UNIVERSITÉ DE MONTRÉAL

De la pertinence de la congruence globale en analyse phylogénétique

par

CLAUDINE LEVASSEUR

Département de sciences biologiques
Faculté des Arts et des sciences

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

Mai 2005

Page d'identification du jury

UNIVERSITÉ DE MONTRÉAL

Faculté des études supérieures

Cette thèse intitulée :

De la pertinence de la congruence globale en analyse phylogénétique

présentée par :

CLAUDINE LEVASSEUR

a été évaluée par un jury composé des personnes suivantes :

Pierre Legendre, président-rapporteur

François-Joseph Lapointe, directeur de recherche

Bernard Angers, membre du jury

Olaf Bininda-Emonds, examinateur externe

RÉSUMÉ

L'avènement des techniques moléculaires modernes a donné lieu à une véritable révolution en analyse phylogénétique. Parce qu'elles offrent la possibilité d'obtenir une grande quantité de caractères en peu de temps, ces nouvelles méthodes s'avèrent des outils appréciables pour générer les données qui serviront à estimer les relations évolutives entre les organismes vivants. Mais cette accumulation rapide des données moléculaires a suscité plusieurs questionnements quant à la façon d'analyser ces différents gènes et les autres types de données déjà existants, comme la morphologie. Un débat oppose principalement deux opinions: combiner les données (congruence des caractères) ou les analyser séparément pour ensuite combiner les arbres à l'aide d'une méthode de consensus (congruence taxonomique).

Devant la controverse qui existe encore sur la meilleure stratégie à adopter, une nouvelle position a été présentée. Plutôt que choisir l'une ou l'autre, l'approche de congruence globale préconise l'utilisation conjointe des analyses combinée et séparées. Elle stipule que l'utilisation d'une méthode de consensus qui tient compte des longueurs de branches (consensus moyen) permet d'obtenir des résultats semblables ou identiques à l'aide des deux approches.

Cette thèse présente divers articles qui étudient la pertinence de cette nouvelle approche, d'abord dans un cadre de consensus et ensuite pour la reconstruction de super-arbres. Dans le premier volet, il est montré que les arbres issus du consensus moyen sont souvent identiques à la phylogénie provenant de l'analyse combinée. De plus, dans les cas où le consensus moyen diffère de cette dernière, la validation de l'arbre de congruence des caractères révèle que les zones de désaccord sont généralement peu supportées par les données. Des études de simulations indiquent également que l'approche de congruence globale permet d'améliorer la qualité des estimations phylogénétiques, particulièrement lorsque le consensus moyen est utilisé pour combiner les arbres obtenus par des analyses séparées. Enfin, les arbres produits à l'aide de la méthode de MRP (matrix representation with parsimony), qui semble partager les attributs des approches de congruence des caractères et taxonomique, ont été comparés à ceux résultant de la combinaison des données et du consensus moyen. Malgré son caractère hybride, les résultats ont montré que

les arbres des analyses combinée et séparées sont plus semblables entre eux qu'ils ne le sont des arbres obtenus avec la méthode de MRP.

Dans un deuxième volet, la généralisation de la congruence globale au cas particulier de super-arbres (où les jeux de données à combiner ne possèdent que quelques espèces en commun) a été testée. Dans ces situations, certaines distances entre les taxons sont inconnues. Une étude par simulation a permis de montrer qu'il est pertinent d'estimer ces dernières, puisque cela augmente la fiabilité des arbres obtenus. Finalement, le taux de succès du consensus moyen a été mesuré dans le cadre des super-arbres. D'après les résultats des simulations, il reste encore certains problèmes à étudier avant de pouvoir utiliser l'approche de congruence globale pour la reconstruction de super-arbres.

Parce que l'analyse phylogénétique est constamment en développement, de nouvelles méthodes sont régulièrement proposées. Il est impératif de tester la justesse de ces approches pour optimiser leur utilisation. La présente recherche s'inscrit dans cette visée et a permis de montrer que l'approche de congruence globale, telle que présentée dans cette thèse, représente une solution intéressante pour reconstruire des phylogénies plus justes. Par contre, les résultats montrent également que certains problèmes méthodologiques restent à investiguer avant de pouvoir généraliser cette approche au cas particulier de super-arbres.

Mots-clés : analyse phylogénétique, congruence des caractères, congruence globale, congruence taxonomique, consensus, données manquantes, « *matrix representation with parsimony* » (MRP), simulations, super-arbres, « *total evidence* »

ABSTRACT

With the growing development of modern molecular techniques, phylogenetic analysis has undergone a true revolution. Because they allow collecting many characters very rapidly, these methods represent valuable tools to generate data that can be used for estimating the evolutionary relationships among living organisms. However, the exponential accumulation of data gave rise to many questions about how to analyze the different genes with other types of data already at hand, like morphology. A current debate opposes two distinct approaches: combining all the data to generate one phylogenetic hypothesis (character congruence) or analyze each data partition separately before combining the resulting phylogenies with a consensus method (taxonomic congruence).

Yet no agreement has been reached on the best strategy to adopt, and a new position has been recently presented. Rather than choosing one or the other, the global congruence approach advocates the joint use of combined and separate analyses. It also proposes that combining trees with a consensus technique that takes into account branch length (average consensus) may provide similar or identical results to the character congruence approach.

This thesis presents different articles that study the pertinence of this new approach in the consensus and supertree settings. In the first section, we show that average consensus trees are often identical to the combined analysis phylogeny. Moreover, in cases where those results differ, validation of the character congruence tree generally reveals that parts of the trees that are incongruent are not well supported by the data. Simulation studies also show that accuracy can be increased by the global congruence approach, especially when the average consensus is used to combine trees from the separate analyses. Lastly, trees constructed with the MRP (matrix representation with parsimony) method, which seems to be a hybrid of the character and taxonomic congruence approaches, have been compared to the ones obtained with the average consensus and combined analysis. These results indicate that the two competing methods are closer to one another than to the hybrid MRP method.

