UNIVERSITÉ DU QUÉBEC À MONTRÉAL

ALGORITHME POUR LA RECONSTRUCTION DE SÉQUENCES ANCESTRALES EN INCLUANT LES ÉVÉNEMENTS DE TRANSFERTS HORIZONTAUX DE GÈNES

MÉMOIRE

PRÉSENTÉ

COMME EXIGENCE PARTIELLE

DE LA MAÎTRISE EN INFORMATIQUE

PAR

MOUHAMADOU MOUSTAPHA MBAYE

JUILLET 2019

REMERCIEMENTS

La réalisation de ce mémoire a été possible grâce au soutien de plusieurs personnes aux quelles je voudrais témoigner toute ma reconnaissance.

D'abord, je tiens à remercier chaleureusement mon directeur de recherche, Pr Abdoulaye Baniré Diallo, pour son encadrement, son soutien, sa disponibilité et surtout ses précieux conseils.

J'adresse mes sincères remerciements à tous mes collègues du laboratoire de bio-informatique et à toutes les personnes qui, par leurs paroles, leurs écrits, leurs conseils et leurs critiques, ont guidé mes réflexions, plus particulièrement Nadia Tahiri.

Je tiens à remercier ma famille, mon père pour les valeurs qu'il m'a transmises. Je remercie tous mes frères et soeurs Cheikh Mbaye, Khadim Mbaye pour leurs soutiens.

Je tiens à remercier tous mes amis qui ont rendu mon expérience plus riche : Ndiaye Diop, Gor Mack, Mouhamad Dieye pour leur support constant.

Enfin, ce travail est dédié à ma mère, qui m'a toujours poussé et motivé. J'espère qu'elle appréciera cet humble geste comme preuve de reconnaissance de la part d'un fils.

Mouhamadou Moustapha Mbaye

TABLE DES MATIÈRES

LISTE DES FIGURES

LISTE DES TABLEAUX

RÉSUMÉ

L'analyse comparative des séquences génomiques nous permet d'utiliser des outils et algorithmes bio-informatiques pour la reconstruction des séquences génomiques ancestrales. Ce dernier consiste à identifier les séquences du passé qui ont permis d'obtenir les descendants observés en ce moment. Le problème de la reconstruction de séquences ancestrales a été abordé dans le passé en tenant compte des phénomènes de substitution, d'insertion, de délétion et de duplication. Cependant, très peu d'études ont été consacrées à l'inclusion des transferts horizontaux de gènes (THG). Or ces derniers sont prépondérants dans le règne microbien. Dans ce mémoire, nous nous intéressons à l'inclusion de ce dernier dans un cadre de maximum de vraisemblance. Nous utilisons un type de structure de données qui est la combinaison d'un modèle de Markov caché et d'un arbre phylogénétique permettant de capturer les dimensions temporelles et spatiales. Ainsi la nouvelle procédure de reconstruction des génomes ancestraux impliquera la reconstruction de séquences pour chaque bloc aligné incluant l'ensemble des substitutions, des insertions, des suppressions et des transferts horizontaux de gènes qui peuvent produire un ensemble de données de séquences existantes en fonction d'un arbre phylogénétique.

Cette méthode est basée sur la réconciliation des topologies entre un arbre d'espèces fixé et les arbres des différents blocs alignés (considéré comme des arbres de gènes). Elle exploite la distance de Robinson et Foulds (RF) comme critère métrique de comparaison topologique.

Le cadre construit a été testé à travers des simulations et les résultats préliminaires montrent l'importance de cette inclusion pour la compréhension des scénarios d'évolution.

Mots clés : Transfert horizontal de gènes - Algorithmes - Reconstruction de génomes ancestraux - Insertions et suppressions - Maximum de vraisemblance - Arbre phylogénétique.

INTRODUCTION

0.1 Introduction

La reconstruction est l'étude des relations évolutives entre individus, populations ou espèces et leurs ancêtres. L'hypothèse d'une parenté d'espèces représentée sous la forme d'un arbre nommé arbre phylogénétique remonte au 19e siècle. Depuis quelques années, de plus en plus de génomes d'espèces sont séquencés en totalité ou partiellement. Maintenant, il est intéressant de savoir les rôles de chaque région du génome. Pour cela, il est important de comprendre les évènements qui sont arrivés dans le passé à travers les principaux scénarios d'évolution. Ainsi, la conservation des séquences contemporaines permet par exemple, de formuler des hypothèses sur la fonction des protéines (Sunyaev *et al.*, 2001; Ng et Henikoff, 2002) et d'identifier les séquences codantes ou sous sélection négative.

Beaucoup d'efforts sont aussi fournis pour comprendre l'évolution des nouveaux traits humains, tous ces efforts dépendent directement ou indirectement de la reconstruction de l'histoire de l'évolution des bases dans le génome humain, et donc de la reconstruction de séquences des génomes de nos lointains ancêtres (Blanchette *et al.*, 2004a).

L'évolution des génomes provient d'un processus impliquant des évènements de mutations à plusieurs échelles : incluant des insertions, des délétions, des substitutions de nucléotides, des duplications et des transferts horizontaux de gènes (Bouchard-Côté et Jordan, 2013). D'un côté, les méthodes de reconstruction phylogénétique se sont concentrées sur l'évolution des séquences, certaines intégrant

l'évolution des familles de gènes. D'un autre côté, les réarrangements chromosomiques ont été très largement étudiés, ce qui a donné naissance à de nombreux modèles d'évolution décrivant l'architecture des génomes. Ces deux voies de modélisation se sont rarement rencontrées jusqu'à ses dernières années (Semeria, 2015).

Cependant, le problème de la reconstruction des scénarios d'évolution n'est certainement pas nouveau, il a été appliqué dans plusieurs contextes. Par exemple, la méthode de parcimonie (Fitch, 1971; Swofford, 1993) a été utilisée pour déduire des séquences d'acides aminés ancestrales du lysozyme en utilisant les substitutions. La difficulté du problème de reconstruction ancestrale de séquence est due en grande partie au fait que des insertions et des délétions affectent souvent plusieurs nucléotides consécutifs. Cela implique que les colonnes de l'alignement ne peuvent pas être traitées de façon autonome, par opposition au problème du maximum de vraisemblance pour les substitutions (Felsenstein, 1981; Blanchette *et al.*, 2004b).

Par ailleurs, l'échange de matériels génétiques entre des espèces qui ne partagent pas de relation de parenté est connu depuis plusieurs années sous le phénomène de transferts horizontaux de gènes. D'après la littérature scientifique, il est raisonnable de considérer que le transfert horizontal de gènes est un mécanisme habituel d'adaptation des organismes biologiques aux stress environnementaux. Si deux espèces sont issues d'un ancêtre commun qui a reçu un gène par transfert, leurs génomes vont garder une trace de ce transfert commun à leur ancêtre, et donc conserver une preuve d'une ascendance partagée. Le transfert horizontal de gène (THG) aussi appelé « Transfert latéral » est un phénomène très bien décrit chez les bactéries sous trois mécanismes que sont la transformation, la conjugaison et la transduction. En particulier, de nombreuses bactéries deviennent résistantes à des antibiotiques en acquérant les gènes de résistance à partir d'espèces éloignées d'un point de vue phylogénétique (Prescott *et al.*, 2018). Dans ce mémoire, nous abordons le problème de la reconstruction de séquences ancestrales dans un cadre

qui inclut les transferts horizontaux de gènes. Pour mieux cerner la problématique du mémoire, nous avons divisé ce dernier en quatre chapitres :

Dans le premier chapitre, nous introduirons le sujet et ses motivations. Puis nous détaillerons les objectifs du travail ainsi que notre contribution.

Dans le deuxième chapitre, nous présenterons une revue de la littérature qui est associée au domaine de la reconstruction ancestrale et du transfert horizontal de gène.

Le troisième chapitre introduira l'algorithme permettant de résoudre la reconstruction ancestrale de séquence, basée sur la distance de Robinson et Foulds (RF) pour la détection de transfert horizontal de gène.

Nous expliciterons la procédure d'identification du nombre optimal de transferts dans chaque paire d'arbres, ainsi que la manière dont la probabilité de transition entre les différentes topologies d'arbres est établie. Enfin, nous finirons par examiner les résultats.

Dans le quatrième chapitre, nous effectuerons une conclusion générale et une synthèse du travail accompli tout en mettant en évidence les améliorations possibles d'une nouvelle version.

0.2 Motivation et Objectif

Les séquences observées actuelles des espèces contemporaines peuvent être utilisées pour reconstruire des séquences ancestrales sur la base d'un modèle d'évolution de gène. Une telle connaissance des séquences ancestrales est utile pour comprendre les processus évolutifs ainsi que les aspects fonctionnels de composants tels que les familles de gènes qui composent ces espèces. Cependant, il existe un besoin urgent de procédures d'inférence capables de gérer des données provenant d'un

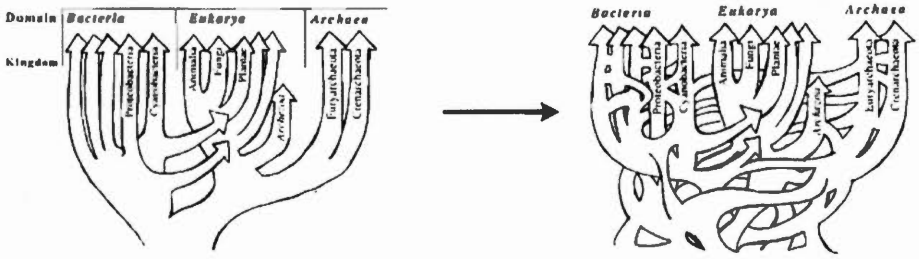


Figure 0.1: L'évolution de la perception du monde procaryote (Doolittle, 1999).

grand nombre de taxons et pouvant fournir des inférences pour des états ancestraux et des paramètres évolutifs sur des périodes de plus en plus grandes. Il existe deux grandes approches de reconstruction de séquences ancestrales : les méthodes basées sur le maximum de vraisemblance et les méthodes de parcimonie (Farris, 1970; Pupko *et al.*, 2000). Dans le cadre de ces deux approches, les principaux modèles actuels se consacrent exclusivement aux phénomènes de substitution, d'insertion et de délétion de fragments géniques. Or les mécanismes tels que les transferts horizontaux de gènes ont un impact réel dans les processus d'adaptation des bactéries et archéobactéries (BOC, 2012; Philippe et Douady, 2003). Ce mécanisme constitue le moyen d'adaptation le plus connu.

0.3 Contribution

En utilisant les séquences (ADN, ARN ...) existantes et les relations phylogénétiques entre elles, il est possible de déduire les séquences ancestrales les plus plausibles dont elles sont issues. Les types fondamentaux de mutations qui modifient les séquences biologiques sont la substitution, l'insertion et la délétion. Bien que les insertions et les délétions jouent un rôle important dans l'évolution des séquences (Diallo *et al.*, 2007), l'inférence phylogénétique est souvent effectuée en

tenant compte uniquement des substitutions. De plus, les analyses basées sur un seul alignement multiple peuvent être trompeuses et les méthodes d'alignement multiples ont tendance à apporter plus de variations à l'analyse phylogénétique que les méthodes de construction d'arbres (Liò et Goldman, 1998). Or ses deux étapes sont importantes pour la reconstruction ancestrale. Par conséquent, afin d'établir une bonne reconstruction ancestrale il est souhaitable d'incorporer des insertions et des délétions dans l'analyse phylogénétique, et la co-estimation de la phylogénie et l'alignement à partir de leur distribution postérieure conjointe (Liu *et al.*, 2011). Notre objectif est de créer un algorithme avec un modèle d'insertion, de délétion et de substitution pouvant être couplé à un modèle de transfert horizontal de gène (Westesson *et al.*,).

Nous fournissons une description détaillée dans le chapitre 3 et étudierons la possibilité de faire évoluer les probabilités de transition associées à un modèle probabiliste.

CHAPITRE I

REVUE DE LA LITTÉRATURE

1.1 Séquences biologiques

Après la découverte de l'ADN, plusieurs équipes de chercheurs ont tenté des milliers d'études pour reconstruire l'évolution des êtres vivants au cours du temps, et étudier la différence de composition des séquences entre les espèces afin de mieux connaître l'histoire évolutive. Darwin (darwin 1859) a proposé une théorie solide de l'origine et de la diversité des espèces dans laquelle il conclut que les mutations d'ADN peuvent aboutir à la formation d'une nouvelle espèce.

Nucléotide : Un nucléotide est une molécule organique qui est l'élément de base d'un acide nucléique tel que l'ADN ou l'ARN. Chaque nucléotide renferme dans sa structure une base azotée qui l'identifie. Il existe 4 types de nucléotides qui sont l'adénine (A), la guanine (G), la cytosine (C) et la thymine (T). Pour les ARN, la thymine est remplacée par l'uracile (U), parmi ces quatre nucléotides, deux sont des purines (adénine et guanine) et les deux autres sont des pyrimidines (cytosine et thymine (uracile pour l'ARN)).

Séquence : Une séquence d'acide nucléique ADN ou ARN est formée d'une succession de nucléotides. Cette succession contient l'information génétique portée par ces poly-nucléotides. Il nécessite aussi une interprétation supplémentaire pour

étudier les instructions et pour en comprendre le sens biologique.

ADN : L'ADN est une macromolécule biologique au coeur de toutes nos cellules à l'exception des globules rouges ainsi que chez de nombreux virus. Elle est une molécule très longue, composée d'une succession de nucléotides accrochés les uns aux autres par des liaisons. L'ADN contient toute l'information génétique, elle est aussi appelée génome.

ARN : L'ARN est chimiquement très proche de l'ADN, il est en général synthétisé dans les cellules à partir d'une matrice d'ADN dont il est une copie. Il est généralement utilisé dans les cellules comme intermédiaire des gènes pour fabriquer les protéines.

ADN	**A**	**C**	**C**	**T**	**G**	**C**	**C**	**A**	**G**	**A**	**T**
ARN	**A**	**C**	**C**	**U**	**G**	**C**	**C**	**A**	**G**	**A**	**U**

Figure 1.1: Séquences d'ADN et d'ARN

1.2 Alignement de séquence

La comparaison des séquences nécessite la réalisation d'un alignement. Un alignement de séquences est une superposition deux à deux entre les caractères de la première séquence et ceux de la deuxième de manière à pouvoir comparer les caractères qui proviennent de la même histoire évolutive. Cette opération consiste à disposer les unes en dessous des autres des portions de séquences similaires tout en minimisant leurs différences (on peut aligner entre eux des gènes d'une même famille multigénique ou des gènes d'espèces différentes).

Les mutations sont des processus qui affectent la séquence d'ADN à une petite échelle en modifiant le contenu d'un nucléotide.

Ces mutations sont classées de trois formes :

1. la substitution qui remplace un nucléotide par un autre ;
2. l'insertion qui insère un nucléotide dans la séquence d'ADN ;
3. la délétion qui supprime un nucléotide de la séquence d'ADN

Ces événements peuvent aussi faire intervenir des écarts (*gaps*). Lorsque nous exécutons un logiciel d'alignement de séquences, nous remarquons que le nombre d'écarts est limité (Larkin *et al.*, 2007; Katoh et Standley, 2013; Edgar, 2004).

L'alignement des séquences définit la correspondance entre leurs nucléotides et les séquences alignées constituent les données de la reconstruction phylogénétique. L'alignement est donc une étape cruciale qui conditionne la qualité de la reconstruction phylogénétique. Une fois les séquences alignées, une méthode de reconstruction d'arbres phylogénétiques peut être appliquée pour obtenir un arbre qui reflète une structure visuelle donnée.

S1	**A**	**C**	**C**	**T**	**-**	**-**	**T**	**G**	**G**	**C**	**C**
S2	**T**	**T**	**G**	**A**	**A**	**C**	**-**	**-**	**C**	**G**	**G**
S3	**-**	**T**	**T**	**G**	**-**	**T**	**T**	**A**	**G**	**C**	**C**
S4	**T**	**A**	**-**	**-**	**G**	**A**	**A**	**C**	**T**	**A**	**A**

Figure 1.2: Alignement de séquences multiples de quatre séquences, avec des gaps

1.3 Arbre phylogénétique

La phylogénie est un domaine d'étude associé au développement des êtres vivants. Il a fallu attendre l'arrivée des données moléculaires, vers la fin des années 1970

et au début des années 1980, pour accélérer l'identification et la quantification d'importants mécanismes d'évolution, tels que l'échange de matériel génétique entre des individus appartenant à différentes espèces. La phylogenèse étudie la reconstruction de l'histoire évolutive des êtres vivants. Le terme phylogenèse a été introduit par Haeckel en 1860 (Darlu et Tassy, 1993).

Un arbre phylogénétique est composé de quatre éléments principaux :

1. les nœuds de l'arbre représentent l'ancêtre commun de ses descendants. Le nom qu'il porte est celui du clade formé des groupes frères qui lui appartiennent ;
2. les branches (ou arêtes) définissent les relations entre les taxons en termes de descendance. À chacune des branches, on peut associer une mesure, par exemple une distance, génétique ou autre, une durée, une quantité d'évolution ou un nombre de mutations ;
3. les nœuds internes associés à des ancêtres virtuels ;
4. la racine représente l'ancêtre commun de toutes les espèces considérées.

L'arbre peut être enraciné ou pas, selon qu'on parvienne à identifier l'ancêtre commun à toutes les feuilles. Lorsque l'ancêtre commun de toutes les espèces de l'arbre est identifié, l'arbre est enraciné. Il est orienté dans le sens du temps d'évolution des espèces. L'arbre enraciné permet de définir une relation de descendance entre deux nœuds successifs. L'arbre non enraciné (*unrooted tree*) est un graphe connexe non cyclique, il n'existe donc pas une seule boucle, c'est-à-dire qu'un seul et unique chemin permet de passer d'un sommet à un autre. Généralement, il est impossible d'établir l'origine de diversification des espèces au niveau d'un arbre non enraciné (Dobson, 1974). L'arbre non enraciné est habituellement utilisé pour la classification d'un groupe d'espèces sans considérer le sens d'évolution.

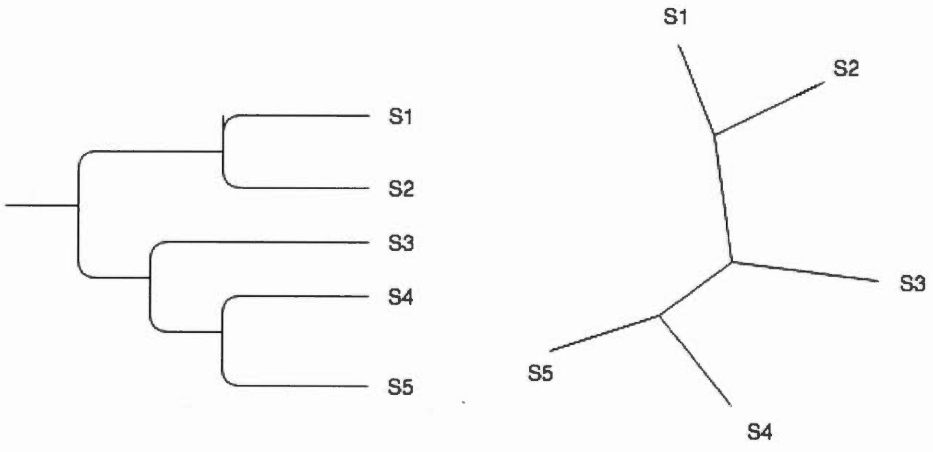


Figure 1.3: Arbres enracinées (à gauche) et non enracinées (à droite).

Les arbres phylogénétiques peuvent être inférés à partir d'un ensemble varié de données telles que les séquences moléculaires, les fréquences des gènes, les traits quantitatifs et l'ordre des gènes. À partir d'analyses comparatives des séquences nucléotidiques des gènes codants pour les ARNs ribosomiques et de plusieurs protéines, les phylogénéticiens moléculaires ont construit un "arbre universel de la vie" en le prenant comme base pour une classification hiérarchique "naturelle" de tous les êtres vivants(Doolittle, 1999).

On peut avoir deux formes de représentations de longueurs de branches d'arbres :

1. représentation sans échelles : Les longueurs de branches ne sont pas proportionnelles au nombre de changements évolutifs ;

2. représentation avec échelles : les longueurs de branches sont proportionnelles au nombre de substitutions ou nombre de substitutions/sites qui se sont produites le long de la branche.

