Inférence phylogénétique : méthodes basées sur les distances

Fabio Pardi

6

Inférence phylogénétique : méthodes basées sur les distances

Fabio PARDI
LIRMM, CNRS, Université de Montpellier, Montpellier, France

Une approche populaire pour l'inférence phylogénétique consiste à estimer une matrice de distances évolutives entre paires de taxons, et ensuite à utiliser ces informations pour inférer leur arbre phylogénétique. Dans ce chapitre, nous expliquerons d'abord comment les distances doivent être définies et estimées, et ensuite nous nous concentrerons sur la tâche consistant à construire un arbre phylogénétique qui correspond bien aux distances estimées. Plus précisément, nous soulignerons les propriétés que les distances doivent satisfaire en principe, et les principales approches classiques d'inférence d'arbres, les *moindres carrés* et le *minimum d'évolution*. Enfin, nous nous concentrerons sur les méthodes plus connues, en particulier *neighbor joining* et un large éventail d'algorithmes qui s'en inspirent.

6.1. Introduction

L'inférence phylogénétique basée sur les distances repose sur une observation simple mais très importante : supposons que nous puissions connaître le nombre précis de changements qui se sont produits sur le chemin évolutif entre deux taxons quelconques i et j dans une collection de n taxons ($1 \leq i \leq j \leq n$) (voir figure 6.1) Ces « changements » peuvent être définis de plusieurs manières, mais le raisonnement vaut pour tout ce qui peut être compté, par exemple le nombre de substitutions survenues

dans un gène ou le nombre de fois qu'il y a eu un changement dans l'ordre des gènes entre le génome de i et celui de j. Il s'avère que, tant que ces événements sont assez fréquents, cette information est tout ce dont nous avons besoin pour reconstituer la topologie de l'arbre non raciné décrivant l'évolution de nos n taxons.

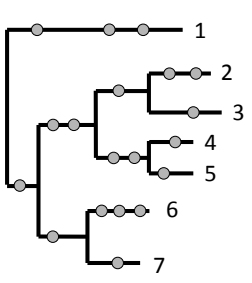
Cette idée (formalisée dans un résultat classique et bien connu que nous énoncerons plus tard dans le théorème 6.2) constitue le fondement théorique des approches basées sur les distances pour l'inférence phylogénétique. En pratique, malheureusement, le nombre de changements qui se sont produits entre les taxons, leur *distance évolutive*, ne peut pas être observé, mais peut seulement être estimé. C'est sur la base des distances *estimées* que la reconstitution de l'arbre est alors effectuée, le but étant de trouver l'arbre qui correspond le mieux aux distances estimées, en prenant parfois en compte l'incertitude dans ces estimations.

L'utilisation de distances ou de similitudes entre taxons est l'une des approches les plus anciennes en biologie systématique (Sneath 1957 ; Sokal et Michener 1958). Les premiers travaux ont vu une distinction très claire entre, d'une part les techniques de *clustering* hiérarchique pour la classification taxonomique (Sokal et Sneath 1963), et d'autre part les méthodes d'optimisation visant directement l'inférence phylogénétique (Cavalli-Sforza et Edwards 1967 ; Fitch et Margoliash 1967). Ces deux lignes de travail ont finalement convergé vers la fin des années 1980, menant à la méthode très populaire de *neighbor joining* (NJ) (Saitou et Nei 1987) qui combinait des idées algorithmiques de classification (par exemple UPGMA, (Sokal et Michener 1958) ; ADDTREE, (Sattath et Tversky 1977)) aux principes d'optimisation de la phylogénétique (par exemple ME (Kidd et Sgaramella-Zonta 1971) ; OLS (Cavalli-Sforza et Edwards 1967) voir *infra*). Depuis NJ, la phylogénétique a assisté à une renaissance des méthodes basées sur les distances, souvent liées ou inspirées par NJ. Nous en verrons plusieurs dans la section 6.4.3.

Ces méthodes sont largement utilisées pour leur efficacité de calcul, un avantage qui les rend particulièrement adaptées à l'inférence de très grandes phylogénies ou de grands nombres de phylogénies (par exemple, pour le bootstrap). Elles sont aussi utilisées pour fournir un arbre guide pour l'alignement progressif (Larkin *et al*. 2007) ou un arbre de départ pour des approches d'inférence plus sophistiquées telles que celles basées sur le maximum de vraisemblance (Guindon et Gascuel 2003). Plus généralement, l'utilisation de distances estimées entre les séquences biologiques est une réponse évidente pour faire face aux énormes jeux de données générés par les techniques de séquençage modernes, toujours plus rapides et moins chères. En témoigne le succès continu de NJ qui reste à ce jour l'algorithme le plus cité en phylogénétique (environ 3 000 citations par an depuis 2010 pour l'article original de Saitou et Nei 1987).

Dans ce chapitre, nous exposerons les grandes lignes des idées qui sous-tendent la méthodologie basée sur les distances. Nous commencerons par expliquer l'importance

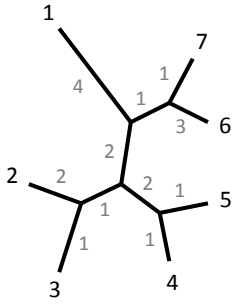
d'estimer des distances qui reflètent le nombre de changements qui se sont réellement produits entre deux taxons i et j plutôt que le nombre le plus parcimonieux de changements qui sont nécessaires pour expliquer les différences entre i et j. Après avoir décrit de manière concise la tâche d'estimation des distances évolutives, l'accent sera mis ici sur la deuxième étape cruciale : l'inférence d'un arbre qui correspond bien aux distances estimées.



a)

	1	2	3	4	5	6	7
1	0	9	8	9	9	8	6
2		0	3	6	6	9	7
3			0	5	5	8	6
4				0	2	9	7
5					0	9	7
6						0	4
7							0

b)



c)

Figure 6.1. *L'idée fondamentale de la reconstitution d'un arbre à partir de distances*

COMMENTAIRE SUR LA FIGURE 6.1.– *Un certain nombre de changements (cercles gris) se produisent le long des branches d'une phylogénie inconnue (a). Supposons que nous soyons capables de connaître le nombre de changements intervenus entre chaque paire de taxons et que nous ayons enregistré ces informations dans une matrice (b). Ensuite, tant qu'au moins un changement s'est produit sur chaque branche interne, nous sommes capable de reconstituer avec précision la topologie de l'arbre non raciné sous-jacent à la vraie phylogénie, y compris le nombre de changements qui se sont produits sur chaque branche (c). Ceci est une conséquence du théorème 6.2.*

6.2. Fondements mathématiques

Pour énoncer les résultats classiques sur l'utilisation des distances pour l'inférence phylogénétique, nous devons définir quelques notions fondamentales. Tout au long de ce chapitre, nous identifierons les taxons d'intérêt avec des indices $1, 2, \ldots, n$. Ces « taxons » peuvent être tout ce qui est le produit final d'un processus évolutif en forme d'arbre, des séquences génétiques aux recettes de cuisine transmises de génération en génération.

Une *dissimilarité* sur les taxons $1, 2, \ldots, n$ est une collection (une matrice, un vecteur, une fonction, peu importe) $\mathbf{d} = (d_{ij})$ telle que $d_{ij} \geq 0$, $d_{ij} = d_{ji}$ et $d_{ii} = 0$ pour tout $i, j \in \{1, 2, \ldots n\}$. Une *distance métrique* est une dissimilarité pour laquelle on a

aussi $d_{ij} \leq d_{ik} + d_{kj}$[1]. Nous remarquons qu'en phylogénétique le mot « distance » est souvent utilisé pour signifier une dissimilarité, comme par exemple quand on parle de « matrice de distance ». La confusion se résout facilement si l'on se rend compte que ce qu'on entend habituellement par « distance » est une distance *estimée* (c'est-à-dire quelque chose qui se rapproche d'une distance mais qui n'en est pas une). Ici nous ne parlerons donc pas de « distance » tout court pour éviter toute confusion.

Un *arbre phylogénétique* T sur $1, 2, \ldots, n$ est composé (1) d'une topologie d'arbre τ, c'est-à-dire un arbre non raciné sans nœuds de degré 2, dont les feuilles sont étiquetées de manière unique avec $1, 2, \ldots, n$; et (2) des longueurs $\lambda_e \geq 0$ affectées à chaque branche e de τ. Les longueurs de branche sont utilisées pour modéliser le changement évolutif le long de chaque branche. (Dans la section 6.3.1, nous verrons que pour les séquences moléculaires, λ_e est généralement défini comme l'espérance mathématique du nombre de substitutions par site survenant sur e.)

En raison de la longueur des branches, un arbre phylogénétique T détermine naturellement une distance métrique $\mathbf{d}^T = (d_{ij}^T)$ sur ses taxons $1, 2, \ldots, n$, où d_{ij}^T est simplement définie comme la longueur du chemin reliant i et j dans T. Autrement dit, si τ_{ij} représente l'ensemble des branches entre i et j dans τ :

$$d_{ij}^T = \sum_{e \in \tau_{ij}} \lambda_e \qquad [6.1]$$

Quand $\mathbf{d} = \mathbf{d}^T$ pour un arbre phylogénétique T, on dit que $\mathbf{d}$ est une *distance d'arbre* (sur T). Les distances d'arbres sont parfois appelées distances *patristiques* (parce qu'elles suivent le chemin ancestral entre i et j) ou encore distances *additives* (ce qui évoque la somme en [6.1]).