In the second section, we tested the applicability of the global congruence to the more general cases of supertrees (when different data sets partially overlap). In those situations, distances among some taxa are undefined. A simulation study showed that it is judicious to estimate those missing distances since it increases the accuracy of the trees.

Finally, success rate of the average consensus was measured in the supertree setting. Based on the simulation results, there are still many problems to overcome before the global congruence approach could be used to reconstruct supertrees.

Phylogenetic analysis is a field in constant evolution. Thus, new methods are proposed regularly. It is necessary to assess the reliability of these approaches to optimize their use. This research allowed showing that the global congruence approach, as presented in this thesis, represents an appealing alternative for estimating more accurate phylogenies. However, results also show that some methodological issues need to be investigated before generalizing this approach to the supertree setting.

Keywords: character congruence, consensus, global congruence, matrix representation with parsimony (MRP), missing data, phylogenetic analysis, taxonomic congruence, simulations, supertrees, total evidence

TABLE DES MATIÈRES

LISTE DES TABLEAUX

LISTE DES FIGURES

À tous les créateurs

«Lorsque je lis le dictionnaire dans l'ordre, je lis tous les livres dans le désordre.»

La Rage
Louis Hamelin

REMERCIEMENTS

Mes premiers remerciements s'adressent à mon directeur de recherche. Tout d'abord, pour avoir piqué ma curiosité : François, ton enseignement hors du commun m'a vraiment touchée et tu as su me contaminer de la passion qui t'anime. Ensuite pour m'avoir fait confiance, moi qui étais une étudiante inconnue venue d'un autre monde, celui de l'anthropologie. Tu as su me faire une place au sein de ton laboratoire et faire preuve de toute la patience voulue pour m'aider à m'intégrer dans ce merveilleux univers de la biologie. Je te suis également très reconnaissante pour les nombreuses opportunités que tu m'as offertes qui m'ont permis d'acquérir toute l'expérience nécessaire pour réussir dans le domaine compétitif de la recherche : publications, congrès, expériences d'enseignement. Tu as su me laisser apprendre tout en me guidant et ce, malgré la lenteur que cela impliquait parfois! Merci pour tous ces exercices formateurs! Je te remercie aussi pour ton dynamisme excessif, ton dévouement envers tes étudiants, ta motivation débordante et ta grande disponibilité. Également pour les rires, les anecdotes, et tous les Kilos, Souvenirs d'Afrique, Nonya et autres. Sans oublier les permissions de décoration du labo, les après-midi non-productifs et les chocolats chauds dans les moments de découragement. J'ai toujours senti ta présence à mes côtés. Tu as été une oreille attentive, compréhensive et ouverte. Tu m'as fourni tous les outils nécessaires au niveau académique, mais surtout au niveau humain. J'ai senti que tu me considérais comme ton égale, et non comme ton disciple (mais tu resteras toujours Le Guru dans mon cœur!). Tu as aussi permis à la vie de continuer pendant ces cinq années de labeur. Voyages au bout du monde, famille, enfants, passions connexes ou éloignées n'ont jamais eu à être délaissés pour la science.

J'aimerais aussi porter une attention toute spéciale aux autres membres du LEMEE. Pierre-Alexandre qui a été le premier à tracer la voie vers le fameux Ph.D., Olivier avec ses mille talents et son rire exubérant, Nathalie, maman du labo et de toutes les mamans, Sarah avec ses Cherry Blossom maison, et aussi les nouveaux arrivés, Catherine, Sébastien, Anaïs, Véronique et Yong, qui ont apporté un vent de fraîcheur (et qui ont également fait baisser la moyenne d'âge radicalement!). Merci pour les escapades au bilboquet, les fondues au chocolat, les petites surprises sucrées, les diaporamas de voyage, les histoires à dormir debout et les rires qui en ont rendu plusieurs jaloux! La vie de labo n'aurait jamais

été aussi agréable sans vous tous! J'ai également une pensée pour nos visiteuses assidues : Louise qui a toujours gardé un œil sur la tribu et Jacinthe qui, par son exemple acharné, m'a donné le courage de continuer… et de terminer!

Je ne peux passer sous silence la contribution des professeurs Pierre Legendre, Anne Bruneau, Guy Cucumel et Bernard Angers qui ont accepté de siéger à l'un ou l'autre des comités qui m'ont évaluée au cours de mes études. Merci à vous tous d'avoir pris le temps de m'écouter et de me conseiller malgré vos horaires chargés. J'apprécie que vous l'ayez fait avec tant de professionnalisme.

Un merci particulier à ma famille et ma belle-famille pour leurs encouragements, leur patience, leur comprehension et leur grande disponibilité dans les moments de crise. Votre aide à un moment ou l'autre a toujours été précieuse et c'est aussi grâce à votre soutien si j'ai pu terminer ma thèse. Je dois également souligner la présence assidue de mes parents qui m'a permis de concilier études et famille dans le respect de mes valeurs.

Je désire remercier spécialement Pierre Alexandre, mon amoureux et meilleur ami, sans qui rien de ceci n'aurait été possible. Ton support inconditionnel, ta confiance, tes questionnements, tes remises en question, ta passion, ta vision du monde marginale, ta spontanéité, ta créativité, ta simplicité, tes rêves, tes taquineries et ta bonne nature sont autant de choses qui m'ont permis d'avancer sereinement et de réaliser toutes les étapes qui m'ont menée au dépôt de cette thèse. Tu as toujours fait preuve de beaucoup de patience avec moi et ta présence aura été indispensable jusqu'à la fin. Merci d'avoir été là, d'être là et d'exister. Je ne peux me permettre d'oublier ma fille, Maxime, qui a inondé les deux dernières années d'étude d'éclat de rires et de bonheur, et qui m'a souvent permis de décrocher pour mieux travailler. Tes finesses, tes sourires et tes douceurs ont ajouté une dimension exceptionnelle à ma vie d'étudiante. Merci également à mon fils, Édouard, qui par sa venue a fixé un échéancier tangible à la fin de ce projet, sans qui je serais sans doute encore en pleine rédaction!

