De la Robustesse des Arbres Phylogénétiques

Mahendra Mariadassou
sous la direction de Avner Bar-Hen

29 janvier 2007

Introduction

Le premier génome à avoir jamais été séquencé est celui du phage ϕ-X174, en 1980. Ce génome fait 5368 kilo-bases de long. Depuis cette époque, les techniques de séquençage se sont améliorées et de nouveaux génomes sont séquencés dans leur intégralité à un taux rapide. Le génome humain, dont une première version a été complétée en 2001 dans le cadre du Human Genome Project, en constitue un remarquable exemple. Cette révolution des techniques de biologie moléculaire a considérablement augmenté la taille des bases de données à partir desquelles on peut, entre autre, faire de la phylogénie.

La phylogénie est l'étude de la formation et de l'évolution des organismes vivants en vue d'établir leur parenté. La phylogenèse est le terme le plus utilisé pour décrire la généalogie d'un ensemble d'organismes, généralement des espèces, issu d'un organisme commun.

Un arbre phylogénétique est un arbre binaire qui représente une phylogénie et montre le degré de parenté entre différents organismes, censés avoir un ancêtre commun. Un noeud à deux descendants représente le plus proche ancêtre commun de ces deux descendants, et la longueur d'une branche entre un ancêtre et un descendant représente le temps ou, dans un sens à définir plus précisément, la quantité d'évolution entre l'apparition de ces deux espèces. Chaque noeud est un *taxon*. Les feuilles de l'arbre représentent les taxons encore existant tandis que les noeuds internes représentent de taxons hypothétiques qui ont disparu ou n'ont jamais existé.

L'inférence d'un arbre phylogénétique s'apparente à la construction d'un coalescent : on part des feuilles de l'arbre (le groupe d'espèces auxquelles on s'intéresse) et on tente de positionner les «parents». Bien entendu, deux espèces proches ne doivent avoir un parent proche que si elles sont fortement similaires. La difficulté majeure de cette reconstruction est l'absence d'informations sur les parents. Par définition, les parents sont des noeuds internes de l'arbre, c'est à dire des taxons hypothétiques qui peuvent avoir existé ou pas, sur lesquels on ne dispose d'aucune information. Les relations de parenté ne peuvent donc pas être déterminées de manière certaines, mais sont plus ou moins probables.

Deux écoles de pensée existent en phylogénie : la phénétique et la cladistique. La cladistique s'appuie sur l'identification de l'homologie des caractères communs à deux

espèces : tel caractère commun est-il une homologie ou bien une homoplasie, c'est à dire hérité d'un ancêtre commun ou non ? Cette identification est difficile et n'est utilisée que pour les espèces fossiles dont l'ADN est rarement conservé. La phénétique, quand à elle, se base sur la ressemblance de deux espèces pour déterminer leur niveau de parenté. Avec l'avènement de la génomique, cette ressemblance est maintenant calculée sur les séquences d'ADN. Cette méthode possède l'avantage relatif de se baser sur des jeux de données importants, ce qui permet d'espérer des estimateurs relativement fiables.

1 Méthodes classiques de phylogénie

Formalisons en termes mathématiques le problème du choix de la phylogénie. Soit $X = (X_1, \ldots, X_n)$ un ensemble d'espèces. On dispose pour chaque espèce i d'une séquence $X_i^1, \ldots, X_i^p$ de p caractères. Le plus souvent, il s'agit de la séquence d'un gène de taille p et X_i^j encode alors la base j de ce gène dans l'espèce i, c'est à dire un des 4 nucléotides A, C, T, G. Soit T un arbre binaire à n feuilles dont les n feuilles sont étiquetées par les n espèces de X. L'arbre T est entièrement caractérisé par sa topologie et par la longueur de ses branches. La plupart des méthodes d'inférence se déroule en trois étapes :

- choix d'un critère à minimiser ;
- à topologie fixée, calcul des longueurs de branches qui minimise le critère, et mise en mémoire de la valeur optimale du critère et des longueurs correspondantes ;
- choix, parmi toutes les topologies d'arbre à n noeuds, de celle qui minimise le critère, et des longueurs de branches correspondantes.

