Université de Montréal

Chimères, données manquantes et congruence: validation de différentes méthodes par simulations et application à la phylogénie des mammifères

par
Véronique Campbell

Département de sciences biologiques
Faculté des arts et sciences

Thèse présentée à la Faculté des études supérieures
en vue de l'obtention du grade de Philosophiae doctor (Ph. D.)
en sciences biologiques

Août, 2009

Université de Montréal
Faculté des études supérieures

Cette thèse intitulée :

Chimères, données manquantes et congruence: validation de différentes méthodes par simulations et application à la phylogénie des mammifères

présentée par :
Véronique Campbell

a été évaluée par un jury composé des personnes suivantes :

M. Mohamed Hijri, président-rapporteur
M. François-Joseph Lapointe, directeur de recherche
Mme. Anne Bruneau, membre du jury
M. Pierre Darlu, examinateur externe
Mme. Ariane Burke, représentante du doyen de la FESP

RÉSUMÉ

Avec l'essor de la phylogénomique, le défi n'est plus de générer suffisamment de données moléculaires nécessaires à l'inférence phylogénétique, mais plutôt de développer des moyens adaptés au traitement de ces données. Entre autres, de nouvelles méthodes phylogénétiques qui permettent de réduire le temps de calcul doivent être développées et validées. De plus, de nouveaux outils sont indispensables pour déterminer statistiquement la meilleure façon de combiner les données tout en réduisant les données manquantes propres à certains taxons. Dans cette perspective, ma thèse a pour but d'explorer des alternatives qui pourraient se révéler efficaces pour faciliter les analyses phylogénomiques.

En premier lieu, il est impératif de déterminer si toutes les données peuvent être combinées dans une analyse unique. Dans l'éventualité où les jeux de données soutiennent des phylogénies significativement similaires (c.-à-d. la congruence des données), il peut être préférable de les réunir dans une seule analyse, pour permettre d'augmenter le signal phylogénétique global. Dans le cas contraire, c'est-à-dire face à des données incongruentes, il est plus judicieux d'analyser les jeux de données séparément pour éviter d'inclure des signaux phylogénétiques contradictoires. Dans le cadre du premier volet de ma thèse, des simulations ont été réalisées afin de tester la validité du test de Congruence Entre Matrices de Distance. Ce test mesure le degré de congruence entre des matrices qui peuvent provenir de nombreuses sources. Les résultats montrent que le test a une erreur de type I adéquate et une bonne puissance lorsqu'il est appliqué à des matrices de distance ultramétrique et additive.

Lorsqu'il est recommandé de combiner différents jeux de données dans une seule analyse, l'utilisation de taxons chimères peut être une option intéressante. Cette approche réduit le nombre de données manquantes en combinant les séquences d'ADN de différentes espèces qui appartiennent à un même groupe taxonomique pour ne former qu'un seul taxon représentatif. Le deuxième volet de ma thèse évalue la validité de cette méthode à l'aide de simulations et de données empiriques moléculaires provenant d'espèces de mammifères. La comparaison de l'exactitude phylogénétique révèle une performance égale des deux approches et, même, une supériorité pour les taxons chimères, dans certaines conditions. Étant donné la réduction non négligeable

en temps de calcul et en puissance informatique liée à l'utilisation de taxons chimères, cette dernière méthode s'avère être idéale pour les analyses phylogénomiques.

Dans certains cas, il peut sembler plus approprié d'analyser différents jeux de données séparément au lieu de les inclure dans une même analyse. Pour faciliter une vue d'ensemble, les arbres phylogénétiques obtenus peuvent être combinés en utilisant une méthode de consensus ou encore une méthode de super-arbre. La congruence globale soutient que la concordance des résultats obtenus suite à des analyses séparées et combinées permet d'augmenter la fiabilité des inférences phylogénétiques. Le troisième volet compare donc différentes méthodes de reconstruction phylogénétique de types consensus et super-arbre avec l'approche de super-matrice lorsque appliqué à des séquences mitogénomiques échantillonnées parmi 93 espèces de mammifères. Les super-arbres obtenus étaient congruents à l'arbre inféré à partir de la supermatrice.

S'ouvrant sur des pistes de recherche future, mon étude constitue un apport au domaine de la phylogénomique en validant certaines approches qui se sont révélés être des outils efficaces dans un contexte de méta-analyses.

Mots-clés : CEMD, consensus, erreur de type I, incongruence, exactitude, mitogénomique, phylogénomique, puissance, super-arbre, super-matrice

ABSTRACT

With the rise of phylogenomics, the challenge is no longer to generate sufficient molecular data for phylogenetic inference, but rather to develop appropriate means to process the enormous amount of data produced. New phylogenetic algorithms with better computation time and complexity have to be developed and validated. In addition, new tools are needed to statistically determine the best approach to combine datasets while reducing the amount of missing data. Some aspects of these problems motivated my thesis, which explore alternatives that could facilitate phylogenomic analyses, while contributing to improving knowledge about evolutionary relationships among mammal families.

It is imperative to determine if all data can be combined in a single analysis. In the event that different datasets support significantly similar phylogenies (i.e., congruent phylogenetic trees), it may be preferable to include all datasets in a single analysis, in order to increase the overall phylogenetic signal. However, when faced with incongruent datasets, it is generally recommended to analyze each dataset separately to avoid conflicting signals. In the first part of my thesis, simulations were conducted to verify the validity of the Congruence Among Distance Matrices (CADM) test. This test measures the degree of congruence among distance matrices from different sources. The results showed that the test has an adequate Type I error and good power when applied to ultrametric and additive distance matrices.

When it is appropriate to combine different datasets in a single analysis, the use of composite taxa may represent a valuable option. This approach reduces the number of missing entries in a supermatrix by combining the DNA sequences of different species belonging to the same taxonomic group. Therefore, a single composite sequence is defined by more than one species (or taxon). Following criticisms about the use of composite taxa, the second part of my thesis assessed the validity of the composite approach through simulations and empirical data from mammal species. Comparisons of phylogenetic accuracy of trees inferred from incomplete and composite matrices revealed an equal performance of both approaches, and even superiority of composite taxa under certain conditions. Given high phylogenetic accuracy and significant

reduction in computing time when analyzing composite matrices (due to reduced taxon number), this approach will probably be increasingly used in phylogenomic analyses. In some cases, it may seem more appropriate to analyze datasets separately rather than in a single analysis (separate vs. combined analysis). The phylogenetic trees inferred from each dataset can then be combined using a consensus or supertree method. A global congruence framework suggests that increased phylogenetic accuracy can be obtained if the results from the separate and combined analyses are congruent. In this perspective, the third part of my thesis compared different consensus and supertree methods to a supermatrix analysis of mitogenomic sequences representing 93 mammal families. Supertrees congruent to the tree inferred from the supermatrix were obtained.

The results presented in my thesis provide an important contribution towards the resolution of different methodological debates through the validation of different phylogenetic approaches that will facilitate phylogenomic analyses.

Keywords: accuracy, CADM, composite taxon, consensus, incongruence, mitogenomic, phylogenomic, type I error, supermatrix, supertree

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

LISTE DES SIGLES ET ABRÉVIATIONS

A	adenine
AC	average consensus
ADN	acide désoxyribonucléique
AIC	Akaike information criterion
ARN	acide ribonucléique
ATP6	ATP synthase F0 subunit 6
ATP8	ATP synthase F0 subunit 8
BB	model tree backbone
bp	base pair
BBP	Bayesian posterior probability
BioNJ	modified NJ algorithm
BML	Bayesian maximum likelihood
BS	non-parametric bootstrap support
C	cytosine
CADM	Congruence Among Distance Matrices
CEMD	Congruence Entre des Matrices de Distance
CI	confidence interval
CI_I	Rohlf's consensus information index
CM	congruent matrices
CM_{DPT}	CM generated with DPT permutations
CM_M	CM generated with Mantel permutations
CM_P	partly similar matrices
CM_I	identical matrices
COX1	cytochrome oxidase subunit I
COX2	cytochrome oxidase subunit II
COX3	cytochrome oxidase subunit III
CRSNG	Conseil de Recherches en Sciences Naturelles et en Génie
CYTB	cytochrome b
D	jeu de données hypothétique
D_1	agreement subtrees
DNA	desoxyribonucleic acid
DPT	double permutation test

e.g.	exemplum gratia
ESTs	Expressed sequence tags
et al.	et alia
FESP	Faculté des études supérieures et postdoctorales
FQRNT	Fonds Québécois de Recherches sur la Nature et les Technologies
G	guanine
GTR	General Time Reversible model
GTR3	General Time Reversible three-state model
GTR4	General Time Reversible four-state model
hLRTs	hierarchical likelihood ratio tests
i.e.	id est
ILD	Incongruence Length Difference
IM	incongruent or independent matrices
J	jumble option in the Fitch program (PHYLIP)
JC	Jukes-Cantor model
LBA	long-branch attraction
LEMEE	Laboratoire d'Écologie Moléculaire et d'Évolution
LINEs	long interspersed nuclear elements
M	number of matrices
Ma	million d'années
ML	maximum likelihood
ML-JC	maximum likelihood using a JC model of nucleotide substitution
ML-TVM	maximum likelihood using a TVM model of nucleotide substitution
MR	majority rule consensus
MRC	majority rule consensus with compatible groupings
MRP	matrix representation with parsimony
MSS	most similar supertree
mt	mitochondrial or mitogenomic
MT	model tree
MT_S	MT with short terminal branches
MT_L	MT with long terminal branches
NAD1	NADH dehydrogenase subunit 1
NAD2	NADH dehydrogenase subunit 2
NAD3	NADH dehydrogenase subunit 3

NAD4	NADH dehydrogenase subunit 4
NAD4L	NADH dehydrogenase subunit 4L
NAD5	NADH dehydrogenase subunit 5
NJ	neighbor-joining
NJ-JC	neighbor-joining with a JC correction
NJ-TVM	neighbor-joining with a TVM correction
NNI	nearest neighbor interchange
NSERC	Natural Sciences and Engineering Research Council of Canada
PM	partition metric
RQCHP	Réseau Québécois de Calcul de Haute Performance
SDM	unweighted super distance matrix
SDMw	weighted super distance matrix
SFIT	maximum splits fit
SINEs	short interspersed nuclear elements
SPR	subtree pruning and regrafting
T	thymine
TIM	transitional model
T-PTP	topology-dependent permutation tail probability test
TrN	Tamura-Nei model
TVM	transversional model
UE	unité évolutive
UEH	unite évolutive hypothétique
Y	pyrimidine

LISTE DES SYMBOLES

α	type I error
α	substitution rates among sites
β	type II error
Γ	gamma distribution
gA, gC, gG, gT	equilibrium frequencies of nucleotides (A, C, G, T)
H_0	null hypothesis
H_1	alternate hypothesis
i	row "i"
I	invariant site
I	level of matrix incompleteness
j	column "j"
L	DNA sequence length
m	number of ties
n	number of objects
n_{perm}	number of permuted objects
p	probability value
p	number of matrices
p	p distance
p	percentage of correctly inferred trees
p_{BB}	percentage of correctly inferred backbone trees
p_n	number of permutations
rAC	relative substitution rate between two nucleotides (A, C, G or T)
R_j	sum of ranks
$\overline{R}$	mean of the sum of ranks
s	mutation rate
T	correction factor for tied ranks
t_k	number of tied ranks for each k
W	Kendall's coefficient of concordance
χ^2	chi-square
χ^{2*}	permuted chi-square distribution
χ^{2*}_{DPT}	permuted chi-square distribution using DPT randomization
χ^{2*}_{M}	permuted chi-square distribution using Mantel randomization
χ^2_{ref}	observed Friedman's chi-square

À tous les chercheurs...

REMERCIEMENTS

Au cours de mes études universitaires, j'ai eu l'opportunité de rencontrer des personnes extraordinaires qui ont su me communiquer leurs passions, partager leurs intérêts, m'encourager et m'aider sans hésiter, autant sur le plan académique que personnel. Je ne peux malheureusement pas tous les mentionner ici, tant ils sont nombreux. Je restreindrai donc mes remerciements aux personnes directement impliquées par ma thèse de doctorat.

Trois femmes m'ont inspirée et encouragée à entreprendre des études supérieures. Jean Harris, une femme incroyable qui m'a chaleureusement accueillie en Afrique du Sud lors d'un stage en biologie marine et qui m'a initiée au monde de la biologie. Daphne Fairbairn, que j'ai rencontrée durant mon bac en biologie à l'Université Concordia et qui, en plus de m'initier à la recherche, a toujours cru en mes capacités et m'a poussée à aller plus loin. Finalement, ma mère, consultante en éducation, et pour qui, de toute évidence, les études et l'éducation occupent une place importante. C'est, entre autres, grâce à ces trois femmes que je termine aujourd'hui un doctorat en biologie. Je les en remercie infiniment.

J'ai complété la première année de mon doctorat à l'Université de Californie, à Riverside. J'aimerais souligner la gentillesse des personnes que j'ai rencontrées en Californie et qui ont contribué à rendre mon séjour aussi plaisant qu'il l'a été. " Thanks to Alysa, Denise, Reuben and Rob." Mais plus particulièrement, je tiens à remercier Mark Springer, mon premier superviseur, pour m'avoir donné la chance d'entreprendre mon doctorat en Californie.

De retour à Montréal, j'ai eu la chance d'entrer en contact avec François-Joseph Lapointe qui a, sans hésiter, accepté de me prendre comme étudiante au doctorat. François, parmi tes nombreuses qualités, j'apprécie énormément ta compréhension et ta joie de vivre. Merci d'avoir été constamment présent tout en me laissant la liberté dont j'avais besoin. Merci également pour ta confiance du début à la fin et ton soutien durant les moments de découragement. Tu es un exemple pour plusieurs et je te remercie de m'avoir permis de cheminer à tes côtés. Et SURTOUT, merci pour les nombreuses heures passées à écrire des programmes, sans lesquels cette thèse

n'aurait pas été possible. Les années passées sous ta supervision resteront à jamais gravées dans ma mémoire et cela aussi, bien sûr, grâce à tous ceux qui, comme moi, ont fait partie de ton laboratoire, le LEMEE. Entre autres, Nathalie, Olivier, Sarah, et Sébastien avec qui j'ai partagé la quasi-totalité de mes années d'études doctorales ainsi que de nombreuses heures de rires, de discussions, de bonheur mais aussi de frustrations. Merci à tous d'avoir su créer un environnement sain, propice à la recherche tout comme à l'épanouissement personnel. Merci également à Anaïs qui a fait partie du labo à mes débuts et qui a continué à s'enquérir des progrès de ma thèse tout au long de cette grande aventure.

Je tiens aussi à souligner les nombreuses heures passées au téléphone, avec ma sœur, Geneviève, qui fait un doctorat à l'autre bout du monde. Le partage de ses anecdotes et histoires incroyables dans la brousse africaine m'a permis de voyager tout en restant à Montréal.

J'aimerais également remercier tous les professeurs qui ont contribué à ma formation scientifique, lors de rencontres et comités ou en tant que membres du jury de cette thèse (en ordre alphabétique): Anne Bruneau, Pierre Darlu, Mohamed Hijri, Pierre Legendre et Hervé Philippe. Aussi, plusieurs personnes ont permis l'avancement de ma recherche par leurs appuis techniques. Merci à Stéphane Guindon, Antoine Lapointe et plus particulièrement à Marie-Hélène Duplain qui m'a toujours accueilli avec le sourire. Je remercie aussi le Réseau Québécois de Calcul de Haute Performance (RQCHP) pour m'avoir donné l'accès à leur infrastructure de calcul scientifique de haute performance ainsi qu'à Daniel Stubbs pour son soutien informatique.

Finalement, je voudrais remercier ceux qui m'ont octroyé un soutien financier, sans lequel je n'aurais pas pu mener mon doctorat à terme. Entre autres, le Conseil de Recherches en Sciences Naturelles et en Génie (CRSNG) et le Fonds Québécois de Recherches sur la Nature et les Technologies (FQRNT), ainsi que le Département de sciences biologiques et la Faculté des études supérieures et postdoctorales (FESP) de l'Université de Montréal.

CHAPITRE 1:

INTRODUCTION GÉNÉRALE

1.1. L'ANALYSE PHYLOGÉNÉTIQUE

1.1.1. Le contexte actuel

La connaissance de l'histoire évolutive des espèces vivantes est essentielle pour déterminer les relations de parenté entre les espèces présentes et éteintes ainsi que pour comprendre et expliquer leur biologie, écologie et physiologie. La branche scientifique dédiée à l'étude de l'évolution et des relations de parenté entre les organismes est l'analyse phylogénétique. Une phylogénie décrit « le cours historique de la descendance des êtres organisés » (Darlu & Tassy 1993). Durant les 40 dernières années, l'intérêt pour le domaine de l'analyse phylogénétique s'est accentué et de nombreuses méthodes statistiques et algorithmiques ont été développées afin d'analyser adéquatement les données disponibles (voir Felsenstein 2004 pour une revue bibliographique des méthodes disponibles). Parallèlement, le choix des données à utiliser s'est modifié avec le perfectionnement de nombreuses techniques moléculaires. Désormais, en plus des caractères morphologiques, les caractères moléculaires sont fréquemment utilisés pour inférer l'histoire évolutive des espèces. La disponibilité grandissante de séquences d'ADN combinée à l'utilisation d'ordinateurs de plus en plus puissants permettent désormais d'aborder de façon plus efficace la reconstruction phylogénétique (ex.: Smith & Donoghue 2008, Smith et al. 2009).

En outre, une nouvelle discipline a vu le jour : la phylogénomique (Eisen 1998, Eisen & Fraser 2003). Au cours des dernières années, un nombre impressionnant d'études phylogénétiques basées sur des caractères génomiques et étudiant divers groupes d'organismes ont été publiées (ex.: Fitzpatrick et al. 2006, Shedlock et al. 2007, Hackett et al. 2008, Horvath et al. 2008, Kuo et al. 2008, Struck & Fisse 2008, Hallström & Janke 2009) et même, des études visant à reconstruire l'arbre évolutif de toutes les espèces, c.-à-d. l'Arbre de la Vie ou *Tree of Life* (Wolf et al. 2002, Driskell et al. 2004, Delsuc et al. 2005, Ciccarelli et al. 2006, Philippe & Telford 2006, Dunn et al. 2008). De nouvelles approches ont été proposées afin de faciliter ces méta-analyses. Aussi, de nombreuses chaînes de traitement informatique (*pipeline*) sont maintenant disponibles. Ces dernières requièrent un minimum d'intervention de la part de l'usager puisque la sélection des taxons, l'alignement des séquences et l'inférence phylogénétique sont automatisés (ex.: PhyloBuilder : Glanville et al. 2007; Phylemon : Tárraga et al. 2007; GreenPhylDB : Conte et al. 2008; AMPHORA : Wu & Eisen 2008; mega-phylogeny : Smith et al. 2009). Toutefois, il subsiste encore de nombreux débats quant à la

méthodologie optimale à employer. Ainsi, la recherche n'est plus limitée par le nombre de données, mais plutôt par un manque de méthodes d'analyse adaptées (Philippe et al. 2005a, Rokas & Caroll 2006, Nahum & Pereira 2008). Entre autres, les méta-analyses sont souvent caractérisées par un nombre important de données manquantes qui peuvent nuire à la reconstruction phylogénétique (Hartmann & Vision 2008). De plus, l'analyse d'un nombre élevé de caractères peut mener à une augmentation de signaux non-phylogénétiques (Delsuc et al. 2005, Rodriguez-Ezpeleta et al. 2007) et contradictoires entre différents jeux de données (Jeffroy et al. 2006, Galtier & Daubin 2008, Kuo et al. 2008). Enfin, la meilleure stratégie à adopter pour combiner la quantité grandissante de données moléculaires reste à déterminer (ex.: le débat super-arbre versus super-matrice: Gatesy et al. 2002, 2004, Bininda-Emonds et al. 2004a, b, c, de Queiroz & Gatesy 2007, Ren et al. 2009, Cotton & Wilkinson 2009). Ma thèse vise à évaluer différentes approches phylogénétiques afin de contribuer à l'élaboration de stratégies efficaces pour analyser le nombre grandissant de données moléculaires.

1.1.2. Les arbres phylogénétiques

L'analyse phylogénétique tente de reconstruire l'histoire évolutive entre les espèces ou, plus généralement, entre les taxons. Les taxons sont des groupes qui incluent l'ensemble des organismes d'une catégorie de la classification biologique (ex.: un genre, une famille, ou un ordre). Les résultats de l'analyse phylogénétique sont communément exposés sous forme d'arbre, appelé arbre phylogénétique (Fig. 1.1). Les feuilles de l'arbre, ou nœuds externes, correspondent à des taxons terminaux aussi appelés unités évolutives (UE). Les nœuds internes de l'arbre représentent les ancêtres hypothétiques (UEH). Contrairement aux taxons terminaux qui sont existants et donc observables, leur histoire évolutive est inconnue et ne peut-être qu'inférée (Darlu & Tassy 1993). Ainsi, l'arbre phylogénétique ne représente qu'une hypothèse quant aux relations de parenté entre les taxons étudiés. L'ordre des différents branchements définit les relations de parenté entre les taxons et correspond à la topologie de l'arbre. Les descendants d'un même ancêtre forment un clade ou groupe monophylétique. La phylogénie permet donc de classifier les organismes selon leur histoire évolutive. En principe, la phylogénie inférée devrait être binaire, c.-à-d. que chaque nœud représente le point de spéciation des taxons issus d'un ancêtre commun. Si plus de trois branches sont connectées à un même noeud, il y a présence d'une polytomie. Il existe deux types de polytomies : dure (*hard)* ou molle (*soft)*, tout dépendant de l'interprétation choisie

(Maddison 1989). Une polytomie dure indique un événement de spéciation multiple où plus de deux espèces sont apparues simultanément. Une polytomie molle représente l'incertitude des relations évolutives entre les taxons et ce manque de résolution est souvent dû à un nombre insuffisant de caractères (Driskell et al. 2004, Dunn et al. 2008).

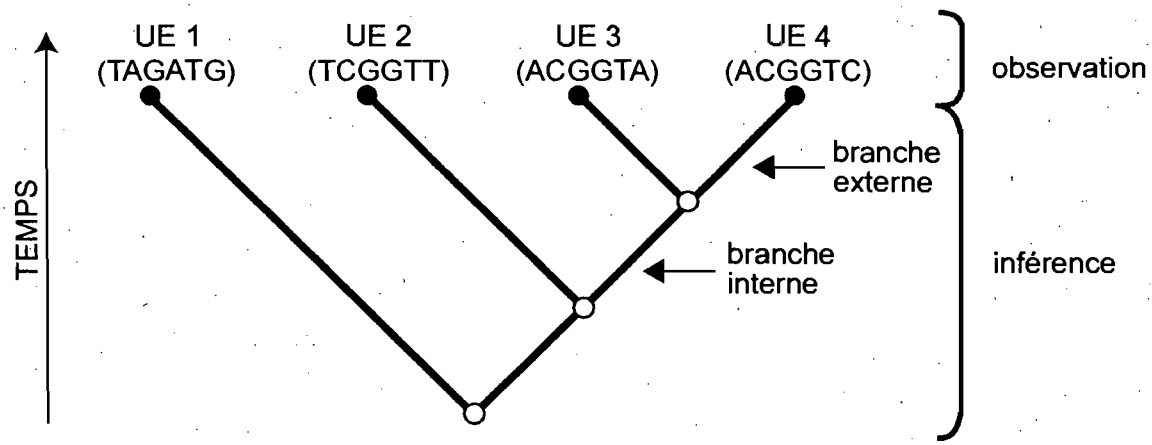

Figure 1.1. Arbre phylogénétique illustrant l'hypothèse des relations évolutives entre quatre taxons (UE 1 à UE 4).
L'étude des caractères moléculaires (dans cet exemple, une séquence d'ADN de six nucléotides) permet d'inférer l'histoire évolutive des taxons comparés, au cours du temps. L'arbre est complètement résolu puisqu'un maximum de trois branches sont connectées à chacun des nœuds. Les bifurcations représentent les événements de spéciation. Chaque nœud interne (cercle blanc) représente un ancêtre hypothétique (UEH) alors qu'un nœud externe (cercle noir) représente une unité évolutive (UE) existante et observable (illustration modifiée de Darlu & Tassy 1993).

1.1.3. La reconstruction phylogénétique

La reconstruction de l'histoire évolutive s'appuie sur des analyses mathématiques de données (ou caractères) provenant d'espèces contemporaines ou fossiles. Les caractères étudiés doivent être homologues, c'est-à-dire provenir d'un ancêtre commun (de Pinna 1991). L'analyse des changements d'état de ces caractères permet d'identifier les relations de descendance entre les taxons. Idéalement, les caractères étudiés auront changé peu de fois d'état au cours de l'évolution ou, dans le meilleur des cas, seulement une fois, afin de faciliter la reconstruction des relations de parenté.

Malheureusement, les multiples changements d'état sont fréquents, surtout lorsque les séquences d'ADN sont utilisées, puisque seuls quatre états, correspondant aux quatre nucléotides, sont possibles (A, C, G et T) et que les mutations (changement d'état) sont nombreuses. Les changements multiples causent du bruit dans les données sous forme d'homoplasie. L'homoplasie existe lorsque deux caractères sont identiques par le biais de processus évolutifs autres que la descendance, comme les réversions ou la convergence (Doyle & Davis 1998). Ces changements ont eu lieu indépendamment dans différents ancêtres et ne reflètent donc pas les liens de parenté évolutive.

Les méthodes d'analyse phylogénétique actuelles sont diverses : les méthodes phénétiques (ou de distances : voir Swofford et al. 1996 pour une revue des nombreuses méthodes de distances), cladistiques (reposant sur le maximum de parcimonie : Edwards & Cavalli-Sforza 1963) et probabilistes (le maximum de vraisemblance : Felsenstein 1973, 1981 et l'inférence bayésienne : Rannala & Yang 1996, Huelsenbeck et al. 2001). Les méthodes de distances se basent sur la ressemblance globale entre paires de taxons; plus les ressemblances sont élevées, plus les taxons sont apparentés. Par exemple, la distance entre deux taxons peut être calculée par le nombre de sites possédant un état différent dans une séquence d'ADN. Pour sa part, la parcimonie repose sur le principe du plus petit nombre de changements et donc choisit l'hypothèse évolutive qui minimise le nombre de changements évolutifs lors de l'analyse d'un jeu de données. Contrairement aux méthodes de distance, cette approche se base uniquement sur les synapomorphies. Les synapomophies sont des homologies dérivées et partagées. En somme, un état de caractère dérivé diffère de l'état ancestral et cet état est partagé par les membres inclus dans le groupe (défini par la synapomorphie) ainsi que leur plus récent ancêtre. Pour distinguer l'état ancestral et l'état dérivé, les caractères doivent être polarisés. La polarité peut être déterminée en utilisant un groupe externe au groupe d'intérêt, le groupe externe étant plus éloigné évolutivement (Maddison et al. 1984). Contrairement aux deux méthodes précédentes, les méthodes probabilistes permettent d'incorporer un modèle d'évolution moléculaire spécifique aux données lors de l'inférence phylogénétique et elles utilisent le concept de vraisemblance (la probabilité d'observer les données à partir d'une hypothèse évolutive). Bien qu'il existe différentes écoles de pensée quant à la meilleure méthode pour analyser l'alignement de séquences moléculaires, les méthodes probabilistes sont les plus utilisées. En effet, plusieurs études utilisant des simulations par ordinateurs ont démontré que les méthodes probabilistes sont plus adéquates pour analyser les

données moléculaires ayant évolué selon un modèle complexe (Huelsenbeck & Hillis 1993, Yang 1996a, Steel & Penny 2000, mais voir Goloboff 2003 pour une opinion différente). Cependant, l'émergence des méta-analyses a créé un regain d'intérêt pour les méthodes de distances, puisque celles-ci sont plus rapides que les méthodes probabilistes (ex.: Desper & Gascuel 2006). Également, le maximum de parcimonie semble être mieux adapté pour l'analyse de certains caractères génomiques qui sont codés de façon similaire aux caractères morphologiques (Albert 2005).

1.1.4. Les sources d'incongruence

Lorsque différents jeux de données (ou différentes partitions d'un même jeu de données) sont analysés, il est possible que les phylogénies inférées pour chacun de ces jeux diffèrent les unes des autres. On dit alors que les arbres inférés sont incongruents, c'est-à-dire qu'ils représentent différentes relations de parenté entre les taxons. Johnson & Soltis (1998) définissent la congruence de façon générale comme étant l'accord entre des arbres, des caractères ou des jeux de données. Ainsi, deux caractères (ou plus) seront congruents s'ils proposent les mêmes relations phylogénétiques. Or, il est plutôt rare que tous les caractères d'un ou de plusieurs jeux de données soutiennent exactement les mêmes relations phylogénétiques. La fiabilité des relations évolutives est souvent mesurée par le pourcentage de *bootstrap* (Felsenstein 1985) ou les probabilités postérieures bayésiennes (Huelsenbeck et al. 2001) qui sont des mesures statistiques de la robustesse des noeuds reflétant le signal phylogénétique présent dans les données. Phillips et al. (2004) ont obtenu des valeurs de *bootstrap* de 100% appuyant des relations évolutives contradictoires en utilisant des jeux de données différents pour les mêmes espèces. Puisque deux arbres différents ne peuvent tous deux décrire la phylogénie des espèces qui est unique, un de ces deux arbres (ou les deux) doit nécessairement présenter des relations artéfactuelles. De nombreux facteurs peuvent être à la base d'un désaccord entre deux phylogénies qui possèdent des taxons communs. Ces facteurs peuvent être regroupés en trois catégories : (1) les erreurs stochastiques, (2) des histoires évolutives différentes entre les gènes et, (3) les erreurs systématiques.

1.1.4.1. Les erreurs stochastiques

Les erreurs stochastiques sont des erreurs aléatoires dues à un nombre limité de caractères (c.-à-d. un échantillonnage incomplet). En général, un nombre accru de

caractères informatifs permet d'augmenter la résolution d'un arbre. Lorsqu'un arbre est complètement résolu, il est binaire et donc exempt de polytomie. Un faible nombre de sites, notamment lorsqu'un seul gène est utilisé, ne fournit que peu d'information pour reconstruire la phylogénie (Rokas et al. 2003a, b) alors que l'inclusion de plusieurs gènes permet d'augmenter le signal phylogénétique (de Queiroz 1993, Murphy et al. 2001a, Eisen & Fraser 2003, Philippe et al. 2004). Driskell et al. (2004) ont observé des relations évolutives qui n'apparaissaient pas dans les jeux de données analysés individuellement à cause d'un signal trop faible. En somme, l'ajout de caractères (ou de gènes) permet d'établir une phylogénie mieux résolue et diminue les erreurs dues à l'échantillonnage. De récentes études phylogénomiques ont obtenu des proportions de *bootstrap* de 100% à chaque nœud, dues au grand nombre de caractères utilisés (Rokas et al. 2003b, Dopazo et al. 2004). Le pouvoir de résolution est donc grandement amélioré dans les études phylogénomiques, surtout pour inférer des relations plus anciennes (Dunn et al. 2008). Par contre, même des jeux de données de grande échelle peuvent ne pas suffire à résoudre des événements de spéciation extrêmement rapides caractérisés par de courtes branches (Rokas & Caroll 2006, Wiens et al. 2008).

1.1.4.2. L'histoire évolutive des gènes

Outre un faible signal phylogénétique, le risque encouru lors de l'utilisation d'un seul ou de quelques gènes dans une analyse est de reconstruire l'histoire évolutive du ou des gènes et non celle des espèces (Page & Charleston 1997, Page 2000, Rokas et al. 2003b). Les analyses séparées de différentes séquences d'ADN (ou de gènes) peuvent proposer des phylogénies différentes pour les mêmes espèces, si un des gènes a une histoire évolutive différente des espèces qui le portent. Plusieurs processus évolutifs peuvent être à l'origine de la non-concordance entre l'histoire des gènes et celles des espèces : par exemple, les transferts horizontaux de gènes qui sont très répandus chez les bactéries et qui ont été observés chez les eucaryotes (Doolittle 1999) et les gènes paralogues qui sont le résultat de la duplication d'un gène (Fitch 1970). Les gènes ayant subi un transfert horizontal ou une duplication peuvent être identifiés en les comparant avec d'autres gènes, d'où l'avantage d'inclure de nombreux gènes dans une analyse (de Queiroz 1993, Lerat et al. 2003, Rokas et al. 2003b). En plus des transferts horizontaux et des duplications, un grand nombre de processus évolutifs peuvent expliquer l'incongruence entre différents jeux de données, tels que: l'hybridation, l'introgression, le polymorphisme ancestral et les recombinaisons génomiques (Wendel

& Doyle 1998, Degnan & Rosenberg 2006, Rannala & Yang 2008, Rosenberg & Tao 2008).

1.1.4.3. Les erreurs systématiques

Bien que les deux sources d'erreurs citées précédemment soient diminuées lors de l'utilisation d'un grand nombre de données, les erreurs systématiques sont souvent accentuées (Rokas et al. 2003a, Delsuc et al. 2005, Nishihara et al. 2007, Rodriguez-Ezpeleta et al. 2007). Les erreurs systématiques sont dues à la présence de signaux qui ne reflètent pas la parenté des espèces (c.-à-d. des signaux non-phylogénétiques) et sont intimement liées à la qualité des données et des méthodes de reconstruction (Delsuc et al. 2005). Les erreurs systématiques peuvent être diminuées en utilisant une méthode de reconstruction phylogénétique qui modélise de façon appropriée l'hétérogénéité présente dans les données observées (Lio & Goldman 1998, Rodriguez-Ezpeleta et al. 2007).

Une forte hétérogénéité dans les données, que ce soit un biais dans la composition des nucléotides (ex.: Gibson et al. 2005), une interdépendance entre les sites d'un codon (Whelan 2008) ou encore une différence du taux de substitutions entre les sites ou entre les taxons, peut créer un artéfact dans l'inférence des relations évolutives, lorsqu'une méthode d'analyse ou un modèle inapproprié est utilisé (Pupko et al. 2002, Rokas et al. 2003a, Delsuc et al. 2005, Philippe et al. 2005b, Jeffroy et al. 2006). Delsuc et al. (2002) ont démontré que les processus évolutifs responsables des fréquences de nucléotides et des taux et types de mutations pouvaient être très variables entre différents gènes. Ainsi, il est possible que des taxons qui possèdent des fréquences de nucléotides similaires soit regroupé sur cette base, même s'ils ne partagent pas d'ancêtre commun récent (ex. : Steel et al. 1993, Phillips et al. 2004, Collins et al. 2005, Jeoffroy et al. 2006). D'un autre côté, la vitesse d'évolution d'un nucléotide dépend de sa position dans le génome ou dans un gène (première, deuxième ou troisième position d'un codon), du type de locus (gène nucléaire, mitochondrial, pseudogène, microsatellitte, etc.) et de sa fonction (Bofkin & Goldman 2007). Dans chacun de ces cas, les nucléotides seront soumis à des pressions sélectives qui leur sont propres, créant ainsi des partitions naturelles au sein des données. Lorsque les pressions sélectives sont très divergentes, il est probable que le signal phylogénétique soit significativement différent d'une partition à l'autre (Bull et al. 1993). En effet, Dopazo &

Dopazo (2005) et Philippe et al. (2005b) ont observé que le taux d'évolution des gènes choisis pour une analyse phylogénétique influençait les groupements phylogénétiques inférés. Un des dangers lié à l'utilisation de gènes ou d'espèces qui évoluent rapidement est l'attraction des longues branches lors de la reconstruction phylogénétique.

Le phénomène de l'attraction des longues branches (*long-branch attraction*, LBA: Felsenstein 1978) est un problème très fréquent en analyse phylogénétique (Philippe & Laurent 1998, Gribaldo & Philippe 2002, Bergsten 2005, Delsuc et al. 2005). Cet artéfact est observé lorsque deux taxons sont regroupés indépendamment de leurs liens de parenté parce qu'ils partagent un grand nombre d'états de caractère identique qui résultent d'une évolution convergente à ces sites, plutôt que d'une descendance commune. Par exemple, il arrive que le groupe externe, qui est par définition plus éloigné évolutivement, attire des taxons qui ont évolué rapidement et pour qui de nombreuses substitutions ont eu lieu, ce qui place ces taxons illégitimement à la base de l'arbre (Philippe 1997, Philippe & Laurent 1998). Les espèces qui évoluent très rapidement ou celles qui ont divergé depuis longtemps sont caractérisées par de longues branches qui reflètent les multiples substitutions d'états de caractère, d'où le nom d'attraction des longues branches. Philippe & Laurent (1998) ont proposé différentes approches afin d'éviter un artéfact dans la reconstruction de la phylogénie lorsque des espèces ont un taux d'évolution rapide :

(1) choisir une méthode de reconstruction moins sensible à ce problème;

(2) choisir des gènes qui évoluent plus lentement;

(3) utiliser des signatures moléculaires rares telles que des insertions;

(4) améliorer l'échantillonnage des taxons pour couper les branches longues et choisir un groupe externe moins éloigné.

Ces quatre approches sont discutées dans les trois sections suivantes.

1.1.4.3.1. Le choix d'une méthode de reconstruction

Lorsque les données ne correspondent pas aux prémisses d'une méthode de reconstruction, celle-ci peut être incohérente, c'est-à-dire que l'appui en faveur d'une phylogénie erronée augmentera en fonction du nombre de caractères utilisés (Felsenstein 1978, Kim 1996, Lockhart et al. 1996, Phillips et al. 2004). Toutes les méthodes d'inférence phylogénétique sont sensibles au problème de LBA mais les

méthodes probabilistes et de distances corrigées à l'aide d'un modèle évolutif le sont moins (Lockhart et al. 1996). Avec les approches probabilistes, l'exactitude de l'inférence dépend largement de la qualité du modèle évolutif choisi *a priori* et du degré de déviation des données par rapport à ce modèle. Il est donc important d'utiliser un modèle évolutif qui représente adéquatement celui des caractères. Chang (1996) a démontré que la méthode du maximum de vraisemblance est cohérente lorsque l'inférence phylogénétique utilise un modèle d'évolution identique à celui qui caractérise l'évolution des données observées. Dans un effort pour améliorer les modèles évolutifs, plusieurs paramètres ont été ajoutés dans différents modèles afin de tenir compte de divers biais observés dans les données. À titre d'exemple, on peut tenir compte de l'inégalité entre les taux de transitions et de transversions (Kimura 1980), ou bien du taux de substitutions hétérogènes entre les sites (Yang 1993). Les modèles évolutifs sont plus ou moins complexes, cependant même le modèle le plus complexe est loin de représenter toute la complexité inhérente à l'évolution des nucléotides.

1.1.4.3.2. Le choix du type de caractères

Le choix et la stratégie d'échantillonnage des caractères sont importants en analyse phylogénétique. Le type de marqueurs utilisé doit être approprié à la question posée. Le taux d'évolution diffère souvent d'un marqueur à un autre. Par conséquent, différents marqueurs peuvent être utilisés pour élucider des relations plus ou moins anciennes. Lorsqu'un gène évolue trop rapidement, le signal phylogénétique disparaît en raison des substitutions multiples (saturation mutationnelle) alors qu'un gène qui évolue trop lentement offre peu de résolution puisque la majorité des sites sont identiques entre les espèces (Moritz et al. 1987, Graybeal 1994). Par exemple, les mitochondries animales sont caractérisées par un taux de mutation plus élevé que celui de l'ADN nucléaire dans la majorité des espèces. De plus, différents taux de substitution sont observés pour différentes parties de la mitochondrie et pour différents gènes nucléaires. Dans le génome mitochondrial, la région de contrôle évolue rapidement alors que les gènes codant pour l'ARN ribosomal évoluent beaucoup plus lentement (Moritz et al. 1987).

Alors que les analyses phylogénétiques moléculaires traditionnelles s'appuient le plus souvent sur des séquences de gènes nucléaires ou mitochondriaux, les analyses phylogénomiques utilisent des séquences à plus grande échelle de l'ordre de mégakilobases. Ces méga-matrices sont obtenues par la combinaison de nombreux

gènes complets ou partiels (ex.: ESTs, *Expressed Sequence Tags*: Rudd 2003). Outre la comparaison de séquences d'ADN, d'autres types de caractères phylogénétiques sont utilisés en phylogénomique. Parmi ceux-ci, citons l'ordre des gènes, le contenu en gènes, la comparaison de signatures génomiques et des changements génomiques rares tels que l'intégration de LINEs ou de SINEs (*Long and Short Interspersed Nuclear Elements*) et la fission/fusion de gènes (voir Delsuc et al. 2005 pour une description de ces méthodes). Ces caractères sont analysés à l'aide de méthodes phylogénétiques traditionnelles. Ainsi le contenu en gènes de différents génomes est codé dans une matrice binaire reflétant la présence ou l'absence du gène et analysée par des méthodes de distance ou de maximum de parcimonie (Snel et al. 1999, Korbel et al. 2002). De façon similaire, la méthode utilisée pour analyser les signatures génomiques calcule la fréquence d'apparition de certains motifs oligonucléotidiques et la matrice de fréquence est analysée avec des méthodes de distance (Edwards et al. 2002). Les intégrations de LINEs et de SINEs sont également codées selon leur présence/absence et analysées avec le maximum de parcimonie (Nikaido et al. 1999). Un avantage lié à l'utilisation de changements génomiques rares est l'absence d'homoplasie, puisque les événements sont pour la plupart peu fréquents, aléatoires et irréversibles (Rokas & Holland 2000, Boore 2006, mais voir aussi Hillis 1999, Snel et al. 2000 et Pecon-Slattery et al. 2004 pour une opinion différente).

1.1.4.3.3. L'échantillonnage des taxons

Dans le but d'augmenter la justesse (ou l'exactitude) des reconstructions phylogénétiques, plusieurs auteurs ont suggéré d'accroître le nombre de taxons inclus dans les analyses (Lecointre et al. 1993, Hillis 1996, 1998, Graybeal 1998, Lin et al. 2002a, Pollock et al. 2002, Zwickl & Hillis 2002, DeBry 2005, Leebens-Mack et al. 2005, Baurain et al. 2007, Telford 2008). Un des avantages liés à la sélection d'un plus grand nombre de taxons est la diminution de l'attraction des longues branches (Hendy & Penny 1989, Philippe 1997, Hillis 1998, Philippe & Laurent 1998, Philippe et al. 2005c, Hedtke et al. 2006). Et même, certains auteurs ont affirmé qu'il est plus judicieux d'augmenter le nombre de taxons plutôt que le nombre de caractères dans les analyses phylogénétiques (Lecointre et al. 1993, Hillis 1996, 1998, Graybeal 1998). À l'opposé, d'autres études ont démontré que l'ajout de taxons pouvait diminuer l'exactitude des arbres inférés dans certaines situations (Kim 1996, Poe & Swofford 1999, Rokas & Carroll 2005, Gatesy et al. 2007) et que l'ajout de caractères devait être préféré à l'ajout

de taxons (Rosenberg & Kumar 2001, 2003, Rokas & Carroll 2005). Pour éviter des erreurs stochastiques et systématiques, il apparaît évident que la situation idéale est d'analyser un maximum de caractères pour l'ensemble des espèces (Pollock et al. 2002, Rosenberg & Kumar 2003, Delsuc et al. 2005) en utilisant une méthode de reconstruction cohérente (Hillis et al. 1996). Hillis et al. (2003) concluent qu'avec des ressources limitées, le choix d'inclure plus de taxons ou plus de caractères dans une analyse dépend de plusieurs facteurs, notamment, les objectifs de l'étude, le type ainsi que le nombre de taxons et de caractères déjà échantillonnés. D'un autre côté Smith et al. (2009) ont reconstruit une phylogénie des Viridiplantae (i.e., les plantes vertes) et ont retrouvé plusieurs relations bien établies en utilisant un seul gène, mais en augmentant le nombre d'espèces échantillonnées (i.e., 13 533 espèces). Cependant, quelques relations incongruentes ont été notées qui sont probablement le reflet de l'histoire évolutive du gène utilisé. Il serait donc à propos d'inclure plus d'un gène pour éviter ce type d'artéfact. Dans cette optique, la révolution phylogénomique devrait permettre de bénéficier également d'un grand nombre de taxons et de caractères dans une même étude (Philippe & Telford 2006, Baurain et al. 2007, Heath et al. 2008).

1.1.5. La combinaison des données

Lorsqu'on choisit d'inclure un grand nombre de caractères et de taxons, plusieurs approches existent pour combiner les données dans l'intention de produire une phylogénie unique. Avec l'apparition des méga-analyses, le débat concernant la méthodologie optimale pour combiner les données s'est transformé. Alors qu'il opposait les partisans de l'analyse séparée et de l'analyse combinée de différents jeux de données (Huelsenbeck et al. 1996a, b), il confronte maintenant les supporters de l'analyse de type super-matrice et super-arbre (Bininda-Emonds 2004a, de Queiroz & Gatesy 2007). Ce nouveau débat est une généralisation de l'ancien, en ce sens qu'avec l'approche de type super-arbre, l'identité des taxons se chevauche partiellement au lieu d'être identique pour tous les jeux de données. Une troisième approche existe également: l'approche combinée conditionnelle, qui est à mi-chemin entre l'analyse séparée et combinée (Fig. 1.2).

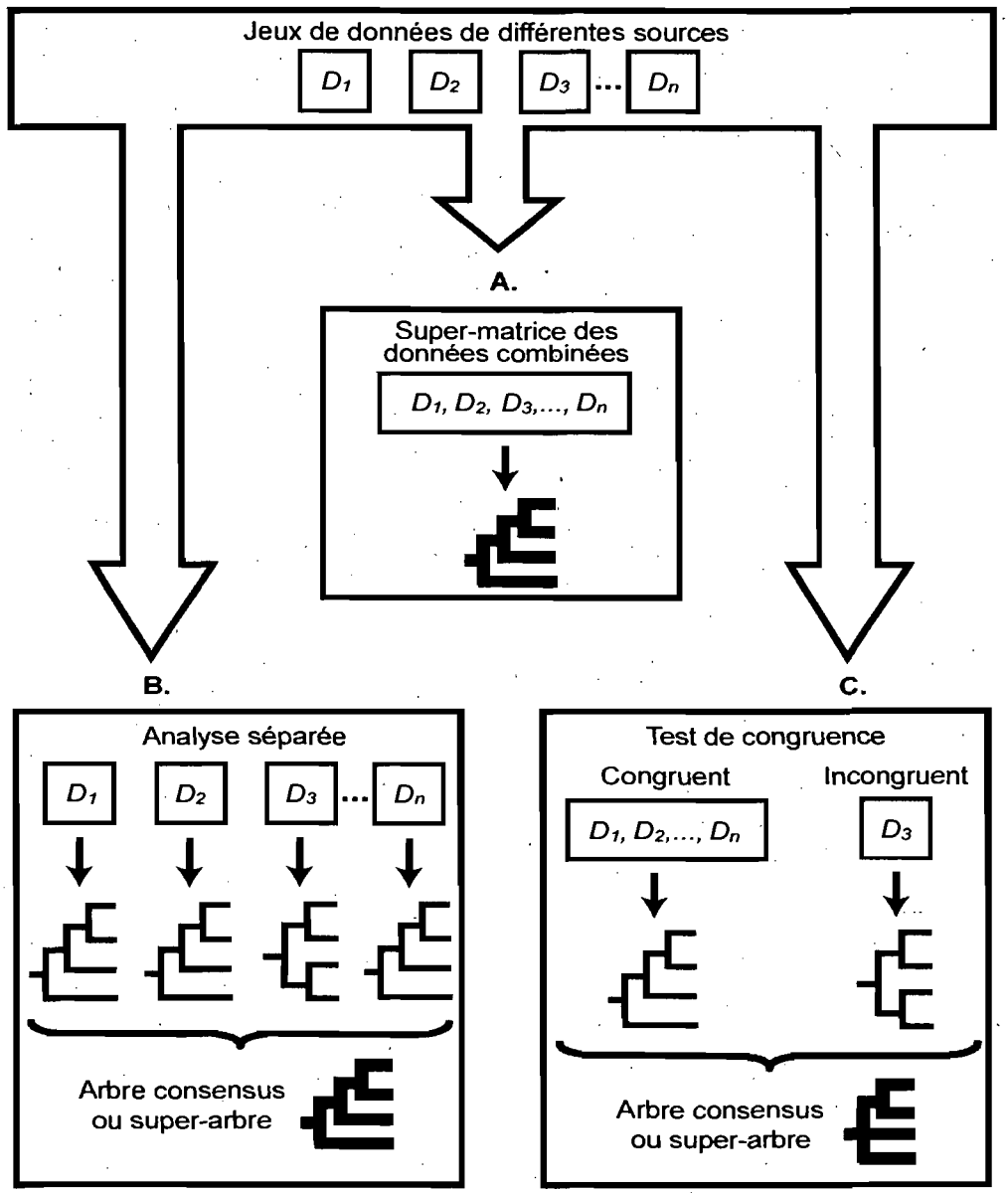

Figure 1.2. La combinaison des jeux de données en analyse phylogénétique et phylogénomique.

À partir de n jeux de données, trois méthodes d'analyse sont communément utilisées. A. L'approche d'évidence totale (ou analyse combinée) inclut les différents jeux de données dans une seule matrice ou super-matrice avant l'analyse. B. L'approche de l'analyse séparée effectue les analyses sur chaque jeu de données séparément et compare les arbres résultants à l'aide d'une méthode de consensus ou de super-arbre. C. L'approche combinée conditionnelle inclut seulement les jeux de données qui sont congruents dans une même analyse (ex.: D1, D2, ..., Dn) alors que les jeux de données incongruents (ex.: D3) sont analysées séparément (adaptée de Huelsenbeck et al. 1996b).

1.1.5.1. L'analyse combinée ou de type super-matrice

L'analyse conjointe des données est souvent comparée à l'approche de l'évidence totale (*total evidence*, *sensu* Kluge 1989) puisqu'elle incorpore toute l'information disponible dans une même analyse (Fig. 1.2A). Les partisans de cette approche s'appuient sur les avantages de la congruence des caractères et affirment que l'incorporation simultanée de toutes les données dans une analyse unique permet d'accentuer le signal phylogénétique et ainsi d'obtenir des relations évolutives qui ne pourraient pas être inférées dans le cas où les jeux de données seraient analysés séparément (Kluge 1989, Barrett et al. 1991, Eernisse & Kluge 1993, Jones et al. 1993, Kluge & Wolf 1993). De nos jours, les jeux de données analysés contiennent de plus en plus de caractères et de taxons et les matrices de caractères sont généralement appelées des super-matrices (*sensu* Sanderson et al. 1998). Les analyses incluant plus de 100 gènes sont maintenant courantes (ex.: Fitzpatrick et al. 2006, Nishihara et al. 2007, Wildman et al. 2007, Dunn et al. 2008, Zou et al. 2008) alors qu'elles étaient encore rares il y a quelques années. Par contre, les matrices comprenant de nombreux caractères pour de nombreux taxons sont souvent dominées par les données manquantes (Driskell et al. 2004, Philippe et al. 2005a, Wiens 2006, Telford 2008).

Alors que certains auteurs affirment qu'une inférence juste peut être obtenue avec un grand nombre de données manquantes s'il reste suffisamment de caractères informatifs (Wiens 2003a, 2006, Driskell et al. 2004, Philippe et al. 2004, Wiens & Moen 2008), d'autres concluent au contraire que l'inclusion de données manquantes peut mener à une diminution de la résolution et de l'exactitude de l'inférence phylogénétique (Kearney 2002, Hartmann & Vision 2008). Pour diminuer le nombre de données manquantes, certains chercheurs ont recours aux taxons chimères (*composite taxa*), où les séquences d'ADN de deux ou plusieurs taxons sont combinées pour ne former qu'une seule séquence ou taxon chimère (Springer et al. 2004a). Une autre méthode phylogénétique qui contourne le problème des données manquantes est l'approche de type super-arbre.

1.1.5.2. L'analyse séparée ou de type super-arbre

À l'opposé de l'analyse combinée, Miyamoto & Fitch (1995) suggèrent d'analyser les différents jeux de données individuellement. Ainsi, les jeux de données de différents types ou hétérogènes peuvent être analysés avec des algorithmes ou modèles

d'évolution adaptés. Par la suite, les phylogénies obtenues sont comparées à l'aide de différents indices de congruence ou combinées à l'aide de méthodes de consensus (Fig. 1.2B). Cette approche correspond à la congruence taxonomique (*sensu* Mickevich 1978) où l'appui pour des groupes monophylétiques est obtenu en comparant les résultats d'analyses indépendantes. L'arbre consensus est obtenu en utilisant différentes méthodes, plus ou moins conservatrices (voir Swofford 1991 pour une revue des méthodes de consensus). Par exemple, le consensus strict (Sokal & Rohlf 1981) est la méthode la plus conservatrice puisqu'elle n'inclut que les groupes monophylétiques présents dans toutes les phylogénies comparées. Les autres méthodes sont moins conservatrices et, par le fait même, l'arbre consensus sera souvent plus résolu. Certains préfèrent une solution plus conservatrice où la probabilité de représenter de vrais groupes monophylétiques est plus grande (Swofford 1991) alors que d'autres favorisent des méthodes de consensus qui offrent plus de résolution et dont l'arbre consensus contient plus d'information phylogénétique (Kluge & Wolf 1993). Cependant, plusieurs critiques ont été émises à l'égard des méthodes de consensus (Barrett et al. 1991, Chippindale & Wiens 1994).

Lorsque les taxons des différentes phylogénies présentent un chevauchement partiel, les phylogénies sources sont combinées à l'aide de méthodes de super-arbre (Baum 1992, Doyle 1992, Ragan 1992a, b). Plusieurs raisons peuvent être invoquées pour expliquer le chevauchement partiel des taxons entre différentes études ou différents jeux de données. Certaines données peuvent ne tout simplement pas être disponibles pour certains taxons (représentées par des données manquantes dans une matrice), ou encore la sélection des taxons peut différer d'un groupe de recherche à un autre. La première méthode de super-arbre, développée en 1992, et appelée *matrix representation with parsimony* ou MRP (Baum 1992, Ragan 1992a, b), représente les nœuds (ou sous-arbres) des différentes phylogénies dans une matrice, à l'aide d'un codage additif binaire (Farris et al. 1970). Depuis l'avènement du MRP, plusieurs autres techniques de super-arbre ont été proposées (voir Baum & Ragan 2004, Bininda-Emonds 2004b, Wilkinson et al. 2005, Cotton & Wilkinson 2007, 2009 pour une liste des méthodes disponibles). L'approche de type super-arbre a été utilisée, entres autres, pour résoudre les relations entre les différentes espèces de primates (Purvis 1995), de carnivores (Bininda-Emonds et al. 1999), et également, de toutes les espèces de mammifères (Bininda-Emonds et al. 2007). Cette approche pourrait d'ailleurs faciliter la

reconstruction de l'Arbre de la Vie (Sanderson et al. 1998, Bininda-Emonds et al. 2002, Pennisi 2003).