Finalement, j'aimerais remercier le Conseil de Recherche en Sciences Naturelle et en Génie du Canada (CRSNG), le Fonds pour la formation de l'Aide à la Recherche du Québec (FCAR), le Département de sciences biologiques, la Faculté des études supérieures

et François-Joseph Lapointe pour leur soutien financier qui m'a permis de réaliser ce projet tout en restant saine d'esprit.

INTRODUCTION

Une véritable révolution s'est amorcée dans le domaine de l'analyse phylogénétique dans les dernières années (Nei & Kumar 2000). Il y a quelques décennies, les chercheurs qui tentaient de découvrir les relations évolutives entre les espèces disposaient de peu de moyens pour accomplir cette lourde tâche. La morphologie était de loin le type de données le plus utilisé pour reconstruire l'histoire évolutive (Miyamoto & Cracraft 1991) et chacun de ces caractères était soigneusement choisi et mesuré (Hillis 1987). Cette opération, souvent longue et fastidieuse, ne permettait l'analyse que d'un petit nombre de caractères. De plus, les méthodes d'analyse phylogénétique et les outils informatiques de l'époque étaient très limitants et peu performants.

Depuis l'avènement des techniques moléculaires modernes (notamment le séquençage) et l'accès facile aux ressources informatiques de pointe, la recherche dans le domaine a beaucoup changé. Pratiquement toutes les hypothèses phylogénétiques désormais proposées s'appuient sur des données moléculaires, en tout ou en partie. Le nombre de caractères dans chaque analyse a augmenté considérablement et l'utilisation de plusieurs gènes est maintenant fortement prescrite (Helm-Bychowski & Cracraft 1993; Hillis 1995; Lanyon 1993; Miyamoto & Fitch 1995; Sanderson & Shaffer 2002; Sheldon & Bledsoe 1993; Wendel & Doyle 1998). C'est que les données moléculaires s'accumulent à un rythme affolant. Par exemple, nous connaissons à ce jour le génome complet de plusieurs espèces (Nei & Kumar 2000). De plus, les méthodes d'analyse phylogénétique se sont multipliées et elles sont maintenant très conviviales et largement distribuées. La matière première est donc à notre disposition et les outils nécessaires à son analyse sont de plus en plus adaptés à cette nouvelle réalité.

Néanmoins, plusieurs problèmes et questionnements subsistent encore malgré cette situation en apparence idéale. Par exemple, les années transitoires entre l'ère des taxonomistes classiques et l'ère des biologistes moléculaires a donné naissance à un grand débat qui a suscité de nombreuses réactions et qui n'a toujours pas fait de consensus au sein

de la communauté scientifique. Comme les nouvelles hypothèses phylogénétiques soutenues par les données moléculaires contredisent parfois celles déjà proposées (dérivées des caractères morphologiques), la question se posa d'abord sur la pertinence de l'un et l'autre type de données (Donoghue & Sanderson 1992; Doyle 1992; Patterson *et al.* 1993; Systma 1990). Comme il semblait absurde de devoir mettre de côté l'un ou l'autre type de données, un nouveau questionnement sur la meilleure façon de réconcilier ces hypothèses contradictoires donna naissance à un débat qui dure depuis plus de 15 ans. À ce jour, malgré l'utilisation presque exclusive des données moléculaires, le débat est toujours d'actualité puisqu'il est fréquent que les gènes utilisés dans les analyses phylogénétiques produisent différentes interprétations des relations évolutives entre les organismes (Wendel & Doyle 1998).

DÉFINIR L'INCONGRUENCE

Les arbres phylogénétiques

Les différents caractères (morphologiques, moléculaires ou autres) prélevés des organismes sont analysés de manière à reconstruire leur histoire évolutive. Ces relations sont représentées sous la forme d'un arbre, où chacune des branches terminales correspond à un de ces dits organismes (voir Figure 1). Ces branches terminales peuvent représenter n'importe quel niveau taxonomique (espèce, genre, famille) mais sont le plus souvent des espèces. Plus généralement, on nomme ces entités des taxons (ou taxa) ou des unités taxonomiques. Le terme topologie désigne l'arrangement de ces relations. Deux arbres sont topologiquement identiques si les taxons y sont regroupés de la même manière. Il est également possible de mesurer la distance (génétique par exemple) qui sépare ces taxons les uns des autres à l'aide de la longueur des branches qui les relient. On parle alors de distances d'arbre. Pour ce faire, il suffit de faire la somme de toutes les branches sur le chemin entre les deux unités taxonomiques (par exemple, entre les taxons B et D dans la Figure 1). Les nœuds, c'est-à-dire les points de bifurcations d'une branche, représentent ultimement le point de séparation entre deux taxons. L'ensemble des relations et des distances entre les taxons représente l'hypothèse phylogénétique ou la phylogénie.