Précisons un peu ces trois étapes pour deux grandes familles de méthodes : les méthodes de parcimonie et les méthodes de maximum de vraisemblance. Il en existe aussi tout un pan de la phylogénie, s'appuyant sur les matrices de distances entre espèces, que je n'aborderai pas.

1.1 Parcimonie

À tout seigneur, tout honneur, les méthodes de parcimonie sont les plus simples à expliquer et les premières à avoir été utilisées en phylogénie. L'idée générale des méthodes de parcimonie est due à Cavalli et Safarzo [1], lorsqu'ils ont déclaré qu'on doit privilégier l'arbre d'évolution qui «nécessite le moins de quantité d'évolution» : on cherche l'arbre qui nécessite le moins d'événements (mutation, substitution,...) pour aboutir à nos données. Le critère à minimiser est donc le nombre d'événements évolutifs. On doit en plus être capable, à topologie fixée, de compter ce nombre d'événements et de chercher parmi tous les arbres, celui qui le minimise.

1.1.1 Présentation sur un exemple simple

Illustrons ces différentes étapes par un exemple simple. On considère un ensemble de 5 espèces pour lesquelles on connaît un gène de taille 6, répertoriées dans la table 1.

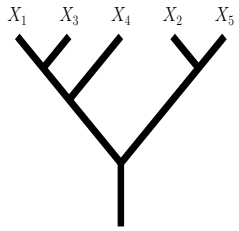

FIG. 1 – Topologie considérée

On autorise tous les changements de nucléotide $A \leftrightarrow C$, $A \leftrightarrow G$, $A \leftrightarrow T$, ... On autorise aussi l'état initial, à la racine de l'arbre, à présenter n'importe séquence.

Espèce	Séquence					
X_1	A	T	T	A	A	T
X_2	T	T	A	T	T	T
X_3	A	A	T	T	T	T
X_4	A	A	T	A	A	A
X_5	T	T	A	A	A	T

TAB. 1 – Les 5 espèces et leur gène de taille 6

On considère enfin une topologie, par exemple celle de Fig. 1 pour laquelle on veut calculer le nombre minimum d'événements nécessaires pour aboutir aux espèces 1 à 5. Ce calcul se fait nucléotide par nucléotide. Pour le nucléotide 1, les espèces présentent la séquence ATAAT. Si l'état initial du nucléotide 1 est C ou G, il faut au moins deux événements pour arriver à cette séquence. Si en revanche, l'état initial est T ou A, il suffit d'un événement, comme le montre Fig. 2. Ainsi pour le nucléotide 1, le nombre minimum d'événements compatible avec nos 5 espèces est 1. En faisant exactement le même calcul pour les nucléotides 2 à 6. On trouve que, pour cette topologie, il faut au moins 1+2+1+2+2+1 = 9 événements pour arriver à nos espèces.

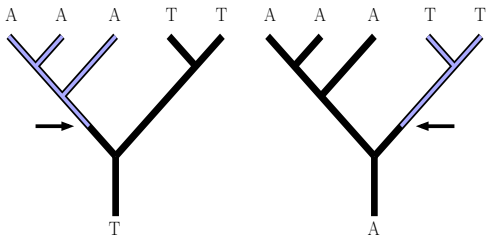

FIG. 2 – Arbre le plus parcimonieux pour le premier nucléotide, avec une racine à l'état T (gauche) ou A (droit). La flèche indique le changement d'état.

On est ainsi capable de calculer le nombre minimum d'événements pour cette topologie. En effectuant ce calcul pour toutes les topologies d'arbres binaires à 5 feuilles, on peut se convaincre que l'arbre du maximum de parcimonie ne correspond pas à la topologie de Fig. 1 mais à celle de Fig. 3, qui ne nécessite que 8 changements.

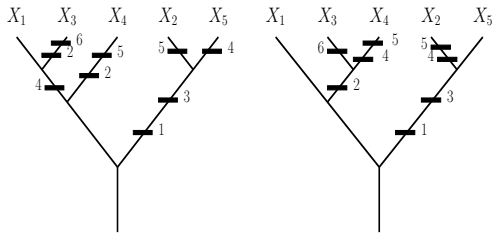

FIG. 3 – Arbre le plus parcimonieux pour la topologie T (gauche) et pour celle du maximum de parcimonie (droite). Les changements d'états sont indiqués par un rectangle, annoté du caractère qu'ils concernent.