Toutes les dissimilarités ne sont pas des distances d'arbre. Une observation intéressante est que si nous prenons quatre taxons quelconques i, j, h, k et considérons les sommes :

$$d_{ij} + d_{hk}, \quad d_{ih} + d_{jk}, \quad d_{ik} + d_{jh} \qquad [6.2]$$

alors, quand $\mathbf{d}$ est une distance d'arbre, deux de ces sommes doivent être égales et la troisième somme doit être inférieure ou égale aux deux autres sommes. Cela peut être compris en considérant un exemple concret : si nous prenons les taxons $1, 2, 3$ et 6 dans l'arbre de la figure 6.1c, alors nous avons $d_{12} + d_{36} = d_{13} + d_{26} = 17$, ce qui est supérieur à $d_{16} + d_{23} = 11$. Les deux premières sommes sont égales car ce sont des sommes sur le même ensemble de branches, alors que $d_{16} + d_{23}$ n'est que

1. Les définitions standard de distance exigent également que $d_{ij} = 0$ implique $i = j$, mais nous ne ferons pas cette hypothèse ici.

sur un sous-ensemble de ces branches. Les relations entre les sommes dans l'équation [6.2] peuvent être exprimées succintement en ce qui est connu comme la *condition des quatre points* : pour tout i, j, h, k :

$$d_{ij} + d_{hk} \leq \max\{d_{ih} + d_{jk}, d_{ik} + d_{jh}\}$$

Étonnamment, la condition des quatre points est non seulement nécessaire, mais aussi suffisante pour que $\mathbf{d}$ soit une distance d'arbre :

THÉORÈME 6.1.– $\mathbf{d}$ *est une distance d'arbre si et seulement si* $\mathbf{d}$ *satisfait la condition des quatre points.*

L'étude des distances d'arbres est un sujet ancien et est l'un des fondements de la phylogénétique mathématique (Zaretskii 1965 ; Pereira 1969 ; Buneman 1971). L'un des résultats les plus importants pour l'inférence phylogénétique basée sur les distances est le suivant.

THÉORÈME 6.2.– *Soit* T *un arbre phylogénétique ayant des longueurs de branches strictement positives. Alors* T *est l'unique arbre phylogénétique ayant des longueurs de branches strictement positives ayant* $\mathbf{d}^T$ *comme distance d'arbre.*

Autrement dit, il ne peut pas exister un autre arbre phylogénétique T' ayant une topologie différente (non isomorphe) de celle de T ou ayant des longueurs de branches strictement positives et différentes de celles de T, tel que $\mathbf{d}^T = \mathbf{d}^{T'}$.

Pour prouver ce théorème, l'idée principale est que la condition des quatre points pour un quartet $\{i, j, h, k\}$ induit une topologie d'arbre sur ce quartet et impose également certaines contraintes sur un sous-ensemble de longueurs de branches. Par exemple, dans la figure 6.1, le fait que $d_{16} + d_{23} < d_{12} + d_{36} = d_{13} + d_{26}$ implique que le chemin entre 1 et 6 ne partage aucune branche avec le chemin entre 2 et 3, et ceci détermine la topologie du sous-arbre reliant les taxons 1, 2, 3 et 6. Par contre le chemin entre 1 et 2 et le chemin entre 3 et 6 ont des branches en commun, et la longueur totale de ces branches partagées est déterminée par l'expression $1/2 \cdot (d_{12} + d_{36} - d_{16} - d_{23})$ (égale à 3 dans la figure 6.1). Collectivement, ces informations pour tous les quartets $\{i, j, h, k\}$ sont largement suffisantes pour déterminer T, y compris ses longueurs de branches. Nous renvoyons le lecteur à (Barthélémy et Guénoche 1991) ou (Semple et Steel 2003) pour un traitement complet de ce résultat.

Une première question naturelle suivant le théorème 6.2 est : est-ce que tous les éléments de $\mathbf{d}^T$ sont nécessaires pour déterminer T, ou le résultat d'unicité tient-il également lorsque seules certaines des distances sont connues ? Il s'avère qu'une

grande partie du contenu de $\mathbf{d}^T$ est effectivement redondante. Par exemple, pour un arbre phylogénétique entièrement résolu (c'est-à-dire binaire) T avec des longueurs de branche strictement positives, la connaissance d'un sous-ensemble soigneusement choisi de $2n - 3$ distances de $\mathbf{d}^T$ est suffisante pour déterminer de manière unique T (encore une fois topologie *et* longueurs de branche) (Dress *et al*. 2012). L'intuition derrière cette idée (pour les longueurs de branches) apparaîtra plus tard dans la figure 6.3. Voir aussi (Huber et Kettleborough 2015 ; Kettleborough *et al*. 2015) pour des travaux récents sur ce sujet.

Bien entendu, la question qui nous intéresse le plus dans ce chapitre est : comment retrouver l'unique arbre T à partir de $\mathbf{d}^T$? En fait, il s'agit d'un problème facile et de nombreux algorithmes peuvent le résoudre de manière efficiente avec des complexités en temps d'exécution aussi basses que $O(n \log n)$ (Hein 1989). Bien que cela puisse sembler une très bonne nouvelle pour l'inférence phylogénétique, nous devons garder à l'esprit qu'en pratique $\mathbf{d}^T$ n'est pas directement observable à partir de données biologiques. Le mieux que nous puissions faire est de produire une estimation $\boldsymbol{\delta}$ de $\mathbf{d}^T$. En général, $\boldsymbol{\delta}$ ne sera pas une distance d'arbre (parfois même pas une distance métrique) mais juste une dissimiliarité qui, dans le meilleur des cas, se rapproche de $\mathbf{d}^T$. Concevoir des méthodes pour inférer T à partir d'une *approximation* d'une distance d'arbre est loin d'être trivial.

Une méthode d'inférence phylogénétique basée sur les distances peut alors être décrite comme une fonction qui prend une dissimilarité en entrée et renvoie un arbre phylogénétique. Une bonne méthode basée sur les distances est celle qui est capable de construire un arbre T' qui est proche de T lorsque la dissimilarité d'entrée $\boldsymbol{\delta}$ est proche de $\mathbf{d}^T$. Par exemple, un prérequis minimal pour toute méthode basée sur les distances est la convergence (ou consistance) topologique.

DÉFINITION 6.1.– *Une méthode d'inférence d'arbre basée sur les distances $\mathcal{M}$ est* topologiquement convergente *si, pour tout arbre binaire T avec des longueurs de branches strictement positives, pour $\boldsymbol{\delta}$ suffisamment proche de $\mathbf{d}^T$, l'utilisation de $\mathcal{M}$ sur $\boldsymbol{\delta}$ donne un arbre T' avec la même topologie que T.*

Des études avancées sur les propriétés théoriques de ces méthodes cherchent à caractériser la robustesse de l'inférence de la topologie d'arbre aux écarts de $\boldsymbol{\delta}$ par rapport à $\mathbf{d}^T$. Suite à un article pionnier de (Atteson 1999), des résultats intéressants continuent d'être obtenus sur ce sujet (Mihaescu *et al*. 2009 ; Gascuel et Steel 2016). Une description précise de cette ligne de travail dépasse le cadre de ce chapitre.

Nous reviendrons sur l'inférence d'arbre proprement dite dans la section 6.4. Avant cela, donnons un bref aperçu des approches les plus courantes pour produire les estimations de distance $\boldsymbol{\delta} = (\delta_{ij})$.

6.3. Estimation des distances

Le premier composant d'une approche basée sur les distances pour l'inférence d'arbre est l'estimation des distances. Le message le plus important à retenir des sections précédentes est que toutes les dissimilarités ne peuvent pas être adoptées en phylogénétique. Seules les dissimilarités qui peuvent être considérées comme des estimations d'une distance d'arbre ont une justification rigoureuse.

Idéalement, à mesure que de plus en plus de données sont collectées, les distances estimées deviendront de plus en plus proches d'une distance d'arbre $\mathbf{d}^T$ pour un arbre T qui représente correctement l'histoire évolutive des taxons considérés. Ceci, combiné à l'utilisation de toute méthode topologiquement convergente (voir définition 6.1), garantit en principe la convergence de l'arbre inféré vers la topologie d'arbre correcte.

6.3.1. *Distances estimées à partir de séquences alignées*

La méthode la mieux étudiée pour estimer des distances évolutives est à partir de séquences (nucléotidiques ou protéiques) alignées. Ce dont nous avons besoin est une collection de n séquences homologues[2] (une pour chaque taxon $1, 2, \ldots, n$) et pour chaque paire de taxons i et j un alignement des séquences de i et j. (Nous n'avons pas besoin d'un alignement multiple contenant toutes les séquences.)

Comment calculer ensuite les estimations de distance $\boldsymbol{\delta}$? D'après tout ce que nous avons dit précédemment, il devrait être intuitif que compter le nombre de différences entre chaque paire de séquences (ou même prendre une version pondérée d'un tel décompte) n'est pas une bonne idée. Bien que cette approche naïve produise effectivement une distance métrique (le nombre de différences est également connu sous le nom de *distance de Hamming*) elle aboutit souvent à une distance qui s'écarte considérablement d'une distance d'arbre.

Autrement dit, puisque ce que nous devrions essayer d'estimer est le nombre de substitutions qui se sont effectivement produites entre i et j (voir à nouveau la figure 6.1), le nombre de différences m_{ij} entre i et j est en fait une très mauvaise estimation : comme plusieurs substitutions peuvent se produire au même site, m_{ij} sous-estime souvent le nombre réel de substitutions (m_{ij} évoque le mot mésappariements ou *mismatches* en anglais, voir la figure 6.3.1 pour un exemple).

