Une nouvelle méthode efficace pour l'estimation des données manquantes dans les séquences des nucléotides avant la reconstruction phylogénétique

Abdoulaye Baniré Diallo[1], Vladimir Makarenkov[1], Alix Boc[1] et François-Joseph Lapointe[2]

1. Département d'informatique, Université du Québec à Montréal, C.P. 8888, Succ. Centre-Ville, Montréal (Québec), Canada, H3C 3P8

2. Département de sciences biologiques, Université de Montréal, C.P. 6128, Succ. Centre-ville, Montréal (Québec), H3C 3J7, Canada

Courriels : diallo.abdoulaye_banire@courrier.uqam.ca, makarenkov.vladimir@uqam.ca, boc.alix@courrier.uqam.ca et Francois-Joseph.Lapointe@UMontreal.CA

Résumé

Dans cet article nous abordons le problème de reconstruction d'arbres phylogénétiques à partir de séquences de nucléotides contenant des données manquantes. Nous introduisons une nouvelle façon de considérer des bases manquantes dans les séquences alignées avant de calculer les distances d'évolution. La nouvelle méthode d'estimation des valeurs manquantes dans les séquences d'ADN ou d'ARN permet d'améliorer la qualité d'inférence phylogénétique par rapport aux méthodes existantes "*ignore missing sites*" (*IMS*) et "*proportional distribution of missing and ambiguous bases*" (*PDMAB*) incluses dans le logiciel PAUP (Swofford, 2001). La technique décrite est basée sur des formules probabilistes applicables dans le cadre du modèle de Jukes-Cantor (Jukes et Cantor, 1969). Les performances de la nouvelle méthode s'améliorent avec la diminution du nombre des bases dans les séquences et l'augmentation du pourcentage des valeurs manquantes.

Introduction

La présence des données manquantes et ambiguës dans les séquences d'ADN représente l'un des plus grands obstacles lors de la reconstruction phylogénétique. Les données manquantes peuvent être dues à la difficulté de séquencer certaines régions du génome d'un spécimen donné ou a une mauvaise conservation des spécimens. Les bases dans la séquence d'une espèce qui sont incertaines ou qui ne peuvent pas être déterminées sont considérées comme manquantes. Huelsenbeck (1991) a indiqué que les taxons qui contiennent beaucoup de caractères inconnus diminuent considérablement la qualité de l'inférence phylogénétique. L'une des questions qui revient souvent est la suivante : faut-il inclure ou ignorer les taxons avec des données manquantes dans une analyse phylogénétique, de même que, est-il nécessaire de considérer les sites avec des caractères manquants? Dans cette étude nous nous intéressons surtout à la seconde question qui consiste en la nécessité d'utiliser des sites avec des données manquantes dans la reconstruction phylogénétique. Le logiciel PAUP de Swofford (2001) propose deux méthodes de calcul des distances évolutives entre les espèces à partir des données de séquences incomplètes. La première méthode, appelée *IMS* (« *Ignoring missing sites* »), incluse dans PAUP préconise l'élimination des sites incomplets lors du calcul de la matrice de distances. Selon Wiens (2003), une telle approche représente une solution viable seulement lorsque les séquences considérées sont longues. La seconde méthode de PAUP, appelée *PDMAB* (« *proportionnal distribution of missing and ambiguous bases* »), consiste à considérer les sites incomplets en estimant les valeurs des données manquantes.

Dans cet article, nous proposons une nouvelle méthode, appelée PEMV (*probabilistic estimation of missing values*), qui vise à réestimer de données manquantes avant le calcul des distances d'évolution. Contrairement à PDMAB, notre méthode considère un facteur de rapprochement entre tous les taxons en tenant compte de la similarité entre l'espèce comportant la base manquante et les autres. La nouvelle méthode utilise une approche probabiliste pour déterminer la probabilité de la base manquante d'être soit A, C, G, ou T pour les séquences d'ADN, ou A, C, G, ou U pour les séquences d'ARN. Elle s'applique dans le cadre du modèle d'évolution de Jukes-Cantor (1969) et peut être généralisée pour toute autre transformation séquences-distances.

Dans la section suivante nous introduirons la nouvelle méthode d'estimation des valeurs manquantes dans les données séquentielles et la comparons aux méthodes disponibles dans le logiciel PAUP. Par la suite, nous discuterons des résultats de simulations statistiques permettant de mesurer les performances des trois méthodes en fonction du pourcentage de valeurs manquantes dans les séquences et de la longueur des séquences générées. La qualité d'inférence phylogénétique est estimée à l'aide de la distance topologique de Robinson et Foulds (1981). La métrique de Robinson et Foulds est égale au nombre minimum d'opérations élémentaires, de contractions et d'expansions de branches de l'arbre, nécessaires pour transformer un arbre phylogénétique en un autre. La méthode Neighbor Joining (NJ) de Saitou et Nei (1987) a été utilisée dans les simulations pour reconstruire les arbres phylogénétiques.

La nouvelle méthode d'estimation des nucléotides manquants PEMV

La nouvelle méthode d'estimations de données manquantes *PEMV* est présentée ici dans le cadre du modèle d'évolution de Jukes-Cantor (1969). Ce modèle assume que les nucléotides A, C, G et T sont équifréquents et que les substitutions sont équiprobables (i.e. la probabilité d'une transition est égale à celle d'une transversion). Pour calculer les distances de Jukes-Cantor, la formule de correction standard est utilisée : $D = -3/4 \ ln(1 - 4/3d)$, où d représente la distance observée, qui est la proportion des nucléotides différents dans les séquences considérées. La nouvelle méthode utilise les principes de base du modèle de Jukes-Cantor.