1.1.2 Remarques au sujet de l'arbre inféré

Une question importante est celle de la présence/absence d'une racine dans l'arbre. En effet, dans notre exemple, il existe plusieurs arbres du maximum de parcimonie, qui dépendent de l'état et de la position de la racine dans l'arbre, comme illustré dans Fig. 4. Lorsqu'on enlève cette racine, tous ces arbres se confondent. Il y a beaucoup de tels arbres enracinés, un pour chaque branche de l'arbre non raciné, et ils ont tous le même nombre de changements. En fait, on peut même montrer que le nombre de changements ne dépend que l'arbre non raciné, et aucunement de la position ou de l'état de la racine dans l'arbre raciné. Cette propriété est vraie dès lors que l'on autorise tous les changements d'un nucléotide à un autre. Il suffit donc de considérer uniquement les arbres non racinés. Cette remarque est intéressante dans la mesure où il existe moins d'arbres non racinés que d'arbres enracinés, réduisant d'autant le travail de parcours exhaustif des topologies d'arbres.

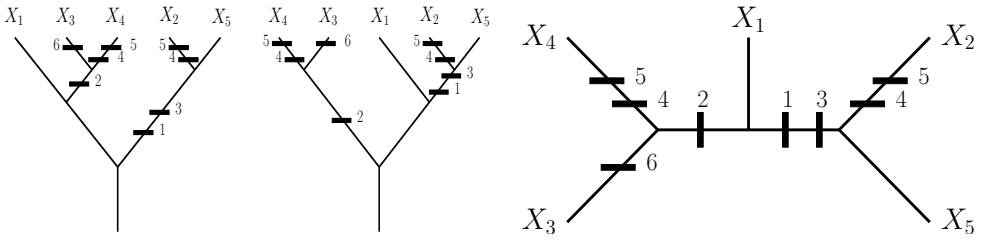

FIG. 4 – Deux arbres qui réalisent le maximum de parcimonie (gauche, centre), ici 8. L'arbre non-raciné correspondant (droite)

Il n'a jusqu'à présent nul part été mention des longueurs des branches de l'arbre. La longueur qu'on souhaite naturellement attribuer à une branche est le nombre d'événements qui y prennent place. Après avoir trouvé la topologie du maximum de parcimonie, on aimerait donc situer les changements de nucléotides sur les branches de celle-ci. Le problème est que pour chaque nucléotide, on peut avoir plusieurs possibilités pour placer ces changements. La méthode standard consiste alors à faire pour chaque branche la moyenne, sur toutes les reconstructions possibles, du nombre de changements qui y prennent place et de lui attribuer cette moyenne comme longueur. On est ainsi sûr que la somme des longueurs des branches est le nombre total d'événements dans l'arbre, 8 dans notre exemple, mais ces longueurs peuvent désormais être non-entières.

Une autre spécificité de la parcimonie est la possibilité de tomber sur des arbres ex-aequo. Cette possibilité est gênante si deux arbres non racinés différents réalisent le maximum de parcimonie. Le choix de l'arbre final relève en effet alors de l'arbitraire.

1.1.3 Remarques au sujet du parcours exhaustif des arbres

Si la méthode à mettre en oeuvre pour reconstruire l'arbre du maximum de parcimonie est simple à comprendre, sa mise en pratique n'en est pas moins ardue. Deux grands algorithmes, l'algorithme de Fitch [5] et celui de Sankoff [10], existent pour compter de manière efficace le nombre de changements le long d'une topologie. Mais le problème essentiel est plutôt celui du parcours des arbres. On doit parcourir toutes les topologies d'arbre à n espèces mais le nombre de telles topologies croît de manière astronomique avec le nombre d'espèces considérées : lorsqu'on rajoute une espèce à un arbre à $n-1$ espèces, on peut ancrer la branche qui y conduit sur chacune des $2n-3$ branches déjà présentes dans l'arbre, ce qui signifie qu'il existe :