1.1.5.3. L'analyse combinée conditionnelle

L'analyse combinée conditionnelle teste la congruence ou l'homogénéité des différents jeux de données avant de les inclure dans une analyse conjointe (Bull et al. 1993, de Queiroz 1993; Fig. 1.2C). Plusieurs méthodes sont disponibles pour tester statistiquement l'incongruence ou la congruence de différents jeux de données en analyse phylogénétique (Bull et al. 1993, Rodrigo et al. 1993, Farris et al. 1994, 1995, Huelsenbeck & Bull 1996). Cependant, la plupart de ces méthodes ont été sévèrement critiquées (Huelsenbeck et al. 1996a, b, Cunningham 1997a, b, Barker & Lutzoni 2002, Darlu & Lecointre 2002, Leigh et al. 2008). Plus récemment, des auteurs ont suggéré des tests dérivés de modèles probabilistes (Bapteste et al. 2005, Brochier et al. 2005, Suchard 2005, Susko et al. 2006, Ané et al. 2007, Leigh et al. 2008). Suite à l'analyse statistique, les jeux de données qui ne sont pas congruents avec les autres sont analysés séparément et les phylogénies obtenues peuvent ensuite être comparées avec des méthodes de consensus ou de super-arbre.

1.1.5.4. Le débat super-matrice versus super-arbre

Plusieurs avantages et inconvénients ont été avancés lors de polémiques entourant l'utilisation des super-matrices ou des super-arbres (ex.: Sanderson et al. 1998, Springer & de Jong 2001, Gatesy et al. 2002, Bininda-Emonds 2004a, b, Gatesy & Springer 2004, Gatesy et al. 2004, de Queiroz & Gatesy 2007). Certains des avantages de l'approche de type super-matrice ont déjà été mentionnés dans cette introduction. Par exemple, l'incorporation d'un grand nombre de jeux de données dans une analyse permet de diminuer les erreurs stochastiques et d'augmenter le signal phylogénétique par rapport au bruit. Cependant, il peut être fastidieux d'assembler une super-matrice et celle-ci est souvent caractérisée par un grand nombre de données manquantes. Pour sa part, une analyse de type super-arbre permet de tirer profit du travail effectué par des équipes de recherche qui utilisent différents types de caractères et différents taxons (Bininda-Emonds & Sanderson 2001, Bininda-Emonds 2004c). Aussi, puisque l'information nécessaire à la construction d'un super-arbre est incluse dans les arbres phylogénétiques, les analyses peuvent être effectuées sans avoir la lourde tâche d'obtenir les données originales. De plus, comme les méthodes de super-arbre utilisent

des phylogénies au lieu de caractères, différents types de données peuvent être combinés pour une même inférence phylogénique. En effet, les analyses des jeux de données ont été effectuées au préalable, par des méthodes choisies en fonction du type de caractères, afin de produire chaque phylogénie. Ceci représente un autre avantage sur les super-matrices où l'analyse simultanée de tous les caractères limite grandement la possibilité d'utiliser différents critères d'optimisation ou différentes méthodes d'analyse dans le cas où les données sont hétérogènes ou incompatibles (Tateno et al. 1994, Bininda-Emonds 2004b, Edwards et al. 2007, Kubatko & Degnan 2007). Cependant, les méthodes probabilistes appliquées aux super-matrices permettent, dans une même analyse, l'utilisation de différents modèles appropriés aux différents gènes ou positions dans l'alignement (Yang 1996b, Ronquist & Huelsenbeck 2003, Nylander et al. 2004, Bofkin & Goldman 2007). Venant appuyer l'approche super-matrice, des simulations par ordinateur ont démontré que certains algorithmes de type super-arbre n'étaient pas performants, et ce même dans une situation idéale où tous les gènes évoluaient selon un taux et un modèle d'évolution identique (Eulenstein et al. 2004). Lorsque les taux de substitutions entre les gènes sont hétérogènes, Ren et al. (2009) ont observé que l'approche de type super-matrice était supérieure à l'approche de type super-arbre, mais seulement lorsque des paramètres différents, qui reflètent l'hétérogénéité dans les données, étaient utilisés pour chaque gène.

Il arrive que la construction de super-arbre produise des groupes monophylétiques qui ne sont pas présents dans les phylogénies sources. Pisani & Wilkinson (2002) ont décrit des cas où les super-arbres générés ne correspondaient pas aux phylogénies inférées à partir d'une super-matrice ou encore étaient en conflit avec les arbres sources (Wilkinson et al. 2007). Ces nouveaux groupes ne sont pas nécessairement basés sur la congruence des caractères, ils seraient plutôt un artéfact de la méthode puisqu'il n'y a aucune interaction entre les données brutes des différents jeux de données. Les caractères qui appuient des hypothèses phylogénétiques différentes (*subsignals*, sensu Pisani & Wilkinson 2002) sont exclus par la méthode de super-arbre et ne peuvent donc interagir pour soutenir de nouveaux groupes comme lorsque combinés dans une super-matrice. Cette absence d'interaction entre les différents jeux de données à l'état primaire a d'ailleurs été sévèrement critiquée par plusieurs (Rodrigo 1993, 1996, Gatesy et al. 2002, Springer & de Jong 2001, Wilkinson et al. 2001, Gatesy & Springer 2004, Ross & Rodrigo 2004). Une autre critique formulée à l'égard des super-arbres a trait au choix des arbres sources inclus dans l'analyse (Springer & de

Jong 2001, Gatesy et al. 2002). En combinant plusieurs phylogénies provenant de différentes sources, le risque d'inclure des caractères qui ont été utilisés dans plus d'une phylogénie est augmenté. Si tel est le cas, la prémisse selon laquelle les caractères sont indépendants est violée et le poids donné à ces caractères est augmenté de façon injustifiée (Springer & de Jong 2001, Gatesy et al. 2002, Bininda-Emonds 2004c).

Par le passé, plusieurs auteurs (de Queiroz 1993, 1995, Larson 1994) ont suggéré de contraster les résultats d'analyses séparées et combinées. En ce sens, Lapointe et al. (1999) ont démontré par l'approche de congruence globale que les méthodes des super-matrices et des super-arbres peuvent produire des résultats équivalents lorsqu'une méthode de consensus qui tient compte des longueurs de branches est utilisée (par exemple: le consensus moyen; Lapointe & Cucumel 1997). Ces super-arbres sont mieux résolus que ceux qui sont inférés à partir de méthodes qui n'utilisent que la topologie des arbres sources (Lapointe 1998a).

1.2. LES MAMMIFÈRES

1.2.1. La classification des mammifères

Traditionnellement, les espèces de mammifères sont regroupées en deux sous-classes : Prototheria et Theria. La sous-classe Prototheria aussi appelé Monotremata, comprend les monotrèmes alors que la sous-classe Theria est subdivisée en deux grands groupes : Metatheria (Marsupialia) et Eutheria (Placentalia). Les monotrèmes sont ovipares alors que les placentaires et les marsupiaux sont vivipares. Chez les marsupiaux, le développement de l'embryon se termine dans la poche marsupiale. À l'intérieur de ces trois groupes, plusieurs classifications biologiques ont été proposées selon différentes interprétations et différents choix de critères et de caractères phylogénétiques (Tableau 1.1). En effet, la nomenclature et les divisions taxonomiques ont changé ainsi que le nombre total d'espèces. À titre d'exemple, alors que Gill (1872) avait défini 13 ordres et 105 familles de mammifères, plus récemment Wilson & Reeder (2005) ont déterminé 29 ordres et 153 familles. Les espèces de mammifères sont distribuées de façon très inégale à tous les niveaux hiérarchiques de la classification (ex.: genres, familles et ordres). Les mammifères actuels sont largement dominés par les placentaires (5080 espèces, environ 95% des espèces). Par comparaison, les

marsupiaux comptent 331 espèces, et les monotrèmes ne comptent que cinq espèces (Wilson & Reeder 2005).

Idéalement, la classification devrait être naturelle, c.-à-d. correspondre aux liens de parenté entre les espèces (Darwin 1859, Hennig 1966). Elle est donc de plus en plus influencée par les études phylogénétiques (Lecointre & Le Guyader 2006). La classification des familles de mammifères utilisée dans le cadre de ma thèse est celle de Wilson & Reeder (2005), qui compte 150 familles de mammifères (excluant les familles éteintes) comprenant deux familles de monotrèmes, 19 familles de marsupiaux et 129 familles de placentaires. Cependant, alors que Wilson & Reeder (2005) reconnaissent 29 ordres de mammifères, la majorité des études phylogénétiques récentes n'en compte que 26. Le nombre d'ordres plus élevé proposé par Wilson & Reeder (2005) est dû à la séparation de trois ordres : (1) Cetartiodactyla subdivisé en Cetacea et Artiodactyla, (2) Eulipotyphla subdivisé en Erinaceomorpha et Soricomorpha et (3) Xenarthra subdivisé en Cingulata et Pilosa. La nomenclature des ordres de mammifères utilisée dans ma thèse reflète le consensus moléculaire actuel qui comprend 26 ordres.

Tableau 1.1. Nombre d'ordres, de familles, de genres et d'espèces de mammifères selon différentes sources.
(Adapté de Shoshani & McKenna 1998).

Auteurs	Ordres	Familles	Genres	Espèces
Gill (1872)	13	105	-	-
Gregory (1910)	17	-	-	-
Simpson (1945)	18	118	932	-
Wilson & Reeder (1993)	26	135	1135	4629
McKenna et al. (1997)	23	125	1083	-
Wilson & Reeder (2005)	29	153	1229	5416

La connaissance de l'histoire évolutive des mammifères est importante non seulement à des fins de classification et d'un point de vue anthropocentrique (notre propre histoire évolutive), mais aussi afin de mieux interpréter et comprendre la morphologie, la physiologie, le comportement et le génome des différentes espèces de mammifères (Springer et al. 2004b). De plus, la phylogénie des mammifères est un outil important dans le domaine médical, que ce soit pour l'identification de régions régulatrices dans le génome humain ou encore pour prédire les conséquences fonctionnelles de mutations qui surviennent dans les gènes associés à de nombreuses maladies (Springer & Murphy 2007). Il est donc primordial d'établir avec justesse les relations de parenté entre les espèces de mammifères, néanmoins plusieurs incertitudes subsistent quant aux relations évolutives entre les espèces de mammifères.

1.2.2. La phylogénie des mammifères

Bien que la majorité des espèces de mammifères soit extrêmement bien connue et étudiée, leur histoire évolutive demeure nébuleuse. La classe des mammifères comprend 5416 espèces (selon Wilson & Reeder 2005) et serait apparue il y a de 217 à 328 Ma, durant la période du Crétacé (Springer & Murphy 2007). Plusieurs hypothèses ont été proposées pour décrire de façon temporelle la radiation des mammifères (revue par Archibald & Deutschman 2001, Archibald 2003). Ces hypothèses diffèrent quant à la vitesse et à l'ère géologique auxquelles ont eu lieu les radiations inter et intraordinales. L'hypothèse qui est la plus appuyée par les données moléculaires correspond au « long fuse model » où la radiation interordinale s'est effectuée au Crétacé alors que la radiation intraordinale s'est effectuée à la suite de la crise qui est survenue à la limite du Crétacé-Tertiaire (Springer & Murphy 2007). Peu importe l'hypothèse choisie, la radiation des mammifères a eu lieu durant une période relativement courte, ce qui explique en partie la difficulté à reconstruire leur phylogénie avec exactitude (Simpson 1945, Miyamoto & Goodman 1986, Hallström 2008, Nishihara et al. 2009).

1.2.2.1. Les phylogénies estimées à partir de caractères morphologiques

Les études phylogénétiques qui utilisent des caractères morphologiques ont permis de définir les deux sous-classes de mammifères ainsi que la plupart des ordres et des espèces de mammifères (Springer & Murphy 2007). Aussi, certains clades tels que (1) les glires, regroupant les lagomorphes et les rongeurs (Simpson 1945), (2) les

paenungulates, regroupant les ordres : Proboscidae, Hyracoidae et Sirenia (Simpson 1945) et (3) les xénarthres (Cope 1889) ont été maintenus jusqu'à ce jour. Par contre, les études morphologiques se sont avérées moins efficaces pour résoudre les relations de parenté entre les groupes et plus particulièrement pour résoudre les relations anciennes puisque peu de caractères morphologiques sont discriminants à cette échelle (Graur 1993a, b). En effet, les phylogénies inférées présentent de nombreuses polytomies au niveau interordinal (par exemple, voir la phylogénie proposée par Novacek 1992). En outre, l'existence de nombreuses radiations adaptatives parallèles complique l'élucidation des relations de parenté lorsque les études sont basées uniquement sur la morphologie (Madsen et al. 2001, Springer et al. 2004b et Springer et al. 2007). Ainsi, des espèces partageant une niche semblable subiront des pressions évolutives comparables et présenteront certaines similitudes morphologiques ayant évolué indépendamment dans différents groupes. Ces homoplasies morphologiques sont souvent non-détectées et suggèrent des groupes qui ne sont pas monophylétiques.

Par exemple, il est maintenant reconnu que le groupe des édentés, tel que présenté par Benton (1988), est un groupe artificiel qui regroupe les xénarthres et les pangolins, entre autres dû à l'absence d'incisives et de molaires. Plusieurs autres groupes, formés sur la base de similitudes morphologiques (Novacek 1992, 2001, Shoshani & McKenna 1998, Asher et al. 2003), sont fortement contredits par les études moléculaires (Springer et al. 2004b, 2007). Citons: (1) les ongulés qui présentent un ou plusieurs sabots et qui regroupent une multitude d'ordres dont les périssodactyles et les artiodactyles (qui sont eux-mêmes un groupe artificiel puisqu'ils excluent les cétacés), (2) les archontes, caractérisés par un pénis pendant et qui incluent : les chauves-souris, les primates, les lémurs volants et les toupayes, (3) les volitantiens qui possèdent une membrane alaire entre les doigts et qui regroupent les chiroptères et les dermoptères, (4) les anagalides regroupant les rongeurs, les lagomorphes et les macroscelidés et (5) les insectivores (ex.: le tenrec et le hérisson) forment tous des groupes non-monophylétiques lorsqu'ils sont comparés aux plus récentes études (voir la Figure 1.3 qui illustre l'hypothèse phylogénétique moléculaire concernant l'histoire évolutive de ces différents groupes et espèces). De plus, les études basées sur la morphologie et les fossiles supposent une origine des mammifères beaucoup plus récente que les études moléculaires. En se basant sur de nouvelles données fossiles, Wible et al. (2007) ont estimé l'origine de la radiation placentaire à la limite du Crétacé-Tertiaire il y

a environ 65 Ma, alors que les études moléculaires récentes la situent entre 105 et 120 Ma (Biriinda-Emonds et al. 2007, Murphy et al. 2007, Nishira et al. 2009). Malgré le débat qui oppose les partisans des études morphologiques et moléculaires (voir Springer et al. 2004b, 2007), la méthodologie utilisée dans ma thèse repose uniquement sur des caractères moléculaires, étant donné leur accessibilité et le nombre croissant de séquences d'ADN disponibles.

1.2.2.2. Les phylogénies estimées à partir de caractères moléculaires

Les mammifères représentent un groupe intéressant pour l'étude des relations évolutives à l'aide de marqueurs moléculaires puisque de nombreuses controverses phylogénétiques restent à résoudre. D'un autre côté, certaines portions de la phylogénie des mammifères sont extrêmement bien soutenues et peuvent servir de modèles pour tester différentes approches phylogénétiques. Un nombre impressionnant d'études moléculaires ont été publiées durant les dernières années, ce qui a permis d'établir un consensus au sujet de la phylogénie des ordres de mammifères, que ce soit en analysant des gènes nucléaires ou mitochondriaux, ou encore, en combinant dans une même analyse les deux types de marqueurs (Madsen et al. 2001, Murphy et al. 2001a, b, Arnason et al. 2002, 2008, Delsuc et al. 2002, Jow et al. 2002, Lin et al. 2002b, Amrine-Madsen et al. 2003, Hudelot et al. 2003, Nikaido et al. 2003, Waddell & Shelley 2003, Reyes et al. 2004, Phillips & Pratt 2008). Plus récemment, de nombreuses études phylogénomiques ont également vu le jour. Celles-ci utilisent différents marqueurs phylogénétiques pour établir les relations de parenté, comme l'analyse génomique de séquences nucléotidiques (ex.: Hallström et al. 2007, 2008, Nikolaev et al. 2007, Nishihara et al. 2007, Wildman et al. 2007, Prasad et al. 2008), l'analyse d'insertions ou de délétions (ex.: Thomas et al. 2003, Murphy et al. 2007) ou encore l'analyse d'éléments transposables (ex.: Kriegs et al. 2006, Nishihara et al. 2006, Murphy et al. 2007, Waters et al. 2007, Nishihara et al. 2009).

Les monotrèmes représentent le groupe qui a divergé en premier au sein des mammifères. Quant à l'évolution des deux autres groupes, les marsupiaux et les placentaires, deux hypothèses principales ont été avancées (voir Phillips & Penny 2003). La plupart des études phylogénétiques soutiennent l'hypothèse du groupe Theria selon laquelle les marsupiaux représentent le groupe frère des placentaires. La deuxième hypothèse, appelée l'hypothèse Marsupionta, suggère que les marsupiaux

sont plutôt le groupe frère des monotrèmes (Gregory 1947). Bien que généralement réfutée par les études morphologiques et nucléaires, cette deuxième hypothèse a été appuyée par des séquences mitochondriales (Janke et al. 1996, 1997, 2002, Penny & Hasegawa 1997). Dans une étude subséquente, Phillips & Penny (2003) ont démontré que l'inférence d'une phylogénie en accord avec l'hypothèse Marsupionta était en fait un artéfact des méthodes de reconstruction utilisées, causé par un biais des fréquences de nucléotides entre les séquences mitochondriales de différentes espèces ainsi qu'un taux de substitution hétérogène entre les différentes partitions (ou gènes). D'autres études plus récentes qui ont analysé des séquences de gènes nucléaires viennent aussi appuyer l'hypothèse Theria (Baker et al. 2004, van Rheede et al. 2006, Kullberg et al. 2008, Prasad et al. 2008).

Les études nucléaires, mitochondriales et phylogénomiques sont majoritairement en accord sur la phylogénie des ordres de mammifères placentaires et leur division en quatre grands groupes monophylétiques: Xenarthra, Afrotheria, Laurasiatheria et Euarchontoglires (Fig. 1.3: voir Springer et al. 2004b et Springer & Murphy 2007 pour une revue). Laurasiatheria et Euarchontoglires sont des groupes frères formant le clade Boreoeutheria. Le groupe Euarchontoglires est subdivisé en deux clades : Glires et Euarchonta (Waddell et al. 1999a), ce dernier étant différent de l'hypothèse morphologique Archonta qui incluait également les chauves-souris (Gregory 1910, Novacek 1992). Cependant, certaines analyses phylogénomiques doutent de la validité du groupe Euarchontoglires (Kullberg et al. 2006, Cannarozzi et al. 2007, Hughes & Friedman 2007, Huttley et al. 2007). Il est à noter que la discipline de la phylogénomique est très récente et souvent limitée par des contraintes méthodologiques et informatiques dues au grand nombre de caractères à analyser. Pour compenser, un nombre restreint de taxons est analysé afin de réduire la taille de la matrice, au risque d'inclure des biais systématiques tel que l'attraction des longues branches. Les études mentionnées ci-haut, où le clade Euarchontoglires n'est pas retrouvé, ont probablement été affectées par l'attraction de longues branches à cause d'un faible nombre de taxons échantillonné, en particulier chez les rongeurs, qui sont reconnus pour leur taux d'évolution rapide (Philippe 1997, Lin et al. 2002b). D'ailleurs, de subséquentes études génomiques qui ont utilisé un échantillonnage taxonomique plus complet ont retrouvé le groupe Euarchontoglires (Nikolaev et al. 2007, Ranwez et al. 2007, Prasad et al. 2008).

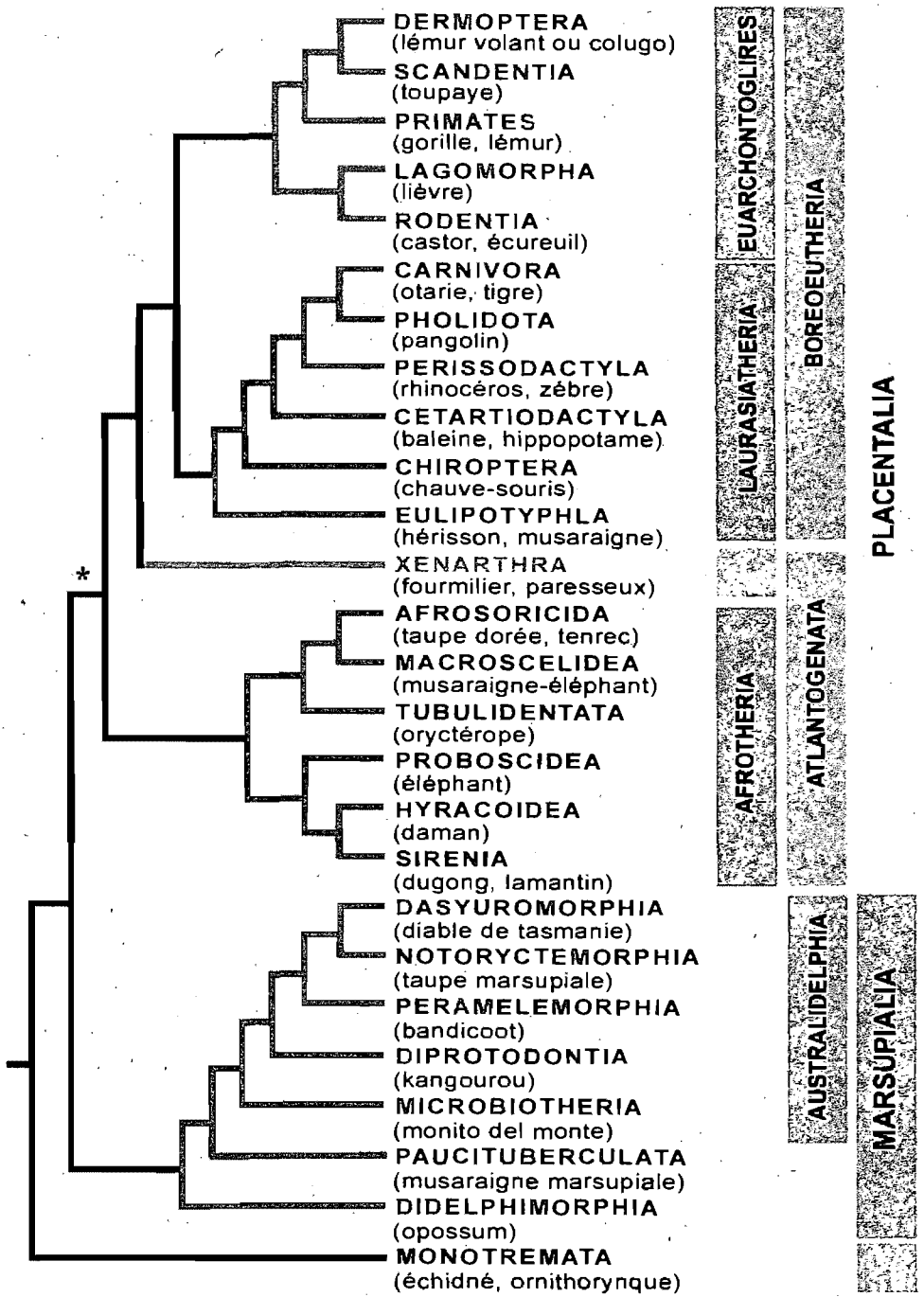

Figure 1.3. La phylogénie des 26 ordres de mammifères telle que suggérée par la majorité des études moléculaires.

D'après: Murphy et al. (2004), Springer et al. (2004b), et Springer & Murphy (2007). Les grands groupes de mammifères placentaires sont identifiés par différentes couleurs. Un ou quelques représentants sont énumérés entre parenthèses sous chacun des ordres. * Voir le texte pour une discussion de l'ordre de branchements à la base des placentaires.

Un aspect de la phylogénie des mammifères placentaires qui a été abondamment étudié récemment concerne l'ordre des embranchements à la base du groupe (voir Springer et al. 2007 et Wildman et al. 2007 pour une revue). Hedges et al. (1996) ont été les premiers à observer la concordance entre l'origine des groupes de placentaires et la tectonique des plaques. D'après Smith et al. (1994), la Laurasie (supercontinent reliant l'Eurasie et l'Amérique du Nord) se serait séparée de la Pangée il y a environ 140 Ma alors que le supercontinent Gondwana se serait divisé pour donner l'Afrique et l'Amérique du Sud, il y a environ 105 Ma. Des études de datation moléculaire de la diversification des placentaires ont permis de proposer plusieurs hypothèses quant à l'origine des groupes de mammifères. Puisque les groupes Afrotheria, Xenarthra et Boreoeutheria sont respectivement originaires d'Afrique, d'Amérique du Sud et de Laurasie, plusieurs scénarios liant l'évolution de ces groupes à la séparation et la dérive des continents sont possibles. Quatre hypothèses ont été émises à partir d'études moléculaires : la présence du clade (1) Afrotheria (Fig. 1.4A), ou (2) Xenarthra (Fig. 1.4B) à la base des autres mammifères placentaires ou encore (3) une division Boreoeutheria/Atlantogenata (Fig. 1.4C) ou finalement (4) une radiation simultanée des clades Afrotheria, Xenarthra et Boreotheria (Fig. 1.4D). Cependant, aucune de ces hypothèses ne s'accorde entièrement avec une origine purement septentrionale des mammifères, tel que soutenu par les analyses morphologiques (Archibald 2003, Hunter et al. 2006, Wible et al. 2007).

Selon la première hypothèse, il y a environ 105 Ma, le territoire de Gondwana s'est scindé ce qui aurait entraîné une première isolation du groupe Afrotheria (en Afrique) du reste des mammifères (en Amérique du Sud). Par la suite, une dispersion des placentaires de l'Amérique du Sud vers l'hémisphère Nord se serait produite, il y a environ 95 Ma, formant le groupe Boreotheria dans l'hémisphère Nord et laissant le groupe Xenarthra en Amérique du Sud (Fig. 1.4A : Murphy et al. 2001b, Amrine-Madsen et al. 2003, Nikolaev et al. 2007, Nishira et al. 2007). La deuxième hypothèse a été appuyée par l'insertion de deux rétroposons qui n'étaient pas présents chez les Xénathres et dans le groupe externe (Fig. 1.4B : Kriegs et al. 2006). Par contre, Kriegs et al. (2006) et Murphy et al. (2007) ont émis des doutes sur la validité statistique de l'analyse de ces deux rétroposons.

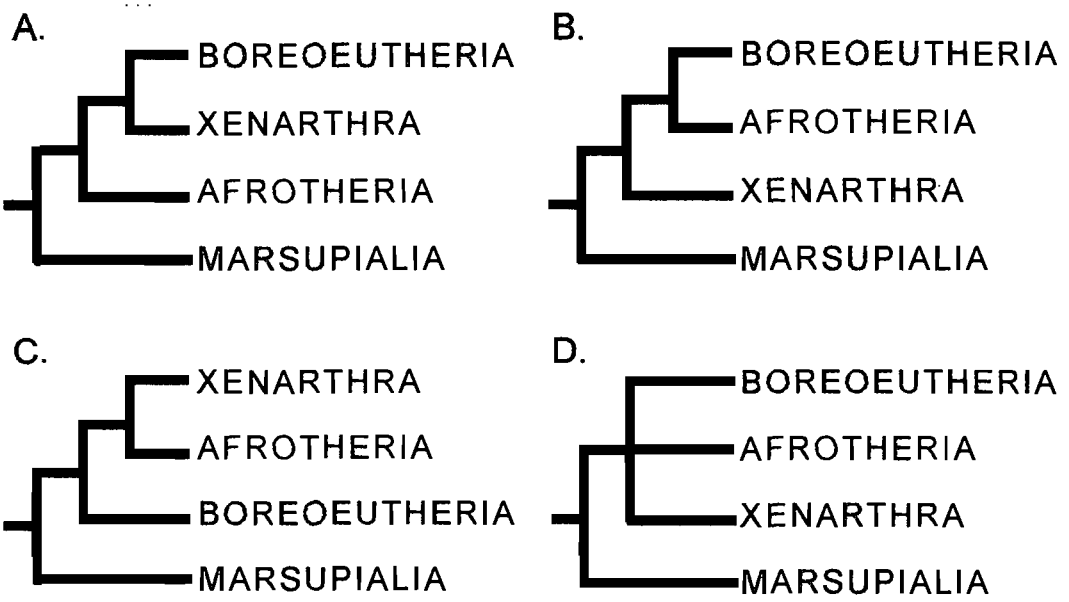

Figure 1.4. Les hypothèses de radiation des grands groupes de placentaires illustrées avec les marsupiaux comme groupe externe.
Boreoeutheria comprend les groupes Laurasiatheria et Archontoglires. Selon les hypothèses : A. Afrotheria, B. Xenarthra, ou C. Boreoeutheria ont divergé les premiers au sein des placentaires, ou alors D. une radiation simultanée de ces trois groupes a eu lieu.

La troisième hypothèse (Fig. 1.4C) suggère plutôt une première division des mammifères placentaires en deux lignées lorsque la Pangée s'est divisée en Laurasie au nord et Gondwana au sud, formant les clades Boreoeutheria et Atlantogenata (composé des Xénarthres et des Afrothériens) respectivement (Douady et al. 2002, Hallström et al. 2007, 2008, Kjer & Honeycutt 2007, Murphy et al. 2007, Waters et al. 2007, Wildman et al. 2007, Prasad et al. 2008). Tout récemment, une analyse génomique de rétroposons sur dix espèces de mammifères a identifié une vingtaine de locus informatifs pour chacune des trois premières hypothèses d'évolution, suggérant une quatrième hypothèse (Fig. 1.4D) c'est-à-dire, une radiation presque simultanée des trois lignées, Xenarthra, Afrotheria et Boreotheria (Churakov et al. 2009, Nishihara et al. 2009). De plus, l'intégration de nouvelles données paléogéographiques a permis à ces auteurs d'établir que les trois lignées auraient divergé, il y a 120 Ma, alors que les trois continents (correspondant aux endroits de diversification des trois lignées) réunis par les Détroits de Gibraltar et Brésilien étaient en cours de séparation.

À l'intérieur des quatre grands clades placentaires, certaines relations interordinales sont moins bien établies que d'autres. Au sein des Euarchontoglires, la position phylogénétique de l'ordre Scandentia est incertaine et, tout dépendant des études, il peut être groupe-frère des Dermoptères (Murphy et al. 2001a), ou encore groupe-frère des primates (Janecka et al. 2007, Hallström et al. 2008). Parmi le groupe Laurasiatheria, l'embranchement des ordres varie selon les analyses. Nishihara et al. (2006) proposent même un nouveau groupe : Pegasoferae, basé sur quatre insertions de rétroposons caractérisant les ordres Chiroptera, Perissodactyla et Carnivora. Une insertion supplémentaire caractérise le sous-groupe Perissodactyla/Carnivora. À l'intérieur du clade Eulipotyphla, les musaraignes se regrouperaient avec les tenrecs plutôt qu'avec les taupes (Douady et al. 2002, Waddell & Shelley 2003). De plus, un patron d'insertion de rétroposons supporte une phylogénie où l'ordre Eulipotyphla est basal au clade Laurasiatheria. Également, au sein du groupe Afrotheria, Nishihara et al. (2005) ont identifié deux insertions de rétroposons favorisant un groupe composé des Tubulidentata et Afrosoricida.

Les relations interordinales entre les sept ordres de marsupiaux ont fait l'objet d'un nombre d'études considérablement moins élevé que celles réalisées pour les mammifères placentaires. Se fondant sur des caractères morphologiques, Szalay (1982) avait séparé les ordres de marsupiaux en deux groupes : Australidelphia et Ameridelphia. Le groupe Australidelphia comprend les quatre ordres retrouvés en Australasie, soit Dasyuromorphia, Diprotodontia, Notoryctemorphia et Peramelemorphia ainsi que Microbiotheria dont l'unique représentant, le monito del monte, se retrouve en Amérique. Les deux autres ordres retrouvés en Amérique, Didelphimorphia et Paucituberculata formant le groupe Ameridelphia. Alors que le groupe Australidelphia forme un groupe monophylétique qui a été corroboré par de nombreuses études morphologiques, nucléaires et mitochondriales (Kirsch et al. 1991, 1997, Horovitz & Sánchez-Villagra 2003, Nilsson et al. 2003, 2004, Phillips et al. 2006, Beck 2008, Meredith et al. 2008a), Ameridelphia s'est révélé être un groupe paraphylétique puisque l'ordre Paucituberculata serait en fait plus proche du groupe Australidelphia qu'il ne l'est de l'ordre Didelphimorphia (Fig. 1.3: Nilsson 2003, 2004, Amrine-Madsen et al. 2003, Beck 2008, Meredith et al. 2008a). Les relations phylogénétiques entre les ordres du groupe Australidelphia sont aussi débattues avec, entre autres, une position interne des microbiothères au sein du groupe (Nilsson et al. 2003, 2004, Cardillo et al. 2004, Munemasa et al. 2006) ou encore une position basale

(Amrine-Madsen et al. 2003, Phillips et al. 2006, Beck 2008, Meredith et al. 2008a). La position des taupes marsupiales (Notoryctemorphia), des marsupiaux carnivores (Dasyuromorphia) ainsi que celles des bandicoots (Peramelemorphia) varie également selon les études. Les plus récentes études moléculaires favorisent un groupement de ces trois ordres formant un clade polyprotodonte australien (Phillips et al. 2006, Beck 2008, Meredith et al. 2008a).

1.2.2.3. La congruence entre les phylogénies mitochondriales et nucléaires

Bien que les résultats des récentes analyses nucléaires et mitochondriales (mt) soient en grande partie congruents, il n'en a pas toujours été ainsi. En effet, certains auteurs affirmaient que le taux d'évolution des séquences mitochondriales n'était pas adapté pour résoudre des questions évolutives anciennes telle que la radiation des mammifères (Curole & Kocher 1999, Springer et al. 2001, Corneli 2002). Le génome mitochondrial de la plupart des espèces de mammifères varie de 16 500 à 17 200 paires de base selon le nombre et la longueur des régions répétées présentes dans la région de contrôle (Arnason & Janke 2002). Cette région est la plus variable de la mitochondrie et comprend l'origine de réplication de ce génome. La mitochondrie est constituée de 37 gènes qui codent pour 22 ARN de transfert, 2 ARN ribosomaux et 13 protéines impliquées dans la chaîne de transport des électrons et la phosphorylation oxydative (Anderson et al. 1981, Chomyn et al. 1985, 1986). Des études phylogénétiques utilisant le génome mitochondrial ont obtenu des résultats venant en contradiction avec les quatre grands clades placentaires (ex.: Cao et al. 2000, Nikaido et al. 2000) mais également en contradiction avec les analyses nucléaires des marsupiaux (ex.: Nilsson et al. 2004). De plus, les analyses de séquences mitochondriales sont à l'origine d'affirmations plutôt surprenantes. Par exemple, « le cochon d'inde n'est pas un rongeur » (D'Erchia et al. 1996), « les hérissons et/ou les rongeurs sont les premiers à avoir divergé au sein des placentaires » (Cao et al. 2000, Mouchaty et al. 2000, Arnason et al. 2002) ou encore « les lémurs volants se regroupent avec les primates et donc ces derniers forment un groupe paraphylétique » (Arnason & Janke 2002, Arnason et al. 2002). Ces résultats sont contredits par la majorité des analyses de gènes nucléaires et peuvent être expliqués en partie par un taux d'évolution rapide (ex.: les rongeurs et les primates), un biais compositionnel entre les espèces, un modèle d'évolution inapproprié ou encore un échantillonnage inadéquat

des taxons (Huchon et al. 2002, Lin et al. 2002a, b, Nikaido et al. 2003, Phillips et al. 2006).

Ainsi, des études mitochondriales récentes ont démontré qu'il est possible de réconcilier les arbres nucléaires et mitochondriaux lorsqu'un modèle d'évolution approprié et un échantillonnage adéquat sont utilisés. Jow et al. (2002) et Hudelot et al. (2003) ont appliqué des modèles d'évolution spécifiques pour analyser des gènes mitochondriaux codant pour des ARN de transfert et ribosomaux. Leurs modèles tiennent compte de la structure secondaire de l'ARN, par exemple le motif « tige-boucle », où la dépendance entre les nucléotides varie selon leur position. Les bases situées dans la double hélice (ou tige) sont appariées entre elles, ce qui influence le taux et le patron de substitutions. Reyes et al. (2004) ont obtenu une phylogénie mitochondriale congruente aux phylogénies nucléaires en incluant des taxons supplémentaires pour briser les longues branches. Aussi, pour éviter un biais dans la composition de nucléotides ainsi qu'un biais dans l'usage des codons synonymes entre les différentes espèces, plusieurs études mitochondriales excluent la troisième position des codons des analyses puisque cette position évolue plus rapidement, et les premières positions synonymes pour la leucine (ex.: Reyes et al. 2004, Phillips et al. 2006, Arnason et al. 2008). Afin d'éviter l'exclusion d'un grand nombre de sites, Gibson et al. (2005) ont développé un modèle qui tient compte de l'hétérogénéité observée dans la fréquence des nucléotides T et C entre les espèces et qui cause un biais dans l'usage des codons synonymes. Leur modèle, le GTR3, est semblable au modèle GTR4 (*General Time Reversible four-state model*), mais seules trois catégories sont utilisées puisque les nucléotides C et T sont combinés en une seule catégorie Y (pour pyrimidine). En utilisant un modèle GTR3-4, où la première position des codons est analysée avec un modèle GTR3 et la deuxième position avec un modèle GTR4, ils ont obtenu une phylogénie mitochondriale en accord avec les phylogénies nucléaires.

1.2.2.4. Les phylogénies interfamiliales

S'il existe désormais un grand nombre d'études moléculaires incluant tous les ordres de mammifères, très peu d'entre elles ont entrepris l'analyse de toutes les familles de mammifères. Probablement freinées par le nombre relativement élevé de familles au sein de la classe des mammifères (c.-à-d. 150 familles) et de la faible proportion de séquences nucléaires pour certaines familles, la majorité des études phylogénétiques

se concentrent plutôt sur les relations à l'intérieur de certains ordres ou de certaines familles (ex.: l'ordre des cétacés: Xiong et al. 2009, et des cétartiodactyles: Agnarsson & May-Collado 2008; la famille des soricidés: Dubey et al. 2007, et des ursidés: Krause et al. 2008; ou encore, trois familles de chiroptères: Gu et al. 2008). Ces études, plus restreintes à l'échelle taxonomique, mais qui incluent souvent plus d'espèces par clades ont permis d'obtenir des résultats très intéressants : par exemple, la séparation des hippopotames des autres suiformes formant un groupe monophylétique avec les cétacés appelé Cetartiodactyla (Irwin & Arnason 1994, Ursing & Arnason 1998).

Pourtant, les études interfamiliales qui incluent la majorité des familles de mammifères pourraient révéler des relations qui n'apparaissent pas sur des phylogénies individuelles de chacune ou de quelques familles. Une des premières phylogénies moléculaires incluant de nombreuses familles de mammifères (42 familles) fut l'étude de type super-matrice de Murphy et al. (2001b) qui était basée sur 19 gènes nucléaires et trois gènes mitochondriaux. Indépendamment, Liu et al. (2001) ont entrepris de résoudre l'histoire évolutive de la majorité des familles de mammifères (86 familles) dans une étude regroupant 430 phylogénies sources combinées pour former un super-arbre. Malheureusement certaines critiques remettent en question la validité de leurs résultats (Springer & de Jong 2001, Gatesy et al. 2002). Parmi les lacunes observées, citons, entre autres, la non-indépendance des arbres sources, le manque de fiabilité de certains arbres sources, des prémisses injustifiées de monophylie au niveau des ordres, ainsi qu'un poids égal à tous les nœuds, peu importe leur support original. Étant donné la rivalité entre les partisans de l'approche de type super-arbre et super-matrice, Beck et al. (2006) a repris l'analyse de Liu et al. (2001) avec un échantillonnage plus complet (113 familles placentaires), une sélection plus stricte des arbres sources et l'absence de prémisse dictée *a priori* de groupement monophylétique. Contrairement à Liu et al. (2001), le super-arbre des familles de mammifères obtenu par Beck et al. (2006) s'accorde mieux au consensus moléculaire actuel des relations évolutives entre les placentaires. Avec l'accroissement exponentiel du nombre de séquences d'ADN disponibles et l'augmentation de la puissance informatique, les analyses de type super-matrice seront plus facilement réalisables, ce qui risque d'alimenter d'autant plus le débat super-matrice/super-arbre. À ce jour, on connaît la séquence mitochondriale complète de plus de 200 espèces de mammifères ainsi que les séquences de nombreux gènes nucléaires. D'ailleurs une étude mitogénomique de 103 espèces de mammifères (dont 80 familles) a permis de réunifier les phylogénies nucléaires et

mitochondriales en ce qui a trait aux relations évolutives à la base des mammifères (Arnason et al. 2008). Néanmoins, il est nécessaire de continuer et d'élargir l'effort de séquençage à des groupes moins bien représentés. Ainsi, seulement 96 des 150 familles de mammifères ont une espèce pour laquelle la séquence mitochondriale est disponible. De plus, aucun gène nucléaire n'a été séquencé pour la totalité de ces 96 familles.

1.3. ORGANISATION ET FONDEMENTS DE LA THÈSE

Dans le cadre de mon doctorat, j'ai mené à terme des études ayant comme axe de recherche la validation de méthodes phylogénétiques applicables en phylogénomique, c'est-à-dire adaptées à l'analyse d'un nombre élevé de jeux de données moléculaires combiné à une représentation taxonomique de plus en plus complète. Différentes approches ont été testées à l'aide de données simulées et empiriques, contribuant ainsi à la connaissance des relations évolutives entre les différentes familles de mammifères. La méthodologie utilisée pour atteindre les objectifs de recherche et les résultats obtenus sont présentés sous forme d'articles scientifiques correspondant aux différents chapitres de cette thèse. Ces articles répondent à des objectifs pouvant se diviser en trois catégories, représentant les volets principaux de mon doctorat. Ces trois parties et les objectifs qui y sont rattachés sont brièvement introduits ci-dessous. Finalement, une conclusion générale offre une vue d'ensemble des résultats obtenus dans les différents chapitres et discute de l'importance de mes recherches dans le contexte actuel de l'analyse phylogénomique.

1.3.1. La validation du test de CEMD à partir de matrices de distances ultramétriques et additives

Plusieurs tests statistiques ont été proposés pour tester la congruence entre différents jeux de données avant de les combiner dans une analyse phylogénétique des taxons. Malheureusement, ces tests ne sont pas adaptés aux analyses phylogénomiques et ont été vivement critiqués pour diverses raisons. Un test statistique, le test de CEMD (Congruence Entre des Matrices de Distance, ou CADM en anglais), a été proposé pour mesurer la congruence entre matrices de distances calculées à partir d'ensembles de caractères distincts mais portant sur les mêmes objets (Legendre & Lapointe 2004). Ce test est une généralisation du test de Mantel (Mantel 1967) et il est applicable à plus de

deux matrices. L'hypothèse nulle est définie comme étant l'incongruence de toutes les matrices de distance. Lorsque l'hypothèse nulle est rejetée, certaines ou toutes les matrices sont congruentes. Des tests d'incongruence *a posteriori* peuvent être effectués dans le but de déterminer quelles matrices sont incongruentes. Le premier objectif de ma thèse consistait donc à évaluer statistiquement la validité du test CEMD lorsqu'il est appliqué dans un cadre phylogénomique.

Dans les chapitres 2 et 3, des simulations de jeux de données ont été réalisées dans le but de déterminer l'erreur de type I et la puissance du test lorsqu'il est utilisé pour tester la congruence entre des matrices de distances ultramétriques (qui peuvent, par exemple, représenter des arbres phylogénétiques où toutes les espèces évoluent à un taux constant) et additives (qui sont obtenues à partir d'arbres phylogénétiques où les espèces présentent une hétérogénéité de leur taux d'évolution). Lorsque le test révèle que les différents jeux de données sont congruents, il est alors possible de les réunir dans une analyse combinée sous forme de super-matrice. Dans le cas contraire, il est préférable d'analyser séparément les jeux de données qui sont incongruents.

Puisque les résultats des chapitres 2 et 3 indiquent que le test de CEMD est une option valable pour tester la congruence entre plusieurs jeux de données, une application de ce test est également présentée au chapitre 5, où les séquences d'ADN des 12 gènes mitochondriaux du brin lourd (*H-strand*) de 102 espèces de mammifères sont analysées.

1.3.2. La validation de l'approche par taxons chimères à l'aide de simulations et de données empiriques provenant d'espèces de mammifères

Une approche utilisée pour éviter les données manquantes dans les super-matrices de séquences d'ADN est la construction de taxons chimères (Shoshani & McKenna 1998, Madsen et al. 2001, Murphy et al. 2001a, b, Scally et al. 2001, Asher et al. 2004, Springer et al. 2004a). Les super-matrices construites à partir de données phylogénomiques comportent un nombre important de données manquantes et donc l'approche des taxons chimères est de plus en plus utilisée (Delsuc et al. 2006, Telford 2007, Beck 2008; Bourlat et al. 2008, Duvall et al. 2008). Dans une étude empirique, Malia *et al.* (2003) ont analysé à nouveau la matrice de séquences utilisée par Madsen *et al.* (2001), mais en conservant les données manquantes plutôt que de former des

taxons chimères. Ils ont conclu que l'utilisation des taxons chimères introduit des relations évolutives qui ne sont pas appuyées par les données originales. Ils suggèrent donc d'analyser les matrices en incluant les données manquantes, même s'il en résulte une perte de résolution, plutôt que d'inférer de fausses relations de parenté. Cependant, à ce jour, aucune étude n'a été faite pour déterminer statistiquement la validité des relations évolutives obtenues suite à l'analyse de taxons chimères. Le second objectif de ma thèse vise donc à établir laquelle des deux approches, données manquantes ou taxons chimères, est préférable lors d'une inférence phylogénétique.

L'utilisation de simulations a l'avantage de permettre de tester des méthodes phylogénétiques sous différents modèles d'évolution et avec des paramètres prédéterminés et contrôlés. Alors que les modèles utilisés représentent, pour la plupart, une simplification des processus évolutifs réels, ils permettent néanmoins d'évaluer adéquatement l'exactitude des méthodes phylogénétiques, puisque l'histoire évolutive est connue (Hillis 1995, Huelsenbeck 1995). J'ai donc utilisé des simulations pour mesurer l'exactitude des arbres inférés avec ou sans taxons chimères, en prenant comme arbre de référence (ou « vrai » arbre) celui utilisé pour l'évolution des séquences d'ADN. Pour différentes conditions de simulations, j'ai calculé le pourcentage d'arbres ayant une topologie identique à celle du « vrai » arbre. Le quatrième chapitre expose les résultats obtenus quant à la performance de l'approche par taxons chimères qui est comparé à la performance de l'analyse des mêmes matrices où les données manquantes ont été conservées.

Alors que les chapitres cités précédement ont été réalisés à l'aide de données simulées, la méthodologie utilisée au cinquième chapitre utilise des données empiriques pour mesurer la performance de l'approche par taxons chimères. Les séquences d'ADN provenant d'individus ont évolué selon un modèle biologique qui leur est propre et dont la complexité risque de ne pas être adéquatement représentée lors de simulations. De plus, il est difficile de reproduire exactement toutes les subtilités évolutives qui peuvent se retrouver dans un jeu de données et qui dépendent entre autres des taxons choisis. Par exemple, certaines espèces évoluent plus rapidement que d'autres ou encore diffèrent dans leur composition moléculaire. En l'occurrence, l'approche par taxons chimères a également été testée en utilisant des séquences d'ADN d'espèces de mammifères pour lesquels le génome mitochondrial est disponible. Suite à l'alignement des séquences de 12 gènes mitochondriaux, des matrices

comprenant des taxons chimères ont été formées. Les résultats de ce cinquième chapitre permettent de corroborer ou de réfuter les résultats obtenus au chapitre précédent.

1.3.3. La comparaison des méthodes de types consensus et super-arbre pour inférer la phylogénie des familles de mammifères

Jusqu'à tout récemment, les études phylogénétiques regroupant tous les groupes de mammifères se sont limitées principalement à résoudre les relations interordinales (ex.: Madsen et al. 2001, Murphy et al. 2001a, b, 2007, Springer et al. 2004b, Nishihara et al. 2006, 2009, Springer et al. 2007, Hallström et al. 2008, Prasad et al. 2008) aux dépens des relations interfamiliales (Liu et al. 2001, Beck et al. 2006). Les analyses de type super-matrice menées spécifiquement pour inférer les relations interfamiliales mt au sein des mammifères sont encore rares. Notons entre autres l'étude mitochondriale d'Arnason et al. (2008) qui inclut 80 des 150 familles de mammifères, mais dont le but premier était d'augmenter l'échantillonnage d'espèces afin d'obtenir une phylogénie interordinale plus compatible aux phylogénies nucléaires. Dû au grand nombre de familles de mammifères et à une insuffisance du nombre de séquences d'ADN pour certains groupes, les études interfamiliales sont plutôt de type super-arbre: par exemple, l'étude de Liu et al. (2001) inclut 86 familles placentaires contemporaines alors que l'analyse de Beck et al. (2006) inclut les 113 familles placentaires décrites par Wilson & Reeder (1993). Bien que les méthodes de type super-arbre facilitent les méta-analyses, elles demeurent peu étudiées en comparaison aux méthodes « conventionnelles » d'analyse phylogénétique et un débat subsiste toujours quant à leur validité (Cotton & Wilkinson 2009, Steel & Rodrigo 2008). Par ailleurs, l'approche de congruence globale suggère que lorsque des méthodes de super-arbre et de super-matrice produisent des arbres congruents, cela permet une crédibilité accrue de cette estimation phylogénétique (Lapointe et al. 1999).

Le troisième objectif de ma thèse est de comparer l'approche de types consensus et super-arbre en utilisant comme référence une phylogénie mitochondriale des familles de mammifères inférée à partir d'une super-matrice. Pour construire la super-matrice, une espèce par famille a été sélectionnée pour laquelle la séquence mitochondriale complète était disponible dans la banque de données de GenBank. À partir des données moléculaires recueillies, différentes approches de consensus et de super-

arbre ont été comparées au résultat obtenu suite à l'analyse de la super-matrice. Quatre méthodes de consensus ont été testées: (1) le consensus strict (Sokal & Rohlf 1981, Page 1989), (2) le consensus majoritaire (MR: Margush & McMorris 1981, Swofford 1991), (3) le consensus majoritaire avec groupements compatibles (MRC), et (4) le consensus d'Adams (Adams 1972, 1986). Une description de ces méthodes est diponible dans Swofford (1991) et dans le chapitre 5. Les méthodes de type super-arbre qui sont comparées dans ce chapitre ont été choisies afin de refléter les différents critères d'optimisation possibles lors de la construction d'un super-arbre (revue par Bininda-Emonds 2004c). Par conséquent, cinq méthodes ont été sélectionnées: (1) *Matrix Representation with Parsimony* (MRP: Baum 1992, Ragan 1992a, b) qui est la méthode la plus couramment utilisée, (2) *maximum splits fit* (SFIT: Creevey & McInerney 2005) et trois méthodes de distances: (3) *Most Similar Supertree* (MSS: Creevey et al. 2004), (4) *Average Consensus* (AC: Lapointe & Cucumel 1997), et (5) *Super Distance Matrix* (SDM: Criscuolo et al. 2006). Alors que la majorité des méthodes de consensus et de super-arbre utilisent seulement la topologie des arbres sources (ex.: toutes les méthodes de consensus et les méthodes 1 à 3 de super-arbre), les méthodes AC et SDM tiennent compte des longueurs de branches et donc devraient proposer une solution plus résolue.

CHAPITRE 2:

ASSESSING CONGRUENCE AMONG ULTRAMETRIC DISTANCE MATRICES

Cet article est publié sous la référence :

Campbell V., Legendre P., & Lapointe F.-J. 2009. Assessing congruence among ultrametric distance matrices. Journal of Classification. 26: 103 - 117.

2.1. RÉSUMÉ

Récemment, un test de la Congruence Entre des Matrices de Distance (CEMD) a été développé. L'hypothèse nulle stipule une incongruence entre les matrices de données. Il a déjà été démontré que le test CEMD a une erreur de type I adéquate et une bonne puissance lorsque le test est appliqué à des matrices de dissimilarité. Dans cette étude, nous étudions la pertinence du test CEMD pour comparer des matrices de distance ultramétrique. Nous avons testé l'erreur de type I et la puissance du test sur des matrices de distance calculées à partir de dendrogrammes générés de façon aléatoire. Nous montrons que le test a une erreur de type I correcte et une bonne puissance. Afin de produire la distribution de référence nécessaire au test de signification de la statistique, des permutations simples (comme dans le test de Mantel) ou doubles (comme dans le test DPT) ont été utilisées. La puissance du test CEMD est identique quel que soit le type de permutation utilisé. Cette étude démontre clairement que le test CEMD peut être utilisé pour déterminer la congruence entre plusieurs dendrogrammes.

2.2. ABSTRACT

Recently, a test of Congruence Among Distance Matrices (CADM) has been developed. The null hypothesis is the incongruence among all data matrices. It has been shown that CADM has a correct type I error rate and good power when applied to independently-generated distance matrices. In this study, we investigate the suitability of CADM to compare ultrametric distance matrices. We tested the type I error rate and power of CADM with randomly generated dendrograms and their associated ultrametric distance matrices. We show that the test has correct type I error rates and good power. To obtain the significance level of the statistic, a single (as in the Mantel test) or a double (as in the double permutation test, DPT) permutation procedure was used. The power of CADM remained identical when the two permutation methods were compared. This study clearly demonstrates that CADM can be used to determine whether different dendrograms convey congruent information.

2.3. INTRODUCTION

Often, in classification studies, different sets of variables are used to derive dendrograms for the same set of objects. Depending on the set of variables, classifications may differ. Therefore, it is important to know to which extent the information conveyed by each set of variables is congruent to the others. Also, when different classifications of the same objects are available but the information on the variables used is not, assessing the degree of resemblance or congruence of different classifications may be of interest.

Congruence or incongruence tests, depending on how the null hypothesis is postulated, have been extensively studied. Planet (2006) has recently classified congruence tests in two categories: those based on character information and those based on tree shape or topology. Character congruence tests compare the fit of the data on two competing trees (e.g., ILD: Mickevich & Farris 1981, Farris et al. 1994; Templeton Test: Templeton 1983; and T-PTP: Faith 1991). These tests will not be reviewed further here, given that they present a different approach than the one discussed in this study. In contrast, topological congruence tests compare the branching pattern (topology) of different trees without considering the underlying data. Such tests are based on numerical measurements of topological difference obtained from indices calculated on consensus trees (e.g., Consensus Fork Index: Colless 1980; Rohlf Consensus Index: Rohlf 1982) or from tree distances (e.g., Partition Metric: Robinson & Foulds 1981; Quartets Distance: Estabrook et al. 1985; and Path Difference Metric: Steel & Penny 1993). Significance testing is generally possible by comparing the statistic to a reference distribution generated by permutations (e.g., Steel & Penny 1993) or by using non-parametric bootstrap of the original data (Page 1996).

Along with different indices that have been proposed to quantify the similarity between dendrograms, a classical approach is to calculate a cophenetic (or matrix) correlation coefficient between two ultrametric matrices representing dendrograms (Sokal & Rohlf 1962). If the dendrograms come from independent data tables, the null hypothesis of a correlation equal to zero can be tested using a Mantel's generalized permutation test strategy, where only the object labels are permuted (Mantel 1967). Amongst other, a double-permutation test (DPT: Lapointe & Legendre 1990), which takes into account the topology, label positions and cluster heights of the dendrograms, has also been

proposed (Lapointe & Legendre 1995; see also Podani 2000). It has been shown that only DPT provides correct rates of type I error when a correlation between a pair of dendrograms is used as the test statistic (Lapointe & Legendre 1995).