Supposons que la base k dans la séquence i est inconnu (voir Figure 1). Pour calculer la distance entre la séquence i et toutes les autres séquences dans la matrice **C**, *PEMV* estime, à l'aide de l'équation 1 ci-dessous, les probabilités $P_{ik}(A)$, $P_{ik}(C)$, $P_{ik}(G)$ et $P_{ik}(T)$ d'avoir respectivement le nucléotide A, C, G ou T à la position k de la séquence i.

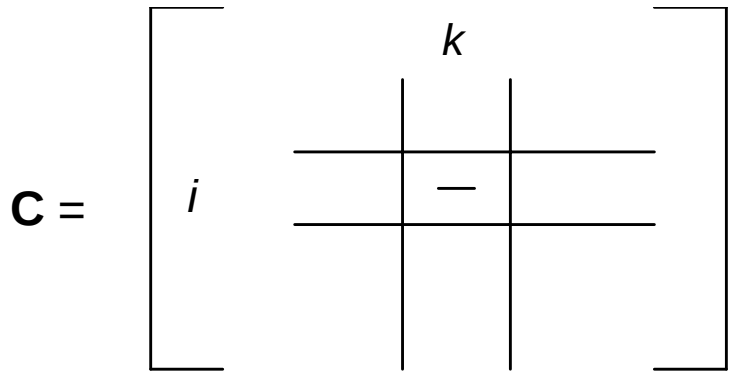

Figure 1: Le caractère k est manquant dans la séquence i; il est représenté par le symbole "−".

La probabilité qu'une base manquante corresponde à un nucléotide spécifique dépend du nombre de séquences ayant le même nucléotide à la position *k* de même que de la distance (en ignorant les sites manquants) entre *i* et toutes les autres séquences ayant des nucléotides connus sur le site *k*. Pour faire ceci, il faudra évaluer en premier lieu la distance δ contenant le score de parité entre toutes les séquences. Le score de parité δ est calculé en ignorant les données manquantes. Cette distance est calculée comme le rapport entre le nombre de paires de nucléotides distincts dans une paire de séquences et le nombre de sites comparables dans ces séquences.

$$P_{ik}(\mathrm{V}) = \frac{1}{N_k}\Big(\sum_{\substack{j,\ \text{tels que} \\ C_{jk}=\mathrm{V}}} \delta_{ij} \; + \; \frac{1}{3} \sum_{\substack{j,\ \text{tels que} \\ C_{jk}\neq\mathrm{V}}} (1-\delta_{ij}) \Big), \qquad (1)$$

où le caractère V remplace l'une des quatre nucléotides A, C, G ou T; N_k – est le nombre de valeurs existantes dans la colonne *k*; δ_{ij} – est le score de parité entre les séquences *i* et *j* calculé en ignorant les sites manquants; **C** – est la matrice des séquences des nucléotides alignées, P_{ik}(A), P_{ik}(C), P_{ik}(G) et P_{ik}(T) – est la probabilité d'avoir, respectivement, le nucléotide A, C, G ou T à la position *k* de la séquence *i*.

Le théorème caractérisant les probabilités P_{ik}(A), P_{ik}(C), P_{ik}(G) et P_{ik}(T) pour une séquence *i* et un site donné *k* peut être formulée de la façon suivante:

Théorème. *Pour toute séquence i, tout site k de la matrice* **C**, *tels que* c_{ik} *est un nucléotide manquant, nous avons* : $P_{ik}(A) + P_{ik}(C) + P_{ik}(G) + P_{ik}(T) = 1$.

Preuve : En remplaçant la somme P_{ik}(A) + P_{ik}(C) + P_{ik}(G) + P_{ik}(T) par les expressions équivalentes de l'équations 1, nous aurons à démonter que :

$$\frac{1}{N_k}\Big(\sum_{\substack{j,\ \text{tels que} \\ C_{jk}=\mathrm{A,C,G\ ou\ T}}} \delta_{ij} \; + \; \frac{1}{3} \sum_{\substack{j,\ \text{tels que} \\ C_{jk}\neq\mathrm{A}}} (1-\delta_{ij}) \; + \; \frac{1}{3} \sum_{\substack{j,\ \text{tels que} \\ C_{jk}\neq\mathrm{C}}} (1-\delta_{ij}) \; +$$

$$+ \; \frac{1}{3} \sum_{\substack{j,\ \text{tels que} \\ C_{jk}\neq\mathrm{G}}} (1-\delta_{ij}) \; + \; \frac{1}{3} \sum_{\substack{j,\ \text{tels que} \\ C_{jk}\neq\mathrm{T}}} (1-\delta_{ij}) \; = 1. \qquad (2)$$

La partie gauche de l'équation 2 est équivalente à l'expression suivante :

$$\frac{1}{N_k}\Big(\sum_{\substack{j,\text{ tels que}\\ C_{jk}=A}}\delta_{ij}+\sum_{\substack{j,\text{ tels que}\\ C_{jk}=C}}\delta_{ij}+\sum_{\substack{j,\text{ tels que}\\ C_{jk}=G}}\delta_{ij}+\sum_{\substack{j,\text{ tels que}\\ C_{jk}=T}}\delta_{ij}-$$

$$-\frac{1}{3}\Big(3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=A}}\delta_{ij}+3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=C}}\delta_{ij}+3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=G}}\delta_{ij}+3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=T}}\delta_{ij}\Big)+$$

$$+\frac{1}{3}\Big(3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=A}}1+3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=C}}1+3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=G}}1+3\sum_{\substack{j,\text{ tels que}\\ C_{jk}=T}}1\Big)\Big)=$$

$$=\frac{1}{N_k}\Big(\sum_{\substack{j,\text{ tels que}\\ C_{jk}=A}}1+\sum_{\substack{j,\text{ tels que}\\ C_{jk}=C}}1+\sum_{\substack{j,\text{ tels que}\\ C_{jk}=G}}1+\sum_{\substack{j,\text{ tels que}\\ C_{jk}=T}}1\Big)=\frac{N_k}{N_k}=1.$$

Ce qu'il fallait démontrer □.