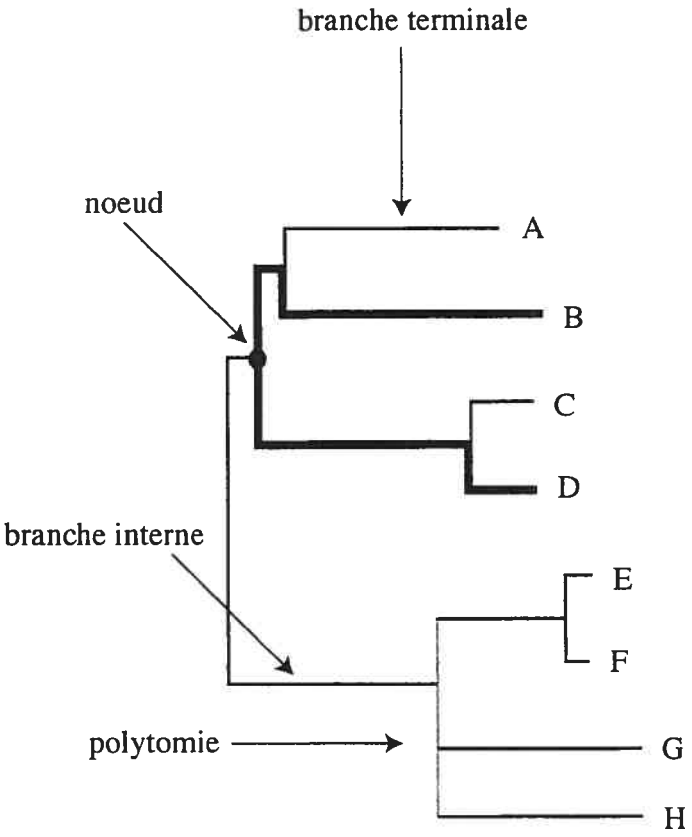

Figure 1. Arbre phylogénétique illustrant la terminologie utilisée

Un arbre est parfaitement résolu s'il y a bifurcation à chaque branche interne et n'est pas complètement résolu si une ou plusieurs de ces branches se séparent en trois branches ou plus. Cette séparation en plusieurs branches est souvent appelée une polytomie (voir Figure 1). Dans le cas où toutes les branches terminales se trouvent au même niveau, l'arbre n'a aucune résolution et l'on parle d'un buisson ou d'un arbre étoilé (voir Figure 2 A). Il existe deux façons d'interpréter les polytomies (Maddison 1989). Tout d'abord, on peut l'expliquer de la même manière que les bifurcations, c'est-à-dire par un événement de spéciation, qui est multiple au lieu de binaire. On parle dans ce cas de « vraies » polytomies (*hard polytomies*). Deuxièmement, il est possible que les données utilisées ne permettent pas de déterminer la relation entre deux taxons et, pour cette raison, que ces derniers soient représentés au même niveau. Ce sont les « fausses » polytomies (*soft polytomies*). Alors que les premières sont directement reliées à l'histoire phylogénétique des organismes étudiés, les dernières ne sont que le résultat d'un manque d'information et n'impliquent qu'une incertitude quant aux relations entre les taxons.

L'incongruence

Les termes congruence et incongruence sont utilisés de manière particulière en analyse phylogénétique. Suivant la définition qu'en fait Johnson & Soltis (1998), la congruence est un descripteur général de l'accord entre des arbres, des caractères ou des jeux de données. Dans cette thèse, je fais une utilisation plus stricte de ces deux termes où la congruence et l'incongruence se limitent à décrire l'accord ou le désaccord entre des phylogénies. Je définirai plus loin l'hétérogénéité, que j'utilise dans le cas des caractères et des jeux de données.

Lorsque deux arbres sont topologiquement identiques, ils sont congruents. Ils représentent les mêmes relations phylogénétiques. Des phylogénies peuvent également être congruentes sans être identiques. Dans le cas où un des arbres (ou les deux) présente une ou plusieurs polytomies, il y aura congruence si les parties résolues des deux arbres sont identiques. On dit généralement que ces arbres sont compatibles, c'est-à-dire que les relations de l'un ne contredisent pas les relations de l'autre. Par exemple, dans la Figure 2, les arbres A et B sont compatibles avec l'arbre C. Lors de l'analyse de plusieurs gènes ou