Comment estimer le nombre de substitutions qui se sont produites sur le chemin reliant i et j ? Nous décrivons maintenant, en termes très généraux, l'approche du

2. Si l'on souhaite inférer une phylogénie d'espèces, les séquences doivent également être orthologues (c'est-à-dire qu'aucune duplication n'est nécessaire pour expliquer l'évolution des séquences). Sinon, l'arbre de ces séquences doit être interprété comme une phylogénie de gènes.

maximum de vraisemblance (ML, *maximum likelihood*) pour résoudre ce problème d'estimation, où le nombre de substitutions est généralement normalisé par la longueur des séquences. Le lecteur intéressé est renvoyé à d'autres manuels (Felsenstein 2004 ; Yang 2006) pour un traitement détaillé de l'estimation des distances à partir de séquences alignées.

i GA**AT**A**C**TC**AA**A
  ||.|.|.|.||
k GACTGCCCGAA
  ||.|||.||.|
j GA**TT**G**C**TC**GG**A

Figure 6.2. *Le nombre de différences sous-estime le nombre de substitutions. Supposons que la séquence k existe à un stade intermédiaire sur le chemin évolutif entre i et j. Alors que le nombre de différences m_{ij} entre i et j est de 4 (lettres en gras), le nombre de substitutions qui se sont produites entre i et j est au moins $m_{ik} + m_{kj} = 7$.*

Les modèles de substitution (voir chapitre 2) déterminent une matrice de probabilités de substitution $\mathbf{P}(\lambda) = (p_{xy}(\lambda))$, où x et y désignent des nucléotides, des acides aminés ou d'autres caractères biologiques, et $p_{xy}(\lambda)$ désigne la probabilité que x devienne y après avoir évolué le long d'une branche de longueur λ. Nous soulignons que λ n'est pas exprimé en unités de temps : le taux des modèles de substitution est généralement défini de sorte que, en moyenne, il y ait λ substitutions par site le long d'une branche de cette longueur.

Soit π_x la probabilité stationnaire de x. Cela peut être défini comme la limite de $p_{xy}(\lambda)$ quand λ tend vers l'infini, et elle est parfois estimée en prenant simplement la fréquence de x dans le jeu de données. Dans un modèle réversible, la probabilité d'observer deux caractères x et y aux deux extrémités d'une branche de longueur λ est donnée par $\pi_x p_{xy}(\lambda) = \pi_y p_{yx}(\lambda)$.

Maintenant, soit i_k et j_k les k-ièmes caractères alignés des séquences i et j, respectivement. La vraisemblance de d_{ij} en tant que longueur du chemin entre i et j peut alors être exprimée comme suit :

$$L(d_{ij}) = \prod_{k=1}^{m} \pi_{i_k} p_{i_k j_k}(d_{ij}) \qquad [6.3]$$

où m est le nombre de caractères alignés dans l'alignement de i et j (les sites présentant des « – » dans l'alignement sont ignorés). L'estimation ML de la distance d_{ij} est la valeur qui maximise $L(d_{ij})$, et peut être obtenue numériquement ou analytiquement selon le modèle.

Pour illustration, nous considérons le modèle le plus simple de substitution pour l'ADN, le modèle de Jukes-Cantor (Jukes et Cantor 1969), (voir chapitre 2). Dans ce modèle on a $\pi_A = \pi_C = \pi_G = \pi_T = 1/4$ et, si $x \neq y$:

$$p_{xy}(\lambda) = \frac{1}{4}\left(1 - e^{-\frac{4}{3}\lambda}\right), \qquad p_{xx}(\lambda) = \frac{1}{4}\left(1 + 3e^{-\frac{4}{3}\lambda}\right)$$

Rappelons maintenant que m_{ij} désigne le nombre de différences dans l'alignement de i et j. La vraisemblance [6.3] peut alors s'écrire :

$$L(d_{ij}) = \frac{1}{4^{2m}}\left(1 - e^{-\frac{4}{3}d_{ij}}\right)^{m_{ij}}\left(1 + 3e^{-\frac{4}{3}d_{ij}}\right)^{m-m_{ij}} \qquad [6.4]$$

La valeur de d_{ij} qui maximise [6.4] est donnée par :

$$\delta_{ij} = -\frac{3}{4}\ln\left(1 - \frac{4}{3}\frac{m_{ij}}{m}\right)$$

Autrement dit, l'estimation δ_{ij} de la distance entre les deux séquences est une fonction strictement croissante de la proportion de différences m_{ij}/m. De plus, $\delta_{ij} \geq m_{ij}/m$, ce qui correspond au fait que le nombre de substitutions est toujours supérieur ou égal au nombre de mésappariements. D'une certaine manière, δ_{ij} « corrige » m_{ij}/m en tenant compte des substitutions non observées entre i et j.

D'autres modèles aboutissent à des estimations de distance qui sont des fonctions de plusieurs caractéristiques de l'alignement par paires (par exemple des fonctions traitant différemment les transitions et les transversions) donc, en général, les estimations de distance ne sont pas simplement des versions corrigées de la proportion m_{ij}/m. Comme l'estimation de distance par ML est équivalente à l'inférence par ML d'un arbre à 2 taxons non enraciné (donc avec une seule longueur de branche à optimiser), les techniques numériques d'estimation de distance sont largement les mêmes que celles utilisées pour l'optimisation des longueurs de branche.

6.3.2. *Approches alternatives pour estimer les distances*

Historiquement, l'utilisation de séquences alignées n'a pas été la première façon d'estimer les distances évolutives. Par exemple, les plus anciennes méthodes basées sur les distances utilisaient des données telles que les caractères morphologiques (Sokal et Michener 1958), les fréquences d'allèles (Cavalli-Sforza et Edwards 1967), la force des réactions immunologiques croisées (Sarich et Wilson 1967) et les données d'hybridation ADN-ADN (Sibley et Ahlquist 1984). De nouvelles méthodes de calcul de distances évolutives apparaissent fréquemment dans la littérature. Si un examen approfondi des approches proposées dépasse le cadre de ce chapitre, nous donnons ici brièvement un aperçu de deux grandes familles d'idées.

Une ligne de travail est motivée par l'observation que l'alignement des séquences est une tâche complexe nécessitant l'utilisation d'algorithmes qui demandent beaucoup de ressources, qui sont inexacts, sensibles aux erreurs de séquençage et difficiles à appliquer sur des génomes ayant subi des réarrangements structurels. Pour résoudre ces problèmes, les méthodes *alignment-free* (AF) utilisent le même type de données que celles basées sur l'alignement (des séquences nucléotidiques ou protéiques) mais visent à éviter complètement l'étape d'alignement. Une vaste littérature a émergé sur ces idées au cours des 20 dernières années avec des applications qui vont bien au-delà du calcul des dissimilarités en phylogénétique. Les méthodes AF quantifient la similarité/dissimilarité entre les séquences sur la base de caractéristiques telles que (1) les fréquences de certaines classes de mots (par exemple de k-mers qui sont des séquences contiguës de longueur k) (Sims *et al*. 2009 ; Sims et Kim 2011), (2) la longueur des sous-chaînes partagées (Ulitsky *et al*. 2006 ; Leimeister et Morgenstern 2014), (3) les micro-alignements (Haubold *et al*. 2015 ; Leimeister *et al*. 2017). Beaucoup plus d'approches que celles déjà citées ont été proposées, mais un thème récurrent est qu'elles ont souvent un lien avec la compression de données et la théorie de l'information (Vinga 2014). En effet, la similarité entre deux séquences est liée au degré auquel la connaissance de l'une informe sur l'autre (par exemple en permettant de raccourcir la description de l'autre). Voir (Haubold 2014 ; Zielezinski *et al*. 2017 ; 2019 ; Leimeister 2018) pour des aperçus récents de ce sujet de recherche d'actualité.

Un deuxième axe de travail repose sur l'observation que les génomes n'évoluent pas uniquement par des substitutions et par des petites insertions et délétions nucléotidiques (qui agissent « microscopiquement » au niveau des séquences). Les génomes évoluent également du point de vue de leur structure macroscopique par le biais de réarrangements qui changent l'ordre et l'orientation de grandes régions génomiques (par exemple : inversions, duplications, translocations, fusions, fissions, etc.). La comparaison de la structure à grande échelle des génomes de plusieurs espèces peut être très instructive sur leur histoire évolutive, mais, à ce jour, la méthodologie standard en phylogénie n'exploite pas ces informations. De nombreux travaux sur la modélisation de l'évolution du génome à grande échelle ont vu le jour au cours des 30 dernières années, dont David Sankoff et Pavel Pevzner ont été les pionniers. Beaucoup de ces travaux se concentrent sur le calcul du plus petit nombre d'événements de réarrangement nécessaires pour convertir un génome en un autre. Ce nombre peut être utilisé dans un contexte basé sur la distance (mais ce n'est pas idéal, voir *infra*). Nous renvoyons le lecteur au chapitre 5 pour un aperçu plus détaillé de ce domaine.

Nous concluons par une mise en garde. Étant donné que l'inférence phylogénétique n'est souvent pas leur principale application, bon nombre des approches présentées précédemment conduisent à des dissimilarités ou des distances métriques qui ne peuvent pas être interprétées comme des estimations de distances évolutives. (Autrement dit, on ne peut pas s'attendre à ce qu'elles convergent vers une distance d'arbre $\mathbf{d}^T$.) Dans ce cas, la construction d'un arbre sur la base de ces dissimilarités doit être

considérée comme une forme de *clustering*, plutôt que de l'inférence phylogénétique. Aucune garantie théorique telle que la convergence topologique n'est valable dans ce cas.

6.4. L'inférence d'arbre

Nous organisons notre étude des méthodes d'inférence d'arbre autour de trois questions bien distinctes : Toute combinaison de réponses à ces questions définit une méthode différente basée sur les distances.

Q1 – Comment ajuster les longueurs de branche ? Il faut décrire une méthode pour ajuster les longueurs des branches de toute topologie d'arbre τ, de sorte que les distances d'arbre résultantes d_{ij}^T soient aussi proches que possible des distances estimées δ_{ij}. Ceci est généralement réalisé en utilisant les techniques des moindres carrés empruntées à la régression (Cavalli-Sforza et Edwards 1967 ; Fitch et Margoliash 1967).