Une fois les différentes probabilités P_{ik} calculées, nous calculons la matrice de distances **D** entre toutes les séquences en appliquant l'équation 3. La distance *PEMV* entre les séquences *i* et *j* se calcule donc comme suit :

$$d_{ij}=\frac{N_{ij}^{c}-N_{ij}^{m}+\sum_{k\in\Lambda_{ij}}(1-P_{ij}^{k})}{N}, \qquad (3)$$

où d_{ij} – est la distance observée entre les séquences *i* et *j* calculée en ignorant les sites manquants, N – est le nombre de sites (i.e. nombre de colonnes de la matrice **C**) ayant au moins un nucléotide présent, N_{ij}^{m} - est le nombre de paires de nucléotides identiques dans les séquences *i* et *j*, N_{ij}^{c} - est le nombre de paires de nucléotides comparables (i.e. quand les deux nucléotides sont présents dans les sites correspondants de *i* et *j*) dans les séquences *i* et *j*, Λ_{ij} est un ensemble de sites dans les séquences *i* et *j* tels que au moins un nucléotide est manquant dans leurs sites correspondants, et P_{ij}^{k} - la probabilité, calculée en utilisant l'équation 1, d'avoir une paire de nucléotides identiques au site *k* dans les séquences *i* et *j*. La distance de Jukes-Cantor s'obtient à partir de *d* par l'application de la formule logarithmique.

Un exemple d'application

Dans cette section, nous présentons un exemple d'application de notre méthode (*PEMV*) ainsi que des deux méthodes incluses dans PAUP (*IMS* et *PDMAB*) à une matrice **C** de 3 séquences à 8 nucléotides chacune :

$$\mathbf{C} = \begin{array}{r} 1\,2\,3\,4\,5\,6\,7\,8 \\ \text{Séquence 1 : ACGGTAAG} \\ \text{Séquence 2 : ACGTAAAA} \\ \text{Séquence 3 : ACGT – AGC} \end{array}$$

Le caractère '–' représente une base manquante dans la séquence 3. La distance entre les paires de séquences complètes (sans bases manquantes) est évidemment la même pour les 3 méthodes. Par exemple, en considérant la matrice **C**, nous remarquons qu'entre la séquence 1 et la séquence 2 il y a trois différences entre les paires de bases aux sites 4, 5 et 8. La distance entre les deux séquences est donc 3/8 ou 0.375. Nous allons maintenant appliquer les 3 méthodes de transformation séquences-distances pour calculer les valeurs de d_{13} et d_{23} qui représentent la distance entre les séquences 1 et 3 et 2 et 3, respectivement. Voici comment le calcul se ferait pour les 3 méthodes *IMS*, *PDMAB* et *PEMV* :

1) En utilisant la méthode *IMS*, qui ignore les sites incomplets pour trouver d_{23}, nous ne considérons que les 7 sites présents dans les deux séquences. Il y a deux différences entre elles, aux sites 7 et 8. Ce qui fait que la distance entre les séquences 2 et 3 sera 2/7. De la même manière, nous obtenons la valeur de d_{13} qui est de 3/7.

2) En utilisant la méthode *PDMAB* pour calculer d_{23}, les bases manquantes seront évaluées en fonction des changements non ambigus entre les 2 séquences (voir le manuel de PAUP pour plus de détails). En considérant les séquences 2 et 3, nous regardons toutes les paires de bases correspondantes telles que le nucléotide de la séquence 2 soit un A (car le nucléotide manquant doit être comparée à un A). Quatre paires de bases, A-A, A-A, A-G, A-C, sont donc à considérer. La matrice de comparaisons entre les séquences 2 et 3 est la suivante :

	A	C	G	T
A	2.5	1.25	1.25	0.0
C		1.0	0.0	0.0
G			1.0	0.0
T				1.0

Table 1: La matrice de comparaisons entre les séquences 2 et 3 utilisée dans la méthode *PDMAB*.

La distance est par la suite calculée en faisant la sommation sur toutes les cases en excluant la diagonale. Par conséquent, la distance entre les séquences 2 et 3 est égale à 2.5/8. Suivant le même principe, nous obtenons la valeur de 3/7 pour la distance entre les séquences 1 et 3.

3) En utilisant la nouvelle méthode, *PEMV*, nous déterminons tout d'abord les probabilités P_{ik} que la base manquante soit un A, C, G ou T. Le calcul nécessite de connaître le score de parité δ, qui ne considère que les sites complets (7 dans ce cas). On obtient donc $\delta_{13} = 4/7$ et $\delta_{23} = 5/7$. Par la suite nous déterminons le nombre de bases présentes au site contenant la base manquante. Ce nombre est deux dans notre exemple (pour le site 5). Nous calculons par la suite P_{35}(A), P_{35}(C), P_{35}(G), P_{35}(T) selon l'équation 1. P_{35} représente la probabilité que la base manquante soit un nucléotide indiqué entre parenthèses dans la séquence 3 et à la position 5. Dans notre exemple on aura :