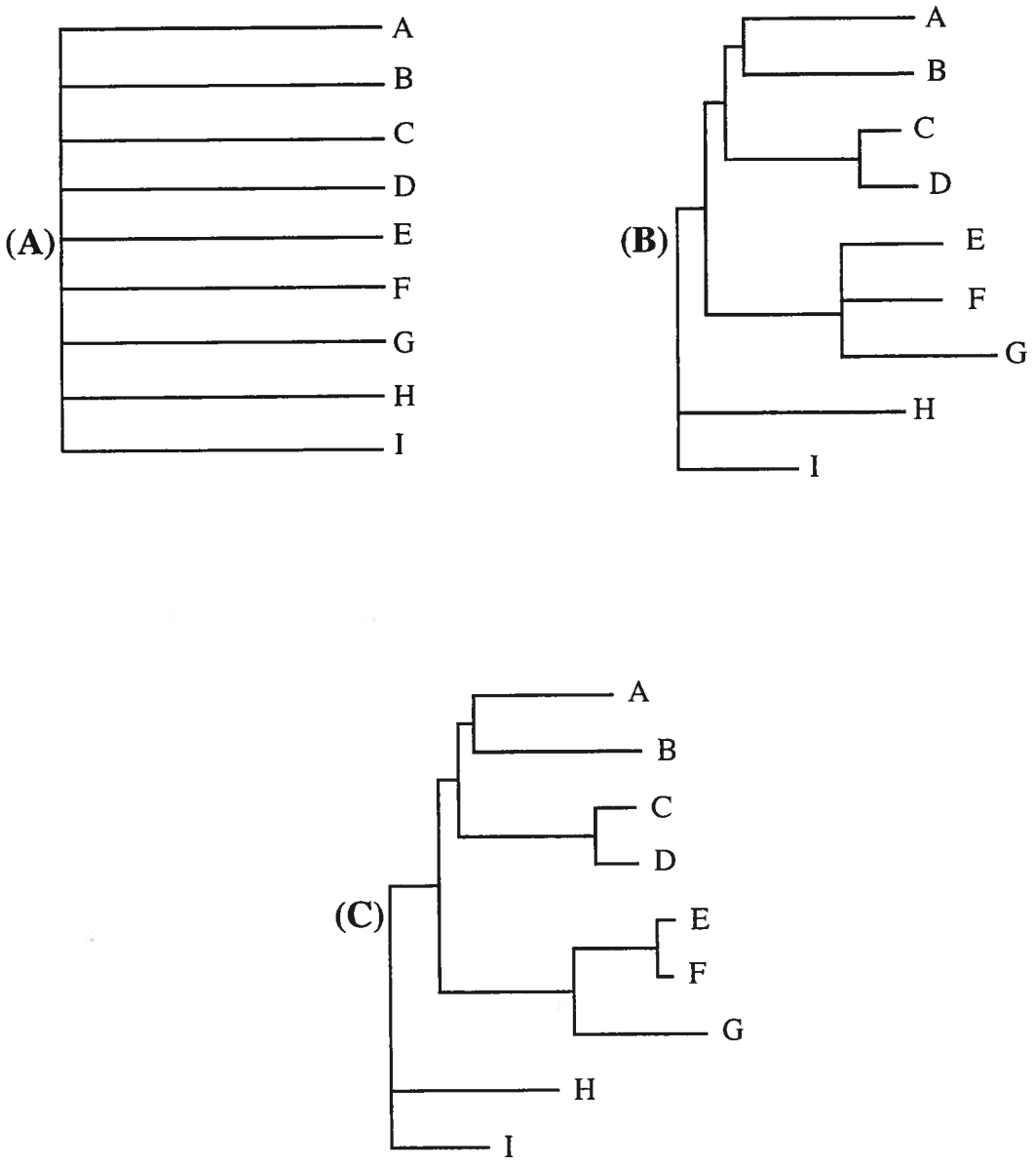

Figure 2. Représentation d'un arbre irrésolu (A), partiellement résolu (B) et complètement résolu (C). L'arbre A est compatible avec les arbres B et C, et l'arbre B est compatible avec l'arbre C.

de différents types de données, il arrive rarement que les arbres phylogénétiques estimés soient congruents. L'incongruence est plutôt la norme.

POURQUOI Y A-T-IL INCONGRUENCE?

Toutes les données mesurées sur des organismes ont la même histoire évolutive : celle des taxons. Comment est-il possible alors que l'analyse de plusieurs gènes, par exemple, nous mène à des hypothèses phylogénétiques différentes, où les relations évolutives entre les taxons ne sont pas les mêmes? Il est maintenant reconnu que différents types de données comme les gènes ou les données morphologiques d'un même groupe d'organismes peuvent avoir des patrons d'évolution différents. Plusieurs phénomènes peuvent expliquer l'incongruence. Wendel & Doyle (1998) donnent une description exhaustive des différentes sources d'incongruence possible. Ils invoquent d'abord des causes techniques comme le manque de données, le choix d'un gène évoluant trop vite ou trop lentement pour le niveau taxonomique étudié, un nombre insuffisant de nucléotides, des erreurs de séquençage ou un mauvais échantillonnage taxonomique. Ils suggèrent également des causes qui résultent de processus d'évolution au niveau des organismes (évolution morphologique convergente, diversification rapide, hybridation et introgression, transfert horizontal) et au niveau des gènes et du génome (recombinaison intragénique, interactions entre loci et évolution concertée, taux d'évolution hétérogène entre les taxons ou entre les gènes, nonindépendance des sites). Quoiqu'il en soit, l'incongruence est une réalité lorsque plusieurs sources d'informations sont utilisées pour reconstruire l'histoire phylogénétique et cette situation peut être vue comme un phénomène désirable permettant de comprendre les différents mécanismes sous-jacents à l'évolution des organismes (Wendel & Doyle 1998).

QUE FAIRE? LE DÉBAT

L'incongruence entre les phylogénies dérivées des données morphologiques et moléculaires a ouvert la porte à un vigoureux débat qui suscite encore aujourd'hui diverses réactions dans la communauté scientifique. Étant donné que (1) il est préférable de combiner différents types de données (moléculaires, morphologiques, etc.) mais aussi de multiples jeux de données d'un même type (par exemple, plusieurs gènes), et que (2) ces

différents jeux de données, pris séparément sont susceptibles de produirent des arbres incongruents, quelle est la meilleure manière d'analyser ces données? Trois approches contradictoires ont alimenté le débat. Suivant le principe de non-spécificité de Adanson (Sneath & Sokal 1973) selon lequel il est préférable d'utiliser un grand nombre de caractères, Kluge (1989) présenta le principe de *total evidence* : tous les caractères, quel que soit le type de données, devraient être utilisés simultanément dans une même analyse pour produire une hypothèse phylogénétique globale. C'est ce qu'on appelle aussi la congruence des caractères. À l'opposé, l'approche de congruence taxonomique (*sensu* Mickevich 1978) repose sur l'indépendance des divers types de données et propose que l'analyse des caractères provenant de sources différentes (morphologiques, différents gènes) soient la base de phylogénies séparées qui pourront par la suite être combinées à l'aide de méthodes de consensus. L'approche conditionnelle (Bull *et al.* 1993; De Queiroz 1993; Miyamoto & Fitch 1995; Rodrigo *et al.* 1993), quant à elle, propose une stratégie modérée impliquant les deux précédentes. En présence de données hétérogènes, c'est-à-dire dont les différences entre les partitions ne peuvent pas être attribuées qu'à une erreur d'échantillonnage, les données devraient être traitées suivant les principes de la congruence taxonomique. Dans le cas inverse, les données pourront servir à une analyse simultanée comme le prescrit la congruence des caractères.

