, P., H. Boedtker, J. R. Fresco, R. Haselkorn, et M. Litt. 1959. Secondary structure in ribonucleic acids. Proc Natl Acad Sci U S A 45 :482–99.

Dutheil, J. et N. Galtier. 2007. Detecting groups of coevolving positions in a molecule : a clustering approach. BMC Evol Biol 7 :242.

Dutheil, J., T. Pupko, A. Jean-Marie, et N. Galtier. 2005. A model-based approach for detecting coevolving positions in a molecule. Mol Biol Evol 22 :1919–28.

Dutheil, J. Y. 2012. Detecting coevolving positions in a molecule : why and how to account for phylogeny. Briefings in bioinformatics 13 :228–243.

Ehrlich, P. R. et P. H. Raven. 1964. Butterflies and plants : a study in coevolution. Evolution Pages 586–608.

Engelen, S., L. A. Trojan, S. Sacquin-Mora, R. Lavery, et A. Carbone. 2009. Joint evolutionary trees : a large-scale method to predict protein interfaces based on sequence sampling. PLoS Comput Biol 5 :e1000267.

Felsenstein, J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous characters. American journal of human genetics 25 :471.

Felsenstein, J. 1981. Evolutionary trees from dna sequences : a maximum likelihood approach. J Mol Evol 17 :368–76.

Felsenstein, J. 1983. Statistical inference of phylogenies. Journal of the Royal Statistical Society. Series A (General) Pages 246–272.

Felsenstein, J. 2004. Inferring Phylogenies. Sinauer Associates, Inc.

Fisher, R. A. 1930. The genetical theory of natural selection : a complete variorum edition. Oxford University Press.

Fraser, H. B., A. E. Hirsh, D. P. Wall, et M. B. Eisen. 2004. Coevolution of gene expression among interacting proteins. Proceedings of the National Academy of Sciences of the United States of America 101 :9033–9038.

Guindon, S., J.-F. Dufayard, V. Lefort, M. Anisimova, W. Hordijk, et O. Gascuel. 2010. New algorithms and methods to estimate maximum-likelihood phylogenies : assessing the performance of phyml 3.0. Syst Biol 59 :307–21.

Haldane, J. B. S. 1949. Suggestions as to quantitative measurement of rates of evolution. Evolution Pages 51–56.

Huxley, J. et al. 1942. Evolution. the modern synthesis. Evolution. The Modern Synthesis. .

Kauffman, S. et S. Levin. 1987. Towards a general theory of adaptive walks on rugged landscapes. Journal of theoretical Biology 128 :11–45.

Kern, A. D. et F. A. Kondrashov. 2004. Mechanisms and convergence of compensatory evolution in mammalian mitochondrial trnas. Nature genetics 36 :1207–1212.

Kimura, M. 1968. Evolutionary rate at the molecular level. Nature 217 :624–6.

King, J. L. et T. H. Jukes. 1969. Non-darwinian evolution. Science 164 :788–98.

Kondrashov, A. S., S. Sunyaev, et F. A. Kondrashov. 2002. Dobzhansky-muller incompatibilities in protein evolution. Proc Natl Acad Sci U S A 99 :14878–83.

Kulathinal, R. J., B. R. Bettencourt, et D. L. Hartl. 2004. Compensated deleterious mutations in insect genomes. Science 306 :1553–4.

Lamarck, J.-B.-P. 1809. Philosophie zoologique.

Lartillot, N. 2006. Conjugate gibbs sampling for bayesian phylogenetic models. Journal of computational biology 13 :1701–1722.

Leontis, N. B. et E. Westhof. 1998. A common motif organizes the structure of multi-helix loops in 16 s and 23 s ribosomal rnas. Journal of molecular biology 283 :571–583.

Leontis, N. B. et E. Westhof. 2001. Geometric nomenclature and classification of rna base pairs. RNA 7 :499–512.

Mangeart, M. 1843. Du Jour Natal, Censorinus, traduction de M.J. Mangeart. Paris.

Martin, L. C., G. B. Gloor, S. D. Dunn, et L. M. Wahl. 2005. Using information theory to search for co-evolving residues in proteins. Bioinformatics 21 :4116–24.