$$(2n-3)!! = \frac{(2n-2)!}{2^{n-1}(n-1)!}$$

différentes possibilités d'ajouter une espèce pour former un arbre à n espèces. Chaque possibilité conduit à un arbre différent, qui tous ensembles, constituent l'ensemble des arbres à n espèces. Cette formule donne donc le nombre d'arbres binaires, enracinés et dont le feuilles sont étiquetées. On sait cependant qu'il suffit de considérer les arbres non racinés. Malheureusement, même s'ils sont moins nombreux que les arbres enracinés, ils restent beaucoup trop nombreux. Fondamentalement, passer d'un arbre raciné à un arbre non raciné revient à enlever le noeud étiqueté «racine», qu'on peut assimiler à une espèce. Il y a donc autant d'arbres non racinés à n espèces que de racinés à $n+1$ espèces. L'argument simple de comptage est du à Cavalli-Sforza et Edwards [2]. Plus que le nombre exact d'arbres, c'est l'ordre de grandeur qui est intéressant. On se rend compte que même pour seulement 20 espèces, le parcours exhaustif de tous les arbres n'est pas envisageable (on est face à un problème NP-dur). De nombreuses heuristiques de recherche permettent de s'affranchir du parcours exhaustif, mais elles n'assurent malheureusement que de trouver l'arbre le plus parcimonieux dans un certain voisinage d'arbres, pas l'arbre le plus parcimonieux dans l'absolu.

1.2 Maximum de Vraisemblance

L'ennui majeur des méthodes de parcimonie est leur absence de justification mathématique. Elles sont souvent utilisées comme des heuristiques, non comme des méthodes statistiques de reconstruction d'arbre. Et lorsqu'on se penche sur leurs propriétés statistiques, elles peuvent ne pas constituer un estimateur consistant de la vraie phylogénie ! Les méthodes de maximum de vraisemblance présentent une solution à ce problème. On considère que les versions du gènes qu'on observe chez les n espèces sont une réalisation d'une variable aléatoire dont la loi dépend, entre autres paramètres, d'une phylogénie. On se retrouve alors dans un cadre classique où la phylogénie n'est rien d'autre qu'un paramètre du modèle qu'on peut estimer par maximum de vraisemblance. L'estimateur du maximum de vraisemblance de la phylogénie ainsi obtenu hérite alors des propriétés classiques des estimateurs du maximum de vraisemblance, en particulier il est consistant.

Il nous faut donc choisir un cadre paramétrique convenable, et expliquer comment calculer la phylogénie du maximum de vraisemblance dans ce cadre-là.

1.2.1 Choix d'un cadre paramétrique et calcul de vraisemblance

On commence par se donner un modèle d'évolution des séquences d'ADN [9]. Le plus simple, et celui que je me contenterai de présenter, est le modèle de Jukes-Cantor [7], qui stipule que pendant un instant dt, chaque nucléotide change d'état avec une probabilité dt et passe dans un des 4 états possibles (y compris celui qu'il occupe déjà) avec probabilité $dt/4$. Dans ce modèle, tous les changements d'état d'un nucléotide vers un autre sont autorisés. D'autres modèles, plus raffinés, existent mais constituent essentiellement des variations de Jukes-Cantor.

Une fois le modèle d'évolution donné, on se donne un arbre T, *i.e.* une topologie et les longueurs t_j des branches correspondantes. On fait ensuite deux hypothèses un peu restrictives : on suppose que

- les nucléotides évoluent indépendamment les uns des autres
- l'évolution est indépendante dans chaque branche de l'arbre, sachant l'état des nucléotides au début de cette branche.

Ces deux hypothèses rendent le calcul et la maximisation de la vraisemblance beaucoup plus faciles. Elles constituent la plupart du temps des hypothèses de travail raisonnables même si on peut toujours des situations où elles sont mises en défaut.