$P_{35}(A) = 1/2 * (1 – 4/7) * 1/3 + 1/2 * 5/7 = 18/42,$
$P_{35}(C) = 1/2 * (1 – 4/7) * 1/3 + 1/2 * (1 – 5/7) * 1/3 = 5/42,$
$P_{35}(G) = 1/2 * (1 – 4/7) * 1/3 + 1/2 * (1 – 5/7) * 1/3 = 5 /42,$
$P_{35}(T) = 1/2 * 4/7 + 1/2 * (1\text{-}5/7) * 1/3 = 14/42.$

Une fois les probabilités connues, nous calculons la distance d_{23} selon l'équation 3. La distance entre les séquences 2 et 3 se calcule donc comme suit : $d_{23} = (2 + 1 – P_{35}(A))/8 = 0.3214$. De la même

manière, nous calculons la distance d_{13} qui est égale à 0.4583. Le tableau ci-dessous présente les trois matrices de distance obtenues avec les méthodes *IMS*, *PDMAB* et *PEMV* :

IMS			
Séq	1	2	3
1	0	0.375	0.42857
2		0	0.28571
3			0

PDMAB			
Séq	1	2	3
1	0	0.375	0.42857
2		0	0.3125
3			0

PEMV			
Séq	1	2	3
1	0	0.375	0.4583
2		0	0.3214
3			0

Table 2: Les matrices de distances obtenues par les méthodes *IMS*, *PDMAB* et *PEMV*.

Simulations Monte Carlo

Une étude Monte Carlo a été menée pour tester la performance de la nouvelle méthode pour l'inférence phylogénétique. Nous examinerons ici comment la méthode *PEMV* réagit en fonction de la longueur des séquences de nucléotides et du pourcentage des bases manquantes dans ces séquences. Tous les résultats obtenus ci-dessous proviennent des simulations faites avec 1000 arbres phylogénétiques aléatoires à 8, 16, 24 et 32 feuilles. Dans chaque cas, une vraie phylogénie, notée *T*, a été obtenue en utilisant la procédure de génération aléatoire de phylogénies décrite par Kuhner et Felsenstein (1994). Les longueurs des branches de l'arbre sont obtenues à partir d'une distribution exponentielle. En suivant l'approche de Guindon et Gascuel (2002), nous avons ajouté du bruit aux longueurs des branches sous la forme de déviation de l'hypothèse d'horloge moléculaire. Toutes les longueurs de branche de *T* sont multipliées par $1+ax$, où la variable x est obtenue aléatoirement d'une distribution exponentielle standard ($P(x>k) = \exp(-k)$). La constante a est un facteur de réglage qui ajuste l'intensité de la déviation. Comme dans Guindon et Gascuel (2002), a a été initialisée à 0.8. Les arbres générés par cette procédure sont supposés être de profondeur de $O(\log(n))$, où n est leur nombre des feuilles. Le code source de notre programme de génération d'arbre (écrit en C) est disponible à l'adresse suivante : http://www.info.uqam.ca/~makarenv/tree_generation.cpp.

Chaque arbre aléatoire obtenu a été par la suite soumis au logiciel SeqGen de Rambaut et Grassly (1996). Ce logiciel permet de simuler l'évolution de séquences de nucléotides le long d'une ou de plusieurs phylogénies, en utilisant plusieurs modèles de substitution des nucléotides. Nous avons utilisé SeqGen pour obtenir un alignement de séquences de longueur l (l = 125, 250, 500 et 1000 nucléotides) pour le modèle de substitution de Jukes et Cantor (1969). Une fois les séquences alignées obtenues, nous y avons introduit aléatoirement des bases manquantes. De 0 à 50% des nucléotides ont été retirés des séquences dans nos simulations. Les séquences incomplètes obtenues, ont été soumises aux trois méthodes de calcul des matrices de distances d'évolution présentées ci-dessus. Pour chaque matrice de distances ainsi obtenue, nous avons reconstruit une phylogénie *T'* en utilisant la méthode NJ de Saitou et Nei (1987). L'arbre phylogénique *T'* a été ensuite comparé à l'arbre initial *T* à l'aide de la distance de Robinson et Foulds (1981). Dans le cadre de notre simulation, nous avons évalué la performance des méthodes de calcul des distances *IMS*, *PDMAB* et *PEMV* et vérifié l'influence du nombre de taxons, de la taille des séquences et du pourcentage de données manquantes sur la qualité de l'inférence phylogénétique.

Résultats des simulations

Les résultats que nous présentons dans cette section comparent les trois méthodes pour un nombre de taxons égal à 8, 16, 24 et 32. Pour chaque situation, nous testons l'influence de la longueur de la séquence (pour des longueurs de 125, 250, 500 et 1000 nucléotides) et du pourcentage des données

manquantes (qui varient de 0 à 50%) sur la performance des 3 méthodes. Les performances des méthodes sont comparées à l'aide de la distance topologique de Robinson et Foulds (1981). Elle est mesurée en pourcentage de sa valeur maximale qui est 2*n*-6 pour un arbre phylogénétique à *n* feuilles. Plus ce pourcentage est petit, plus l'arbre *T'* obtenu se rapproche de l'arbre initial (i.e. meilleure est l'estimation des données manquantes).