Although many congruence tests were developed in a phylogenetic context, they are often used in other fields such as ecology, anthropology, archaeology, sociology and classification (Legendre & Lapointe 2004). Unfortunately, the majority of these tests only apply to the comparison of two datasets (or matrices) at a time. Legendre & Lapointe (2004) described a test of congruence among distance matrices (CADM) that is applicable to more than two matrices. Based on Kendall's W concordance statistic, CADM is an extension of the Mantel test that can be used to assess congruence of multiple matrices. CADM presents several advantages with respect to other congruence tests. (1) The statistic is calculated directly from the distance matrices; thus different types of data can be compared if converted to distance matrices using an appropriate function. (2) The matrices can be weighted differentially if needed. (3) *A posteriori* tests can be performed to discriminate incongruent from congruent matrices. Previous simulations have shown that the global and *a posteriori* CADM tests have a correct rate of type I error and good power when applied to independently-generated distance matrices (Legendre & Lapointe 2004). In this study, we have tested the type I error rate and power of the global and *a posteriori* tests of CADM using randomly generated dendrograms and their associated ultrametric distance matrices. To assess the significance of the statistic, we tested two different permutation procedures: a simple Mantel's permutation test (Mantel 1967) and a double-permutation test (DPT: Lapointe & Legendre 1990).

2.4. CADM TEST

The null hypothesis (H_0) for the global CADM test is the incongruence of all distance matrices (Legendre & Lapointe 2004). That is, matrices are statistically independent from each other and convey distinct information about the relationships among the objects under study. Rejecting H_0 indicates that at least two matrices contain congruent information. In those cases, *a posteriori* CADM tests can be performed to determine the contribution of each matrix to the overall congruence. *A posteriori* tests can be used to identify incongruent and congruent matrices in a set, but it does not specify the pairs or groups of congruent matrices. To this end, complementary Mantel tests based upon

ranks can be used. Following that, congruent matrices can be combined in a classification analysis. A summary of the computations to perform the CADM test follows:

1. The upper off-diagonal section of each distance matrix is unfolded and written into a vector corresponding to row *i* in a worktable.
2. The entries of each row are transformed into ranks according to their values.
3. The sum of ranks R_j is calculated for each column *j* of the table.
4. The mean $\overline{R}$ of all R_j values is calculated.
5. Kendall's coefficient of concordance (*W*) is computed using the following formula:

$$W = \frac{12S}{p^2(n^3 - n) - pT}$$

where *p* is the number of matrices, n is the number of objects in each matrix, *S* is obtained using:

$$S = \sum_{j=1}^{n} (R_j - \overline{R})^2$$

and *T* is a correction factor for tied ranks:

$$T = \sum_{k=1}^{m} (t_k^3 - t_k)$$

in which t_k is the number of tied ranks for each (*k*) of *m* groups of ties.

6. *W* is transformed into a Friedman's χ^2, which is a pivotal statistic appropriated for testing, using the following formula:

$$\chi^2 = p(n-1)W$$

The observed Friedman's χ^2 (χ^2_{ref}) is tested against a distribution of the statistic obtained under permutation (χ^{2*}). For the global CADM test, all matrices are permuted at random, whereas for *a posteriori* tests only the matrix tested is permuted. After p_n (*n* permutations), the one-tailed probability of the data under H_0 is computed as the number of χ^{2*} values greater than or equal to χ^2_{ref} divided by (p_n -1). In *a posteriori* comparisons, the p-value should be adjusted to maintain an adequate experimentwise error rate using a method designed specifically to correct for multiple testing (e.g., Holm 1979). Two different permutation models were compared in this study (see section 2.5). More details about the CADM method can be found in Legendre & Lapointe (2004).

2.5. SIMULATION PROCEDURE

Computer simulations were performed to assess the Type I error (α) rate and power of CADM when applied to ultrametric distance matrices. The type I error rate is the probability of incorrectly rejecting a true H_0 and should not be larger than the nominal significance level (α) of the test (Edgington 1995). The type II error (β) rate refers to the probability of failing to reject a false H_0. The power of the test is the rate of rejection of a false H_0 (i.e., 1 - β).

2.5.1. Global CADM test

We generated, at random, independent ultrametric distance matrices (IM) representing dendrograms, according to the completely random ultrametric matrix algorithm described by Lapointe & Legendre (1991, section 7). To examine a range of different parameter values, the number of objects within each matrix (n = 5, 10, 20 and 50) as well as the number of independent ultrametric matrices (IM = 2, 3, 4, 5 and 10) varied. Comparing IM corresponds to a situation where H_0 is "true" by construct (i.e., all the dendrograms are incongruent, see Figure 2.1A). To estimate the type I error rate, the rejection rate (i.e., the proportion of replicates for which the "true" H_0 was rejected) was calculated at different significance levels (α = 0.01, 0.05, 0.10, 0.20, 0.30, 0.40, 0.50, 0.60, 0.70, 0.80, 0.90) along with its 95% confidence interval (CI) computed according to a binomial distribution. A thousand replicate simulations were performed in each case. For each replicate, 999 random permutations of the dendrograms were computed to construct the reference distribution for significance testing. Two different permutation models were compared to obtain the reference distribution: Mantel (χ^{2*}_{M}) and DPT (χ^{2*}_{DPT}). For this study, we included the DPT procedure as an option in the CADM program (available at www.bio.umontreal.ca/casgrain/en/labo/cadm.html).

To estimate the power of CADM, congruent ultrametric distance matrices (CM) were generated (see Fig. 2.1B). CM are partially similar distance matrices that were generated by permutation (described below) of an original random ultrametric distance matrix. Figure 2.1 illustrates the steps involved in simulations to test the type I error and power of CADM with a set of five ultrametric distance matrices. Two different permutation procedures were used to generate CM. (1) A fixed number of randomly chosen objects were permuted on the dendrogram, corresponding to the permutation of

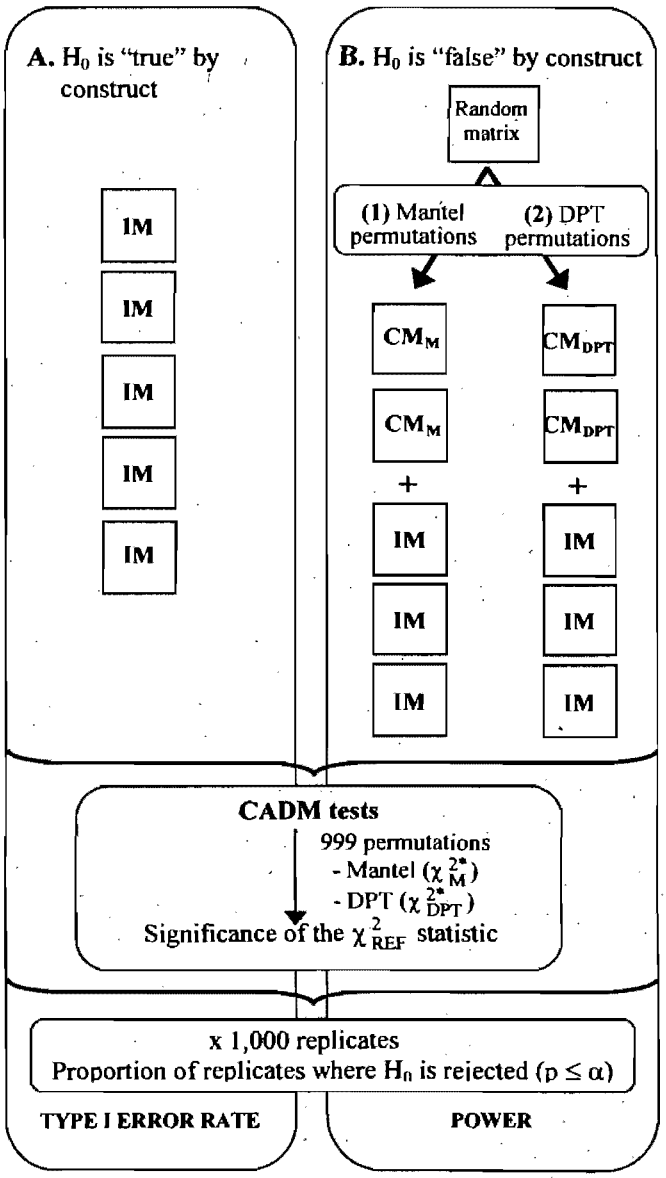

Figure 2.1. Flowchart of the protocol used to estimate type I error rate and power of CADM, for five ultrametric distance matrices.

A. H_0 is "true" by construct, and it includes five incongruent ultrametric distance matrices (IM). B. H_0 is "false" by construct, and it includes two partially similar matrices (CM) and three randomly generated matrices (IM). (1) In Mantel type permutations, CM were generated by permuting rows and columns of an initial matrix (CM_M). (2) For DPT, CM were generated by permuting rows and columns as well as cluster heights of an initial matrix (CM_{DPT}). CADM tests were performed on each set separately to estimate the type I error rate in A and power in B.

rows and columns (labels) within the ultrametric matrix. This is similar to the type of permutation that is performed in the Mantel test but restricted to some objects only (CM_M). (2) A fixed number of randomly chosen objects and cluster heights were permuted so that the dendrogram topology and the objects were permuted. This second permutation approach to construct CM is similar to that used in the DPT test but restricted to some objects only (CM_{DPT}). The generated CM were more or less congruent depending on the number of objects that were permuted (see Fig. 2.2). For power simulations, the proportion of permuted objects was identical regardless of the matrix sizes (i.e., 40% with n_{perm} = 2, 4, 8 or 20 for matrices of size n = 5, 10, 20 and 50 respectively). In each trial, the total number of distance matrices was fixed to either five or ten, but the number of CM and IM varied. When CM = 0 (IM = 5 or 10), H_0 is "true", otherwise H_0 is "false" by construct (CM ≥ 2). The power of the test, which corresponds to the proportion of replicates where H_0 is rejected when false, was calculated for each combination of parameters.

For each replicate of the CADM test, 999 random permutations were computed to estimate the reference distribution using the Mantel (χ^{2*}_{M}) and DPT (χ^{2*}_{DPT}) randomization procedures. H_0 was rejected when χ^2_{ref} was greater than or equal to 95% of the χ^{2*} (which corresponds to a one-tailed test with an alpha level of 5%). The rejection rate of H_0 (out of 1000 replicates) was calculated along with its 95% confidence interval (CI).

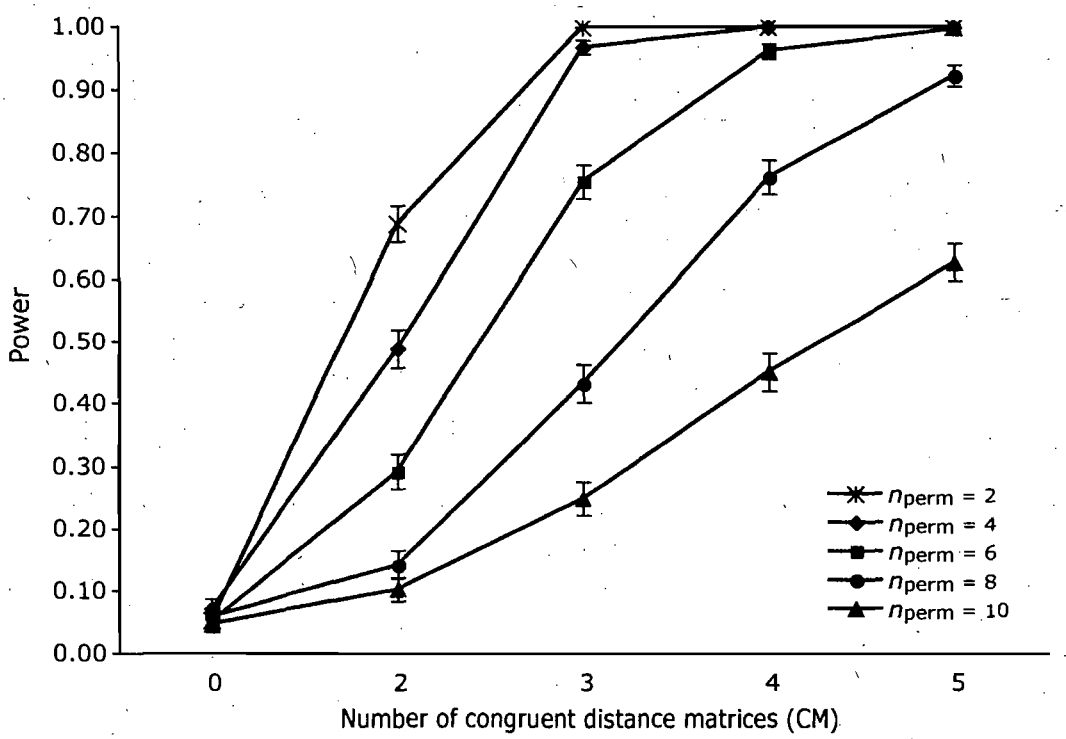

Figure 2.2. Estimated power (mean and 95% CI) obtained in simulations of global CADM tests using Mantel permutations, with different levels of congruence.
When CM = 0, all matrices are incongruent, thus H_0 is "true" by construct. For CM ≥ 2, H_0 is "false". For $\alpha = 0.05$, and n = 20, different levels of congruence among CM were achieved by permuting a different number of objects (n_{perm} shown with different symbols).

2.5.2. *A posteriori* CADM tests

Simulations have also been performed to assess the type I error rate and power of *a posteriori* CADM tests. The H_0 in such cases is the incongruence of the matrix subjected to the test with respect to all other matrices. Therefore, only the matrix subjected to the test is permuted. The sets of five ultrametric distance matrices generated to assess power of the global test were also used for *a posteriori* CADM tests (n = 5, 10 and 20). Again, the rejection rate of H_0, out of 1000 replicates, was calculated along with its 95% confidence interval (CI) for an alpha level of 0.05. For each replicate of the CADM test, 999 random permutations were computed to estimate the reference distribution using Mantel (χ^{2*}_{M}) and DPT (χ^{2*}_{DPT}) randomization procedures.

2.6. SIMULATION RESULTS

The results in Table 2.1 show that the CADM test underestimated the number of cases where H_0 should have been rejected when IM = 2 and n = 5; the 95% CI of the type I error rate did not include the nominal significance level (α) when compared to a χ^{2*}_{M} distribution. Similar results were observed when using a χ^{2*}_{DPT} distribution; however the type I error rate and its 95% CI were closer to the nominal values. Nevertheless, a test whose error rate under H_0 is lower than the alpha level remains valid (Edgington 1995). When IM = 2 and n > 5 or when IM was larger than two (i.e., IM = 3, 4, 5 and 10), the global CADM test had an adequate estimated Type I error rate for both types of permutations. Type I error rates obtained with 10 IM (n = 5, 10, 20 and 50) are shown in Figure 2.3. The 95% CI of the rejection rates included the nominal significance level (α) in nearly all cases.

Table 2.1. CADM type I error rates obtained for the global tests on pairs of ultrametric distance matrices (IM = 2), for different numbers of objects (n).
The corresponding 95% confidence intervals are in parentheses.

Permutation models	n	Significance levels		
		0.01	0.05	0.10
Mantel	5	0.0003 (0.0002 – 0.0005)	0.033 (0.032 – 0.035)	0.079 (0.077 – 0.081)
	10	0.001 (0.0005 – 0.002)	0.053 (0.049 – 0.058)	0.100 (0.094 – 0.106)
	20	0.010 (0.008 – 0.013)	0.050 (0.044 – 0.056)	0.100 (0.092 – 0.109)
DPT	5	0.008 (0.007 – 0.009)	0.042 (0.040 – 0.043)	0.095 (0.092 – 0.097)
	10	0.009 (0.007 – 0.011)	0.050 (0.045 – 0.054)	0.100 (0.095 – 0.110)
	20	0.013 (0.010 – 0.016)	0.054 (0.048 – 0.060)	0.100 (0.093 – 0.110)

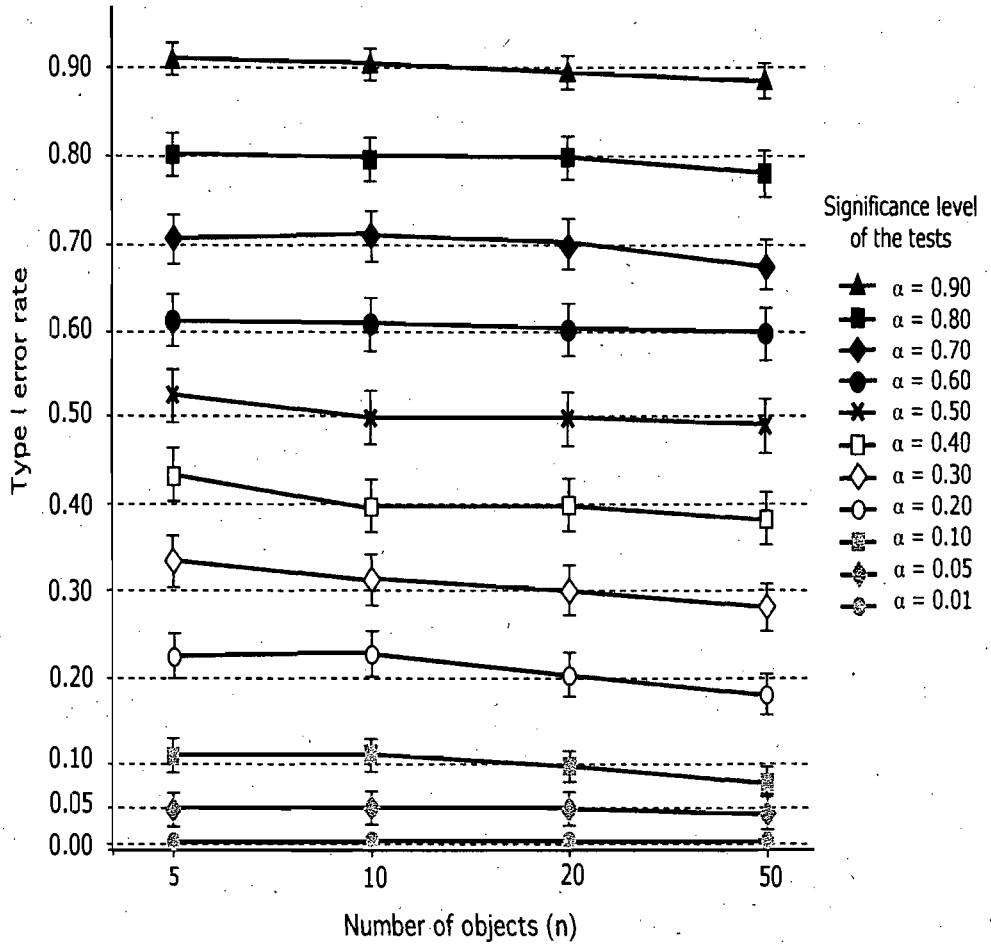

Figure 2.3. Type I error rates obtained in simulations of global CADM tests using Mantel permutations.

Results are shown for tests involving ten IM and varying numbers of objects (n = 5, 10, 20 and 50) along with different significance levels (α). The corresponding 95% CI (bars) are provided for each rejection rate (based on 1000 replicates).

Power was estimated by calculating the proportion of replicates where H_0 was rejected when H_0 was "false" by construct. Simulation results were nearly identical when CADM was tested using CM_M or CM_{DPT}, thus only the CM_M results will be reported here. Power obtained with CADM when different numbers of matrices were included in the analysis and for matrices with varying number of objects is shown in Figure 2.4. For IM + CM = 5 (Fig. 2.4A) and IM + CM = 10 (Fig. 2.4B), an increase in power was observed with (1)

an increase in the number of objects and (2) an increase in the number of CM relative to the total number of matrices. For sets of five matrices, a power of 1.0 was obtained only when matrices of 50 objects were used, whereas a power of 1.0 was obtained with smaller size matrices (i.e., 20 objects) for sets of ten matrices. Thus, power was higher when the total number of matrices included in the analysis was larger. Also, a comparison of Figures 2.4A and 2.4B reveals that for a given number of CM, power was higher when the number of IM was lower. Hence, it was easier to detect four congruent matrices out of five than four out of ten. This trend was accentuated when the size of the matrices increases.

Simulation results of *a posteriori* CADM tests are presented in Figure 2.5. In *a posteriori* comparisons, only the matrix that was subjected to the test was permuted. Therefore, rejection rates were obtained for each matrix permuted individually. For simplicity, power is only shown for matrix number 1 (i.e., only the first matrix was permuted, Fig. 2.5A) and matrix number 5 (i.e., only the fifth matrix was permuted, Fig. 2.5B). When H_0 was "true" (CM = 0), the rejection rate for the permuted matrix was close to 0.05, which was expected at the α level of 0.05 used to perform the tests. When CM $\geq$ 2, the rejection rate was greater than 0.05 when the permuted matrix was a CM but it was near 0.05 when the permuted matrix was an IM. Consequently, matrix number 1, which was congruent by construct with matrix number 2 in all cases except when CM = 0, showed a rejection rate greater than 0.05 when CM = 2, 3, 4 or 5 (Fig. 2.5A). In contrast, matrix number 5, which was congruent with the other matrices in the set only when CM = 5, showed a rejection rate greater than 0.05 only when CM = 5 (Fig. 2.5B). Figure 2.5 illustrates that *a posteriori* CADM tests have an accurate type I error rate when tested at an α level of 0.05, regardless of the number of CM versus IM, as observed when H_0 was "true" by construct. Similarly to the results obtained for the global CADM test, power was good and increased with the number of objects and with the number of CM.

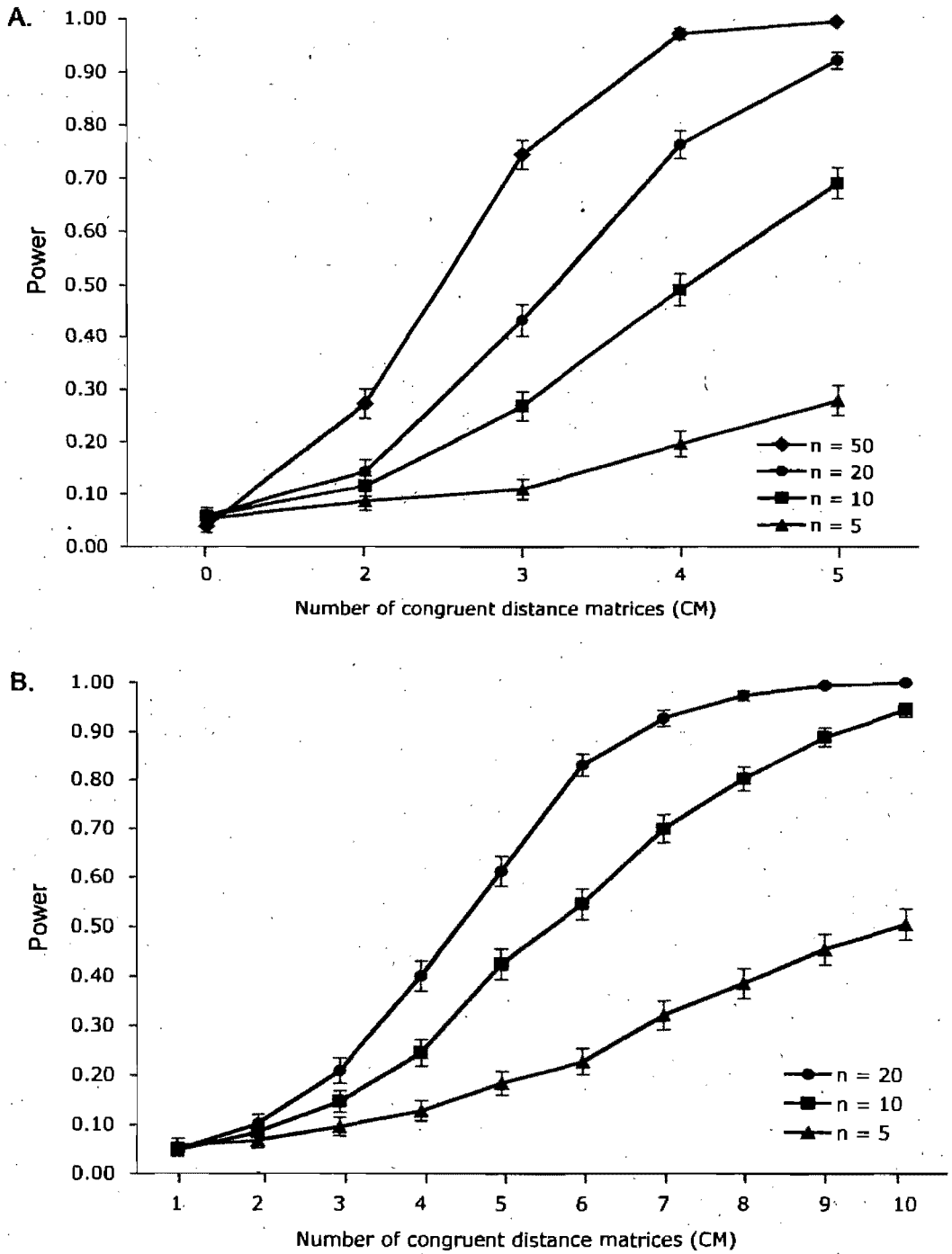

Figure 2.4. Estimated power (mean and 95% CI) obtained in simulations of global CADM tests using Mantel permutations, for different numbers of objects (n).

When CM = 0, all matrices are incongruent and thus H_0 is "true" by construct. For CM ≥ 2, H_0 is "false" (at $\alpha = 0.05$). A. Five distance matrices (IM + CM = 5) in each set with different numbers of objects. B. Ten distance matrices (IM + CM = 10) in each set with different numbers of objects.

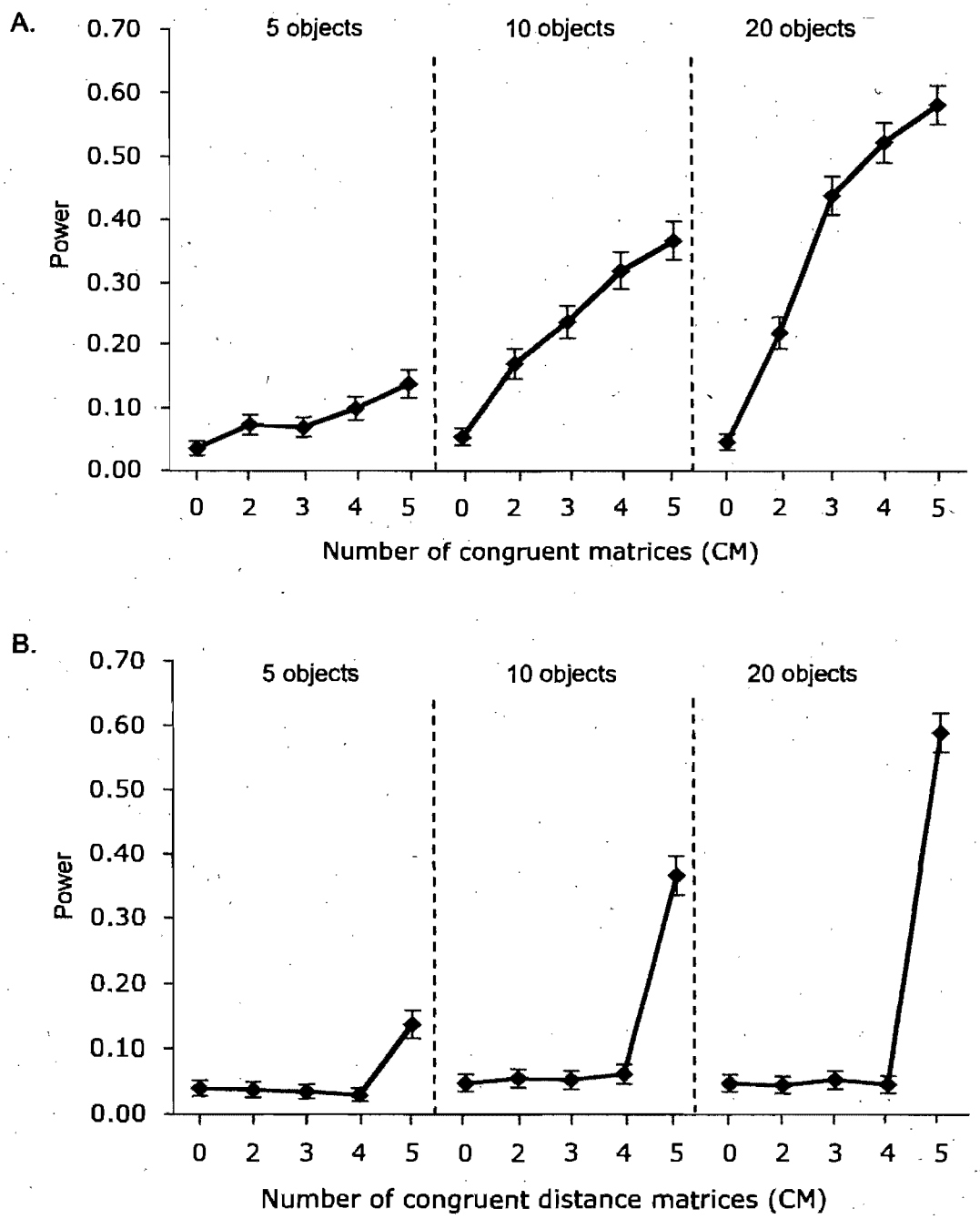

Figure 2.5. Estimated power (mean and 95% CI) obtained in simulations of *a posteriori* CADM tests, for different numbers of objects (n).

CM is the number of congruent matrices (in all cases IM + CM = 5). When CM=0, all matrices are incongruent and thus H_0 is "true" by construct. For CM ≥ 2, H_0 is "false", at a level α = 0.05. CM are numbered so that when CM = 2, matrices 1 and 2 are congruent. The symbols correspond to rejection rates of H_0 obtained when a given matrix is permuted. A. Matrix number 1 is permuted. B. Matrix number 5 is permuted.

2.7. DISCUSSION

CADM represents a powerful approach to test whether some matrices are incongruent to others. As opposed to the Mantel test, which compares matrices in a pairwise fashion, the global CADM test allows for comparisons among multiple matrices in a single analysis, without the need for a multiple testing correction (Legendre & Lapointe 2004). Our results support and generalize those of Legendre & Lapointe (2004), where the type I error and power of CADM was tested using random distance matrices. The results presented in this paper clearly show that CADM can be used to assess congruence among ultrametric distance matrices associated to dendrograms. The simulation results revealed that both the global and *a posteriori* CADM tests have correct type I error rates and good power when tested on ultrametric distance matrices, using either Mantel permutations or DPT. In classification studies, it can be used to determine if dendrograms defined on the same set of objects are congruent and thus support a similar classification. In cases where some dendrograms are incongruent, *a posteriori* tests can be used to determine which ones provide different information about the classification. Incongruent dendrograms can thus be compared to identify conflicting parts of the classification. Otherwise, dendrograms can be combined to derive a global classification using a consensus method.

The difference between the Mantel and the DPT randomization procedures rely on different aspects of the dendrogram being permuted (Lapointe & Legendre 1995). The Mantel procedure permutes the label positions on a fixed topology whereas DPT also permutes cluster heights, thus changing the tree topology. Therefore, the permutation set is larger when using DPT as a randomization procedure. Hence, for five objects, there are 60 different labelings of a topology, while 180 different dendrograms are possible when cluster heights are also randomized. For the global CADM significance test, all the dendrograms are randomized and the number of permutation possibilities increases exponentially with the number of dendrograms compared. Lapointe & Legendre (1995) have suggested that the DPT test might be more appropriate to compare dendrograms since a statistical bias may be introduced when using Mantel randomization, which samples only a subset of the reference distribution. While evaluating the Mantel test and DPT as testing procedures to compare correlations between random pairs of dendrograms, Lapointe & Legendre (1995) have shown that the Mantel test was more conservative than DPT and that only the latter test provided

unbiased type I error. They concluded that the Mantel randomization procedure was more likely to overlook congruent matrices. However, their study did not compare matrices that included more than five objects.

In our study, a Friedman's χ^2 statistic was used instead of a correlation coefficient because the test of concordance compares several matrices simultaneously, and matrices of different sizes were tested. Similarly to Lapointe & Legendre (1995), DPT randomization provided an improved type I error rate over the Mantel permutation when matrix pairs of five objects where compared. However, our results suggest that both permutation methods produce unbiased type I error rates when larger matrices ($n > 5$) are compared or when more than two matrices are included in the analysis, probably because a relatively small subset of the tree space is sampled with both types of randomizations when using 1000 permutations.

No significant difference in power was observed between the Mantel and DPT randomization procedures. Power curves were similar to those obtained by Legendre & Lapointe (2004) where CADM was tested with random distance matrices. That is, an increase in power was observed with (1) an increase in matrix size and (2) an increase in the number of CM relative to the total number of matrices. Also, for a given number of CM, power was higher when the number of IM was lower. In this study, good power was generally achieved, even though 40% of objects were permuted in each matrix. Power was further increased when comparing more congruent CM, i.e., CM that were generated using fewer permuted objects (as shown on Fig. 2.2). Although DPT is included as an option in the CADM program (see the method section), we recommend to use Mantel permutations even when comparing ultrametric distance matrices since it is less time consuming and performs identically to the DPT randomization, except in the particular case of two IM and five objects.

In an effort to further demonstrate the utility of CADM in different fields, we are currently testing its performance with additive distance matrices corresponding to phylogenetic trees. Further development may also include a generalization to allow comparisons of partially overlapping trees and matrices, which could be used for supertree methods.

2.8. ACKNOWLEDGMENTS

We would like to thank the members of the LEMEE (Laboratoire d'Écologie Moléculaire et d'Évolution) of Université de Montréal for their constructive comments on a prelirminary version of this manuscript as well as three anonymous reviewers. This study was supported by NSERC and FQRNT scholarships to VC and by NSERC grants OGP0007738 to PL and OGP0155251 to FJL.

CHAPITRE 3:

THE PERFORMANCE OF THE CONGRUENCE TEST AMONG DISTANCE MATRICES (CADM) IN PHYLOGENETIC ANALYSIS

Cet article sera soumis prochainement:

Campbell V., Legendre P., & Lapointe F.-J. The performance of the Congruence test Among Distance Matrices (CADM) in phylogenetic analysis. Journal of Mathematical Biology.

3.1. RÉSUMÉ

CEMD est un test statistique utilisé pour estimer le niveau de Congruence Entre des Matrices de Distance. Des études de simulations ont démontré que l'erreur de type I est adéquate et que la puissance est bonne lorsque le test est appliqué à des matrices de dissimilarité et de distance ultramétrique. Contrairement à la plupart des tests de congruence utilisés en analyse phylogénétique, l'hypothèse nulle suppose l'incompatibilité des jeux de données. Dans cette étude, nous avons effectué des simulations pour évaluer l'erreur de type I et la puissance du test de CEMD lorsqu'il est appliqué à des matrices de distance additive qui représentent des arbres phylogénétiques. Des séquences d'ADN de diverses longueurs ont été simulées sur des arbres de différentes tailles, générés au hasard et avec plusieurs paramètres d'évolution. Nos résultats ont montré que le test a une erreur de type I appropriée et une bonne puissance. La puissance augmente avec le nombre d'objets (taxons), le nombre de matrices congruentes ainsi que le degré de congruence entre les matrices de distance. Le test CEMD s'avère donc un candidat idéal pour déterminer la congruence des jeux de données avant de les combiner dans une analyse phylogénétique ou phylogénomique.

3.2. ABSTRACT

CADM is a statistical test used to estimate the level of Congruence Among Distance Matrices. It has been shown to have a correct rate of type I error and good power when applied to dissimilarity matrices and to ultrametric distance matrices. Contrary to most other tests of congruence used in phylogenetic analysis, the null hypothesis assumes the incongruence of the data matrices. In this study, we performed computer simulations to assess the type I error rate and power of the test when applied to additive distance matrices representing phylogenies, under a wide range of conditions. DNA sequences of different lengths were simulated on randomly generated trees of varying sizes, and under different evolutionary conditions. Our results showed that the test has an accurate type I error rate and good power. As expected, power increased with the number of objects (i.e., taxa), the number of congruent matrices under comparison and the level of congruence among distance matrices. Based on these results, CADM is an ideal candidate to test incongruence in phylogenomic studies where numerous datasets are analyzed simultaneously.

3.3. INTRODUCTION

In phylogenetics, data matrices are assembled and analyzed to infer evolutionary relationships among species or higher taxa. Depending on the study, character-state or distance matrices may be used and several different types of data may be available to estimate the phylogeny of a particular group (Swofford et al. 1996). An increasing number of phylogenomic studies are published for datasets including more than 100 genes (e.g., Lerat et al. 2003, Rokas et al. 2003b, Driskell et al. 2004, Philippe et al. 2005a, Fitzpatrick et al. 2006, Nishihara et al. 2007, Wildman et al. 2007, Dunn et al. 2008, Zou et al. 2008). Whereas character state data (e.g., DNA sequences) are commonly used for parsimony, maximum likelihood or Bayesian analyses, distance methods can be selected as an alternate option to decrease computing time when analyzing large datasets or else, in comparative studies where primary data are not available.

Different approaches have been proposed as to how to analyze the growing amount of information that may originate from different sources. The total evidence approach (Kluge 1989), also called character congruence approach (*sensu* Mickevich 1978) or combined analysis (*sensu* de Queiroz 1993), combines different datasets in a single supermatrix (Eernisse & Kluge 1993, Kluge & Wolf 1993, Gatesy et al. 2004, de Queiroz & Gatesy 2007). The taxonomic congruence approach (*sensu* Mickevich 1978), or consensus approach (*sensu* de Queiroz 1993), analyzes each matrix separately, and the resulting trees are combined *a posteriori* using a consensus (Swofford 1991, Farris et al. 1995, Miyamoto & Fitch 1995, Huelsenbeck et al. 1996a) or a supertree method (Sanderson et al. 1998, Bininda-Emonds et al. 2002, Bininda-Emonds 2004b, c). The pros and cons of these competing approaches have been debated at length (de Queiroz et al. 1995, Huelsenbeck et al. 1996a, b, Wiens 1998a, Bininda-Emonds 2004a, Crandall & Buhay 2004, Gadagkar et al. 2005, Philippe et al. 2005a, de Queiroz & Gatesy 2007, Nishihara et al. 2007). An intermediate approach, referred to as conditional data combination, consists in testing *a priori* the level of congruence of different datasets, that is the level of phylogenetic agreement, and only those considered statistically congruent are combined in a supermatrix. The remaining incongruent datasets are analyzed separately (Bull et al. 1993, de Queiroz 1993, Rodrigo et al. 1993, Farris et al. 1995, Huelsenbeck & Bull 1996).

Incongruence among datasets is observed when the corresponding trees support incompatible clades. Two main causes can be invoked to explain incongruence; either different trees were inferred due to random chance or sampling error, or else, the datasets truly support different groupings (Planet 2006). Numerous factors have been described to explain differences in phylogenetic trees obtained from the analysis of datasets sampled in identical species. A wide range of evolutionary processes may cause nucleotides at different sites to evolve differently, for examples due to their codon positions or to different functional constraints (Stewart et al. 1987, Luo et al. 1989, Wolfe et al. 1989). Also, various parts of the genome may have experienced different phylogenetic histories (e.g., mitochondrial vs. nuclear genes) and trees inferred from different data types (e.g., morphological or molecular data) may support very different phylogenies (Springer & de Jong 2001). Furthermore, the use of an inappropriate method to analyze a given dataset may lead to a spurious phylogeny, which would erroneously be incongruent to another phylogeny that has been accurately estimated (Bull et al. 1993, Huelsenbeck et al. 1996a, Barker & Lutzoni 2002). Thus, given two datasets, for which only one has parameters prone to long-branch attraction (Felsenstein 1978, Hendy & Penny 1989), the choice of an inconsistent phylogenetic method to analyze both datasets will produce different trees. Other evolutionary processes can explain incongruence between datasets such as horizontal gene transfer, gene duplications, hybridization and introgression (see Wendel & Doyle 1998, Planet 2006, for an exhaustive list).

Numerous statistical procedures have been developed to test whether the observed incongruence among datasets is due to chance or not (see Huelsenbeck et al. 1996a, Planet 2006 for reviews). The null hypothesis of such tests states that the observed difference between a pair of trees is the result of stochastic variation among the corresponding datasets. The most commonly used of these tests is certainly the incongruence length difference test (ILD: Farris et al. 1994), despite numerous problems that have been associated to it. Type I error rates were shown to be well above the nominal significance level when datasets with great differences in substitution rates among sites were compared (Barker & Lutzoni 2002, Darlu & Lecointre 2002). Cunningham (1997b) suggested that a nominal significance level of 0.01 or 0.001 would be more appropriate. Power was also low when short nucleotide sequences simulated on different tree structures were compared (Darlu & Lecointre 2002).

Given two or more datasets (partitions) sampled on identical species, a concordance statistic can be calculated and tested against a distribution of permuted values to assess whether the partitions are in agreement or if they differ significantly. Legendre & Lapointe (2004) have described a test of congruence among distance matrices (CADM) that is applicable to more than two matrices. Based on Kendall's W concordance statistic (Kendall 1955), CADM is an extension of the Mantel test, which can be used to test the degree of congruence among multiple matrices. The Kendall coefficient assesses whether the concordance (or congruence) among the matrices is not simply due to chance. It is part of the class of conformity tests and the null hypothesis (H_0) is the independence or incongruence of the distance matrices. Thus, CADM differs from most other available phylogenetic tests of incongruence, which assume that the datasets have a common evolutionary history (H_0: congruence), and tests the alternate hypothesis of different histories among the datasets (H_1: incongruence). CADM assumes that the datasets have different evolutionary histories (H_0: incongruence) and tests the hypothesis that the datasets have a common evolutionary history (H_1: congruence).

Previous simulations have shown that the global and *a posteriori* CADM tests have a correct type I error rate and good power when applied to dissimilarity matrices computed from independently-generated raw data (Legendre & Lapointe 2004). Identical results were obtained in simulations involving ultrametric distance matrices (Campbell et al. 2009: chapter 2). CADM has also been successfully used to detect incongruence among phylogenetic trees obtained from different gene sequences (Legendre & Lapointe 2005). In this paper, we expand on previous CADM simulations to assess the performance of the test when applied to phylogenetic data. Specifically, the type I error rate and power of the global and *a posteriori* CADM tests were measured using distance matrices obtained from DNA sequences simulated on additive trees under various phylogenetic conditions.

3.4. METHODS

3.4.1. CADM test

The null hypothesis (H_0) of the global CADM test is the incongruence of all distance matrices (Legendre & Lapointe, 2004). That is, matrices are statistically independent from each other and convey distinct information about evolutionary relationships, which is expressed as different rankings of the distances among the taxa. Rejecting H_0 indicates that at least two matrices contain congruent information and thus support similar evolutionary histories. In those cases, *a posteriori* CADM tests can be performed to determine the contribution of each matrix to the overall congruence. *A posteriori* tests can be used to identify incongruent and congruent matrices in a set, but it does not specify the pairs or groups of congruent matrices. To this end, complementary Mantel tests based upon ranks can be used. The congruent matrices can then be combined in a supermatrix analysis. A summary of the computations to perform the CADM test follows:

1. The upper off-diagonal section of each distance matrix is unfolded and written into a vector corresponding to row *i* in a worktable.
2. The entries of each row are transformed into ranks according to their values.
3. The sum of ranks (R_j) is calculated for each column *j* of the table.
4. The mean ($\overline{R}$) of all R_j values is calculated.
5. Kendall's coefficient of concordance (*W*) is computed using the following formula:

$$W = \frac{12S}{p^2\left(n^3 - n\right) - pT}$$

where *p* is the number of matrices, *n* is the number of objects in each matrix, *S* is obtained using:

$$S = \sum_{j=1}^{n} (R_j - \overline{R})^2$$

and *T* is a correction factor for tied ranks:

$$T = \sum_{k=1}^{m} \left(t_k^3 - t_k\right)$$

in which t_k is the number of tied ranks for each *k* of *m* groups of ties. Thus, Kendall's *W* statistic is simply the variance of the row sums of ranks R_j divided by the maximum possible value that this variance can take, which occurs when all data matrices are in

total agreement. Thus, W ranges from 0 to 1, where 0 represents a complete disagreement in the rankings of the distances among the different matrices, and a value of 1 is observed when the distance matrices are in complete agreement.

6. W is transformed into a Friedman's χ^2, which is a pivotal statistic appropriated for testing, using the following formula:

$$\chi^2 = p(n-1)W$$

The observed Friedman's χ^2 (χ^2_{ref}) is tested against a distribution of the statistic obtained under permutation (χ^{2*}). For the global CADM test, all matrices are permuted at random, whereas for *a posteriori* tests, each matrix is permuted alternatively. A matrix that is not congruent to any other will have a small impact on the statistic once permuted. After a number of permutations (p_n), the one-tailed probability of the data under H_0 is computed as the number of χ^{2*} values greater than or equal to χ^2_{ref} divided by (p_n -1). In *a posteriori* comparisons, the p-value should be adjusted to maintain an adequate experimentwise error rate using a method designed specifically to correct for multiple testing (e.g., Holm, 1979). More details about the CADM procedure can be found in Legendre & Lapointe (2004) and Campbell et al. (2009: chapter 2). A version of CADM is now available in R 2.9.0 (Ihaka & Gentleman 1996, R Development Core Team 2009), within the Ape 2.3 package (Paradis et al. 2004, Paradis 2006).

For the simulations described below, one thousand replicates were simulated for each combination of parameters, unless stated otherwise. For each replicate, 999 random permutations were computed to estimate the reference distribution of the CADM statistic. We calculated the rate of rejection of H_0 with its 95% confidence interval (CI), at a nominal significance level of 0.05, for cases where H_0 was true (type I error rate) and for cases where H_0 was false (power). All the analyses were performed on ten Power Mac G5, with PowerPC 970MP processors (2 x 2.5 GHz).

3.4.2. Type I error rate

The type I error rate, that is the probability of rejecting H_0 when the data conform to this hypothesis, was assessed for both the global and *a posteriori* CADM tests. A statistical test is valid if the rejection rate of H_0 is smaller or equal to the nominal significance level of the test (Edgington 1995). Given that H_0 postulates that all distance matrices are

incongruent, we considered H_0 to be true by construct when distance matrices calculated on DNA sequences simulated on independently-generated phylogenetic trees were compared. To do so, random additive distance matrices were obtained using the method proposed by Lapointe & Legendre (1992). Phylogenetic trees were computed from the distance matrices using a neighbor-joining algorithm (NJ: Saitou & Nei 1987) in PAUP* 4.0 (Swofford 1998). DNA sequences were simulated on the phylogenetic trees using Seq-Gen 1.3.2 (Rambaut & Grassly 1997). To reproduce the complexity of DNA substitutions observed in real sequence data, we used a general time-reversible model (GTR: Lanave et al. 1984, Tavaré 1986, Rodriguez et al. 1990) following a gamma distribution (Γ: Yang 1993), with invariant sites (I). Parameters were identical to those used by Zwickl & Hillis (2002). Accordingly, the equilibrium frequencies of nucleotides A, C, G, and T were: gA = 0.1776, gC = 0.3336, gG = 0.2595, gT = 0.2293, the relative substitution rates were: rAC = 3.297, rAG = 12.55, rAT = 1.167, rCG = 2.060, rCT = 13.01, rGT = 1.0, and parameters α and I were 0.8168 and 0.5447 respectively. Distance matrices were calculated from the DNA sequence matrices using a *p* distance (Kumar et al. 1993), corrected with the same parameters as those used to simulate DNA sequences. Given that the DNA sequences were simulated on randomly-generated phylogenetic trees, the distance matrices obtained are incongruent matrices (IM). In order to explore various situations that might be encountered in phylogenetic analysis, different conditions were tested: different number of independent distance matrices (IM = 2, 3, 4, 5 and 10), different number of taxa in each matrix (n = 10, 25, 50 and 100) and varying lengths of DNA sequence (L = 1000, 5000, 10 000 and 20 000 bp).

3.4.3. Power

Power, which is the rate of rejection of a false H_0, was evaluated for different conditions of application of CADM. Rejection rates of H_0 were calculated with sets of distance matrices that included varying numbers of congruent matrices (CM) with different levels of similarity and different evolutionary parameters. The number of matrices (M) varied in a set and included incongruent matrices (IM) in addition to CM, for cases where CM < M.

3.4.3.1. Different levels of congruence among matrices

DNA sequences were simulated under a GTR + Γ + I model on the NJ trees obtained from partly similar matrices (CM_P) and from identical matrices (CM_I). CM_P were generated by random permutations of different numbers of taxa and branch lengths from a random additive distance matrix. As the number of permuted taxa increases, so does the distortion of the original matrix, whereas the level of congruence among matrices decreases. The number of taxa permuted varied according to the total number of taxa (n) included in each matrix, in order to maintain the same proportion of the taxa permuted regardless of the matrix size. The effect of the level of congruence on power was tested for $CM_P = 3$, out of a total of five matrices (M = 5), with n = 10 or 50, and L = 10 000bp. Power of *a posteriori* tests was also investigated with the same sets of CM_P. The number of taxa permuted varied from 0 to 60% of the total number of taxa. Additional simulations were performed to compare the particular case of 0% permuted taxa, which correspond to CM_I (i.e., near 100% congruence among matrices) to CM_P with 40% permuted taxa. For these analyses, a total of five distance matrices were compared (M = 5) but with varying number of CM_I or CM_P (i.e., 0, 2, 3, 4 or 5); n = 10, 25, 50 or 100; and L = 1000, 5000, 10 000 or 20 000bp. When CM_I or $CM_P = 0$, only incongruent matrices (IM) were included in the set of five matrices, which corresponds to a true H_0. A false H_0 was constructed when CM_I or $CM_P \geq 2$, and all matrices were congruent when CM_I or $CM_P = 5$.

3.4.3.2. Effect of different evolutionary parameters

Because genes controlled by different evolutionary processes can share an identical evolutionary history (i.e., branching pattern), we investigated the effect of different evolutionary parameters on the power of CADM. Following Darlu & Lecointre (2002), DNA sequences were simulated under the GTR + Γ + I model described above but with different mutation rates (s = 0.02 and 0.4) and different heterogeneity levels of substitution rates among sites (α = 0.06 and 0.8168). Homogeneity of substitution rates among sites were simulated using α = 200. The same phylogenetic tree was used to simulated DNA sequence matrices representing different partitions within a replicate, but different tree topologies were used for each replicate. DNA sequence matrices simulated with *identical* or *different* evolutionary parameters on an identical tree were compared for M = 2 or 5; s = 0.02 or 0.4; α = 0.06, 0.8168 or 200; $CM_I = 2$ or 5; n = 10, 25, 50 or 100; and L = 1000, 5000, 10 000 or 20 000bp. DNA sequences were

simulated under the same GTR parameters as above, except for s and α that varied. Thus, in addition to comparing datasets that evolved under identical conditions, we also compared datasets that were simulated with different s or α values. Thus, for s = 0.02 and 0.4, we compared datasets characterized by heterogeneity of substitution among sites vs. datasets with a homogeneous substitution rate (α = 0.06 vs. α = 200, and α = 0.8168 vs. α = 200); and for α = 0.06, 0.8168 and 200, we compared datasets with a low mutation rate vs. a high mutation rate (s = 0.02 vs. s = 0.4).

3.5. RESULTS

3.5.1. Type I error rate

Type I error rate was evaluated by calculating the number of replicates that rejected the null hypothesis when H_0 was true by construct. To construct datasets under a true H_0 of incongruence among matrices, IM were compared using CADM. Table 3.1 presents type I error rates of the global CADM test, at a nominal significance level of 0.05, obtained for different numbers of IM (2, 3, 4, 5 and 10); n (10, 25, 50 and 100); and L (1000, 5000, 10 000 and 20 000bp). In all cases, the 95% CI of the rejection rates included the nominal 0.05 alpha level, suggesting an adequate type I error rate when CADM is applied to compare distance matrices in a phylogenetic context. Type I error rates were also investigated for *a posteriori* CADM test, where each matrix involved in a set of matrices under comparisons are permuted one at a time. As for the global test, in all cases and for each matrix, the 95% CI included the nominal 0.05 alpha level suggesting an adequate type I error rate (results not shown).

3.5.2. Power

3.5.2.1. Different levels of congruence among matrices

The estimated power is the proportion of replicates for which the null hypothesis is rejected when H_0 is false by construct. For 1000 replicates, a power of 1.0 (i.e., rejection rates of 1.0) indicates that all replicates rejected the false null hypothesis, and thus power is maximal. Figure 3.1 shows power curves obtained when different numbers of taxa were permuted to construct congruent matrices (CM = 3, M = 5), of L = 10 000bp and n = 10 and 50 taxa.

Table 3.1. Type I error rates for CADM simulations with DNA sequences matrices simulated on independently-generated additive trees under a GTR + Γ + I model of evolution.

Rejection rate are given at a significance level of 0.05, with 95% confidence intervals in parentheses, calculated from 1000 replicates, except for shaded cells (100 replicates). IM = incongruent matrix, n = number of taxa, L = DNA sequence length.