La probabilité des observations sous cet arbre, ou encore la vraisemblance de l'arbre étant données les observations, est alors donnée par

$$L = \prod_{j=1}^{p} \text{Prob}(X_1^j, \ldots, X_n^j | T) := \prod_{j=1}^{p} \text{Prob}(X^{(j)} | T)$$

où p est le nombre de caractères (ici de nucléotides) sur lequel on se base et $X^{(j)}$ est la donnée du nucléotide j. Par exemple, dans l'arbre de Fig. 5, on a pour le caractère

représenté :

$$\text{Prob}(X^{(j)}|T) = \sum_x \sum_y \sum_z \sum_w \text{Prob}(A, C, C, C, G, x, y, z, w|T) \tag{1}$$

où chaque somme se fait sur les 4 nucléotides.

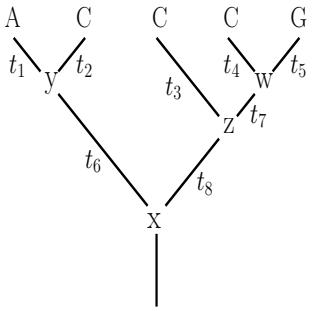

FIG. 5 – Une topologie, avec la longueur des branches et l'état d'un nucléotide aux feuilles, utilisée pour illustrer le calcul de la vraisemblance.

Les deux hypothèses d'indépendance permettent de factoriser chaque probabilité du second membre en un produit de termes :

$$\begin{aligned}
&\text{Prob}(A, C, C, C, G, x, y, z, w|T) = \\
&\quad \text{Prob}(x) \times \text{Prob}(y|x, t_6) \times \text{Prob}(A|y, t_1) \times \text{Prob}(C|y, t_2) \\
&\qquad \times \text{Prob}(z|x, t_8) \times \text{Prob}(C|z, t_3) \\
&\qquad \times \text{Prob}(w|z, t_7) \times \text{Prob}(C|w, t_4) \times \text{Prob}(G|w, t_5)
\end{aligned} \tag{2}$$

En combinant les équations 1 et 2, et en déplaçant les sommes le plus à droite possible, on obtient :

$$\begin{aligned}
\text{Prob}(X^{(j)}|T) = &\\
&\sum_x \text{Prob}(x) \Big(\sum_y \text{Prob}(y|x, t_6)\text{Prob}(A|y, t_1)\text{Prob}(C|y, t_2) \Big) \\
&\quad \times \Big(\sum_z \text{Prob}(z|x, t_8)\text{Prob}(C|z, t_3) \\
&\qquad (\sum_w \text{Prob}(w|z, t_7)\text{Prob}(C|w, t_4)\text{Prob}(G|w, t_5)) \Big)
\end{aligned} \tag{3}$$

On remarque que le schéma de parenthèse dans cette expression est $(A, C)(C, (C, G))$, c'est-à-dire exactement la même structure que dans l'arbre. En partant de cette constatation, Felsenstein [3] a développé un algorithme récursif pour calculer la vraisemblance de l'arbre en $O(np)$ opérations, au lieu de $O(4^{np})$ (toutes les configurations possibles pour les noeuds non observés).

1.2.2 Racine de l'arbre et parcours exhaustif de l'arbre

Une particularité que le modèle de Jukes Cantor partage avec ses camarades est la *réversibilité*. Un nucléotide a autant de chances de passer de l'état x à l'état y en un temps t que l'inverse. Cette réversibilité assure :

$$\text{Prob}(x)\text{Prob}(y|x, t_6) = \text{Prob}(y)\text{Prob}(x|y, t_6)$$

qui, en substituant dans 3, montre que la racine peut être déplacé de x à y. En fait, on peut même montrer que la racine peut être déplacé n'importe où entre x et y. Comme la racine peut être n'importe où sur la branche, et une fois arrivée en bout de branche se déplacer sur la branche suivante, elle peut être n'importe où. De sorte que notre arbre est en fait un arbre *non raciné*, qui ne donne aucune information sur la position de la racine. C'est une bonne nouvelle dans la mesure où on peut donc se contenter de considérer les arbres non racinés.

Malheureusement, et comme pour les méthodes de parcimonie, les méthodes de maximum de vraisemblance n'échappent pas au parcours de tous les arbres. Bien qu'on se puisse se contenter des arbres non racinés, ils en reste beaucoup trop à examiner (le problème est encore une fois NP-complet), ce qui nous force une fois de plus soit à adopter des heuristiques (au risque de tomber dans un maximum local de la vraisemblance), soit à nous restreindre à de petits ensembles d'espèces.