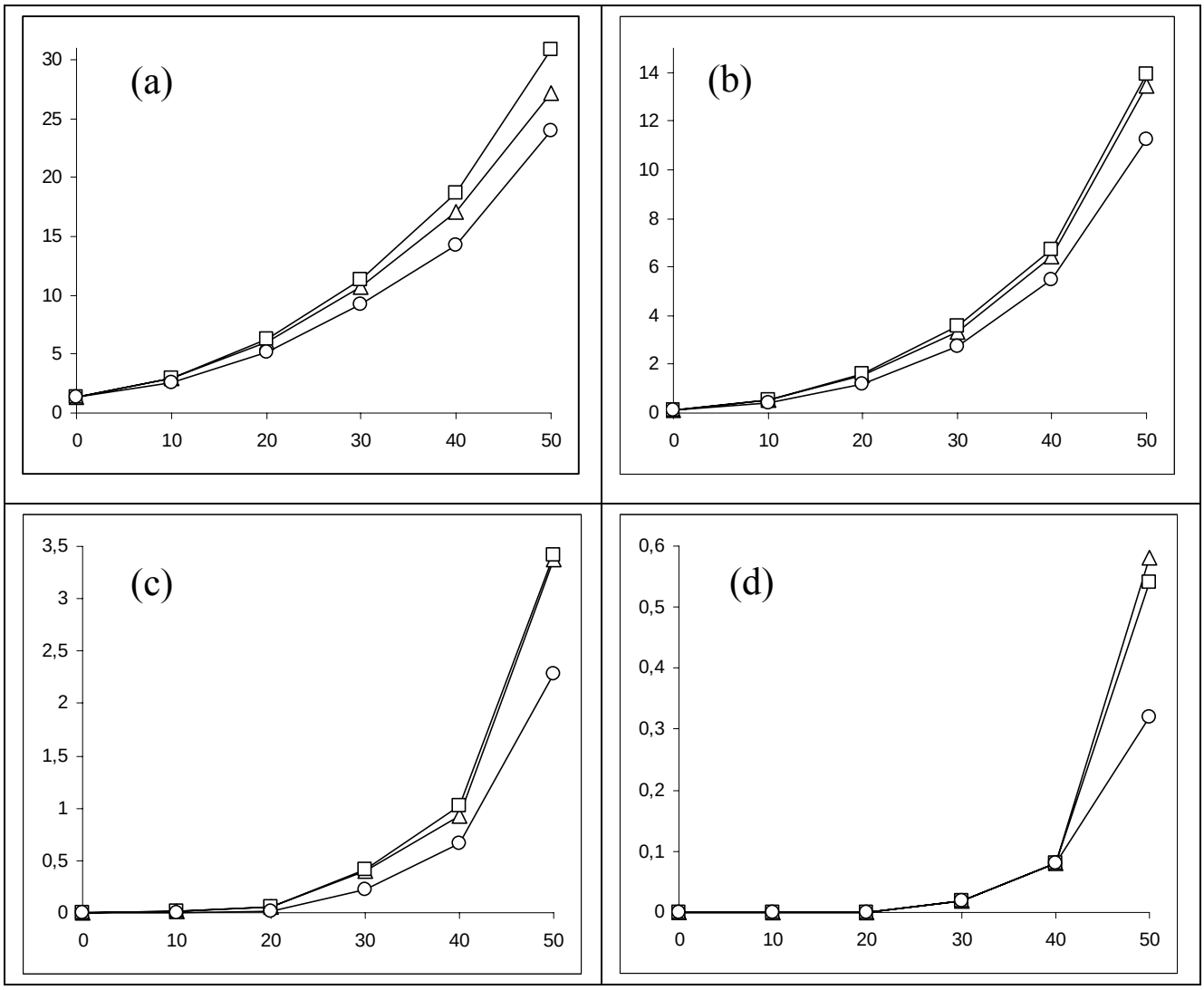

Figure 2: Les résultats moyens obtenus pour 1000 arbres phylogénétiques aléatoires à 8 feuilles. Le pourcentage des données manquantes varie de 0 à 50% (l'axe des abscisses). Les courbes traduisent les variations de la distance de Robinson et Foulds pour les méthodes *IMS* (Δ), *PDMAB* (□) et *PEMV* (○). L'influence des tailles des séquences est représentée sur les quatre figures : (a) pour les séquences à 125 caractères, (b) 250 caractères, (c) 500 caractères et (d) 1000 caractères.

En observant les figures 2, 3 et 4 illustrant les résultats des simulations pour 8, 16, 24 espèces, nous remarquons que *PEMV* donne de meilleurs résultats pour tous les pourcentages de données manquantes et pour toutes les longueurs de séquences. Les résultats de la méthode *IMS* sont très similaires à ceux de *PDMAB* lorsque la longueur des séquences est différente de 125 caractères. La figure 5 illustrant les résultats pour les arbres à 32 feuilles montre que pour des longueurs des séquences inférieures à 500 la nouvelle méthode est plus performante que *IMS* et *PDMAB*. Au delà de cette taille, les trois méthodes donnent des résultats très similaires.