Q2 – Quoi optimiser ? Il faut définir un critère attribuant un score à tout arbre T obtenu après ajustement des longueurs de branche (question Q1). Ce score doit représenter la plausibilité de T compte tenu des distances estimées. Un choix évident pour cela est le même critère des moindres carrés que l'on a utilisé pour ajuster les longueurs de branche, mais comme nous le décrirons *infra*, de nombreuses méthodes récentes sont basées sur un critère différent, celui du minimum d'évolution (Kidd et Sgaramella-Zonta 1971). Certaines approches (par exemple ADDTREE ; (Sattath et Tversky 1977)) optimisent directement les critères topologiques et contournent ainsi Q1.

Q3 – Comment optimiser ? Il faut concevoir un algorithme pour rechercher l'arbre optimal par rapport au critère en Q2. Comme il s'agit généralement d'un problème d'optimisation difficile, les algorithmes pour cette tâche sont heuristiques et basés sur des idées simples mais efficaces telles que l'ajout de taxons un par un, le *clustering* agglomératif (décrit en section 6.4.3) ou le *hill-climbing* (Swofford *et al*. 1990 ; Felsenstein 2004).

Dans ce qui suit, nous commencerons par étudier la méthodologie pour répondre à Q1 (section 6.4.1). Ensuite, nous introduirons l'un des critères les plus populaires pour Q2, le minimum d'évolution (section 6.4.2) qui est à la base de ce qui reste la plus connue parmi les méthodes basées sur les distances, *neighbor joining*. Nous décrirons cette méthode avec d'autres algorithmes basés sur les mêmes idées (section 6.4.3). Enfin, nous décrirons brièvement des approches récentes qui, à proprement parler, ne sont pas des méthodes basées sur les distances, mais qui partagent avec elles plusieurs idées et la même emphase sur l'efficacité de calcul (section 6.4.4). Étant donné que les techniques heuristiques employées pour répondre à Q3 sont souvent les mêmes que celles utilisées dans d'autres approches pour l'inférence d'arbre telles que le maximum

de vraisemblance et de parcimonie, nous ne nous concentrerons pas sur ces techniques ici.

6.4.1. *Ajustement des longueurs de branches par les moindres carrés*

Compte tenu des distances estimées (δ_{ij}), le but de l'inférence phylogénétique par les moindres carrés (Cavalli-Sforza et Edwards 1967 ; Fitch et Margoliash 1967) est de trouver un arbre phylogénétique T qui minimise une fonction quadratique des écarts entre les estimations δ_{ij} et les distances d'arbre d_{ij}^T. Différents choix pour cette fonction quadratique sont possibles (voir équations [6.5] et [6.6]). Alors que de nombreuses versions de ce problème se sont avérées difficiles (Day 1987), nous nous concentrons ici sur le problème plus simple de l'attribution de longueurs de branche à un arbre de topologie fixe τ. Comme nous le montrerons, des solutions exactes et calculables de manière polynomiale sont disponibles pour cette tâche.

La méthode des moindres carrés a été introduite en phylogénétique dans les années 1960. La fonction-objectif proposée était :

$$\sum_{i<j} w_{ij}(\delta_{ij} - d_{ij}^T)^2 \qquad [6.5]$$

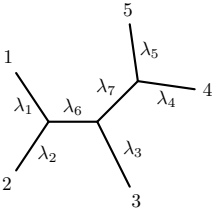
avec $w_{ij} = 1$ (Cavalli-Sforza et Edwards 1967) ou $w_{ij} = {}^1\!/_{\delta_{ij}^2}$ (Fitch et Margoliash 1967). Le choix $w_{ij} = 1$ définit les moindres carrés ordinaires (ou OLS, *ordinary least squares*), et les deux choix sont des cas particuliers des moindres carrés pondérés (ou WLS, *weighted least squares*). En WLS, le poids w_{ij} représente le degré de confiance que l'on peut attacher à l'estimation de distance δ_{ij}. Dans l'idéal, w_{ij} doit être inversement proportionnel à la variance de δ_{ij}, mais en pratique le choix des poids est délicat, car ces variances sont difficiles à évaluer. Un choix possible sur lequel nous reviendrons plusieurs fois par la suite est celui des poids *équilibrés* où w_{ij} est proportionnel à $2^{-|\tau_{ij}|}$ (on rappelle que τ_{ij} désigne l'ensemble des branches entre i et j dans τ), ce qui attribue une confiance plus faible aux distances entre taxons qui sont séparés par plusieurs branches dans τ.

Les approches WLS ne prennent pas en compte les corrélations entre les estimations δ_{ij}, δ_{kl} pour différentes paires de taxons. Ces corrélations peuvent être significatives lorsque les chemins τ_{ij}, τ_{kl} partagent des branches. Pour tenir compte de ces corrélations, les moindres carrés généralisés (ou GLS, *generalized least squares*) (Chakraborty 1977 ; Bulmer 1991) visent à minimiser :

$$\sum_{i<j}\sum_{k<l} w_{ij,kl}(\delta_{ij} - d_{ij}^T)(\delta_{kl} - d_{kl}^T) \qquad [6.6]$$

où $w_{ij,kl}$ doivent (dans l'idéal) être les éléments de l'inverse de la matrice de variance-covariance pour les estimations de distance δ_{ij}. Tout comme OLS est un cas particulier

de WLS, WLS est un cas particulier de GLS obtenu en fixant $w_{ij,kl} = 0$ chaque fois que $\{i, j\} \neq \{k, l\}$. Dans la pratique, GLS est rarement utilisé pour l'inférence phylogénétique en raison de la difficulté d'évaluer les covariances et en raison de ses coûts de calcul élevés par rapport à ceux de OLS et WLS.



$$\begin{bmatrix} 1&1&0&0&0&0&0\\ 1&0&1&0&0&1&0\\ 1&0&0&1&0&1&1\\ 1&0&0&0&1&1&1\\ 0&1&1&0&0&1&0\\ 0&1&0&1&0&1&1\\ 0&1&0&0&1&1&1\\ 0&0&1&1&0&0&1\\ 0&0&1&0&1&0&1\\ 0&0&0&1&1&0&0 \end{bmatrix} \times \begin{bmatrix} \lambda_1\\ \lambda_2\\ \lambda_3\\ \lambda_4\\ \lambda_5\\ \lambda_6\\ \lambda_7 \end{bmatrix} = \begin{bmatrix} \lambda_1+\lambda_2\\ \lambda_1+\lambda_3+\lambda_6\\ \lambda_1+\lambda_4+\lambda_6+\lambda_7\\ \lambda_1+\lambda_5+\lambda_6+\lambda_7\\ \lambda_2+\lambda_3+\lambda_6\\ \lambda_2+\lambda_4+\lambda_6+\lambda_7\\ \lambda_2+\lambda_5+\lambda_6+\lambda_7\\ \lambda_3+\lambda_4+\lambda_7\\ \lambda_3+\lambda_5+\lambda_7\\ \lambda_4+\lambda_5 \end{bmatrix} = \begin{bmatrix} d_{12}^T\\ d_{13}^T\\ d_{14}^T\\ d_{15}^T\\ d_{23}^T\\ d_{24}^T\\ d_{25}^T\\ d_{34}^T\\ d_{35}^T\\ d_{45}^T \end{bmatrix}$$

Figure 6.3. *Les distances d'arbre sont des fonctions linéaires des longueurs de branche. La matrice A_τ peut être utilisée pour exprimer $\mathbf{d}^T$ comme $A_\tau \boldsymbol{\lambda} = \mathbf{d}^T$. Remarquons qu'il s'agit d'un système surdéterminé de 10 équations linéaires. Tout sous-ensemble de 7 équations linéairement indépendantes sur les 10 d'origine est suffisant pour déterminer les 7 inconnues.*

Les longueurs de branche optimales pour les critères WLS et GLS (voir équations [6.5] et [6.6]) peuvent être exprimées de manière concise en notation matricielle. Pour cela, nous allons considérer $\boldsymbol{\delta}$ et $\mathbf{d}^T$ comme des vecteurs colonnes dont les éléments correspondent à des sous-ensembles de deux taxons, et sont classés dans l'ordre lexicographique habituel (voir par exemple le vecteur à droite de la figure 6.3). De plus, étant donné toute topologie τ, soit $A_\tau = (a_{ij,e})$ une matrice dont les lignes correspondent à des sous-ensembles de deux taxons (dans l'ordre lexicographique) et dont les colonnes correspondent aux branches de τ. Ses éléments sont tels que $a_{ij,e} = 1$ si e est sur le chemin entre i et j dans τ, et $a_{ij,e} = 0$ sinon. Compte tenu de ces notations, on peut écrire :

$$A_\tau \boldsymbol{\lambda} = \mathbf{d}^T$$

où $\boldsymbol{\lambda} = (\lambda_e)$ désigne les longueurs de branche de T sous forme vectorielle (voir figure 6.3 pour un exemple illustrant ces notations). Les fonctions-objectif de OLS, WLS et GLS peuvent alors être écrites sous forme matricielle :

$$(\boldsymbol{\delta} - A_\tau \boldsymbol{\lambda})^t \, W \, (\boldsymbol{\delta} - A_\tau \boldsymbol{\lambda}) \qquad [6.7]$$

où $W = (w_{ij,kl})$ contient les poids et t désigne la transposée de la matrice. WLS et OLS sont obtenus lorsque W est une matrice diagonale et la matrice identité, respectivement. Les longueurs de branche qui minimisent [6.7] peuvent alors être exprimées comme :

$$\widehat{\boldsymbol{\lambda}}_\tau = \left(A_\tau^t W A_\tau\right)^{-1} A_\tau^t W \boldsymbol{\delta} \qquad [6.8]$$

Les calculs matriciels dans l'équation [6.8] sont coûteux, mais plusieurs propriétés des matrices impliquées peuvent être exploitées pour accélérer les calculs (Gascuel 1997b ; Bryant et Waddell 1998). Pour WLS, la complexité de calcul est dominée par l'inversion de la matrice dans l'équation [6.8], ou de manière équivalente par la résolution du système d'équations linéaires correspondant. Ces algorithmes peuvent calculer $\widehat{\boldsymbol{\lambda}}_\tau$ pour WLS en temps $O(n^3)$ où nous rappelons que n désigne le nombre de taxons (Bryant et Waddell 1998).