1.4 Reconstruction des arbres phylogénétiques

La reconstruction d'arbres phylogénétiques s'appuie sur l'aspect des différentes modifications réalisées sur certains critères à travers la descendance, son but est de construire l'arbre (topologie) et définir les longueurs des arêtes. Ces longueurs approximent le nombre d'événements mutationnels qui se sont produits le long des arêtes. Un arbre muni de longueurs de branches positives définit une mesure de distances particulières appelée distance arborée ou distance additive d'arbre (Saitou et Nei, 1987). Lors d'une reconstruction d'arbres phylogénétiques, la première étape, nommée "alignement", consiste à mettre en correspondance les sites des séquences de manière à pouvoir comparer ce qui est comparable (Stamatakis, 2006). Les séquences utilisées pour la reconstruction peuvent être de l'ADN ou de l'ARN. Cette étape est suivie par la reconstruction de l'arbre. Ici, nous décrivons par la suite le principe des méthodes de distances, de parcimonie et du maximum de vraisemblance.

1.4.1 Approche basée sur les distances

Les méthodes de distances aussi appelées méthode phénétique reconstruisent la phylogénie d'un ensemble de taxons en s'intéressant à leur degré de ressemblance à partir de l'ensemble des distances évolutives qui séparent chaque couple de séquences. Cette méthode se limite sur le calcul de la distance évolutive entre les différentes séquences puis calcule une instance d'arbre par des algorithmes d'agglomération. L'approche basée sur les distances utilise une matrice d'estimation de la distance évolutive appelée matrice de dissimilarité, introduite par (Cavalli-Sforza et Edwards, 1967).

1.4.1.1 Matrice de dissimilarité

Un critère de ressemblance est essentiel pour regrouper les individus qui se ressemblent. Dans le domaine biologique, une matrice est obtenue en comparant des séquences deux à deux. Cette méthode calcule la distance de dissimilarité entre chaque paire d'espèces avant d'inférer l'arbre phylogénétique dont les longueurs des branches se rapprochent le mieux aux distances calculées. La somme des longueurs de branches entre deux espèces A et B, de l'arbre phylogénétique T est appelée distance évolutive entre les espèces A et B. La distance $\delta(A, B)$ entre deux sommets A et B dans un arbre est définie comme étant la somme de toutes les longueurs d'arêtes du chemin unique menant de A à B. La distance observée $\delta(AB)$ est calculée à partir des séquences alignées ou d'autres types d'informations. Pour les séquences nucléotidiques, toutes les substitutions observables dans l'alignement de deux séquences peuvent être comptabilisées avec un même poids. Puis le nombre de substitutions est divisé par le nombre total des sites comparés. La distance observée obtenue par une simple comparaison des séquences est considérée comme non corrigée, car elle comporte deux biais majeurs.

Tableau 1.1: Exemple d'alignement de séquences multiples pour cinq espèces.

S1	AGCTAT
S2	AGAATC
S3	AACCGT
S4	TGGAAG
S5	CCGGAT

À partir de ces alignements, nous pouvons calculer la matrice de dissimilarités en appliquant une transformation Séquences Distances appropriée.

Tableau 1.2: Matrice de dissimilarités pour cinq espèces.

	S1	S2	S3	S4	S5
S1	0				
S2	8	0			
S3	8	6	0		
S4	7	7	6	0	
S5	9	5	7	7	0

Cette matrice de dissimilarité de cinq espèces a été construite à partir des séquences du tableau 2.1. Une fois ces mesures établies, les méthodes de reconstruction d'arbre par agglomération peuvent être appliquées.

1.4.1.2 Méthodes d'agglomération

Ces méthodes choisissent la paire de nœuds à agglomérer à partir d'une matrice de distances. Pour rechercher l'arbre phylogénétique ayant les longueurs de branches qui se rapprochent au mieux des distances mesurées, on peut utiliser plusieurs méthodes :

Les méthodes ADDTREE (Orter, 1982) et Neighbor Joining (NJ) (Saitou et Nei, 1987). Le critère d'agglomération utilisé par ADDTREE repose sur la condition des quatre points. Sa tâche consiste à sélectionner la configuration la plus appropriée sur la base d'une mesure de dissemblance. Les distances évolutives dont on dispose ne sont que des approximations des distances réelles.

Les méthodes de reconstruction d'arbre par agglomération, Neighbor Joining, sont les méthodes les plus fréquemment utilisées pour l'inférence d'arbres phylogénétiques à partir des distances évolutives. Elles reconstruisent la structure d'un arbre

phylogénétique en commençant par un arbre en étoile qui contient n feuilles associées aux objets et n-1 branches. À la différence des méthodes précédentes, elle n'impose pas aux distances estimées d'être ultra-métriques. L'hypothèse évolutive d'horloge moléculaire n'est donc pas posée. En revanche, les distances doivent être métriques et additives pour avoir l'assurance d'obtenir l'arbre de longueurs minimum. C'est-à-dire l'arbre dont la somme des longueurs estimées est minimale, et qui permet une estimation correcte des longueurs des branches (Darlu et Tassy, 1993).

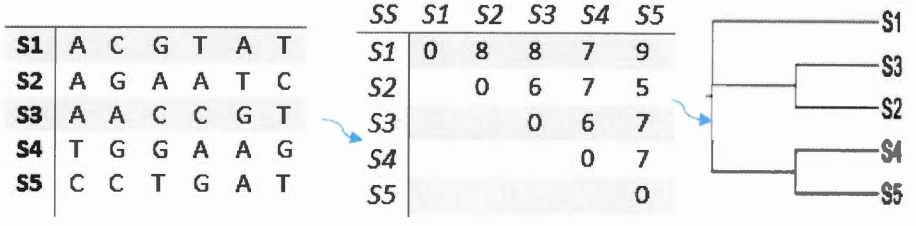


Figure 1.4: Processus d'inférence d'arbres phylogénétiques par une méthode de distances

1.4.2 Arbre parcimonieux

Les méthodes de parcimonie sont des méthodes statistiques non paramétriques basées sur le principe de maximum de parcimonie. Elles sont notamment utilisées pour l'inférence phylogénétique. La meilleure hypothèse pour expliquer un processus est celle qui fait appel au plus petit nombre d'événements. Il est généralement impossible d'examiner l'ensemble des topologies d'arbres, car leur nombre croît exponentiellement avec le nombre d'UEs (feuilles). Le recours à des heuristiques est donc indispensable, ces derniers permettent de « parcourir » la fonction à minimiser. Pour la parcimonie, l'objectif est de trouver la topologie qui minimise le nombre d'opérations permettant de partir d'une séquence à une autre. Cette recherche s'achève lorsqu'un extremum est atteint. Le ou les arbres obtenus cor-

respondent alors à des extremums locaux de la fonction. Par ailleurs, il existe plusieurs stratégies de réarrangement ou de transformation de branches qui sont :

- le réarrangement local (NNI) : Cet étape est l'échange de la plus proche voisin (NNI), pour réorganiser la topologie locale des arbres dans les régions de supports faibles ;
- le réarrangement global (SPR) : Élagage et regreffage de sous-arbres (SPR). Une branche de l'arbre est coupée et insérée dans une autre branche pour créer une topologie alternative ;
- la bissection et reconnexion d'arbres (TBR) : Bissection et reconnexion des arbres (TBR). Une branche est coupée pour former deux sous-arbres. Une nouvelle topologie est créée en reconnectant les branches de chaque sous-arbre ;
- les fusions d'arbres (tree-fusing) : La fusion d'arbres permute deux solutions entre deux arbres par ailleurs quasi optimaux ;
- les algorithmes génétiques : est définie pour permuter les sous-arbres à score élevé en arbres principaux à score élevé.

Pour rechercher l'arbre le plus parcimonieux, plusieurs approches peuvent être utilisées. Lorsque le nombre de taxons est inférieur à 10, il est possible d'effectuer une recherche exhaustive. *Ranwez et al.* (Ranwez, 2002) propose une méthode simple de calcul de parcimonies où l'on considère quatre séquences de deux nucléotides. Dans la figure 2.5, les substitutions requises sont indiquées, les solutions optimales nécessitent quatre substitutions. Les solutions optimales (1) et (2) sont différenciées par le choix des nucléotides ancestraux.

Le calcul de la parcimonie d'un arbre dont la topologie est fixée se fait en parcourant une seule fois l'arbre. Pour cela, on choisit arbitrairement une branche de l'arbre sur laquelle se trouve la racine r de l'arbre. On oriente ainsi l'arbre étudié et chacun de ses nœuds internes possèdent alors deux nœuds fils. En fixant la longueur d'un arc dans l'arbre avec une distance de Hamming, on peut définir le score de parcimonie d'un arbre de séquences comme la somme des longueurs des arcs.

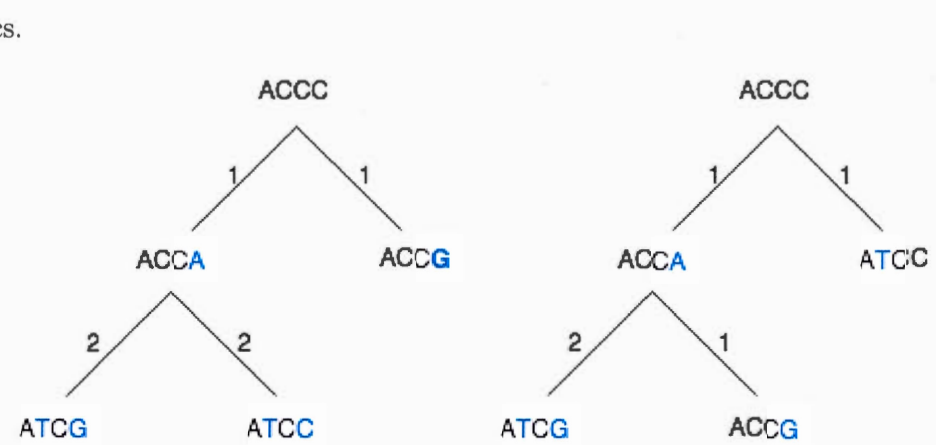


Figure 1.5: Deux histoires évolutives d'arbre ayant un score de parcimonie différent

1.4.3 Reconstruction d'arbre par Maximum de vraisemblance

L'approche de maximum de vraisemblance est utilisée par beaucoup de chercheur pour estimer des phylogénies. DNAML et PHYML sont des algorithmes rapides et précis, proposés par (Felsenstein et Felenstein, 2004; Guindon *et al.*, 2005) pour estimer les phylogénies du maximum de vraisemblance à partir de vastes ensembles de données de séquences d'ADN et de protéines. L'approche de maximum de vraisemblance introduite par (Strimmer et Von Haeseler, 1996), utilise des arbres basés sur des données de séquence d'ADN ou d'acides aminés. Cette

méthode applique la reconstruction par arbre de vraisemblance maximale à tous les quartets possibles pouvant être formés à partir de n séquences. Les arbres du quartet ne servent que de points de départ pour reconstruire un ensemble optimal d'arbres à n taxons. Cette méthode compare les arbres phylogénétiques possibles sur la base de leur habileté à prédire les données observées. L'application répétée de l'étape déroutante permet d'attribuer des valeurs de fiabilité aux regroupements dans l'arbre de puzzle final, un arbre de consensus construit à partir de tous les arbres intermédiaires. La vraisemblance est la probabilité conditionnelle d'observer les données sous un modèle particulier. Étant donné un modèle qui spécifie les probabilités d'observer différents événements. Pour une vraisemblance L l'obtention des données observées peut être calculée en suivant ce processus avec les hypothèses suivant : Paramètres du modèle Θ , topologie T et longueurs de branches L.

1. considérer une topologie, un site (les paramètres du modèle d'évolution des séquences) et un ensemble de longueurs de branches ;

2. calculer la vraisemblance des paramètres : probabilité d'observer les états de caractères au site en fonction des paramètres d'états de caractères au site en fonction des paramètres ;

3. faire le calcul pour tous les caractères ;

4. calculer les longueurs de branches et les paramètres du modèle qui maximisent la vraisemblance ;

5. maximiser ensuite cette vraisemblance sur l'ensemble des topologies ;

6. l'arbre qui a la vraisemblance maximale est l'arbre de maximum de vraisemblance.

Les méthodes de reconstruction phylogénétique reposent presque toujours sur l'optimisation d'un critère permettant de comparer les phylogénies possibles d'un ensemble de taxons. Les méthodes classiques pour rechercher la phylogénie de vraisemblance maximale deviennent très coûteuses en temps de calcul lorsque le nombre de séquences augmente. Lorsque l'on souhaite reconstruire la phylogénie d'un grand nombre de séquences, il est donc difficile d'utiliser directement ce type de méthodes (Ranwez, 2002).

1.5 Reconstruction ancestrale

1.5.1 Reconstruction ancestrale basée sur les méthodes de parcimonie

L'idée d'une évolution minimale ou d'une parcimonie est proposée par Edwards et Cavalli-Sforza (Beanland et Howe, 1992). Des algorithmes pratiques ont été développés pour de nombreux cas particuliers de parcimonie. Étant donné un arbre phylogénétique et les séquences alignés des espèces contemporaines, la reconstruction ancestrale basée sur la parcimonie consiste à identifier les séquences aux noeuds ancestraux tout en minimisant le nombre d'opérations de mutations permettant d'expliquer la diversité observée dans les séquences contemporaines. Ce problème a été largement étudié dans le cadre de la reconstruction phylogénétique (Fitch, 1971).

1.5.2 Reconstruction ancestrale basée sur les méthodes de Maximum de vraisemblance

La méthode de maximum de vraisemblance (MV) est une approche statistique courante utilisée pour inférer les paramètres de la distribution de probabilité d'un échantillon donné. La méthode MV utilise aussi un modèle mathématique du processus d'évolution des séquences pour définir la probabilité qu'une phylogé-

nic puisse produire les séquences observées, et cherche ensuite la phylogénie pour laquelle cette probabilité est maximale. Pour l'utilisation du maximum de vraisemblance, il faut donc être capable de choisir un modèle d'évolution et d'estimer ses paramètres. L'estimation des paramètres prenant en compte tous les alignements possibles entre les séquences et calcule la vraisemblance d'un arbre pour ce modèle.

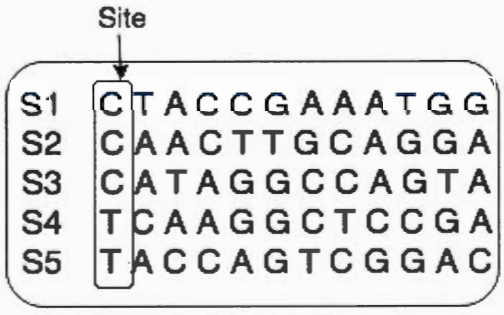


Figure 1.6: Site dans un alignement de séquence d'ADN.

L'alignement de toutes les séquences de nœuds de feuilles générés par le processus de simulation de la moyenne $\alpha(AB)$, a été trouvée à partir d'un très grand nombre de simulations. Ainsi une constante β a été introduite pour corriger cette sous-estimation de $\alpha(AB)$ faible(voir formule ci-dessous).

$$\alpha_i = K * \frac{C_i^\beta}{\sum_i C_i^\beta} \tag{1.1}$$

αi est le facteur de taux basé sur l'alignement du site i , k est le nombre de sites dans un alignement donné et Ci est la valeur attribuée au site i. D'autres études ont fourni de nouveaux algorithmes pour la reconstruction ancestrale ML (Pupko *et al.*, 2000; Koshi et Goldstein, 1996).

Le calcul de la fonction de vraisemblance dans sa forme la plus simple nécessite

un modèle décrivant les probabilités de substitution d'un nucléotide par un autre. Cependant, pour être statistiquement robustes et performantes, les méthodes probabilistes nécessitent l'incorporation de modèles qui décrivent de la façon la plus réaliste possible les processus biologiques d'évolution des séquences. La possibilité de définir explicitement un modèle d'évolution est un atout non seulement pour la reconstruction ancestrale, mais également pour mieux comprendre le processus d'évolution des séquences. Le modèle d'évolution de Jukes-Cantor (Tamura *et al.*, 2011) représente le modèle d'évolution de séquences nucléotidiques le plus simple. Ce modèle propose pour tous les nucléotides d'une séquence, des chances égales de changement et des chances égales de transition vers les trois autres bases.

1.6 la Reconstruction ancestrale avec insertion et suppression

1.6.1 Méthode de parcimonie

L'approche de parcimonie pour le problème de la reconstruction indel, a été introduite par (Ekelund *et al.*, 2004) et (Blanchette *et al.*, 2004a). Une analyse de qualité nous permet de dire que certaines reconstructions sont exactes et nous permettent d'identifier des caractéristiques chez les espèces modernes, telles que les restes d'anciennes insertions de transposons, qui n'ont pas été identifiées par analyse directe. En remontant l'histoire prédite de l'évolution des nucléotides dans la région reconstruite, on estime la quantité de renouvellement d'ADN dû à l'insertion, la délétion et la substitution dans les différentes lignées de mammifères placentaires depuis l'ancêtre euthérien, montrant des variations considérables entre les lignées. Le cas le plus simple est que tous les sites ont le même poids et toutes les mutations ont le même coût.

1.6.2 Méthode de Maximum de vraisemblance avec insertion et suppression

La reconstruction de séquences ancestrales avec indel est essentielle à diverses études évolutives. Des algorithmes performants comme FastML (Ashkenazy *et al.*, 2012) sont faits pour la reconstruction de séquences ancestrales intégrant des indels et des caractères ; en implémentant diverses fonctionnalités innovantes. Il utilise une méthode de codage indel, dans laquelle plusieurs sites sont codés sous forme de données binaires, il reconstruit ensuite les états indel ancestraux en supposant un processus de Markov à temps continu. D'autres algorithmes efficaces ont également été développés pour différents types de reconstructions basées sur MV, avec une complexité temporelle linéaire avec le nombre de séquences pour la plupart des algorithmes. Les méthodes de reconstruction de séquences ancestrales nécessitent en entrée à la fois un alignement de séquences multiples de séquences existantes et un arbre phylogénétique correspondant, les écarts sont généralement traités comme des caractères inconnus, une donnée manquante ou comme un caractère supplémentaire. Ainsi (Diallo *et al.*, 2007) qui est notre article de référence, présente un outil convivial pour la reconstruction de séquences ancestrales. Il résout un problème qu'on appelle "Indel Maximum Likelihood Problem" (IMLP) qui est une étape importante pour la reconstruction de la génomique ancestrale séquences, en étudiant le problème de la reconstruction du scénario le plus probable des insertions et des suppressions capables d'expliquer les écarts observés dans l'alignement.

Les concepts de vraisemblances sont fondés sur les méthodes probabilistes, un modèle d'évolution des séquences sous la forme d'un processus de Markov à temps continu pour la reconstruction baptisée *links*(Thorne *et al.*, 1991). La modélisation prend en compte des substitutions entre caractères, mais également l'apparition d'événements de mutations : insertions et délétions. La vraisemblance est la

probabilité conditionnelle d'observer les données sous un modèle particulier. Étant donné un modèle qui spécifie les probabilités d'observer différents événements, la vraisemblance L d'obtenir les données observées peut être calculée (Entfellner, 2011).

$$L = Pr(X|H) = \prod_{i}^{m} Pr(X^m|H) \tag{1.2}$$

est la probabilité conditionnelle d'observer les données $\mathbf{X}$ sous l'hypothèse $\mathbf{T}$, où $\mathbf{X(i)}$ représente les données au site $\mathbf{i}$. Ainsi pour chaque site, la probabilité d'observer l'état actuel sera calculée en fonction de chacun des états possibles au site et à chacun des nœuds internes de l'arbre. On recherche la vraisemblance des données $\mathbf{X}$ sous différentes hypothèses évolutives $\mathbf{H}$ d'un modèle d'évolution $\mathbf{M}$ et l'on retient les hypothèses qui rendent cette vraisemblance maximale.

L'approche de maximum de vraisemblance est plus fiable que les méthodes de distances et de maximum de parcimonie. Théoriquement, c'est l'approche qui conduit au résultat le plus proche de l'arbre évolutif réel (Ashkenazy *et al.*, 2012). Comparée au maximum de parcimonie, elle est aussi beaucoup plus robuste. Cette approche permet également d'appliquer les différents modèles d'évolution et d'estimer la longueur des branches en fonction des changements évolutifs. Dans la figure ci-dessous nous avons différents modelés d'évolution de séquence.