1.3 Autour de la parcimonie et du maximum de vraisemblance

1.3.1 Lien entre les deux méthodes

Le maximum de parcimonie est séduisant par sa simplicité mais peut échouer à fournir un estimateur consistant de l'arbre [4]. Le maximum de vraisemblance fournit à coup sûr un estimateur consistant mais sa mise en oeuvre est plus délicate. Ces deux méthodes semblent donc étrangères, il existe cependant des situations où elles coïncident. Faire du maximum de parcimonie revient implicitement à supposer que les événements évolutifs (changements de nucléotides) sont rares. En effet, dans ce cas, chaque événement a une probabilité très faible et, intuitivement l'arbre «le plus probable», *i.e.* du maximum de vraisemblance est celui qui nécessite le moins d'événements évolutifs, *i.e.* du maximum de parcimonie.

1.3.2 Amélioration des deux méthodes

On peut améliorer les deux méthodes de façon analogue en raffinant un peu le modèle sous-jacent d'évolution de la séquence d'ADN. On sait en effet que dans un gène, tous les nucléotides n'ont pas la même importance. En particulier, la redondance du code génétique fait que la troisième base de chaque codon est souvent peu contrainte et donc plus sujette à changements les deux autres. On peut introduire cette connaissance dans nos méthodes ; pour la parcimonie en diminuant le poids attribués aux changements qui affectent une troisième base et, pour la vraisemblance, en augmentant la probabilité

de tels changements. De même on sait que les événements de type $\{A, T\} \leftrightarrow \{C, G\}$, dits *transversions* sont plus rares que les événements $A \leftrightarrow T$ ou $C \leftrightarrow G$, dits *transitions*. On peut de façon analogue en tenir compte en attribuant des poids/probabilités différents à chaque type d'événements.

2 La place des mathématiques en phylogénie

Nous avons vu dans la Section 1 comment reconstruire une phylogénie à partir de séquences d'ADN. C'est la partie centrale, la plus connue de la phylogénie, celle qui vient à l'esprit en premier. Il existe cependant au moins deux autre étapes, toutes aussi importantes : une d'alignement des séquences en amont et une en aval de quantification de l'incertitude de la phylogénie inférée. Cette quantification est spécifique à un cadre de travail statistique. C'est à celle-ci que je souhaite m'intéresser durant ma thèse.

2.1 Avant et pendant l'estimation

2.1.1 Alignement de séquence

L'étape préliminaire à toute inférence de phylogénie en est aussi une des plus complexes. Sans s'attarder sur les détails techniques, l'alignement est une étape cruciale. Les séquences de gènes présentes dans différentes espèces présentent en effet un défaut majeur : des insertions et/ou des délétions ont pu décaler les nucléotides les uns par rapport aux autres de sorte que la j-ème base chez une espèce ne corresponde pas à la j-ème base d'une autre espèce mais plutôt à la $j+5$-ème. Les transversions et transcriptions compliquent encore la tâche puisqu'une base A chez une espèce ne correspond plus forcément à la même base chez les autres espèces. Le but de l'alignement est donc d'apparier toutes les séquences, de sorte que la j-ème base correspondent chez *toutes* les séquences alignées à la j-ème base de l'espèce originale (la Table 1 correspond à un exemple de séquences alignéees). C'est à cette condition et à cette condition seule que les modèles de phylogénie ont du sens.

2.1.2 Définition du modèle et calcul des paramètre du modèle

La Section 1 présente les deux modèles les plus utilisés en phylogénie. De nombreuses études ont été réalisées pour décrire le comportement de l'une et l'autre des deux vis-à-vis de la topologie de la vraie phylogénie. On sait par exemple que si la vraie topologie ne contient que des branches courtes et que le taux d'évolution n'est pas très élevé, il est préférable d'utiliser l'arbre du maximum de parcimonie, équivalent dans ce cas à celui du maximum de vraisemblance mais plus facile à trouver. En revanche, si la vraie phylogénie contient deux longues branches qui ne sont pas voisines ces deux longues branches deviennent voisines dans la phylogénie du maximum de parsimonie, comme l'illustre Fig. 6. Ainsi, en fonction des données dont on dispose et de la nature de la vraie phylogénie, il vaut mieux utiliser l'une ou l'autre des deux méthodes.