Dans certains cas, les calculs peuvent être accélérés davantage en notant l'existence de formules analytiques beaucoup plus simples que [6.8]. Cela a été initialement remarqué pour OLS (Vach 1989 ; Rzhetsky et Nei 1993), puis pour WLS avec des poids équilibrés (Desper et Gascuel 2004), et plus tard étendu à une sous-classe plus vaste de WLS, les cas où les poids w_{ij} sont *multiplicatifs* (Mihaescu et Pachter 2008 ; Pardi et Gascuel 2012). L'importance pratique de ces résultats algorithmiques réside dans le fait que le calcul de $\widehat{\boldsymbol{\lambda}}_\tau$ est souvent appelé comme sous-procédure des méthodes d'inférence d'arbres basées sur les distances.

Nous concluons en remarquant qu'aucune des approches décrites ici ne garantit que les longueurs de branche ajustées $\widehat{\boldsymbol{\lambda}}_\tau$ sont toutes positives. Les approches des moindres carrés conduisent parfois à des longueurs négatives, ce qui ne correspond à aucune réalité biologique. Bien entendu, on peut toujours contraindre les longueurs de branche à être non négatives dans un critère des moindres carrés (NNLS, *non-negative least squares*), une approche qui, cependant, augmente davantage les coûts de calcul (Lawson et Hanson 1974).

6.4.2. *Des moindres carrés au minimum d'évolution*

Une fois qu'un moyen d'ajuster les longueurs de branche a été déterminé, la question suivante (Q2) est de savoir comment mesurer la plausibilité des arbres qu'on obtient. Une possibilité est d'adopter le même critère des moindres carrés que celui utilisé pour ajuster les longueurs de branche et de rechercher des arbres qui minimisent ce critère. Ainsi, une topologie est considérée comme plausible si ses distances d'arbre peuvent être ajustées étroitement aux distances estimées. Étant donné que ces approches utilisent les moindres carrés comme réponse aux deux questions Q1 et Q2, nous les appellerons « LS+LS ».

Il est intéressant de noter que toute méthode fournissant l'arbre optimal par rapport à un critère LS+LS est topologiquement convergente (voir définition 6.1). Cela découle du théorème 6.2 qui implique que si $\boldsymbol{\delta} \approx \mathbf{d}^T$, le critère des moindres carrés [6.7] ne converge à 0 que lorsque τ est la topologie de T (voir par exemple (Pardi et Gascuel 2012) pour plus de détails sur cet argument). En pratique, les approches LS+LS sont connues pour être précises lorsque les longueurs de branche négatives

sont interdites (Kuhner et Felsenstein 1994), une contrainte qui peut être imposée dans des programmes tels que FITCH (Felsenstein 1997) et PAUP* (Swofford 2003).

Une approche alternative (le *minimum d'évolution* (ME)) utilise comme score pour un arbre la somme des longueurs de branche ajustées $\widehat{\boldsymbol{\lambda}}_\tau$. Cette somme, que nous noterons ici $S(\widehat{\boldsymbol{\lambda}}_\tau)$, est simplement la longueur totale de l'arbre. L'intuition sous-jacente à ME est la même que dans le maximum de parcimonie : de la même manière que les explications simples sont préférables aux explications complexes, ici les arbres les plus courts sont considérés comme les plus plausibles.

Un détail important sur ME est la façon dont les longueurs de branche négatives sont comptées dans $S(\widehat{\boldsymbol{\lambda}}_\tau)$. Cela peut être fait de différentes manières : elles peuvent être simplement exclues de la somme (Swofford *et al*. 1990), ou $S(\widehat{\boldsymbol{\lambda}}_\tau)$ peut être défini comme la somme des valeurs absolues des longueurs de branche (Kidd et Sgaramella-Zonta 1971). Actuellement, l'approche la plus courante consiste à définir $S(\widehat{\boldsymbol{\lambda}}_\tau)$ comme la somme de toutes les longueurs de branche, quel que soit leur signe (Saitou et Imanishi 1989 ; Rzhetsky et Nei 1993). Bien que cela semble favoriser les longueurs de branches négatives, en pratique les longueurs de branches négatives sont souvent largement compensées par des longueurs de branches positives qui accroissent la longueur totale de l'arbre.

D'un point de vue théorique, pourquoi ME fonctionne n'est pas aussi clair qu'on pourrait le penser. Dans les années 1990, on savait que l'inférence de l'arbre de longueur minimale $S(\widehat{\boldsymbol{\lambda}}_\tau)$ avec des longueurs de branche ajustées par OLS (c'est-à-dire la méthode « OLS+ME ») est topologiquement convergente (Rzhetsky et Nei 1993). Cependant, il a été montré plus tard que ce résultat ne vaut pas plus généralement pour WLS+ME (Gascuel *et al*. 2001). Il s'avère que la convergence topologique n'est valable que pour un sous-ensemble d'approches WLS+ME. À ce jour, cette condition préalable importante de l'inférence basée sur les distances a été démontrée pour toutes les instances de WLS+ME avec des poids multiplicatifs, pour les trois façons de traiter les longueurs de branche négatives décrites précédemment (Pardi et Gascuel 2012).

Un cas particulier important de ceci est WLS+ME avec des poids équilibrés qui est également connu sous le nom de BME (*balanced minimum evolution*) (Pauplin 2000 ; Desper et Gascuel 2002 ; 2004 ; Lefort *et al*. 2015). Ce critère d'optimisation présente plusieurs caractéristiques mathématiques intéressantes (Semple et Steel 2004), notamment le fait que pour les arbres entièrement résolus (binaires), la longueur de l'arbre ajusté peut être exprimée de façon élégante comme une simple fonction des distances estimées :

$$S(\widehat{\boldsymbol{\lambda}}_\tau) = \sum_{i<j} 2^{1-|\tau_{ij}|}\delta_{ij} \qquad [6.9]$$

BME est important pour mieux comprendre certaines des méthodes les plus centrales de la phylogénétique basée sur les distances, comme nous l'expliquerons ci-dessous.

6.4.3. *NJ et autres algorithmes agglomératifs*

Neighbor joining (NJ) est un algorithme de *clustering* agglomératif, c'est-à-dire qu'il construit un arbre de façon *bottom-up* (ascendante) en alternant les deux étapes suivantes jusqu'à ce que l'arbre soit complet :

– étape de sélection : en fonction des distances estimées, on choisit deux taxons *actifs* i et j à « agglomérer ». C'est-à-dire qu'on les relie à un nouveau taxon (ij) représentant leur ancêtre commun immédiat. (Au départ, tous les taxons sont actifs.) ;

– étape de réduction : on retire i et j de la liste des taxons actifs, et on définit de nouvelles distances entre le nouveau taxon actif (ij) et les taxons actifs restants.

Ces deux étapes sont communes à tous les algorithmes agglomératifs. Pour NJ, il existe plusieurs manières équivalentes de spécifier ces étapes (Saitou et Nei 1987 ; Studier et Keppler 1988 ; Gascuel 1994). Ici, nous décrivons l'algorithme le plus rapide (Studier et Keppler 1988). Dans l'étape de sélection, NJ agglomère les taxons i et j qui minimisent :

$$(|A| - 2)\delta_{ij} - \sum_{k \in A} (\delta_{ik} + \delta_{jk}) \qquad [6.10]$$

où A est l'ensemble des taxons actifs. Quant à l'étape de réduction, la nouvelle distance entre (ij) et tout autre taxon actif k est définie par :

$$\delta_{(ij)k} = \frac{1}{2}(\delta_{ik} + \delta_{jk} - \delta_{ij}) \qquad [6.11]$$

Le critère [6.10] est mystérieux à première vue, mais il est possible de montrer que si $\boldsymbol{\delta} = \mathbf{d}^T$ pour un arbre phylogénétique T, alors les taxons i et j qui minimisent [6.10] doivent former une « cerise » dans T (c'est-à-dire que i et j sont reliés par un chemin d'exactement 2 branches ; voir aussi (Bryant 2005) pour une étude approfondie de [6.10]). De plus, il est facile de voir que si $\boldsymbol{\delta} = \mathbf{d}^T$, la valeur de $\delta_{(ij)k}$ obtenue avec [6.11] est égale à la longueur du chemin entre k et le seul nœud séparant i et j. De plus, ces observations sur [6.10] et [6.11] impliquent la convergence topologique de NJ.

Différents choix pour les deux étapes ci-dessus conduisent à d'autres algorithmes agglomératifs bien connus : le *clustering* à saut minimum (ou *single-linkage clustering* (Sneath 1957)), le *clustering* à saut moyen (ou *group-average linkage clustering*) communément appelé UPGMA (Sokal et Michener 1958), WPGMA (Sokal et Sneath

1963), ADDTREE (Sattath et Tversky 1977), UNJ (Gascuel 1997b) et BIONJ (Gascuel 1997a). La dernière méthode est un cas particulier de l'approche MVR (Gascuel 2000a) qui adapte l'étape de réduction pour tenir compte des variances et des covariances des estimations de distance δ_{ij}. Il est similaire dans l'esprit à un autre algorithme agglomératif, Weighbor (Bruno *et al*. 2000), qui modifie également le critère de sélection et utilise une formule différente (Bulmer 1991) pour évaluer les variances des distances.