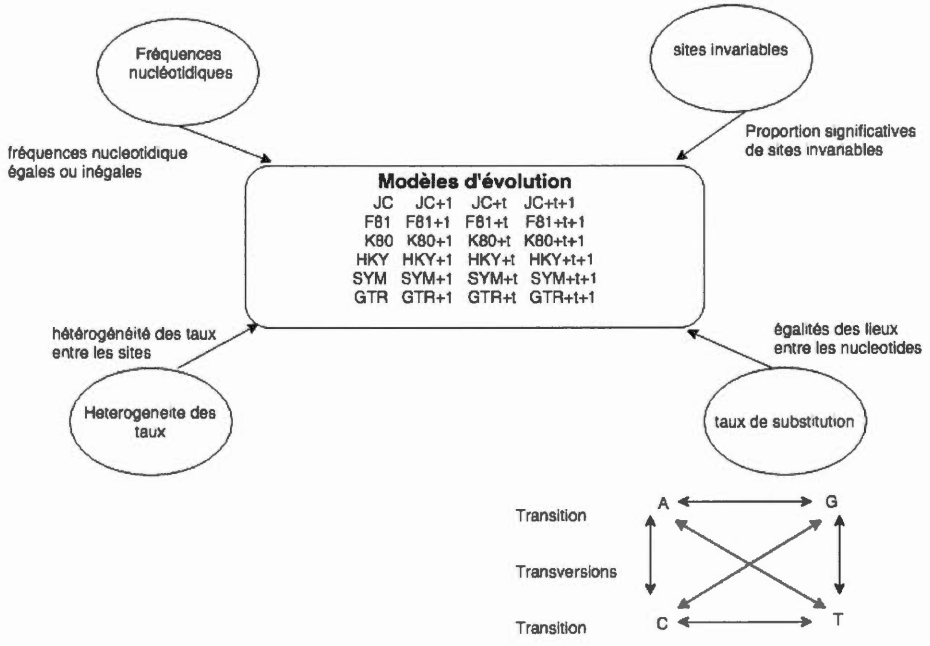


Figure 1.7: Constitution des principaux modèles d'évolution des séquences nucléotidiques.

1.7 Modèle de Markov

La modélisation stochastique permet l'utilisation de modèles probabilistes pour résoudre des problèmes avec des informations incertaines ou incomplètes. Ainsi, les modèles de Markov suscitent un intérêt pour les aspects théoriques et appliqués, le modèle de Markov permet de modéliser des systèmes qui changent aléatoirement. On suppose que les états futurs ne dépendent que de l'état actuel et non des événements survenus avant lui.

En biologie moléculaire, le modèle de Markov est très utilisé, souvent pour inférer différents taux d'évolution pour différents sites sur une séquence d'ADN (Felsenstein et Churchill, 1996). La probabilité à priori de chacun des sites, la longueur moyenne du chemin des sites ayant tous le même taux. La probabilité totale de la

phylogénie est calculée comme une somme de termes, chaque terme étant la probabilité des données étant donné une affectation particulière des taux aux sites, multipliés par la probabilité antérieure de cette combinaison particulière de taux.

1.7.1 Chaîne de Markov

Les chaînes de Markov sont des suites de variables aléatoires qui sont caractérisées par un aspect particulier de dépendance, qui leur attribue des propriétés singulières et un rôle important en modélisation de problème complexe. Beaucoup d'études testent la validité du calcul statistique basé sur l'utilisation des chaînes de Markov de premier, deuxième et troisième ordre. Certains problèmes peuvent être résolus en le modélisant comme une chaîne de Markov.

Exemple définition : Soit S, un ensemble fini et X = Xn un processus stochastique à valeurs dans S. On dit que X est une chaîne de Markov si pour tout $n \in N$, et pour tout$x_0, x_1, , , x_n + 1 \in s$ tel que $P(X_0 = x_0, X_1 = x_1, ..., X_n = x_n) > 0$

Si l'on considère E un espace d'état S(A,B,C) avec des probabilités de transition P, la représentation graphique de la chaîne de Markov est définie par les états qui sont représentés par des cercles numérotés et les probabilités de transitions par des flèches(voir figure ci-dessous).

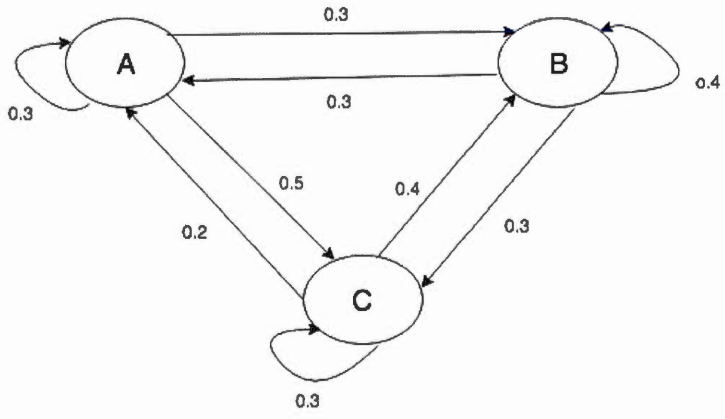


Figure 1.8: Exemple de chaîne de Markov

La sélection du modèle est conduite dans la classe des chaînes de Markov à longueur variable par l'algorithme du contexte qui génère des séquences dont la probabilité de chaque symbole dans la séquence dépend du symbole précédent(chaîne de Markov pour la modélisation de séquences biologiques (Bourguignon et Robelin, 2004)). Une chaîne de Markov est composée de quatre états : A, G, C et T. La probabilité de passer d'un état à l'autre ou d'un résidu d'ADN à un autre. Ce paramètre est défini comme la probabilité de transition(voir figure ci-dessous).

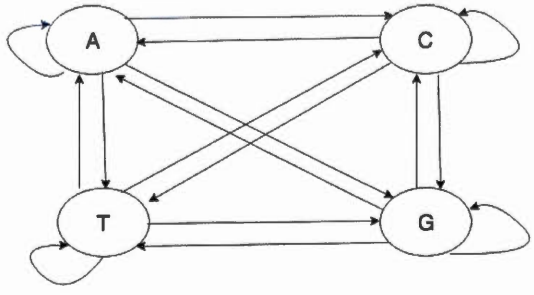


Figure 1.9: Exemple de chaîne de Markov pour l'ADN.

1.7.2 Modèle de Markov caché

Un modèle de Markov caché (MMC) est un modèle statistique permettant de représenter un processus de Markov dont l'état est non observable. Les modèles de Markov cachés sont des modèles probabilistes pour les séquences de symboles. Ils sont très utilisés pour l'analyse des séquences biologiques (ADN, ARN et protéines). Le modèle de Markov est composé de quatre états cachés : A, G, C et T, chacun de ces états peut émettre un chiffre (probabilité d'émission) et la probabilité de transition, de passer d'un état à l'autre. Pour faciliter l'analyse, un état de début et un état de fin sont rajoutés, une exécution du processus commence dans l'état *start*, qui est composée d'une séquence d'états cachés et d'émissions et se termine lorsque l'état *end* est atteint (Popier, 2005).

Le type de questions répondues par les MMC est les suivants :

- est-ce qu'une séquence particulière appartient à une famille spécifique ?
- quelles informations pouvons-nous tirer de sa structure interne ?

Cependant, de l'information sur ces états cachés peut être déduite de données observées. En effet, la fréquence d'observation de ces données dépend de l'état du processus dans lequel celles-ci sont émises.

Le modèle de Markov caché d'arbre est utilisé pour résoudre le IMLP, dont l'arbre - HMM est défini par trois éléments : l'ensemble des états, l'ensemble des probabilités d'émission, et l'ensemble des probabilités de transition. Il se traduit par la spécification de modèles de Markov cachés du second ordre, leur apprentissage et leur utilisation pour permettre une segmentation de grandes séquences d'ADN en différentes classes qui traduisent chacune un état organisationnel et structural des

motifs d'ADN locaux sous-jacents.

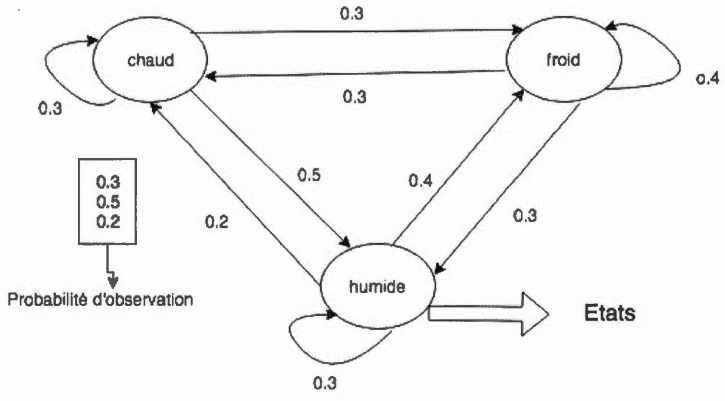


Figure 1.10: Exemple de modèle de Markov caché.

Le modèle de Markov est un modèle statistique composé d'états, d'émissions et de transitions.

a) États

Un modèle de Markov caché est basé sur un modèle de Markov, sauf qu'on ne peut pas observer directement les séquences d'états : les états sont cachés, chaque état émet des "observations" qui, elles, sont observables. On travaille sur la séquence d'observations générée par les états. On note la présence d'un état "Début" qui sert à présenter les probabilités de commencer dans chacun des états du modèle. Par définition, on ne revient jamais à l'état de départ, raison pour laquelle il n'y a jamais de transition vers cet état. Les chaînes de Markov admettent plusieurs types d'états. Leur classification permet de mieux étudier les propriétés asymptotiques des chaînes de Markov et est essentielle pour la compréhension.

- Un état récurrent : Un état i est dit récurrent si la probabilité d'un éventuel retour à l'état i vaut 1, sachant que la chaîne a commencé à l'état i.

- Un état transitoire : Un état i est transitoire si la probabilité d'un éventuel retour à l'état i est inférieur 1, sachant que la chaîne a commencé à l'état i.

- Un état périodique : Un état i est dit périodique si l'état j est accessible depuis l'état i et si la probabilité d'atteindre j en n transitions depuis i est strictement positive.

- Un état apériodique : Un état i est dit apériodique si il est contraire d ;un états périodique.

Les états sont définis comme un scénario indel à colonnes uniques(Diallo *et al.*, 2007). Étant donné un arbre binaire enraciné T avec n feuilles, chaque état correspond à un étiquetage différent des arêtes et avec l'un des événements possibles. Les méthodes directes ne sont pas utilisables pour des modèles de grande taille (nombre d'états)(Garcia et Martin-Vide, 1993).

b) Émissions

Chaque état peut émettre chacune des observations avec une certaine probabilité que nous appelons "probabilité d'émission". Cette probabilité peut éventuellement être nulle, ce qui signifie que l'état ne peut pas émettre l'observation concernée. Une exécution du processus commence dans l'état *start*, composé d'une séquence d'états cachés et d'émissions, et se termine lorsque l'état *end* est atteint.

Un ensemble de probabilités d'émission p(s :e) représentant (en excluant l'état initial et les états finaux) la probabilité d'émettre e de l'état caché s (Bréhélin et Gascuel, 2000).

c) Transition

Une transition matérialise la possibilité de passer d'un état à un autre. Dans le modèle de Markov, les transitions sont unidirectionnelles : une transition de l'état A vers l'état B ne permet pas d'aller de l'état B vers l'état A. Tous les états ont des transitions vers tous les autres états, y compris vers eux-mêmes. Chaque transition est associée à sa probabilité positive d'être empruntée et cette probabilité peut éventuellement être nulle (Vouma Lekoundji, 2014).

Soit X= $(X_n)n \in N$ une chaîne de Markov d'espace d'états S.

Soit $(u, v) \in S$ deux états.

La probabilité de transition de u à v est notée :

$$P(uv) = P(x_n + 1 = u|X_n = v) \tag{1.3}$$

Le modèle de Markov a été utilisé ces dernières années dans divers contextes, pour modéliser le déséquilibre de liaison dans les données génétiques des populations et de cartographie des maladies. L'identification de segments chromosomiques d'ascendance distincts a un large éventail d'applications allant de la cartographie des maladies à l'apprentissage de l'histoire. (Price *et al.*, 2009) décrivent une méthode HAPMIX, qui utilise un modèle génétique de population explicite pour effectuer une telle inférence d'ascendance locale basée sur des données de variation à échelle précise. La recombinaison se produisant aux ascendances du modèle est des processus de Poisson le long du génome. Supposons que nous ayons s sites, et une carte donnant les distances génétiques r1,r2,... r(S-1) entre des paires de sites adjacentes. Compte tenu des paramètres essentiels, et pour un chromosome mélangé haploïde, ils formalisent les probabilités de transition par rapport aux états (caché) pour la position. Pour le modèle original de Li et Stephens, il est

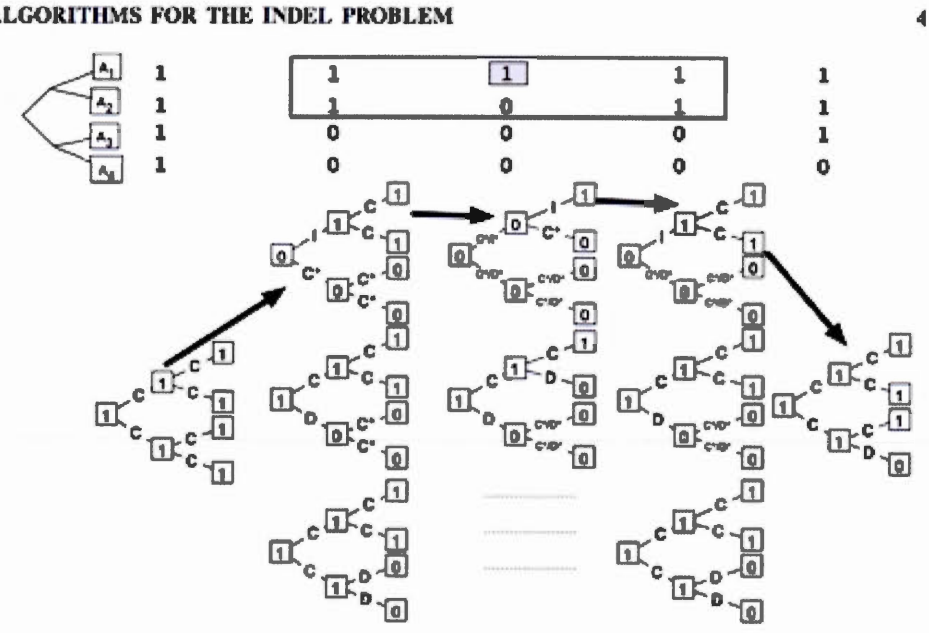


Figure 1.11: Exemple de modèle de Markov avec des états de probabilité valides, non nuls, associés à l'alignement multiple donné en haut de la figure.(Diallo et al, 2007).

possible de réduire sensiblement le temps de calcul en utilisant le fait que de nombreuses paires de probabilités de transition entre états sont identiques, ce qui permet de regrouper les termes dans l'algorithme terme unique partagé entre tous les états de destination comme dans le cas de (Diallo *et al.*, 2007)voir figure ci-dessus.

1.8 Chemin le plus probable : Algorithme de Viterbi

L'algorithme de Viterbi (Forney, 1973) permet de calculer la séquence la plus probable d'états cachés étant donné une séquence d'émissions connue. Bien qu'il ne soit plus possible pour une séquence d'indiquer de quelle chaîne d'états elle provient, l'information la plus souvent demandée est la chaîne des états. Ainsi il faudrait pour une séquence donnée déterminer quel est le chemin le plus probable d'émettre une séquence donnée.

Définition de l'algorithme de Viterbi :

Nous calculons la probabilité conjointe de la séquence d'observation avec la meilleure séquence d'état.

$$\delta_t(i) = \max_{q_0, q_1, .. q_{t-1}} P(q_0 q_1 q_{t-1}, o_1, o_2, o_t, q_t = i|\lambda) \tag{1.4}$$

Comme étant la probabilité conjointe des observations o_1 à o_t et la séquence d'états $q_1 q_2 q_3 \ldots q_t$ se terminant à l'état $q_t = s_i$ étant donné l'HMM λ.

$$\delta_t(i) = \max_{i=1}^{N} \delta_{t-1} a_i j b_j(o_t) \tag{1.5}$$

La programmation par contraintes permet de formuler des problèmes classiques associés aux MMC comme des problèmes à contraintes (Petit et Henning, 2009). Plus précisément de modéliser le calcul du chemin de Viterbi comme un problème d'optimisation pour effectuer efficacement ce calcul. L'algorithme calcule pour chaque état caché du modèle **s** le chemin le plus probable atteignant cet état à l'étape **i** et la probabilité de ce chemin partiel. Les valeurs cumulées (métriques au sens de Viterbi) calculées en chacun de ces états dans l'un des sens sont mémorisées afin qu'à l'étape suivante, dans l'autre sens, des décisions puissent être prises sur l'ensemble des traitements du bloc. Cette structure de données associe à chaque

état caché le chemin le plus probable atteignant cet état ou la probabilité de celui-ci. L'algorithme permet le calcul du chemin de Viterbi pour un MMC classique, ce calcul est effectué linéairement par rapport à la taille de la séquence d'émissions. Le calcul du chemin de Viterbi ne nécessite que de se remémorer de différents chemins les plus probables atteignant chaque état caché pour chaque étape du processus. Cependant, le cadre des MMC contraints introduit une nouvelle complexité dans leur calcul. En effet, il est nécessaire de prendre en compte la validité du chemin suivi, c'est-à-dire prendre en compte la satisfiabilité des contraintes associées à ce chemin.

1.9 Algorithme Forward, Backward

Les modèles de Markov cachés (HMM) sont principalement basés sur les algorithmes forward, backward et de Viterbi. Des combinaisons de ces algorithmes sont utilisées dans l'apprentissage des paramètres de HMM. L'algorithme forward, backward commence par effectuer un calcul progressif des probabilités, un calcul «en avant», qui donne la probabilité d'obtenir n premières observations dans une séquence donnée, en terminant dans chaque état possible du modèle de Markov. Il effectue ensuite un calcul rétrogressif des probabilités, qui représente la probabilité d'obtenir les autres observations ultérieures à un état initial. Ces deux ensembles de probabilités peuvent être combinés pour obtenir à tout instant la probabilité de l'état caché.

Définition algorithme Forward :

$$\alpha_t(i) = P(o_1 o_2 .. o_t, q_t = i|\lambda) \tag{1.6}$$

Comme étant la probabilité d'observations o_1 à o_t avec la séquence d'états qui se termine à l'état $q_t = s_i$ pour l'HMM λ

Définition algorithme Backward :

$$\beta_t(i) = P(o_1 o_2 .. o_t + 1, q_t = i|\lambda) \tag{1.7}$$

Comme étant la probabilité d'observations $o_t + 1$ à o_T avec la séquence d'états qui se termine à l'état $q_t = s_i$ pour l'HMM λ.

La combinaison de modèles de Markov phylogénétique est cachée dans l'analyse de Biosequence (Siepel et Haussler, 2004). Ces modèles combinent des modèles phylogénétiques de l'évolution moléculaire, qui s'appliquent à des sites individuels, et des modèles de Markov cachés, qui permettent des changements d'un site à l'autre. Beaucoup d'auteurs ont montré que la probabilité totale d'un alignement peut être calculée efficacement en utilisant un algorithme de programmation dynamique récursif, qui est équivalent à l'algorithme forward. Le principal résultat est une méthode simple et efficace pour prendre en compte les états d'ordre supérieur dans le HMM. Une implémentation naïve de l'algorithme avant / arrière nécessiterait un temps de calcul proportionnel.

1.10 Topologie d'arbre

La structure d'arbre intervient naturellement dans de nombreux algorithmes pratiques. Les algorithmes d'optimisation combinatoire, de type diviser pour régner, d'évaluation d'arbres de jeux et d'évaluation d'expressions fonctionnelles possèdent très souvent une structure intrinsèque en arbre dans laquelle les sommets représentent les tâches calculatoires et les arêtes les communications nécessaires aux calculs. Mais, il est souvent difficile de connaître à priori la forme de l'arbre ou sa taille avant que les calculs ne soient quasiment tous effectués.