Mateiu, L. et B. Rannala. 2006. Inferring complex dna substitution processes on phylogenies using uniformization and data augmentation. Systematic biology 55 :259–269.

Metropolis, N., A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, et E. Teller. 1953. Equation of state calculations by fast computing machines. The journal of chemical physics 21 :1087–1092.

Metropolis, N. et S. Ulam. 1949. The monte carlo method. Journal of the American statistical association 44 :335–341.

Moore, P. B. 1999. Structural motifs in rna. Annual review of biochemistry 68 :287–300.

Pagel, M. 1994. Detecting correlated evolution on phylogenies : a general method for the

comparative analysis of discrete characters. Proceedings of the Royal Society of London B : Biological Sciences 255 :37–45.

Pagel, M. 1999. The maximum likelihood approach to reconstructing ancestral character states of discrete characters on phylogenies. Systematic Biology 48 :612–622.

Pagel, M. 2000. Phylogenetic–evolutionary approaches to bioinformatics. Brief Bioinform 1 :117–30.

Pagel, M. et A. Meade. 2006. Bayesian analysis of correlated evolution of discrete characters by reversible-jump markov chain monte carlo. Am Nat 167 :808–25.

Pagel, M., A. Meade, et D. Barker. 2004. Bayesian estimation of ancestral character states on phylogenies. Syst Biol 53 :673–84.

Phillips, P. C. 2008. Epistasis–the essential role of gene interactions in the structure and evolution of genetic systems. Nat Rev Genet 9 :855–67.

Pollock, D. D., W. R. Taylor, et N. Goldman. 1999. Coevolving protein residues : maximum likelihood identification and relationship to structure. J Mol Biol 287 :187–98.

Ravasi, T., H. Suzuki, C. V. Cannistraci, S. Katayama, V. B. Bajic, K. Tan, A. Akalin, S. Schmeier, M. Kanamori-Katayama, N. Bertin, et al. 2010. An atlas of combinatorial transcriptional regulation in mouse and man. Cell 140 :744–752.

Saitou, N. et M. Nei. 1987. The neighbor-joining method : a new method for reconstructing phylogenetic trees. Molecular biology and evolution 4 :406–425.

Saporta, G. 2006. Probabilités, analyse des données et statistique. Editions Technip.

Shindyalov, I. N., N. A. Kolchanov, et C. Sander. 1994. Can three-dimensional contacts in protein structures be predicted by analysis of correlated mutations ? Protein Eng 7 :349–58.

Szendro, I. G., M. F. Schenk, J. Franke, J. Krug, et J. A. G. De Visser. 2013. Quantitative analyses of empirical fitness landscapes. Journal of Statistical Mechanics : Theory and Experiment 2013 :P01005.

Watson, J. D., F. H. Crick, et al. 1953. Molecular structure of nucleic acids. Nature 171 :737–738.

Weinreich, D. M. et L. Chao. 2005. Rapid evolutionary escape by large populations from local fitness peaks is likely in nature. Evolution 59 :1175–82.

Weinreich, D. M., N. F. Delaney, M. A. Depristo, et D. L. Hartl. 2006. Darwinian evolution can follow only very few mutational paths to fitter proteins. Science 312 :111–4.

Weinreich, D. M., Y. Lan, C. S. Wylie, et R. B. Heckendorn. 2013. Should evolutionary geneticists worry about higher-order epistasis ? Current opinion in genetics & development 23 :700–707.

Weinreich, D. M., R. A. Watson, et L. Chao. 2005. Perspective : Sign epistasis and genetic constraint on evolutionary trajectories. Evolution 59 :1165–74.

Wittkopp, P. J., B. K. Haerum, et A. G. Clark. 2004. Evolutionary changes in cis and trans gene regulation. Nature 430 :85–88.

Wright, S. 1931. Evolution in mendelian populations. Genetics 16 :97–159.

Zirkle, C. 1941. Natural Selection before the "Origin of Species". Proceedings of the American Philosophical Society Pages 71–123.

Zuckerkandl, E. et L. Pauling. 1962. Molecular disease, evolution and genetic heterogeneity. Horizons in Biochemistry .

Zuckerkandl, E. et L. Pauling. 1965. Evolutionary divergence and convergence in proteins. Evolving genes and proteins 97 :97–166.