		Number of IM				
n	L	2	3	4	5	10
10	1000	0.052 (0.038, 0.066)	0.047 (0.034, 0.060)	0.042 (0.030, 0.054)	0.049 (0.036, 0.062)	0.046 (0.033, 0.059)
	5000	0.050 (0.036, 0.064)	0.050 (0.036, 0.064)	0.038 (0.026, 0.050)	0.046 (0.033, 0.059)	0.046 (0.033, 0.059)
	10 000	0.049 (0.036, 0.062)	0.048 (0.035, 0.061)	0.046 (0.033, 0.059)	0.046 (0.033, 0.059)	0.047 (0.034, 0.060)
	20 000	0.047 (0.034, 0.060)	0.047 (0.034, 0.060)	0.039 (0.027, 0.051)	0.045 (0.032, 0.058)	0.043 (0.030, 0.056)
25	1000	0.054 (0.040, 0.068)	0.056 (0.042, 0.070)	0.054 (0.040, 0.068)	0.056 (0.042, 0.070)	0.04 (0.028, 0.052)
	5000	0.053 (0.039, 0.070)	0.048 (0.035, 0.061)	0.046 (0.033, 0.059)	0.05 (0.036, 0.064)	0.042 (0.030, 0.054)
	10 000	0.046 (0.033, 0.059)	0.054 (0.040, 0.068)	0.05 (0.036, 0.064)	0.049 (0.036, 0.062)	0.050 (0.036, 0.064)
	20 000	0.043 (0.030, 0.056)	0.050 (0.036, 0.064)	0.054 (0.040, 0.068)	0.047 (0.034, 0.060)	0.040 (0.028, 0.052)
50	1000	0.048 (0.035, 0.061)	0.062 (0.047, 0.077)	0.059 (0.044, 0.074)	0.050 (0.036, 0.064)	0.049 (0.036, 0.062)
	5000	0.056 (0.042, 0.070)	0.049 (0.036, 0.062)	0.055 (0.041, 0.069)	0.053 (0.039, 0.070)	0.050 (0.007, 0.093)
	10 000	0.041 (0.029, 0.053)	0.048 (0.035, 0.061)	0.053 (0.039, 0.067)	0.051 (0.037, 0.065)	0.050 (0.007, 0.093)
	20 000	0.050 (0.036, 0.064)	0.053 (0.039, 0.067)	0.050 (0.036, 0.064)	0.056 (0.042, 0.070)	0.060 (0.012, 0.107)
100	1000	0.051 (0.037, 0.065)	0.042 (0.030, 0.054)	0.040 (0.028, 0.052)	0.044 (0.031, 0.057)	0.030 (-0.004,0.064)
	5000	0.066 (0.051, 0.081)	0.040 (0.001, 0.079)	0.030 (-0.004,0.064)	0.050 (0.007, 0.093)	0.060 (0.013, 0.107)
	10 000	0.030 (-0.004,0.064)	0.060 (0.013, 0.107)	0.070 (0.019, 0.120)	0.050 (0.007, 0.093)	0.040 (0.001, 0.079)
	20 000	0.060 (0.013, 0.107)	0.050 (0.007, 0.093)	0.040 (0.001, 0.079)	0.060 (0.013, 0.107)	0.070 (0.019, 0.120)

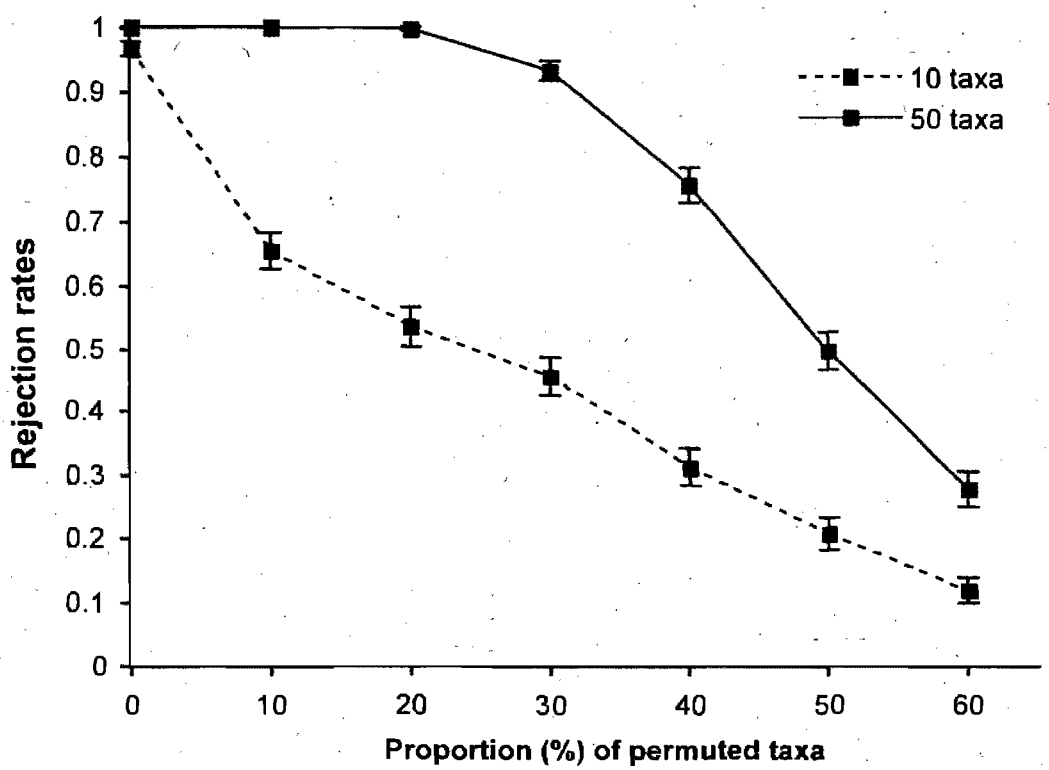

Figure 3.1. Rejection rates of H_0 for the global CADM test, comparing datasets simulated on partly similar trees, with identical evolutionary parameters (GTR+ Γ+ I).
Three congruent matrices (CM_P) and two incongruent matrices (IM) were included in each test, for a total of five distance matrices (M = 5). CM_P were generated by permuting an increasing number of taxa from a total of 10 taxa (dashed line) and 50 taxa (solid line), and for L = 10 000bp. Rejection rates are given at a significance level of 0.05, with 95% confidence intervals represented by vertical lines, calculated from 1000 replicates.

When the proportion of permuted taxa is equal to 0, the distances matrices were obtained from DNA sequences simulated on identical trees (CM_I). When the proportion of permuted taxa is greater than 0, the distance matrices were obtained from DNA sequences simulated on partly similar trees (CM_P). Power decreased with a decrease in the level of congruence among the three matrices (i.e., with an increase in the number of taxa permuted). A power close to 1.0 was observed when identical trees were used (CM_I), regardless of matrix sizes (n). Reduced power (i.e., less than 0.5) was observed for matrices with 25% or more permuted taxa, when n = 10 taxa; whereas it was observed for matrices with 50% or more permuted taxa, when n = 50 taxa.

In *a posteriori* CADM tests, the rejection rate of each individual matrix was similar to the power level obtained in the global test. Figure 3.2 presents the rejection rates obtained for each matrix tested individually for M = 5 (same matrices as those used for simulations presented in Fig. 3.1). The three partly similar matrices (CM_P) were obtained from DNA sequences simulated under identical evolutionary conditions. In situations where not all matrices are congruent in a replicate, incongruent matrices are expected to fail to reject the null hypothesis of incongruence, thus corresponding to type I error rate. Whereas, congruent matrices should reject H_0, corresponding to power.

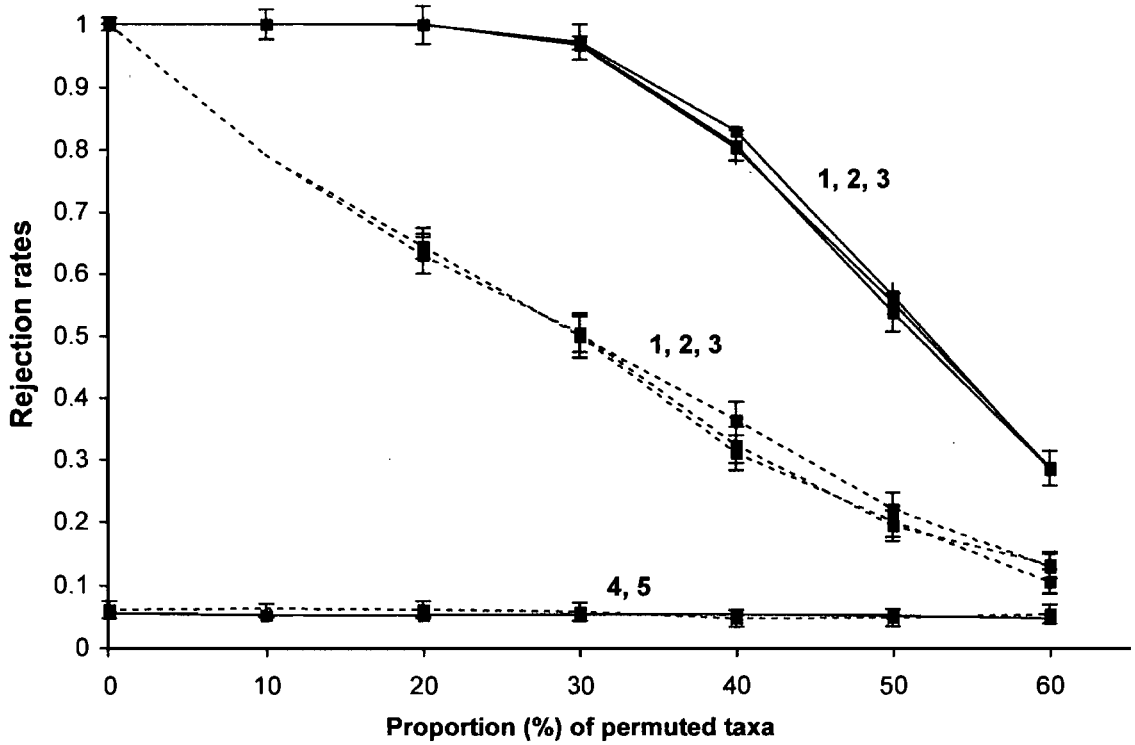

Figure 3.2. Rejection rates of H_0 for *a posteriori* CADM tests, comparing datasets simulated on partly similar trees, with identical evolutionary parameters (GTR+ Γ+ I).

The five distance matrices are the same as those used in simulations reported in Figure 3.1 for the global CADM test. For a posteriori tests, the power curves are given for each of the five matrices permuted separately, numbered from 1 to 5, for datasets with 10 taxa (dashed lines) and 50 taxa (solid lines). Rejection rates are given at a significance level of 0.05, with 95% confidence intervals represented by vertical lines, calculated from 1000 replicates.

This is exemplified in Figure 3.2, where congruent matrices (matrices 1, 2 and 3) presented rejection rates above 0.05, thus above the significance level alpha that was used for each test. Rejection rates for incongruent matrices 4 and 5 were near 0.05, for both n = 10 and 50 taxa. The power curves in Figure 3.2 are nearly identical to those in Figure 3.1 for datasets of the same size n when individual congruent matrices were tested (matrices 1, 2 and 3).

Tables 3.2 and 3.3 present rejection rates for M = 5, and that included a varying number of CM. In Table 3.2, CM corresponded to distance matrices obtained from DNA sequences simulated on identical trees (CM_I), whereas in Table 3.3, CM corresponded to distance matrices obtained from DNA sequences simulated on partly similar trees with 40% of permuted taxa (CM_P). For both tables, power increased with (1) an increase in the number of objects (n); (2) an increase in the number of congruent matrices (CM); and (3) an increase in sequence lengths (L). However, power increased much more rapidly in Table 3.2, with maximum power when three or more CM were included in a replicate, regardless of matrix sizes (n). In Table 3.3, maximum power was observed only for four matrices, or more, included in a set of M = 5, and with larger matrices (n = 100 taxa).

3.5.2.2. Effect of different evolutionary parameters

Power was also calculated for distance matrices obtained from DNA sequences simulated on identical trees under a GTR model, with identical or different evolutionary parameters. In Table 3.4, distance matrices obtained from DNA sequences simulated with *identical* evolutionary parameters were compared, for $CM_I = 2$. Rejection rates of pairwise comparisons tested for different values of mutation rates (s) and heterogeneity of substitution rates among sites (α) are presented. In Table 3.5, distance matrices obtained from DNA sequences simulated with *contrasting* evolutionary parameters were compared. In both cases, and for every condition tested, rejection rates were high, with at least 78.9% of the replicates that rejected the H_0. The lowest rejection rate was observed when matrices obtained from DNA sequences simulated under more extreme parameters of evolution (i.e., s = 0.02; α = 0.06) were tested. But in general, most cases rejected the null hypothesis of incongruence (i.e., power of 1.0). Identical simulations were also performed for $CM_I = 5$ (M = 5), and rejection rates were of 1.0 for every case (results not shown).

Table 3.2. Rejection rates of H_0 for CADM comparing datasets simulated on *identical* trees and with *identical* evolutionary parameters (GTR+ Γ + I) and with M = 5.

A false H_0 was constructed by including a different number of congruent matrices (CM_I) together with a different number of incongruent matrices (IM), for a total of five distance matrices (M = 5). When CM_I = 5, all matrices included in the test are congruent. Rejection rates are given at a significance level of 0.05, with 95% confidence intervals in parentheses, calculated from 1000 replicates, except for shaded cells (100 replicates). Dashes (-) correspond to a CI of 1.000 - 1.000.

		CM_I			
n	L	2	3	4	5
10	1000	0.308 (0.279-0.337)	0.928 (0.912-0.944)	1.000 -	1.000 -
	5000	0.363 (0.333-0.393	0.973 (0.963-0.983)	1.000 -	1.000 -
	10 000	0.383 (0.353-0.413)	0.966 (0.955-0.977)	1.000 -	1.000 -
	20 000	0.380 (0.350-0.410)	0.974 (0.964-0.984)	1.000 -	1.000 -
25	1000	0.569 (0.538-0.600)	1.000 -	1.000 -	1.000 -
	5000	0.662 (0.633-0.691)	1.000 -	1.000 -	1.000 -
	10 000	0.675 (0.646-0.704)	1.000 -	1.000 -	1.000 -
	20 000	0.682 (0.653-0.711)	1.000 -	1.000 -	1.000 -
50	1000	0.740 (0.715-0.769)	1.000 -	1.000 -	1.000 -
	5000	0.851 (0.829-0.873)	1.000 -	1.000 -	1.000 -
	10 000	0.869 (0.848-0.890)	1.000 -	1.000 -	1.000 -
	20 000	0.898 (0.880-0.917)	1.000 -	1.000 -	1.000 -
100	1000	0.890 (0.828-0.952)	1.000 -	1.000 -	1.000 -
	5000	0.970 (0.936-1.000)	1.000 -	1.000 -	1.000 -
	10 000	0.970 (0.936-1.000)	1.000 -	1.000 -	1.000 -
	20 000	0.970 (0.936-1.000)	1.000 -	1.000 -	1.000 -

Table 3.3. Rejection rates of H_0 for CADM comparing datasets simulated on *partly similar* trees and with *identical* evolutionary parameters (GTR+ Γ + I).

A different number of congruent matrices (CM_P) and a different number of incongruent matrices (IM) were included in each test, for a total of five distance matrices (M = 5). To generate CM_P, DNA sequences were simulated on partly similar trees (with permutations of 40% of n). Rejection rates are given at a significance level of 0.05, with 95% confidence intervals in parentheses, calculated from 1000 replicates, except for shaded cells (100 replicates).

		CM_P			
n	L	2	3	4	5
10	1000	0.106 (0.087-0.125)	0.263 (0.236-0.290)	0.523 (0.492-0.554)	0.802 (0.777-0.827)
	5000	0.105 (0.086-0.124)	0.300 (0.272-0.328)	0.586 (0.555-0.617)	0.866 (0.845-0.887)
	10 000	0.113 (0.093-0.133)	0.311 (0.282-0.340)	0.608 (0.578-0.638)	0.872 (0.851-0.893)
	20 000	0.122 (0.102-0.142)	0.314 (0.285-0.343)	0.615 (0.585-0.645)	0.875 (0.854-0.896)
25	1000	0.130 (0.109-0.151)	0.409 (0.378-0.440)	0.805 (0.780-0.830)	0.977 (0.968-0.986)
	5000	0.158 (0.135-0.181)	0.495 (0.464-0.526)	0.893 (0.874-0.912)	0.993 (0.988-0.998)
	10 000	0.151 (0.129-0.173)	0.508 (0.477-0.539)	0.902 (0.884-0.920)	0.997 (0.994-1.000)
	20 000	0.153 (0.131-0.175)	0.514 (0.483-0.545)	0.907 (0.889-0.925)	0.996 (0.992-1.000)
50	1000	0.163 (0.140-0.186)	0.560 (0.529-0.591)	0.960 (0.948-0.972)	1.000 -
	5000	0.206 (0.181-0.231)	0.701 (0.673-0.729)	0.991 (0.985-1.000)	1.000 -
	10 000	0.218 (0.192-0.244)	0.730 (0.702-0.758)	0.996 (0.992-1.000)	1.000 -
	20 000	0.229 (0.203-0.255)	0.748 (0.721-0.775)	0.997 (0.994-1.000)	1.000 -
100	1000	0.210 (0.129-0.291)	0.730 (0.641-0.819)	0.990 (0.970-1.000)	1.000 -
	5000	0.260 (0.173-0.347)	0.880 (0.815-0.945)	1.000 -	1.000 -
	10 000	0.270 (0.181-0.359)	0.900 (0.840-0.960)	1.000 -	1.000 -
	20 000	0.310 (0.218-0.402)	0.920 (0.866-0.974)	1.000 -	1.000 -

Table 3.4. Rejection rates of H_0 for CADM comparing datasets simulated on *identical* trees (CM_I = 2, M = 2) and with *identical* evolutionary parameters.

Results are shown for a GTR + Γ + I model with different s and α. Rejection rates are given at a significance level of 0.05, with 95% confidence intervals in parentheses, calculated from 1000 replicates.

		α = 0.06		α = 0.8168		α = 200	
n	L	s = 0.02	s = 0.4	s = 0.02	s = 0.4	s = 0.02	s = 0.4
10	1000	0.789	0.958	0.944	1.000	0.940	1.000
		(0.764-0.814)	(0.946-0.970)	(0.930-0.958)	-	(0.925-0.955)	-
	5000	0.999	1.000	1.000	1.000	1.000	1.000
		(0.997-1.000)	-	-	-	-	-
	10 000	1.000	1.000	1.000	1.000	1.000	1.000
		-	-	-	-	-	-
	20 000	1.000	1.000	1.000	1.000	1.000	1.000
		-	-	-	-	-	-
50	1000	0.891	0.997	0.976	1.000	0.978	1.000
		(0.872-0.910)	(0.994-1.000)	(0.966-0.986)	-	(0.969-0.987)	-
	5000	1.000	1.000	1.000	1.000	1.000	1.000
		-	-	-	-	-	-
	10 000	1.000	1.000	1.000	1.000	1.000	1.000
		-	-	-	-	-	-
	20 000	1.000	1.000	1.000	1.000	1.000	1.000
		-	-	-	-	-	-

Table 3.5. Rejection rates of H_0 for CADM comparing datasets simulated on *identical* trees (CM_I = 2, M = 2), with *contrasting* evolutionary parameters (GTR model with different s or α, for each dataset).

Rejection rate are given at a significance level of 0.05, with 95% confidence intervals in parentheses, calculated from 1000 replicates.

		s = 0.02		s = 0.4		α = 0.06	α = 0.8168	α = 200
n	L	α: 200 vs. 0.06	α: 200 vs. 0.8168	α: 200 vs. 0.06	α: 200 vs. 0.8168	s: 0.02 vs. 0.4	s: 0.02 vs. 0.4	s: 0.02 vs. 0.4
10	1000	0.866 (0.845-0.887)	0.939 (0.924-0.954)	0.993 (0.988-0.998)	1.000 -	0.949 (0.935-0.963)	0.998 (0.995-1.000)	0.999 (0.997-1.000)
	5000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
	10 000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
	20 000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
25	1000	0.927 (0.911-0.943)	0.965 (0.954-0.976)	1.000 -	1.000 -	0.992 (0.986-0.998)	1.000 -	1.000 -
	5000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
	10 000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
	20 000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
50	1000	0.945 (0.931-0.959)	0.980 (0.971-0.989)	1.000 -	1.000 -	0.999 (0.997-1.000)	1.000 -	1.000 -
	5000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
	10 000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -
	20 000	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -	1.000 -

3.6. DISCUSSION

Our simulations clearly demonstrate the validity of CADM to test for congruence among different data partitions. However, in comparison to other tests used in phylogenetic analysis, the null and alternate hypotheses are reversed. Most phylogenetic incongruence tests assess that the datasets share an identical evolutionary history and are therefore congruent (Planet 2006). On the contrary, the null hypothesis of CADM assumes that the datasets support different phylogenetic hypotheses. Thus, type I error and power studies are not directly comparable between phylogenetic incongruence tests such as the incongruence length difference test (ILD: Farris et al. 1994) and CADM, given that the tests are designed differently. Kendall's coefficient of concordance is widely used in other fields, especially psychology, where it is used to assess the degree of correspondence or strength of association among different estimators (Siegel & Castellan 1988). In a phylogenetic context, CADM evaluates the degree of agreement among different estimators (i.e., data partitions) of a common phylogeny (as represented by evolutionary distances).

In order to investigate type I error rates, that is the number of replicates that rejected H_0 when it is true by construct, incongruent distance matrices were compared. In every case, the 95% CI of the rejection rate included the nominal significance level of 0.05 used for the test (Table 3.1). Hence, CADM accurately detects incongruent matrices even when multiple data partitions are tested simultaneously. In comparison, ILD produces inflated type I error rates under particular conditions. Darlu & Lecointre (2002) designed computer simulations to assess the performance of ILD under different conditions of rejection of the null hypothesis (i.e., congruence between datasets). They observed rejection rates well above the alpha level when sequences were simulated on identical trees, but with important difference in the substitution rates among sites. Furthermore, the rejection rates increased for longer sequences and for assymetrical trees. Similarly, ILD was shown to be strongly biased in detecting topological congruence (Barker & Lutzoni 2002), and to be negatively influenced by the presence of a substantial number of noisy characters (Dolphin et al. 2000). Different methods have been proposed to alleviate this problem, such as using an alternative null model (Dolphin et al. 2000) or an arcsine transformation of the standardized length of the trees in order to linearize the relationship between noise and tree length (Quicke et al. 2007).

Numerous congruence tests have also been designed recently such as principal components analysis on log-likelihoods or p-values (Brochier et al. 2005) or heat maps to identify groups of congruent markers (Bapteste et al. 2005, Susko et al. 2006). Bayesian approaches have also been suggested by others (Suchard 2005, Ané et al. 2007). Leigh et al. (2008) have discussed caveats associated to each method, and they proposed a hierarchical clustering method based on log-likelihood ratios (Huelsenbeck & Bull 1996) to test congruence. However, these tests are still dependent upon tree inference (Leigh et al. 2008), and thus, might produce spurious results when inadequate models of evolution are used (Sullivan & Swofford 2001, Ripplinger & Sullivan 2008). CADM is not specifically designed to compare phylogenetic trees, and thus is not directly affected by the choice of a substitution model. Interestingly, because it relies on distance matrices, CADM could also apply to pathlength distance matrices obtained from phylogenetic trees (Legendre & Lapointe 2005).

As observed in previous simulation studies (Legendre & Lapointe 2004, Campbell et al. 2009: chapter 2), the power of CADM increased with the number of taxa, and with the number of congruent matrices within a set of distance matrices (Tables 3.2 and 3.3). Thus, the test performs according to expectations. Indeed, the power of a test should increase with the number of objects and with effect size, that is the degree to which congruence is present (Cohen 1988). Interestingly, power also tends to increase with longer DNA sequences from which distance matrices are calculated. This novel observation is opposed to the prediction of Legendre & Lapointe (2004), who stated that power is not affected by the number of variables in the raw data. They argue that the number of variables should not affect the outcome of the test since data partitions are converted into distance matrices prior to computing the test. However, Legendre & Lapointe (2004) proposed a weighted version of CADM, which can be used to assign weights to each matrix in the global analysis. Comparisons of distance matrices obtained from DNA sequences is a particular application of the CADM test. We believe that the better power observed for longer DNA sequences can be explained by the number of informative sites. However, it appears that power increases more rapidly with the number of taxa than with the number of characters, and even more rapidly with the number of congruent matrices under comparison.

Tables 3.4 and 3.5 present the rejection rates for two congruent matrices, and nearly all cases tested rejected the null hypothesis of incongruence. However, when all matrices

under comparison are congruent, power is always maximal regardless of the evolutionary parameters (results not shown). When congruent and incongruent matrices were compared, most cases rejected the null hypothesis, with maximal power (Table 3.2).

When the overall level of congruence decreases among congruent matrices, so does power (Fig. 3.1). For DNA sequence matrices simulated on phylogenetic trees with 40% permuted taxa, a drastic decrease in power was observed (Table 3.2 vs. Table 3.3). The greater the effect size, the greater the power of the test will be (Cohen 1988). In this study, topological differences are reflected by a decrease in congruence among the DNA sequence matrices, and this can be interpreted as noise in the data. Indeed, power decreases quite abruptly with an increase in topological differences (Fig. 3.1). The level of congruence among distance matrices is indicated by the p-values of *a posteriori* tests. The higher the probability, the higher the weight of evidence that this matrix differs from the other matrices.

One of the main advantages of CADM lies in its ability to test for multiple matrices in a single analysis, and identify congruent and incongruent matrices in a set of matrices. This is achieved through *a posteriori* testing, which compares each matrix to all other matrices by permuting a single matrix at a time. Our results show that power of *a posteriori* CADM tests is equivalent to the power observed for the global test (Fig. 3.1 and 3.2). Once the null hypothesis of incongruence is rejected, *a posteriori* tests should be used to identify the matrices that can be combined in a supermatrix, and those that should be analyzed separately. Other tests, such as ILD, can be modified to test for incongruence among multiple matrices. However, when the null hypothesis of congruence is rejected, it is impossible to know which matrices are incongruent (Planet 2006). Different approaches have been proposed to identify individual incongruent matrix within a set of multiple matrices, and the methods and problems associated are discussed by Planet (2006). The Concaterpillar program also allows to test for incongruence among multiple matrices through pairwise comparisons (Leigh et al. 2008). However, the number of tests increases exponentially with the number of datasets, and it becomes excessively computationally demanding when numerous datasets have to be compared.

In light of our results, CADM has proven to be statistically valid for detecting congruence among distance matrices in a phylogenetic context. One important advantage of this permutation method is its computational efficiency in significance testing. CADM offers several other advantages with respect to previously described incongruence tests: (1) The statistic is calculated directly from distance matrices, thus any data that can be transformed into a distance matrix can be analyzed (e.g., DNA or amino acid sequences). (2) Different types of data can be compared if converted to distance matrices using an appropriate function. (3) Data that readily come in the form of distance matrices do not have to be further transformed into character-state data matrices. (4) Given that distances can be calculated directly from the raw data, possible biased introduced by the use of an inappropriate phylogenetic method are avoided. (5) Also, appropriate distances can be chosen for each individual dataset to accurately model its evolutionary parameters. (6) If needed, pathlength distances calculated on phylogenetic trees can also be used, which provide an interesting method to test for congruence among different trees in a supertree approach. (7) Distance matrices can be weighted differentially to account for different numbers of characters. (8) *A posteriori* tests can be performed to identify which particular matrices are congruent among all datasets tested. With the growing amount of taxa and sequences that are used in phylogenomics, CADM offers a simple alternative to compare multiple matrices and identify congruent data partitions.

3.7. ACKNOWLEDGMENTS

We would like to thank the members of the Laboratoire d'Écologie Moléculaire et d'Évolution (LEMEE) for their constructive comments on a preliminary version of this manuscript. For the phylogenetic analyses, we used the computational resources located in the Laboratoire Interfacultaires de Micro-Informatique de l'Université de Montréal and we thank Marie-Hélène Duplain for granting access to the lab outside business hours. This study was supported by NSERC and FQRNT scholarships to VC and by NSERC grant OGP0155251 to FJL.

CHAPITRE 4:

THE USE AND VALIDITY OF COMPOSITE TAXA IN PHYLOGENETIC ANALYSIS

Cet article est publié sous la référence :

Campbell V. & Lapointe F.-J. In Press. The Use and Validity of Composite Taxa in Phylogenetic Analysis. Systematic Biology.

4.1. RÉSUMÉ

En analyse phylogénétique, il est proposé de réduire le nombre de données manquantes dans les super-matrices en combinant les séquences de différentes espèces pour obtenir une séquence d'ADN complète: ces séquences hybrides sont appelées chimères. Une autre approche consiste à analyser les super-matrices incomplètes en conservant les données manquantes. L'exactitude des arbres phylogénétiques estimés à l'aide de matrices qui comprennent des taxons chimères a récemment été remise en question, et la meilleure stratégie pour l'analyse de super-matrices incomplètes est débattue. Grâce à des simulations numériques, nous avons comparé l'exactitude phylogénétique des deux approches concurrentes. Plus précisément, nous avons exploré la validité de l'utilisation de taxons chimères pour déduire des relations de haut niveau, c'est-à-dire les relations entre groupes monophylétiques. Des séquences d'ADN ont été simulées sur un arbre modèle de 42 taxons et des super-matrices comprenant différents pourcentages de données manquantes ont été générées. Ces super-matrices incomplètes ont été analysées en codant chaque position manquante avec un "?" ou en réduisant la quantité de données manquantes grâce à la combinaison de deux ou plusieurs taxons pour générer des séquences chimères. Un total de 180 combinaisons de paramètres ont été analysées, c.-à.-d. 18 situations différentes pour lesquelles deux méthodes d'inférence ont été utilisées et cinq pourcentages de données manquantes ont été générés. Nous avons observé une exactitude significativement plus élevée pour les matrices chimères dans 46 des 180 combinaisons, alors que les matrices avec données manquantes étaient significativement plus performantes dans huit cas seulement. Dans toutes les autres situations, l'exactitude phylogénétique obtenue n'était pas significativement différente pour les matrices chimères et les données manquantes. Cette étude démontre que l'utilisation de séquences chimères représente une stratégie optimale pour réduire la quantité de données manquantes dans les super-matrices et de plus, cette approche réduit grandement le temps de calcul, un aspect crucial à considérer dans les études phylogénomiques.

4.2. ABSTRACT

In phylogenetic analysis, one possible approach to minimize missing data in DNA supermatrices consists in sampling sequences from different species to obtain a complete sequence for all genes included in the study. Those complete sequences are composites, since DNA sequences that are combined belong to different species. An alternative approach is to analyze incomplete supermatrices by coding unavailable DNA sequences as missing. The accuracy of phylogenetic trees estimated using matrices that include composite taxa has recently been questioned and the best approach for analyzing incomplete supermatrices is highly debated. Through computer simulations, we compared the phylogenetic accuracy of the two competing approaches. We explored the effect of composite taxa when inferring higher-level relationships, i.e., relationships between monophyletic groups. DNA sequences were simulated on a 42-taxon model tree and incomplete supermatrices containing different percentages of missing data were generated. These incomplete supermatrices were analyzed either by coding the missing data with "?" or by reducing the amount of missing data through the combination of two or more taxa to generate composite taxa. Out of 180 comparisons (18 simulation cases with two different inference methods and five levels of incompleteness), we observed significantly higher phylogenetic accuracies for composite matrices in 46 comparisons, whereas missing data matrices outperformed composites in eight comparisons. In all other cases, the phylogenetic accuracy obtained with composite matrices was not significantly different from that of missing data matrices. This study demonstrates that composite taxa represent a useful approach to minimize the amount of missing data in supermatrices and we suggest that it is the optimal approach to use in phylogenomic studies to reduce computing time.

4.3. INTRODUCTION

With advances in molecular techniques, a large number of DNA sequences are rapidly becoming available for an increasing number of species. This wealth of genetic information can be used to improve the accuracy of phylogenies, given that consistent phylogenetic methods are applied (Hillis et al. 1996). Inclusion of longer DNA sequences (Huelsenbeck & Hillis 1993, Hillis et al. 1994, Wiens 2003b, de Queiroz & Gatesy 2007, Telford 2007) and increased taxon sampling (Graybeal 1998, Hillis 1998, Rannala et al. 1998, Zwickl & Hillis 2002, Delsuc et al. 2005, Leebens-Mack et al. 2005,

Hedtke et al. 2006, Telford 2008) provide more power to infer the "correct" evolutionary tree (Sanderson et al. 1998, Telford 2008). The simultaneous analysis of a large number of genes also minimizes the adverse effects of lateral gene transfers (Doolittle 1999, Lerat et al. 2003) and duplications (Page 2000). An increasing number of phylogenomic studies are published for datasets including more than 100 genes (e.g., Lerat et al. 2003, Rokas et al. 2003, Driskell et al. 2004, Philippe et al. 2005b, Fitzpatrick et al. 2006, Nishihara et al. 2007, Wildman et al. 2007, Dunn et al. 2008, Zou et al. 2008).

Two opposite views have been proposed as to how to incorporate the growing amount of data to infer evolutionary relationships. Whereas the combined approach (*sensu* de Queiroz 1993) combines different datasets in a supermatrix (Eernisse & Kluge 1993, Kluge & Wolf 1993, Gatesy et al. 2004, de Queiroz & Gatesy 2007), the consensus approach (*sensu* de Queiroz 1993) analyzes datasets separately, and the resulting trees are then combined with a consensus (Swofford 1991, Farris et al. 1995, Huelsenbeck & Bull 1996) or a supertree method (Sanderson et al. 1998, Bininda-Emonds et al. 2002, Bininda-Emonds 2004a, b). The pros and cons of these competing approaches have been debated at length in the literature (de Queiroz et al. 1995, Huelsenbeck et al. 1996a, b, Wiens 1998a, Bininda-Emonds 2004a, b, Crandall & Buhay 2004, Gadagkar et al. 2005, Philippe et al. 2005a, de Queiroz & Gatesy 2007, Nishihara et al. 2007).

When the combined approach is used, the concatenation of numerous genes from different species often results in a supermatrix with missing data. Indeed, a taxon bias is observed in sequence databases, with a large number of genes (or whole genome) sequenced for a few key species thus leading to large supermatrices dominated by missing data (Crandall & Buhay 2004, Driskell et al. 2004, Philippe et al. 2005a, Wiens 2006, Telford 2008). Different methods have been employed to handle missing data, such as removing incomplete taxa or simply coding the data as missing (see reviews by Wiens & Reeder 1995, Wiens 2006). Whereas the former method discards a large number of potentially informative characters, the second may cause a decrease in phylogenetic resolution (Huelsenbeck 1991, Wiens & Reeder 1995, Flynn et al. 2005). However, recent computer simulations have shown that the misplacement of an incomplete taxon on a phylogenetic tree is often due to poor character sampling rather than the amount of missing data, and that this effect can be alleviated by adding

characters (Wiens 2003b, Philippe et al. 2004). Also, while an unbalanced distribution of missing data within a matrix can bias the estimation of model parameters, this effect is less important than the benefit gained from adding an incomplete taxon that breaks a long branch (Wiens 2005).

An alternative approach is now often used to circumvent the presence of missing data in supermatrices: the construction of composite taxa (e.g., Shoshani & McKenna 1998, Murphy et al. 2001b, Scally et al. 2001, Asher et al. 2004, Springer et al. 2004a, Poux et al. 2006, Telford 2007, Beck 2008). To obtain such composite taxa (or chimeric taxa), sequences from different species are combined within a monophyletic group, defined *a priori*, to form a complete sequence for all the genes included in the analysis (Shoshani & McKenna 1998). When using composites, taxonomic relationships are usually inferred at a level higher than the one of the taxa used to create composites. Hence, species from different genera (e.g., Beck 2008) or even above the generic level (e.g., Scally et al. 2001) may be combined to form composite taxa when inferring relationships among orders. An important assumption when creating composite taxa is that each composite is monophyletic relative to the other taxa included in the analysis (Nixon & Davis 1991, Prendini 2001, Scally et al. 2001, Malia et al. 2003, Springer et al. 2004a). Composite matrices include a different number of composite taxa depending on the number of available sequences and the amount of missing data in each sequence. Although some studies included only one composite taxon in their analyses (e.g., Flynn et al. 2005, Marek & Bond 2006), most studies incorporated a substantial amount of composite taxa (e.g., 12/28 taxa: Madsen et al. 2001; 25/52 taxa: Philippe et al. 2007; 6/58 taxa: Duvall et al. 2008). In the field of phylogenomics, the composite approach will certainly be opted more often given that an increasing number of genes will become available for an increasing number of taxa. Phylogenomic studies of higher-level relationships that include a single composite taxon to represent each terminal can readily be found. For examples, Delsuc et al. (2006) combined 38 species to form 14 composite taxa and Bourlat et al. (2008) combined 168 species into 37 composite taxa.

Malia et al. (2003) have evaluated the effect of the composite approach by reanalyzing Madsen et al.'s (2001) data and concluded that the use of composite taxa can suggest evolutionary relationships that are not supported when the matrix is analyzed with missing data. Therefore, they recommend analyzing incomplete data matrices as is, i.e., with missing data, although they observed a decrease in phylogenetic resolution

(i.e., polytomies within lower-level groups due to the inclusion of taxa with no overlapping sequences). Even if composite taxa have been increasingly used in recent studies, the performance of this approach has never been assessed with respect to the analysis of complete data matrices. Also, the phylogenetic accuracy obtained with matrices that include composite taxa has not been directly compared to that of missing data matrices in a simulation framework. The simulation approach represents a powerful tool to investigate the accuracy of phylogenetic methods under controlled conditions and with fixed parameters (Hillis 1995, Huelsenbeck 1995, Wiens 1998c). Even though simulations represent a simplified version of the reality and cannot encompass the full range of possible cases, they can be used to predict the accuracy of the results when the actual phylogeny is unknown (Hillis 1995, Wiens 1998c). The main objective of this study is to explore the phylogenetic accuracy of composite taxa in retrieving the "true" relationships among clades. Through simulations, we compared the relative performance of composite matrices versus missing data matrices and also, to the ideal situation of a complete data matrix.

4.4. METHODS

In this study, we compared the performance of two competing approaches when analyzing incomplete matrices: either by coding missing data as "?" or by forming composite taxa to reduce the amount of missing data. As is usually the case when composite taxa are used, we were interested in recovering higher-level relationships, i.e., relationships among monophyletic groups. The datasets analyzed consisted in DNA sequence matrices simulated on a known phylogenetic tree (referred to as the model tree). The following parameters, involved in the formation or in the analyses of the datasets, were investigated:

1. Model Tree (MT). Two different branch length ratios of the model tree were used: one with short terminal branches (MT_S) and the other with long terminal branches (MT_L).
2. DNA Sequence Length (L). Three different sequence lengths were used: 1500, 8362 and 20 000bp.
3. Model of DNA Evolution. A simple model (Jukes-Cantor: JC) and a more complex model (Transversional: TVM) were used as models of substitution to simulate the evolution of DNA sequences on the MT.

4. Level of Matrix Incompleteness (I). DNA sequence matrices were analyzed with five different levels of matrix incompleteness: 5, 15, 30, 50 and 75% of missing data.
5. Inference Method. Neighbor-joining (NJ) and maximum likelihood (ML) methods were used to infer phylogenetic trees:
   a. For DNA sequences simulated under a JC model: NJ distances were corrected with a JC model (NJ-JC), and ML analyses were performed using a JC substitution model (ML-JC).
   b. For DNA sequences simulated under a TVM model: NJ was corrected either with a JC or a TVM + Γ + I model (NJ-JC and NJ-TVM), and ML analyses assumed a JC or a TVM + Γ + I model (ML-JC and ML-TVM).
6. Matrix Size. The formation of composite taxa automatically results in a reduction of matrix size. Therefore, in addition to comparing composite matrices to missing data matrices that included all taxa, we also compared composite matrices to missing data matrices with a reduced number of taxa corresponding to the sizes of composite matrices.
7. Non-monophyletic Composites. One important assumption of composite matrices is that all taxa are combined within a monophyletic group. We tested the effect of violating the monophyly assumption by creating non-monophyletic composites that combined sequences of taxa from different clades.

4.4.1. Simulation of DNA sequences

All simulations were performed using a model tree (MT) of 42 taxa representing ten monophyletic groups of four taxa and two outgroups (Fig. 4.1). The relationships among the ten monophyletic groups are represented by the internal branches among clades or "backbone" of the tree (BB: bolded lines in Fig. 4.1). For all simulations, two different ratios of between-clade to within-clade branch lengths were used: a 1:3 ratio (MT_L), corresponding to longer terminal branch lengths with respect to internal branch lengths, and a 1:0.6 ratio (MT_S), corresponding to terminal branch lengths five times shorter than those on MT_L. Branch lengths on MT_L were derived from a maximum-likelihood analysis of BRCA1 sequences from 52 species representing all orders of Eutherian mammals (Madsen et al. 2001), and thus approximate branch lengths obtained with real data. Internal branches among clades were relatively short (0.005 - 0.1 substitutions per site), whereas the majority of branches within clades were relatively long (0.01 - 0.3 substitutions per site).

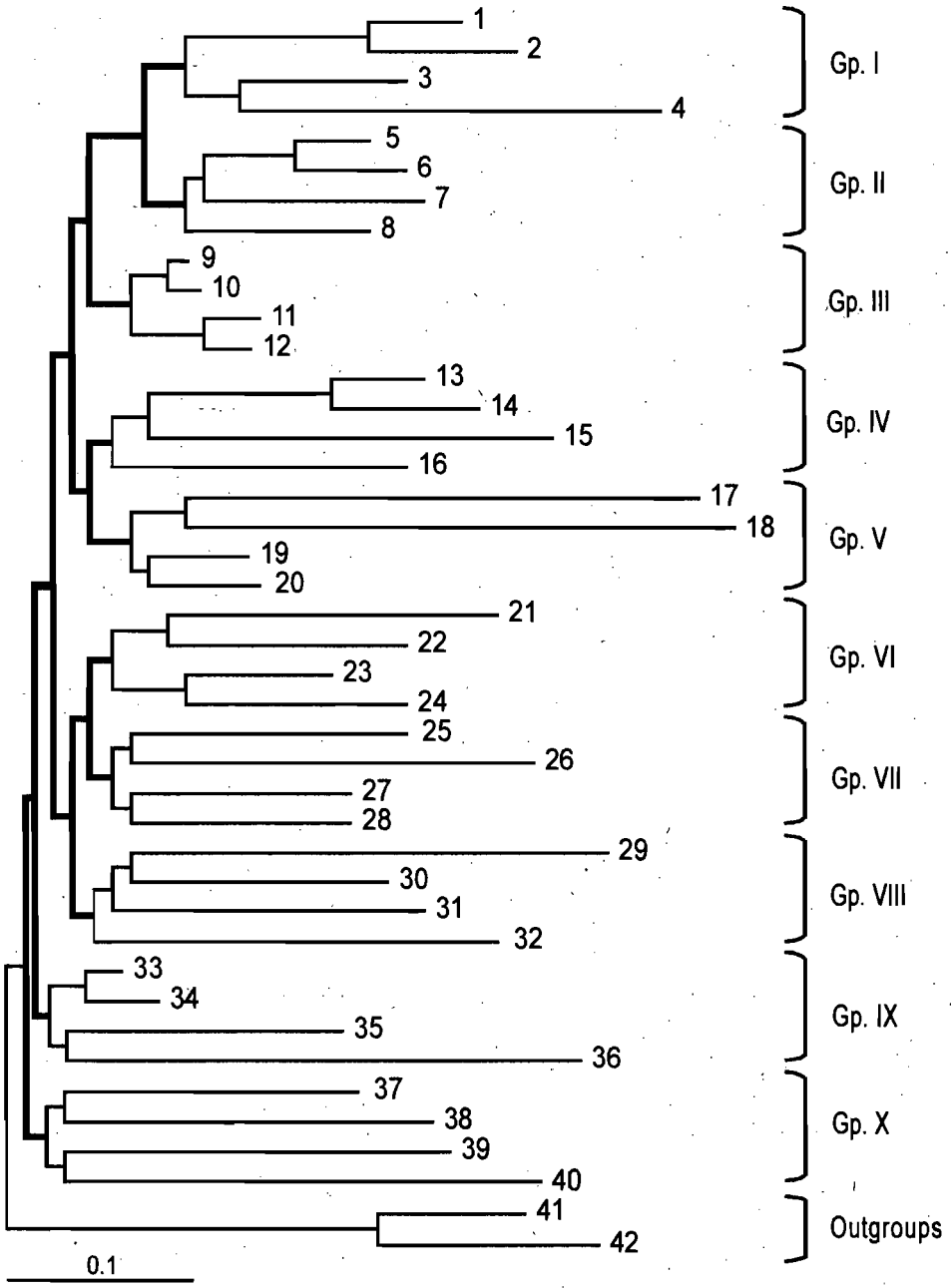

Figure 4.1. Model tree of 42 taxa with a branch length ratio of 1:3 (MT_L). The ten monophyletic groups (Gp.) of four species are indicated by numbers I to X.
Phylogenetic relationships among the groups are represented by the backbone (BB) of the tree (in bold). Branch lengths correspond to the number of substitutions per site and are derived from the Eutherian mammal tree published by Madsen et al. (2001).

To generate complete data matrices (Fig. 4.2A), DNA sequences were simulated on MT_S and MT_L, using Seq-Gen 1.3.2 (Rambaut & Grassly 1997). A simple and a more complex model of sequence evolution were employed. The first model of evolution assumes equal base frequencies and equal substitution rates (JC: Jukes & Cantor 1969), whereas the second model was selected to represent more accurately the complexity of DNA substitutions observed in real sequence data. Namely, a transversional model (TVM: Posada & Crandall 1998) with a gamma distribution (Γ: Yang 1993) and with invariant sites (I) was used (TVM + Γ + I). The equilibrium frequencies of nucleotides A, C, G, and T were: $g_A = 0.4054$, $g_C = 0.3160$, $g_G = 0.052$, $g_T = 0.2243$, the relative substitution rates were: $r_{AC} = 0.1450$, $r_{AG} = 4.5789$, $r_{AT} = 0.3872$, $r_{CG} = 0.4051$, $r_{CT} = 4.5789$, $r_{GT} = 1.0$, and parameters α and I were 0.4367 and 0.3088 respectively. To test the effect of character sampling, three different matrix sizes were simulated: 1500 bp (five hypothetical genes of 300bp), 8362bp (15 genes with lengths corresponding to those used by Murphy et al. 2001a) and 20 000bp (20 genes of 1000bp). One thousand replicates were simulated for each combination of parameters.

4.4.2. Missing data matrices

Missing data matrices (Fig. 4.2B) were derived from the complete matrix by deleting, at random, an increasing number of genes, i.e., 5, 15, 30, 50 or 75% of the total number of genes in the ingroup taxa. We ensured that all species had at least one gene in common to avoid undefined distances and to increase resolution when inferring the phylogenetic tree. In all cases, the two outgroup sequences were kept complete.

4.4.3. Composite matrices

The missing data matrices were used to generate composite taxa (Fig. 4.2C and Table 4.1) using the following criteria:

1. When all taxa within a monophyletic group had complete sequences, no composite taxon was formed.
2. When only one taxon within a monophyletic group had missing data, no composite taxon was formed in that clade and all taxa were used in the analyses, including the taxon with missing data.

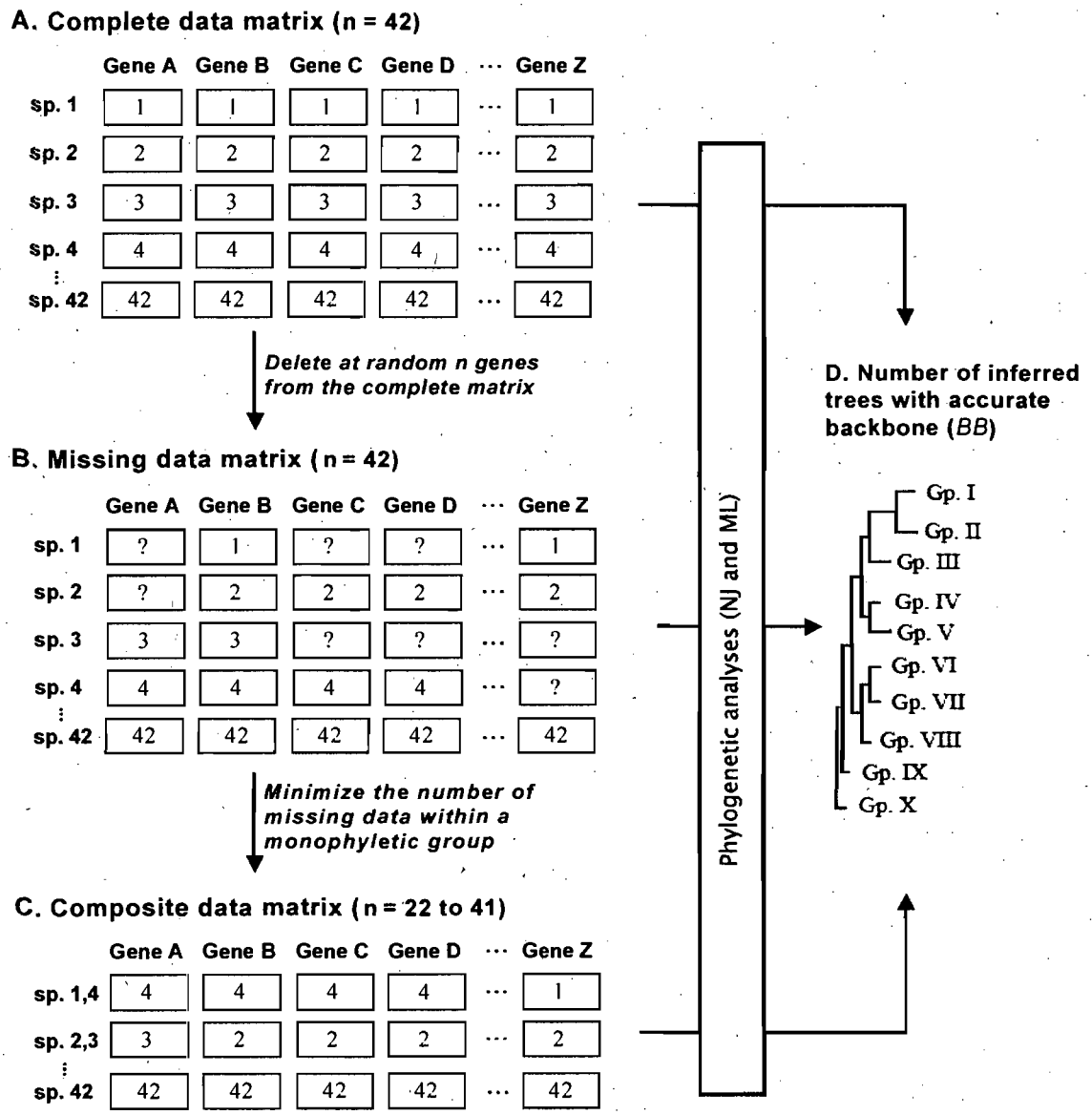

Figure 4.2. Schematic representation of the simulation procedure to create missing data and composite matrices.

A. One thousand complete DNA sequence matrices of different sizes (L = 1500, 8362 and 20 000bp) were simulated on a model tree (MT_S or MT_L) using a JC or a TVM+Γ+I model of evolution. B. A number of genes were randomly deleted from the complete matrix to generate matrices with 5, 15, 30, 50 or 75% of missing data. C. From the missing data matrices, composite matrices were generated by combining taxa within each monophyletic group in order to minimize the number of missing data. D. For each type of matrix, the number of replicates that inferred a phylogenetic tree with a backbone identical to that of the model tree was calculated using two different phylogenetic methods: neighbor-joining (NJ) and maximum likelihood (ML).

3. When two or more taxa had missing data within a monophyletic group, all possible composites of two or three species were formed (given the four species A, B, C and D, the thirteen possible combinations are: AB/CD, AC/BD, AD/BC, AB/C/D, AC/B/D, AD/B/C, BC/A/D, BD/A/C, CD/A/B, A/BCD, B/ACD, C/ABD, and D/ABC). The combination chosen to generate the composite was the one that minimized the amount of missing data that remained in the matrix, once the composites were formed.

Matrix sizes varied according to the number of composites created. Average number of taxa and proportion of missing data included in composite matrices are given in Table 4.1, for composite matrices of different sequence lengths and levels of incompleteness.

Table 4.1. Properties of composite matrices of different sequence lengths (L), originating from missing data matrices of different levels of incompleteness (I).
Proportion of missing data (expressed as % of character missing), average and range of composite matrix sizes (i.e., number of taxa) are calculated from 1000 replicates.

L (bp)	I (%)	Missing data (%)	Average matrix size	Range
1500	5	2.7	39.4	37.0 - 41.0
	15	3.3	29.9	26.0 - 34.0
	30	7.1	23.1	22.0 - 27.0
	50	26.2	22.0	22.0 - 23.0
	75	64.0	22.0	22.0 - 22.0
8362	5	0.8	30.7	26.0 - 36.0
	15	0.9	22.4	22.0 - 25.0
	30	6.3	22.0	22.0 - 22.0
	50	23.4	22.0	22.0 - 22.0
	75	56.1	22.0	22.0 - 22.0
20 000	5	0.4	27.7	23.0 - 32.0
	15	0.9	22.1	22.0 - 24.0
	30	6.4	22.0	22.0 - 22.0
	50	23.4	22.0	22.0 - 22.0
	75	55.3	22.0	22.0 - 22.0

4.4.4. Matrix size

In order to untangle the effect of composite formation from that of matrix size reduction inherent to composite formation, we compared composite matrices to missing data matrices of equal sizes. To generate missing data matrices of reduced sizes, the taxa with the highest proportion of missing data were deleted, until we obtained a matrix with a size equal to that of the corresponding composite matrix. At least one taxon was kept in each monophyletic group.

4.4.5. Non-monophyletic composites

Non-monophyletic composites were created by selecting at random genes from two species belonging to sister groups, while keeping the other taxa complete, thus generating matrices with 41 taxa. To reproduce the unconscious error that is made when non-monophyletic composites are used, we arbitrarily labeled the non-monophyletic composite with one of the two species and thus decided that it belong to one of the two clades. However, when inferring a phylogenetic tree from such a non-monophyletic composite matrix, the non-monophyletic taxa would be, in theory, correctly placed in any two of the "parental" clades. Preliminary analyses have shown that accuracy values were indeed very low when the non-monophyletic composite taxon was kept as a terminal taxon when assessing accuracy (see results). Therefore, we removed the non-monophyletic composite taxon following phylogenetic inference, but before assessing phylogenetic accuracy of the inferred trees. Thereby, we investigated the potential effect that a non-monophyletic composite might have had on higher-level relationships during phylogenetic inference.

4.4.6. Topological accuracy of inferred trees

A phylogenetic tree was estimated from each complete, missing data or composite dataset, using two different phylogenetic methods: neighbor-joining (NJ: Saitou & Nei 1987) and maximum likelihood (ML). Missing data matrices were analyzed with the missing data coded as "?" in the matrix. NJ distances were corrected with a JC or a TVM model, with parameters identical to those used to simulate DNA sequence data, in PAUP* 4.0 (Swofford 1998). ML was used either with a JC or a TVM model, with base frequencies, proportion of invariable sites and gamma distribution parameters estimated for each dataset, using the optimized BIONJ input tree option in

PHYML 2.4.4 (Guindon & Gascuel 2003). The NJ and ML analyses were performed on ten Power Mac G5, with PowerPC 970MP processors (2 x 2.5 GHz).

To compute phylogenetic accuracy, the estimated phylogenies were compared to the model tree. We used the concept of accuracy as defined by Hillis (1995), that is, the percentage of correctly inferred trees (p). To facilitate comparison across matrices with different numbers of taxa, and because we were only interested in higher-level relationships, accuracy was restricted to the percentage of correctly inferred backbone (p_{BB}), i.e., the number of replicates for which: (1) the ten monophyletic groups were recovered (regardless of the relationships within each clade), and (2) the relationships among the clades were correctly inferred (bolded part of Fig.4.1, also shown on Fig. 4.2D). In order to determine if the inferred tree had a backbone identical to the *MT*, we constructed two different constraints in PAUP*. The first constraint verified that all ten groups were monophyletic, and the second, that the relationships among the ten groups were accurately recovered. To accommodate matrices of different sizes, the constraints had to be modified for each replicate, to represent the taxa in the composite matrix and in missing data matrices of reduced sizes. Phylogenetic accuracy was calculated as the number of inferred trees that satisfied both constraints. Significant differences in accuracy values between different datasets were assessed using a Pearson's Chi-squared test in the R Stats Package (R Development Core Team 2009). In order to keep the experimentwise error rate at a 0.05 level, we applied a Bonferroni correction for multiple tests (Rice 1989), by dividing the α value by the number of comparisons made in each of the 18 simulation cases (i.e., ten or 20 comparisons: see Table 4.2). One thousand replicates were analyzed except for larger matrices for which only 100 replicates were analyzed with ML (see results).

4.5. RESULTS

To compare missing and composite matrices, 18 different cases were analyzed (Table 4.2). For clarity, these different cases are labeled 1 to 18. Each case represents one of the three sequence lengths simulated under a JC or ML model (Table 4.2A, B vs. 4.2C, D), either on MT_s or MT_L (Table 4.2A, C vs. 4.2B, D). Both NJ and ML were used as inference methods. In Table 4.2A and B, where DNA sequences were simulated under a JC model, a JC correction was used to infer the tree (i.e., NJ-JC and ML-JC). In table 4.2C and D, where DNA sequences were simulated under a TVM model, both the JC

and TVM corrections were used to analyze the datasets (i.e., NJ-JC, ML-JC, NJ-TVM and ML-TVM).