Deux aspects essentiels de la discipline nommée topologie sont les notions de voisinage et d'équivalence de formes. Ce sont ces idées que nous voulons évoquer

lorsque nous utilisons le mot topologie à propos d'arbre (Paulin, 1988). Il a longtemps été supposé que l'évolution des espèces est un processus de branchement qui peut être représenté seulement par une topologie d'arbre. Une espèce est liée à son ancêtre le plus proche et toutes les autres relations entre les espèces ne peuvent être prises en compte. Ce type d'étude a d'abord été développée dans le cadre de la biologie, c'est pourquoi on parle d'arbre phylogénétique. Cette rapide augmentation du nombre de topologies possibles en fonction du nombre de séquences étudiées est un problème majeur en reconstruction phylogénétique.

Dans une topologie arbre raciné binaire T = (VT , ET) avec des longueurs de branche λ. Si n est le nombre de feuilles de T, il existe n - *1* des nœuds internes et *2n* - *2* bords et la formule pour trouver le nombre de topologies est :

$$\hat{N} = \frac{(2n-3)!}{2^{n-2}(n-2)!} \tag{1.8}$$

Dans une topologie arbre non racinée on a

$$\hat{N} = \frac{(2n-5)!}{2^{n-3}(n-3)!} \tag{1.9}$$

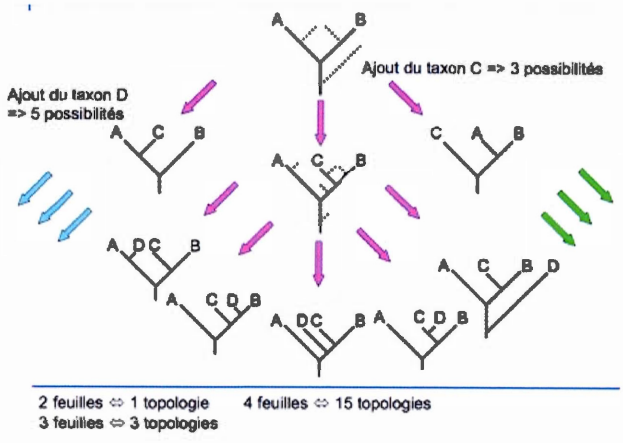


Figure 1.12: Arbres racinés avec différentes topologies (Céline Brochier-Armanet 2015-2016) .

Pour résoudre les conflits topologiques entre les topologies d'espèces et de gène, le scénario obtenu repose sur les transferts horizontaux de gène (THG).

1.11 Transfert horizontal de gène

Le transfert horizontal de gènes (THG) est un mécanisme d'évolution naturel qui consiste à faire un transfert direct du matériel génétique d'une espèce à une autre, ils correspondent à la possibilité naturelle ou artificielle qu'un organisme capte un fragment d'ADN présent dans son environnement de proximité ce qui fait ce mécanisme est comme un événement de suppression et d'insertion successif. La possibilité que le transfert horizontal de gènes puisse jouer un rôle clé dans l'évolution biologique est un changement fondamental dans notre perception des aspects généraux de la biologie évolutive survenue ces dernières années (BOC, 2012).

C'est un phénomène où il arrive par exemple que des espèces échangent une partie de leur matériel génétique, on parle alors de transferts horizontaux. Lors de la reconstruction phylogénétique, on suppose que la similitude entre les séquences est due au fait qu'elles possèdent un ancêtre commun proche ; or les transferts horizontaux augmentent la similitude entre des séquences qui n'ont pas forcément de parents communs proches. Les transferts horizontaux peuvent jouer un rôle considérable dans l'évolution des séquences et fausser complètement les reconstructions phylogénétiques issues d'analyses de séquences moléculaires.

Le transfert horizontal de gènes (THG) est très fréquent chez les procaryotes, Bactéries et Archéobactéries, des mécanismes sophistiqués ont été développés pour acquérir rapidement de nouveaux gènes(Courvalin, 1998). Les bactéries ont développé divers mécanismes de résistance pour survivre en présence d'antibiotiques, un des plus répandus est la synthèse d'enzymes qui inactivent les antibiotiques.

La production de ces enzymes est généralement due à l'acquisition de gènes provenant d'autres bactéries, par transfert « horizontal »(Faure, 2009). Le risque principal de la présence de gènes de résistance modifiée génétiquement est de contribuer à la dissémination de la résistance aux antibiotiques chez les bactéries pathogènes. Le mode transfert de résistance aux antibiotiques vers les bactéries pourrait résulter d'un transfert horizontal d'ADN (Simonet, 2000). L'incidence élevée de la résistance aux antibiotiques chez les bactéries pathogènes est principalement due au fait qu'elles ont développé des systèmes de transfert d'ADN extrêmement efficaces et à très large spectre d'hôte. Le Transfert horizontal, aussi appelé « Transfert latéral », est un phénomène très bien décrit chez les bactéries sous les trois mécanismes : la Transformation, la Conjugaison et la Transduction.

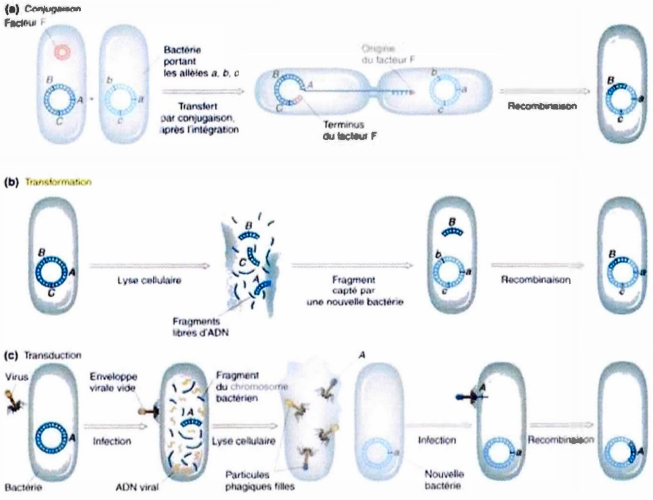
La conjugaison : phénomène qui implique un contact physique direct entre bactérie donatrice et bactérie réceptrice, et par laquelle le plasmide passe de l'une à l'autre ; le processus est le suivant : deux cellules entrent en contact et s'échangent toute une partie de leur matériel génétique. Ce mécanisme est particulièrement important à l'intérieur d'une même espèce, mais le phénomène de conjugaison peut également avoir lieu entre des organismes qui ne font pas partie de la même espèce.

La transformation : phénomène par lequel une bactérie dite " compétente " incorpore de l'ADN nu présent dans l'environnement ; cet ADN nu peut être intégré à l'intérieur d'une cellule particulière, puis être intégré au génome de cet organisme. Il y a donc un transfert horizontal entre deux espèces qui peuvent être tout à fait différentes.

La transduction : phénomène au cours duquel l'ADN est véhiculé par un bactériophage. Certains types de virus sont des organismes capables d'intégrer leurs matériels génétiques dans l'hôte et s'inspirent du génome de l'hôte pour donner

naissance à de nouvelles particules virales.

C) Transfert horizontal de gènes



Ces vecteurs sont utilisés en biotechnologie pour transférer des gènes d'une espèce à une autre (transgénèse) afin d'obtenir des OGM.

Figure 1.13: Mécanisme de transfert horizontal de gène chez les bactéries.

Le problème de la construction de scénarios parcimonieux pour des ensembles individuels de gènes orthologues donnés pour un arbre d'espèces(Mirkin *et al.*, 2003). L'analyse comparative des génomes séquencés révèle de nombreux cas de transfert de gène horizontal apparent, au moins chez les procaryotes, et indique que la perte de gènes spécifiques à la lignée aurait pu être encore plus fréquente dans l'évolution. Cela complique la notion d'arbre des espèces, qui doit être réinterprété comme une tendance évolutionnaire dominante et fait de la reconstruction des génomes ancestraux une tâche non triviale. Des algorithmes ont été développés pour réconcilier les modèles phylétiques avec l'arbre des espèces en postulant la perte de gènes, l'émergence du COG et le HGT. Ils prouvent que chacun de ces algorithmes produit un scénario évolutif parcimonieux, qui peut être repré-

senté comme une cartographie des événements de perte et de gain sur l'arbre des espèces. ECCETERA (Anselmetti, 2017) est un algorithme qui utilise un modèle de parcimonie avec certain paramètre pour la réconciliation d'arbres de gènes avec l'arbre des espèces utilisant un modèle d'évolution des mutations prenant en compte des événements de duplications, pertes et transferts latéraux de gènes, a l'ensemble des gènes ancestraux d'un arbre de gènes.

Avec un nombre élevé de gènes, le nombre de transferts de gène horizontaux et les événements de perte de gène seront presque identiques. L'approche de la reconstruction des scénarios évolutifs est conservatrice en ce qui concerne la détection de HGT, car seuls les modèles de présence / absence de gènes dans les génomes séquencés sont pris en compte. On peut avoir d'autre type de transfert dans le cas de la résistance acquise par mutation ponctuelle, par hyper-production d'un gène déjà présent ou par acquisition de gènes de résistance exogènes peut être transférée verticalement (à la descendance) ou horizontalement (aux autres lignées bactériennes).

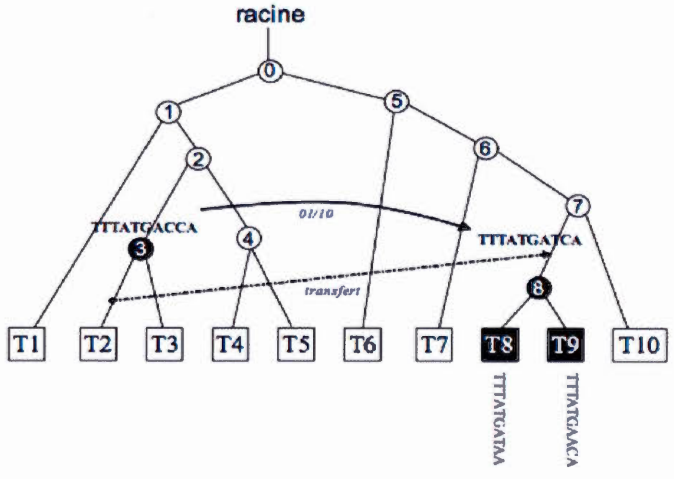


Figure 1.14: Transfert horizontal entre les branches *(3, 2)* et *(7, 8)* (Nguyen *et al.*, 2005) .

1.11.1 Détection du transfert horizontal de gène

Le transfert horizontal de gènes peut rapprocher deux feuilles d'arbres lointains. Donc pour la réconciliation de deux arbres il nous faut des méthodes de détection de transfert. Les approches pour détecter des transferts horizontaux de gènes peuvent être divisées en deux parties principales :

1. l'analyse du génome ancestral peut révéler des séquences avec un taux d'évolution anormal. En supposant que ces séquences ne soient pas apparues par un processus de sélection, on peut en déduire qu'elles sont le résultat d'un transfert horizontal de gènes. L'ADN introduit par un transfert horizontal récent présente souvent des caractéristiques de séquence inhabituelles et peut être distingué de l'ADN ancestral (Lawrence et Ochman, 1997) ;
2. la comparaison d'un arbre phylogénétique d'espèces avec une phylogénie basée sur un gène observé, et inféré pour le même ensemble d'espèces, peut révéler des conflits topologiques qui sont explicables par des transferts horizontaux.

Certains arbres particulièrement aberrants au niveau de leur topologie peuvent introduire du bruit dans l'analyse et fausser la reconstruction. Il nous faut donc un critère pour mesurer la similarité des arbres, on peut citer trois critères qui permettent de déterminer les meilleurs transferts à chaque étape du processus de réconciliation : la distance de Robinson et Foulds (Robinson et Foulds, 1981) ; distance des moindres carrés et la dissimilarité de bipartition. La méthode la plus utilisée est la distance topologique de Robinson et Foulds (RF) qui mesure la différence entre les phylogénies d'espèces et de gène.

La méthode de RF, pour la détection de transfert horizontal de gène est basée sur les topologies des arbres, il est donc judicieux de sélectionner les arbres sur un

critère de similarité topologique. La distance de RF représente le nombre minimum d'opérations élémentaires de contraction et d'expansion de nœuds nécessaires pour transformer un arbre phylogénétique en un autre. La distance topologique de RF entre le vrai arbre généré et l'arbre solution correspondant à la distance d'arbre obtenue. Cette distance, qui est souvent utilisée pour comparer les topologies de deux arbres additifs différents pour transformer une structure d'arbre en une autre. Robinson et Foulds ont montré que cette distance est aussi égale au nombre de bipartitions, ou scissions, qui sont présentes dans un arbre et absentes à l'autre. En calculant la distance topologique, nous avons toujours supposé qu'une arête de longueur nulle induit une bipartition une fois la distance de Robinson et Foulds obtenue. Cette distance varie donc de *0* pour deux arbres identiques, à *1* pour deux arbres ne possédant aucune bipartition commune.

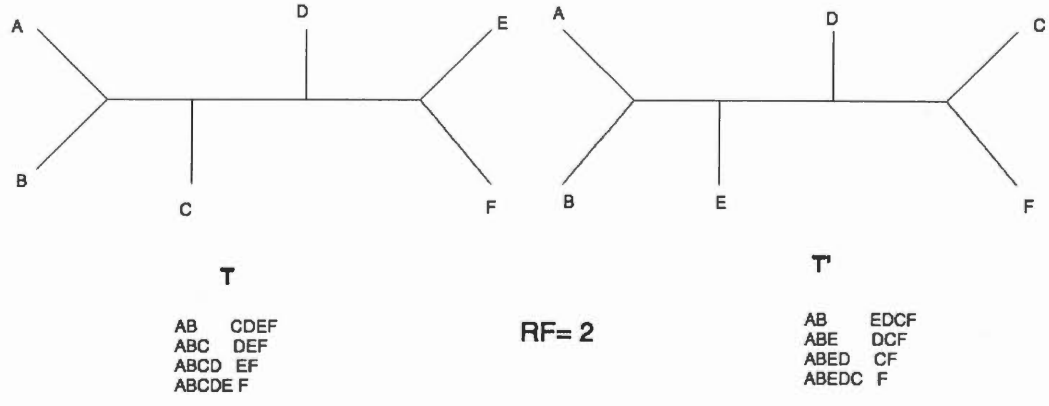


Figure 1.15: La distance de Robinson et Foulds.

La méthode qui est basée sur la réconciliation des topologies de l'arbre du gène considéré (Boc *et al.*, 2004) et de l'arbre d'espèces. Cette méthode considère un modèle d'évolution phylogénétique en réseau. Le critère d'optimisation utilisé est la distance topologique de RF, qui permet de mesurer la similarité entre deux arbres phylogénétiques pour pouvoir détecter les transferts horizontaux. La

nouvelle méthode pour générer un scénario de propagation du gène testé pour un ensemble d'espèces considérées. Ils montrent aussi comment les différences topologiques entre les arbres de gène et d'espèces peuvent être exploitées pour déterminer un scénario possible de transferts latéraux du gène considéré survenus au cours de l'évolution.

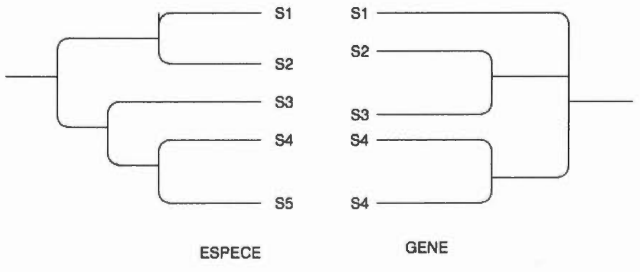


Figure 1.16: La phylogénie d'espèces est différente du gène.

Certaines méthodes de détection de transfert utilisent des critères pour identifier les gènes transférés ou perdus récemment (Daubin, 2002). Certaines méthodes ne détectent qu'un sous-ensemble des gènes transférés horizontalement par une succession d'événements gain et perte de gène acquis récemment en utilisant des génomes proches. L'hypothèse de parcimonie est utilisée pour discriminer entre une acquisition récente et des pertes chez les espèces voisines. Aussi, une solution efficace au problème de la détection de transferts horizontaux de gène pour un groupe d'espèces observées proposer par (BOC, 2012), leur approche algorithmique à ce problème utilise le principe de réconciliation d'arbres d'espèces et de gène, tout en considérant des contraintes d'évolution biologiques et algorithmique. La méthode proposée, traite la question du transfert complet qui suggère que la séquence transférée entre deux espèces représente un gène au complet. Elle décrit un algorithme, une nouvelle mesure de comparaison d'arbres phylogénétiques et un procédé de validation des résultats.

L'inférence phylogénétique du calcul d'une distance évolutive entre des paires d'es-

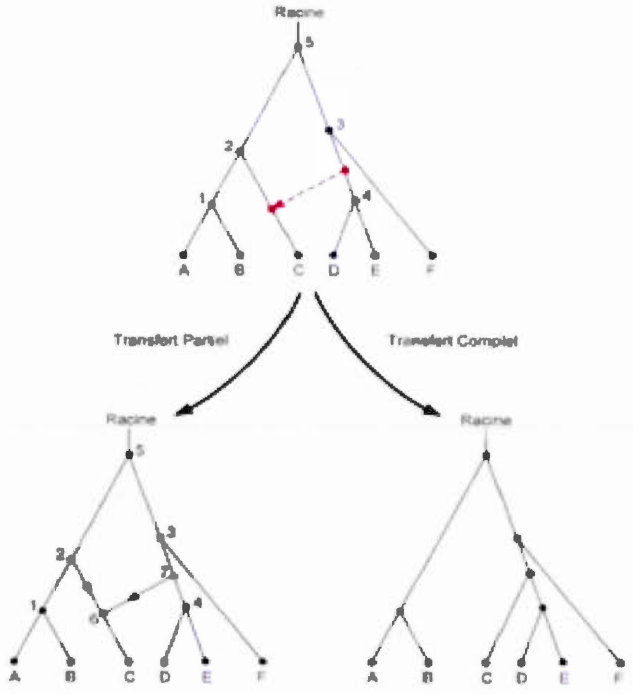


Figure 1.17: Deux modèles de transferts horizontaux(Boc, 2016).

pèces définies par les gènes sources a été étudié dans (Criscuolo, 2006). Ces méthodes permettent de comparer des arbres topologiquement. Sachant que chaque branche interne d'un arbre phylogénétique induit une bipartition de l'ensemble de ses feuilles, la distance de bipartition *dRF* mesure le nombre de bipartitions induites par un arbre, mais pas par l'autre. Un arbre phylogénétique non enraciné dénombrant au plus n - 3 branches internes, la distance *dRF* est couramment normalisée par $2n$ - 6 pour la situer dans l'intervalle [0, 1]. Bien que ce soit la distance topologique la plus utilisée, ils sont souvent recommandés pour comparer des arbres relativement proches. Elle utilise aussi d'autres méthodes de comparaison comme la distance d'agrément *dMAST(T, T')* qui n'est pas du tout sensible à ce biais, mais elle dénombre les feuilles qui ne sont pas présentes dans la plus grande restriction commune à T et T'. La distance de quadruplets *dquad* qui dénombre

tous les sous-arbres de quatre feuilles qui sont présents dans un arbre, mais pas dans l'autre. Cette distance topologique d'arbre phylogénétique ne présente aucun des biais propres à *dRF* et *dMAST* dans la comparaison d'arbre.

1.11.2 Nombre optimal de transferts

Le transfert horizontal de gène est un mécanisme qui permet de mettre en clair l'échange de matériel génétique entre individus ; et l'acquisition rapide de nouvelles fonctions permettant de s'adapter à un nouveau changement génomique. Un gène transféré doit notamment interagir avec de nouveaux partenaires génomiques. Plusieurs études ont montré des méthodes de détections de transfert horizontal avec des modélisations mathématiques permettant de trouver le nombre optimal de transferts. Cependant, les travaux d'Hao et Golding (Hao et Golding, 2004) suggèrent qu'un grand nombre de transferts horizontaux sont présents aux extrémités de l'arbre de la vie. L'algorithme de détection (HGT) dans le cadre du modèle de transfert de gène complet (Boc et Makarenkov, 2003), cette méthode utilise la topologie d'un arbre phylogénétique comme une structure de base sur laquelle on ajoute, au fur et à mesure en suivant un critère d'optimisation pour trouver le nombre optimal de transferts. Chaque transfert horizontal est mis à jour, permettant une meilleure compréhension de ce phénomène de réconciliation. Deux critères d'optimisation seront considérés : le premier d'entre eux est la fonction des moindres carrés (LS) comme critère d'optimisation métrique, et la distance topologique de Robinson et Foulds (RF) est utilisée comme critère d'optimisation topologique. L'arbre phylogénétique d'espèces original T est graduellement transformé en l'arbre phylogénétique du gène T' par une série de mouvements SPR (THG). Le but de ces opérations est de trouver la plus courte séquence possible d'arbres.