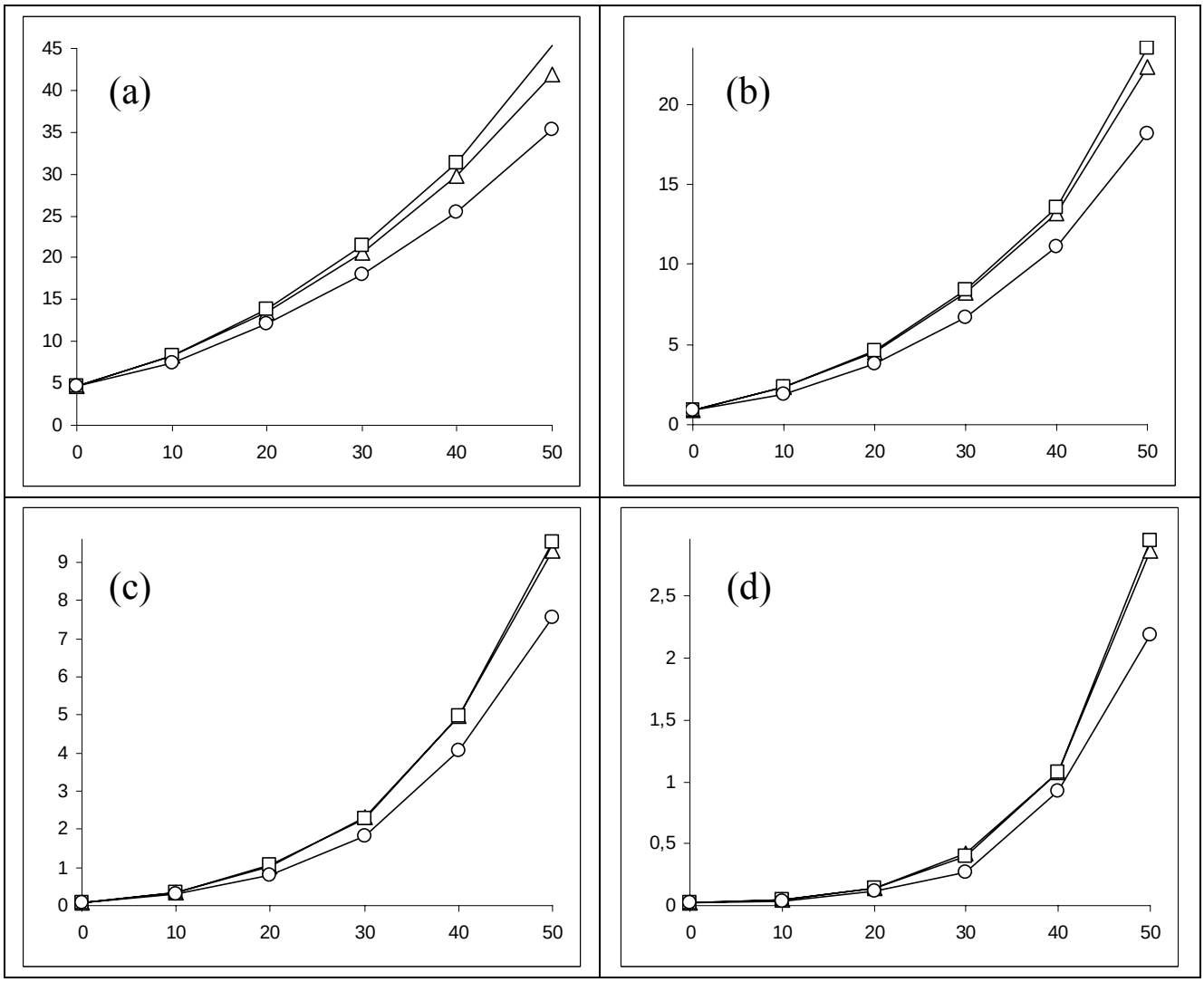

Figure 3: Les résultats moyens obtenus pour 1000 arbres phylogénétiques aléatoires à 16 feuilles. Le pourcentage des données manquantes varie de 0 à 50% (l'axe des abscisses). Les courbes traduisent les variations de la distance de Robinson et Foulds pour les méthodes *IMS* (Δ), *PDMAB* (□) et *PEMV* (○). L'influence des tailles des séquences est représentée sur les quatre figures : (a) pour les séquences à 125 caractères, (b) 250 caractères, (c) 500 caractères et (d) 1000 caractères.

Ceci est certainement dû à la présence d'un nombre de caractères suffisants permettant de reconstruire une phylogénie correcte. Dans le dernier cas, il serait possible d'appliquer seulement IMS qui est la méthode la plus simple et rapide des trois méthodes comparées.