Table 4.2. Phylogenetic accuracy (ρ_{BB}) of phylogenetic trees inferred from complete, missing data and composite matrices of different sizes (1500, 8362 and 20 000bp) and different levels of incompleteness (I).

Complete (0% missing) and missing data (5 to 75% missing) matrices included 42 taxa while composite matrices were of varying sizes (as presented in Table 4.1). DNA sequences were simulated under : A. a Jukes-Cantor (JC) model on a model tree with short terminal branch lengths (MT_S), B. a JC model on a model tree with long terminal branch lengths (MT_L), C. a transversional model following a gamma distribution with invariant sites (TVM) model on MT_S, and D. a TVM model on MT_L. Two different phylogenetic methods (NJ: neighbor-joining, and ML: maximum likelihood) were used to analyze the datasets with a JC or TVM correction (i.e., NJ-JC, NJ-TVM, ML-JC, and ML-TVM). The 18 different simulation cases are identified by circled numbers to facilitate their references in the text. Accuracy values for composite matrices were compared to corresponding missing data matrices, using a Chi-squared test adjusted with a Bonferroni correction. Shaded cells represent values that were significantly higher. Accuracy values that did not remain significant when composite matrices were compared to missing data matrices of equal sizes are marked with an asterisk. Phylogenetic accuracy was calculated from 1000 replicates except for cases 9, 12, 14, 15, 17 and 18, where only 100 replicates were analyzed with the ML method.

A. JC model on MT_S

		① 1500bp		② 8362bp		③ 20 000bp	
	I (%)	Missing	Compo	Missing	Compo	Missing	Compo
NJ-JC	0	98.5	-	100.0	-	100.0	-
	5	96.2	96.8	100.0	100.0	100.0	100.0
	15	87.8	95.9	100.0	100.0	100.0	100.0
	30	63.0	87.8	99.9	100.0	100.0	100.0
	50	33.6	47.6	94.7	98.2	99.9	100.0
	75	25.7	22.1	59.0	54.6	93.7	87.5
ML-JC	0	100	-	100.0	-	100.0	-
	5	99.6	99.5	100.0	100.0	100.0	100.0
	15	99.0	99.3	100.0	100.0	100.0	100.0
	30	96.1	99.5	100.0	100.0	100.0	100.0
	50	82.4	87.1	99.7	99.9	100.0	100.0
	75	47.2	43.7	91.9	92.2	99.9	100.0

B. JC model on MT_L

		④ 1500bp		⑤ 8362bp		⑥ 20 000bp	
	I (%)	Missing	Compo	Missing	Compo	Missing	Compo
NJ-JC	0	49.0	-	99.1	-	100.0	-
	5	37.7	39.5	98.0	97.6	100.0	100.0
	15	17.8	34.3	95.9	94.9	99.9	99.2
	30	4.2	16.6	82.5	82.0	99.5	96.2
	50	0.3	0.8	38.4	30.4	88.4	59.9
	75	0.3	0.1	8.2	0.4	34.3	2.8
ML-JC	0	89.9	-	99.9	-	100.0	-
	5	86.8	85.8	99.9	99.9	100.0	100.0
	15	74.7	71.6	99.9	99.5	100.0	100.0
	30	49.7	49.5	99.7	98.8	100.0	100.0
	50	17.6	16.8	92.2	88.4	99.9	99.9
	75	5.9	0.8	52.5	20.8	92.3	96.8

C. TVM model on MT_S

		⑦ 1500bp		⑧ 8362bp		⑨ 20 000bp	
	I (%)	Missing	Compo	Missing	Compo	Missing	Compo
NJ-JC	0	55.7	-	99.7	-	99.9	-
	5	46.9	51.9	97.3	94.8	99.7	98.5
	15	28.5	45.9	94.9	93.5	99.8	99.0
	30	10.0	32.3	91.1	91.4	99.1	97.8
	50	1.2	6.6	64.9	79.5	93.6	93.6
	75	0.4	0.4	17.0	16.6	59.0	53.9
ML-JC	0	68.7	-	100.0	-	100.0	-
	5	65.1	66.3	100.0	99.9	100.0	100.0
	15	55.1	64.3	99.9	99.8	100.0	100.0
	30	31.2	56.2	99.6	99.7	100.0	100.0
	50	7.7	19.7	90.9	95.9	100.0	100.0
	75	0.3	0.9	39.8	38.0	88.0	90.0
		⑩		⑪		⑫	
NJ-TVM	0	59.2	-	99.6	-	100.0	-
	5	48.4	52.2	99.5	99.6	100.0	100.0
	15	27.8	49.2	98.7	99.6	100.0	100.0
	30	8.5	35.3	94.3	97.7	100.0	99.9
	50	0.4	6.4	65.4	86.9	96.0	99.7
	75	0.1	0.5	16.2	15.8	57.9	54.6
ML-TVM	0	84.6	-	100.0	-	100.0	-
	5	77.4	82.0	100.0	100.0	100.0	100.0
	15	71.6	77.5 *	100.0	100.0	100.0	100.0
	30	48.3	69.4	99.8	99.6	100.0	100.0
	50	15.3	26.6	96.1	98.1	100.0	100.0
	75	1.5	2.0	50.9	47.9	97.0	93.0

D. TVM model on MT_L

		(13) 1500bp		(14) 8362bp		(15) 20 000bp	
	I (%)	Missing	Compo	Missing	Compo	Missing	Compo
NJ-JC	0	0.0	-	0.0	-	0.1	-
	5	0.0	0.0	0.0	1.0	0.1	3.4 *
	15	0.0	0.0	0.0	4.0	0.3	8.3
	30	0.0	0.0	0.0	4.0	0.6	15.5
	50	0.0	0.0	0.0	0.0	0.0	2.1
	75	0.0	0.0	0.0	0.0	0.0	0.0
ML-JC	0	0.0	-	4.0	-	14.0	-
	5	0.0	0.0	3.0	16.0	10.0	27.0
	15	0.0	0.0	1.0	26.0	8.0	36.0 *
	30	0.0	0.0	2.0	28.0	2.0	47.0
	50	0.0	0.0	0.0	8.0	0.0	16.0
	75	0.0	0.0	0.0	0.0	0.0	1.0
		(16)		(17)		(18)	
NJ-TVM	0	0.2	-	52.6	-	87.1	-
	5	0.1	0.2	43.2	41.9	84.8	89.3
	15	0.0	0.3	25.6	38.4	73.4	80.2
	30	0.0	0.0	5.8	28.4	42.4	67.3
	50	0.0	0.0	0.1	2.6 *	8.1	18.4
	75	0.0	0.0	0.0	0.0	0.0	0.0
ML-TVM	0	5.4	-	94.0	-	98.0	-
	5	3.7	3.4	90.0	86.0	98.0	98.0
	15	0.9	3.8 *	84.0	75.0	95.0	90.0
	30	0.0	2.0	53.0	67.0	89.0	93.0
	50	0.0	0.1	13.0	30.0	69.0	72.0
	75	0.0	0.0	2.0	0.0	10.0	2.0

In addition, for each of these 18 cases, five different levels of incompleteness (I) were analyzed with both NJ and ML, for a total of 180 comparisons between composite and missing data matrices.

4.5.1. Model tree (MT_S vs. MT_L)

For complete, missing data and composite matrices, phylogenetic accuracy values (p_{BB}) obtained with DNA sequences simulated on MT_S (model tree with short terminal branch lengths) were similar to those obtained from datasets simulated on MT_L with a JC model of evolution (Table 4.2A, C vs. 4.2B). However, when matrices simulated under an identical model of evolution were compared, accuracy values were higher for MT_S

datasets than for MT_L datasets (Table 4.2A vs. B, and 4.2C vs. D). Accuracy values were much lower for MT_L datasets simulated under a more complex model of DNA substitution (Table 4.2D), except for larger datasets (8362 and 20 000bp) analyzed with ML-TVM (and to a lesser extent with NJ-TVM: cases 17 and 18). Shorter sequence datasets (i.e., 1500bp) generally failed to recover the MT backbone, regardless of the inference method used (ρ_{BB} ranging from 0 to 5.4%: cases 13 and 16).

Preliminary analyses have revealed that branch length variations had an impact on phylogenetic accuracy, especially at more extreme branch length ratios (results not shown). For both complete and composite matrices, the phylogenetic accuracy decreased as a function of branch length ratios (i.e., internal branch length/terminal branch length). For more extreme ratios (1:8 to 1:10), accuracy values were very low (smaller than 15%). On the contrary, when branches among clades were longer relative to those within clades, the backbone of the model tree was always recovered.

4.5.2. DNA sequence length (L)

In all cases, we observed an increase in ρ_{BB} with an increase in DNA sequence lengths (Table 4.2). For shorter sequences (i.e., 1500 bp), ρ_{BB} ranged from 0 to 100%. For matrices of that length, the highest accuracy was observed for DNA sequences simulated under a JC model on MT_S (case 1), whereas the lowest accuracy was observed with DNA sequences simulated under TVM on MT_L (cases 13 and 16). In such cases, complete, composite and missing data matrices could barely recover the MT backbone, regardless of the inference method (ρ_{BB} ranging from 0 to 3.8%). Much higher ρ_{BB} were observed for longer sequence lengths. Close to 100% accuracies were obtained for sequences of 20 000bp (cases 3, 6, 9 and 12). Accuracy values obtained for case 18 were also close to 100%, but only when datasets were analyzed with ML-TVM (up to 98.0% accuracy), and to a lesser extent with NJ-TVM (up to 89.3% accuracy). Datasets with an intermediate sequence length (8362bp) showed accuracy values similar or slightly lower than those obtained with sequence length of 20 000bp.

4.5.3. Model of DNA evolution

We observed similar trends in phylogenetic accuracies for datasets simulated under both evolutionary models (JC and TVM: Table 4.2A, B, C and D). However, ρ_{BB} for datasets simulated using the JC model were higher than ρ_{BB} values for the

corresponding datasets simulated using the TVM model. The decrease in accuracy for datasets simulated under the TVM model was even more pronounced for datasets of shorter sequence lengths and simulated on MT_L.

4.5.4. Level of matrix incompleteness (I)

In general, complete matrices (I = 0%) showed higher accuracy values than the corresponding composite and missing data matrices. In some cases (i.e., I = 5%), p_{BB} obtained for complete matrices were very close or equal to that of composite and missing data matrices. At best (cases 2, 3, 6, 9 and 12), 100% accuracy was recorded for nearly every level of incompleteness and for both inference methods. For composite and missing data matrices, p_{BB} decreased as a function of the proportion of missing data, except in two cases (14 and 15), where accuracy values significantly increased with missing data (at I = 5 to 30%). Interestingly, this only occurred for DNA sequences simulated under a TVM model on MT_L, and analyzed with NJ-JC (the increase observed with ML-JC is not significant). In most cases, accuracy dropped drastically at higher level of incompleteness (I = 50 or 75%).

4.5.5. Inference method

In general, NJ and ML results exhibited identical trends for the different parameters. For cases where accuracy values were not 100% (i.e., datasets with shorter DNA sequences or high proportion of missing data), ML performed better than NJ. However, in Table 4.2D, NJ performed better than ML, but only when NJ-TVM results are compared to ML-JC. Accuracy values obtained for ML analyses using a more complex evolutionary model (TVM) were always greater than when the JC model was used (Table 4.2C and D: ML-TVM vs. ML-JC).

4.5.6. Comparison of composite and missing data matrices

Significantly higher p_{BB} were obtained for composite matrices relative to the corresponding missing data matrices, mostly at intermediate levels of incompleteness (46 comparisons out of 180: shaded cells in Table 4.2). On the other hand, missing data matrices performed significantly better than composite matrices in only eight comparisons, which were all simulated under a simple model of evolution and with high levels of missing data (Table 4.2A and B). It appears that composites outperformed missing data matrices more often in cases where shorter sequences were analyzed

(cases 1, 4, 7 and 10). For larger datasets (i.e., 8362 and 20 000 bp), composite matrices significantly outperformed missing data matrices for at least one level of incompleteness in most cases. Significant comparisons for TVM datasets simulated on MT_S involved primarily composite matrices at I = 50%, whereas significant comparisons were observed for a greater range of I (from 5 to 50%) for TVM datasets simulated on MT_L. In Table 4.2C and D, the benefit gained from the analysis of composites disappeared when the matrices were analyzed with ML-TVM.

4.5.7. Matrix size

For all of the 46 comparisons where composite matrices performed significantly better, we reanalyzed missing data matrices by deleting some taxa in order to compare matrices of equal sizes (see Table 4.1). When the reduced missing data matrices were compared to the original missing data matrices that included all taxa (n = 42), accuracy values increased significantly in 11 cases (out of 46), while it significantly decreased in eight cases. When the reduced missing data matrices were compared to the corresponding composite matrices, 41 comparisons out of 46 remained significantly different. From the five comparisons that did not remain significant (marked with an asterisk in Table 4.2), four were observed for datasets simulated under a TVM model on MT_L, where both the missing data and composite approaches performed rather poorly (accuracy values ranging from 0 to 36%: Table 4.2D).

4.5.8. Non-monophyletic composite

As predicted, we noted an important decrease in accuracy values when the monophyly condition was violated, i.e., when the non-monophyletic composite was kept as a terminal taxon when assessing accuracy. Even though only one non-monophyletic composite was created per matrix, accuracies only reached a maximum of 55% for the 18 cases tested (results not shown). However, a large part of the reduction in accuracy can be explained by the incompatibility of the inferred trees with the first constraint (testing the monophyly of the clades). When the non-monophyletic composite taxon was deleted before scoring the inferred trees for phylogenetic accuracy, the values greatly increased. Table 4.3 reports accuracy values obtained from non-monophyletic composite matrices, where the non-monophyletic composite taxon was pruned from the inferred trees. Accuracy values in such cases are highly similar to values obtained from

the analysis of composite matrices with low level of incompleteness (i.e., from 5 to 30%).

Table 4.3. Phylogenetic accuracy (ρ_{BB}) of phylogenetic trees inferred from matrices containing one non-monophyletic composite.

The non-monophyletic taxon was pruned from the tree before assessing accuracy. Average accuracy values were calculated from 1000 replicates except for datasets corresponding to cases 9, 12, 14, 15, 17 and 18, in Table 4.2, where only 100 replicates were analyzed for the ML analysis. Abbreviations as in Table 4.2.

Evolution Parameters	Inference Method	L (bp)		
		1500	8362	20 000
JC model on MT_S	NJ-JC	96.1	99.9	100.0
	ML-JC	99.1	100.0	100.0
JC model on MT_L	NJ-JC	41.7	93.6	97.3
	ML-JC	86.1	98.3	98.8
TVM model on MT_S	NJ-JC	50.0	90.1	92.8
	ML-JC	64.7	99.6	100.0
	NJ-TVM	54.3	97.8	99.5
	ML-TVM	82.6	99.0	100.0
TVM model on MT_L	NJ-JC	0.0	0.0	1.0
	ML-JC	0.0	2.0	10.0
	NJ-TVM	0.0	42.3	79.2
	ML-TVM	7.0	88.0	96.0

4.6. DISCUSSION

4.6.1. Model tree (MT_S vs. MT_L)

In our study, two different branch length ratios were used on the same model tree topology (MT_S and MT_L). Our results indicate a better phylogenetic accuracy for trees inferred from datasets simulated on MT_S, when an identical model of DNA evolution

was used (Table 4.2A vs. B, and 4.2C vs. D). This increase in accuracy values may be explained by the stemminess of the tree, which is the relative length of internal to external branches (Fiala & Sokal 1985). It has been shown that trees with low stemminess (i.e., longer terminal branch lengths) are harder to estimate. Indeed, lower accuracy values were observed in previous simulation studies for such trees (Fiala & Sokal 1985, Rokas et al. 2005, Weisrock et al. 2005). This trend was also observed in our study, with a decrease in phylogenetic accuracy when external branch lengths were increased relative to internal branch lengths (i.e., MT_L). The decrease in accuracy observed with MT_L datasets might also be due to long-branch attraction (Felsenstein 1978, Hendy & Penny 1989), which is most likely to occur for datasets simulated on MT_L.

4.6.2. DNA sequence length (L)

Our results support studies suggesting that the number of characters included in the analysis is more important than the amount of missing data (e.g., Wiens 2003b, Philippe et al. 2004). For identical level of incompleteness, we observed a significant increase in phylogenetic accuracy for larger datasets (i.e., values in each row of Table 4.2). For a matrix size of 20 000bp, we obtained accuracy values close to 100% at all levels of incompleteness, except in two cases (cases 15 and 18, TVM model on MT_L). As the number of characters increases, the impact of missing data decreases, as long as there is sufficient phylogenetic information in the incomplete taxa (Delsuc et al. 2005).

4.6.3. Model of DNA evolution

Two different models of DNA evolution were used to simulate datasets (JC, and TVM + Γ + I). Matrices with DNA sequences simulated under a JC model of evolution always provided higher accuracy values relative to TVM datasets (Table 4.2A, B vs. 4.2C, D). These results support the claim that phylogenetic inference is more difficult with sequences that evolved under a more complex model (e.g., Yang 1996a, Pollock & Bruno 2000). Based on this observation, Zwickl and Hillis (2002) disapproved the use of overly simplistic models of DNA substitutions in simulation studies. Simpler models of DNA evolution usually produce accuracy values close to 100%, unless simulations also involve other parameters such as an extremely low stemminess of the tree. Additionally,

more complex models are certainly more appropriate to represent real DNA sequence evolution.

4.6.4. Level of matrix incompleteness (I)

The proportion of missing data in phylogenetic datasets varies from study to study and the distribution of missing data is matrix-specific (see review by Kearney 2002). In our study, different levels of incompleteness were explored (5 to 75%), which appropriately covers the proportion of missing data observed with real sequence data. The DNA sequence matrix analyzed by Malia et al. (2003) was characterized by a fairly large amount of missing data (40%), but numerous phylogenomic studies have been published recently with an even greater proportion of missing data. Philippe et al. (2004) have observed maximal accuracy values (100%), when 50% of the data matrix was missing, but with a large dataset (30000 amino acid positions). Studies with as much as 92% missing data can still contain significant information about evolutionary relationships (Driskell et al. 2004), although some taxonomic groups may be more affected by large amount of missing data (Philippe et al. 2005b). Wiens et al. (2005) have obtained strong support for the placement of incomplete species in their expected clade, with as much as 90% missing data. Furthermore, Wiens (2006) suggested that, with a 2000bp matrix, accuracy would remain high "even when half of the taxa have 90% of their data cells lacking data". In our study, smaller size matrices (i.e., 1500bp) failed to recover the MT backbone at every levels of incompleteness, when matrices were simulated under a TVM model on MT_L.

Still, for datasets of equal sizes, an increase in the amount of missing data will negatively affect accuracy, given that phylogenetic information is reduced. Indeed, a negative correlation between the proportion of missing data and accuracy values has often been reported in the literature (see review by Wiens 2006). We have also observed a decrease in phylogenetic accuracy with increasing levels of incompleteness, both for missing data and composite matrices. Phillips et al. (2004) have shown that longer DNA sequences amplified the potential for systematic errors, in which case, missing data could contribute positively to accuracy by decreasing dataset sizes. However, we do not believe that this happened in our analyses, since accuracy increased with DNA sequence lengths and because we did not observe higher accuracy values when missing data matrices were compared to complete matrices.

4.6.5. Inference method

Two different inference methods were used (JC and ML). Similar trends were obtained with both methods, although accuracy values were higher with the ML method for shorter sequence lengths or at high levels of missing data. Both methods performed well with longer DNA sequences (i.e., 8362 and 20 000bp), except when the datasets were simulated under a TVM model on a MT_L tree (Table 4.2D). In this later case, both JC and ML failed to recover the MT backbone, except when the appropriate model (TVM) was used to infer the tree. NJ-TVM performed better than ML-JC, most likely because the appropriate corrected distance was used whereas the simpler model was unsuitable to analyze the TVM data. Indeed, Posada & Crandall (2001) argued that a primordial aspect of phylogenetic inference is to use an optimized model that fits adequately the model of nucleotide substitution of the data.

4.6.6. Comparison of composite and missing data matrices

Composite taxa are widely used in phylogenetic analysis (e.g., Shoshani & McKenna 1998, Madsen et al. 2001, Murphy et al. 2001b, Scally et al. 2001, Asher et al. 2004). However, Malia et al.'s (2003) study was the first to directly compare trees estimated from a data matrix with and without missing data, where missing data were avoided by generating composite taxa. Their study demonstrated that missing data matrices produced a less resolved tree, a phenomenon also observed in other studies (e.g., Wiens & Reeder 1995, Kearney 2002, Flynn et al. 2005), whereas composite matrices suggested relationships not supported by the underlying data. To the contrary, our simulations clearly demonstrate the validity of composite taxa in phylogenetic analysis. In the vast majority of comparisons analyzed, no significant difference in phylogenetic accuracy was observed between composite and missing data matrices. Furthermore, in some cases, composite matrices performed significantly better than missing data matrices. The reverse is also true, although it occurred less frequently.

Composite matrices generally outperformed missing data matrices for comparisons of datasets with short DNA sequences (i.e., 1500bp), but also in many cases with longer DNA sequences (i.e., 8362 and 20 000 bp). The significant differences were observed mostly at intermediate levels of incompleteness. At higher levels, accuracy was so low for both approaches (close to 0%) that we were not able to detect any differences. Also, the proportion of missing data remained very high for most composite datasets at

I = 75% (from 55 to 64%: Table 4.1), and thus might conceal the benefit of composite formation. On the other hand, at a lower level of incompleteness (I = 5%), we rarely observed any significant difference between the two approaches. At that low level of incompleteness, fewer composites are formed and both approaches are expected to provide accuracy values similar to that of complete matrices given the small amount of missing data. Indeed, this was observed for most cases in Table 4.2 (when comparing accuracy values at I = 0% vs. 5%).

4.6.7. Matrix size

It has been suggested that increased taxon sampling has a positive effect on phylogenetic accuracy (Hillis et al. 1993, Lecointre et al. 1993, Graybeal 1998, Hillis 1998, Rannala et al. 1998, Pollock et al. 2002, Zwickl & Hillis 2002). However, the reverse trend has also been suggested (Poe & Swofford 1999, Rokas et al. 2003b, Rokas & Carroll 2005). Therefore, it was important to discriminate the effect of matrix size reduction to that of composite formation. Out of 46 comparisons where composites outperformed missing data matrices, 41 comparisons remained significant when missing data matrices were reduced to the size of the corresponding composite matrices. Four of the five comparisons that did not remain significant were cases for which phylogenetic accuracy was very low and where both methods performed poorly. In light of these results, we do not believe that the better performance of composite matrices (with respect to missing data matrices) can be explained by a reduction in matrix size.

4.6.8. Non-monophyletic composite

As strongly advocated by Scally et al. (2001) and Springer et al. (2004a), the composite approach will infer correct relationships, as long as the combined species belong to a true monophyletic group relative to the level at which relationships are inferred (monophyly assumption). However, except for simulations or experimental studies (e.g., Hillis et al. 1992, Sanson et al. 2002), phylogeneticists do not know the "true" phylogenetic tree and have no guarantee of monophyly for any groups. When creating composite taxa, monophyletic groups are defined based on general knowledge and on previous phylogenetic studies. As a result, it may occur that a non-monophyletic composite is created involuntarily based on erroneous knowledge, and that the non-monophyletic composite will not be noticed. Thus, we tested the effect of violating the

monophyly condition when constructing composite taxa by including one non-monophyletic composite in the dataset. We have observed a marked decrease in accuracy when non-monophyletic composites were included in the analysis: a maximal accuracy of 55%, when only one non-monophyletic composite was included in the matrix. Although, these simulations confirmed the importance of sampling species within a monophyletic group, we also observed that the problem caused by the non-monophyletic taxon seems to be restricted to the location of the "parental" clades on the tree. Indeed, when the non-monophyletic taxon was deleted before assessing accuracy, we did not observe a drastic decrease in accuracy when compared to composite matrices. Thus, the effect of including a non-monophyletic taxon in an analysis is probably not too problematic when one is interested mainly in higher-level relationships. However, we have created non-monophyletic taxa from sister clades and thus the effect on higher-level relationships might be more pronounced when more distant clades are combined.

4.6.9. Pros and cons of both approaches

Besides reducing the number of missing entries, the most important advantage of using composite matrices is certainly the reduction in computing time. As reported in Table 4.1, composite matrices contain fewer taxa, and the speed of the analysis is greatly increased, especially for ML analyses. In our simulations, the analysis of one dataset (i.e., one replicate) with the ML-TVM method lasted about 2 hours for missing data matrices, compared to 45 minutes for the corresponding composite matrix (I = 15%, L = 20 000bp).

On the other hand, one advantage of the missing data approach is that it does not require any assumption of monophyly at a level higher than the species. With the composite approach, prior knowledge of phylogenetic relationships is required and there is no guarantee that the species combined truly belong to a monophyletic group. Whereas the composite approach is restricted to the inference of relationships above the taxonomic level at which composites are formed, relationships within higher-level groups can be resolved with the missing data approach.

When forming composite taxa, the number of species representing a higher-level taxon automatically decreases. It is probably important to ensure that more than one

representative per clade are included in the analysis. In a simulation study, Wiens (1998b) has shown that sampling a single species to represent higher-level taxa often causes a drastic reduction in phylogenetic accuracy, and he strongly recommended sampling multiple species per taxa when inferring higher-level relationships.

4.7. CONCLUSIONS

In summary, composite matrices performed significantly better than missing data matrices in 46 situations: with matrices having few characters and to some extent with larger matrices depending on the phylogenetic method and/or the phylogenetic model used to infer the tree. Given sufficient data and an adequate inference method, both missing data and composite matrix approaches exhibited similar accuracy values, for all levels of incompleteness. In a situation that would most likely represent a real dataset (e.g., 20 000bp matrix simulated on the MT_L and analyzed with ML-TVM: case 18) both approaches showed reasonable phylogenetic accuracy values for up to 50% of missing data (69% and 72%, at I = 50%). Overall, we have shown that composite matrices perform as well as missing data matrices under various evolutionary conditions, and that they generally outperformed missing data matrices in sub-optimal phylogenetic conditions (e.g., short DNA sequences). The lower performance observed by Malia et al. (2003) in their reanalysis of Madsen et al.'s (2001) composite matrix might be explained by the inclusion of species that did not have any common sequences. In our simulations, we ensured that all species shared at least one gene. Also, a problem, in Madsen et al.'s (2001) analyses, was the inclusion of non-monophyletic composites. Although, our simulations tend to show that including one non-monophyletic composite do not strongly affect higher-level relationships.

With increasing number of sequences available in public databases, different taxa can be selected to generate composites. In light of our results, we believe that the creation of composite taxa represents a valuable approach to reduce the amount of missing data in DNA sequence matrices. The composite approach will probably be increasingly used as more taxa and sequences become available for large-scale phylogenomic studies (Eisen & Fraser 2003, Telford 2007). We suspect that our conclusions could apply to other types of matrices as well (e.g., morphological data) although it should be further investigated.

4.8. ACKNOWLEDGMENTS

We would like to thank three anonymous reviewers and members of the Laboratoire d'Écologie Moléculaire et d'Évolution (LEMEE) for their constructive comments on a preliminary version of this manuscript. For the phylogenetic analyses, we used the computational resources located in the Laboratoire Interfacultaires de Micro-Informatique de l'Université de Montréal and we thank Marie-Hélène Duplain for granting access to the lab outside business hours. This study was supported by NSERC and FQRNT scholarships to VC and by NSERC grant OGP0155251 to FJL.

CHAPITRE 5:

HIGHER-LEVEL PHYLOGENIES: APPLICATION OF COMPOSITE TAXA AND SUPERTREE TO MAMMALIAN MITOGENOMIC SEQUENCES

Cet article sera soumis prochainement:

Campbell V. & Lapointe F.-J. Higher-level phylogenies: Application of composite taxa and supertree to mammalian mitogenomic sequences. BMC Evolutionary Biology.

5.1. RÉSUMÉ

Deux approches différentes peuvent être utilisées en phylogénomique: l'analyse combinée ou séparée. Avec la première approche, les jeux de données sont combinés dans une super-matrice alors qu'avec la deuxième approche, les jeux de données sont d'abord analysés séparément et les arbres phylogénétiques estimés sont ensuite combinés pour former un super-arbre. Toutefois, les deux approches ont des lacunes. Les super-matrices sont souvent caractérisées par un grand nombre de données manquantes. Une approche qui permet de réduire le nombre de données manquantes consiste à créer des taxons chimères. Ces taxons sont formés par la combinaison de séquences provenant de différentes espèces formant ainsi une séquence chimère qui maximize le nombre de gènes. Bien que cette approche soit de plus en plus utilisée, sa performance a rarement été testée. La méthode de super-arbre est une alternative intéressante pour éviter les données manquantes puisque les jeux de données, qui sont analysés séparément, n'ont pas besoin de représenter des taxons identiques. Toutefois, les approches de super-arbre et de consensus ont été très critiquées; on leur reproche de ne pas fournir d'hypothèses phylogénétiques valables.

Dans cette étude, la congruence des arbres phylogénétiques estimés par différentes approches a été comparée. Des arbres modèles ont été obtenus à partir de séquences mitochondriales complètes de 102 espèces de mammifères, représentant 93 familles. À partir des séquences complètes, des matrices avec des données manquantes ont été générées et analysées telles quelles ou alors en réduisant la quantité de données manquantes en formant des taxons chimères. Les 12 gènes mitochondriaux du brin lourd ont également été analysés séparément et les arbres phylogénétiques obtenus ont été combinés avec une méthode de consensus ou de super-arbre. En moyenne, les matrices avec données manquantes et les matrices avec taxons chimères présentaient un même niveau de congruence lorsque comparées avec les arbres modèles. De plus, la congruence diminuait avec l'augmentation du nombre de données manquantes et ce, avec les deux approches. Parmi les trois types de méthodes de consensus, les méthodes de super-arbre qui tiennent compte des longueurs de branches étaient supérieures à quelques méthodes de consensus topologiques (qui ne tiennent pas compte des longueurs de branches) et qui produisent des arbres peu résolus (i.e, le consensus strict, le consensus d'Adams et le consensus majoritaire qui inclut uniquement les groupes présents dans plus de 50% des arbres). La méthode de MRP

qui est communément utilisée dans les analyses de type super-arbre était équivalente aux autres méthodes de type super-arbre (sauf pour la méthode du *most similar supertree* (MSS) qui était relativement peu performante). Certaines des méthodes comparées dans cette étude sont très efficaces d'un point de vue de leur exactitude et de leur rapidité de calcul et seraient donc mises à profit dans un contexte phylogénomique.

5.2. ABSTRACT

Two different approaches can be used in phylogenomics: the combined or the separate analysis. In the first approach, different datasets are combined in a supermatrix. In the second, datasets are analyzed separately and the phylogenetic trees are combined in a supertree. However, both approaches have caveats. Supermatrices are often caracterized by a large amount of missing data. One possible approach to minimize missing data is to create composite taxa. These taxa are formed by sampling sequences from different species in order to obtain a composite sequence that includes a maximum number of genes. Although this approach is increasingly used, its accuracy has rarely been tested and some authors prefer to analyze incomplete supermatrices by coding unavailable sequences as missing. The supertree method is an interesting alternative to avoid missing data since datasets are analyzed separately and do not need to represent identical taxa. However, the supertree approach and the corresponding consensus methods have been highly criticized for not providing valid phylogenetic hypotheses.

In this study, congruence of trees estimated by different approaches were compared. Model trees were obtained from the analysis of complete mitochondrial sequences of 102 mammal species representing 93 families. From the complete DNA sequence matrix (i.e., the 12 mitochondrial genes of the H-strand), missing data matrices were generated and analyzed as is or by reducing the amount of missing data through the formation of composite taxa. Individual gene datasets were also analyzed separately and the resulting phylogenetic trees were combined with a consensus or a supertree method. On average, missing data and composite matrices showed similar congruence to model trees, which decreased as missing data increased. Among the three types of consensus methods compared, supertree methods that take into account branch lengths were superior to some but not all other consensus methods (i.e., topological

methods). Most consensus methods produced poorly resolved consensus trees (i.e., strict, Adams and majority rule that only includes groupings above 50% frequency) and did not performed well. Supertree methods accounting for branch lengths produced fully resolved trees highly congruent with model trees. Among topological supertree methods, matrix representation with parsimony (MRP) performed equally well to branch-length supertree methods. The most similar supertree method (MSS) was the least congruent with model trees. We conclude that some of the methods tested are worth considering in a phylogenomic context since they performed well and reduce computing time.

5.3. INTRODUCTION

The phylogenomic era has brought a shift from single to multiple datasets (or genes) to study phylogenetic relationships (Rokas et al. 2003b). Inferring phylogenies from larger matrices (i.e. supermatrices) decreases stochastic errors since the increased number of characters better represents the source population. Larger matrices also contain increased phylogenetic signal, which minimizes the effect of homoplasious characters (Driskell et al. 2004, Rodriguez-Ezpeleta et al. 2005, Dunn et al. 2008). However, phylogenomic studies present numerous methodological challenges (see review by Delsuc et al. 2005). For example, systematic errors (such as long-branch attraction), which are due to non-phylogenetic signal can be exacerbated when a large number of characters are used (Hillis et al. 2003, Delsuc et al. 2005, Rokas & Carroll 2005, Rodriguez-Ezpeleta et al. 2007, Nishihara et al. 2007).

Another drawback of analyzing supermatrices is the amount of missing data. The concatenation of a large number of genes from different species often results in a supermatrix with incomplete data for some species. A taxon bias is observed in sequence databases, with a large number of genes (or whole genome) sequenced for a few key species (Crandall & Buhay 2004, Driskell et al. 2004, Philippe et al. 2005a, Wiens 2006, Telford 2008). Different methods have been proposed to analyze matrices with rnissing data, e.g., to remove taxa with the largest number of missing entries or to code the data as missing (see reviews by Wiens & Reeder 1995, Wiens 2006). Other approaches have been developed to estimate missing nucleotides in DNA sequences prior to the analysis, as implemented in the probabilistic estimation of missing value (PEMV: Diallo et al. 2006). Missing data has been described as a source of systematic

error (Huelsenbeck 1991, Kearney 2002). However, computer simulations have shown that the misplacement of an incomplete taxon on a phylogenetic tree is often due to the reduced number of characters (and thus are described as stochastic errors) rather than to the missing data. Therefore, supermatrices should not be affected by missing data as long as a sufficient number of informative characters is sampled (Rosenberg & Kumar 2003, Wiens 2003b, 2005, 2006, Driskell et al. 2004, Philippe et al. 2004, 2005a, Wiens et al. 2008).

Another approach to decrease the amount of missing data in supermatrices is to construct composite taxa (e.g., Shoshani and McKenna 1998, Murphy et al. 2001a, Scally et al. 2001, Asher et al. 2004, Springer et al. 2004a, Poux et al. 2006, Telford 2007, Beck 2008). In order to create such composite taxa (or chimeric taxa), the sequences from different species are combined within a monophyletic group defined *a priori* (Shoshani and McKenna 1998, Springer et al. 2004a, Campbell & Lapointe In press: chapter 4). The species are chosen according to the genes that have been sequenced or that are lacking, so as to minimize the amount of missing data once the composite sequence is formed. Therefore, supermatrices may include a different number of composite taxa, depending on the number of available sequences and the amount of missing data in each sequence. For example, some studies only included one composite taxon in their analyses (e.g., Flynn et al. 2005, Marek & Bond 2006), whereas others incorporated a large number of composite taxa (e.g., 12 composites /28 taxa: Madsen et al. 2001; 25 /52: Philippe et al. 2007; 6 /58: Duvall et al. 2008) or even, only composite taxa (e.g. 38 species to form 14 composite taxa; Delsuc et al. 2006, and 168 species to form 37 composite taxa; Bourlat et al. 2008). In these studies, phylogenetic relationships are inferred at a higher level than the one corresponding to the composite taxa.

In an evaluation of the effect of composite taxa on phylogenetic accuracy, Malia et al. (2003) concluded that the use of composite taxa can suggest misleading evolutionary relationships. Therefore, they advised analyzing incomplete data matrices with missing data instead of forming composite taxa. However, a simulation study that compared phylogenetic accuracy obtained from the two types of matrices revealed that equivalent phylogenetic accuracy was obtained when using either one of the two competing approaches (Campbell & Lapointe In press: chapter 4). The same study suggested that composite taxa might be preferred in a phylogenomic context, since it reduces

drastically the computing time of the phylogenetic analyses. The opposing conclusions obtained by Malia et al. (2003) and Campbell & Lapointe (In press: chapter 4) may be because the first study used DNA sequences from mammalian species, whereas the second relied upon simulated DNA sequences. Simulation studies are useful to test the accuracy of phylogenetic methods under controlled conditions and with fixed parameters (Hillis 1995, Huelsenbeck 1995, Wiens 1998c). However, the complexity of the underlying substitution processes is still not fully understood and theoretical models of evolution often represent a simplified version of DNA sequence matrices. Also, only a limited number of parameters can be tested in any given study.

Another approach that can be applied to deal with incomplete matrices is to use a supertree method (Sanderson et al. 1998, Bininda-Emonds et al. 2002, Bininda-Emonds 2004a, b). In the likely event that some gene sequences are not available for all species, it is possible to estimate a phylogenetic tree for each gene separately and then combine the resulting trees with a supertree approach. Supertrees include all taxa and thus allow datasets with overlapping sets of taxa to be combined. Numerous supertree methods have been developed, the most familiar being matrix representation with parsimony (MRP: Baum 1992, Ragan 1992a, b). They can be defined as a generalization of consensus methods, which only apply to trees defined on an identical set of taxa (de Quieroz 1993, Swofford 1991, Farris et al. 1995, Miyamoto & Fitch 1995, Huelsenbeck et al. 1996b). Interestingly, it is possible to compare the performance of more classical consensus methods to supertree methods in a consensus setting, where all datasets have identical taxa (Bininda-Emonds 2003). Since supertree methods are often developed from existing consensus methods, a setting that allow both types of method to be directly compared is desirable to quantify their relative accuracies. The supertree strategy seems to be increasingly used in phylogenomics, where large amounts of data can be subdivided to facilitate phylogenetic analyses (i.e., divide-and-conquer strategy: Bininda-Emonds 2004b, Wilkinson & Cotton 2006). Furthermore, supertree methods have been proposed as representing the optimal solution to reconstruct the Tree of Life (Sanderson et al. 1998, Bininda-Emonds et al. 2002, Bininda-Emonds 2004c, Wilkinson & Cotton 2006). Consensus and supertree methods are similar in design, and both can be referred to as separate analyses, by opposition to a combined analysis (*sensu* de Queiroz 1993) where sequence datasets are concatenated in a single supermatrix. A heated debate between those in favour and those opposed to the use of consensus has been raging for the last decade (Barrett et

al. 1991, Chippindale & Wiens 1994, de Queiroz et al. 1995, Huelsenbeck et al. 1996a, b, Wiens 1998c). The same debate has recently been extended to supermatrices and supertrees (e.g., Sanderson et al. 1998, Springer & de Jong 2001, Gatesy et al. 2002, Bininda-Emonds 2004a, Crandall & Buhay 2004, Gatesy et al. 2004, Bininda-Emonds 2005, Gadagkar et al. 2005, de Queiroz & Gatesy 2007, Nishihara et al. 2007).

In this study, four different approaches that are commonly used in phylogenomics to analyze DNA sequence matrices were studied in a consensus setting (*sensu* Bininda-Emonds 2003): (1) missing data matrices, (2) composite matrices, (3) consensus methods, and (4) supertree methods. Congruence among these competing approaches was compared in relation to model trees that were obtained from a complete matrix of mitogenomic mammalian sequences. The objectives of this study are twofold. First, to compare the congruence of trees inferred from matrices with missing data and from composite taxon matrices to a tree estimated from a complete dataset. Second, to compare the congruence of different consensus and supertree methods. To do so, complete mitogenomic sequences of 102 mammalian species (representing 93 families) were aligned to generate two model tree topologies. Then, different types of matrices were generated from the complete matrix: missing data, composite and individual gene matrices (that were subsequently combined with a consensus or a supertree method). Congruence was quantified by comparing the trees inferred using the different approaches to the model trees.

5.4. METHODS

The next sections present the methodology that was used to construct the model trees, and the simulation protocol employed to generate missing data and composite matrices from the complete matrix. Then, we present the different consensus and supertree methods that were applied to analyze the individual gene matrices. Finally, details are provided about the distance metrics that were used to measure topological differences between inferred trees and model trees.

5.4.1. Model tree

5.4.1.1. DNA sequence alignments

In February 2009, 96 mammal families had at least one species with a complete mitochondrial (mt) sequence in GenBank. When more than one taxon was available, only one species was selected for each family (with a few exceptions, see below) and the entire mt sequence was downloaded. The 12 mitochondrial genes of the H-strand were aligned using ClustalX 2.0.10 (Higgins & Sharp 1988) and the alignment was further verified by eye in SeAl 2.0a11. Ambiguous sites and overlapping regions of the ATP6-ATP8 and NAD4-NAD4L were removed. Stationarity of base frequencies across taxa was tested using the chi-square test of homogeneity of base frequencies implemented in PAUP* 4.0 (Swofford 1998). A test of congruence among distance matrices (CADM: Legendre & Lapointe 2002) was used to determine the congruence among the 12 mt genes in R 2.9.0 (Ihaka & Gentleman 1996, R Development Core Team 2009), within the Ape 2.3 package (Paradis et al. 2004, Paradis 2006), with 9999 permutations used for significance testing.

A well supported tree, compatible with the current molecular phylogeny of mammals (Springer et al. 2004b, Springer & Murphy 2007), was required to represent interfamilial mitogenomic relationships. However, systematic errors mainly caused by reconstruction artifacts, can produce a biased tree topology (Delsuc et al. 2005). Among potential systematic errors, heterogeneity in base composition (e.g., Penny & Hasegawa 1997, Waddell et al. 1999a, Schmitz et al. 2002, Reyes et al. 2004, Gibson et al. 2005, Montgelard et al. 2008) and different evolutionary rates among species (e.g., Philippe 1997, Delsuc et al. 2002, Lin et al. 2002a, b, Douzery et al. 2003, Huttley et al. 2007) have often been cited as confounding factors affecting the inference of mammalian mitogenomic relationships. Indeed, preliminary analyses of our dataset revealed the presence of systematic biases and different strategies were used to reduce their effect (as in Reyes et al. 2004, Arnason et al. 2008). For one, the third codon position was removed since it evolves more rapidly, especially in the mitochondrial genome (Kjer & Honeycutt 2007, Montgelard et al. 2008), and is often saturated for higher-level relationships (Gibson et al. 2005). Then, the first codon position of leucine (C and T) was recoded as pyrimidine (Y). Three problematic species were also removed (*Anomalurus sp.*, Anomaluridae, *Erinaceus europaeus*, Erinaceidae and *Manis tetradactyla*, Manidae). These species are known to be affected by either reduced or

accelerated evolutionary rates which can lead to long-branch attraction or placement uncertainty due to short branches (Waddell et al. 1999a, Douzery et al. 2003, Bergsten 2005, Horner et al. 2007, Kjer & Honeycutt 2007, Arnason et al. 2008). Finally, nine extra species were added to break long branches within the following six families: Chrysochloridae, Elephantidae, Macroscelidae, Procaviidae, Soricidae and Talpidae (see Arnason et al. 2008). Consequently, a total of 102 complete mt sequences, representing 93 mammalian families, were included (Appendix 5.1).

5.4.1.2. Phylogenetic inference

Modeltest 3.7 was used to identify the best model of nucleotide substitution (Posada & Crandall 1998). Both the hierarchical likelihood ratio tests (hLRTs) and Akaike information criterion (AIC) suggested a general time-reversible model (GTR: Lanave et al. 1984, Tavaré 1986, Rodriguez et al. 1990) following a gamma distribution (Γ: Yang 1993) and with invariant sites (I). The equilibrium frequencies of nucleotides A, C, G, and T were: gA = 0.3452, gC = 0.2054, gG = 0.0901, gT = 0.3593, the relative substitution rates were: rAC = 1.1083, rAG = 6.7749, rAT = 1.1934, rCG = 1.3020, rCT = 3.9717, rGT = 1.0000, and parameters α and I were 0.6762 and 0.4437 respectively. Phylogenetic trees were estimated using two different methods: maximum likelihood (ML: Felsenstein 1973, 1981) and Bayesian maximum likelihood (BML: Rannala & Yang 1996, Huelsenbeck et al. 2001). ML analysis was performed in PhyML 3.0 (Guindon & Gascuel 2003), with a GTR + Γ + I model, where base frequencies, proportion of invariable sites and gamma shape distribution parameter were estimated from the data. The number of categories for the gamma distribution was set to six. A subtree pruning and regrafting (SPR) algorithm was selected, starting from a BioNJ tree, and ten additionnal random starting trees. Non-parametric bootstrap support (BS) was assesed using identical settings in PhyML for 100 replicates. BML was performed with MRBAYES 3.1.2 (Huelsenbeck & Ronquist 2001) on a shared-memory multiprocessor computer (Altix 4700). Two MCMC analyses were run for 5 000 000 generations each, using the same GTR+Γ+I model. The Metropolis coupling used eight chains, starting from a random tree and eight swaps with Markov chains sampled every 100th generation, and with a burn-in of 10%. The majority-rule consensus tree and Bayesian posterior probabilities (BPP) were obtained from the tree distribution.

5.4.1.3. Model tree topology

The ML and BML tree topologies were identical, except at two nodes. These incongruent clades were represented by polytomies in order to render the ML and BML trees completely congruent (Fig. 5.1). This topology was used as the first model tree (MT1). A second model tree (MT2) was then constructed by collapsing all branches that were not supported by the current molecular phylogeny of mammals. In this second tree (Fig. 5.2), eight extra polytomies were added to the first model tree to ensure that all clades were compatible with recent molecular studies (as discussed in Appendix 5.2). MT2 provided a reference for assessing mammalian mitogenomic relationships and clades. Finally, all taxa were assigned to 49 strongly supported clades (Fig. 5.1 and 5.2) derived from support measures obtained in this and other studies. These clades were used to create monophyletic composite taxa (see below).

5.4.2. Simulations

5.4.2.1. Missing data

Missing data matrices were obtained from the complete matrix of 12 mt genes, following Campbell & Lapointe (In press: chapter 4). An increasing number of genes was deleted at random from the matrix (i.e. 5, 15, 30, 50 or 75% of the total number of genes in the ingroup taxa) and coded as "?" in the matrix. We ensured that all species had at least one gene in common to avoid undefined distances and to increase resolution when inferring the phylogenetic tree. Genes were not deleted from the two monotreme sequences, which represent the most recent relatives of placentals and marsupials and which were used as outgroup. One hundred replicates were simulated for each level of incompleteness.

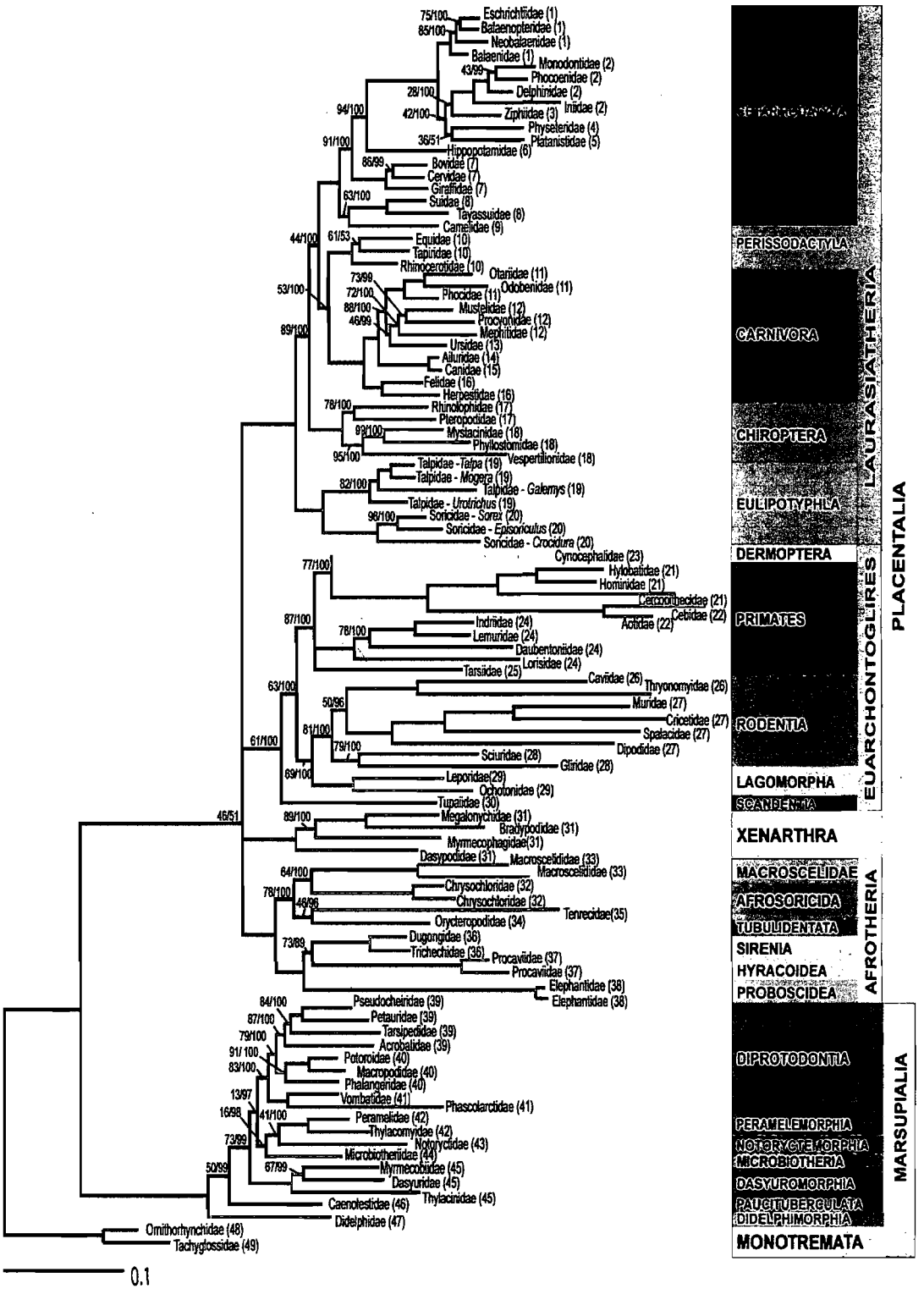

Figure 5.1. First model tree (MT1) representing mitogenomic relationships among 93 mammalian families.

Bootstrap values (BS) and Bayesian posterior probabilities (BPP) are indicated on branches (BS/BPP). Branches without values correspond to BS/BPP = 100/100.

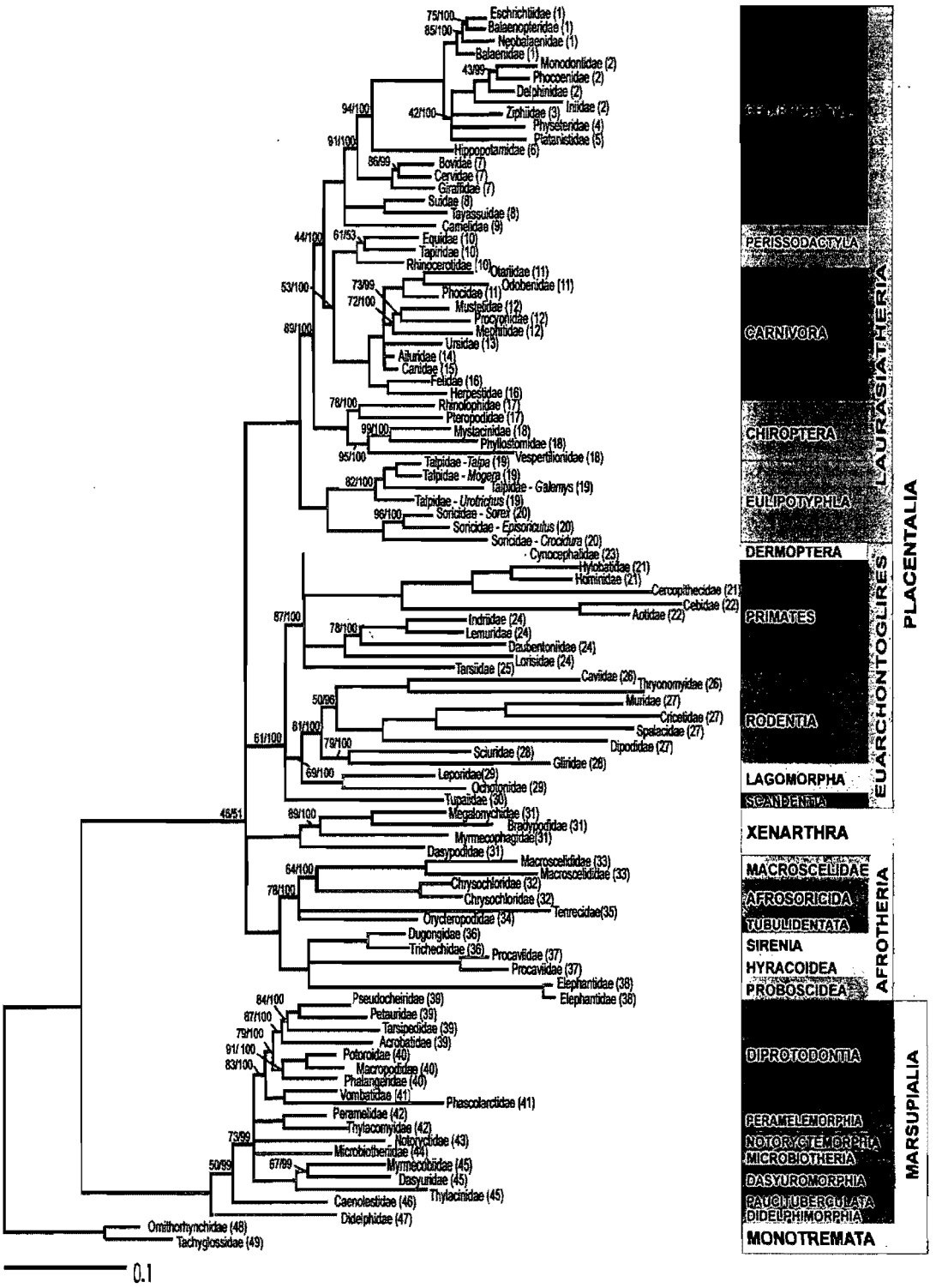

Figure 5.2. Second model tree (MT2) representing mitogenomic relationships among mammalian families.

Bootstrap values (BS) and Bayesian posterior probabilities (BPP) are indicated on branches (BS/BPP). Branches without values correspond to BS/BPP = 100/100.