Congruence des caractères

Les termes congruence des caractères, analyse simultanée ou combinée et l'anglicisme *total evidence* sont tous des synonymes qui ont été utilisés à un moment ou l'autre dans la littérature. Ces locutions réfèrent toutes à la même approche selon laquelle tous les caractères, peu importe leur type, sont analysés ensemble, c'est-à-dire dans une seule et même matrice, pour reconstruire une phylogénie. Invoqué par Kluge (1989), le principal argument pour justifier cette approche est celui de l'évidence totale (d'où le terme *total evidence*) proposé par Carnap (1950). L'hypothèse phylogénétique doit être estimée à partir de toutes les sources de données disponibles et ces dernières doivent nécessairement être combinées dans une même analyse. Ceci aurait pour effet de maximiser le pouvoir explicatif des caractères (Barrett *et al.* 1991; Eernisse & Kluge 1993; Jones *et al.* 1993; Kluge 1989; Kluge & Wolf 1993).

Alors que l'approche présentée par Kluge (1989) doit nécessairement faire appel à un algorithme de parcimonie (une des nombreuses méthodes d'analyse phylogénétique), il est aujourd'hui admis que différentes méthodes d'analyse phylogénétique peuvent être utilisées dans le cadre d'une analyse de congruence des caractères. D'ailleurs, certains auteurs ont proposé des phylogénies obtenues à l'aide d'analyses simultanées avec des méthodes de distances (Lapointe *et al.* 1999) et de maximum de vraisemblance (Hasegawa *et al.* 1997; Sallum *et al.* 2002).

Congruence taxonomique

Telle que définie par Mickevich (1978), la congruence taxonomique mesure la similarité des relations phylogénétiques entre plusieurs phylogénies. Puisque les jeux de données sont analysés séparément, on peut soit mesurer la congruence entre les arbres issus des diverses sources de données à l'aide de différents indices (voir Colless 1980; Mickevich 1978; Mickevich & Platnick 1989; Rohlf 1982) ou tenter de les combiner. Les méthodes de consensus représentent une des nombreuses façons de synthétiser les hypothèses proposées par ces analyses séparées et elles représentent généralement l'option la plus utilisée.

Méthodes de consensus

Plusieurs méthodes de consensus sont largement utilisées, principalement parce qu'elles sont accessibles dans plusieurs logiciels, dont PAUP* (Swofford 1999), qui est un des plus utilisés en analyse phylogénétique. Pour les besoins de cette thèse, je présente seulement les méthodes disponibles dans ce logiciel (pour une revue extensive des méthodes de consensus existantes, voir (Bryant 2003). Ces techniques synthétisent de manière différente le type d'information (groupement, groupe monophylétique) et le degré d'accord entre les arbres initiaux (strict, majoritaire) (Page 1992; Wilkinson 1994). Alors que certaines nécessitent qu'un clade (groupe monophylétique) soit présent dans tous les arbres initiaux (consensus strict; Sokal & Rohlf 1981), d'autres permettent d'inclure un groupe qui est présent dans un seul de ces arbres (consensus semi-strict; Bremer 1990). Il est à noter que les quatre méthodes présentées ci-après sont des méthodes de consensus topologique, c'est-à-dire qu'elles résument l'information sur les relations entre les taxons,

sans utiliser la longueur des branches (qui renseigne sur la distance entre ceux-ci). La Figure 3 montre un exemple où trois arbres (A, B, C) sont combinés à l'aide de ces différentes méthodes de consensus (D, E, F, G). Il est intéressant de noter que les quatre méthodes produisent des solutions différentes.

consensus d'Adams

Adams (1972) fut le premier à proposer une méthode pour combiner dans un seul arbre l'information contenue dans plusieurs (Swofford 1991). C'est une méthode d'intersection qui tient compte des emboîtements (*nestings*). On dit qu'un groupe est emboîté dans un plus grand lorsque le plus récent ancêtre commun du plus petit groupe est un descendant du plus récent ancêtre commun du plus grand groupe. Par exemple, dans l'arbre présenté dans la Figure 3C, {A, F} est emboîté dans le groupe {A, D, E, F}. Suivant cette définition, on ne doit pas interpréter les groupes présents dans l'arbre consensus comme étant des groupes monophylétiques. Le consensus d'Adams préserve les emboîtements communs à tous les arbres initiaux. La figure 4 illustre comment obtenir le consensus de Adams pour les 3 arbres présentées à la Figure 3. Après avoir fait la liste des groupes retrouvés dans chaque arbre à combiner, on retient les ensembles de taxons qui sont toujours groupés ensemble, seuls ou au sein d'un plus grand groupe. Dans cet exemple, ce sont les groupes {A, F}, {D, E} et {B, C}. Évidemment la méthode se complexifie à mesure que le nombre de taxons augmente. Cette méthode a été critiquée parce qu'elle peut produire des groupes qui ne sont retrouvés dans aucun des arbres initialement combinés (Rohlf 1982; Rohlf *et al.* 1983; Sokal & Rohlf 1981), ce qui complique l'interprétation de l'arbre consensus.

consensus strict

Cette méthode, présentée par Sokal & Rohlf (1981) est probablement la plus simple et la plus utilisée. Seuls les groupes monophylétiques retrouvés dans tous les arbres initiaux sont inclus dans la solution. C'est le type de méthode qui produit l'arbre consensus le plus conservateur pour un ensemble d'arbres donné. Il permet de mettre en relief les relations phylogénétiques communes à *tous* les arbres comparés. En contrepartie, les arbres consensus strict sont souvent peu résolus (voir Figure 3D) lorsque les phylogénies