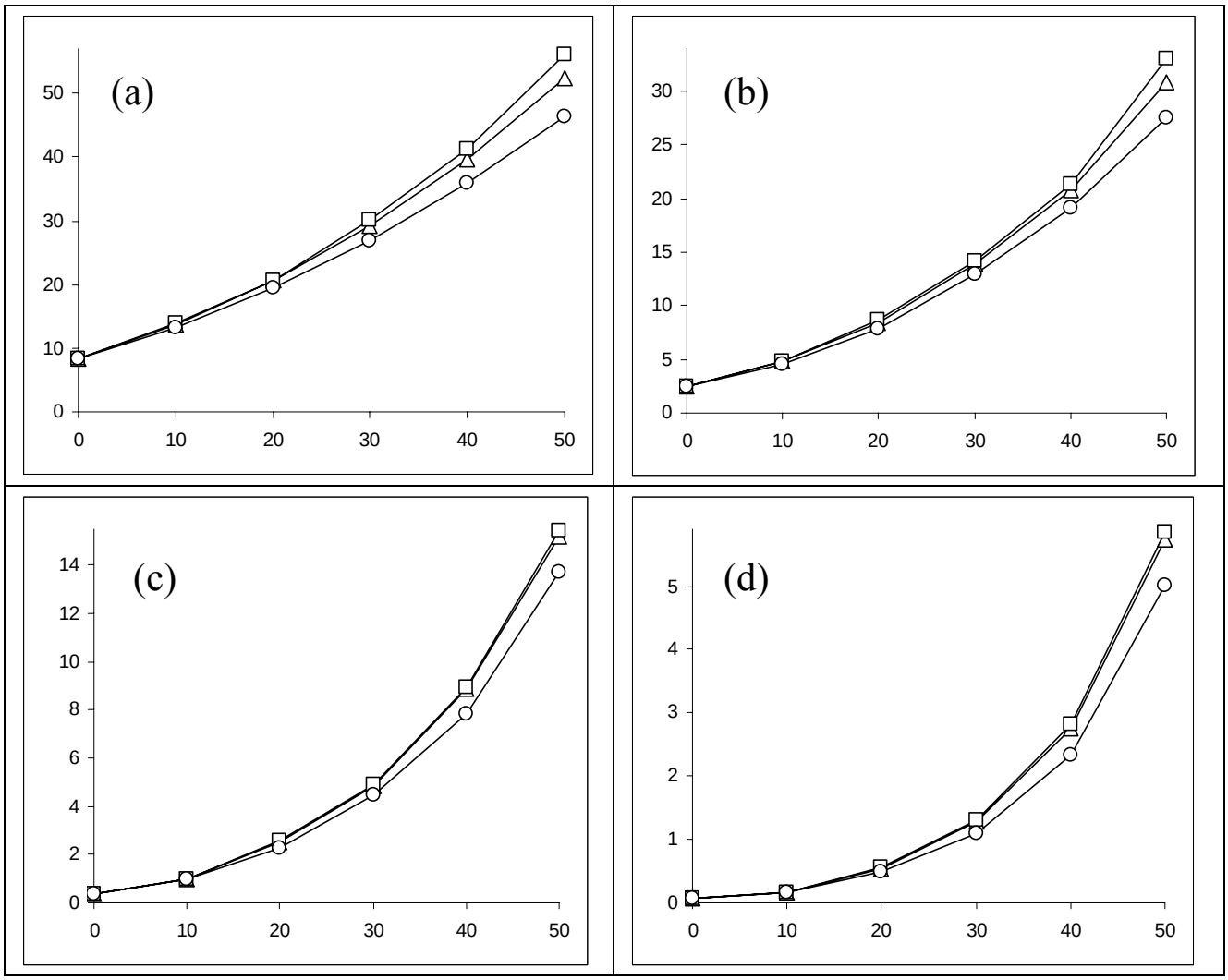

Figure 4: Les résultats moyens obtenus pour 1000 arbres phylogénétiques aléatoires à 24 feuilles. Le pourcentage des données manquantes varie de 0 à 50% (l'axe des abscisses). Les courbes traduisent les variations de la distance de Robinson et Foulds pour les méthodes *IMS* (Δ), *PDMAB* (□) et *PEMV* (○). L'influence des tailles des séquences est représentée sur les quatre figures : (a) pour les séquences à 125 caractères, (b) 250 caractères, (c) 500 caractères et (d) 1000 caractères.

La distance de Robinson et Foulds croît lorsqu'on augmente le nombre de taxons dans toutes les méthodes. Elle diminue lorsque la longueur des séquences augmente. La dernière tendance se maintient même avec l'augmentation du pourcentage des bases manquantes (la tendance similaire était observée par Wiens, 1998). Remarquons que la distance de Robinson et Foulds pour 0% de bases manquantes n'est pas toujours nulle (surtout pour les séquences à 125 et 250 bases). Ce fait est dû à des artefacts de la procédure de génération des séquences, SeqGen, et à ceux de la méthode de reconstruction d'arbre, NJ, utilisées dans nos simulations. Comme ce n'était pas l'objectif de cet article, ici nous n'avons pas investigué laquelle de ces deux procédures est responsable de cette déviation des résultats.