Parmi ces algorithmes, les plus rapides sont les algorithmes de *clustering* classiques (*single-linkage*, UPGMA et WPGMA) qui construisent un arbre en temps $O(n^2)$ (Sibson 1973 ; Murtagh 1984 ; Gronau et Moran 2007). Cependant, ces approches de *clustering* ne sont précises (en fait, topologiquement convergentes) que lorsque les distances estimées se rapprochent d'une distance d'arbre pour un arbre dont les feuilles sont toutes à la même distance d'un point central, à interpréter comme leur dernier ancêtre commun. Cela se produit lorsque les taux d'évolution sont constants et que la longueur des branches est directement proportionnelle au temps, une hypothèse connue sous le nom d'*horloge moléculaire*. Pour cette raison, ces algorithmes rapides ne sont pas un bon choix quand l'hypothèse de l'horloge moléculaire est suspectée d'être violée.

Outre ces algorithmes de *clustering* rapide (et ADDTREE et MVR qui sont plus lents, ils s'exécutent en temps $O(n^4)$), tous les algorithmes agglomératifs présentés ici construisent un arbre en temps $O(n^3)$. Cela inclut NJ où chacune des $O(n)$ étapes de sélection est effectuée en temps $O(n^2)$ *via* le précalcul des sommes $\sum_{k \in A} \delta_{ik}$ (Studier et Keppler 1988). L'une des raisons du succès continu de NJ est le fait qu'il a longtemps été considéré comme réalisant un très bon compromis entre précision et rapidité.

Récemment, beaucoup de travail a été consacré à la mise au point d'implémentations efficaces de NJ nécessitant moins de mémoire et temps d'exécution que l'implémentation standard. Parmi ces travaux, on trouve QuickTree (Howe *et al*. 2002), QuickJoin (Mailund et Pedersen 2004), la méthode de (Zaslavsky et Tatusova 2008), NINJA (Wheeler 2009), RapidNJ et ERapidNJ (Simonsen *et al*. 2010). Une idée commune de ces approches est d'accélérer l'identification de la paire de taxons optimale par rapport au critère [6.10] en limitant la recherche à un sous-ensemble de paires garanti pour contenir la paire optimale. Bien que la complexité temporelle reste de $O(n^3)$ dans le pire cas, ces approches entraînent une amélioration spectaculaire de l'efficacité permettant l'inférence d'arbres avec plus de 50 000 taxons en quelques heures sur un PC normal (Wheeler 2009 ; Simonsen *et al*. 2010).

Des algorithmes encore plus rapides peuvent être obtenus en employant des heuristiques au lieu d'algorithmes exacts pour trouver une bonne paire de taxons par rapport au critère [6.10]. C'est le cas de FastNJ (Elias et Lagergren 2009) et RelaxedNJ (Evans *et al*. 2006), qui est implémenté dans Clearcut (Sheneman *et al*. 2006). Dans le cas de

FastNJ, les taxons à agglomérer sont sélectionnés parmi une liste de $O(n)$ paires, ce qui implique un temps d'exécution de $O(n^2)$. Il faut s'attendre à une certaine perte de précision pour FastNJ et RelaxedNJ, car, en général, ces approches ne construisent pas le même arbre que NJ. Malgré cela, ces méthodes restent topologiquement convergentes (Elias et Lagergren 2009).

Une question qui a intrigué les phylogénéticiens pendant un certain temps est de savoir si NJ est lié à l'un des critères d'optimisation (voir section 6.4.2), et plus généralement s'il peut être décrit comme une combinaison de réponses aux questions Q1 et Q3 (voir section 6.4). Comme on peut le lire dans (Felsenstein 2004, p. 170), « il a souvent été suggéré que NJ ait une certaine relation avec les moindres carrés ordinaires et avec le minimum d'évolution, sans être définissable comme un algorithme pour l'un ou l'autre ». D'autres liens similaires ont été signalés par d'autres auteurs (Saitou et Nei 1987 ; Gascuel 1994). Cependant, si la bonne performance de NJ était due à sa relation avec OLS+ME, alors nous nous attendrions à ce que de meilleurs arbres (plus courts) par rapport au critère OLS+ME soient également plus précis, c'est-à-dire plus proches du vrai arbre, que les arbres inférés avec NJ. Ceci est en fait contredit par plusieurs expériences basées sur des données simulées (Saitou et Imanishi 1989 ; Kumar 1996 ; Gascuel 2000b ; Desper et Gascuel 2002).

Une solution à cette énigme est venue de la prise de conscience que le véritable critère d'optimisation derrière NJ est BME (Desper et Gascuel 2005 ; Gascuel et Steel 2006). Pour comprendre ce résultat, il faut d'abord noter que BME attribue également une longueur $S(\widehat{\boldsymbol{\lambda}}_\tau)$ aux arbres qui ne sont pas entièrement résolus (non binaires). Pour ces topologies d'arbres, il existe une formule légèrement plus complexe que celle de l'équation [6.9] (Semple et Steel 2004 ; Gascuel et Steel 2006). Considérons maintenant un algorithme agglomératif tel que, à tout moment, tous les nœuds actifs sont connectés à un nœud central (l'algorithme commence par l'arbre en étoile sur $\{1, 2, \ldots, n\}$ lorsque tous les taxons sont actifs). À chaque étape de sélection, cet algorithme choisit la paire de taxons actifs (i, j) dont l'agglomération produit l'arbre avec la plus petite longueur de BME $S(\widehat{\boldsymbol{\lambda}}_\tau)$. Il s'avère que cet algorithme glouton fait *exactement* la même séquence de choix que NJ (Gascuel et Steel 2006).

Motivées par l'observation que NJ peut être vu comme un algorithme glouton pour BME, d'autres méthodes guidées par BME ont été proposées (Desper et Gascuel 2002 ; Catanzaro *et al*. 2012). Ces méthodes diffèrent de NJ par leur réponse à la question Q3, c'est-à-dire par l'algorithme pour rechercher l'arbre optimal par rapport au critère BME. Par exemple, FastME (Desper et Gascuel 2002 ; Lefort *et al*. 2015) implémente des heuristiques pour construire un arbre initial et pour effectuer une recherche locale dans l'espace des topologies d'arbre. Bien que le temps d'exécution de FastME soit comparable à celui de NJ, sa précision de reconstruction est supérieure (Desper et Gascuel 2002 ; 2004 ; Vinh et von Haeseler 2005), confirmant ainsi l'attractivité de BME comme critère d'optimisation pour l'inférence phylogénétique basée sur les distances.

6.4.4. *Au-delà des distances*

Un goulot d'étranglement de calcul important de toutes les méthodes que nous avons présentées jusqu'à présent est qu'elles nécessitent l'estimation initiale et le stockage des distances pour toutes les paires de taxons, ce qui demande un temps $O(m \cdot n^2)$, en supposant que les distances sont estimées à partir de séquences alignées de longueur m, et un espace $O(n^2)$. Cependant, comme nous l'avons dit dans la section 6.2, toutes les distances ne sont pas nécessaires pour reconstruire une phylogénie, ce qui signifie, qu'en principe, il devrait être possible d'améliorer ces bornes. Certaines distances, celles avec de fortes variances, peuvent même être trompeuses pour l'inférence. De plus, pour les grands jeux de données avec des centaines de milliers de taxons, réduire l'utilisation de la mémoire peut être une nécessité : le simple stockage de la matrice de distance entière pour 100 000 taxons nécessite généralement environ 20 Go de mémoire, ce qui peut être problématique pour de nombreux utilisateurs.

Au cours des dernières années, un certain nombre d'approches ont été proposées pour contourner ce goulot d'étranglement. L'une des plus largement utilisées est FastTree (Price *et al*. 2009) dont la stratégie de construction d'un arbre initial est inspirée de NJ, mais avec deux différences fondamentales, la première visant une amélioration de la précision d'inférence, et la seconde l'efficacité du calcul. Nous décrivons ci-après brièvement les principaux points distinctifs de FastTree, car ils expliquent ses bonnes performances et donc sa popularité.

La première différence avec NJ et d'autres méthodes basées sur les distances est que FastTree stocke des profils de séquence pour les taxons actifs (au lieu d'une matrice de distance) et ne calcule la distance entre deux profils que si la paire correspondante est candidate à l'agglomération. Après chaque agglomération, un profil de séquence pour le nouveau nœud (ij) est calculé comme la moyenne arithmétique des profils pour i et j. Les paires de taxons à joindre sont sélectionnées sur la base d'un critère similaire à [6.10], mais où les distances ne sont pas corrigées et sont définies sur la base des profils stockés pour les taxons actifs.

La deuxième différence consiste à maintenir une liste de $O(\sqrt{n})$ « top-hits » pour chaque taxon (c'est-à-dire des voisins putatifs) qui est combinée aux stratégies de FastNJ (Elias et Lagergren 2009) et RelaxedNJ (Evans *et al*. 2006) pour réduire le nombre de paires de taxons à considérer pour l'agglomération. À la fin de son exécution, FastTree n'aura considéré que $O(n^{1.5} \log n)$ paires, contre $O(n^3)$ pour NJ et $O(n^2)$ pour FastNJ. La construction d'un arbre initial, avec une complexité revendiquée de $O(m \cdot n^{1.5} \log n)$ en temps et $O(mn + n^{1.5})$ en espace, est ensuite suivie d'une recherche locale basée sur des réarrangements de topologie (NNI, SPR) en utilisant des techniques similaires à celles implémentées dans FastME (Desper et Gascuel 2002).

FastTree est plus rapide et demande moins de mémoire que les méthodes basées sur les distances évoquées jusqu'à présent, et il peut facilement gérer des jeux de données contenant des centaines de milliers de séquences. Étant donné le rôle central que jouent les profils de séquence, il est difficile de voir FastTree comme une méthode purement basée sur les distances. FastTree partage des idées avec des méthodes basées sur les caractères telles que le maximum de parcimonie bénéficiant en particulier d'informations sur les séquences ancestrales.