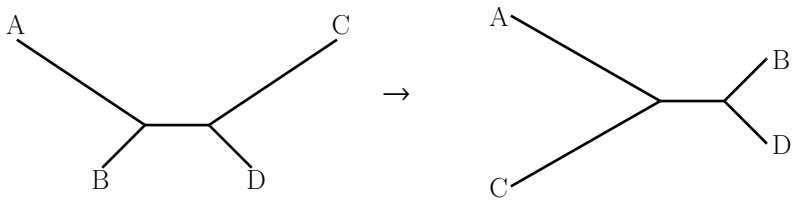

FIG. 6 – Illustration du phénomène d'attraction des longues branches, l'arbre réel (gauche) donne après inférence, l'arbre inféré (droite)

2.2 Après l'estimation

2.2.1 Robustesse de l'arbre inféré

Une étape postérieure à l'estimation consiste à mesurer la robustesse de la phylogénie obtenue ; bien que digne d'intérêt, cette estimation est rarement faite : c'est celle qui m'intéresse le plus. Le cadre statistique dans lequel on travaille nous assure que la phylogénie estimée est la réalisation d'une variable aléatoire, fonction des données de départ. Elle est sujette à ce titre à une variabilité d'estimation. On se propose de quantifier cette incertitude d'estimation et d'en tenir compte dans l'interprétation des résultats. Ce problème, non spécifique à la phylogénie, a donné naissance à de nombreux outils : test de deux topologies pour déterminer si l'une a significativement plus de chance que l'autre d'être la vraie phylogénie [8][6] , méthodes de rééchantillonnage (bootstrap) pour déterminer la pertinence des noeuds internes, méthodes de jackknife pour tester des phylogénies dans un cadre moins paramétrique que celui du maximum de vraisemblance,. . .Je présente deux questions plus ciblées qui m'intéressent tout particulièrement.

2.2.2 Homoplasie

Lorsqu'on compare deux séquences d'ADN, ils existent deux sources majeures de ressemblance : la ressemblance héritée d'un ancêtre commun et l'*homoplasie*, la ressemblance obtenue par convergence. En phylogénie, seule la première nous intéresse ; c'est en effet la seule qui nous apporte de l'information sur la "vraie" phylogénie, celle selon laquelle les espèces ont évolués. À ce titre, non seulement l'homoplasie ne nous intéresse pas mais elle est en plus indésirable : elle a pour effet de rapprocher dans l'arbre des espèces qui devraient normalement y être éloignées, éloignant ainsi considérablement la topologie inférée de la vraie. On cherche donc à se débarrasser de cette information "parasite". Un moyen d'y parvenir est d'estimer, à l'aide du jackknife, l'influence de chaque nucléotide sur la phylogénie inférée et de retirer ceux qui ont une influence trop forte. Cette technique, complètement balisée, a fait ses preuves dans de nombreux domaines. Formellement : on estime l'arbre du maximum de vraisemblance (MV) T sur les p caractères dont on dispose. Pour chaque caractère j, on estime l'arbre MV T_j sur les $p-1$ *autres* caractères, puis on considère l'arbre T_j qui s'éloigne le plus de T. Si cette

différence est significative, on élimine le caractère j puis on réitère le processus jusqu'à ce qu'aucun noeud n'ait une influence disproportionnée sur la phylogénie inférée. Bien sûr, un problème crucial dans cet exemple est la calibration de la méthode : quelle distance adopter pour quantifier l'«éloignement» entre T_j et T ? Comment la calculer de manière non prohibitive ? Comment définir le seuil à partir duquel elle est significative ? Il est évident que moins on dispose de caractères pour l'inférence, plus l'arbre inféré est sensible à chacun d'entre eux. Comment faire alors varier le seuil avec le nombre de caractères ? Et de manière plus générale, comment peut-on détecter et prendre en compte intelligemment l'homoplasie de certains caractères ?