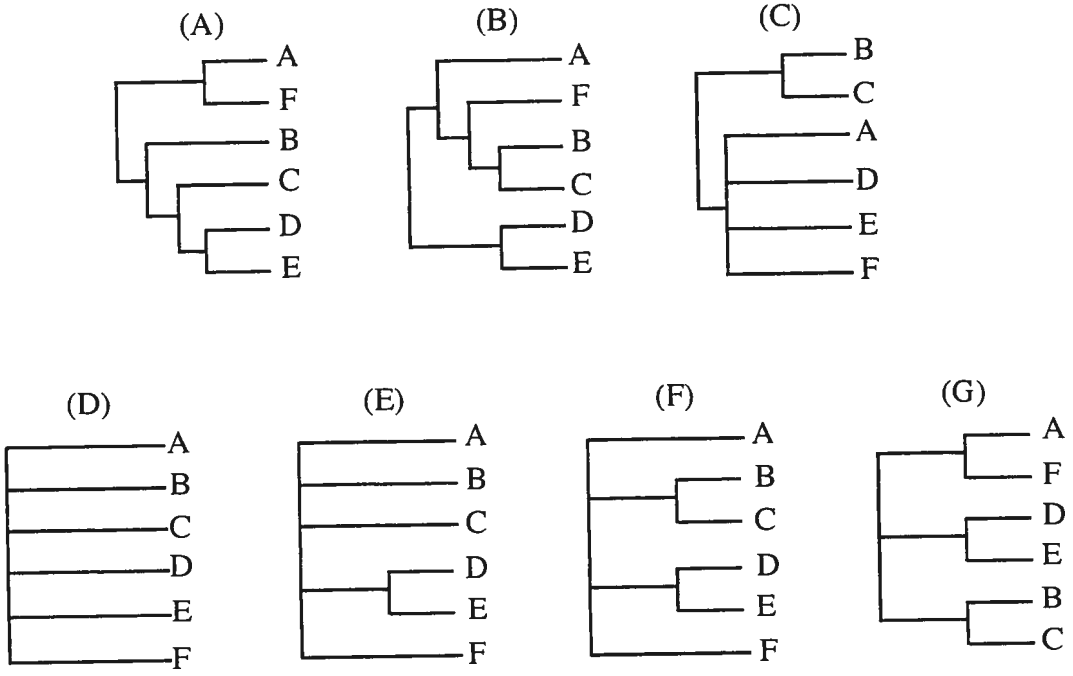

Figure 3. Résultats de la combinaison de trois arbres (A,B,C) à l'aide des méthodes de consensus strict (D), semi-strict (E), majoritaire (F) et Adams (G).

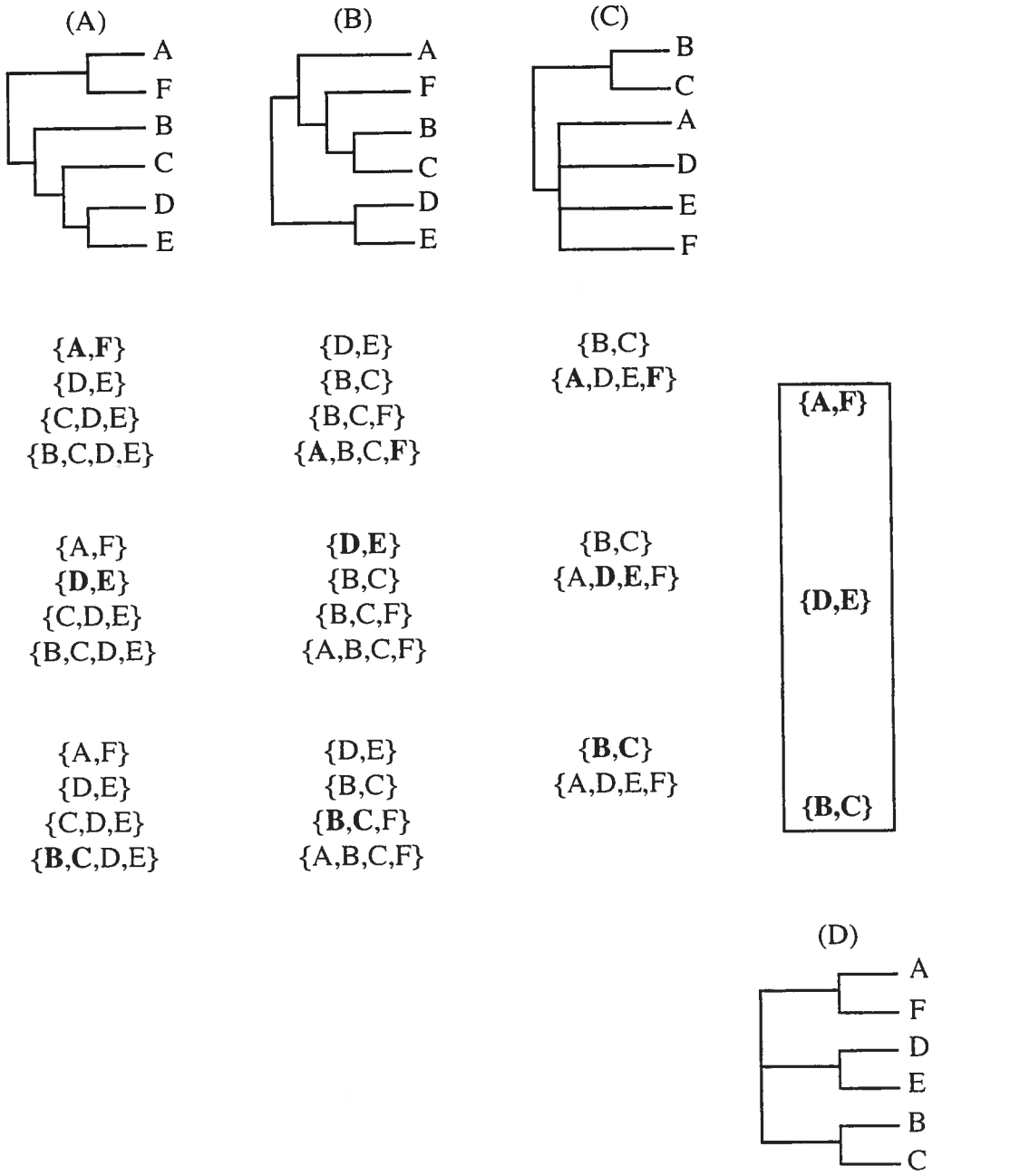

Figure 4. Exemple illustrant la combinaison de trois arbres (A,B,C) à l'aide de la méthode de consensus de Adams.