Lorsque la distance de RF est considérée, on peut l'utiliser comme critère d'optimisation comme suit : Toutes les transformations possibles de l'arbre des espèces, consistant à transférer l'un de ses sous-arbres d'un bord à l'autre, sont évaluées la distance de RF entre l'arbre d'espèces transformé T et l'arbre génique T1 est calculé. Le transfert de sous-arbre (le déplacement de SPR) fournissant le minimum de la distance RF entre T et T1 est retenu comme une solution. Dans le processus s'il existe des sous-arbres identiques avec deux feuilles ou plus dans T et T' alors on diminue la taille du problème en les écrasant dans T et T' pour faciliter le calcul, car plus l'arbre est petit plus le calcul sera plus facile. Notez que le problème de trouver le nombre minimum d'opérations de transfert de sous-arbres nécessaires pour transformer un arbre en un autre s'est avéré être NP-difficile, mais peut être approché d'un facteur 3. D'autres études ont parlé du nombre optimal de transferts pour la réconciliation de l'arbre d'espèce en arbre de gène (Boc, 2016) illustrant dans la figure 2 18 ci-dessous.

Une fois que les familles de transferts ont été détectées, il faut reconstruire chaque arbre et le comparer avec l'arbre de référence afin de détecter les incongruences. Ces incompatibilités avec la phylogénie des espèces sont causées par le fait qu'une espèce est plus similaire au gène de l'espèce donneuse qu'aux espèces proches de l'espèce receveuse. La méthode de RIATA-HGT a été étudiée par : (Nakhleh *et al.*, 2005), la première heuristique polynomiale à gérer tous les scénarios THG, sans aucune restriction. La méthode infère avec précision des événements de THG basés sur l'analyse de l'incongruence entre les espèces et les arbres génétiques, le nombre optimal d'événements HGT est surestimé. Pour chaque paire d'espèces, il existe trois manières possibles de les concilier, chacune de ces manières est appelée une sous-résolution. Chaque sous-résolution est un ensemble d'événements, qui dans ce cas n'est qu'un événement.

Des implémentations permettent de spécifier un ensemble d'arbres génétiques et

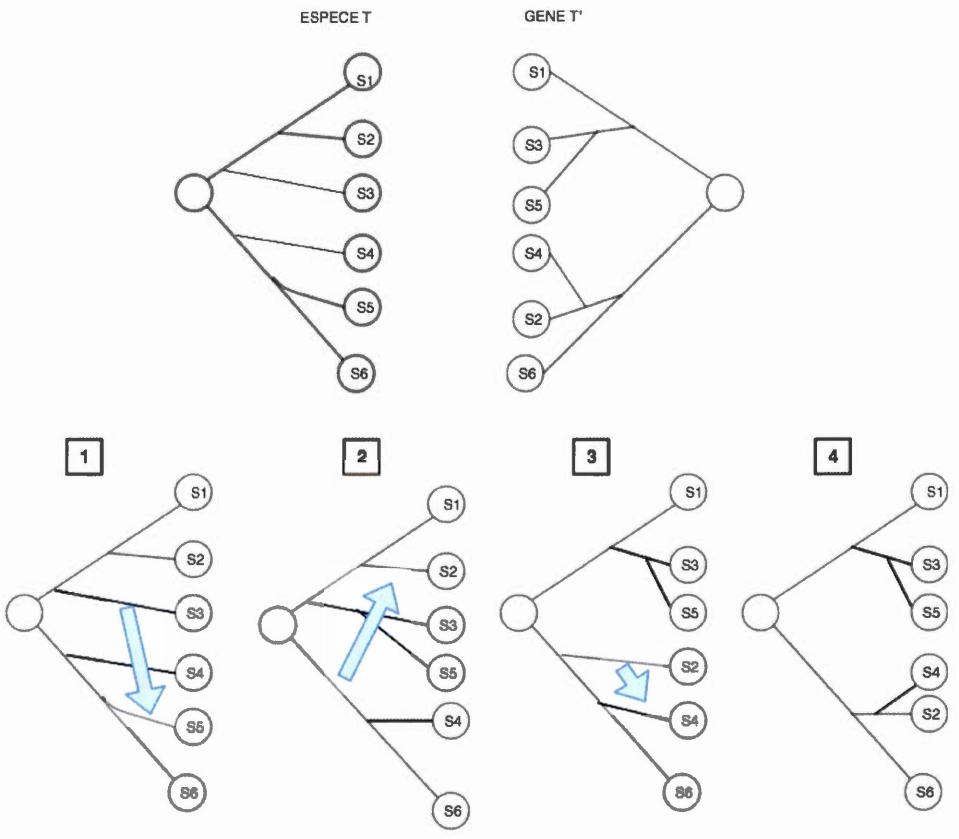


Figure 1.18: Fonctionnement de l'algorithme de détection de THGs procédant à la réconciliation des arbres d'espèces T et de gène T'

itèrent sur chaque paire de l'arbre des espèces et d'un arbre génétique, et résumé les résultats pour tous les arbres. Le cas de RIATA commence par calculer le MAST(T') de l'arbre des espèces et de l'arbre du gène ; l'arbre T' forme la base pour détecter et reconstruire les événements HGT.

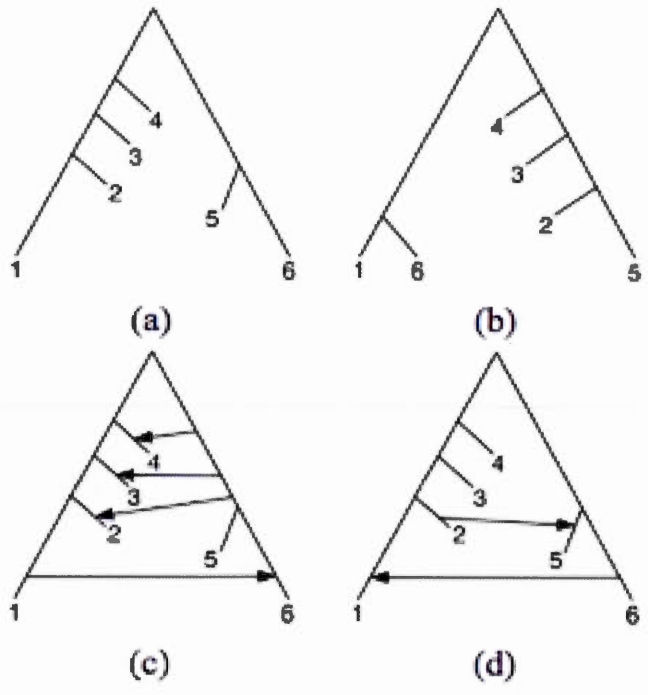


Figure 1.19: (a) Arbre des espèces (a) et un arbre de gène (b) transféré horizontalement(Nakhleh *et al.*, 2005).

Les deux réseaux dans (c) et (d) sont formés sur la base de l'arbre d'espèce ST, et les deux induisent l'arbre de gène GT. Cependant, même si les scénarios de HGT réels qui ont eu lieu sont décrits par le réseau en (c), le problème de reconstruction du HGT cherche la solution en (d).

1.12 Méthode probabiliste à la phylogénie

La théorie moderne des probabilités utilise le langage des ensembles pour modéliser une expérience aléatoire, il est toujours possible d'associer à une variable aléatoire une probabilité et définir ainsi une loi de probabilité. Lorsque le nombre d'épreuves augmente indéfiniment, les fréquences observées pour le phénomène

étudié tendent vers les probabilités et les distributions observées vers les distributions de probabilité ou loi de probabilité. Identifier la loi de probabilité suivie par une variable aléatoire donnée est essentiel, car cela conditionne le choix des méthodes employées pour répondre à une question biologique donnée.

1.12.1 Loi de poisson

La loi de Poisson, du nom de Siméon Denis Poisson (années 1800), intervient essentiellement dans deux configurations : Elle s'applique aux phénomènes considérés comme rares comme des mutations sur des régions fortement conservées d'un génome. La distribution de Poisson est foncièrement asymétrique, mais de moins en moins il tend à ressembler à la loi binomiale lorsque son paramètre λ augmente. Plus précisément, on démontre qu'une suite de variables aléatoires Xn obéissant à la loi binomiale $B(n, p)$ où p varie avec n de façon que np tend vers une limite λ pour n infini, converge en loi vers la loi de Poisson de paramètre λ.

Ainsi, si l'on admet (i) que la fréquence de mutations par locus et par génération est faible ; (ii) que le nombre de générations considéré est grand, alors le nombre de mutations attendues dans l'arbre est donné par une loi de Poisson. La loi de Poisson dérive de la loi binomiale lorsque le nombre d'essais est très grand et la probabilité d'observer un succès (ici le taux de mutation) est très petite. Ainsi, on peut montrer que, sachant μ le taux de mutations par génération et par locus, le nombre k de mutations après t générations suit une distribution décrite par une loi de Poisson.

L'application de ces méthodes probabilistes à la phylogénie fut suggérée tout d'abord par Edwards et Cavalli-Sforza en 1964. Mais à cette époque, les problèmes de calcul posés étaient insolubles et on s'en tenait à des approximations simplifiées. Au fur et à mesure que les capacités de calcul croissaient, les modèles

devenaient plus proches de la réalité biologique. (Price *et al.*, 2009) L'identification de l'ascendance des segments chromosomiques et l'ascendance distincte a un large éventail d'applications de la cartographie des maladies à l'apprentissage de l'histoire. Ainsi certaines propriétés sont exploitées et HAPMIX implémente des techniques HMM standard pour déduire efficacement les probabilités postérieures des états sous-jacents. Comme la recombinaison se produit à chaque génération, l'ascendance du modèle passe comme un processus de Poisson le long du génome. Puisque l'ascendance ne peut pas être observée directement, il est naturel de considérer le statut d'ascendance sous-jacente comme l'information « cachée » dans un HMM. Cette approche infère de manière probabiliste un état caché à chaque position le long d'un chromosome. La loi de poisson permet de calculer la probabilité de transition entre deux états ce qui permet de calculer la vraisemblance des données observées dans la progéniture pour chaque état sous-jacent possible.

La probabilité d'observer les données expérimentales sous le modèle choisi dépend de la valeur des paramètres de ce modèle. La méthode du maximum de vraisemblance détermine ces paramètres pour que cette probabilité soit la plus grande possible.(L'application de ces méthodes probabilistes à la phylogénie "université tours") L'évolution des caractères dans des branches d'arbre. Avec la méthode du maximum de vraisemblance, en tenant compte de la longueur des branches, néanmoins avec quelques hypothèses supplémentaires, d'autres informations peuvent être utilisées pour inférer ce qu'il a pu se passer à un site donné. Sur l'ensemble d'une séquence assez longue, ce sont soit des automorphismes soit des homoplasies qui sont requises par l'arbre le plus parcimonieux. Autrement dit, la longueur des branches et donc l'espérance pour chaque caractère sont basées sur l'information collectée sur l'ensemble des caractères. Ce qui permet de calculer la probabilité de changement par site sur chaque branche de l'arbre, en admettant que ces changements soient rares et que leur probabilité est donnée par une loi de Poisson :

$$P(X = k) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{1.10}$$

Qui a pour moyenne Λ de caractère et k nombre de changement. On obtient ainsi la probabilité de l'arbre le plus parcimonieux compte tenu de la longueur des branches estimées sur tous les caractères.

1.12.2 Loi géométrique

Pour deux séquences portées par deux individus différents (deux lignées), on peut à la génération précédente observer un événement de coalescence ou non. La probabilité d'observer un tel événement est de *1/N*. Notons qu'une population de séquences diploïdes est de taille *2N*, car N est le nombre d'individus de la population. Tous les résultats qui vont suivre sont démontres pour une population haploïde, mais en remplaçant N par *2N*, on obtient les résultats pour les diploïdes. Ainsi donc, les deux séquences ont une probabilité $x = \frac{1}{N}$ de coalescer à la génération précédente et $y = (1 - 1/N)$ de ne pas coalescer. Sachant que $\frac{1}{N}$ est une petite valeur (car bien souvent la taille de la population est grande, $N >>> 1$), il y a une forte probabilité pour que les deux lignées ne coalescent pas. Dans ce cas, il existe une nouvelle chance de coalescer une génération encore auparavant. Comme chaque génération est indépendante l'une de l'autre, on peut calculer que la probabilité d'avoir un ancêtre commun deux générations dans le passé est $x*y$ et celle d'avoir un ancêtre commun ni a la première ni à la seconde est $(1 - 1/N)$. Cette distribution de probabilité est bien connue, il s'agit d'une distribution géométrique qui sous sa forme générale est $P(n) = p*(1-p)^(n-1)$. On peut montrer que la moyenne de cette distribution est *E[n]=1/p* et sa variance *V[n]=(1-p)/p2*. Intuitivement, on comprend que s'il y'a 1/6 de chance de sortir un 4 au dé, on attend en moyenne 6 coups de dé pour faire le premier 4.

Les méthodes probabilistes traitent des concepts de distance permettant de quantifier la ressemblance globale(Darlu et Tassy, 1993). Les propriétés des distances qui sont essentielles dans la perspective phylogénétique. Trois méthodes principales sont ensuite évoquées parmi eux : les méthodes de vraisemblance qui analysent des matrices de ressemblance entre UE prises deux à deux seront traitées à propos des méthodes probabilistes. Parmi toutes ces méthodes, on peut distinguer certains cas qui sont abordés comme un processus géométrique. Plusieurs fonctions peuvent être mises en application pour tenter de prendre en compte tous les événements le long des branches, mais cachées à l'observation. Leur finalité est de trouver une relation entre distance patristique et distance observée. Toutes ces fonctions nécessitent des hypothèses sur les probabilités des événements et leur occurrence en fonction du temps. Mais démontrer le fondement de l'hypothèse évolutive est parfois difficile, et des propriétés de la distance utilisée dépendent la qualité de l'inférence des distances patristiques à partir des distances observées. Un exemple simple de correction peut être donné à partir de la distance déduite de l'indice de similitude de Sokal et Michener. Supposons que la probabilité d'observer une différence d'état entre k et i soit égale à π. Soit f la probabilité à priori pour que k soit dans l'état A et $(1 - f)$ dans l'état B. La probabilité d'observer simultanément l'état A chez i et chez j est calculée ainsi :

— si l'ancêtre k est dans l'état a (probabilité f). Alors il ne faut pas observer de changements entre k et i d'une part (probabilité $1 - \pi$) ni entre k et j (probabilité $1 - \pi$).

— si l'ancêtre est dans l'état b (probabilité 1-f). Dans ce cas il faut observer indépendamment une différence entre k et i (probabilité π) et entre k et j (probabilité π). D'où : la probabilité d'observer l'état A chez i et l'état B chez j (ou l'inverse) est donnée par une distribution géométrique : $pab = pba = \pi(1 - \pi)$.

1.13 Conclusion

Avoir des informations précises sur le déroulement de l'évolution naturelle des êtres humains est évidemment intéressant, mais la reconstruction de séquence ancestrale à partir d'arbre phylogénétique peut également avoir des applications pratiques comme dans certaines études environnementales. Par exemple, en biologie, on se sert fréquemment des arbres phylogénétiques pour savoir quelles espèces ou populations doivent être prioritairement sauvées de l'extinction. Dans le cas de la médecine et les bios, la phylogénie peut également permettre d'avoir des informations sur l'émergence et l'évolution d'un agent pathogène, et ainsi donner des solutions pour éradiquer celui-ci, ou réduire le risque d'infection. Un objectif central de l'analyse phylogénétique est d'estimer les relations évolutives et les paramètres dynamiques qui sous-tendent le processus de branchement évolutionnaire comme le cas des maladies épidémiologiques à partir des données moléculaires. Les méthodes statistiques utilisées dans ces analyses nécessitent que le processus de branchement d'arbre sous-jacent soit spécifié.

L'utilisation des différentes méthodes de reconstruction ancestrale présente des avantages et inconvénients dont il faut tenir compte. L'utilisation de chacune de ces méthodes implique certaines hypothèses implicites ou explicite qui peut rendre la reconstruction mauvais. Différentes options sont proposées concernant le modèle d'évolution des séquences. La recherche des relations entre séquences réclame une grande prudence, la convergence des diverses méthodes vers un arbre commun est un signe favorable.

Enfin, le modèle utilisé ne représente que des simplifications de l'évolution biologique réelle. La première méthode de reconstruction ancestrale était incapable de résoudre des événements de transferts horizontaux dont il est très probable qu'on observe actuellement des exemples qui ont pu contribuer à l'évolution pas-

sée. L'accroissement des capacités de calcul permet d'espérer un affinement de ces méthodes et d'apparitions de méthodes nouvelles plus aptes à résoudre certaines questions de reconstruction.

Dans ce troisième chapitre, nous allons montrer une autre méthode de reconstruction ancestrale indel avec des transferts horizontaux de gène.

CHAPITRE II

RECONSTRUCTION ANCESTRALE AVEC THG

2.1 Introduction

Pour pouvoir reconstruire la phylogénie d'un ensemble de séquences moléculaires, il faut que ces séquences soient issues d'une même séquence ancestrale. L'importance de la similitude entre ces séquences est un facteur important pour une bonne qualité de la reconstruction. En effet, il est compliqué de faire la reconstruction ancestrale de ces séquences, si elles sont toutes semblables ou si elles ne partagent aucun point commun. Certains phénomènes de congruences peuvent aussi compliquer la reconstruction comme la duplication de gène, mais aussi le phénomène de la perte ou de transfert horizontal de gène qui peuvent rendre plus correct la reconstruction (Guindon et Gascuel, 2003).

Les deux méthodes de reconstructions de Maximum de vraisemblance ont été mises en œuvre dans les progiciels phylogénétiques PAML : ANCESCON (Strimmer et Von Haeseler, 1996), qui est un ensemble d'inférences phylogénétiques basées sur la distance et la reconstruction de séquences protéiques ancestrales qui tient compte de la variation observée des taux d'évolution entre les positions et qui décrit plus précisément l'évolution des familles de protéines ; et ANCESTOR (Zhang et Nei, 1997) calcule les précisions des séquences d'acides aminés ancestrales. Cependant, la reconstruction conjointe est très inefficace dans certaines

applications qui utilisent des algorithmes intrinsèquement lents. Ces algorithmes évaluent de nombreuses reconstructions possibles un par un. Ainsi, la durée d'exécution des programmes de reconstructions existants est exponentielle, c'est-à-dire que ces programmes ne sont pas applicables lorsque le nombre de taxons est élevé.

Des modèles d'évolutions de séquence sont régis par un modèle réversible probabiliste(Pupko *et al.*, 2000),en considérant les sites indépendamment pour les séquences d'acides aminés. Ce modèle est décrit par une matrice M, indiquant les taux de remplacement relatifs des acides aminés et un vecteur des fréquences d'acides aminés. Pour chaque branche de longueur L, la probabilité de remplacement $i \rightarrow L$, peut être calculée à partir de la décomposition en valeurs propres de la matrice M. La topologie des arbres non racinés et les longueurs de branches sont supposées connues à priori.

Ici, nous fournissons un nouvel algorithme pour la reconstruction ancestrale qui prendra en entrée un alignement de séquences, plusieurs topologies d'arbres différents et les nombres optimaux de transferts horizontaux entre chaque paire de topologies. Nous essayerons de résoudre le problème de tous les passages entre états de mêmes topologies et de topologies différentes, ce problème sera résolu en considérant le transfert horizontal de gène. Et enfin la discussion des simulations obtenues. Le temps d'exécution de notre algorithme s'adapte linéairement au nombre de séquences. Le nouvel algorithme est basé sur le schéma de programmation dynamique dont les topologies d'arbres sont différentes et le nombre optimal de transferts horizontaux est calculé par paire d'arbres.

2.2 Différentes topologies d'arbres

Dans la reconstruction ancestrale, deux aspects de la discipline nommée topologie sont essentiels :les notions de voisinage et l'équivalence de formes. Ce sont ces

idées que nous voulons évoquer lorsque nous utilisons le mot topologie à propos d'arbre. Ce type d'étude a d'abord été développé dans le cadre de la biologie, c'est pourquoi nous parlons d'arbres phylogénétiques. Cette rapide augmentation du nombre de topologies possibles en fonction du nombre de séquences étudiées est un problème majeur en reconstruction phylogénétique et chaque alignement de séquence peut avoir plusieurs topologies différentes. Nous avons trois grandes familles de méthodes de reconstruction phylogénétique que sont :

1. les méthodes de distances,
2. les méthodes de parcimonie,
3. les méthodes de maximum de vraisemblance.