5.4.2.2.Composite taxa

Each missing data matrix was used to generate a corresponding composite taxon matrix using the following criteria (adapted from Campbell & Lapointe In press: chapter 4):

1. When a taxon was not part of a clade (i.e., the 49 groups defined on Fig. 5.1) or was a single representative of its clade, no composite taxon was formed.
2. When all taxa within a clade had complete sequences, no composite taxon was formed.
3. When only one taxon within a clade had missing data, no composite taxon was formed in that clade and all taxa were used in the analyses, including the taxon with missing data.
4. When two taxa had missing data within a clade of size two, the sequence that contained fewer missing data was kept and the missing values were replaced with the sequence of the other taxa, when available.
5. When two or more taxa had missing data within a clade of size three, all possible composites of two or three species were formed (for three species A, B and C, the four possible combinations are: A/BC, AB/C, AC/B, and ABC). The optimal combination selected to generate the composite was the one that minimized the amount of missing data that remained in the matrix, once the composites were formed.
6. When two or more taxa had missing data within a clade of size four, all possible composites of two, three or four species were formed (for four species A, B, C and D, the fourteen possible combinations are: AB/CD, AC/BD, AD/BC, AB/C/D, AC/B/D, AD/B/C, BC/A/D, BD/A/C, CD/A/B, A/BCD, B/ACD, C/ABD, D/ABC, and ABCD). The optimal combination selected to generate the composite was the one that minimized the amount of missing data that remained in the matrix, once the composites were formed.

Composite matrix sizes varied according to the number of composite taxa created. The average number of taxa and proportion of missing data in the composite matrices are given in Table 5.1.

5.4.2.3. Phylogenetic inference

Missing data and composite taxon matrices were analyzed with PhyML 3.0 installed on ten Power Mac G5, with PowerPC 970MP processors (2 x 2.5 GHz). Settings were similar to those used to analyze the complete matrix, except that the nearest neighbor interchange (NNI) algorithm was used instead of the SPR algorithm and that a single BioNJ tree was used as a starting tree.

5.4.3. Consensus and supertree methods

5.4.3.1. Individual datasets

For the consensus and supertree methods, the 12 individual mt genes were analyzed separately. Stationarity of base frequencies across taxa was tested using the chi-square test of homogeneity of base frequencies implemented in PAUP* 4.0, with a Bonferroni correction for multiple tests (Rice 1989). Modeltest 3.7 was used to identify the best substitution model for each dataset. ML analyses were then performed on each dataset with PhyML 3.0, using the model suggested by the AIC criterion. Analytical parameters were identical to those described for the complete matrix analysis, except for the evolutionary model.

Given that all 12 datasets included an identical number of taxa (n = 102), the comparison of consensus and supertree methods was performed in a consensus setting (Bininda-Emonds 2003). Therefore, even though we will maintain the use of "supertree" for methods that have been developed in a supertree context, all methods can be considered as consensus methods and can be divided into three categories: (1) consensus techniques based on topological relationships (topological consensus methods), (2) supertree techniques based on topological relationships (topological supertree methods), and (3) supertree techniques that take into account branch lengths (branch length supertree methods).

5.4.3.2. Topological consensus methods

Four consensus methods were applied to combine the 12 independent gene trees in PAUP* 4.0: (1) strict, (2) majority rule (MR), (3) majority rule with compatible groupings (MRC), and (4) Adams consensus. The strict consensus only retains groups that are identical among all input trees (Sokal & Rohlf 1981, Page 1989). The majority rule consensus (MR) contains groups that are present in more than 50% of input trees

(Margush & McMorris 1981, Swofford 1991), such that groups found in seven or more trees were kept. The second type of majority rule consensus (MRC) retains all compatible groupings below 50% of occurrence in addition to those above 50%. The Adams consensus presents groups that are nested within another without necessarily including identical taxa (Adams 1972, 1986). Therefore, Adams consensus does not only propose monophyletic groups. A more complete description of these methods can be found in Swofford (1991).

5.4.3.3. Topological supertree methods

Three different optimality criteria were used to construct supertrees (consensus) from the 12 independent gene trees in CLANN 3.0.2 (Creevey & McInerney 2005): (1) matrix representation with parsimony (MRP), (2) most similar supertree (MSS), and (3) maximum splits fit (SFIT). In MRP, nodes present in each tree are coded into a binary matrix using the Baum and Ragan method (Baum 1992, Ragan 1992a, b). The binary matrix is then analyzed with a parsimony algorithm (Edwards & Cavalli-Sforza 1963) using ten TBR searches and a random starting tree. The MSS method calculates the symmetric difference between each gene tree and the supertree and sums these differences to obtain a supertree score (Creevey et al. 2004). The optimal supertree is the one with the best score (smallest distance). For the SFIT method, the splits present in each gene tree and a candidate supertree are recorded and the supertree with the maximum split fit (sharing the greatest number of splits) is selected as optimal (Creevey & McInerney 2005). For MSS and SFIT, a SPR heuristic search using ten repetitions each starting with a different NJ tree was selected to search among all possible supertree topologies (default parameters in CLANN). For all of these methods, a strict consensus was used to combine equally optimal supertrees, if any.

5.4.3.4. Branch-length supertree methods

Three other optimality criteria, which take into account the branch lengths of the input trees, were also employed: (4) average consensus (AC), (5) unweighted super distance matrix (SDM), and (6) weighted super distance matrix (SDMw). These methods are implemented in the *SDM* program (Criscuolo et al. 2006). The AC criterion optimizes the sum-of-squared distances between each source tree and the consensus tree, by averaging the path-length distance matrices computed from each gene tree and then applying a least-squares algorithm to this average matrix (Lapointe & Cucumel 1997).

SDM applies the same criterion, except that path-length distance matrices are first transformed so as to minimize the sum-of-squared distances among them (Criscuolo et al. 2006). The weighted version (SDMw) assigns a weight to each tree prior to computing the average matrix, based on the sequence length of the corresponding gene. All supertrees (consensus) were estimated using an unweighted least squares algorithm (Cavalli-Sforza & Edwards 1967) in PHYLIP 3.68 (Felsenstein 1989) with the FITCH program, using the jumble option (J = 10), which randomize the input order of species for each run and with global rearrangements allowed (SPR algorithm).

5.4.4. Distance metrics

Two dissimilarity measures were computed in PAUP* 4.0 to quantify the congruence between model trees (MT1 and MT2), trees inferred from missing data and composite matrices, and consensus trees. The symmetric-difference or partition metric (PM) counts the number of different splits in the trees being compared (Robinson & Foulds 1979, 1981). In order, to compare model trees (102 taxa) with composite trees of various sizes (49 to 94 taxa), PM was normalized by dividing each value by the maximal possible value (2n - 6). The agreement subtrees (D_1: Gordon 1980, Finden & Gordon 1985, Goddard et al. 1994) calculates the number of taxa that need to be pruned from the trees to obtain a congruent topology. Here again, normalized D_1 are obtained by dividing each value by the maximum possible value (n - 3). Rohlf's (1982) consensus information index (CI_I) was also calculated on the consensus trees to measure their relative resolution (the index ranges from 0 when the consensus is a bush to 1 when the tree is fully resolved).

5.5. RESULTS

5.5.1. Missing data and composite matrices

The construction of composite matrices (from the corresponding missing data matrices) always reduced the number of missing entries in the matrix (Table 5.1). Moreover, the number of taxa included in the composite matrices decreased as missing data increased. To compare missing data to composite matrices, five different levels of incompleteness (I) were considered (i.e., 5, 15, 30, 50 and 75%). The phylogenetic congruence of the competing approaches were highly similar, ranging from 0.300 (I = 5%) to 0.729 (I = 75%) for missing data matrices and from 0.291 (I = 5%) to 0.800

(I = 75%) for composite matrices (Table 5.2). For both approaches similar trends were observed for the two metrics (PM and D_1).

Table 5.1. Properties of composite matrices created from missing data matrices with different levels of incompleteness (I).

Proportion of missing data in composite matrices (expressed as % of character missing), average and range of matrix sizes (number of taxa) are calculated from 100 replicates.

I (%)	Missing data (%)	Average matrix size (nb of taxa)	Range (nb of taxa)
5	2.72	88.25	83 - 94
15	5.97	63.20	58 - 69
30	13.42	51.28	49 - 55
50	28.72	49.05	49 - 50
75	58.05	49.03	49 - 50

Table 5.2. Congruence of phylogenetic trees inferred from missing data and composite matrices of different levels of incompleteness (I).

Missing data matrices included 102 taxa while composite matrices were of varying sizes (see Table 5.1). Normalized dissimilarity values (PM and D_1) were calculated from 100 replicates. MT1: first model tree, MT2: second model tree, PM: partition metric, D_1: agreement subtrees.

Model tree	I (%)	Missing data		Composites	
		PM	D_1	PM	D_1
MT1	5	0.327	0.436	0.323	0.424
	15	0.425	0.508	0.436	0.473
	30	0.589	0.593	0.573	0.527
	50	0.729	0.686	0.724	0.594
	75	0.724	0.688	0.820	0.666
MT2	5	0.300	0.478	0.291	0.470
	15	0.394	0.543	0.383	0.530
	30	0.548	0.624	0.500	0.592
	50	0.681	0.706	0.636	0.650
	75	0.675	0.708	0.716	0.712

5.5.2. Individual datasets

The length (L) of each of the12 aligned mitochondrial genes varied from 90 to 1803bp, when all three codon positions were included. The homogeneity test of base frequencies indicated that seven out of the 12 datasets were heterogeneous (Table 5.3). However, when only the first two codon positions were considered, all datasets were homogeneous. Consequently, all subsequent analyses were performed using alignments with only the first and second codon positions. The two optimality criteria (hLRTs and AIC) implemented in Modeltest suggested different models for some datasets. Indeed, whereas the hLRTs criterion proposed a GTR model for all datasets, AIC suggested varying models depending on the dataset, as listed in Table 5.3. The congruence among distance matrix test (CADM) suggested that all 12 datasets were congruent (Friedman's χ^2 = 44341.5, Kendall's W= 0.7175, p = 0.0001).

5.5.3. Consensus and supertree methods

Important differences were observed between PM and D_1 and among topological consensus methods (Table 5.4). These results may be explained by the fact that some consensus methods were poorly resolved (CI_I = 0.02, 0.10 and 0.18). If we compare a fully resolved tree to a bush, PM will take a value of 0.5 because only the clades in the fully resolved tree are contributing to the distance. On the other hand, D_1, which calculates the number of taxa that have to be pruned from both trees to obtain identical topologies, will exhibit a very small value (n - 2 taxa need to be deleted for both topologies to be compatible). Because the majority rule consensus that included all compatible groupings (MRC) is more resolved than other classical consensus methods (CI_I = 0.94), it provided the best results and was the closest to model tree topologies (PM = 0.22 and 0.23). The majority rule consensus (MR) was the second most congruent consensus method (PM = 0.30 and 0.25), although much less resolved (CI_I = 0.10), and thus D_1 was considerably increased (0.79 and 0.73, compared to 0.39 and 0.46 for MRC).

The topological supertree techniques suggested more than one optimal supertree: 184 (SFIT), five (MRP), and two (MSS) optimal supertrees. These supertrees were first combined using a strict consensus supertree, and thus, were not fully resolved (CI_I = 0.53 to 0.98). The least congruent method was MSS (PM = 0.54 and 0.56; D_1 =0.62 and 0.60). MRP and SFIT performed well (PM = 0.23 and 0.24).

Table 5.3. Statistical description of the 12 genes on the mitochondrial H-strand and concatenated dataset (ALL).

L (bp): length of the gene in base pairs. No cst: number of constant sites in the alignment. No info: number of informative sites in the alignment. AIC: Model selected according to AIC criterion in Modeltest, which always included parameters G (a gamma distribution of substitution rates) and I (a proportion of invariable sites). $\chi 2$ (1, 2): chi square test for homogeneity of base frequencies across species on datasets with third codon position removed (P = 1.0 in every case). $\chi 2$ (1, 2, 3): chi square test on datasets with codon positions 1, 2 and 3 included. * Identifies significant values after a Bonferroni correction, P < 0.004 (0.05/12).

Datasets	L (bp)	No cst (%)	No info (%)	AIC	χ^2 (1, 2)	χ^2 (1, 2, 3)
ATP6	452	200 (44.2)	204 (45.1)	GTR[1]	78.25	419.99 *
ATP8	60	10 (16.7)	47 (78.3)	TrN[2]	145.87	198.83
COX1	1022	780 (76.3)	157 (15.4)	GTR	15.85	573.75 *
COX2	440	247 (56.1)	150 (34.1)	TVM[3]	28.26	290.95
COX3	522	339 (64.9)	137 (26.2)	TVM	33.42	320.19
CYTB	754	402 (53.3)	276 (36.6)	TIM[4]	80.04	525.82 *
NAD1	618	312 (50.5)	239 (38.7)	GTR	69.18	484.12 *
NAD2	690	173 (25.1)	463 (67.1)	GTR	129.20	623.00 *
NAD3	230	109 (47.4)	100 (43.5)	TrN	78.92	251.49
NAD4	918	380 (41.4)	458 (49.9)	GTR	96.29	681.88 *
NAD4L	192	69 (35.9)	105 (54.7)	TVM	65.08	251.49
NAD5	1202	463 (38.5)	619 (51.5)	GTR	136.64	919.59 *
ALL	7100	3484 (49.1)	2955 (41.6)	GTR	267.78	4107.63 *

[1] GTR: General time reversible model (Tavaré 1986)
[2] TrN: Tamura-Nei model (Tamura & Nei 1993)
[3] TVM: Tranversional model (Posada & Crandall 1998)
[4] TIM: Transitional model (Posada & Crandal 1998)

Supertree methods that take branch lengths into account performed relatively well (PM from 0.22 to 0.30, and D_1 ranging from 0.43 to 0.50) and proposed one optimal, fully resolved supertree. AC was slightly more accurate than both SDM and SDMw. The weighted version of SDM (i.e., SDMw) did not improve phylogenetic performance (PM increased slightly and D_1 remained the same).

Table 5.4. Congruence of phylogenetic trees inferred from consensus and supertree methods (that ignore or consider branch lengths).

MT1: first model tree, MT2: second model tree, CI_I: Rohlf's consensus information index, PM: partition metric, D1: agreement subtrees, MR: majority rule consensus, MRC: majority rule consensus with compatible groupings, MRP: matrix representation with parsimony, MSS: most similar supertree, SFIT: maximum splits fit, AC: average consensus, SDM: unweighted super distance matrix, SDMw: weighted super distance matrix.

		CI_I	MT1		MT2	
			PM	D_1	PM	D_1
Topological consensus methods	Strict	0.02	0.46	0.94	0.39	0.94
	MR	0.10	0.30	0.79	0.25	0.73
	MRC	0.94	0.22	0.39	0.23	0.46
	Adams	0.18	0.42	0.78	0.36	0.75
Topological supertree methods	MRP[1]	0.98	0.23	0.47	0.23	0.43
	MSS[2]	0.91	0.54	0.62	0.56	0.60
	SFIT[3]	0.53	0.24	0.51	0.22	0.50
Branch-length supertree methods	AC	1.00	0.25	0.43	0.22	0.49
	SDM	1.00	0.27	0.45	0.24	0.50
	SDMw	1.00	0.30	0.45	0.27	0.50

[1] Strict consensus of five most parsimonious trees.

[2] Strict consensus of two equally optimal supertrees.

[3] Strict consensus of 184 equally optimal supertrees.

5.6. DISCUSSION

With increased sizes of datasets used in phylogenomics, the number of missing entries in character matrices also increased (Sanderson & Driskell 2003, Driskell et al. 2004). However, matrices characterized by a relatively low number of missing data do not seem to exhibit a reduced phylogenetic accuracy. Indeed, Prasad et al. (2008) have observed that up to 25% of missing data did not affect phylogenetic inference, other than decreasing the bootstrap support of a few branches. Although, phylogenomic matrices with more than 50% of missing characters are fairly common (Gatesy et al. 2002, Kearney 2002, Driskell et al. 2004, Philippe et al. 2005b, Hartmann & Vision 2008), recent computer simulations have shown that perfect accuracy can be obtained even in such cases (e.g., Wiens 2003b, Philippe et al. 2004). They concluded that the number of informative characters was more important than the number of missing entries. Similarly, Campbell & Lapointe (In press: chapter 4) noticed an increase in accuracy when longer sequences were analyzed for identical proportion of missing data. Numerous studies have obtained strong bootstrap support when analyzing highly incomplete supermatrices (e.g., Driskell et al. 2004, Philippe et al. 2004, Wiens et al. 2005). Removing incomplete taxa from an analysis to avoid missing taxa may negatively affect phylogenetic accuracy. Indeed, it has been shown that even an incomplete taxon can represent a key taxon in breaking long-branches (Wiens 2005). Also, deleting some taxa may remove potential informative sites, as for example, when removing incomplete fossil taxa (Doyle & Donoghue 1987, Gauthier et al. 1988). However, other simulation studies observed a negative impact of missing data on phylogenetic inference (Huelsenbeck 1991, Wiens & Reeder 1995, Bininda-Emonds & Sanderson 2001, Kearney 2002, Flynn et al. 2005). Phillipe et al. (2005a) mentionned that some taxa may be more affected by a large amount of missing data. Campbell & Lapointe (In press: chapter 4) have observed that different factors (e.g., branch lengths and inference methods) could alter the impact of missing data on phylogenetic inference. Given different opinions on the effect of missing data on phylogenetic accuracy, a possible option is to avoid or reduce the number of missing entries in character matrices.

When multiple incomplete sequences are available within a monophyletic group of species, it is possible to reduce the amount of missing data by creating composite taxa to infer higher-level phylogenies. This strategy is commonly employed in phylogenomic

analyses (e.g., Delsuc et al. 2006, Phillips et al. 2006, Seiffert 2007, Beck 2008, Bourlat et al. 2008, Duvall et al. 2008). Malia et al. (2003) compared the tree inferred from a supermatrix that either contained composite taxa or that did not. They concluded that the formation of composite taxa could impede accurate phylogenetic inference and strongly argued against the use of composite taxa. In contrast, Campbell & Lapointe (In press: chapter 4) reached a different conclusion when comparing the two competing approaches in a simulation study. Instead, missing data matrices and composite matrices had similar phylogenetic accuracy when tested over a wide range of conditions. These contrasting conclusions could be explained by the simulation design used by Campbell & Lapointe (In press: chapter 4) that may not represent typical sampling scheme (i.e., they used ten monophyletic groups of 4 taxa each) or because models of evolution may not entirely capture the complexity of real DNA sequences. Alternatively, Malia et al.'s (2003) study design has been criticized since it included species that did not share any sequences and also because they did not dissociate all composite taxa (see Springer et al. 2004a). Furthermore, they based their conclusion on only one dataset, which included non-monophyletic composites.

In the present study, sampling was done in a similar fashion to other phylogenomic studies. That is, taxa were selected solely on the basis of sequence availability. Thus, one representative taxon from each family of mammals was included (with few exceptions, see methods). Also, both model trees and all simulated datasets were obtained from the complete mitogenomic sequences, thus removing potential biases included by purely theoretical simulation design. The phylogeny inferred from the complete dataset was congruent with recent molecular studies of interordinal relationships (Springer et al. 2004b, Springer & Murphy 2007), but also with various interfamilial studies of mammals (see Appendix 5.2). Therefore, we are confident that composite taxa were created within a monophyletic group, which is an important condition to satisfy when combining sequences from different species (Scally et al. 2001, Springer et al. 2004a, but see Campbell & Lapointe In press: chapter 4). In agreement with Campbell & Lapointe (In press: chapter 4), the results from this study provide further evidence that support the validity of composite taxa. For all levels of incompleteness tested (i.e, from 5 to 75% missing data), missing data and composite matrices were equally congruent to model trees.

Another increasingly popular method to build large supermatrices is to assemble expressed sequence tags (ESTs: Rudd 2003). These often represent a collection of partial gene sequences and the distribution of missing data differs from typical DNA sequence matrices. Hartmann & Vision (2008) have shown through simulations that phylogenetic accuracy was reduced when analyzing incomplete EST matrices that were characterized by 14 to 60% of missing data. Among different phylogenetic methods tested, ML was least affected by missing data. Also, phylogenetic accuracy increased when missing data were distributed randomly in the matrix, rather than in a typical EST fashion. However, the increase in accuracy observed with randomly removed entries was not observed when maximum parsimony was used, and both random and EST missing data distributions performed poorly in that case. Our results are in agreement with Hartmann & Vision (2008) since we observed an increase in the distance metrics between the inferred and model trees when the proportion of missing data increased. In the present study, only ML was used to infer phylogenetic trees, since similar trends were obtained with other phylogenetic methods in previous simulations (i.e., neighbor-joining and Bayesian analysis on selected datatsets: Campbell & Lapointe In press: chapter 4). Jeffroy et al. (2006) have also suggested that accuracy can be more affected by the proportion of phylogenetic signal to noise than by the choice of the phylogenetic method. As in other DNA supermatrices, composite taxa are commonly used in EST studies (Roeding et al. 2007). However, the performance of the composite approach for EST matrices remains to be addressed since our study only compared the missing data and composite taxon approaches with randomly distributed missing data.

Yet, another alternative method to avoid missing data in large-scale studies is the construction of supertrees from individual source trees (Sanderson et al. 1998). Supertree methods combine trees that have overlapping taxon sets, whereas consensus methods summarize trees with identical taxon set. Both approaches have been extensively studied (e.g., de Queiroz 1993, de Queiroz et al. 1995, Bininda-Emonds & Sanderson 2001, Bininda-Emonds 2003, Wilkinson et al. 2005, Criscuolo et al. 2006) and can be compared when identical taxa are used (e.g., Lapointe 1998b, Lapointe et al. 1999, Levasseur & Lapointe 2001, 2006, Bininda-Emonds 2003, Wilkinson et al. 2005, 2007). Among the consensus methods, the majority rule with compatible groupings (MRC) was the most congruent to model trees, when compared to consensus and supertrees. Criticisms of consensus emphasized the poor resolution of consensus trees (e.g., Barrett et al. 1991, Kluge & Wolf 1993, de Queiroz et al.

1995). However, MRC was well resolved (CI_I = 0.938), which may explain its performance relative to other topological consensus methods. Through simulations, Bininda-Emonds (2003) has also observed that MRC provided the highest accuracy amongst consensus methods.

In general, most supertree methods provided similar congruence to model trees (except for MSS, see below). This result was surprising, given that numerous studies have proposed that accounting for branch lengths should give more accurate supertrees (Lapointe et al. 1999, Levasseur & Lapointe 2001, 2006). However, we confirmed that supertree techniques based on topological relationships are unlikely to offer a fully resolved consensus tree. On the other hand, Criscuolo et al. (2006) have shown that MRP and SDM were equally accurate at low levels of missing data (i.e., 25% of deleted taxa), and that the benefit of accounting for branch lengths was only revealed at higher levels of missing data (e.g. 75% of deleted taxa).

MSS was the least accurate of all supertree methods. Creevey & McInerney (2005) have compared their MSS approach to the AC technique, but without branch lengths (i.e., with all branch lengths equal to one). The better result obtained with AC (and SDM) with respect to MSS, might suggest that branch lengths contain information different from topological relationships, when supertree methods are used. A similar result was also observed by Criscuolo et al. (2006). As for the other supertree methods (SFIT and MRP), they were similar to techniques that use branch lengths when measured from their congruence to model trees. This result is consistent with studies that have shown that MRP is accurate under certain conditions (e.g., Bininda-Emonds & Sanderson 2001, Criscuolo et al. 2006, Fitzpatrick et al. 2006, Higdon et al. 2007). Through simulations, Bininda-Emonds & Sanderson (2001) have observed that MRP provided accuracy values comparable to those obtained from a supermatrix analysis (and that accuracy was slightly increased when a weighted MRP was used). Among the supertree methods with branch lengths, SDM outperformed slightly SDMw. The distance matrices are weighted according to sequence lengths in SDMw, with trees inferred from longer sequences contributing more to the "super" distance values. Thus, biases will be amplified if they are associated with longer sequence datasets. This might explain why SDMw might not always provide the optimal solution. Also, AC was slightly more accurate than SDM, in contrast with Criscuolo et al. (2006), who showed the opposite under all conditions tested. Fitzpatrick et al. (2006), in a fungal study

comparing AC, MRP and the supermatrix approaches, reported that AC might be prone to long-branch attraction, but this was not the case here. Supertree methods that account for branch lengths may thus provide additional information, which could help resolve some least-resolved clades.

The results from this study demonstrate that the composite approach is a powerful approach to analyse missing data matrices. As stated by Campbell & Lapointe (In press: chapter 4), including composite taxa greatly reduced the matrix size (i.e., the number of taxa), which drastically decreased the length of the ML analyses, while providing good phylogenetic estimates. This represents a main advantage of the composite taxon approach in a phylogenomic context, where strategies to minimize computing time must be developed. In the second part of this study we have compared different consensus and supertree approaches relative to a supermatrix analysis. Most of the supertree methods tested were highly congruent with both model trees. Interestingly, the majority rule consensus with compatible clades was also highly congruent, which suggest that it represents is an accurate and fast approach to summarize information obtained in separate analyses.

5.7. ACKNOWLEDGMENTS

For the computational resources, we would like to thank Marie-Hélène Duplain for granting access to the Laboratoire Interfacultaires de Micro-Informatique de l'Université de Montréal, outside business hours. We would also like to thank Antoine Lapointe and Daniel Stubbs for their programming skills and the RQCHP (Réseau Québécois de Calcul de Haute Performance) for granting access to its HPC facilities for the Bayesian analyses. Also, we greatly appreciated the help of Stéphane Guindon for running some of the ML analyses. This study was supported by a FESP (Faculté des Études Supérieures de l'Université de Montréal) scholarship to VC and by NSERC grant OGP0155251 to FJL.

APPENDIX 5.1. GenBank accession numbers of complete mitochondrial DNA sequences from 102 species representing 93 mammalian families.

Family and species taxonomy based on Wilson & Reeder (2005).

Order	Family	Species	Complete
MONOTREMATA	Tachyglossidae	*Tachyglossus aculeatus*	NC_003321
	Ornithorhynchidae	*Ornithorhynchus anatinus*	NC_000891
DIDELPHIMORPHIA	Didelphidae	*Didelphis virginiana*	NC_001610
PAUCITUBERCULATA	Caenolestidae	*Caenolestes fuliginosus*	NC_005828
MICROBIOTHERIA	Microbiotheriidae	*Dromiciops gliroides*	NC_005826
DASYUROMORPHIA	Thylacinidae	*Thylacinus cynocephalus*	NC_011944
	Myrmecobiidae	*Myrmecobius fasciatus*	NC_011949
	Dasyuridae	*Phascogale tapoatafa*	NC_006523
PERAMELEMORPHIA	Thylacomyidae	*Macrotis lagotis*	NC_006520
	Peramelidae	*Isoodon macrourus*	NC_002746
NOTORYCTEMORPHIA	Notoryctidae	*Notoryctes typhlops*	NC_006522
DIPROTODONTIA	Phascolarctidae	*Phascolarctos cinereus*	NC_008133
	Vombatidae	*Vombatus ursinus*	NC_003322
	Phalangeridae	*Trichosurus vulpecula*	NC_003039
	Potoroidae	*Potorous tridactylus*	NC_006524
	Macropodidae	*Macropus robustus*	NC_001794
	Pseudocheiridae	*Pseudocheirus peregrinus*	NC_006519
	Petauridae	*Petaurus breviceps*	NC_008135
	Tarsipedidae	*Tarsipes rostratus*	NC_006518
	Acrobatidae	*Distoechurus pennatus*	NC_008145
XENARTHRA	Dasypodidae	*Dasypus novemcinctus*	NC_001821
	Bradypodidae	*Bradypus tridactylus*	NC_006923
	Megalonychidae	*Choloepus didactylus*	NC_006924
	Myrmecophagidae	*Tamandua tetradactyla*	NC_004032
PROBOSCIDEA	Elephantidae	*Elephas maximus*	NC_005129
		Loxodonta africana	NC_000934
SIRENIA	Dugongidae	*Dugong dugon*	NC_003314
	Trichechidae	*Trichechus manatux*	NC_010302
HYRACOIDEA	Procaviidae	*Procavia capensis*	NC_004919
		Dendrohyrax dorsalis	NC_010301
TUBULIDENTATA	Orycteropodidae	*Orycteropus afer*	NC_002078
MACROSCELIDEA	Macroscelididae	*Macroscelides proboscideus*	NC_004026
		Elephantulus sp.	NC_004921
AFROSORICIDA	Tenrecidae	*Echinops telfairi*	NC_002631
	Chrysochloridae	*Chrysochloris asiatica*	NC_004920
		Eremitalpa granti	NC_010304

CETACARTIODACTYLA	Balaenidae	*Balaena mysticetus*	NC_005268
	Balaenopteridae	*Megaptera novaeangliae*	NC_006927
	Eschrichtiidae	*Eschrichtius robustus*	NC_005270
	Neobalaenidae	*Caperea marginata*	NC-005269
	Delphinidae	*Lagenorhynchus albirostris*	NC_005278
	Monodontidae	*Monodon monoceros*	NC_005279
	Phocoenidae	*Phocoena phocoena*	NC_005280
	Physeteridae	*Physeter catodon*	NC_002503
	Iniidae	*Inia geoffrensis*	NC_005276
	Platanistidae	*Platanista minor*	NC_005275
	Ziphiidae	*Berardius bairdii*	NC_005274
	Suidae	*Sus scrofa*	NC_000845
	Tayassuidae	*Pecari tajacu*	NC_012103
	Hippopotamidae	*Hippopotamus amphibius*	NC_000889
	Camelidae	*Lama pacos*	NC_002504
	Giraffidae	*Giraffa camelopardalis*	NC_012100
	Cervidae	*Cervus elaphus*	NC_007704
	Bovidae	*Bos taurus*	NC_001567
PERISSODACTYLA	Equidae	*Equus caballus*	NC_001640
	Tapiridae	*Tapirus terrestris*	NC_005130
	Rhinocerotidae	*Ceratotherium simum*	NC_001808
CARNIVORA	Ailuridae	*Ailurus fulgens*	NC_011124
	Ursidae	*Ursus americanus*	NC_003426
	Canidae	*Vulpes vulpes*	NC_008434
	Felidae	*Felis catus*	NC_001700
	Herpestidae	*Herpestes javanicus*	NC_006835
	Mustelidae	*Gulo gulo*	NC_009685
	Otariidae	*Eumetopias jubatus*	NC_001050
	Odobenidae	*Odobenus rosmarus*	NC_004029
	Phocidae	*Phoca vitulina*	NC_001325
	Procyonidae	*Procyon lotor*	NC_009126
	Mephitidae	*Mephitis mephitis*	1
EULIPOTYPHLA	Soricidae	*Crocidura russula*	NC_006893
		Sorex unguiculatus	NC_005435
		Episoriculus fumidus	NC_003040
	Talpidae	*Talpa europaea*	NC_002391
		Galemys pyrenaicus	NC_008156
		Mogera wogura	NC_005035
		Urotrichus talpoides	NC_005034
CHIROPTERA	Pteropodidae	*Pteropus dasymallus*	NC_002612
	Vespertilionidae	*Chalinolobus tuberculatus*	NC_002626
	Mystacinidae	*Mystacina tuberculata*	NC_006925
	Rhinolophidae	*Rhinolophus monoceros*	NC_005433

	Phyllostomidae	*Artibeus jamaicensis*	NC_002009
RODENTIA	Thryonomyidae	*Thryonomys swinderianus*	NC_002658
	Caviidae	*Cavia porcellus*	NC_000884
	Gliridae	*Myoxus glis*[5]	NC_001892
	Sciuridae	*Sciurus vulgaris*	NC_002369
	Dipodidae	*Jaculus jaculus*	NC_005314
	Spalacidae	*Nannospalax ehrenbergi*[6]	NC_005315
	Cricetidae	*Cricetulus griseus*	NC_007936
	Muridae	*Mus musculus*	NC_005089
LAGOMORPHA	Ochotonidae	*Ochotona princeps*	NC_005358
	Leporidae	*Oryctolagus cuniculus*	NC_001913
PRIMATES	Lemuridae	*Lemur catta*	NC_004025
	Indriidae	*Propithecus coquereli*	NC_011053
	Daubentoniidae	*Daubentonia madagascariensis*	NC_010299
	Lorisidae	*Nycticebus coucang*	NC_002765
	Tarsiidae	*Tarsius bancanus*	NC_002811
	Cebidae	*Cebus albifrons*	NC_002763
	Aotidae	*Aotus trivirgatus*	AY250707
	Cercopithecidae	*Macaca mulatta*	NC_005943
	Hylobatidae	*Hylobates lar*	NC_002082
	Hominidae	*Pan troglodytes*	NC_001643
DERMOPTERA	Cynocephalidae	*Cynocephalus variegatus*	NC_004031
SCANDENTIA	Tupaiidae	*Tupaia belangeri*	NC_002521

[1]Numbers available in Delisle & Strobeck (2005)

APPENDIX 5.2. Description and discussion of mammalian clades and model tree topologies.

MAMMALIA

Within the mammalian class, three infraclasses are defined: Prototheria (or Monotremata), Metatheria (or Marsupialia) and Eutheria (or Placentalia). Recent molecular (Phillips & Penny 2003, van Rheede et al. 2006, Bininda-Edmonds et al. 2007, Prasad et al. 2008) and morphological studies (Luo & Wible 2005, Ji et al. 2006, Rowe et al. 2008) support the so-called Theria hypothesis, where monotremes are the sister clade to therian mammals (i.e., marsupials and placentals).

METATHERIA

Extant marsupials are divided into seven orders and 21 families (Wilson & Reeder 2005) that are classified in two cohorts: Australidelphia and Ameridelphia (Szalay 1982). Australidelphia comprises five orders: Dasyuromorphia, Diprotodontia, Noryctemorphia, Peramelemorphia and Microbiotheria; whereas Ameridelphia consists of the remaining two orders: Didelphimorphia and Paucituberculata. Australidelphia represents a monophyletic group corroborated by numerous molecular and morphological analyses (e.g., Kirsch et al. 1991, 1997, Phillips et al. 2001, 2006, Horovitz & Sánchez-Villagra 2003, Beck 2008, Meredith et al. 2008a, 2009a, b). However, many studies support a paraphyletic Ameridelphia, with Didelphimorphia branching first, and Paucituberculata as a sister group to Australidelphia (e.g., this study, Asher et al. 2004, Beck 2008, Meredith et al. 2008a).

Australidelphia

Australidelphia is composed of four Australasian orders as well as the only representative of the Microbiotheria order, the South American Monito del monte. The evolutionary relationships among these five orders are contentious (Phillips et al. 2006). For example, recent studies support a sister relationship between Microbiotheria and the four Australasian orders (Phillips et al. 2006, Beck et al. 2008, Meredith et al. 2008a), whereas others have suggested that Microbiotheria is nested within Australidelphia (Asher et al. 2004, Nilsson et al. 2004). Also, the placement of Noryctemorphia is controversial and the branching order among Australidelphia differs among studies (e.g., Meredith et al. 2006, 2008a, 2009a, b,

Phillips et al. 2008, Beck et al. 2008). Even though all relationships within Australidelphia were strongly supported with the BML analysis (BPP ≥ 97%), ML analysis provided lower support values, and three nodes were collapsed (BS values from 13 to 41%) in the consensus model tree. This topology implies no resolution among the five australidelphian orders, and thus renders our model tree compatible with different phylogenetic arrangements. Within Dasyuromorphia, the relationships between the three families are congruent with a recent marsupial study, the first to include the mitogenomic sequence of the marsupial wolf (Thylacinidae: Miller et al. 2009). The monophyly of Diprotodontia obtained in this study is congruent with morphological and molecular studies (e.g., Kirsch et al. 1997, Horovitz & Sánchez-Villagra 2003, Asher et al. 2004, Phillips et al. 2006, Beck 2008, Meredith et al. 2008a, 2009a, b). Nine diprotodontian families were included in our model tree, which were sampled within four clades: (1) suborder Macropodiformes (Potoroidea and Macropodidae), (2) suborder Vombatiformes (Phascolarctidae and Vombatidae), (3) superfamily Petauroidea (Acrobatidae, Petauridae, Pseudocheiridae, Tarsipedidae), and (4) superfamily Phalangeroidea (Phalangeridae). The two superfamilies (Petauroidea and Phalangeroidea) are included in the larger suborder Phalangeriformes. Among and within those four clades, similar phylogenetic relationships are recovered in a large number of diprotodontian studies, as well as in our model tree topology, with Phalangeriformes representing a paraphyletic group (e.g., Phillips & Pratt 2008, Meredith et al. 2009a, b). Because the monophyly of Macropodiformes and Phalangeroidea was recovered with high support (BS=91%, BPP=100%) in our study, and because this clade is congruent with recent molecular topologies (Phillips & Pratt 2008, Meredith et al. 2009a, b), it was numbered as a single clade in our model tree.

EUTHERIA

Within Eutheria, many clades have received strong support from molecular studies above the ordinal level. The placental mammals are generally divided into four supra or superorders: Afrotheria (Stanhope et al. 1998), Euarchontoglires (or Supraprimates: Waddell et al. 2001), Laurasiatheria (Waddell et al. 1999b) and Xenarthra (Cope 1889). These clades were also recovered in our analyses with BPP and BS of 100%, except for Euarchontoglires where BS was only 61%. Euarchontoglires and Laurasiatheria are often grouped together in a clade called Boreoeutheria (Murphy et al. 2001a, b, Kriegs et al. 2006, Nishihara et al. 2006, Arnason et al. 2008). However, the relationships

among Boreoeutheria and the two remaining clades is currently debated. All three possible scenarios of divergence have been proposed. Some suggest that Afrotheria is at the base of the eutherian tree (Beck et al. 2006, Nikolaev et al. 2007, Nishihara et al. 2007), others support a basal xenarthran root (Kriegs et al. 2006, Svartman et al. 2006) or a basal split between Xenarthra/Afrotheria and Boreoeutheria (Hallström et al. 2007, Kjer & Honeycutt 2007, Murphy et al. 2007, Waters et al. 2007, Wildman et al. 2007, Prasad et al. 2008). Recently, molecular studies based on retroposons have concluded in a simultaneous divergence of these three clades (Churakov et al. 2009, Nishihara et al. 2009). Given the lack of consensus regarding the order of divergence at the base of Eutheria, and given incongruent ML and BML trees in this study, the relationships among Afrotheria, Euarchontoglires, Laurasiatheria and Xenarthra are depicted by a polytomy on both model trees.

Afrotheria

Afrotheria comprises six orders: Afrosoricida, Hyracoidea, Macroscelidae, Proboscidea, Sirenia and Tubulidentata. The monophyly of Afrotheria is supported by various types of characters: morphology (Sánchez-Villagra et al. 2007, Seiffert 2007), molecular sequences (Arnason et al. 2008), indels (Madsen et al. 2001, Amrine-Madsen et al. 2003), and retroposons (Nishihara et al. 2005, 2006). Two major groups are generally recognized within Afrotheria: Afroinsectiphillia (i.e., Afrosoricida, Macroscelidae and Tubulidentata: Waddell et al. 2001b) and Paenungulata (i.e., Hyracoidea, Proboscidea and Sirenia: Simpson 1945).

Afroinsectiphillia

The monophyly of Afroinsectiphillia (Waddell et al. 2001b) is supported by numerous studies (e.g., Robinson et al. 2004, Nishihara et al. 2005, Waters et al. 2007), but contradicted by others (e.g., Waters et al. 2007, Arnason 2008). The two families that compose Afrosoricida (i.e., Tenrecidae and Chrysochloridae) were not retrieved as a monophyletic group in neither ML nor BML analyses. Rather Chrysochloridae grouped with Macroscelididae (BS=64%, BPP=100%), as obtained in a mitogenomic analysis (Arnason et al. 2008) and in a study of LINEs (Waters et al. 2007). However, the relationship of Tenrecidae and Tubulidentata differs from Arnason et al.'s (2008) study. Therefore, we collapsed the node that grouped Tenrecidae to Tubulidentata (which was poorly supported in the ML analysis: BS=46%).

Paenungulata

Relationships among the three orders that compose paenungulates are still unresolved (Rokas & Caroll 2006, Seiffert 2007) and different types of molecular data suggest contradicting and poorly supported topologies (Nishihara et al. 2005, Kellogg et al. 2007, Pardini et al. 2007, Seiffert 2007). However, morphological studies strongly agree with a Tethytheria hypothesis (i.e., Proboscidea/Sirenia clade: Novacek 1986, Asher et al. 2003). In the present study, the relationships inferred with mitogenomic sequences do not support a Tethytheria clade, which is also in contradiction with recent mitogenomic support in favor of Tethytheria (Kjer & Honeycutt 2007, Arnason et al. 2008). Thus, the node that excluded Proboscidae from the two remaining orders was collapsed in our model tree (this particular node has the lowest BS within Paenungulata: BS=73%, compare to BS=100% for all other nodes).

Laurasiatheria

Laurasiatheria includes six orders: Carnivora, Cetartiodactyla, Chiroptera, Eulipotyphla, Perrisodactyla and Pholidota. However, the order Pholidota was not included in our model tree. The branching order within Laurasiatheria and monophyly of this group has been highly debated (see Springer et al. 2004b, 2007 for a review). However, the Fereuungulata clade (comprised of Perrisodactyla, Carnivora, Cetartiodactyla and Pholidota) seems to be well supported in many studies (e.g., Pumo et al. 1998, Waddell et al. 2001b, Murphy et al. 2004, Kjer & Honeycutt 2007).

Fereuungulata

Perrisodactyla. The exact position of Perrisodactyla within Laurasiatheria remains uncertain. It has been placed as sister group to Cetartiodactyla in some studies (Murphy et al. 2001b, Lin et al. 2002a, Beck et al. 2006, May-Collado & Agnarsson 2006), but the current consensus tends to favor a Perrisodactyla/Pholidota/Carnivora or a Perrisodactyla/Carnivora clade as retrieved in this study (Murphy et al. 2001a, Arnason & Janke 2002, Kjer & Honeycutt 2007, Arnason et al. 2008, Prasad et al. 2008).

Carnivora. The order Carnivora is subdivided in two suborders: Caniformia and Feliformia. The only two feliform families sampled in this study, Felidae and Herpestidae, were retrieved as a monophyletic group. The remaining families of Caniformia are divided into: Arctoidea and Cynoidea. The Cynoidea only includes

Canidae, and the remaining families are placed in Arctoidea, which is further subdivided into: Musteloidea (Procyonidae, Mustelidae, Ailuridae and Mephitidae), Pinnipedia (Otariidae, Odobenidae and Phocidae) and Ursidae. While monophyly and interfamilial relationships of Pinnipedia are well established (Arnason & Janke 2002, Davis et al. 2004, Delisle & Strobeck 2005, Fulton & Strobeck 2006, Higdon et al. 2007), the relative position of families within Musteloidea is currently debated (e.g., Delisle & Strobeck 2005, Flynn et al. 2005, Fulton & Strobeck 2006, Arnason et al. 2007). Recent studies of caniform relationships recovered a Pinnepedia/Musteloidea clade with Ursidae as the next branching lineage (Delisle & Strobeck 2005, Flynn et al. 2005, Fulton & Strobeck 2006, Arnason et al. 2007). In a recent mitogenomic studies, Arnason et al. (2008) left unresolved the relationships among these three groups. Given that ML and BML results did not support the monophyly of Musteloidea, and because it contradicts the currently supported relationships of Caniformia, three nodes were collapsed in the consensus model tree to ensure its compatibility with recent molecular analyses.

Cetartiodacyla. Within Cetartiodactyla, Hippopotamidae is the sister group to Cetacea (Beck et al. 2006, May-Collado & Agnarsson 2006, Kjer & Honeycutt 2007, Agnarsson & May-Collado 2008, O'Leary & Gatesy 2008). The remaining families are grouped into three well supported clades: Tylopoda (Camelidae), Suina (Suidae & Tayassuidea) and six families in Ruminantia, three of which are included in this study: Bovidae, Cervidae and Giraffidae (Agnarsson & May-Collado 2008, O'Leary & Gatesy 2008). These three clades were recovered in both ML and BML analyses. However some studies suggest a basal position for Tylopoda (e.g., Agnarsson & May-Collado 2008, O'Leary & Gatesy 2008), which was not recovered in our analyses. Rather, a Tylopoda/Suina clade was supported (BS=63%, BPP=100%), and also obtained in other studies (Arnason et al. 2002). Given these inconsistencies, the node supporting the Tylopoda/Suina clade was collapsed.

Cetacea is divided in two suborders: Mysticeti, represented by baleen whales and Odontoceti, represented by toothed whales and dolphins (Arnason et al. 2004, May-Collado & Agnarsson 2006, Agnarsson & May-Collado 2008, Xiong et al. 2009). These two clades were recovered in our analyses with 100% BS and BPP for Mysticeti, and 42% BS and 100% BPP for Odontoceti. Although monophyly of Odontoceti was confirmed in recent mitochondrial studies (e.g., Arnason et al. 2004, May-Collado &

Agnarsson 2006, Agnarsson & May-Collado 2008, Xiong et al. 2009), familial relationships within this clade remain uncertain, especially with respect to the placement of Ziphiidae, Physeteridae and Platanistidae (Arnason et al. 2004, May-Collado & Agnarsson 2006, Agnarsson & May-Collado 2008, Xiong et al. 2009). Aside these problematic families, the four remaining families included in this study, i.e., Monodontidae, Phocoenidae, Delphinidae and Iniidae are systematically recovered as a monophyletic group in recent molecular studies (Arnason et al. 2004, May-Collado & Agnarsson 2006, Agnarsson & May-Collado 2008, Xiong et al. 2009). Three of these four families form the superfamily Delphinoidea (excluding Iniidae). In order to make our model tree compatible with the different phylogenies obtained from mitogenomic studies, we collapsed two poorly supported nodes that define the relationships among Ziphiidae, Physeteridae and Platanistidae (BS=28 and 36%, BPP=51 and 100%).

Chiroptera

Chiroptera has recently been placed in a clade with Perrisodactyla and Carnivora (Pegasoferae: Nishihara et al. 2006). This result contrasts with numerous studies where Chiroptera was placed as a sister group to Fereuungulata (e.g., Beck et al. 2006). This latter association, which was retrieved in our model tree, is also supported by other mitogenomic studies (e.g., Pumo et al. 1998, Kjer & Honeycutt 2007, Arnason et al. 2008). Chiroptera encompasses 18 families that have been divided in two suborders: Megachiroptera and Microchiroptera (Simmons & Geisler 1998). However, some molecular studies have questioned microbat monophyly. Rather, most molecular studies support a Pteropodiformes/Vespertilioniformes dichotomy (Hutcheon & Kirsch 2006, Teeling et al. 2005, Kjer & Honeycutt 2007). Our study is consistent with this latter hypothesis with Rhinolophidae and Pteropodidae in one clade (Pteropodiformes: BS=78%, BPP=100%), and Mysticinidae, Phyllostomidae and Vespertilionidae in the other (Vespertilioniformes: BS=95%, BPP=100%).

Eulipotyphla

Eulipotyphla includes three families: Erinaceidae, Soricidae and Talpidae. Some classifications groups the three families into two different orders: Erinaceomorpha and Soricomorpha (e.g., Wilson & Reeder 2005). Whereas a number of studies did recover a monophyletic Eulipotyphla (e.g., Nikaido et al. 2003, Arnason et al. 2008), others did not (e.g., Kjer & Honeycutt 2007). Also, the exact position of Eulipotyphla within Placentalia has been problematic, probably due to long-branch attraction in hedgehogs

(Erinaceidae) and shrews (Soricidae) as well as to their unusually high mitogenomic AT content. Indeed, some studies placed Eulipotyphla outside of Laurasiatheria, at the base of Placentalia (Mouchaty et al. 2000, Arnason et al. 2002). However, analyses that used more appropriate models of evolution or phylogenetic methods, and greater taxon sampling, recovered a monophyletic Laurasiatheria with a basal position of Eulipotyphla (Murphy et al. 2001a, b, Amrine-Madsen et al. 2003, Nikaido et al. 2003, Beck et al. 2006, Nishihara et al. 2006, Nikolaev et al. 2007, Arnason et al. 2008, Prasad et al. 2008). In the present study, a monophyletic Eulipotyphla was recovered in both ML and BML analyses (100% BS and BPP) at the base of Laurasiatheria, probably because Erinaceidae was removed from the analysis and that extra species were added within Talpidae and Soricidae families.

Xenarthra

Monophyly of the order Xenarthra and its four constituent families is well established in molecular studies (Delsuc et al. 2002). Morevover, the relationships among the four families are also strongly supported by different datasets (e.g., Arnason et al. 2008). The molecular consensus suggests a basal position for the family Dasypodidae (armadillos), sometimes elevated to its own order, Cingulata (Wilson & Reeder 2005). The remaining three families form a clade, referred to as the order Pilosa by some authors (Wilson & Reeder 2005), where Myrmecophagidae (anteaters) is the sister group to sloths (Megalonychidae and Bradypodidae). Monophyly of Xenarthra was confirmed in both ML and BML analyses with 100% support.

Euarchontoglires

Euarchontoglires comprises five orders: Dermoptera, Lagomorpha, Primates, Rodentia and Scandentia, that are divided in two groups: Euarchonta (Dermoptera, Primates and Scandentia) and Glires (Lagomorpha and Rodentia). Although some mitochondrial and nuclear studies had challenged the monophyly of Glires (D'Erchia et al. 1996, Arnason et al. 2002, Horner et al. 2007, Wildman et al. 2007), the current molecular consensus, including mitogenomic studies, support a monophyletic Glires (e.g., Lin et al. 2002b, Beck et al. 2006, Nishihara et al. 2006, Kjer & Honeycutt 2007, Kriegs et al. 2007, Arnason et al. 2008, Prasad et al. 2008).

Primates

Primates are divided in two suborders: Strepsirrhini and Haplorrhini. Within primates, numerous higher-level clades are well supported by molecular datasets (see review by Disotell 2008, Prasad et al. 2008).

Suborder Haplorrhini. Haplorrhine primates are divided in three groups: (1) the parvorder Platyrrhini (New World monkeys), (2) the parvorder Catarrhini (Old World monkeys and apes) and (3) the infraorder Tarsiiformes (tarsiers). Platyrrhine and catarrhine primates are further grouped into Anthropoidea (humans, apes and monkeys). Within Haplorrhini, the phylogenetic position of Tarsiiformes (family Tarsidae) remains controversial. Nuclear and mitochondrial studies have supported a close relationship of tarsiers and Anthropoidea (Poux and Douzery 2004, Gibson et al. 2005, Beck et al. 2006, Matsui et al. 2009), although alternative topologies could not be statistically rejected (Poux and Douzery 2004, Matsui et al. 2009). Also, different analytical methods or different datasets have proposed a closer relationship with Strepsirrhini, thus making the suborder Haplorrhini polyphyletic (Hudelot et al. 2003). Others placed Tarsiiformes at the base of the primate group (Arnason et al. 2002, Matsui et al. 2009) or were unable to provide resolution at that node (Herke et al. 2007, Arnason et al. 2008). In this study, the phylogenetic position of Tarsidae differed in the ML and BML analyses and thus, was represented by a polytomy at the base of Primates in both model trees.

Parvorder Catarrhini. A number of studies support a monophyletic Catarrhini (e.g., Poux et al. 2006, Herke et al. 2007, Matsui et al. 2009). In this study, three catarrhine families were included: Cercopithecidae, Hylobatidae (the gibbons) and Hominidae (the great apes and humans).

Parvorder Platyrrhini. A large number of phylogenetic studies support the monophyly of platyrrhine primates (e.g., Springer et al. 2003, Ray et al. 2005, Poux et al. 2006, Herke et al. 2007, Matsui et al. 2009). Two representative families of the New World monkeys were included in this study: Cebidae (the capuchins) and Aotidae (the owl monkeys).

Suborder Strepsirrhini. The strepsirrhine primates are generally divided into three suborders: (1) Lorisiformes (three families, one of which is included in this study: Lorisidae), (2) Lemuriformes (four families, two of which are included in this study:

Lemuridae and Indriidae) and (3) Chiromyiformes, with a single family: Daubentoniidae and only one extant species: the Aye-aye. Congruent with our analyses, a large number of phylogenetic studies support the monophyly of the strepsirrhine primates (e.g., Poux et al. 2006, Herke et al. 2007, Arnason et al. 2008, Matsui et al. 2009).

Dermoptera

The order Dermoptera consists of only two extant species of flying lemurs or Colugos (family Cynocephalidae). The phylogenetic position of Dermoptera within the placental mammals is still debated (Martin et al. 2008). Some studies support a closer relationship between Dermoptera and Primates (Hudelot et al. 2003, Beck et al. 2006, Bininda-Edmonds et al. 2007, Janecka et al. 2007) or Scandentia (Springer et al. 2004b, Nie et al. 2008). Others suggest a phylogenetic position nested within Primates, as sister group to Anthropoidea (Murphy et al. 2001a, Arnason et al. 2002, 2008, Kjer & Honeycutt 2007). This latter relationship was retrieved in our study in both ML and BML analyses. However, because of the uncertainties of its phylogenetic placement, reflected by the inconsistencies obtained in previous studies, the evolutionary relationship between Dermoptera and Primates was represented by a polytomy in the consensus model tree.

Rodentia

Rodents are the most speciose group of mammals and are divided into 33 families and five suborders: Sciuromorpha, Castorimorpha, Myomorpha, Anomaluromorpha and Hystricomorpha (Wilson & Reeder 2005). Nine rodent families had at least one species with a complete mtgenome sequence available. These nine families fall into three of the five suborders; a complete mt sequence was not available for Castorimorpha and the Anomaluridae representative was removed from the analysis due to its fast evolving rate. Sciuromorpha includes Sciuridae and Gliridae; Myomorpha includes Cricetidae, Dipodidae, Muridae and Spalacidae; and Hystricomorpha includes Caviidae and Thryonomyidae. Although rodent monophyly had been questioned by early molecular studies (D'Erchia et al. 1996), recent analyses generally recognize its monophyly (Lin et al. 2002, Poux et al. 2006, Kjer & Honeycutt 2007, Arnason et al. 2008). In the present study, the phylogenetic relationships among families were congruent to recent nuclear and mitochondrial studies (Montgelard et al. 2008, Blanga-Kanfi et al. 2009) and were strongly supported in both ML and BML analyses.

Scandentia

Scandentia are divided in two families: Tupaiidae and Ptilocercidae, which are composed of different species of tree shrews. Only Tupaiidae is included in this study. Similar to the ambiguous position of Dermoptera, the phylogenetic position of Scandentia changes according to the dataset and the inference method used (Martin et al. 2008). Apart from the Scandentia/Dermoptera clade mentioned in the Dermoptera section, it has also been placed at the base of a Primates/Dermoptera group (e.g., Beck et al. 2006), as a sister clade to Lagormorpha (e.g., Arnason et al. 2002, Lin et al. 2002), at the base of Euarchontoglires (Kjer & Honeycutt 2007), or as a polytomy with primates and glires (Arnason et al. 2008). Given, these inconsistencies, we opted for the topology presented by Arnason et al. (2008), i.e. a polytomy at the base of Euarchontoglires.