6.5. Conclusion

Malgré la simplicité de leur approche et la perte d'informations qui est nécessairement entraînée par la compression des séquences en une matrice numérique, les méthodes basées sur les distances ne sont pas seulement efficaces en termes de calcul, mais aussi remarquablement précises. La précision de l'inférence des méthodes basées sur les distances a fait l'objet de nombreuses études de simulation (Saitou et Imanishi 1989 ; Kuhner et Felsenstein 1994 ; Kumar 1996 ; Gascuel 2000b ; Nakhleh *et al*. 2002 ; Desper et Gascuel 2004). L'idée générale de ces travaux est de générer des séquences en utilisant des modèles de substitution et des arbres modèles connus, et ensuite de comparer les arbres construits par les méthodes testées aux arbres modèles.

D'après les études de simulation, il ressort que, bien que non comparable à celle des méthodes basées sur la vraisemblance, la précision des méthodes basées sur les distances est compétitive avec celle du maximum de parcimonie (MP), MP étant supérieur pour les arbres à branches courtes et inférieur lorsque l'effet de substitutions multiples au même site devient important. La raison en est en grande partie intuitive : alors que les méthodes basées sur les distances tiennent naturellement compte des substitutions multiples dans la façon dont elles estiment les distances, MP ne modélise même pas les longueurs de branche, ce qui entraîne des problèmes tels que la non-convergence statistique dans des cas extrêmes (Felsenstein 1978). Une autre leçon importante de ces études empiriques est l'importance d'améliorer l'arbre reconstruit initialement *via* des réarrangements topologiques tels que NNI et SPR (par exemple (Vinh et von Haeseler 2005)). C'est pourquoi la plupart des méthodes modernes de reconstruction d'arbres incluent cette étape importante, comme nous l'avons vu pour FastME (Desper et Gascuel 2002) et FastTree (Price *et al*. 2009).

Outre le bon compromis entre efficacité et précision, un autre avantage important des méthodes basées sur les distances est leur polyvalence : elles peuvent être utilisées non seulement avec des données de séquence alignées, mais dans tous les contextes où les données permettent d'estimer une mesure de la distance évolutive entre les taxons. (Voir à nouveau section 6.3.2 pour des approches alternatives d'estimation de distances.)

Un domaine où les méthodes de distance prouvent également leur polyvalence est la phylogénomique où elles sont souvent utilisées (seules ou combinées avec d'autres

méthodes) pour l'inférence d'arbres d'espèces à partir de grandes collections d'alignements provenant de différents loci génomiques. Une approche consiste à résumer les informations provenant de chaque locus dans une matrice de distance distincte. Cela produit généralement des milliers de matrices de distance dont l'analyse combinée est beaucoup plus efficace que l'analyse d'alignements concaténés (Lapointe et Cucumel 1997 ; Criscuolo *et al*. 2006 ; Binet *et al*. 2016).

Une autre ligne de travail pour l'inférence des arbres d'espèces cherche à tenir compte du tri de lignées incomplet (ILS, *incomplete lineage sorting*), qui peut amener les arbres phylogénétiques pour différents loci à avoir non seulement des longueurs de branches différentes, mais aussi des topologies différentes. Plusieurs approches « ILS-compatibles » utilisent l'inférence basée sur les distances comme élément clé. Les distances sont soit dérivées des topologies d'arbres de gènes inférées à chaque locus (voir par exemple STAR (Liu *et al*. 2009), NJst (Liu et Yu 2011), ASTRID (Vachaspati et Warnow 2015)) soit estimées à partir de temps de coalescences interspécifiques (voir par exemple STEAC (Liu *et al*. 2009), MT (Liu *et al*. 2010), GLASS (Mossel et Roch 2008), iGLASS (Jewett et Rosenberg 2012)).

Tout ceci témoigne de l'utilité des méthodes basées sur les distances qui continueront sans doute à trouver de nouvelles applications en biologie évolutive à chaque fois que l'on recherchera une efficacité de calcul combinée à de solides garanties théoriques.

6.6. Bibliographie

Atteson, K. (1999). The performance of neighbor-joining methods of phylogenetic reconstruction. *Algorithmica*, 25(2-3), 251–278.

Barthélémy, J.-P., Guénoche, A. (1991). *Trees and proximity representations*. John Wiley & Sons.

Binet, M., Gascuel, O., Scornavacca, C., Douzery, E.J., Pardi, F. (2016). Fast and accurate branch lengths estimation for phylogenomic trees. *BMC Bioinformatics*, 17(1), 23.

Bruno, W.J., Socci, N.D., Halpern, A.L. (2000). Weighted neighbor joining : a likelihood-based approach to distance-based phylogeny reconstruction. *Molecular Biology and Evolution*, 17(1), 189–197.

Bryant, D.J. (2005). On the uniqueness of the selection criterion in neighbor-joining. *Journal of Classification*, 22(1), 3–15.

Bryant, D.J., Waddell, P.J. (1998). Rapid evaluation of least-squares and minimum-evolution criteria on phylogenetic trees. *Molecular Biology and Evolution*, 15(10), 1346–1359.

Bulmer, M. (1991). Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution*, 8(6), 868–883.

Buneman, P. (1971). The recovery of trees from measures of dissimilarity. Dans *Mathematics in the Archeological and Historical Sciences*, Hodson, F.R., Kendall, D.G., Tautu, P. (dir.). Edinburgh University Press, 387–395.

Catanzaro, D., Labbé, M., Pesenti, R., Salazar-González, J.-J. (2012). The balanced minimum evolution problem. *INFORMS Journal on Computing*, 24(2), 276–294.

Cavalli-Sforza, L.L., Edwards, A.W. (1967). Phylogenetic analysis. models and estimation procedures. *American Journal of Human Genetics*, 19(3 Pt 1), 233.

Chakraborty, R. (1977). Estimation of time of divergence from phylogenetic studies. *Canadian Journal of Genetics and Cytology*, 19(2), 217–223.

Criscuolo, A., Berry, V., Douzery, E.J., Gascuel, O. (2006). Sdm: a fast distance-based approach for (super) tree building in phylogenomics. *Systematic Biology*, 55(5), 740–755.

Day, W.H. (1987). Computational complexity of inferring phylogenies from dissimilarity matrices. *Bulletin of Mathematical Biology*, 49(4), 461–467.

Desper, R., Gascuel, O. (2002). Fast and accurate phylogeny reconstruction algorithms based on the minimum-evolution principle. *Journal of Computational Biology*, 9(5), 687–705.

Desper, R., Gascuel, O. (2004). Theoretical foundation of the balanced minimum evolution method of phylogenetic inference and its relationship to weighted least-squares tree fitting. *Molecular Biology and Evolution*, 21(3), 587–598.

Desper, R., Gascuel, O. (2005). The minimum evolution distance-based approach to phylogenetic inference. Dans *Mathematics of Evolution and Phylogeny*, Gascuel, O. (dir.). Oxford University Press, 1–32.

Dress, A.W., Huber, K.T., Steel, M. (2012). 'lassoing'a phylogenetic tree i : basic properties, shellings, and covers. *Journal of mathematical biology*, 65(1), 77–105.

Elias, I., Lagergren, J. (2009). Fast neighbor joining. *Theoretical Computer Science*, 410(21), 1993–2000 [En ligne]. Disponible à l'adresse : http ://www.sciencedirect.com/science/article/pii/S0304397508009079.

Evans, J., Sheneman, L., Foster, J. (2006). Relaxed neighbor joining : a fast distance-based phylogenetic tree construction method. *Journal of Molecular Evolution*, 62(6), 785–792.

Felsenstein, J. (1978). Cases in which parsimony or compatibility methods will be positively misleading. *Systematic Zoology*, 27(4), 401–410.

Felsenstein, J. (1997). An alternating least squares approach to inferring phylogenies from pairwise distances. *Systematic Biology*, 46(1), 101–111.

Felsenstein, J. (2004). *Inferring Phylogenies*. Sinauer associates Sunderland, Massuchusetts.

Fitch, W.M., Margoliash, E. (1967). Construction of phylogenetic trees. *Science*, 155(3760), 279–284.

Gascuel, O. (1994). A note on sattath and tversky's, saitou and nei's, and studier and keppler's algorithms for inferring phylogenies from evolutionary distances. *Molecular Biology and Evolution*, 11(6), 961–963.

Gascuel, O. (1997a). BIONJ: an improved version of the nj algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7), 685–695.

Gascuel, O. (1997b). Concerning the NJ algorithm and its unweighted version, UNJ. *Mathematical Hierarchies and Biology*, 37, 149–171.

Gascuel, O. (2000a). Data model and classification by trees : the minimum variance reduction (mvr) method. *Journal of Classification*, 17(1), 67–99.

Gascuel, O. (2000b). On the optimization principle in phylogenetic analysis and the minimum-evolution criterion. *Molecular Biology and Evolution*, 17(3), 401–405.

Gascuel, O., Steel, M. (2006). Neighbor-joining revealed. *Molecular Biology and Evolution*, 23(11), 1997–2000.

Gascuel, O., Steel, M. (2016). A 'stochastic safety radius' for distance-based tree reconstruction. *Algorithmica*, 74(4), 1386–1403.

Gascuel, O., Bryant, D., Denis, F. (2001). Strengths and limitations of the minimum evolution principle. *Systematic Biology*, 50(5), 621–627.

Gronau, I., Moran, S. (2007). Optimal implementations of upgma and other common clustering algorithms. *Information Processing Letters*, 104(6), 205–210.

Guindon, S., Gascuel, O. (2003). A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. *Systematic Biology*, 52(5), 696–704.

Haubold, B. (2014). Alignment-free phylogenetics and population genetics. *Briefings in Bioinformatics*, 15(3), 407–418.

Haubold, B., Klötzl, F., Pfaffelhuber, P. (2015). Andi: Fast and accurate estimation of evolutionary distances between closely related genomes. *Bioinformatics*, 31(8), 1169–1175.