Voici un aperçu de ces méthodes, la méthode dite de distance, est une méthode basée sur les similarités entre paires de séquences. Le calcul de la parcimonie d'un arbre dont la topologie est fixée se fait en parcourant une seule fois l'arbre. Pour cela, on choisit, arbitrairement une branche de l'arbre sur laquelle se trouvera la racine r de l'arbre. Les méthodes de maximum de vraisemblance utilisent un modèle mathématique du processus d'évolution des séquences pour définir la probabilité qu'une phylogénie puisse produire les séquences observées. Nous allons étudier cette méthode en utilisant le modèle de Markov caché en intégrant des transferts horizontaux de gènes.

Au travers de ces exemples simples, on a pu voir que la méthode de vraisemblance consiste d'abord à définir un modèle de Markov caché à l'aide d'un certain nombre de paramètres. Plusieurs hypothèses sont ensuite formulées à propos de ces paramètres, le nombre de paramètres à estimer n'augmente pas plus vite, ces hypothèses satisfaites pour les constructions ancestrales. Les paramètres à estimer sont les longueurs de branches et les colonnes d'alignement pour résoudre le

modèle de Markov. Il est clair que si les états des caractères aux nœuds étaient également considérés comme des paramètres à estimer, ceux-ci augmenteraient en même temps que le nombre de caractères et les estimations de tels paramètres par la méthode de vraisemblance ne seraient plus nécessairement consistantes.

Pour résoudre le problème (Diallo *et al.*, 2007) nous utilisons un type spécial de modèle de Markov caché dans l'arbre (arbre-HMM), qui est une combinaison d'un modèle de Markov caché standard et d'un arbre phylogénétique. Dans notre modèle, après l'alignement de séquence on va générer différentes topologies d'arbres enracinées par des logiciels comme Trex (RAxML). Nous montrons comment trouver le chemin le plus probable à travers l'arbre-HMM conduit au scénario le plus probable.

Exemple du problème Indel Maximum de vraisemblance, l'entrée comprend l'alignement multiple (montré à droite au format binaire) dont une colonne binaire est rajoutée au début et à la fin dans le processus complet et les topologies d'arbres avec leurs longueurs de branches. La sortie consiste en un ensemble d'insertions, de suppressions et conservations, placées le long des bords de l'arbre, expliquant les espaces (zéros) dans l'alignement. Les colonnes identiques dans l'alignement sont combinées comme dans l'exemple de la figure 2.1 *C2C4 et C3C6.*

Deux topologies d'arbres avec l'ensemble des états de probabilités valides, non nulles associés à l'alignement multiple indiqué en haut de la figure. Lorsque les bords sont étiquetés avec plus d'un caractère, les arbres représentent plusieurs états possibles.

- I : Insertion
- D : Délétion
- C : Conservation

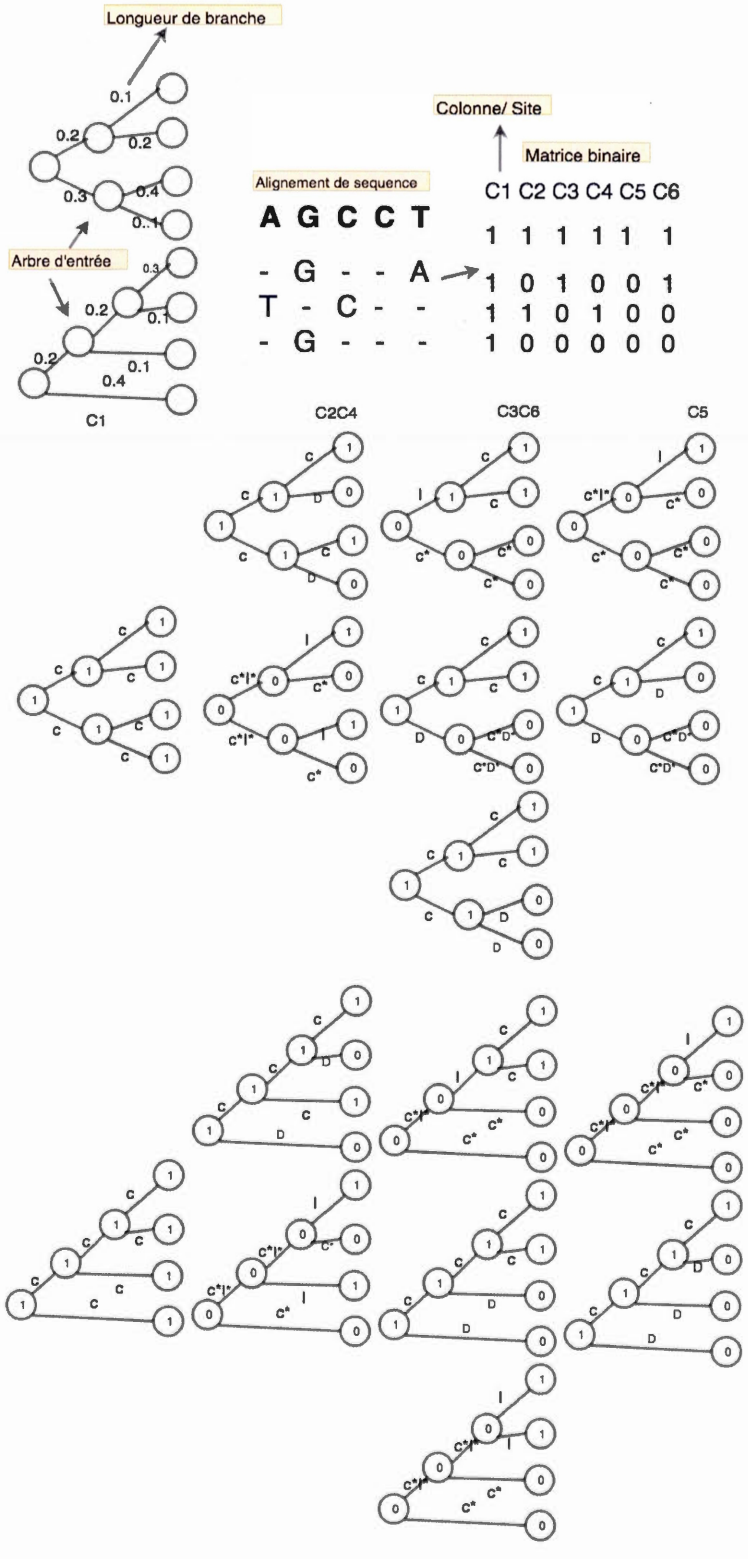


Figure 2.1: Deux topologies d'arbres avec l'ensemble des états.

2.3 Transferts horizontaux

Le transfert horizontal de gènes (THG) est un mécanisme d'évolution naturel qui consiste en un transfert direct en intégrant du matériel génétique d'une espèce à une autre. La possibilité que le transfert horizontal de gènes puisse jouer un rôle clé dans l'évolution biologique est un changement fondamental dans notre perception des aspects généraux de la biologie évolutive survenue ces dernières années.

C'est un phénomène où il arrive, par exemple, que des espèces échangent une partie de leur matériel génétique, on parle alors de transferts horizontaux(transfert latéral de gènes). Lors de la reconstruction phylogénétique, on suppose que la similitude entre les séquences est due au fait qu'elles possèdent un ancêtre commun proche ; or les transferts horizontaux augmentent la similitude entre des séquences qui n'ont pas forcément de parents communs proches. Les transferts horizontaux peuvent jouer un rôle considérable dans l'évolution des séquences et fausser complètement les reconstructions issues d'analyses de séquences moléculaires.

Les différentes étapes du transfert horizontal dans l'étude (Boc *et al.*, 2004) : Considérons deux arbres phylogénétiques enracinés : deux arbres d'espèce binaire T avec des longueurs de branche λ. Si n est le nombre de feuilles de T, il existe n - *1* des nœuds internes et *2n* - *2* bords même chose pour l'arbre T'. Ensuite on détermine un scénario de transferts horizontaux nécessaires pour transformer T en T'. Toutes les THG possibles entre les paires de branches de l'arbre T qui sont en accord avec notre modèle d'évolution sont évaluées. La procédure algorithmique s'arrête lorsqu'un nombre prédéfini de THG est ajouté dans T.

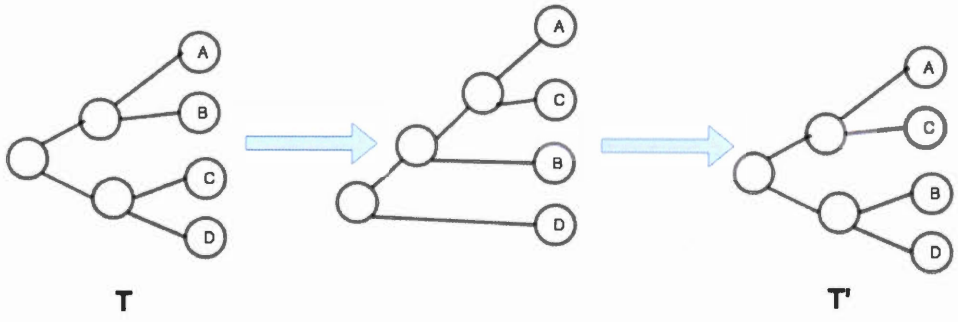


Figure 2.2: L'arbre d'espèces (T) est transformé en l'arbre de gène (T') en deux étapes.

Le calcul du nombre de transferts entre paires d'espèces se fait par beaucoup de logiciels comme HGT détection (Trex Alex). On calcul tous les transferts possibles entre paires de topologies d'arbre et ses résultats sont mis en entrée, le nombre transfert entre paires d'arbres représente k pour calculer la probabilité du nombre de transferts et de la probabilité de transition qu'on verra en dessous.

2.3.1 Nombre optimal de transferts

Les arbres fournis en entrée peuvent également contenir des informations de fiabilité sur chacune de leurs arêtes. Des algorithmes sont utilisés pour reconstruire les arbres portant sur le même ensemble de taxons. Parmi les branches des arbres de gènes incompatibles avec l'arbre des espèces, les plus soutenues correspondent aux transferts horizontaux les plus probables. Comme l'a décrit (Boc *et al.*, 2010), l'arbre phylogénétique d'espèces original T est graduellement transformé en l'arbre phylogénétique du gène T' par une série d'opérations SPR qui sont les transferts. Le but étant de trouver la plus petite séquence possible d'arbres T, T1, T2 ... , T' qui transforme T en T'.

Le processus de transformation se fait en k étape, en considérant toutes les étapes

de transfert possible (déplacement SPR), pour de grands arbres on peut avoir de milliers de déplacements. On utilise le critère d'optimisation sélectionné qui peut être dans notre cas la distance de Robinson et Foulds (RF), dont la direction de chaque THG est déterminée. Si le nombre trouver est égale a 0 même topologie et 1 topologie différente. Mais à chaque itération, le transfert diminuant le plus la distance de Robinson et Foulds entre les deux arbres d'espèce est considéré comme le plus probable, il est par la suite ajouté à l'arbre T. La procédure s'arrête quand le coefficient RF est égal à 0. La complexité temporelle de l'algorithme proposé est O(n) pour inférer k transferts servant à réconcilier une paire de phylogénies d'espèces et de gène avec n feuille.

L'algorithme de détection a au plus n-3 étapes, et nécessite au plus n-3 transferts (n-3 opérations SPR), pour transformer un arbre binaire d'espèces T avec n feuilles en un arbre binaire T' défini sur le même ensemble de feuilles. Il est difficile de trouver un nombre optimal de transferts dépassant cinq déplacements.

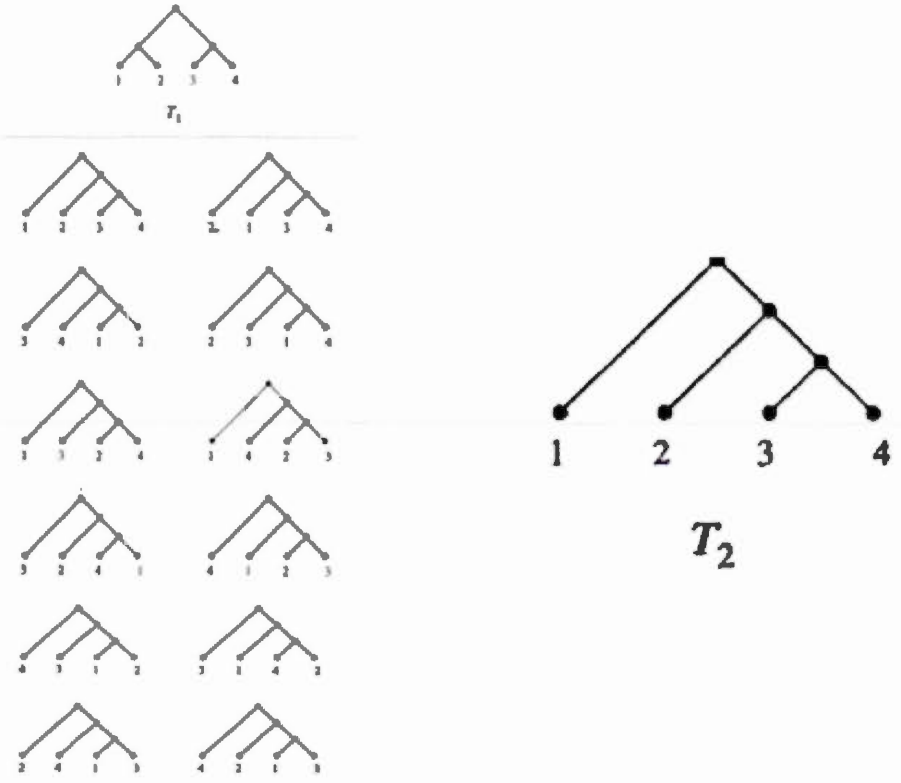


Figure 2.3: Exemple de 12 topologies a quatre espèces

Exemple d'une topologie arborescente qui affecte le nombre d'arbres dans un SPR à partir d'un arbre binaire enraciné. L'arbre T1 a 12 autres arbres à distance de SPR comme illustrant dans la figure 2 3 ci-dessus. Le nombre de transferts permet de calculer la probabilité du nombre de transferts par rapport au nombre de transferts dans le parti de la transition entre états.

2.4 Matériels et Méthodes

2.4.1 Probabilité d'émission

L'émission d'un état observé n'est pas déterministe ! Chaque état caché émet de manière aléatoire *1* parmi N symboles d'un alphabet selon une certaine distribution de probabilité d'émission. Dans nos arborescences HMM (Diallo *et al.*, 2007), chaque état émet une collection de symboles correspondant à l'ensemble des caractères obtenus aux feuillets de T lorsque le scénario indel se produit. Intuitivement, on peut penser qu'un état émette une colonne d'alignement. Chaque topologie va suivre le même scénario de calcul de probabilité d'émission.

La définition suivante officialise cela.

Soit C une colonne d'alignement (c'est-à-dire une affectation de 0 ou 1 à chaque feuille dans T). Nous avons alors la probabilité d'émission dégénérée suivante pour l'état s.

$$Pre(C|s) = \begin{cases} 1 & \text{si } \{etat(x) = col(x)\} \; pour \; tous \; x \in feuille(T) \\ 0 & \text{sinon.} \end{cases}$$

2.4.2 Probabilité de transition

Un problème clé qui rend difficile de la reconstruction de séquence ancestrale est que les données de séquences alignées à priori ayant subi des processus évolutifs impliquant des insertions et des délétions. Considérez la figure 3.4 qui représente un ensemble d'états (arbre) dans lequel chaque feuille est associée à 1 ou 0, où la chaîne évolue par des processus d'insertion, de suppression et de substitution le long de chaque branche de l'arbre. Même si l'on considère des modèles évolutifs qui

sont stochastiquement indépendants le long des branches de l'arbre (conditionnant les états ancestraux), le problème inférentiel consistant à inférer des voies évolutives (conditionnant les données observées aux feuilles de l'arbre) des branches de l'arbre. Plutôt, les décisions de transfert horizontal de gène prises dans tout l'arbre peuvent influencer la distribution des longueurs de branches et la topologie de l'arbre. La probabilité de transition peut se calculer de deux manières différentes : arbre de même topologie et topologie diffèrent (voir figure ci-dessous) ; et pour détecter les topologies d'arbre, il nous faut un algorithme qui nous permet de détecter les transferts horizontaux de gène. Dans cette figure, le C1 .... C6 représentent les colonnes ou sites comme dans la figures 2.1

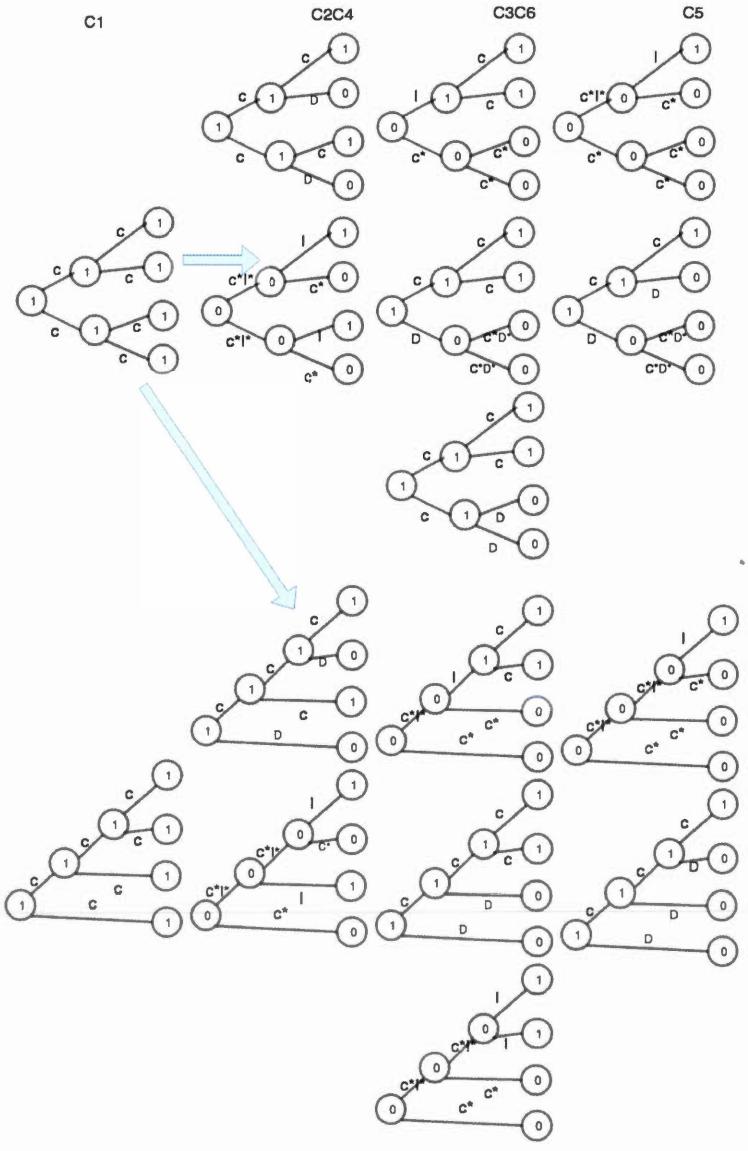


Figure 2.4: Exemple de transition entre états

2.4.3 Détection du transfert horizontal de gène

La présence de transfert peut rendre deux arbres complètement incongruents. En effet, ils peuvent ne pas avoir de nœud interne en commun et donc aucune histoire

évolutive commune. Par conséquent, le mécanisme de transfert horizontal peut complètement altérer la topologie d'un arbre phylogénétique. Pour détecter si on a un transfert de gène, on utilise la distance topologique de Robinson et Foulds qui mesure la différence entre les phylogénies d'espèces dans notre cas.

La distance de Robinson et Foulds (RF) représente le nombre minimum d'opérations élémentaires de contraction et d'expansion de nœuds nécessaires pour transformer un arbre phylogénétique en un autre. On utilise cette distance de Robinson et Foulds comme critère topologique, pour savoir si dans le processus on a un transfert horizontal si oui on utilise la deuxième méthode pour calculer la probabilité de transition sinon on utilise la première méthode.

$$RF = \begin{cases} = 0 & \text{si } T = T' \text{ on utilise la première méthode(1)} \\ \neq 0 & \text{si } T \neq T' \text{ on utilise la deuxième méthode(2)} \end{cases}$$

1- si T=T' : Pour calculer la probabilité de transition on utilise la méthode de (Diallo *et al.*, 2007). Car, dans cette méthode toutes les topologies d'arbres (états) sont identiques.