CHAPITRE 6:

DISCUSSION GÉNÉRALE

6.1. DISCUSSION

Le but premier de ma thèse est de contribuer au domaine de la phylogénomique en validant certaines approches qui ont été proposées et en vérifiant leur utilité dans un contexte phylogénomique. Avec l'accumulation des séquences génomiques, l'analyse phylogénétique subit une révolution où l'emphase n'est plus mise sur la recherche de moyens pour générer des données moléculaires à faible coût, mais plutôt sur le développement de méthodes adaptées au traitement d'un nombre élevé de données. Il est donc crucial de trouver des approches phylogénétiques qui permettent de réduire le temps de calcul et la demande grandissante en puissance informatique. Parallèlement, de nombreux débats méthodologiques qui perdurent ont été transposés aux méta-analyses plutôt que d'être résolus par celles-ci. Par exemple, le débat qui opposait les partisans de l'analyse séparée et de l'analyse combinée (Huelsenbeck et al. 1996b) est maintenant appliqué aux méta-analyses et oppose les partisans des analyses de type super-matrice à ceux des analyses de type super-arbre (Bininda-Emonds 2004a, de Queiroz & Gatesy 2007). Les partisans de l'analyse conditionnelle (Bull et al. 1993, de Queiroz 1993), quant à eux, sont limités par des problèmes méthodologiques. Puisque les tests de congruence les plus utilisés ont été vivement critiqués (Huelsenbeck et al. 1996b, Cunningham 1997a, b, Barker & Lutzoni 2002, Darlu & Lecointre 2002, Leigh et al. 2008), il est impératif de développer de nouveaux tests et de s'assurer qu'ils pourront traiter efficacement et simultanément un grand nombre de jeux de données. Un autre exemple de débat qui ne semble pas vouloir s'éteindre concerne l'effet des matrices avec des données manquantes sur l'inférence phylogénétique. Alors qu'il pourrait sembler que le problème des données manquantes a été enrayé avec l'arrivée de la phylogénomique, il n'en est malheureusement pas ainsi. En effet, certains auteurs affirment que les matrices avec des données manquantes ne réduisent pas la qualité de l'inférence phylogénétique en autant qu'il y ait suffisamment de caractères informatifs (Wiens 2003b, 2006, Driskell et al. 2004, Philippe et al. 2004, Wiens & Moen 2008), ce qui est le cas pour de longues séquences d'ADN ou autres matrices génomiques. Pourtant, une étude récente a démontré que la façon dont les données manquantes sont distribuées dans une matrice influence l'inférence phylogénétique (Hartmann & Vision 2008). Les matrices de EST, qui sont de plus en plus utilisées en phylogénomique, sont souvent caractérisées par des paires de taxons pour lesquelles aucune séquence commune n'est disponible. Hartmann & Vision (2008) ont observé que l'exactitude phylogénétique était grandement diminuée lors de l'analyse de telles

matrices en comparaison avec des matrices où des blocs de données sont manquants (e.g., des gènes complets manquants pour quelques taxons). De plus, les méta-analyses utilisent principalement des méthodes de distance ou de parcimonie qui sont plus affectées par les données manquantes que le sont les méthodes probabilistes (Hartmann & Vision 2008).

Bien qu'il y ait plusieurs autres débats ou méthodes qui mériteraient d'être étudiés (par exemple, l'amélioration des modèles évolutifs de nucléotides), ma thèse est limitée à trois objectifs principaux qui sont en relation avec les thématiques citées au paragraphe précédent. Outre les questions méthodologiques, je voulais également approfondir les relations évolutives entre les familles de mammifères puisque celles-ci sont trop souvent délaissées en faveur des phylogénies interordinales.

On trouvera dans les lignes qui suivent un retour sur les trois objectifs de ma thèse, un bref résumé des résultats obtenus dans chacun des chapitres, ainsi qu'une discussion de la contribution de chaque chapitre au domaine de la phylogénétique et phylogénomique. Une conclusion générale où sont proposées d'éventuelles ouvertures de recherche termine cette discussion.

6.1.1. La validation du test de CEMD à partir de matrices de distances ultramétriques et additives

Mon premier objectif était de vérifier le comportement du test de la Congruence Entre des Matrices de Distance (CEMD) lorsque appliqué à des matrices de distance ultramétrique et additive. Ce test a originalement été proposé pour mesurer la congruence entre des matrices de dissimilarité (Legendre & Lapointe 2004). Dans ces conditions, il a été démontré que le test avait une erreur de type I juste et une bonne puissance. Puisque le test possède plusieurs avantages en comparaison d'autres tests de congruence, je voulais vérifier sa validité dans un contexte phylogénétique ou phylogénomique. Ainsi, le chapitre 2 décrit les résultats de l'analyse de l'erreur de type I et de la puissance du test de CEMD lorsque utilisé pour comparer des matrices de distance ultramétrique calculées à partir de dendrogrammes générés de façon aléatoire. Les distances et les arbres ultramétriques sont répandus en phylogénétique, entre autres, dans les études de datation des phylogénies (ex.: Hallström et al. 2007, Higdon et al. 2007, Arnason et al. 2008, Beck 2008, Meredith et al. 2008a, b, Xiong et

al. 2009) ou encore dans le cas de consensus ou de méthodes de distance qui ne tiennent pas compte des longueurs de branches (ex.: MSS: Creevey et al. 2004). Dans le chapitre 3, le test de CEMD a été appliqué à des matrices de distance calculées à partir de séquences d'ADN simulées sur des arbres additifs générés de façon aléatoire. En accord avec les résultats de Legendre & Lapointe (2004), les résultats présentés aux chapitres 2 et 3 indiquent que le test de CEMD a une erreur de type I adéquate et une bonne puissance lorsque utilisé dans un contexte phylogénétique. Plus précisément, la puissance du test de CEMD augmente avec le nombre de taxons, et avec le nombre de matrices congruentes incluses dans un ensemble de matrices. Ce comportement de la puissance est conforme aux attentes. En effet, la puissance d'un test est fonction du nombre d'objets et du nombre d'événements attendus, c'est-à-dire, dans ce cas ci, le nombre de matrices congruentes (Cohen 1988). Puisque j'ai démontré que le test de CEMD pouvait être appliqué dans un contexte phylogénétique, je l'ai utilisé dans le chapitre 5 afin de tester la congruence des séquences des 12 gènes présents sur le brin-lourd de l'ADN mitochondrial de 102 espèces de mammifères. Le test de CEMD a révélé que les 12 gènes étaient congruents, ce qui est conforme aux attentes puisque ces gènes proviennent du même brin de l'ADN mitochondrial et sont donc soumis à un patron d'évolution identique.

Le test de CEMD présente de nombreux avantages qui le qualifient comme étant un candidat idéal pour déterminer la congruence entre des jeux de données dans un contexte phylogénomique. Par exemple, la statistique est calculée directement à partir de matrices de distance, ce qui permet la comparaison entre différents types de données lorsque celles-ci sont converties en distances en utilisant une fonction appropriée. Puisque les méta-analyses incluent souvent des données provenant de plusieurs sources, il est possible de comparer toutes les matrices en même temps (ex.: utilisation de caractères morphologiques et moléculaires dans une analyse combinée: Asher et al. 2004, Asher 2007, Seiffert 2007, O'Leary & Gatesy 2008). En outre, les analyses phylogénomiques ont souvent recours à des méthodes de distance pour inférer les arbres phylogénétiques puisqu'elles sont beaucoup plus rapides que les méthodes probabilistes. Ainsi, le test peut être appliqué directement sur ces matrices de distance. Par ailleurs, les distances utilisées peuvent être corrigées avec un modèle d'évolution adapté à chaque jeu de données (dans le cas de séquences moléculaires). De plus, les distances peuvent être calculées sur les arbres phylogénétiques afin d'obtenir les distances patristiques de l'arbre ce qui offre une méthode intéressante

pour tester la congruence lorsqu'une approche de type super-arbre est utilisée. Aussi, les matrices de distance peuvent être pondérées différemment pour tenir compte du nombre de caractères inclus dans les jeux de données. La pondération des jeux de données par le nombre de caractères est d'ailleurs déjà utilisée dans plusieurs autres tests statistiques et dans certaines méthodes de super-arbre (ex.: SDM: Criscuolo et al. 2006). Pour départager les matrices congruentes des matrices incongruentes, des tests *a posteriori* peuvent être effectués. Finalement, étant donné la façon dont le test est construit (voir sections 2.4 et 3.4.1), il est extrêmement rapide, ce qui procure un avantage non-négligeable en phylogénomique où la taille et le nombre des jeux de données est considérable. Avec la quantité croissante de séquences disponibles et les analyses qui tendent à inclure de plus en plus de taxons, le test de CEMD offre une alternative efficace pour comparer plusieurs matrices et identifier si elles peuvent être combinées dans une analyse unique de type super-matrice ou si une méthode de type super-arbre devrait être préférée.

6.1.2. La validation de l'approche par taxons chimères à l'aide de simulations et de données empiriques provenant d'espèces de mammifères

Le deuxième objectif avait pour but de comparer l'exactitude des estimations phylogénétiques provenant de matrices où une proportion des données est manquante à des matrices où le nombre de données manquantes a été réduit par la formation de taxons chimères. En effet, il est possible de réduire le nombre de données manquantes dans les super-matrices en combinant les séquences de différentes espèces pour obtenir une séquence d'ADN complète: ces séquences hybrides sont appelées chimères. L'exactitude des arbres phylogénétiques estimés à l'aide de matrices qui comprennent des taxons chimères a été remise en question (Malia et al. 2003), et la meilleure stratégie pour l'analyse de super-matrices incomplètes est débattue (Malia et al. 2003, Springer et al. 2004a). Aussi, alors que certains auteurs estiment que la présence de données manquantes ne réduit pas la qualité de l'estimation phylogénétique (Wiens 2003b, 2006, Driskell et al. 2004, Philippe et al. 2004, Wiens & Moen 2008), d'autres considèrent que l'exactitude de l'inférence phylogénétique dépend de la façon dont les données manquantes sont distribuées dans une matrice (Hartmann & Vision 2008).

Grâce à des simulations numériques, nous avons comparé l'exactitude phylogénétique des deux approches concurrentes dans différentes conditions. Des séquences d'ADN ont été simulées sur un arbre modèle de 42 taxons représentant dix groupes monophylétiques de 4 taxons. À partir des matrices de données complètes, différents pourcentages de gènes ont été enlevés au hasard pour générer des matrices incomplètes. Ces super-matrices incomplètes ont été analysées de deux façons : soit en codant chaque position manquante avec un "?", ou encore en réduisant la quantité de données manquantes en créant des taxons composites. Un total de 180 combinaisons de paramètres ont été analysées, c.-à.-d. une combinaison de (1) différentes longueurs de branches de l'arbre modèle, (2) différentes longueurs de jeux de données, (3) différents pourcentages de données manquantes, (4) différents modèles d'évolution pour simuler les séquences d'ADN et (5) différentes méthodes d'inférence phylogénétique. Tous ces paramètres ont jusqu'à un certain degré influencé l'exactitude des arbres inférés. En effet, une meilleure exactitude phylogénétique a été retrouvée lorsque l'arbre modèle présentait des longueurs des branches externes réduites par rapport aux branches internes. Ainsi, la *stemminess* de l'arbre est augmentée (Fiala & Sokal 1985). Plusieurs auteurs ont démontré qu'un arbre caractérisé par une faible *stemminess* est plus difficile à estimer (Fiala & Sokal 1985, Rokas et al. 2005, Weisrock et al. 2005). Ensuite nous avons observé de meilleures valeurs d'exactitude lorsque des séquences d'ADN plus longues étaient utilisées. De plus, de façon similaire à ce qui avait déjà été proposé (Wiens 2003b, 2006, Wiens & Moen 2008), pour un même pourcentage de données manquantes, les matrices avec un plus grand nombre de caractères présentaient une exactitude accrue. Par contre, pour des jeux de données de même taille, une diminution de l'exactitude phylogénétique a été observée avec l'augmentation de la proportion de données manquantes. Cet effet négatif des données manquantes a d'ailleurs été noté dans d'autres études (voir les études citées par Wiens 2006). En ce qui a trait aux différents modèles d'évolution, les relations entre les taxons présentées sur l'arbre modèle étaient plus difficilement obtenues lorsque les jeux de données étaient simulés avec un modèle d'évolution plus complexe (i.e., TVM vs. JC). Ceci est en accord avec l'affirmation selon laquelle l'inférence phylogénétique est plus difficile lorsque les séquences évoluent dans le cadre d'un modèle plus complexe (par exemple: Yang 1996a, Pollock & Bruno 2000). Cependant, lorsque l'inférence phylogénétique utilise un modèle d'évolution correspondant au modèle choisi pour simuler les données, l'exactitude de l'inférence est augmentée et même drastiquement dans certaines situations (en accord avec

Posada & Crandall 2001). La méthode d'inférence joue aussi un rôle, en général les méthodes probabilistes retrouvent l'arbre modèle plus souvent que les méthodes de distance. Cette tendance a été observée dans les simulations du chapitre 4 (ML vs. NJ), lorsque les séquences d'ADN étaient courtes ou lorsque la proportion de données manquantes était élevée.

Pour tous les paramètres cités précédemment, les matrices présentant des données manquantes ont été comparées aux matrices où les données manquantes ont été réduites par la formation de taxons chimères. Dans la grande majorité des cas, aucune différence significative quant à l'exactitude phylogénétique n'a été observée entre les matrices de données manquantes et les chimères. Néanmoins, nous avons observé une exactitude significativement plus élevée pour les matrices chimères dans 46 des 180 combinaisons, alors que les matrices avec données manquantes étaient significativement plus performantes dans huit cas seulement. Généralement, les matrices chimères étaient supérieures dans les cas où le pourcentage de données manquantes était intermédiaire (de 15 à 50 % de données manquantes). Lorsque le niveau de données manquantes était plus faible (5 %), des résultats optimaux étaient obtenus avec les deux approches alors qu'à un niveau plus élevé (75 %), une baisse marquée de l'exactitude était observée pour les deux approches. D'ailleurs, le pourcentage de données manquantes était moins réduit par la formation de taxons chimères à des niveaux plus élevées. D'après ces résultats, la formation de taxons chimères est bénéfique, pour l'inférence phylogénétique, à des niveaux intermédiaires de données manquantes.

Outre les simulations numériques, la comparaison entre les deux approches a aussi été étudiée avec des séquences d'ADN provenant d'espèces de mammifères et avec un arbre modèle inféré à partir de séquences complètes. Les espèces de mammifères incluses dans l'étude ont été choisies afin de représenter le plus grand nombre de familles possible, selon la disponibilité des séquences de génomes mitochondriaux. Un total de 102 espèces représentant 93 familles de mammifères ont été alignées. La topologie de l'arbre obtenu suite aux analyses ML et BML est en accord avec le consensus moléculaire actuel des relations évolutives entre les espèces et groupes taxonomiques de mammifères. Les résultats, présentés au chapitre 5 sont similaires à ceux obtenus suite aux simulations numériques (chapitre 4) puisque l'exactitude phylogénétique des arbres inférés par les deux approches est similaire.

La création de taxons chimères est couramment utilisée pour réduire le pourcentage de données manquantes dans une matrice de séquences d'ADN (e.g., Shoshani & McKenna 1998, Madsen et al. 2001, Murphy et al. 2001a, Scally et al. 2001, Asher et al. 2004, Springer et al. 2004, Flynn et al. 2005, Delsuc et al. 2006, Marek & Bond 2006, Poux et al. 2006, Philippe et al. 2007, Telford 2007, Beck 2008, Bourlat et al. 2008, Duvall et al. 2008). Il était donc important de déterminer la validité de cette approche. Les résultats obtenus aux chapitres 4 et 5 appuient l'utilisation de séquences chimères pour réduire la quantité de données manquantes dans une super-matrice. De par la réduction du nombre de taxons dans les matrices, l'approche des séquences chimères réduit grandement le temps de calcul, un aspect favorable pour les études phylogénomiques, où un nombre élevé de taxons et de caractères est utilisé (Eisen & Fraser 2003, Telford 2007). À titre d'exemple, une analyse de maximum de vraisemblance telle que celle décrite au chapitre 5 (p.132) pour le jeu de données complet (102 espèces et 7100bp.), prend environ huit heures de calcul sur un Power Mac G5, avec processeurs PowerPC 970MP (2 x 2.5 GHz), lorsque seulement cinq arbres de départ aléatoires sont utilisés au lieu de dix. Pour les jeux de données ayant 30% de données manquantes ou plus, le nombre de taxons est réduit de 102 à 55 ou moins lorsque l'approche des taxons chimères est choisie. Le temps de calcul pour une matrice de 51 taxons est d'environ 50 minutes pour une analyse identique à celle décrite précédemment, soit environ huit fois plus rapide. De plus, chacun des réplicats de *bootstrap* demande le même temps de calcul lorsque les paramètres de recherche d'arbres sont identiques, et donc 800 heures sont nécessaires pour obtenir 100 réplicats de *bootstrap* pour l'arbre comprenant 102 espèces, contre environ 83 heures pour l'arbre de 51 espèces chimères. Bien entendu, de telles analyses peuvent difficilement être réalisées sur un seul ordinateur et nous avons recours à des grappes informatiques où les tâches peuvent être subdivisées pour accélérer les calculs. Il est par exemple facile d'utiliser un processeur différent pour chacune des cinq analyses débutant avec un arbre de départ aléatoire différent (diminuant ainsi le temps de calcul par un facteur de 5).

Cependant, malgré les outils informatiques à la fine pointe de la technologie dont nous disposons, le temps de calcul reste un élément contraignant, surtout lorsque plusieurs analyses doivent être réalisées et que le nombre de taxons augmente. Ce problème est de toute évidence amplifié dans le cadre d'étude par simulations où plusieurs jeux de données représentant différents paramètres et comprenant de nombreux réplicats

doivent être analysés. Ainsi, afin de réduire le temps de calcul, souvent compté en jours, voir en semaines, j'ai dû utiliser des paramètres de recherche d'arbres moins poussés et/ou moins de réplicats pour certaines analyses. Par exemples, lorsque plusieurs réplicats devaient être analysés: (1) un algorithme de nearest neighbor interchange (NNI) a été utilisé plutôt qu'un algorithme de subtree pruning and regrafting (SPR) qui est plus performant mais aussi plus lent, (voir chapitre 5, section 5.4.2.3), (2) le nombre d'arbre de départ a été diminué de dix à un seul, (3) dans certains cas, le nombre de réplicats a été réduit de 1000 à 100 pour les jeux de données de plus grosses tailles afin de garder le temps de calcul sous le seuil des dix jours par paramètres testés (voir les différents cas analysés au Tableau 4.2). Finalement, afin de donner une idée du temps qui peut être requis pour faire les analyses phylogénétiques; une période de 365 jours aurait été nécessaire pour obtenir une analyse Bayésienne de tous les jeux de données présents dans le Tableau 4.2 avec une grappe informatique Altix 4700 et avec des paramètres standards (similaires à ceux mentionnés au chapitre 5, section 5.4.1.2). La généralisation de ces résultats à d'autres types de matrices, par exemple, aux données morphologiques, mériterait d'être étudiée davantage.

6.1.3. La comparaison des méthodes de types consensus et super-arbre pour inférer la phylogénie des familles de mammifères

Le troisième objectif avait pour but de comparer la performance des méthodes de consensus et de type super-arbre dans un contexte où tous les arbres de départ possèdent des taxons identiques (*consensus setting*, *sensu* Bininda-Emonds 2003). Les méthodes de super-arbre sont une généralisation du cas particulier de consensus où les arbres de départ (*source trees*) ont les mêmes taxons. Ces deux types de techniques ont été abondamment étudiés séparément (ex.: de Queiroz 1993, de Queiroz et al. 1995, Bininda-Emonds & Sanderson 2001, Bininda-Emonds 2003, Wilkinson et al. 2005, Criscuolo et al. 2006) mais ont aussi été comparés dans un contexte de consensus (ex.: Lapointe 1998b, Lapointe et al. 1999, Levasseur & Lapointe 2001, 2006, Bininda-Emonds 2003, Wilkinson et al. 2005, 2007). Les méthodes de consensus offrent une solution rapide, alors que les méthodes de super-arbre utilisent un critère d'optimalité et sont donc plus longues. Alors que certains chercheurs affirment que les méthodes de consensus sont peu intéressantes puisqu'elles procurent peu de résolution (ex.: Barrett et al. 1991, Kluge & Wolf 1993, de Queiroz et al. 1995), d'autres soutiennent que les méthodes de consensus qui tiennent

compte des longueurs de branches sont plus résolues et plus exactes (Lapointe et al. 1999, Levasseur & Lapointe 2001, 2006).

Puisque les relations phylogénétiques des mammifères sont assez bien connues, une topologie modèle peut être obtenue et différentes méthodes peuvent être testées. Les gènes de la mitochondrie ont été choisis comme marqueurs phylogénétiques puisque de nombreux génomes mitochondriaux sont disponibles pour les mammifères. Presque toutes les familles de mammifères pour lesquelles une séquence mitochondriale complète était disponible ont été inclus dans l'analyse afin d'avoir un échantillonnage complet. Les 12 gènes mitochondriaux du brin lourd (H) ont été analysés séparément et combinés par la suite par une méthode de consensus ou de super-arbre pour produire un arbre final. Ces arbres finaux ont été comparés à l'arbre modèle obtenu par l'analyse de la super-matrice de tous les gènes. Ces résultats sont présentés au chapitre 5 de ma thèse.

Parmi les méthodes de consensus, le consensus majoritaire, qui inclut tous les groupements compatibles, était le plus résolu et le plus similaire aux arbres modèles. Alors que les arbres consensus obtenus par les quatre autres méthodes étaient peu résolus et par le fait même présentaient une congruence aux arbres modèles relativement faible, le consensus majoritaire avec groupements compatibles (MRC) était plus résolu et plus proche des arbres inférés par la super-matrice. La résolution obtenue avec MRC était comparable à la résolution obtenue avec les méthodes de super-arbre. Cette similarité entre le consensus majoritaire compatible et les méthodes de super-arbre a d'ailleurs été observée par Bininda-Emonds (2003) qui a comparé les méthodes de consensus à la méthode de MRP par simulations. De toutes les méthodes de consensus qu'il a étudiées, le consensus majoritaire compatible était le plus résolu et le plus proche de la solution proposée par MRP. Les méthodes de super-arbre étaient sensiblement égales, sauf pour la méthode de MSS dont la performance était nettement inférieure aux autres. Contrairement à ce qui a été suggéré par Lapointe et al. (1999) et Levasseur & Lapointe (2001, 2006), les méthodes de consensus qui tiennent compte des longueurs de branches (c.-à-d., AC et SDM) ne semblent pas être favorisées par rapport aux autres (c.-à-d. MRC, MRP et SFIT), dans cette application particulière des méthodes de super-arbre. Il serait intéressant de vérifier la validité de ces résultats dans un contexte de super-arbre. Des simulations pourraient être effectuées, à partir des séquences mitochondriales complètes, où certains taxons

seraient enlevés pour certains gènes pour représenter une situation où il y a un chevauchement partiel des taxons.

6.2. CONCLUSION

Les simulations présentées dans cette thèse ont été réalisées selon des paramètres précis qui peuvent ne pas englober tous les cas possibles rencontrés par les chercheurs. De plus, les modèles évolutifs utilisés pour simuler les séquences d'ADN ne peuvent pas représenter intégralement la complexité inhérente aux « vraies » séquences d'ADN. Ceci dit, étant donné que le but premier de la phylogénétique est d'estimer un arbre représentant des relations évolutives et des spéciations qui ont eu lieu dans le passé, il n'y a aucun moyen de s'assurer que le bon arbre a été retrouvé. Ainsi, les simulations représentent un cadre idéal pour tester différentes approches puisque la phylogénie de départ (qui est celle utilisée pour simuler les données) est connue. D'ailleurs, de grandes avancées méthodologiques ont été permises en simulant des cas simples, par exemple, la découverte du phénomène de l'attraction des longues branches (Felsenstein 1978). C'est pour ces raisons que mes recommandations sont de poursuivre sur le chemin des simulations et d'inciter les chercheurs à être plus nombreux à s'attarder aux problèmes méthodologiques.

Plus spécifiquement, et en lien avec ma thèse, une avenue de recherche qu'il me semble important d'approfondir est la comparaison entre la méthode du consensus moyen (AC) et la méthode de *super distance matrix* (SDM), puisque mes résultats sont en désaccord avec ceux présentés par Criscuolo et al. (2006). Les résultats présentés au chapitre 5 de ma thèse suggèrent un léger avantage de la méthode AC lorsque comparé à la méthode de SDM, pourtant Criscuolo et al. (2006) prétendent le contraire. Une différence notable entre mon étude et la leur est que mon application a été faite dans un contexte de consensus où l'identité des taxons était identique pour toutes les matrices, alors qu'ils ont plutôt utilisé un contexte de super-arbre, où les taxons se chevauchaient partiellement. Une possibilité serait de reprendre le jeu de données des séquences mitochondriales complètes des familles de mammifères et d'éliminer au hasard les séquences de certains gènes pour pouvoir obtenir des matrices individuelles de gènes où certains taxons ne sont pas présents. Cela permettrait de déterminer laquelle des deux méthodes est la plus optimale lorsque appliquées à des séquences mitochondriales de mammifères et ce, dans un contexte de super-arbre. En lien avec la

façon d'analyser les matrices où une proportion des données est manquante, je considère important de poursuivre les simulations visant à déterminer les conditions optimales d'utilisation des séquences chimères. Bien que deux chapitres (3 et 4) de ma thèse présentent des résultats confirmant la validité de l'utilisation des matrices qui comportent des séquences chimères pour l'inférence phylogénétique, il serait à propos d'étudier leur performance lorsque utilisées pour des matrices de séquences provenant de EST. L'étude récente de Hartmann & Vision (2008) a démontré que la distribution des données manquantes était différente dans les matrices de EST et dans une matrice typique où des séquences de gènes complets sont incluses et que cela engendrait un impact négatif sur l'inférence phylogénétique. Ainsi, une matrice de EST pourrait être ré-analysées en utilisant l'approche des taxons chimères, ce qui pourrait non seulement permettre d'augmenter l'exactitude phylogénétique des arbres inférés mais aussi de diminuer grandement le temps de calcul.

Il est intéressant de noter que, déjà en 1904, Karl Pearson a effectué une méta-analyse pour augmenter la puissance statistique d'une corrélation qu'il jugeait limitée par des échantillons de petites tailles (Pearson 1904). Il peut donc sembler étonnant qu'autant de problèmes méthodologiques et statistiques, en lien avec les analyses de grandes échelles, restent à résoudre. S'il est vrai que certains débats en analyse phylogénétique ont été caractérisés de « stagnants », le domaine de la phylogénétique a évolué plutôt rapidement et favorablement. En effet, les progrès accomplis ont été énormes depuis la découverte de l'ADN (Crick & Watson 1953), du séquençage (Gilbert & Maxam 1973, Sanger et al. 1977) et des premières analyses phylogénétiques (Hennig 1966). Alors qu'il était impensable, il y a quelques années à peine, de pouvoir reconstituer l'Arbre de la Vie, plusieurs chercheurs s'attèlent maintenant à cette tâche (Wolf et al. 2002, Driskell et al. 2004, Delsuc et al. 2005, Ciccarelli et al. 2006, Philippe & Telford 2006, Dunn et al. 2008, Nahum & Pereira 2008). Je suis heureuse d'avoir pu participer, par le biais de cette thèse, au développement de cette discipline en plein essor et j'espère que ma thèse contribuera de façon positive au domaine de la phylogénomique.

BIBLIOGRAPHIE

Adams E. N. 1972. Consensus techniques and the comparison of taxonomic trees. Systematic Zoology. 21: 390-397.

Adams E. N. 1986. N-trees as nestings: Complexity, similarity, and consensus. Journal of Classification. 3: 299-317.

Agnarsson I. & May-Collado L. J. 2008. The phylogeny of Cetartiodactyla: The importance of dense taxon sampling, missing data, and the remarkable promise of cytochrome b to provide reliable species-level phylogenies. Molecular Phylogenetics and Evolution. 48: 964-985.

Albert V. A. 2005. Parsimony and phylogenetics in the genomic age. Pp. 1-11 *in* Parsimony, phylogeny, and genomics (V. A. Albert, ed.). Oxford University Press.

Amrine-Madsen H., Koepfli K. P., Wayne R. K. & Springer M. S. 2003. A new phylogenetic marker, apolipoprotein B, provides compelling evidence for eutherian relationships. Molecular Phylogenetics and Evolution. 28: 225-240.

Anderson S., Bankier A. T., Barrell B. G., de Bruijn M. H. L., Coulson A. R., Drouin J., Eperon I. C., Nierlich D. P., et al. 1981. Sequence and organization of the human mitochondrial genome. Nature. 290: 457-465.

Ané C., Larget B., Baum D. A., Smith S. D. & Rokas A. 2007. Bayesian estimation of concordance among gene trees. Molecular Biology and Evolution. 24: 412-426.

Archibald J. D. & Deutschman D. 2001. Quantitative analysis of the timing of origin of extant placental orders. Journal of Mammalian Evolution. 8: 107-124.

Archibald J. D. 2003. Timing and biogeography of the eutherian radiation: Fossils and molecules compared. Molecular Phylogenetics and Evolution. 28: 350-359.

Arnason U., Adegoke J. A., Bodin K., Born E. W., Esa Y. B., Gullberg A., Nilsson M., Short R. V., et al. 2002. Mammalian mitogenomic relationships and the root of the eutherian tree. Proceedings of the National Academy of Sciences of the United States of America. 99: 8151-8156.

Arnason U. & Janke A. 2002. Mitogenomic analyses of eutherian relationships. Cytogenetic and Genome Research. 96: 20-32.

Arnason U., Gullberg A. & Janke A. 2004. Mitogenomic analyses provide new insights into cetacean origin and evolution. Gene. 333: 27-34.

Arnason U., Gullberg A., Janke A. & Kullberg M. 2007. Mitogenomic analyses of caniform relationships. Molecular Phylogenetics and Evolution. 45: 863-874.

Arnason U., Adegoke J. A., Gullberg A., Harley E. H., Janke A. & Kullberg M. 2008. Mitogenomic relationships of placental mammals and molecular estimates of their divergences. Gene. 421: 37-51.

Asher R. J., Novacek M. J. & Geiser J. H. 2003. Relationships of endemic African mammals and their fossil relatives based on morphological and molecular evidence. Journal of Mammalian Evolution. 10: 131-194.

Asher R. J., Horovitz I. & Sánchez-Villagra M. R. 2004. First combined cladistic analysis of marsupial mammal interrelationships. Molecular Phylogenetics and Evolution. 33: 240-250.

Asher R. J. 2007. A web-database of mammalian morphology and a reanalysis of placental phylogeny. BMC Evolutionary Biology. 7: 108.

Baker M. L., Wares J. P., Harrison G. A. & Miller R. D. 2004. Relationships among the families and orders of marsupials and the major mammalian lineages based on recombination activating gene-1. Journal of Mammalian Evolution. 11: 1-16.

Bapteste E., Susko E., Leigh J., MacLeod D., Charlebois R. L. & Doolittle W. F. 2005. Do orthologous gene phylogenies really support tree-thinking? BMC Evolutionary Biology. 5: 33.

Barker F. K. & Lutzoni F. M. 2002. The utility of the incongruence length difference test. Systematic Biology. 51: 625-637.

Barrett M., Donoghue M. J. & Sober E. 1991. Against Consensus. Systematic Zoology. 40: 486-493.

Baum B. R. 1992. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon. 41: 3-10.

Baum B. R. & Ragan M. A. 2004. The MRP method. Pp. 17-34 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.) Kluwer Academic Publishers, Dordrecht, the Netherlands.

Baurain D., Brinkmann H. & Philippe H. 2007. Lack of resolution in the animal phylogeny: Closely spaced cladogeneses or undetected systematic errors? Molecular Biology and Evolution. 24: 6-9.

Beck R. M. D., Bininda-Emonds O. R. P., Cardillo M., Liu F. G. R. & Purvis A. 2006. A higher-level MRP supertree of placental mammals. BMC Evolutionary Biology. 6: 93.

Beck R. M. D. 2008. A dated phylogeny of marsupials using a molecular supermatrix and multiple fossil constraints. Journal of Mammalogy. 89: 175-189.

Benton M. J. 1988. The relationships of the major group of mammals: New approaches. Trends in Ecology and Evolution. 3: 40-45.

Bergsten J. 2005. A review of long-branch attraction. Cladistics. 21: 163-193.

Bininda-Emonds O. R. P., Gittleman J. L. & Purvis A. 1999. Building large trees by combining phylogenetic information: A complete phylogeny of the extant Carnivora (Mammalia). Biological Reviews. 74: 143-175.

Bininda-Emonds O. R. P. & Sanderson M. J. 2001. Assessment of the accuracy of matrix representation with parsimony analysis supertree construction. Systematic Biology. 50: 565-579.

Bininda-Emonds O. R. P., Gittleman J. L. & Steel M. A. 2002. The (Super)tree of life: Procedures, problems, and prospects. Annual Review of Ecology and Systematics. 33: 265-289.

Bininda-Emonds O.R.P. 2003. MRP supertree construction in the consensus setting. Pp. 231-242 *in* Bioconsensus (Janowitz, M.F., F.-J. Lapointe, F.R. McMorris, B. Mirkin, and F.S. Roberts, eds.). DIMACS: Series in Discrete Mathematics and Theoretical Computer Science, volume 61. American Mathematical Society-DIMACS, Providence, Rhode Island.

Bininda-Emonds O. R. P. 2004a. Trees versus characters and the supertree/supermatrix "paradox". Systematic Biology. 53: 356-359.

Bininda-Emonds O. R. P. 2004b. The evolution of supertrees. Trends in Ecology and Evolution. 19: 315-322.

Bininda-Emonds O. R. P. 2004c. Phylogenetic supertrees: Combining information to reveal the Tree of Life. Kluwer Academic Publishers, Dordrecht, the Netherlands.

Bininda-Emonds O. R. P. 2005. Supertree construction in the genomic age. Molecular Evolution: Producing the Biochemical Data, Part B. 395: 745-757.

Bininda-Emonds O. R. P., Beck R. M. D. & Purvis A. 2005. Getting to the roots of matrix representation. Systematic Biology. 54: 668-672.

Bininda-Emonds O. R. P., Cardillo M., Jones K. E., MacPhee R. D. E., Beck R. M. D., Grenyer R., Price S. A., Vos R. A., et al. 2007. The delayed rise of present-day mammals. Nature. 446: 507-512.

Blanga-Kanfi S., Miranda H., Penn O., Pupko T., DeBry R. W. & Huchon D. 2009. Rodent phylogeny revised: Analysis of six nuclear genes from all major rodent clades. BMC Evolutionary Biology. 9: 71.

Bofkin L. & Goldman N. 2007. Variation in evolutionary processes at different codon positions. Molecular Biology and Evolution. 24: 513-521.

Boore J. L. 2006. The use of genome-level characters for phylogenetic reconstruction. Trends in Ecology and Evolution. 21: 439-446.

Bourlat S. J., Nielsen C., Economou A. D. & Telford M. J. 2008. Testing the new animal phylogeny: A phylum level molecular analysis of the animal kingdom. Molecular Phylogenetics and Evolution. 49: 23-31.

Brand C. J. & Keith L. B. 1979. Lynx demography during a snowshoe hare decline in Alberta. Journal of Wildlife Management. 43: 827-849.

Brochier C., Forterre P. & Gribaldo S. 2005. An emerging phylogenetic core of Archaea: Phylogenies of transcription and translation machineries converge following addition of new genome sequences. BMC Evolutionary Biology. 5: 36.

Bull J. J., Huelsenbeck J. P., Cunningham C. W., Swofford D. L. & Waddell P. J. 1993. Partitioning and combining data in phylogenetic analysis. Systematic Biology. 42: 384-397.

Campbell V. & Lapointe F.-J. In press. The use and validity of composite taxa in phylogenetic analysis. Systematic Biology.

Campbell V., Legendre P., & Lapointe F.-J. 2009. Assessing congruence among ultrametric distance matrices. Journal of Classification. 26: 103-117.

Cannarozzi G., Schneider A. & Gonnet G. 2007. A phylogenomic study of human, dog, and mouse. PLoS Computational Biology. 3: 9-14.

Cao Y., Fujiwara M., Nikaido M., Okada N. & Hasegawa M. 2000. Interordinal relationships and timescale of eutherian evolution as inferred from mitochondrial genome data. Gene. 259: 149-158.

Cardillo M., Bininda-Emonds O. R. P., Boakes E. & Purvis A. 2004. A species-level phylogenetic supertree of marsupials. Journal of Zoology. 264: 11-31.

Cavalli-Sforza L. L. & Edwards A. W. 1967. Phylogenetic analysis: Models and estimation procedures. Evolution. 32: 550-570.

Chang J. T. 1996. Full reconstruction of Markov models on evolutionary trees: Identifiability and consistency. Mathematical Biosciences. 137: 51-73.

Chippindale P. T. & Wiens J. J. 1994. Weighting, partitioning, and combining characters in phylogenetic analysis. Systematic Biology. 43: 278-287.

Chomyn A., Mariottini P., Cleeter M. W. J., Ragan C. I., Matsuno-Yagi A., Hatefi Y., Doolittle R. F. & Attardi G. 1985. Six unidentified reading frames of human mitochondrial DNA encode components of the respiratory-chain NADH dehydrogenase. Nature. 314: 592-597.

Chomyn A., Cleeter M. W. J., Ragan C. I., Riley M., Doolittle R. F. & Attardi G. 1986. URF6, last unidentified reading frame of human mtDNA, codes for an NADH dehydrogenase subunit. Science. 234: 614-618.

Churakov G., Kriegs J. O., Baertsch R., Zemann A., Brosius J. & Schmitz J. 2009. Mosaic retroposon insertion patterns in placental mammals. Genome Research. 19: 868-875.

Ciccarelli F. D., Doerks T., von Mering C., Creevey C. J., Snel B. & Bork P. 2006. Toward automatic reconstruction of a highly resolved tree of life. Science. 311: 1283-1287.

Cohen J. 1988. Statistical power analysis for the behavioral sciences. Second edition. Lawrence Erlbaum Associates, Hillsdale, New Jersey.

Colless D. H. 1980. Congruence between morphometric and allozyme data for Menidia species: A reappraisal. Systematic Zoology. 29: 289-299.

Collins T. M., Fedrigo O. & Naylor G. J. P. 2005. Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenetics. Systematic Biology. 54: 493-500.

Conte M. G., Gaillard S., Lanau N., Rouard M. & Perin C. 2008. GreenPhylDB: A database for plant comparative genomics. Nucleic Acids Research. 36: D991-D998.

Cope E. D. 1889. The Edentata of North America. American Naturalist. 23: 657-664.

Corneli P. S. 2002. Complete mitochondrial genomes and eutherian evolution. Journal of Mammalian Evolution. 9: 281-305.

Cotton J. A. & Wilkinson M. 2007. Majority-rule supertrees. Systematic Biology. 56: 445-452.

Cotton J. A. & Wilkinson M. 2009. Supertrees join the mainstream of phylogenetics. Trends in Ecology and Evolution. 24: 1-3.

Crandall K. A. & Buhay J. E. 2004. Genomoc databases and the tree of life. Science. 306: 1144-1145.

Creevey C. J., Fitzpatrick D. A., Philip G. K., Kinsella R. J., O'Connell M. J., Pentony M. M., Travers S. A., Wilkinson M., et al. 2004. Does a tree-like phylogeny only exist at the tips in the prokaryotes? Proceedings of the Royal Society B. 271: 2551-2558.

Creevey C. J. & McInerney J. O. 2005. Clann: Investigating phylogenetic information through supertree analyses. Bioinformatics. 21: 390-392.

Crick F. & Watson J. 1953. A structure for deoxyribose nucleic acid. Nature. 171: 737.

Criscuolo A., Berry V., Douzery E. J. P. & Gascuel O. 2006. SDM: A fast distance-based approach for (super) tree building in phylogenomics. Systematic Biology. 55: 740-755.

Cunningham C. W. 1997a. Is congruence between data partitions a reliable predictor of phylogenetic accuracy? Empirically testing an iterative procedure for choosing among phylogenetic methods. Systematic Biology. 46: 464-478.

Cunningham C. W. 1997b. Can three incongruence tests predict when data should be combined? Molecular Biology and Evolution. 14: 733-740.

Curole J. P. & Kocher T. D. 1999. Mitogenomics: Digging deeper with complete mitochondrial genomes. Trends in Ecology and Evolution. 14: 394-398.

D'Erchia A. M., Gissi C., Pesole G., Saccone C. & Arnason U. 1996. The guinea-pig is not a rodent. Nature. 381: 597-600.

Darlu P. & Tassy P. 1993. La reconstruction phylogénétique. Concepts et méthodes. Collection Biologie Théorique, Masson, Paris.

Darlu P. & Lecointre G. 2002. When does the incongruence length difference test fail? Molecular Biology and Evolution. 19: 432-437.

Darwin C. 1859. The origin of species. Mentor, New York.

Davis C. S., Delisle I., Stirling I., Siniff D. B. & Strobeck C. 2004. A phylogeny of the extant Phocidae inferred from complete mitochondrial DNA coding regions. Molecular Phylogenetics and Evolution. 33: 363-377.

de Pinna M. C. C. 1991. Concepts and tests of homology in the cladistic paradigm. Cladistics. 7: 367-394.

de Queiroz A. 1993. For consensus (sometimes). Systematic Biology. 42: 368-372.

de Queiroz A., Donoghue M. J. & Kim J. 1995. Separate versus combined analysis of phylogenetic evidence. Annual Review of Ecology and Systematics. 26: 657-681.

de Queiroz A. & Gatesy J. 2007. The supermatrix approach to systematics. Trends in Ecology and Evolution. 22: 34-41.

DeBry R. W. 2005. The systematic component of phylogenetic error as a function of taxonomic sampling under parsimony. Systematic Biology. 54: 432-440.

Degnan J. H. & Rosenberg N. A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genetics. 2: 762-768.

Delisle I. & Strobeck C. 2005. A phylogeny of the Caniformia (order Carnivora) based on 12 complete protein-coding mitochondrial genes. Molecular Phylogenetics and Evolution. 37: 192-201.

Delsuc F., Scally M., Madsen O., Stanhope M. J., de Jong W. W., Catzeflis F. M., Springer M. S. & Douzery E. J. P. 2002. Molecular phylogeny of living xenarthrans and the impact of character and taxon sampling on the placental tree rooting. Molecular Biology and Evolution. 19: 1656-1671.

Delsuc F., Brinkmann H. & Philippe H. 2005. Phylogenomics and the reconstruction of the tree of life. Nature Reviews Genetics. 6: 361-375.

Delsuc F., Brinkmann H., Chourrout D. & Philippe H. 2006. Tunicates and not cephalochordates are the closest living relatives of vertebrates. Nature. 439: 965-968.

Desper R. & Gascuel O. 2006. Getting a tree fast: Neighbor Joining, FastME, and distance-based methods. Pp. 6.3.1-6.3.28 *in* Current Protocols in Bioinformatics (A. D. Baxevanis et al., eds.). John Wiley & sons, New York.

Diallo A. B., Lapointe F.-J. & Makarenkov V. 2006. A new effective method for estimating missing values in the sequence data prior to phylogenetic analysis. Evolutionary Bioinformatics. 2: 127-135.

Disotell T. 2008. Primate phylogenetics. Encyclopedia of Life Science. doi: 10.1002/9780470015902.a0005833.pub2.

Dolphin K., Belshaw R., Orme C. D. L. & Quicke D. L. J. 2000. Noise and incongruence: Interpreting results of the incongruence length difference test. Molecular Phylogenetics and Evolution. 17: 401-406.

Doolittle W. F. 1999. Phylogenetic classification and the universal tree. Science. 284: 2124-2128.

Dopazo H., Santoyo J. & Dopazo J. 2004. Phylogenomics and the number of characters required for obtaining an accurate phylogeny of eukaryote model species. Bioinformatics. 20 (Suppl. 1): i116-i121.

Dopazo H. & Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. Genome Biology. 6: R41.

Douady C. J., Chatelier P. I., Madsen O., de Jong W. W., Catzeflis F., Springer M. S. & Stanhope M. J. 2002. Molecular phylogenetic evidence confirming the Eulipotyphla concept and in support of hedgehogs as the sister group to shrews. Molecular Phylogenetics and Evolution. 25: 200-209.

Douzery E. J. P., Delsuc F., Stanhope M. J. & Huchon D. 2003. Local molecular clocks in three nuclear genes: Divergence times for rodents and other mammals and incompatibility among fossil calibrations. Journal of Molecular Evolution. 57: S201-S213.

Doyle J. A., & Donoghue M. J. 1987. The importance of fossils in elucidating seed plant phylogeny and macroevolution. Review of Paleobotany and Palynology. 50: 63-95.

Doyle J. J. 1992. Gene trees and species trees: Molecular systematics as one-character taxonomy. Systematic Botany. 17: 144-163.

Doyle J. J. & Davis J. I. 1998. Homology in molecular phylogenetics: A parsimony perspective. Pp. 102-163 *in* Molecular systematics of plants II DNA sequencing (P. S. Soltis, D. E. Soltis, & J. J. Doyle, eds.). Kluwer Academic Publisher, Dordrecht, The Netherlands.

Driskell A. C., Ané C., Burleigh J. G., McMahon M. M., O'Meara B. C. & Sanderson M. J. 2004. Prospects for building the tree of life from large sequence databases. Science. 306: 1172-1174.

Dubey S., Salamin N., Ohdachi S. D., Barriere P. & Vogel P. 2007. Molecular phylogenetics of shrews (Mammalia: Soricidae) reveal timing of transcontinental colonizations. Molecular Phylogenetics and Evolution. 44: 126-137.

Dunn C. W., Hejnol A., Matus D. Q., Pang K., Browne W. E., Smith S. A., Seaver E., Rouse G. W., et al. 2008. Broad phylogenomic sampling improves resolution of the animal tree of life. Nature. 452: 745-749.

Duvall M. R., Robinson J. W., Mattson J. G. & Moore A. 2008. Phylogenetic analyses of two mitochondrial metabolic genes sampled in parallel from angiosperms find fundamental interlocus incongruence. American Journal of Botany. 95: 871-884.

Edgington E. S. 1995. Randomization tests. Third edition. Marcel Dekker, New York.

Edwards A. W. F. & Cavalli-Sforza L. L. 1963. The reconstruction of evolution. Annals of Human Genetics. 27: 105-106.

Edwards S. V., Fertil B., Giron A. & Deschavanne P. J. 2002. A genomic schism in birds revealed by phylogenetic analysis of DNA strings. Systematic Biology. 51: 599-613.

Edwards S. V., Liu L. & Pearl D. K. 2007. High-resolution species trees without concatenation. Proceedings of the National Academy of Sciences of the United States of America. 104: 5936-5941.

Eernisse D. J. & Kluge A. G. 1993. Taxonomic congruence versus total evidence, and amniote phylogeny inferred from fossils, molecules, and morphology. Molecular Biology and Evolution. 10: 1170-1195.

Eisen J. A. 1998. Phylogenomics: Improving functional predictions for uncharacterized genes by evolutionary analysis. Genome Research. 8: 163-167.

Eisen J. A. & Fraser C. M. 2003. Phylogenomics: Intersection of evolution and genomics. Science. 300: 1706-1707.

Estabrook G. F., McMorris F. R. & Meacham C. A. 1985. Comparison of undirected phylogenetic trees based on subtrees of four evolutionary units. Systematic Zoology. 34: 193-200.

Eulenstein O., Chen D. H., Burleigh J. G., Fernandez-Baca D. & Sanderson M. J. 2004. Performance of flip supertree construction with a heuristic algorithm. Systematic Biology. 53: 299-308.

Faith D. P. 1991. Cladistic permutation tests for monophyly and nonmonophyly. Systematic Zoology. 40: 366-375.

Farris J. S., Kluge A. G. & Eckhardt M. J. 1970. A numerical approach to phylogenetic systematics. Systematic Zoology. 19: 172-189.

Farris J. S., Källersjö M., Kluge A. G. & Bult C. 1994. Testing significance of incongruence. Cladistics. 10: 315-319.

Farris J. S., Källersjö M., Kluge A. G. & Bult C. 1995. Constructing a significance test for incongruence. Systematic Biology. 44: 570-572.

Felsenstein J. 1973. Maximum-likelihood estimation of evolutionary trees from continuous charaters. American Journal of Human Genetics. 25: 471-492.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Systematic Zoology. 27: 401-410.

Felsenstein J. 1981. Evolutionary trees from DNA sequences: A maximum likelihood approach. Journal of Molecular Evolution. 17: 368-376.

Felsenstein J. 1985. Confidence limits on phylogenies: An approach using the Bootstrap. Evolution. 39: 783-791.

Felsenstein J. 1989. PHYLIP: Phylogeny inference package (version 3.2). Cladistics. 5: 164-166.

Felsenstein J. 2004. Inferring phylogenies. Sinauer Associates, Sunderland, Massachusetts.

Fiala K. L. & Sokal R. R. 1985. Factors determining the accuracy of cladogram estimation: Evaluation using computer simulation. Evolution. 39: 609-622.

Finden C. R. & Gordon A. D. 1985. Obtaining common pruned trees. Journal of Classification. 2: 255-276.

Fitch W. M. 1970. Distinguishing homologous from analogous proteins. Systematic Zoology. 19: 99-106.

Fitzpatrick D. A., Logue M. E., Stajich J. E. & Butler G. 2006. A fungal phylogeny based on 42 complete genomes derived from supertree and combined gene analysis. BMC Evolutionary Biology. 6: 1-15.

Flynn J. J., Finarelli J. A., Zehr S., Hsu J. & Nedbal M. A. 2005. Molecular phylogeny of the Carnivora (Mammalia): Assessing the impact of increased sampling on resolving enigmatic relationships. Systematic Biology. 54: 317-337.

Fulton T. L. & Strobeck C. 2006. Molecular phylogeny of the Arctoidea (Carnivora): Effect of missing data on supertree and supermatrix analyses of multiple gene data sets. Molecular Phylogenetics and Evolution. 41: 165-181.

Gadagkar S. R., Rosenberg M. S. & Kumar S. 2005. Inferring species phylogenies from multiple genes: Concatenated sequence tree versus consensus gene tree. Journal of Experimental Zoology Part B-Molecular and Developmental Evolution. 304B: 64-74.

Galtier N. & Daubin V. 2008. Dealing with incongruence in phylogenomic analyses. Philosophical Transactions of the Royal Society B. 363: 4023-4029.

Gatesy J., Matthee C., DeSalle R. & Hayashi C. 2002. Resolution of a supertree/supermatrix paradox. Systematic Biology. 51: 652-664.

Gatesy J., Baker R. H. & Hayashi C. 2004. Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of Crocodylia. Systematic Biology. 53: 342-355.

Gatesy J. & Springer M. S. 2004. A critique of matrix representation with parsimony supertrees. Pp. 369-388 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Edmonds, ed.). Kluwer Academic Publishers, Dordrecht, The Netherlands.

Gatesy J., DeSalle R. & Wahlberg N. 2007. How many genes should a systematist sample? Conflicting insights from a phylogenomic matrix characterized by replicated incongruence. Systematic Biology. 56: 355-363.

Gauthier J., Kluge A. G. & Rowe T. 1988. Amniote phylogeny and the importance of fossils. Cladistics. 4: 105-208.

Gibson A., Gowri-Shankar V., Higgs P. G. & Rattray M. 2005. A comprehensive analysis of mammalian mitochondrial genome base composition and improved phylogenetic methods. Molecular Biology and Evolution. 22: 251-264.

Gilbert W. & Maxam A. 1973. The nucleotide sequence of the *lac* operator. Proceedings of the National Academy of Sciences of the United States of America. 70: 3581-3584.

Gill T. 1872. Arrangement of the families of mammals with analytical tables. Smithsonian Miscellaneous Collections. 11: 1-98.

Glanville J. G., Kirshner D., Krishnamurthy N. & Sjolander K. 2007. Berkeley Phylogenomics Group web servers: Resources for structural phylogenomic analysis. Nucleic Acids Research. 35: W27-W32.

Goddard W., Kubicka E., Kubicki G. & Mcmorris F. R. 1994. The agreement metric for labeled binary trees. Mathematical Biosciences. 123: 215-226.

Goloboff P. A. 2003. Parsimony, likelihood, and simplicity. Cladistics. 19: 91-103.

Gordon A. D. 1980. On the assessment and comparison of classifications. Pp. 149-160 *in* Analyse de données et informatique (R. Tomassone, ed.). I.N.R.I.A., Le Chesnay, France.

Graur D. 1993a. Towards a molecular resolution of the ordinal phylogeny of the eutherian mammals. FEBS Letters. 325: 152-159.

Graur D. 1993b. Molecular phylogeny and the higher classification of eutherian mammals. Trends in Ecology and Evolution. 8: 141-147.

Graybeal A. 1994. Evaluating the phylogenetic utility of genes: A search for genes informative about deep divergences among vertebrates. Systematic Biology. 43: 174-193.

Graybeal A. 1998. Is it better to add taxa or characters to a difficult phylogenetic problem? Systematic Biology. 47: 9-17.

Gregory W. K. 1910. The orders of mammals. Bulletin of the American Museum of Natural History. 27: 1-524.

Gregory W. K. 1947. The monotremes and the palimpsest theory. Bulletin of the American Museum of Natural History. 88: 1-52.

Gribaldo S. & Philippe H. 2002. Ancient phylogenetic relationships. Theoretical Population Biology. 61: 391-408.

Gu X. M., He S. Y. & Lei A. 2008. Molecular phylogenetics among three families of bats (Chiroptera : Rhinolophidae, Hipposideridae, and Vespertilionidae) based on partial sequences of the mitochondrial 12S and 16S rRNA genes. Zoological Studies. 47: 368-378.

Guindon S. & Gascuel O. 2003. A simple, fast, and accurate algorithm to estimate large phylogenies by maximum likelihood. Systematic Biology. 52: 696-704.

Hackett S. J., Kimball R. T., Reddy S., Bowie R. C. K., Braun E. L., Braun M. J., Chojnowski J. L., Cox W. A., et al. 2008. A phylogenomic study of birds reveals their evolutionary history. Science. 320: 1763-1768.

Hallström B. M., Kullberg M., Nilsson M. A. & Janke A. 2007. Phylogenomic data analyses provide evidence that Xenarthra and Afrotheria are sister groups. Molecular Biology and Evolution. 24: 2059-2068.

Hallström B. M. & Janke A. 2008. Resolution among major placental mammal interordinal relationships with genome data imply that speciation influenced their earliest radiations. BMC Evolutionary Biology. 8: 162.

Hallström B. M. & Janke A. 2009. Gnathostome phylogenomics utilizing lungfish EST sequences. Molecular Biology and Evolution. 26: 463-471.

Hartmann S. & Vision T. J. 2008. Using ESTs for phylogenomics: Can one accurately infer a phylogenetic tree from a gappy alignment? BMC Evolutionary Biology. 8: 95.

Heath T. A., Hedtke S. M. & Hillis D. M. 2008. Taxon sampling and the accuracy of phylogenetic analyses. Journal of Systematics and Evolution. 46: 239-257.

Hedges S. B., Parker P. H., Sibley C.G. & Kumar S. 1996. Continental breakup and the ordinal diversification of birds and mammals. Nature. 381: 226-229.

Hedtke S. M., Townsend T. M. & Hillis D. M. 2006. Resolution of phylogenetic conflict in large data sets by increased taxon sampling. Systematic Biology. 55: 522-529.

Hendy M. D. & Penny D. 1989. A framework for the quantitative study of evolutionary trees. Systematic Zoology. 38: 297-309.

Hennig W. 1966. Phylogenetic systematics [Translated by Davis D.D. & Zangerl R. from Hennig W. 1950. Grundzüge einer Theorie der Phylogenetischen Systematik. Deutscher Zentralverlag, Berlin.] University Illinois Press, Urbana.

Herke S. W., Xing J., Ray D. A., Zimmerman J. W., Cordaux R. & Batzer M. A. 2007. A SINE-based dichotomous key for primate identification. Gene. 390: 39-51.