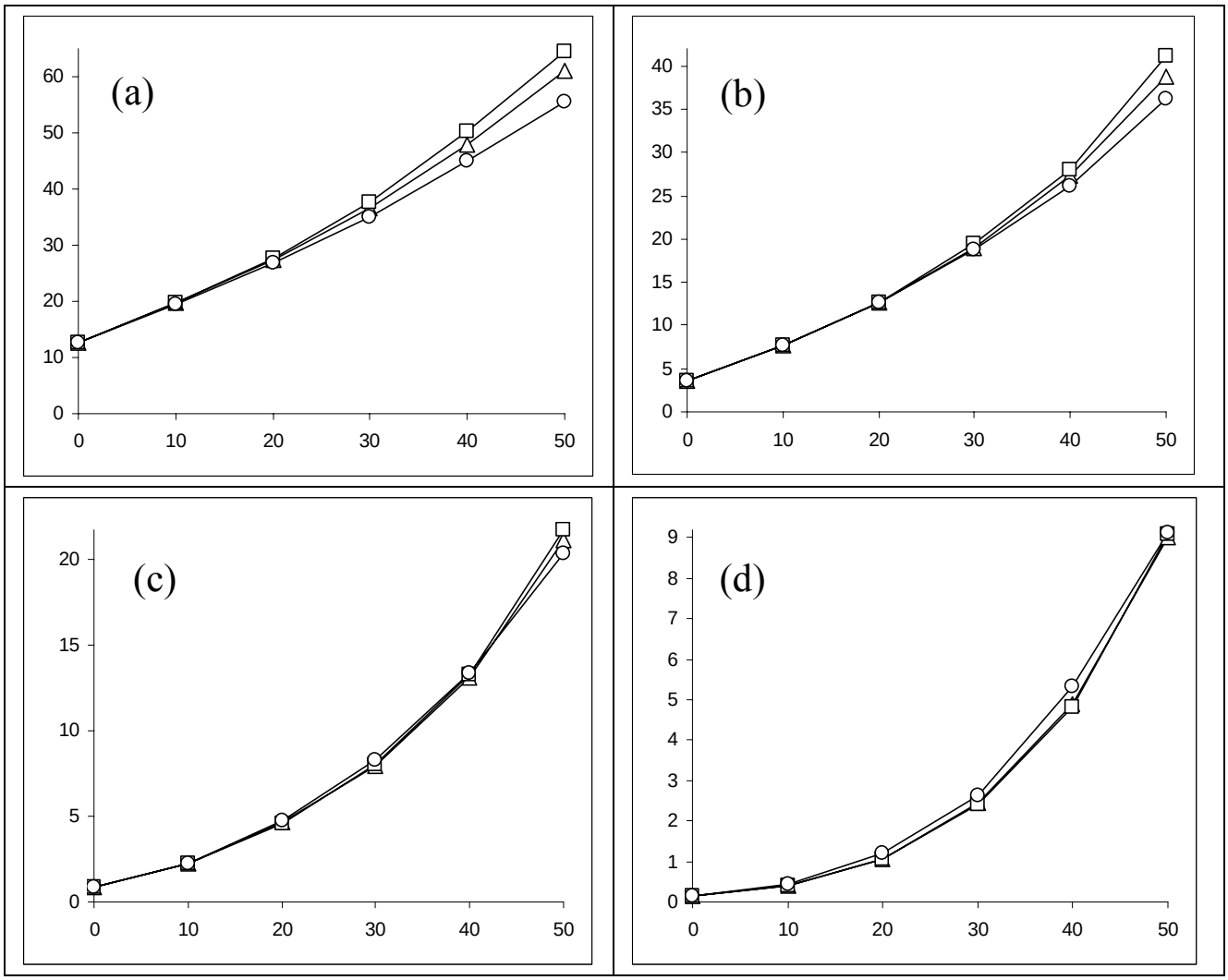

Figure 5: Les résultats moyens obtenus pour 1000 arbres phylogénétiques aléatoires à 32 feuilles. Le pourcentage des données manquantes varie de 0 à 50% (l'axe des abscisses). Les courbes traduisent les variations de la distance de Robinson et Foulds pour les méthodes *IMS* (Δ), *PDMAB* (□) et *PEMV* (○). L'influence des tailles des séquences est représentée sur les quatre figures : (a) pour les séquences à 125 caractères, (b) 250 caractères, (c) 500 caractères et (d) 1000 caractères.

Conclusion

À l'aide des simulations présentées dans la section précédente, nous avons montré l'utilité de la méthode *PEMV* dans la réestimation des données manquantes avant l'inférence d'une phylogénie. La nouvelle méthode fournit de très bons résultats pour les courtes séquences (jusqu'à 250 bases). Les performances de la nouvelle méthode sont les plus importantes avec des pourcentages élevés des nucléotides manquants. Dans ces situations, l'élimination des sites manquants, la méthode *IMS*, ou leur estimation selon la méthode *PDMAB* (les deux modèles disponibles dans PAUP) supprimerait des spécificités importantes des données traitées. Pour les séquences plus longues et pour un nombre élevé de taxons, les trois méthodes examinées dans notre étude donnent des résultats semblables. La méthode PEMV a été présenté ici dans le cadre du modèle de Jukes-Cantor. Il serait intéressant de continuer le développement de cette approche probabiliste en la généralisant pour les modèles d'évolution plus réaliste : Kimura – 2 paramètres (1980), Tajima – Nei (1984) et d'autres. Il serait

également important de comparer les résultats obtenus par Neighbor Joining avec d'autres méthodes de distances, comme par exemple BioNJ de Gascuel (1997), FITCH de Felsenstein (1997) et MW de Makarenkov et Leclerc (1999). Une autre piste d'investigation intéressante serait le test de cette méthode sur des données des séquences des protéines.

Références

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

Gascuel, O. 1997. An improved version of NJ algorithm based on a simple model of sequence Data. *Molecular Biology and Evolution*. 14, 101-111.

Guindon, S., et Gascuel, O. 2002. Efficient biased estimation of evolutionary distances when substitution rates vary across sites. *Molecular Biology and Evolution* 19, 534-543.

Huelsenbeck, J. P. 1991. When are fossils better than existent taxa in phylogenetic analysis? *Sys. Zool*. 40:458-469.

Jukes, T. H. et Cantor, C., 1969. *Mammalian Protein Metabolism*, chapter Evolution of protein molecules, pages 21-132. Academic Press, New York, 1969. T. H. Jukes and C. Cantor. *Mammalian Protein Metabolism*, Academic Press, New York, chapter Evolution of protein molecules, pp. 21-132.

Kimura, M. 1980. A simple method for estimating evolutionary rate of base substitutions through comparative studies of nucleotide sequences. *Journal of Molecular Evolution* 16, 111-120.

Kuhner, M. K. et J. Felsenstein. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution,* 11, 459-68.

Makarenkov, V. et Leclerc, B. 1999. An algorithm for the fitting of a phylogenetic tree according to a weighted least-squares criterion, *Journal of Classification*, 16, 3-26.

Rambaut, A. et Grassly, N. C. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Bioinformatics*, 13, 235-238.

Robinson D. R. et Foulds L. R. 1981. Comparison of phylogenetic trees, *Mathematical Biosciences,* 53, 131-147.

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

Swofford, D. L. 2001. *PAUP*.* Phylogenetic Analysis Using Parsimony (*and Other Methods). Version 4. *Sinauer Associates, Sunderland, Massachusetts*.

Tajima F, Nei M. 1984. Estimation of evolutionary distance between nucleotide sequences. *Mol Biol Evol*. 3, 269-85.

Wiens J.J. 2003. Does adding Characters with missing data increase or decrease phylogenetic accuracy. *Systematic Biology* 47, 625-640.

Wiens J. J. 1998. Missing Data, Incomplete Taxons, and phylogenetic accuracy. *Systematic Biology* 52(4):528-538.