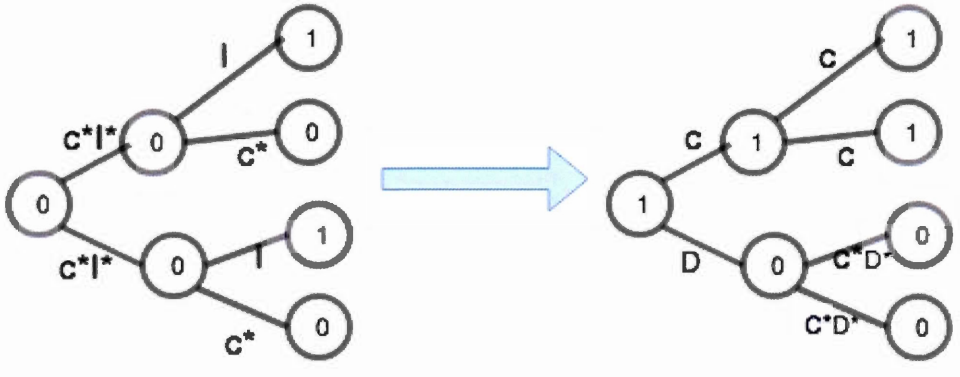


Figure 2.5: transition entre états identique T et T'

la probabilité de transition s'écrit :

$$Ptrans(T, T') = \prod_{k=1} (P(T'(e))|P(T(e)), b) \tag{2.1}$$

2- Si T≠ T' : pour calculer la probabilité de transition on utilise la distribution géométrique et de poisson.

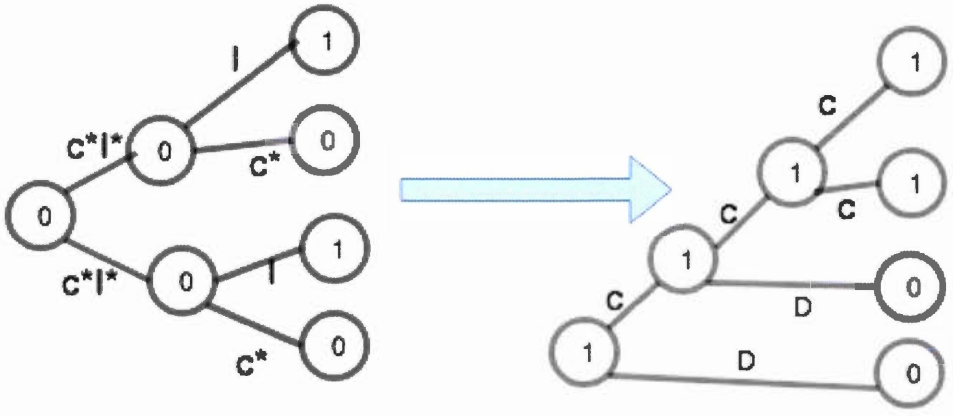


Figure 2.6: transition entre états non identiques T et T'

(a) On utilise une forme de distribution géométrique pour la transition des deux arbres T et T' et la loi de poisson pour calculer la probabilité du nombre de transferts.

2.4.3.1 Distribution géométrique pour la transition

P : la probabilité de transition calculée initialement

q : la probabilité du nombre de transferts

$$Ptrans(T, T') = pq \tag{2.2}$$

La probabilité de transition initiale est celle calculer dans le cas de la même topologie suivant la même colonne que celui de topologie différentes.

2.4.3.2 Loi de poisson pour les transferts

L'utilisation des méthodes probabilistes oblige à émettre de façon explicite des hypothèses a priori sur les probabilités de transformations des arbres.

La loi de poisson s'applique par branche d'arbre ou chaque arbre est appliqué par la loi poisson avec les paramètres λ (nombre de branches) et k (nombre de transfert). Dans cette partie on calcule la probabilité du nombre de transferts par rapport au nombre de branches.

$$P(k;\lambda) = e^{-\lambda}\frac{\lambda^k}{k!} \tag{2.3}$$

Nous utilisons un processus de Poisson pour introduire une variation des probabilités de transition entre les arbres. Les événements se produisent selon un processus de Poisson avec le paramètre λ. Lorsqu'un événement se produit, la probabilité de transition juste avant l'événement de transfert (T) est multipliée par la probabilité du nombre de transferts (T') pour produire une nouvelle probabilité de transition entre deux topologies différentes.

La probabilité de transition des deux états successifs est :

$$Ptrans(T,T') = [\prod_{k=1}(P(T'(e))|P(T(e)),b)][e^{-\lambda}\frac{\lambda^k}{k!}] \tag{2.4}$$

On appelle déplacement le passage d'un état s à un état s' ; la probabilité de se déplacer d'un état s à un état s' est notée Pss' (appelée probabilité de transition), est une fonction de l'ensemble des événements qui se sont produits le long des

bords de T. Cette probabilité de transition est une fonction de deux distributions, à savoir, géométrique et poisson.

L'application au problème de la reconstruction ancestral passe par la construction d'une chaîne de Markov dont chaque pas va impliquer une modification aléatoire de la topologie et des longueurs de branches de l'arbre ainsi que des paramètres du modèle d'évolution des séquences avec les écarts(gaps) dans l'alignement de séquence. Les probabilités généralement utilisées pour estimer la distribution de probabilité postérieure des arbres par comparaison d'hypothèses évolutives du nombre de transferts horizontaux basé sur la différence des topologies d'arbres en compétition pour un même jeu de données. Les probabilités généralement utilisées pour estimer la distribution de probabilité postérieure des arbres phylogénétiques sont la probabilité d'émission et la probabilité de transition pour trouver le meilleur chemin de la reconstruction ancestral.

2.5 Simulation des données

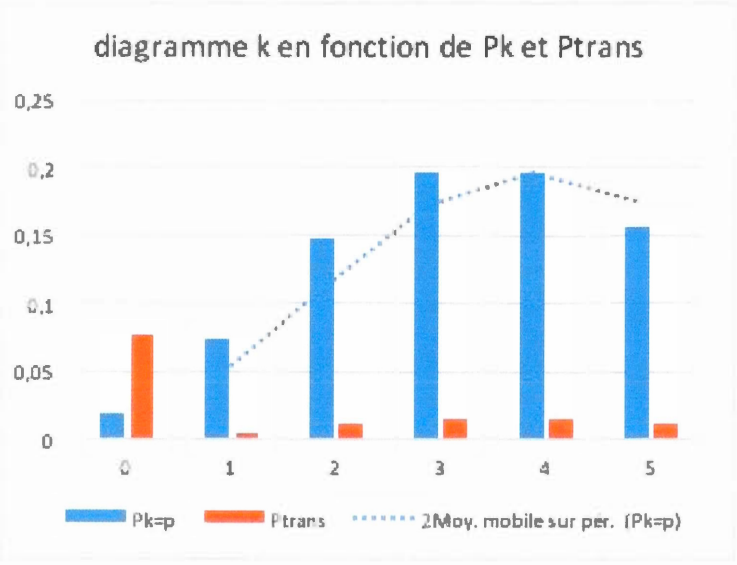
Le changement de l'arbre 1 en l'arbre 2 peut augmenter le nombre de transferts, un tel échange possible augmente la probabilité de l'arbre. Cependant, les arêtes ne sont pas indépendantes et lorsqu'on modifie simultanément deux arêtes avec des valeurs, nous ne pouvons pas être sûrs que la probabilité de l'arborescence augmentera. L'approche standard consiste à effectuer une modification de l'arbre standard. Après chaque modification, le nombre de transferts est mis à jour. Notre approche implique d'abord de calculer de manière indépendante toutes les modifications, c'est-à-dire les longueurs de toutes les branches et les échanges possibles autour de toutes les branches internes, puis d'appliquer simultanément la plupart de ces modifications. Les modifications des longueurs de branches sont prises en compte dans chaque topologie, qui sont générées pour calculer la première mé-

thode de probabilité de transition dont on part par arbre de même topologie.

Plusieurs mécanismes différents ont été utilisés pour mettre à jour l'état de la simulation. Ces mécanismes comprenaient : (1) l'ajout d'un événement à l'arbre (2) la suppression d'un événement de l'arbre (car chaque insertion de branche est suivie d'une délétion) (3) la modification de la position de branche sur l'arbre (4) la modification de la valeur du nombre de transferts (5) changement de la probabilité de transfert pour chaque cas de simulation (6) changer le paramètre de Poisson (λ) (7) mettre à jour les nouvelles probabilités de transfert (P_{trans}). Ainsi on a seulement fait ces trois simulations, car il est difficile de voir plus de cinq transferts entre deux espèces ce qui fait que les deux simulations auraient pu être suffisant pour modéliser le problème, mais on a ajouté une troisième pour voir l'impact quand le nombre de transferts dépasse cinq.

Tableau 2.1: Simulation 1

itération	k	λ	P_k	Ptrans
1	0	4	0,018	0,077
2	1	4	0,073	0,005
3	2	4	0,146	0,011
4	3	4	0,195	0,015
5	4	4	0,195	0,015
6	5	4	0,156	0,012



(a) Diagramme 1

Symbole	Description
k	Nombre de transferts horizontaux
λ	Nombre de branches
Pk	Probabilité du transfert
$Ptrans$	Probabilité de transition

(b) Légende 1

Figure 2.7: Diagramme 1 de k en fonction de P_k et P_{trans}

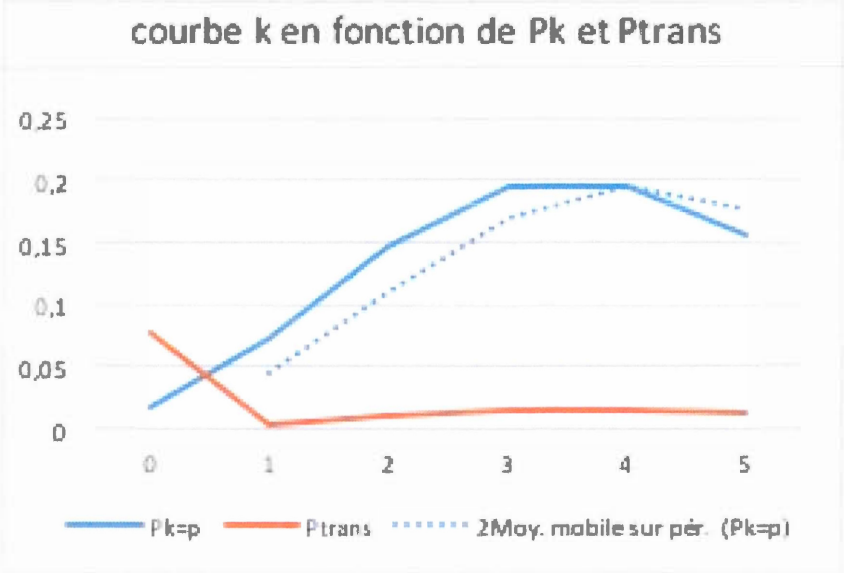
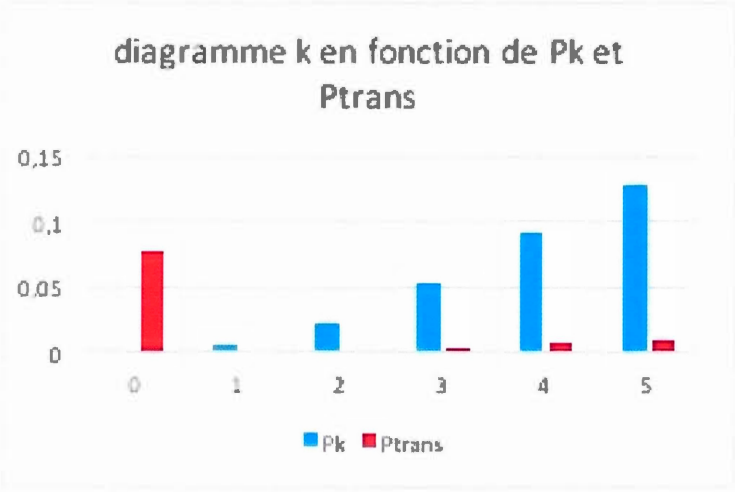


Figure 2.8: Courbe 1 de k en fonction de P_k et P_{trans}

H

Tableau 2.2: Simulation 2

itération	k	λ	P_k	P_{trans}
1	0	7	0,0009	0,0076
2	1	7	0,006	0,00004
3	2	7	0,022	0,00016
4	3	7	0,052	0,00039
5	4	7	0,091	0,00069
6	5	7	0,127	0,00096

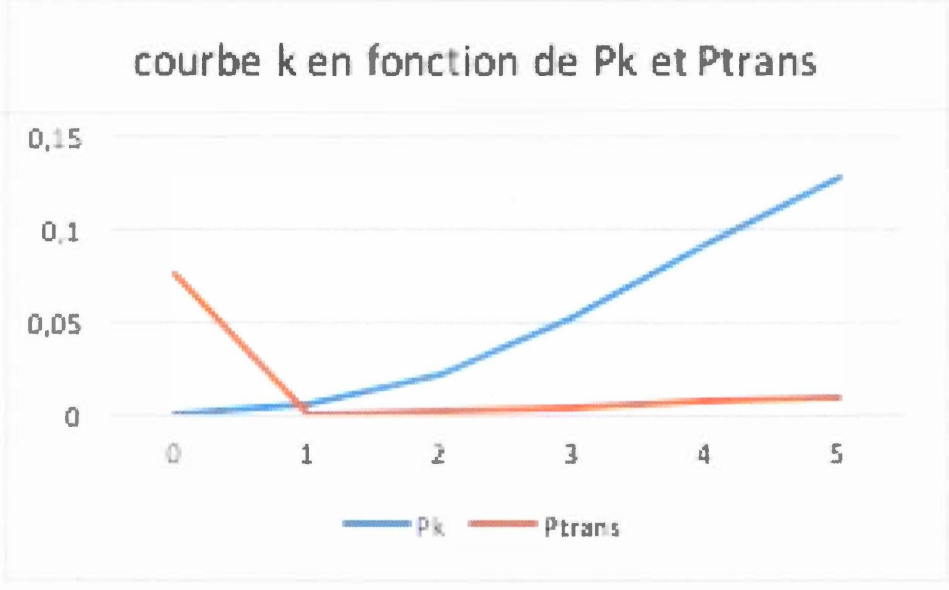


(a) Diagramme 2

Symbole	Description
k	Nombre de transferts horizontaux
λ	Nombre de branches
Pk	Probabilité du transfert
$Ptrans$	Probabilité de transition

(b) Légende 2

Figure 2.9: Diagramme 2 de k en fonction de P_k et P_{trans}



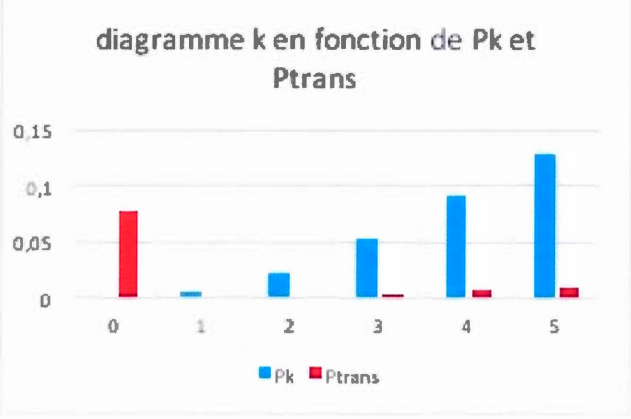
H

Figure 2.10: Courbe 2 de k en fonction de P_k et P_{trans}

H

Tableau 2.3: Simulation 3

itération	k	λ	P_k	Ptrans
1	0	11	0,0004	0,0075
2	1	11	0,005	0,0004
3	2	11	0,017	0,00015
4	3	11	0,04	0,00033
5	4	11	0,085	0,00062
6	5	11	0,09	0,0009
7	6	11	0,082	0,00097
8	7	11	0.076	0,001



(a) Diagramme 3

Symbole	Description
k	Nombre de transferts horizontaux
λ	Nombre de branches
Pk	Probabilité du transfert
$Ptrans$	Probabilité de transition

(b) Légende 3

Figure 2.11: Diagramme 3 de k en fonction de P_k et P_{trans}

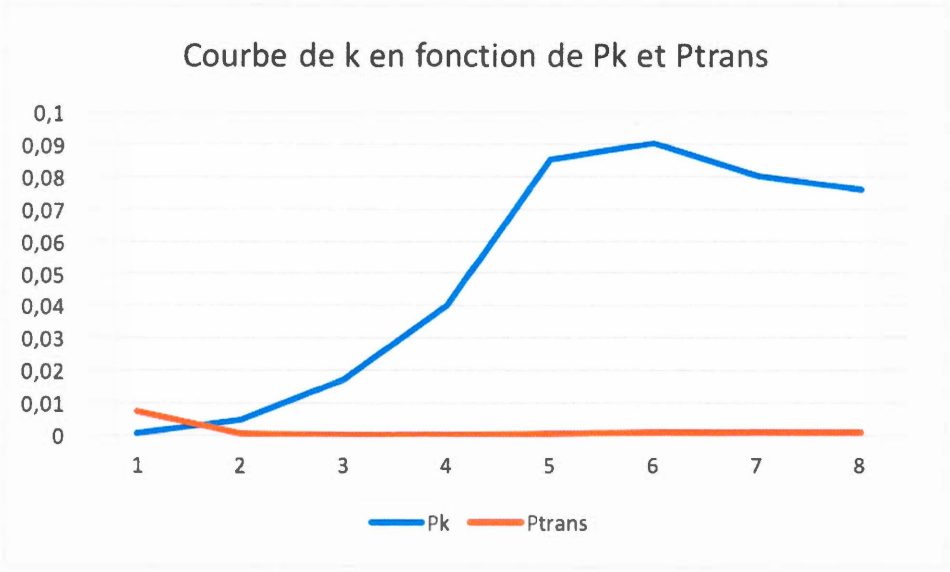


Figure 2.12: Courbe 3 de k en fonction de P_k et P_{trans}

Nous avons présenté un modèle de reconstruction à valeur de chaîne pouvant être utilisé pour l'inférence conjointe des phylogénies et des alignements multiples de séquences. Notre modèle peut être utilisé pour capturer l'homologie des caractères évoluant sur un arbre phylogénétique lors d'événements d'insertion, de suppression et de conservation. L'algorithme de programmation dynamique garantit la recherche de l'ensemble des états (arbres), une colonne par arbre. La complexité de l'algorithme est linéaire. C'est-à-dire que son temps de fonctionnement par site est proportionnel au nombre de branches pour les probabilités de transitions et que son efficacité permet son utilisation pour un nombre illimité d'arbres. Notez que notre algorithme minimise la probabilité sur toutes les combinaisons possibles.

L'avantage de la nouvelle méthode est qu'il permet une représentation sous la forme d'un processus de Poisson sur l'arbre pour évaluer les transferts horizontaux de gènes. Les représentations de Poisson ont joué un rôle important dans les processus de substitution pure ((Alekseyenko *et al.*, 2008; Miklós, 2003)), mais

dans ce travail, nous utilisons des représentations de Poisson pour les transferts horizontaux.

2.6 Discussion

Nous avons initialement étudié deux questions : la manière dont la précision de la reconstruction dépend des branches impliquées et du nombre de transferts nécessaires pour une reconstruction précise. Afin d'explorer ces questions, nous avons préparé des ensembles de données artificiels en modélisant le processus de changement d'arbre. Nous avons commencé avec 0 transfert horizontal jusqu'à 7 transferts pour mieux expliquer le processus.

Les figures présentées ci-dessus montrent les résultats d'analyses : Dans les trois itérations, on a λ constant pour chaque cas. Les estimations de λ (le paramètre de forme du processus de Poisson) et k ont changé lorsque différentes valeurs de paramètres ont été utilisées.