Hein, J.J. (1989). An optimal algorithm to reconstruct trees from additive distance data. *Bulletin of Mathematical Biology*, 51(5), 597–603.

Howe, K., Bateman, A., Durbin, R. (2002). Quicktree : building huge neighbour-joining trees of protein sequences. *Bioinformatics*, 18(11), 1546–1547.

Huber, K.T., Kettleborough, G. (2015). Distinguished minimal topological lassos. *SIAM Journal on Discrete Mathematics*, 29(2), 940–961.

Jewett, E.M., Rosenberg, N.A. (2012). Iglass: an improvement to the glass method for estimating species trees from gene trees. *Journal of Computational Biology*, 19(3), 293–315.

Jukes, T., Cantor, C. (1969). Evolution of protein molecules. Dans *Mammalian Protein Metabolism*, Munro, H. (dir.). Academic Press, 21–132.

Kettleborough, G., Dicks, J., Roberts, I.N., Huber, K.T. (2015). Reconstructing (super)trees from data sets with missing distances : not all is lost. *Molecular Biology and Evolution*, 32(6), 1628–1642.

Kidd, K.K., Sgaramella-Zonta, L.A. (1971). Phylogenetic analysis : concepts and methods. *American Journal of Human Genetics*, 23(3), 235.

Kuhner, M.K., Felsenstein, J. (1994). A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11(3), 459–468.

Kumar, S. (1996). A stepwise algorithm for finding minimum evolution trees. *Molecular Biology and Evolution*, 13(4), 584–593.

Lapointe, F.-J., Cucumel, G. (1997). The average consensus procedure: combination of weighted trees containing identical or overlapping sets of taxa. *Systematic Biology*, 46(2), 306–312.

Larkin, M.A., Blackshields, G., Brown, N.P., 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(21), 2947–2948.

Lawson, C.L., Hanson, R.J. (1974). *Solving least squares problems*. Prentice-hall.

Lefort, V., Desper, R., Gascuel, O. (2015). Fastme 2.0: a comprehensive, accurate, and fast distance-based phylogeny inference program. *Molecular Biology and Evolution*, 32(10), 2798–2800.

Leimeister, C.-A. (2018). Filtered Spaced-word Matches: a Novel Approach to Fast and Accurate Sequence Comparison. Thèse de doctorat, Université Georg-August de Göttingen.

Leimeister, C.-A., Morgenstern, B. (2014). Kmacs: the k-mismatch average common substring approach to alignment-free sequence comparison. *Bioinformatics*, 30(14), 2000–2008.

Leimeister, C.-A., Sohrabi-Jahromi, S., Morgenstern, B. (2017). Fast and accurate phylogeny reconstruction using filtered spaced-word matches. *Bioinformatics*, 33(7), 971–979.

Liu, L., Yu, L. (2011). Estimating species trees from unrooted gene trees. *Systematic Biology*, 60(5), 661–667.

Liu, L., Yu, L., Pearl, D.K., Edwards, S.V. (2009). Estimating species phylogenies using coalescence times among sequences. *Systematic Biology*, 58(5), 468–477.

Liu, L., Yu, L., Pearl, D.K. (2010). Maximum tree: a consistent estimator of the species tree. *Journal of mathematical biology*, 60(1), 95–106.

Mailund, T., Pedersen, C.N. (2004). Quickjoin–fast neighbour-joining tree reconstruction. *Bioinformatics*, 20(17), 3261–3262.

Mihaescu, R., Pachter, L. (2008). Combinatorics of least-squares trees. *Proceedings of the National Academy of Sciences*, 105(36), 13206–13211.

Mihaescu, R., Levy, D., Pachter, L. (2009). Why neighbor-joining works. *Algorithmica*, 54(1), 1–24.

Mossel, E., Roch, S. (2008). Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 7(1), 166–171.

Murtagh, F. (1984). Complexities of hierarchic clustering algorithms: state of the art. *Computational Statistics Quarterly*, 1(2), 101–113.

Nakhleh, L., Moret, B., Roshan, U., St John, K., Sun, J., Warnow, T. (2002). The accuracy of fast phylogenetic methods for large datasets. *Pacific Symposium on Biocomputing*, 211–222.

Pardi, F., Gascuel, O. (2012). Combinatorics of distance-based tree inference. *Proceedings of the National Academy of Sciences*, 109(41), 16443–16448.

Pauplin, Y. (2000). Direct calculation of a tree length using a distance matrix. *Journal of Molecular Evolution*, 51(1), 41–47.

Pereira, J.S. (1969). A note on the tree realizability of a distance matrix. *Journal of Combinatorial Theory*, 6(3), 303–310.

Price, M.N., Dehal, P.S., Arkin, A.P. (2009). Fasttree: computing large minimum evolution trees with profiles instead of a distance matrix. *Molecular Biology and Evolution*, 26(7), 1641–1650.

Rzhetsky, A., Nei, M. (1993). Theoretical foundation of the minimum-evolution method of phylogenetic inference. *Molecular Biology and Evolution*, 10(5), 1073–1095.

Saitou, N., Imanishi, T. (1989). Relative efficiencies of the fitch-margoliash, maximum-parsimony, maximum-likehood, minimum-evolution, and neighbor-joining methods of phylogenetic tree construction in obtaining the correct tree. *Molecular Biology and Evolution*, 6(5), 514–525.

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

Sarich, V.M., Wilson, A.C. (1967). Immunological time scale for hominid evolution. *Science*, 158(3805), 1200–1203.

Sattath, S., Tversky, A. (1977). Additive similarity trees. *Psychometrika*, 42(3), 319–345.

Semple, C., Steel, M. (2003). *Phylogenetics*. Oxford University Press.

Semple, C., Steel, M. (2004). Cyclic permutations and evolutionary trees. *Advances in Applied Mathematics*, 32(4), 669–680.

Sheneman, L., Evans, J., Foster, J.A. (2006). Clearcut: a fast implementation of relaxed neighbor joining. *Bioinformatics*, 22(22), 2823–2824.

Sibley, C.G., Ahlquist, J.E. (1984). The phylogeny of the hominoid primates, as indicated by dna-dna hybridization. *Journal of Molecular Evolution*, 20(1), 2–15.

Sibson, R. (1973). SLINK: an optimally efficient algorithm for the single-link cluster method. *The Computer Journal*, 16(1), 30–34.

Simonsen, M., Mailund, T., Pedersen, C.N. (2010). Inference of large phylogenies using neighbour-joining. Dans *International Joint Conference on Biomedical Engineering Systems and Technologies*. Springer, 334–344.

Sims, G.E., Kim, S.-H. (2011). Whole-genome phylogeny of *Escherichia coli/Shigella* group by feature frequency profiles (ffps). *Proceedings of the National Academy of Sciences*, 108(20), 8329–8334.

Sims, G.E., Jun, S.-R., Wu, G.A., Kim, S.-H. (2009). Alignment-free genome comparison with feature frequency profiles (ffp) and optimal resolutions. *Proceedings of the National Academy of Sciences*, 106(8), 2677–2682.

Sneath, P.H. (1957). The application of computers to taxonomy. *Microbiology*, 17(1), 201–226.

Sokal, R.R., Michener, C.D. (1958). A statistical method for evaluating systematic relationships. *University of Kansas Science Bulletin*, 38, 1409–1438.

Sokal, R.R., Sneath, P.H. (1963). *Principles of Numerical Taxonomy*. W. H. Freeman and Company.

Studier, J.A., Keppler, K.J. (1988). A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5(6), 729–731.

Swofford, D.L. (2003). *'PAUP*: phylogenetic analysis using parsimony (* and other methods)'. Version 4*. Sinauer Associates, Sunderland.

Swofford, D.L., Olsen, G.J., Waddell, P.J., Hillis, D.M. (1990). Phylogeny reconstruction. Dans *Molecular Systematics*, Hillis, D., Moritz, C., Mable, B. (dir.). Sinauer Associates, Sunderland, 411–501.

Ulitsky, I., Burstein, D., Tuller, T., Chor, B. (2006). The average common substring approach to phylogenomic reconstruction. *Journal of Computational Biology*, 13(2), 336–350.

Vach, W. (1989). Least squares approximation of additive trees. Dans *Conceptual and Numerical Analysis of Data*, Opitz, O. (dir.). Springer-Verlag, Berlin, 230–238.

Vachaspati, P., Warnow, T. (2015). Astrid: accurate species trees from internode distances. *BMC Genomics*, 16(S10), S3.

Vinga, S. (2014). Information theory applications for biological sequence analysis. *Briefings in Bioinformatics*, 15(3), 376–389.

Vinh, L.S., von Haeseler, A. (2005). Shortest triplet clustering: reconstructing large phylogenies using representative sets. *BMC Bioinformatics*, 6(1), 92.

Wheeler, T.J. (2009). Large-scale neighbor-joining with NINJA. Dans *International Workshop on Algorithms in Bioinformatics*. Springer, 375–389.

Yang, Z. (2006). *Computational Molecular Evolution*. Oxford University Press.

Zaretskii, K. (1965). Constructing a tree on the basis of a set of distances between the hanging vertices. *Uspekhi Matematicheskikh Nauk*, 20(6), 90–92.

Zaslavsky, L., Tatusova, T.A. (2008). Accelerating the neighbor-joining algorithm using the adaptive bucket data structure. *International Symposium on Bioinformatics Research and Applications*. Springer, 122–133.

Zielezinski, A., Vinga, S., Almeida, J., Karlowski, W.M. (2017). Alignment-free sequence comparison: benefits, applications, and tools. *Genome biology*, 18(1), 186.

Zielezinski, A., Girgis, H.Z., Bernard, G., Leimeister, C.-A., Tang, K., Dencker, T., Lau, A.K., Röhling, S., Choi, J.J., Waterman, M.S. *et al.* (2019). Benchmarking of alignment-free sequence comparison methods. *Genome biology*, 20(1), 144.