Higdon J. W., Bininda-Emonds O. R. P., Beck R. M. D. & Ferguson S. H. 2007. Phylogeny and divergence of the pinnipeds (Carnivora: Mammalia) assessed using a multigene dataset. BMC Evolutionary Biology. 7: 216.

Higgins D. G. & Sharp P. M. 1988. CLUSTAL: A package for performing multiple sequence alignment on a microcomputer. Gene. 73: 237-244.

Hillis D. M., Bull J. J., White M. E., Badgett M. R., & Molineux I. J. 1992. Experimental phylogenetics: Generation of a known phylogeny. Science. 255: 589-592.

Hillis D. M., Allard M. W., & Miyamoto M. M. 1993. Analysis of DNA sequence data: Phylogenetic inference. Methods in Enzymology. 224: 456-487.

Hillis D. M., Huelsenbeck J. P. & Swofford D. L. 1994. Hobgoblin of phylogenetics. Nature. 369: 363-364.

Hillis D. M. 1995. Approaches for assessing phylogenetic accuracy. Systematic Biology. 44: 3-16.

Hillis D. M. 1996. Inferring complex phylogenies. Nature. 383: 130-131.

Hillis D. M., Moritz C. & Mable B. K. 1996. Molecular systematics. Second edition. Sinauer Associates, Sunderland, Massachusetts.

Hillis D. M. 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Systematic Biology. 47: 3-8.

Hillis D. M. 1999. SINEs of the perfect character. Proceedings of the National Academy of Sciences of the United States of America. 96: 9979-9981.

Hillis D. M., Pollock D. D., McGuire J. A. & Zwickl D. J. 2003. Is sparse taxon sampling a problem for phylogenetic inference? Systematic Biology. 52: 124-126.

Holm S. 1979. A simple sequentially rejective multiple test procedure. Scandinavian Journal of Statistics. 6: 65-70.

Horner D. S., Lefkimmiatis K., Reyes A., Gissi C., Saccone C. & Pesole G. 2007. Phylogenetic analyses of complete mitochondrial genome sequences suggest a basal divergence of the enigmatic rodent Anomalurus. BMC Evolutionary Biology. 7: 16.

Horovitz I. & Sánchez-Villagra M. R. 2003. A morphological analysis of marsupial mammal higher-level phylogenetic relationships. Cladistics. 19: 181-212.

Horvath J. E., Weisrock D. W., Embry S. L., Fiorentino I., Balhoff J. P., Kappeler P., Wray G. A., Willard H. F., et al. 2008. Development and application of a phylogenomic toolkit: Resolving the evolutionary history of Madagascar's lemurs. Genome Research. 18: 489-499.

Huchon D., Madsen O., Sibbald M. J. J. B., Ament K., Stanhope M. J., Catzeflis F., de Jong W. W. & Douzery E. J. P. 2002. Rodent phylogeny and a timescale for the evolution of glires: Evidence from an extensive taxon sampling using three nuclear genes. Molecular Biology and Evolution. 19: 1053-1065.

Hudelot C., Gowri-Shankar V., Jow H., Rattray M. & Higgs P. G. 2003. RNA-based phylogenetic methods: Application to mammalian mitochondrial RNA sequences. Molecular Phylogenetics and Evolution. 28: 241-252.

Huelsenbeck J. P. 1991. When are fossils better than extant taxa in phylogenetic analysis. Systematic Zoology. 40: 458-469.

Huelsenbeck J. P. & Hillis D. M. 1993. Success of phylogenetic methods in the four-taxon case. Systematic Biology. 42: 247-264.

Huelsenbeck J. P. 1995. Performance of phylogenetic methods in simulation. Systematic Biology. 44: 17-48.

Huelsenbeck J. P. & Bull J. J. 1996. A likelihood ratio test to detect conflicting phylogenetic signal. Systematic Biology. 45: 92-98.

Huelsenbeck J. P., Bull J. J. & Cunningham C. W. 1996a. Combining data in phylogenetic analysis: Reply. Trends in Ecology and Evolution. 11: 335.

Huelsenbeck J. P., Bull J. J. & Cunningham C. W. 1996b. Combining data in phylogenetic analysis. Trends in Ecology and Evolution. 11: 152-158.

Huelsenbeck J. P. & Ronquist F. 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics. 17: 754-755.

Huelsenbeck J. P., Ronquist F., Nielsen R. & Bollback J. P. 2001. Bayesian inference of phylogeny and its impact on evolutionary biology. Science. 294: 2310-2314.

Hughes A. L. & Friedman R. 2007. The effect of branch lengths on phylogeny: An empirical study using highly conserved orthologs from mammalian genomes. Molecular Phylogenetics and Evolution. 45: 81-88.

Hunter J. P. & Janis C. M. 2006. "Garden of Eden" or "Fool's Paradise"? Phylogeny, dispersal, and the southern continent hypothesis of placental mammal origins. Paleobiology. 32: 339-344.

Hutcheon J. M. & Kirsch J. A. W. 2006. A moveable face: Deconstructing the Microchiroptera and a new classification of extant bats. Acta Chiropterologica. 8: 1-10.

Huttley G. A., Wakefield M. J. & Easteal S. 2007. Rates of genome evolution and branching order from whole genome analysis. Molecular Biology and Evolution. 24: 1722-1730.

Ihaka R. & Gentleman R. 1996. R: A language for data analysis and graphics. Journal of Computational and Graphical Statistics. 5: 299-314.

Irwin D. M. & Arnason U. 1994. Cytochrome b gene of marine mammals: Phylogeny and evolution. Journal of Mammalian Evolution. 2: 37-55.

Janecka J. E., Miller W., Pringle T. H., Wiens F., Zitzmann A., Helgen K. M., Springer M. S. & Murphy W. J. 2007. Molecular and genomic data identify the closest living relative of primates. Science. 318: 792-794.

Janke A., Gemmell N. J., Feldmaier-Fuchs G., von Haeseler A. & Pääbo S. 1996. The mitochondrial genome of a monotreme: The platypus (*Ornithorhynchus anatinus*). Journal of Molecular Evolution. 42: 153-159.

Janke A., Xu X. & Arnason U. 1997. The complete mitochondrial genome of the wallaroo (*Macropus robustus*) and the phylogenetic relationship among Monotremata, Marsupialia, and Eutheria. Proceedings of the National Academy of Sciences of the United States of America. 94: 1276-1281.

Janke A., Magnell O., Wieczorek G., Westerman M. & Arnason U. 2002. Phylogenetic analysis of 18S rRNA and the mitochondrial genomes of the wombat, *Vombatus ursinus*, and the spiny anteater, *Tachyglossus aculeatus*: Increased support for the Marsupionta hypothesis. Journal of Molecular Evolution. 54: 71-80.

Jeffroy O., Brinkmann H., Delsuc F. & Philippe H. 2006. Phylogenomics: The beginning of incongruence? Trends in Genetics. 22: 225-231.

Ji Q., Luo Z. X., Yuan C. X. & Tabrum A. R. 2006. A swimming mammaliaform from the Middle Jurassic and ecomorphological diversification of early mammals. Science. 311: 1123-1127.

Johnson L. A. & Soltis D. E. 1998. Assessing congruence: Empirical examples from molecular data. Pp. 297-348 *in* Molecular systematics of plants II DNA sequencing (P. S. Soltis, D. E. Soltis, & J. J. Doyle, eds.). Kluwer Academic Publisher, Dordrecht, The Netherlands.

Jones T. R., Kluge A. G. & Wolf A. J. 1993. When theories and methodologies clash: A phylogenetic reanalysis of the North American Ambystomatid salamanders (Caudata: Ambystomatidae). Systematic Biology. 42: 92-101.

Jow H., Hudelot C., Rattray M. & Higgs P. G. 2002. Bayesian phylogenetics using an RNA substitution model applied to early mammalian evolution. Molecular Biology and Evolution. 19: 1591-1601.

Jukes T. H. & Cantor C. R. 1969. Evolution of protein molecules. Pp. 21-132 *in* Mammalian protein metabolism (H. N. Munro, ed.). Academic Press, New York.

Kearney M. 2002. Fragmentary taxa, missing data, and ambiguity: Mistaken assumptions and conclusions. Systematic Biology. 51: 369-381.

Kellogg M. E., Burkett S., Dennis T. R., Stone G., Gray B. A., McGuire P. M., Zori R. T. & Stanyon R. 2007. Chromosome painting in the manatee supports Afrotheria and Paenungulata. BMC Evolutionary Biology. 7: 6.

Kendall M. G. 1955. Rank correlation methods. Hafner Publishing Co., New-York.

Kim J. 1996. General inconsistency conditions for maximum parsimony: Effects of branch lengths and increasing numbers of taxa. Systematic Biology. 45: 363-374.

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

Kirsch J. A. W., Dickerman A. W., Reig O. A. & Springer M. S. 1991. DNA hybridization evidence for the Australasian affinity of the American marsupial *Dromiciops*

australis. Proceedings of the National Academy of Sciences of the United States of America. 88: 10465-10469.

Kirsch J. A. W., Lapointe F.-J. & Springer M. 1997. DNA-hybridisation studies of marsupials and their implications for metatherian classification. Australian Journal of Zoology. 45: 211-280.

Kjer K. M. & Honeycutt R. L. 2007. Site specific rates of mitochondrial genomes and the phylogeny of eutheria. BMC Evolutionary Biology. 7: 8.

Kluge A. G. 1989. A concern for evidence and a phylogenetic hypothesis of relationships among *Epicrates* (Boidae, Serpentes). Systematic Zoology. 38: 7-25.

Kluge A. G. & Wolf A. J. 1993. Cladistics: What's in a word? Cladistics. 9: 183-199.

Korbel J. O., Snel B., Huynen M. A. & Bork P. 2002. SHOT: A web server for the construction of genome phylogenies. Trends in Genetics. 18: 158-162.

Krause J., Unger T., Nocon A., Malaspinas A. S., Kolokotronis S. O., Stiller M., Soibelzon L., Spriggs H., et al. 2008. Mitochondrial genomes reveal an explosive radiation of extinct and extant bears near the Miocene-Pliocene boundary. BMC Evolutionary Biology. 8: 220.

Kriegs J. O., Churakov G., Kiefmann M., Jordan U., Brosius J. & Schmitz J. 2006. Retroposed elements as archives for the evolutionary history of placental mammals. PLoS Biology. 4: 537-544.

Kubatko L. S. & Degnan J. H. 2007. Inconsistency of phylogenetic estimates from concatenated data under coalescence. Systematic Biology. 56: 17-24.

Kullberg M., Nilsson M. A., Arnason U., Harley E. H. & Janke A. 2006. Housekeeping genes for phylogenetic analysis of eutherian relationships. Molecular Biology and Evolution. 23: 1493-1503.

Kullberg M., Hallstrom B. M., Arnason U. & Janke A. 2008. Phylogenetic analysis of 1.5 Mbp and platypus EST data refute the Marsupionta hypothesis and unequivocally support Monotremata as sister group to Marsupialia/Placentalia. Zoologica Scripta. 37: 115-127.

Kumar S., Tamura K. & Nei M. 1993. MEGA: Molecular evolutionary genetics analysis, version 1.0. Pennsylvania State University.

Kuo C. H., Wares J. P. & Kissinger J. C. 2008. The apicomplexan whole-genome phylogeny: An analysis of incongruence among gene trees. Molecular Biology and Evolution. 25: 2689-2698.

Lanave C., Preparata G., Saccone C. & Serio G.1984. A new method for calculating evolutionary substitution rates. Journal of Molecular Evolution. 20: 86-93.

Lapointe F.-J. & Legendre P. 1990. A statistical framework to test the consensus of two nested classifications. Systematic Zoology. 39: 1-13.

Lapointe F.-J. & Legendre P. 1991. The generation of random ultrametric matrices representing dendrograms. Journal of Classification. 8: 177-200.

Lapointe F.-J. & Legendre P. 1992. A statistical framework to test the consensus among additive trees (cladograms). Systematic Biology. 41: 158-171.

Lapointe F.-J. & Legendre P. 1995. Comparison tests for dendrograms: A comparative evaluation. Journal of Classification. 12: 265-282.

Lapointe F.-J. & Cucumel G. 1997. The average consensus procedure: Combination of weighted trees containing identical or overlapping sets of taxa. Systematic Biology. 46: 306-312.

Lapointe F.-J. 1998a. How to validate phylogenetic trees? A stepwise procedure. Pp. 71-88 *in* Data science, classification, and related methods: Studies in classification, data analysis, and knowledge optimization (C. Hayashi, H.-H. Bock, K. Yajima, Y. Tanaka, & Y. Baba, eds.). Springer-Verlag, Tokyo.

Lapointe F.-J. 1998b. For consensus (with branch lengths). Pp. 73-80 *in* Advances in data science and classification (A. Rizzi, M. Vichi, & H.-H. Bock, eds.). Springer-Verlag, Berlin.

Lapointe F.-J., Kirsch J. A. W. & Hutcheon J. M. 1999. Total evidence, consensus, and bat phylogeny: A distance-based approach. Molecular Phylogenetics and Evolution. 11: 55-66.

Larson A. 1994. The comparison of morphological and molecular data in phylogenetic systematics. Pp. 371-390 *in* Molecular ecology and evolution: Approaches and applications (B. Schierwater, B. Streit, G. P. Wagner, & R. DeSalle, eds.). Birkhauser Verlag, Basel.

Lecointre G., Philippe H., Vân Lê H. L. & Le Guyader H. 1993. Species sampling has a major impact on phylogenetic inference. Molecular Phylogenetics and Evolution. 2: 205-224.

Lecointre G. & Le Guyader H. 2006. The tree of life, a phylogenetic classification. Belknap Press of Harvard University Press, Cambridge, Massachusetts.

Leebens-Mack J., Raubeson L. A., Cui L. Y., Kuehl J. V., Fourcade M. H., Chumley T. W., Boore J. L., Jansen R. K., et al. 2005. Identifying the basal angiosperm node in chloroplast genome phylogenies: Sampling one's way out of the Felsenstein zone. Molecular Biology and Evolution. 22: 1948-1963.

Legendre P. & Lapointe F.-J. 2004. Assessing congruence among distance matrices: Single-malt Scotch whiskies revisited. Australian and New Zealand Journal of Statistics. 46: 615-629.

Legendre, P., and F.-J. Lapointe. 2005. Congruence entre matrices de distance. Pp. 178-181 *in* Comptes-rendus des 12èmes rencontres de la Société Francophone de Classification (Makarenkov, V., G. Cucumel et F.-J. Lapointe, eds.). Université du Québec, Montréal.

Leigh J. W., Susko E., Baumgartner M. & Roger A. J. 2008. Testing congruence in phylogenomic analysis. Systematic Biology. 57: 104-115.

Lerat E., Daubin V. & Moran N. A. 2003. From gene trees to organismal phylogeny in prokaryotes: The case of the γ-proteobacteria. PLoS Biology. 1: 101-109.

Levasseur C. & Lapointe F.-J. 2001. War and peace in phylogenetics: A rejoinder on total evidence and consensus. Systematic Biology. 50: 881-891.

Levasseur C. & Lapointe F.-J. 2006. Total evidence, average consensus and matrix representation with parsimony: What a difference distances make. Evolutionary Bioinformatics. 2: 1-5.

Lin Y. H., McLenachan P. A., Gore A. R., Phillips M. J., Ota R., Hendy M. D. & Penny D. 2002a. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Molecular Biology and Evolution. 19: 2060-2070.

Lin Y. H., Waddell P. J. & Penny D. 2002b. Pika and vole mitochondrial genomes increase support for both rodent monophyly and glires. Gene. 294: 119-129.

Lio P. & Goldman N. 1998. Models of molecular evolution and phylogeny. Genome Research. 8: 1233-1244.

Liu F.-G. R., Miyamoto M. M., Freire N. P., Ong P. Q., Tennant M. R., Young T. S. & Gugel K. F. 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science. 291: 1786-1789.

Lockhart P. J., Larkum A. W. D., Steel M. A., Waddell P. J. & Penny D. 1996. Evolution of chlorophyll and bacteriochlorophyll: The problem of invariant sites in sequence analysis. Proceedings of the National Academy of Sciences of the United States of America. 93: 1930-1934.

Luo C. C., Li W. H. & Chan L. 1989. Structure and expression of dog apolipoprotein A-I, E, C-I mRNAs: Implications for the evolution and functional constraints of apolipoprotein structure. Journal of Lipid Research. 30: 1735-1746.

Luo Z. X. & Wible J. R. 2005. A Late Jurassic digging mammal and early mammalian diversification. Science. 308: 103-107.

Maddison W. 1989. Reconstructing character evolution on polytomous cladograms. Cladistics. 5: 365-377.

Maddison W. P., Donoghue M. J. & Maddison D. R. 1984. Outgroup analysis and parsimony. Systematic Zoology. 33: 83-103.

Madsen O., Scally M., Douady C. J., Kao D. J., DeBry R. W., Adkins R., Amrine H. M., Stanhope M. J., et al. 2001. Parallel adaptive radiations in two major clades of placental mammals. Nature. 409: 610-614.

Malia M. J. J., Lipscomb D. L. & Allard M. W. 2003. The misleading effects of composite taxa in supermatrices. Molecular Phylogenetics and Evolution. 27: 522-527.

Mantel N. 1967. The detection of disease clustering and a generalized regression approach. Cancer Research. 27: 209-220.

Marek, P. E., and J. E. Bond. 2006. Phylogenetic systematics of the colorful, cyanide-producing millipedes of Appalachia (Polydesmida, Xystodesmidae, Apheloriini) using a total evidence Bayesian approach. Molecular Phylogenetics and Evolution. 41: 704-729.

Margush T. & McMorris F. R. 1981. Consensus n-trees. Bulletin of Mathematical Biology. 43: 239-244.

Martin R. D. 2008. Colugos: Obscure mammals glide into the evolutionary limelight. Journal of Biology. 7: 13.

Matsui A., Rakotondraparany F., Munechika I., Hasegawa M. & Horai S. 2009. Molecular phylogeny and evolution of prosimians based on complete sequences of mitochondrial DNAs. Gene. 441: 53-66.

May-Collado L. & Agnarsson I. 2006. Cytochrome b and Bayesian inference of whale phylogeny. Molecular Phylogenetics and Evolution. 38: 344-354.

McKenna, M. C., Bell, S. K., Simpson, G. G., Nichols, R. H., Tedford, R. H., Koopman, K. F., Musser, G. G., Neff, N. A., Shoshani, J., & McKenna, D. M. 1997. Classification of mammals above the species level. Columbia University Press, New York.

Meredith R., Westerman M., Case J. & Springer M. 2006. A phylogeny and timescale for marsupial evolution. Journal of Vertebrate Paleontology. 26: 99A.

Meredith R., Krajewski C., Westerman M. & Springer M. 2009a. Relationships and divergence times among the orders and families of marsupials. Museum of Northern Arizona Bulletin (In press).

Meredith R., Westerman M. & Springer M. 2009b. A phylogeny of Diprotodontia (Marsupialia) based on five nuclear genes. Molecular Phylogenetics and Evolution. 51: 554-571.

Meredith R. W., Westerman M., Case J. A. & Springer M. S. 2008a. A Phylogeny and timescale for marsupial evolution based on sequences for five nuclear genes. Journal of Mammalian Evolution. 15: 1-36.

Meredith R. W., Westerman M. & Springer M. S. 2008b. A timescale and phylogeny for "Bandicoots" (Peramelemorphia: Marsupialia) based on sequences for five nuclear genes. Molecular Phylogenetics and Evolution. 47: 1-20.

Mickevich M. F. 1978. Taxonomic congruence. Systematic Zoology. 27: 143-158.

Mickevich M. F. & Farris J. S. 1981. The Implications of Congruence in Menidia. Systematic Zoology. 30: 351-370.

Miller W., Drautz D. I., Janecka J. E., Lesk A. M., Ratan A., Tomsho L. P., Packard M., Zhang Y., et al. 2009. The mitochondrial genome sequence of the Tasmanian tiger (*Thylacinus cynocephalus*). Genome Research. 19: 213-220.

Miyamoto M. M. & Goodman M. 1986. Biomolecular systematics of eutherian mammals: Phylogenetic patterns and classification. Systematic Zoology. 35: 230-240.

Miyamoto M. M. & Fitch W. M. 1995. Testing species phylogenies and phylogenetic methods with congruence. Systematic Biology. 44: 64-76.

Montgelard C., Forty E., Arnal V. & Matthee C. A. 2008. Suprafamilial relationships among Rodentia and the phylogenetic effect of removing fast-evolving nucleotides in mitochondrial, exon and intron fragments. BMC Evolutionary Biology. 8: 321.

Moritz C., Dowling T. E. & Brown W. M. 1987. Evolution of animal mitochondrial DNA: Relevance for population biology and systematics. Annual Review of Ecology and Systematics. 18: 269-292.

Mouchaty S. K., Gullberg A., Janke A. & Arnason U. 2000. The phylogenetic position of the Talpidae within eutheria based on analysis of complete mitochondrial sequences. Molecular Biology and Evolution. 17: 60-67.

Munemasa M., Nikaido M., Donnellan S., Austin C. C., Okada N. & Hasegawa M. 2006. Phylogenetic analysis of diprotodontian marsupials based on complete mitochondrial genomes. Genes and Genetic Systems. 81: 181-191.

Murphy W. J., Eizirik E., Johnson W. E., Zhang Y. P., Ryder O. A. & O'Brien S. J. 2001a. Molecular phylogenetics and the origins of placental mammals. Nature. 409: 614-618.

Murphy W. J., Eizirik E., O'Brien S. J., Madsen O., Scally M., Douady C. J., Teeling E., Ryder O. A., et al. 2001b. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science. 294: 2348-2351.

Murphy W. J., Pevzner P. A. & O'Brien S. J. 2004. Mammalian phylogenomics comes of age. Trends in Genetics. 20: 631-639.

Murphy W. J., Pringle T. H., Crider T. A., Springer M. S. & Miller W. 2007. Using genomic data to unravel the root of the placental mammal phylogeny. Genome Research. 17: 413-421.

Nahum L. A. & Pereira S. L. 2008. Phylogenomics, protein family evolution, and the tree of life: An integrated approach between molecular evolution and computational intelligence. Studies in Computational Intelligence (SCI). 122: 259-279.

Nie W. H., Fu B. Y., O'Brien P. C. M., Wang J. H., Su W. T., Tanomtong A., Volobouev V., Ferguson-Smith M. A., et al. 2008. Flying lemurs: The 'flying tree shrews'? Molecular cytogenetic evidence for a Scandentia-Dermoptera sister clade. BMC Biology. 6: 18.

Nikaido M., Harada M., Cao Y., Hasegawa M. & Okada N. 2000. Monophyletic origin of the order Chiroptera and its phylogenetic position among Mammalia, as inferred from the complete sequence of the mitochondrial DNA of a Japanese megabat, the Ryukyu flying fox (*Pteropus dasymallus*). Journal of Molecular Evolution. 51: 318-328.

Nikaido M., Cao Y., Harada M., Okada N. & Hasegawa M. 2003. Mitochondrial phylogeny of hedgehogs and monophyly of Eulipotyphla. Molecular Phylogenetics and Evolution. 28: 276-284.

Nikaido M. R. A. P., Okada N. 1999. Phylogenetic relationships among cetartiodactyls based on insertions of short and long interspersed elements: Hippopotamuses are the closest extant relatives of whales. Proceedings of the National Academy of Sciences of the United States of America. 96: 10261-10266.

Nikolaev S., Montoya-Burgos J. I., Margulies E. H., Rougemont J., Nyffeler B., Antonarakis S. E. & Progra N. C. S. 2007. Early history of mammals is elucidated with the ENCODE multiple species sequencing data. PLoS Genetics. 3: e2.

Nilsson M. A., Gullberg A., Spotorno A. E., Arnason U. & Janke A. 2003. Radiation of extant marsupials after the K/T boundary: Evidence from complete mitochondrial genomes. Journal of Molecular Evolution. 57: S3-S12.

Nilsson M. A., Arnason U., Spencer P. B. S. & Janke A. 2004. Marsupial relationships and a timeline for marsupial radiation in South Gondwana. Gene. 340: 189-196.

Nishihara H., Satta Y., Nikaido M., Thewissen J. G. M., Stanhope M. J. & Okada N. 2005. A retroposon analysis of Afrotherian phylogeny. Molecular Biology and Evolution. 22: 1823-1833.

Nishihara H., Hasegawa M. & Okada N. 2006. Pegasoferae, an unexpected mammalian clade revealed by tracking ancient retroposon insertions. Proceedings of the National Academy of Sciences of the United States of America. 103: 9929-9934.

Nishihara H., Okada N. & Hasegawa M. 2007. Rooting the eutherian tree: The power and pitfalls of phylogenomics. Genome Biology. 8: R199.

Nishihara H., Shigenori M. & Norihiro O. 2009. Retroposon analysis and recent geological data suggest near-simultaneous divergence of the three superorders of mammals. Proceedings of the National Academy of Sciences of the United States of America. 106: 5235-5240.

Nixon K. C. & Davis J. I. 1991. Polymorphic taxa, missing values and cladistic analysis. Cladistics. 7: 233-241.

Novacek M. J. 1986. The skull of leptictid insectivorans and the higher-level classification of eutherian mammals. Bulletin of the American Museum of Natural History. 183: 1-111.

Novacek M. J. 1992. Mammalian phylogeny: Shaking the Tree. Nature. 356: 121-125.

Novacek M. J. 2001. Mammalian phylogeny: Genes and supertrees. Current Biology. 11: R573-R575.

Nylander J. A. A., Ronquist F., Huelsenbeck J. P. & Nieves-Aldrey J. L. 2004. Bayesian phylogenetic analysis of combined data. Systematic Biology. 53: 47-67.

O'Leary M. A. & Gatesy J. 2008. Impact of increased character sampling on the phylogeny of Cetartiodactyla (mammalia): Combined analysis including fossils. Cladistics. 24: 397-442.

Page R. D. M. 1989. Comments on component-compatibility in historical biogeography. Cladistics. 5: 167-182.

Page R. D. M. 1996. On consensus, confidence, and "total evidence". Cladistics. 12: 83-92.

Page R. D. M. & Charleston M. A. 1997. From gene to organismal phylogeny: Reconciled trees and the gene tree/species tree problem. Molecular Phylogenetics and Evolution. 7: 231-40.

Page R. D. M. 2000. Extracting species trees from complex gene trees: Reconciled trees and vertebrate phylogeny. Molecular Phylogenetics and Evolution. 14: 89-106.

Paradis E., Claude J. & Strimmer K. 2004. APE: Analyses of phylogenetics and evolution in R language. Bioinformatics. 20: 289-290.

Paradis E. 2006. Analyses of phylogenetics and evolution with R. Springer, New York.

Pardini A. T., O'Brien P. C. M., Fu B., Bonde R. K., Elder F. F. B., Ferguson-Smith M. A., Yang F. & Robinson T. J. 2007. Chromosome painting among Proboscidea, Hyracoidea and Sirenia: Support for Paenungulata (Afrotheria, Mammalia) but not Tethytheria. Proceedings of the Royal Society B. 274: 1333-1340.

Pearson K. 1904. Report on certain enteric fever inoculation statistics. British Medical Journal. 3: 1243-1246.

Pecon-Slattery J., Wilkerson A. J. P., Murphy W. J. & O'Brien S. J. 2004. Phylogenetic assessment of introns and SINEs within the Y chromosome using the cat family Felidae as a species tree. Molecular Biology and Evolution. 21: 2299-2309.

Pennisi E. 2003. Modernizing the Tree of Life. Science. 300: 1692-1697.

Penny D. & Hasegawa M. 1997. The platypus put in its place. Nature. 387: 549-550.

Philippe H. 1997. Rodent monophyly: Pitfalls of molecular phylogenies. Journal of Molecular Evolution. 45: 712-715.

Philippe H. & Laurent J. 1998. How good are deep phylogenetic trees? Current Opinion in Genetics and Development. 8: 616-623.

Philippe H., Snell E. A., Bapteste E., Lopez P., Holland P. W. H. & Casane D. 2004. Phylogenomics of eukaryotes: Impact of missing data on large alignments. Molecular Biology and Evolution. 21: 1740-52.

Philippe H., Delsuc F., Brinkmann H. & Lartillot N. 2005a. Phylogenomics. Annual Review of Ecology Evolution and Systematics. 36: 541-562.

Philippe H., Lartillot N. & Brinkmann H. 2005b. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Molecular Biology and Evolution. 22: 1246-1253.

Philippe H., Zhou Y., Brinkmann H., Rodrigue N. & Delsuc F. 2005c. Heterotachy and long-branch attraction in phylogenetics. BMC Evolutionary Biology. 5: 50.

Philippe H. & Telford M. J. 2006. Large-scale sequencing and the new animal phylogeny. Trends in Ecology and Evolution. 21: 614-620.

Philippe H., Brinkmann H., Martinez P., Riutort M., & Baguñà J. 2007. Acoel flatworms are not Platyhelminthes: Evidence from phylogenomics. PLoS ONE 2: e717.

Phillips M. J., Lin Y. H., Harrison G. L. & Penny D. 2001. Mitochondrial genomes of a bandicoot and a brushtail possum confirm the monophyly of australidelphian marsupials. Proceedings of the Royal Society of London Series B. 268: 1533-1538.

Phillips M. J. & Penny D. 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Molecular Phylogenetics and Evolution. 28: 171-185.

Phillips M. J., Delsuc F. & Penny D. 2004. Genome-scale phylogeny and the detection of systematic biases. Molecular Biology and Evolution. 21: 1455-1458.

Phillips M. J., McLenachan P. A., Down C., Gibb G. C. & Penny D. 2006. Combined mitochondrial and nuclear DNA sequences resolve the interrelations of the major Australasian marsupial radiations. Systematic Biology. 55: 122-137.

Phillips M. J. & Pratt R. C. 2008. Family-level relationships among the Australasian marsupial "herbivores" (Diprotodontia: Koala, wombats, kangaroos and possums). Molecular Phylogenetics and Evolution. 46: 594-605.

Pisani D. & Wilkinson M. 2002. Matrix representation with parsimony, taxonomic congruence, and total evidence. Systematic Biology. 51: 151-155.

Planet P. J. 2006. Tree disagreement: Measuring and testing incongruence in phylogenies. Journal of Biomedical Informatics. 39: 86-102.

Podani J. 2000. Simulation of random dendrograms and comparison tests: Some comments. Journal of Classification. 17: 123-142.

Poe S. & Swofford D. L. 1999. Taxon sampling revisited. Nature. 398: 299-300.

Pollock, D. D., and W. J. Bruno. 2000. Assessing an unknown evolutionary process: Effect of increasing site-specific knowledge through taxon addition. Molecular Biology and Evolution. 17: 1854-1858.

Pollock D. D., Zwickl D. J., McGuire J. A. & Hillis D. M. 2002. Increased taxon sampling is advantageous for phylogenetic inference. Systematic Biology. 51: 664-671.

Posada D. & Crandall K. A. 1998. MODELTEST: Testing the model of DNA substitution. Bioinformatics. 14: 817-818.

Posada D. & Crandall K. A. 2001. Selecting the best-fit model of nucleotide substitution. Systematic Biology. 50:580-601.

Poux C. & Douzery E. J. P. 2004. Primate phylogeny, evolutionary rate variations, and divergence times: A contribution from the nuclear gene IRBP. American Journal of Physical Anthropology. 124: 1-16.

Poux C., Chevret P., Huchon D., de Jong W. W. & Douzery E. J. P. 2006. Arrival and diversification of caviomorph rodents and platyrrhine primates in South America. Systematic Biology. 55: 228-244.

Prasad A. B., Allard M. W. & Green E. D. 2008. Confirming the phylogeny of mammals by use of large comparative sequence data sets. Molecular Biology and Evolution. 25: 1795-1808.

Prendini, L. 2001. Species or supraspecies taxa as terminals in cladistic analysis? Groundplans versus exemplars revisited. Systematic Biology. 50: 290-300

Pumo D. E., Finamore P. S., Franek W. R., Phillips C. J., Tarzami S. & Balzarano D. 1998. Complete mitochondrial genome of a neotropical fruit bat, *Artibeus jamaicensis*, and a new hypothesis of the relationships of bats to other eutherian mammals. Journal of Molecular Evolution. 47: 709-717.

Pupko T., Huchon D., Cao Y., Okada N. & Hasegawa M. 2002. Combining multiple data sets in a likelihood analysis: Which models are the best? Molecular Biology and Evolution. 19: 2294-2307.

Purvis A. 1995. A composite estimate of Primate phylogeny. Philosophical Transactions of the Royal Society of London Series B. 348: 405-421.

Quicke D. L. J., Jones O. R. & Epstein D. R. 2007. Correcting the problem of false incongruence due to noise imbalance in the incongruence length difference (ILD) test. Systematic Biology. 56: 496-503.

R Development Core Team. 2009. R: A language and environment for statistical computing. R Foundation for Statistical Computing. Vienna, Austria.

Ragan M. A. 1992a. Matrix representation in reconstructing phylogenetic relationships among the Eukaryotes. Biosystems. 28: 47-55.

Ragan M. A. 1992b. Phylogenetic inference based on matrix representation of trees. Molecular Phylogenetics and Evolution. 1: 53-8.

Rambaut A. & Grassly N. C. 1997. Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic frees. Computer Applications in the Biosciences. 13: 235-238.

Rannala B. & Yang Z. 1996. Probability distribution of molecular evolutionary trees: A new method of phylogenetic inference. Journal of Molecular Evolution. 43: 304-311.

Rannala B., Huelsenbeck J. P., Yang Z. H. & Nielsen R. 1998. Taxon sampling and the accuracy of large phylogenies. Systematic Biology. 47: 702-710.

Rannala B. & Yang Z. H. 2008. Phylogenetic inference using whole genomes. Annual Review of Genomics and Human Genetics. 9: 217-231.

Ranwez V., Delsuc F., Ranwez S., Belkhir K., Tilak M. K. & Douzery E. J. P. 2007. OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evolutionary Biology. 7: 241.

Ren F., Tanaka H. & Yang Z. 2009. A likelihood look at the supermatrix-supertree controversy. Gene. 441: 119-125.

Reyes A., Gissi C., Catzeflis F., Nevo E., Pesole G. & Saccone C. 2004. Congruent mammalian trees from mitochondrial and nuclear genes using Bayesian methods. Molecular Biology and Evolution. 21: 397-403.

Rice, W. E. 1989. Analyzing tables of statistical tests. Evolution. 43: 223-225.

Ripplinger J. & Sullivan J. 2008. Does choice in model selection affect maximum likelihood analysis? Systematic Biology. 57: 76-85.

Robinson D. F. & Foulds L. R. 1979. Comparisons on weighted labelled trees. Pp. 119-126 *in* Lecture notes in mathematics. Springer-Verlag, Berlin.

Robinson D. F. & Foulds L. R. 1981. Comparison of phylogenetic trees. Mathematical Biosciences. 53: 131-147.

Robinson T. J., Fu B., Ferguson-Smith M. A. & Yang F. 2004. Cross-species chromosome painting in the golden mole and elephant-shrew: Support for the mammalian clades Afrotheria and Afroinsectiphillia but not Afroinsectivora. Proceedings of the Royal Society of London Series B. 271: 1477-1484.

Rodrigo A. G. 1993. A comment on Baum's method for combining phylogenetic trees. Taxon. 42: 631-636.

Rodrigo A. G., Kelly-Borges M., Bergquist P. R. & Bergquist P. L. 1993. A randomization test of the null hypothesis that two cladograms are sample estimates of a parametric phylogenetic tree. New Zealand Journal of Botany. 31: 257-268.

Rodrigo A. G. 1996. On combining cladograms. Taxon. 45: 267-274.

Rodriguez F., Oliver J. L., Marin A. & Medina J. R. 1990. The general stochastic model of nucleotide substitution. Journal of Theoretical Biology. 142: 485-501.

Rodriguez-Ezpeleta N., Brinkmann H., Burey S. C., Roure B., Burger G., Loffelhardt W., Bohnert H. J., Philippe H., et al. 2005. Monophyly of primary photosynthetic eukaryotes: Green plants, red algae, and glaucophytes. Current Biology. 15: 1325-1330.

Rodriguez-Ezpeleta N., Brinkmann H., Roure B., Lartillot N., Lang B. F. & Philippe H. 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Systematic Biology. 56: 389-399.

Roeding F., Hagner-Holler S., Ruhberg H., Ebersberger I., von Haeseler A., Kube M., Reinhardt R. & Burmester T. 2007. EST sequencing of Onychophora and phylogenomic analysis of Metazoa. Molecular Phylogenetics and Evolution. 45: 942-951.

Rohlf F. J. 1982. Consensus Indices for Comparing Classifications. Mathematical Biosciences. 59: 131-144.

Rokas A. & Holland P. W. H. 2000. Rave genomic changes as a tool for phylogenetics. Trends in Ecology and Evolution. 15: 454-459.

Rokas A., King N., Finnerty J. & Carroll S. B. 2003a. Conflicting phylogenetic signals at the base of the metazoan tree. Evolution and Development. 5: 346-359.

Rokas A., Williams B. L., King N. & Carroll S. B. 2003b. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature. 425: 798-804.

Rokas A. & Carroll S. B. 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Molecular Biology and Evolution. 22: 1337-1344.

Rokas A., Krüger D. & Carroll S. B. 2005. Animal evolution and the molecular signature of radiations compressed in time. Science. 310: 1933-1938.

Rokas A. & Carroll S. B. 2006. Bushes in the tree of life. PLoS Biology. 4: 1899-1904.

Ronquist F. & Huelsenbeck J. P. 2003. MrBayes 3: Bayesian phylogenetic inference under mixed models. Bioinformatics. 19: 1572-1574.

Rosenberg M. S. & Kumar S. 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proceedings of the National Academy of Sciences of the United States of America. 98: 10751-10756.

Rosenberg M. S. & Kumar S. 2003. Taxon sampling, bioinformatics, and phylogenomics. Systematic Biology. 52: 119-124.

Rosenberg N. A. & Tao R. 2008. Discordance of species trees with their most likely gene trees: The case of five taxa. Systematic Biology. 57: 131-140.

Ross H. A. & Rodrigo A. G. 2004. An assessment of matrix representation with compatibility in supertree construction. Pp. 35-63 *in* Phylogenetic supertrees: Combining information to reveal the Tree of Life (O. R. P. Bininda-Emonds, ed.). Kluwer Academic Publishers, Dordrecht, the Netherlands.

Rowe T., Rich T. H., Vickers-Rich P., Springer M. & Woodburne M. O. 2008. The oldest platypus and its bearing on divergence timing of the platypus and echidna clades. Proceedings of the National Academy of Sciences of the United States of America. 105: 1238-1242.

Rudd S. 2003. Expressed sequence tags: Alternative or complement to whole genome sequences? Trends in Plant Science. 8: 321-329.

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

Sánchez-Villagra M. R., Narita Y. & Kuratani S. 2007. Thoracolumbar vertebral number: The first skeletal synapomorphy for afrotherian mammals. Systematics and Biodiversity. 5: 1-7.

Sanderson M. J., Purvis A. & Henze C. 1998. Phylogenetic supertrees: Assembling the trees of life. Trends in Ecology and Evolution. 13: 105-109.

Sanderson M. J. & Driskell A. C. 2003. The challenge of constructing large phylogenetic trees. Trends in Plant Science. 8: 374-9.

Sanger F, Nicklen S. & Coulson A. R. 1977. DNA sequencing with chain-terminating inhibitors. Proceedings of the National Academy of Sciences of the United States of America. 74: 5463-5467.

Sanson G. F. O., Kawashita S. Y., Brunstein A., & Briones M. R. S. 2002. Experimental phylogeny of neutrally evolving DNA sequences generated by a bifurcate series of nested polymerase chain reactions. Molecular Biology and Evolution. 19: 170-178.

Scally M., Madsen O., Douady C. J., de Jong W. W., Stanhope M. J. & Springer M. S. 2001. Molecular evidence for the major clades of placental mammals. Journal of Mammalian Evolution. 8: 239-277.

Schmitz J., Ohme M. & Zischler H. 2002. The complete mitochondrial sequence of Tarsius bancanus: Evidence for an extensive nucleotide compositional plasticity of primate mitochondrial DNA. Molecular Biology and Evolution. 19: 544-553.

Seiffert E. R. 2007. A new estimate of afrotherian phylogeny based on simultaneous analysis of genomic, morphological, and fossil evidence. BMC Evolutionary Biology. 7: 224.

Shedlock A. M., Botka C. W., Zhao S. Y., Shetty J., Zhang T. T., Liu J. S., Deschavanne P. J. & Edward S. V. 2007. Phylogenomics of nonavian reptiles and the structure of the ancestral amniote genorne. Proceedings of the National Academy of Sciences of the United States of America. 104: 2767-2772.

Shoshani J. & McKenna M. C. 1998. Higher taxonomic relationships among extant mammals based on morphology, with selected comparisons of results from molecular data. Molecular Phylogenetics and Evolution. 9: 572-84.

Siegel S. & Castellan N. J. Jr. 1988. Nonparametric statistics for the behavioral sciences. Second edition. McGraw-Hill, New York.

Simmons N. B. & Geisler J. H. 1998. Phylogenetic relationships of Icaronycteris, Archaeonycteris, Hassianycteris, and Palaeochiropteryx to extant bat lineages, with comments on the evolution of echolocation and foraging strategies in Microchiroptera. Bulletin of the American Museum of Natural History. 235: 1-182.

Simpson G. G. 1945. The principles of classification and a classification of mammals. Bulletin of the American Museum of Natural History. 85: 1-350.

Smith A. G., Smith D. G. & Funnell B. M. 1994. Atlas of Mesozoic and Cenozoic coastlines. Cambridge University Press, Cambridge, U. K.

Smith S. A. & Donoghue M. J. 2008. Rates of molecular evolution are linked to life history in flowering plants. Science. 322: 86-89.

Smith S. A., Beaulieu J. M. & Donoghue M. J. 2009. Mega-phylogeny approach for comparative biology: An alternative to supertree and supermatrix approaches. BMC Evolutionary Biology. 9: 37.

Snel B., Bork P. & Huynen M. A. 1999. Genome phylogeny based on gene content. Nature Genetics. 21: 108-110.

Snel B., Bork P. & Huynen M. 2000. Genome evolution: Gene fusion versus gene fission. Trends in Genetics. 16: 9-11.

Sokal R. R. & Rohlf F. J. 1962. The comparison of dendrograms by objective methods. Taxon. 11: 33-40.

Sokal R. R. & Rohlf F. J. 1981. Taxonomic congruence in the Leptopodomorpha re-examined. Systematic Zoology. 30: 309-325.

Springer M. S. & de Jong W. W. 2001. Which mammalian supertree to bark up? Science. 291: 1709-1711.

Springer M. S., DeBry R. W., Douady C., Amrine H. M., Madsen O., de Jong W. W. & Stanhope M. J. 2001. Mitochondrial versus nuclear gene sequences in deep-level mammalian phylogeny reconstruction. Molecular Biology and Evolution. 18: 132-143.

Springer M. S., Scally M., Madsen O., de Jong W. W., Douady C. J. & Stanhope M. J. 2004a. The use of composite taxa in supermatrices. Molecular Phylogenetics and Evolution. 30: 883-884.

Springer M. S., Stanhope M. J., Madsen O. & de Jong W. W. 2004b. Molecules consolidate the placental mammal tree. Trends in Ecology and Evolution. 19: 430-438.

Springer M. S., Burk-Herrick A., Meredith R., Eizirik E., Teeling E., O'Brien S. J. & Murphy W. J. 2007. The adequacy of morphology for reconstructing the early history of placental mammals. Systematic Biology. 56: 673-684.

Springer M. S. & Murphy W. J. 2007. Mammalian evolution and biomedicine: New views from phylogeny. Biological Reviews. 82: 375-392.

Stanhope M. J., Waddell V. G., Madsen O., de Jong W., Hedges S. B., Cleven G. C., Kao D. & Springer M. S. 1998. Molecular evidence for multiple origins of Insectivora and for a new order of endemic African insectivore mammals. Proceedings of the National Academy of Sciences of the United States of America. 95: 9967-9972.

Steel M. & Penny D. 2000. Parsimony, likelihood, and the role of models in molecular phylogenetics. Molecular Biology and Evolution. 17: 839-850.

Steel M. & Rodrigo A. 2008. Maximum likelihood supertrees. Systematic Biology. 57: 243-250.

Steel M. A., Lockhart P. J. & Penny D. 1993. Confidence in evolutionary trees from biological sequence data. Nature. 364: 440-442.

Steel M. A. & Penny D. 1993. Distributions of tree comparison metrics: Some new results. Systematic Biology. 42: 126-141.

Stewart C. B., Schilling J. W. & Wilson A. C. 1987. Adaptive evolution in the stomach lysozymes of foregut fermenters. Nature. 330: 401-404.

Struck T. H. & Fisse F. 2008. Phylogenetic position of nemertea derived from phylogenomic data. Molecular Biology and Evolution. 25: 728-736.

Suchard M. A. 2005. Stochastic models for horizontal gene transfer: Taking a random walk through tree space. Genetics. 170: 419-431.

Sullivan J. & Swofford D. L. 2001. Should we use model-based methods for phylogenetic inference when we know that assumptions about among-site rate variation and nucleotide substitution pattern are violated? Systematic Biology. 50: 723-729.

Susko E., Leigh J., Doolittle W. F. & Bapteste E. 2006. Visualizing and assessing phylogenetic congruence of core gene sets: A case study of the gamma-proteobacteria. Molecular Biology and Evolution. 23: 1119-1030.

Svartman M., Stone G. & Stanyon R. 2006. The ancestral Eutherian karyotype is present in Xenarthra. PLoS Genetics. 2: 1006-1011.

Swofford D. L. 1991. When are phylogeny estimates from molecular and morphological data incongruent? Pp. 295-333 *in* Phylogenetic analyses of DNA sequences (M. M. Miyamoto, & J. Cracraft, eds.). Oxford University Press, Oxford.

Swofford D. L., Olsen G. J., Waddell P. J. & Hillis D. M. 1996. Phylogenetic inference. Pp 407-514 *in* Molecular systematics (D. M. Hillis, C. Moritz, & B. K. Mable, eds.). Sinauer Associates, Sunderland, Massachusetts.

Swofford D. L. 1998. PAUP* Phylogenetic analysis using parsimony (*and other methods). Sinauer Associates, Sunderland, Massachusetts.

Szalay F. S. 1982. A new appraisal of marsupial phylogeny and classification. Pp. 621-640 *in* Carnivorous marsupials (M. Archer, ed.). Royal Zoological Society of New South Wales, Sydney, Australia.

Tamura K. & Nei M. 1993. Estimation of the number of nucleotide sustitutions in the control region of mitochondrial DNA in humans and chimpanzees. Molecular Biology and Evolution. 10: 512-526.

Tárraga J., Medina I., Arbiza L., Huerta-Cepas J., Gabaldon T., Dopazo J. & Dopazo H. 2007. Phylemon: A suite of web tools for molecular evolution, phylogenetics and phylogenomics. Nucleic Acids Research. 35: W38-W42.

Tateno Y., Takezaki N. & Nei M. 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Molecular Biology and Evolution. 11: 261-277.

Tavaré S. 1986. Some probabilistic and statistical problems on the analysis of DNA sequences. Lectures on Mathematics in the Life Sciences. 17: 57-86.

Teeling E. C., Springer M. S., Madsen O., Bates P., O'Brien S. J. & Murphy W. J. 2005. A molecular phylogeny for bats illuminates biogeography and the fossil record. Science. 307: 580-584.

Telford M. J. 2007. Phylogenomics. Current Biology. 17: R945-R946.

Telford M. J. 2008. Resolving animal phylogeny: A sledgehammer for a tough nut? Developmental Cell. 14: 457-459.

Templeton A. R. 1983. Phylogenetic inference from restriction endonuclease cleavage site maps with particular reference to the evolution of human and the apes. International Journal of Organic Evolution. 37: 221-244.

Thomas J. W., Touchman J. W., Blakesley R. W., Bouffard G. G., Beckstrom-Sternberg S. M., Margulies E. H., Blanchette M., Siepel A. C., et al. 2003. Comparative analyses of multi-species sequences from targeted genomic regions. Nature. 424: 788-793.

Ursing B. M. & Arnason U. 1998. Analyses of mitochondrial genomes strongly support a hippopotamus-whale clade. Proceedings of the Royal Society of London Series B. 265: 2251-2255.

van Rheede T., Bastiaans T., Boone D. N., Hedges S. B., de Jong W. W. & Madsen O. 2006. The platypus is in its place: Nuclear genes and indels confirm the sister group relation of monotremes and therians. Molecular Biology and Evolution. 23: 587-597.

Waddell P. J., Cao Y., Hauf J. & Hasegawa M. 1999a. Using novel phylogenetic methods to evaluate mammalian mtDNA, including amino acid-invariant sites-logdet plus site stripping, to detect internal conflicts in the data, with special reference to the positions of hedgehog, armadillo and elephant. Systematic Biology. 48: 31-53.

Waddell P. J., Okada N. & Hasegawa M. 1999b. Towards resolving the interordinal relationships of placental mammals. Systematic Biology. 48: 1-5.

Waddell P. J., Kishino H. & Ota R. 2001. A phylogenetic foundation for comparative mammalian genomics. Genome Informatics. 12: 141-154.

Waddell P. J. & Shelley S. 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, γ-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Molecular Phylogenetics and Evolution. 28: 197-224.

Waters P. D., Dobigny G., Waddell P. J. & Robinson T. J. 2007. Evolutionary history of LINE-1 in the major clades of placental mammals. PLoS ONE. 2: e158.

Weisrock D. W., Harmon L. J. & Larson A. 2005. Resolving deep phylogenetic relationships in salamanders: Analyses of mitochondrial and nuclear genomic data. Systematic Biology. 54: 758-777.

Wendel J. F. & Doyle J. J. 1998. Phylogenetic incongruence: Window into genome history and molecular evolution. Pp. 265-296 *in* Molecular systematics of plants II: DNA sequencing (P. S. Soltis, D. E. Soltis, & J. J. Doyle, eds.). Kluwer Academic Publisher, Dordrecht, The Netherlands.

Whelan S. 2008. The genetic code can cause systematic bias in simple phylogenetic models. Philosophical Transactions of the Royal Society B. 363: 4003-4011.

Wible J. R., Rougier G. W., Novacek M. J. & Asher R. J. 2007. Cretaceous eutherians and Laurasian origin for placental mammals near the K/T boundary. Nature. 447: 1003-1006.

Wiens J. J. & Reeder T. W. 1995. Combining data sets with different numbers of taxa for phylogenetic analysis. Systematic Biology. 44: 548-558.

Wiens J. J. 1998a. Combining data sets with different phylogenetic histories. Systematic Biology. 47: 568-581.

Wiens J. J. 1998b. The accuracy of methods for coding and sampling higher-level taxa for phylogenetic analysis: A simulation study. Systematic Biology. 47: 397-413.

Wiens J. J. 1998c. Does adding characters with missing data increase or decrease phylogenetic accuracy? Systematic Biology. 47: 625-640.

Wiens J. J. 2003a. Incomplete taxa, incomplete characters, and phylogenetic accuracy: Is there a missing data problem? Journal of Vertebrate Paleontology. 23: 297-310.

Wiens J. J. 2003b. Missing data, incomplete taxa, and phylogenetic accuracy. Systematic Biology. 52: 528-538.

Wiens J. J. 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Systematic Biology. 54: 731-742.

Wiens J. J., Fetzner J. W., Parkinson C. L. & Reeder T. W. 2005. Hylid frog phylogeny and sampling strategies for speciose clades. Systematic Biology. 54: 719-748.

Wiens J. J. 2006. Missing data and the design of phylogenetic analyses. Journal of Biomedical Informatics. 39: 34-42.

Wiens J. J., Kuczynski C. A., Smith S. A., Mulcahy D. G., Sites J. W., Townsend T. M. & Reeder T. W. 2008. Branch lengths, support, and congruence: Testing the phylogenomic approach with 20 nuclear loci in snakes. Systematic Biology. 57: 420-431.

Wiens J. J. & Moen D. S. 2008. Missing data and the accuracy of Bayesian phylogenetics. Journal of Systematics and Evolution. 46: 307-314.

Wildman D. E., Uddin M., Opazo J. C., Liu G., Lefort V., Guindon S., Gascuel O., Grossman L. I., et al. 2007. Genomics, biogeography, and the diversification of placental mammals. Proceedings of the National Academy of Sciences of the United States of America. 104: 14395-14400.

Wilkinson M., Thorley J. L., Littlewood D. T. J. & Bray R. A. 2001. Towards a phylogenetic supertree of Platyhelminthes? Pp. 292-301 *in* Interrelationships of the Platyhelminthes (D. T. J. Littlewood, & R. A. Bray, eds.). Taylor & Francis, London, New York.

Wilkinson M., Cotton J. A., Creevey C., Eulenstein O., Harris S. R., Lapointe F.-J., Levasseur C., McInerney J. O., et al. 2005. The shape of supertrees to come: Tree shape related properties of fourteen supertree methods. Systematic Biology. 54: 419-431.

Wilkinson M. & Cotton J. A. 2006. Supertree methods for building the tree of life: divide-and-conquer approaches to large phylogenetic problems. Pp. 61-75 *in* Towards the Tree of Life: Taxonomy and Systematics of Large and Species Rich Taxa (Hodkinson, T., J. Parnell, and S. Waldren, eds.). Systematic Association special volume, CRC Press.

Wilkinson M., Cotton J. A., Lapointe F.-J. & Pisani D. 2007. Properties of supertree methods in the consensus setting. Systematic Biology. 56: 330-337.

Wilson D. E. & Reeder D. M. 1993. Mammal species of the world: A taxonomic and geographic reference. Second edition. Smithsonian Institution, Washington, D.C.

Wilson D. E. & Reeder D. M. 2005. Mammal species of the world: A taxonomic and geographic reference. Third edition. The Johns Hopkins University Press, Baltimore, Maryland.

Wolf Y. I., Rogozin I. B., Grishin N. V. & Koonin E. V. 2002. Genome trees and the Tree of Life. Trends in Genetics. 18: 472-479.

Wolfe K. H., Sharp P. M. & Li W. H. 1989. Mutation rates differ among regions of the mammalian genome. Nature. 337: 283-285.

Wu M. & Eisen J. A. 2008. A simple, fast, and accurate method of phylogenomic inference. Genome Biology. 9: R151.

Xiong Y., Brandley M. C., Xu S. X., Zhou K. Y. & Yang G. 2009. Seven new dolphin mitochondrial genomes and a time-calibrated phylogeny of whales. BMC Evolutionary Biology. 9: 20.

Yang Z. H. 1993. Maximum-likelihood estimation of phylogeny from DNA sequences when substitution rates differ over sites. Molecular Biology and Evolution. 10: 1396-1401.

Yang Z. H. 1996a. Phylogenetic analysis using parsimony and likelihood methods. Journal of Molecular Evolution. 42: 294-307.

Yang Z. H. 1996b. Maximum-likelihood models for combined analyses of multiple sequence data. Journal of Molecular Evolution. 42: 587-596.

Zou X. H., Zhang F. M., Zhang J. G., Zang L. L., Tang L., Wang J., Sang T. & Ge S. 2008. Analysis of 142 genes resolves the rapid diversification of the rice genus. Genome Biology. 9: R49.

Zwickl D. J. & Hillis D. M. 2002. Increased taxon sampling greatly reduces phylogenetic error. Systematic Biology. 51: 588-598.