Il est important de remarquer que ces probabilités de transition sont en fonction de la quantité des probabilités de transferts entre deux arbres, et sont dépendantes du modèle précédent. Autrement dit, on considère dans cet algorithme qu'un événement a autant de chance de produire un meilleur résultat quand le nombre de transferts est petit ce qui implique une probabilité de transfert plus petit. En conséquence, une différence de nombre de transferts entre deux arbres peut correspondre à un nombre petit de changements de branche, sans que la probabilité de transfert et de transition soit différente comme on le voit dans le tableau 1, quatrième et cinquième itération dont les nombres de transferts ne sont pas les mêmes. Les possibilités de « changements multiples, plus de transfert que de branche comme dans le tableau 1 sixième itération » ne sont donc pas prises en compte, car sont des événements rares ou même impossibles, et la probabilité

de réalisation de l'événement dépend des arbres, mais ses événements peuvent donner de bons résultats comme le montre la courbe *3.8* en pointillé. L'analyse de ces figures montre que les probabilités avec des transferts horizontaux de deux états successifs donnent de meilleurs résultats que les probabilités initiales (sans transfert).

La précision topologique de ces différentes méthodes sera mesurée sur l'ensemble des arbres par la distance standard de Robinson et Foulds entre l'arbre inféré et l'arbre réel. Cette distance correspond à la proportion de branches internes trouvées dans un arbre et non dans un autre. Sa valeur va de 0 (les deux topologies sont identiques) à 1 (elles ne partagent aucune branche en commun). La valeur de cette distance a été trouvée en fonction de la divergence maximale par paire dans l'ensemble de données considéré, la divergence entre deux séquences étant simplement la proportion de sites où les deux séquences diffèrent. Dans un ensemble de données de séquence, on fait l'alignement de toutes les séquences et après on infère tous les arbres possibles dans cette séquence. À l'intérieur de l'algorithme, la distance RF nous permet de passer d'une topologie d'arbre à l'autre (Guindon et Gascuel, 2003).

Le processus de Poisson est l'un des modèles les plus utilisés en biologie et il existe plusieurs extensions possibles de l'approche que nous décrivons en reconstruction. Par exemple, on peut placer des événements sur un arbre selon un processus de Markov(Huelsenbeck *et al.*, 2000). Nous pensons que la méthodologie du maximum de vraisemblance présentée dans ce mémoire se révélera être une étape vers une telle base théorique, même si le modèle actuel présente des limitations. Une des forces de cette analyse probabiliste est notre capacité à représenter des degrés de certitude et à inclure diverses possibilités ainsi que les probabilités calculées. Bien que de nombreux progrès aient été accomplis dans ce domaine, de nombreux problèmes restent non résolus. Un modèle incorporant d'autres événements évolu-

tifs communs (par exemple, recombinaison, inversion) en plus des insertions, des délétions et des conservations serait une amélioration. Il serait également utile de permettre au contexte local de l'ADN d'affecter la probabilité des événements évolutifs (Thorne *et al.*, 1992).

2.6.1 complexité

La complexité temporelle de la version interactive de l'algorithme est $O(kn^4)$ pour inférer k transferts servant à réconcilier une paire de phylogénies d'espèces avec n feuilles. L'algorithme proposé est un modèle robuste d'évolution prenant en compte le transfert horizontal de gènes. Ce modèle se basera sur la réconciliation des phylogénies d'espèces, tout en incorporant des règles d'évolution nécessaires pour une meilleure reconstruction ancestrale.

Nous avons identifié les points suivants comme étant les caractéristiques importantes d'un tel algorithme : Complexité algorithmique réduite, on s'assure de pouvoir exécuter l'algorithme dans un temps raisonnable et ainsi effectuer des études à grande échelle. Quand le nombre de feuilles ou de THG augmente, le rendent particulièrement intéressant pour l'analyse de larges phylogénies englobant plusieurs conflits topologiques dus aux transferts horizontaux de gènes. L'algorithme initial est plus économique et moins coûteux puisqu'il utilise moins de paramètres, mais ce nouvel algorithme donne de meilleurs résultats surtout dans le cas des virus.

2.7 Conclusion

Les applications que nous avons vues montrent leur puissance en génomique pour tout ce qui concerne la reconstruction ancestrale de séquence. La reconstruction ancestrale a été appliquée dans plusieurs contextes. Par exemple : Un algorithme de programmation dynamique est développé par (Pupko *et al.*, 2000) pour la re-

construction ancestrale en utilisant un formalisme de maximum de vraisemblance de l'ensemble de toutes les séquences ancestrales d'acides aminés dans un arbre phylogénétique. Aussi la méthode de parcimonie (Fitch, 1971) a été utilisée pour déduire des séquences ancestrales d'acides aminés. Mais la faiblesse de ses algorithmes c'est qu'ils ne tiennent pas compte des transferts horizontaux de gène par rapport à notre algorithme.

Pour la reconstruction de séquences ancestrales, nous avons principalement utilisé des approches standards avec une procédure de détection THG. La présence de deux transferts rend deux arbres complètement incongruents. En effet, ils n'ont aucun nœud interne en commun. Par conséquent, le mécanisme de transfert horizontal peut complètement altérer la topologie d'un arbre phylogénétique. Cette différence de topologie nous permet d'avoir deux méthodes de calcul de transition.

CONCLUSION

Dans ce mémoire, nous avons présenté un nouvel algorithme pour la reconstruction ancestrale de séquences impliquant des transferts horizontaux de gènes qui se base sur l'algorithme décrit dans (Diallo *et al.*, 2007). Nous y avons apporté des améliorations, notamment sur le plan de la complexité algorithmique et l'inclusion des phénomènes de transferts horizontaux de gènes. L'algorithme exploite efficacement les différences topologiques ou métriques entre arbres d'espèces pour un même ensemble d'organismes considérés, et fournit une réponse au problème de la modélisation de reconstruction ancestrale.

Nous avons proposé une modification de la recherche dans l'espace topologique en vue de pouvoir évaluer toutes les possibilités de reconcialiation entre l'arbre d'espèce fixé et les arbres des blocs alignés. Nous avons proposé de modéliser le nombre de transferts le long des branches comme suivant une loi de poisson. Ce qui permet de fixer des taux de transferts et de pouvoir estimer le nombre de transferts possible et de reconstruire adéquatement les scénarios ancestraux. Pour faciliter la reconciliation entre les topologies, nous avons adapté l'heuristique présentée dans (Boc *et al.*, 2010). Tout cela a permis d'avoir un cadre robuste et cohérent qui permet d'étudier les séquences ancestrales dans un cadre plus complet et réaliste. Les simulations présentées dans le mémoire, ont permis de montrer la pertinence de ce cadre.

La complexité algorithmique de cette méthode est $0(m^5)$ pour ajouter r transferts horizontaux de gène dans un arbre phylogénétique d' espèces à n feuilles. Les simulations que nous avons réalisées avec cet algorithme ont montré qu'il est

plus performant que la méthode décrite dans (Diallo *et al.*, 2007), mais généralement moins performant en terme de temps d'exécution car cette nouvelle méthode demande plus réplication d'arbre que celui de Diallo.

Dans le futur, nous comptons l'appliquer sur les données réelles dans la perspective de reconstruction des familles bactériennes. Comme la plupart des organismes, les bactéries subissent beaucoup plus d'échanges génétiques que d'autres, restreints à leur hôte. Leurs génomes incluent quelques portions du chromosome, appelées ilôts génomiques, où de nombreux échanges génétiques et de transferts de gènes peuvent avoir lieu, le transfert jouent un rôle majeur dans leur diversification. Nous comptons aussi inclure l'algorithme dans le programme Ancestors accessible en ligne (Diallo *et al.*, 2009).

RÉFÉRENCES

Alekseyenko, A. V., Lee, C. J. et Suchard, M. A. (2008). Wagner and dollo : a stochastic duet by composing two parsimonious solos. *Systematic biology*, *57*, 772–784.

Anselmetti, Y. (2017). *Reconstruction conjointe de l'ordre des gènes de génomes actuels et ancestraux et de leur évolution structurale dans un cadre phylogénétique.* (Thèse de doctorat). Université de Lyon.

Ashkenazy, H., Penn, O., Doron-Faigenboim, A., Cohen, O., Cannarozzi, G., Zomer, O. et Pupko, T. (2012). Fastml : a web server for probabilistic reconstruction of ancestral sequences. *Nucleic acids research*, *40*, W580–W584.

Beanland, T. J. et Howe, C. J. (1992). The inference of evolutionary trees from molecular data. *Comparative Biochemistry and Physiology Part B : Comparative Biochemistry*, *102*, 643–659.

Blanchette, M., Green, E. D., Miller, W. et Haussler, D. (2004a). Reconstructing large regions of an ancestral mammalian genome in silico. *Genome research*, *14*, 2412–2423.

Blanchette, M., Kent, W. J., Riemer, C., Elnitski, L., Smit, A. F., Roskin, K. M., Baertsch, R., Rosenbloom, K., Clawson, H., Green, E. D. *et al.* (2004b). Aligning multiple genomic sequences with the threaded blockset aligner. *Genome research*, *14*, 708–715.

BOC, A. (2012). Détection des transferts horizontaux de gènes : Modèles et algorithmes appliqués à l'évolution des espèces et des langues.

Boc, A. (2016). Détection des transferts horizontaux de gènes.

Boc, A. et Makarenkov, V. (2003). New efficient algorithm for detection of horizontal gene transfer events. Dans *International Workshop on Algorithms in Bioinformatics*, 190–201. Springer.

Boc, A., Makarenkov, V. et Diallo, A. B. (2004). Une nouvelle méthode pour la détection de transferts horizontaux de gène : la réconciliation topologique d'arbres de gène et d'espèces. *te proceedings of JOBIM.*

Boc, A., Philippe, H. et Makarenkov, V. (2010). Inferring and validating horizontal gene transfer events using bipartition dissimilarity. *Systematic biology*, *59*, 195–211.

Bouchard-Côté, A. et Jordan, M. I. (2013). Evolutionary inference via the poisson indel process. *Proceedings of the National Academy of Sciences*, *110*, 1160–1166.

Bourguignon, P. et Robelin, D. (2004). Modeles de markov parcimonieux : sélection de modele et estimation. *Proceedings of JOBIM. Montréal.*

Bréhélin, L. et Gascuel, O. (2000). Modèles de markov cachés et apprentissage de séquences. *Le temps, l'espace et l'évolutif en sciences du traitement de l'information, Eds. Cépaduès.*

Cavalli-Sforza, L. L. et Edwards, A. W. (1967). Phylogenetic analysis : models and estimation procedures. *Evolution*, *21*, 550–570.

Courvalin, P. (1998). Plantes transgéniques et antibiotiques. *RECHERCHE-PARIS-*, 36–40.

Criscuolo, A. (2006). *Méthodes de distance pour l'inférence phylogénomique.* (Thèse de doctorat). Université Montpellier II-Sciences et Techniques du Languedoc.

Darlu, P. et Tassy, P. (1993). La reconstruction phylogénétique. *Concepts et méthodes. Paris, France.*

Daubin, V. (2002). *Phylogénie et évolution des génomes procaryotes.* (Thèse de doctorat). Université Claude Bernard-Lyon I.

Diallo, A. B., Makarenkov, V. et Blanchette, M. (2007). Exact and heuristic algorithms for the indel maximum likelihood problem. *Journal of Computational Biology*, *14*, 446–461.

Diallo, A. B., Makarenkov, V. et Blanchette, M. (2009). Ancestors 1.0 : a web server for ancestral sequence reconstruction. *Bioinformatics*, *26*(1), 130–131.

Dobson, A. J. (1974). Unrooted trees for numerical taxonomy. *Journal of applied probability*, *11*, 32–42.

Doolittle, W. F. (1999). Phylogenetic classification and the universal tree. *Science*, *284*, 2124–2128.

Edgar, R. C. (2004). Muscle : multiple sequence alignment with high accuracy and high throughput. *Nucleic acids research*, *32*, 1792–1797.

Ekelund, F., Daugbjerg, N. et Fredslund, L. (2004). Phylogeny of heteromita, cercomonas and thaumatomonas based on ssu rdna sequences, including the description of neocercomonas jutlandica sp. nov., gen. nov. *European Journal of Protistology*, *40*, 119–135.

Entfellner, J.-B. D. (2011). *Combinaison de modèles phylogénétiques et longitudinaux pour l'analyse des séquences biologiques : reconstruction de HMM profils ancestraux.* (Thèse de doctorat). Université Montpellier II-Sciences et Techniques du Languedoc.

Farris, J. S. (1970). Methods for computing wagner trees. *Systematic Biology*, *19*, 83–92.

Faure, S. (2009). *Transfert d'un gène de résistance aux beta-lactamines blaCTX-M-9 entre Salmonella et les entérobactéries de la flore intestinale humaine : influence d'un traitement antibiotique.* (Thèse de doctorat). Université Rennes 1.

Felsenstein, J. (1981). Evolutionary trees from dna sequences : a maximum likelihood approach. *Journal of molecular evolution*, *17*, 368–376.

Felsenstein, J. et Churchill, G. A. (1996). A hidden markov model approach to variation among sites in rate of evolution. *Molecular biology and evolution*, *13*, 93–104.

Felsenstein, J. et Felenstein, J. (2004). *Inferring phylogenies*, volume 2. Sinauer associates Sunderland, MA.

Fitch, W. M. (1971). Toward defining the course of evolution : minimum change for a specific tree topology. *Systematic Biology*, *20*, 406–416.

Forney, G. D. (1973). The viterbi algorithm. *Proceedings of the IEEE*, *61*, 268–278.

Garcia, C. C. et Martin-Vide, J. (1993). Analyse par la chaîne de markov de la sécheresse dans le sud-est de l'espagne. *Science et changements planétaires/Sécheresse*, *4*, 123–129.

Guindon, S. et Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic biology*, *52*, 696–704.

Guindon, S., Lethiec, F., Duroux, P. et Gascuel, O. (2005). Phyml online—a web server for fast maximum likelihood-based phylogenetic inference. *Nucleic acids research*, *33*, W557–W559.

Hao, W. et Golding, G. (2004). Patterns of bacterial gene movement. *Molecular biology and evolution*, *21*, 1294–1307.

Huelsenbeck, J. P., Larget, B. et Swofford, D. (2000). A compound poisson process for relaxing the molecular clock. *Genetics*, *154*, 1879–1892.

Katoh, K. et Standley, D. M. (2013). Mafft multiple sequence alignment software version 7 : improvements in performance and usability. *Molecular biology and evolution*, *30*, 772–780.

Koshi, J. M. et Goldstein, R. A. (1996). Probabilistic reconstruction of ancestral protein sequences. *Journal of molecular evolution*, *42*, 313–320.

Larkin, M. A., Blackshields, G., Brown, N., Chenna, R., McGettigan, P. A., McWilliam, H., Valentin, F., Wallace, I. M., Wilm, A., Lopez, R. *et al.* (2007). Clustal w and clustal x version 2.0. *bioinformatics*, *23*, 2947–2948.

Lawrence, J. G. et Ochman, H. (1997). Amelioration of bacterial genomes : rates of change and exchange. *Journal of molecular evolution*, *44*, 383–397.

Liò, P. et Goldman, N. (1998). Models of molecular evolution and phylogeny. *Genome research*, *8*, 1233–1244.

Liu, K., Warnow, T. J., Holder, M. T., Nelesen, S. M., Yu, J., Stamatakis, A. P. et Linder, C. R. (2011). Sate-ii : very fast and accurate simultaneous estimation of multiple sequence alignments and phylogenetic trees. *Systematic biology*, *61*, 90–106.

Miklós, I. (2003). Algorithm for statistical alignment of two sequences derived from a poisson sequence length distribution. *Discrete applied mathematics*, *127*, 79–84.

Mirkin, B. G., Fenner, T. I., Galperin, M. Y. et Koonin, E. V. (2003). Algorithms for computing parsimonious evolutionary scenarios for genome evolution, the last universal common ancestor and dominance of horizontal gene transfer in the evolution of prokaryotes. *BMC evolutionary biology*, *3*, 2.

Nakhleh, L., Ruths, D. et Wang, L.-S. (2005). Riata-hgt : a fast and accurate heuristic for reconstructing horizontal gene transfer. Dans *International Computing and Combinatorics Conference*, 84–93. Springer.

Ng, P. C. et Henikoff, S. (2002). Accounting for human polymorphisms predicted to affect protein function. *Genome research*, *12*, 436–446.

Nguyen, D., Boc, A. et Makarenkov, V. (2005). Hgt-simulator : logiciel pour

simuler des transferts horizontaux de gènes. *proceedings of the SFC2005, Montreal*, 215–219.

Orter, J. G. (1982). Addtree/p : A pascal program for fitting additive trees based on sattath and tversky's addtree algorithm. *Behavior Research Methods*, *14*, 353–354.

Paulin, F. (1988). Topologie de gromov équivariante, structures hyperboliques et arbres réels. *Inventiones mathematicae*, *94*, 53–80.

Petit, M. et Henning, C. (2009). Un calcul de viterbi pour un modele de markov caché contraint. Dans *Cinquièmes Journées Francophones de Programmation par Contraintes, Orléans, juin 2009*, 285–295.

Philippe, H. et Douady, C. J. (2003). Horizontal gene transfer and phylogenetics. *Current opinion in microbiology*, *6*, 498–505.

Popier, A. (2005). Chaînes de markov.

Prescott, L. M., Willey, J. M., Sherwood, L. M. et Woolverton, C. J. (2018). *Microbiologie.* De Boeck Supérieur.

Price, A. L., Tandon, A., Patterson, N., Barnes, K. C., Rafaels, N., Ruczinski, I., Beaty, T. H., Mathias, R., Reich, D. et Myers, S. (2009). Sensitive detection of chromosomal segments of distinct ancestry in admixed populations. *PLoS genetics*, *5*, e1000519.

Pupko, T., Pe, I., Shamir, R. et Graur, D. (2000). A fast algorithm for joint reconstruction of ancestral amino acid sequences. *Molecular biology and evolution*, *17*, 890–896.

Ranwez, V. (2002). *Méthodes efficaces pour reconstruire de grandes phylogénies suivant le principe du maximum de vraisemblance.* (Thèse de doctorat). Université Montpellier II-Sciences et Techniques du Languedoc.

Robinson, D. F. et Foulds, L. R. (1981). Comparison of phylogenetic trees. *Mathematical biosciences*, *53*, 131–147.

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

Semeria, M. (2015). *Évolution de l'architecture des génomes : modélisation et reconstruction phylogénétique.* (Thèse de doctorat). Université Claude Bernard-Lyon I.

Siepel, A. et Haussler, D. (2004). Combining phylogenetic and hidden markov

models in biosequence analysis. *Journal of Computational Biology*, *11*, 413–428.

Simonet, P. (2000). Evaluation de l'impact environnemental : Evaluation des potentialités de transfert de l'adn des plantes transgéniques vers les bactéries du sol. *Oléagineux, Corps gras, Lipides*, *7*, 320–323.

Stamatakis, A. (2006). Raxml-vi-hpc : maximum likelihood-based phylogenetic analyses with thousands of taxa and mixed models. *Bioinformatics*, *22*, 2688–2690.

Strimmer, K. et Von Haeseler, A. (1996). Quartet puzzling : a quartet maximum-likelihood method for reconstructing tree topologies. *Molecular Biology and Evolution*, *13*, 964–969.

Sunyaev, S., Ramensky, V., Koch, I., Lathe III, W., Kondrashov, A. S. et Bork, P. (2001). Prediction of deleterious human alleles. *Human molecular genetics*, *10*, 591–597.

Swofford, D. (1993). Phylogenetic analysis using parsimony (paup), version 3.1. 1. *University of Illinois, Champaign*.

Tamura, K., Peterson, D., Peterson, N., Stecher, G., Nei, M. et Kumar, S. (2011). Mega5 : molecular evolutionary genetics analysis using maximum likelihood, evolutionary distance, and maximum parsimony methods. *Molecular biology and evolution*, *28*, 2731–2739.

Thorne, J. L., Kishino, H. et Felsenstein, J. (1991). An evolutionary model for maximum likelihood alignment of dna sequences. *Journal of Molecular Evolution*, *33*, 114–124.

Thorne, J. L., Kishino, H. et Felsenstein, J. (1992). Inching toward reality : an improved likelihood model of sequence evolution. *Journal of molecular evolution*, *34*, 3–16.

Vouma Lekoundji, J.-B. (2014). Modèles de markov cachés.

Westesson, O., Lunter, G., Paten, B. et Holmes, I. Accurate reconstruction of insertion-deletion histories by statistical phylogenetics. *PLoS One*, *7*.

Zhang, J. et Nei, M. (1997). Accuracies of ancestral amino acid sequences inferred by the parsimony, likelihood, and distance methods. *Journal of molecular evolution*, *44*, S139–S146.