2.2.3 Concaténation de gènes

Comme dans la majorité des méthodes statistiques, on préfère inférer un arbre sur de longues séquences. Une idée pour augmenter la taille des séquences est d'utiliser l'information de plusieurs gènes, et non d'un seul. Le but ultime est évidemment d'utiliser *tout* le génome au lieu de se cantonner à quelques gènes. On se contente actuellement de les concaténer et de faire de l'inférence sur le gène virtuel résultant. On sait cependant que deux gènes peuvent avoir des histoires évolutives très différentes, donnant lieu à des phylogénies tout aussi différentes. La phylogénie inférée sur le gène virtuel peut être sensiblement différente de chacune des deux et donc peu pertinente. Ce constat soulève deux familles de questions, une centrée sur l'agrégation de gènes, une centrée sur leur segmentation. Dans quelles conditions est-il raisonnable d'agréger deux gènes ? On se doute que la réponse réside dans la «distance» entre les deux «vraies» phylogénies : plus elles sont proches, plus l'arbre inféré va leur ressembler et plus la variabilité d'estimation diminue (conséquence presque automatique de l'allongement «virtuel» de la séquence d'estimation). Mais même alors, quel «poids» attribuer dans l'inférence à chacun de ces gènes : proportionnel à sa longueur [11], à un autre critère ? De plus, à l'exemple du SRAS, dont une partie du génome est issu du porc, il n'est pas exclu qu'une séquence inclue du matériel génétique provenant de divers hôtes. On voudrait dans ce cadre introduire un modèle de mélange sur les nucléotides pour les catégoriser : ils sont répartis en Q catégories, chacune permettant d'inférer une phylogénie partielle. Cette idée soulève elle aussi son lot de questions : comment déterminer la catégorie de chaque nucléotide ? Comment choisir le nombre de catégories ? Et surtout que faire des phylogénies partielles ? Une idée tentante est bien sûr de les agréger dans une phylogénie complète. Mais comment fusionner ces phylogénies différentes et surtout est ce raisonnable ?

Je souhaite conclure en soulignant que la phylogénie est un domaine très vaste, qui suscite un enthousiasme énorme tant chez les statisticiens et les informaticiens que chez les biologistes, comme en témoigne la vigueur de la communauté des phylogénistes. De nombreuses questions intéressantes sont ouvertes, notamment sur la pertinence des phylogénies inférées. C'est sur celles-ci que je souhaite me pencher durant ma thèse.

Références

[1] L. L. Cavalli-Sforza and A. W. F. Edwards. Analysis of human evolution. *Genetics Today*, pages 923–933, 1965.

[2] L. L. Cavalli-Sforza and A. W. F. Edwards. Phylogenetics analysis : Models and estimation procedures. *American Journal of man Geneics*, 19 :233–257, 1967.

[3] J. Felsenstein. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. *Systematic Zoology*, 22 :240–249, 1973.

[4] Joseph Felsenstein. *Inferring Phylogenies*. Sinauer Associates, September 2003.

[5] W. M. Fitch. Toward defining the course of evolution : Minimum change for a specified tree topology. *Systematic Zoology*, 20 :406–416, 1971.

[6] Nick Goldman, Jon P. Anderson, and Allen G. Rodrigo. Likelihood-based tests of topologies in phylogenetics. *Systematic Biology*, 49(4) :652–670, 2000.

[7] T.H. Jukes and C.R. Cantor. *Evolution of protein molecules*, volume 3 of *mammalian Protein Metabolism*, chapter 24, pages 21–132. Academic Press, New York, 1969.

[8] H. Kishino and M. Hasegawa. Evaluation of the maximum likelihood estimate of the evolutionary tree topologies from dna sequence data, and the branching order in hominoidea. *J. Mol. Evol.*, 29 :170–179, 1989.

[9] J. Neyman. *Molecular studies of evolution : A source of novel statistical problems*. Statistical Decision Theory and related Topics. Academic Press, New York, 1971.

[10] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal of Applied Mathematics*, 28 :35–42, 1975.

[11] T. Seo, H. Kishino, and J.L. Thorne. Incorporating gene-specific variation when inferring and evaluating optimal evolutionary tree topologies from multilocus sequence data. *PNAS*, 102(12) :4436–4441, 2005.