combinées sont très différentes ou lorsque la position d'un des taxons est instable entre les différents arbres (Swofford 1991).

consensus semi-strict

Connu sous le nom de consensus semi-strict, le *combinable component* de Bremer (1990) est une version plus libérale du consensus strict. Dans le cas où certains des arbres initiaux présentent des polytomies, il est possible que des groupes retrouvés seulement dans quelques arbres ne soient pas contradictoires avec les groupes présents dans tous les arbres. Par exemple, dans la Figure 3, le groupe {D, E} est présent dans deux des trois arbres initiaux. Celui-ci n'est pas inclus dans le consensus strict, mais il fait partie du consensus semi-strict. Ce groupe est compatible avec le dernier arbre, même s'il n'y est pas représenté, puisque le groupe {D, E} est inclus dans le groupe {A, D, E, F}. Notons que dans le cas où les arbres comparés sont dichotomiques (ou complètement résolus, c'est-à-dire où toutes les branches internes se séparent en deux), les résultats des consensus strict et semi-strict sont identiques.

consensus majoritaire

Le consensus majoritaire (Margush & McMorris 1981) est une solution de rechange intéressante au consensus strict lors de la comparaison de plus de deux arbres. Dans pareil cas, ce dernier a plus de chances de produire un arbre peu résolu si la topologie des phylogénies comparées diffère. Les groupes monophylétiques trouvés dans plus de 50% des arbres initiaux sont inclus dans l'arbre consensus majoritaire. Cette méthode donne une représentation intéressante du signal phylogénétique qui est commun à la majorité des jeux de données. Notons que lorsque seulement deux phylogénies sont comparées, les consensus strict et majoritaire présenteront nécessairement la même solution. Il est possible d'obtenir une version plus résolue du consensus majoritaire. Pour ce faire, on pourra inclure dans l'arbre consensus des groupes qui permettraient d'augmenter la résolution des portions non résolues de l'arbre consensus, et ce, même si ces groupes ne sont pas présents dans la majorité des arbres initiaux. Cependant, ces groupes supplémentaires ne devront en aucun cas être contradictoires avec ceux qui sont présents dans la majorité des phylogénies.

Interprétation du consensus

Alors que les arbres de consensus sont utilisées comme façon de synthétiser les résultats d'analyses séparées, la manière d'interpréter ces arbres est encore sujette à de nombreuses discussions. Il est clair que les méthodes de consensus représentent une bonne façon de résumer l'information contenue dans les arbres estimés à l'aide de différents jeux de données en plus de montrer leurs zones d'accord et de désaccord. Par contre, plusieurs auteurs (Bremer 1990; Hillis 1987; Miyamoto 1985; Swofford 1991) croient que les arbres consensus ne devraient pas être interprétés comme des hypothèses phylogénétiques, mais plutôt comme des outils pour mesurer la congruence entre les arbres. D'autre part, il semble que dans certains cas particuliers, il peut être acceptable de considérer l'arbre consensus comme une phylogénie (De Queiroz 1993; Miyamoto & Fitch 1995; Wilkinson 1994).

Résolution

Le manque de résolution des arbres obtenus à l'aide de certaines méthodes de consensus rebute plusieurs chercheurs (Barrett *et al.* 1991; De Queiroz *et al.* 1995; Hillis 1987; Kluge & Wolf 1993; Nixon & Carpenter 1996). Il est vrai que les méthodes classiques produisent des arbres qui sont parfois peu résolus, surtout dans les cas où les phylogénies combinées sont différentes. Le débat est encore ouvert sur la pertinence d'obtenir une phylogénie complètement résolue. Deux opinions s'affrontent sur le terrain : préférer une solution moins résolue avec une plus grande probabilité de contenir la vérité (Swofford 1991) ou préférer une phylogénie complètement résolue qui contient plus d'information (puisque dans le cas extrême où il n'y a aucune résolution, la phylogénie est compatible avec n'importe quel autre arbre), et ce, même si elle est partiellement fausse (Kluge & Wolf 1993).

Choix de la méthode

Comme je l'ai illustré précédemment, les méthodes de consensus peuvent produire des résultats différents pour un même ensemble d'arbres. Pour Eernisse & Kluge (1993), Jones *et al.* (1993) et Kluge & Wolf (1993), il n'existe aucune justification claire sur le choix d'une méthode de consensus, à part les raisons techniques de préférer les résultats d'une méthode à ceux d'une autre. Pour d'autres (Page 1992; Swofford 1991;

Wilkinson 1994), ceci représente plutôt un avantage puisque chacune des méthodes résume différents types d'information (groupes monophylétiques, emboîtements) contenue dans les phylogénies et mesure différents niveaux d'accord (strict, majoritaire) entre ces derniers. Chaque méthode peut donc être utilisée dans des contextes différents. En fait, il est important de souligner que ce genre de choix se pose régulièrement en analyse phylogénétique, où il existe plusieurs méthodes de reconstruction phylogénétique, plusieurs indices de comparaison d'arbres, plusieurs types de distances et de modèles évolutifs. Ce type de décision peut se baser sur des considérations philosophiques, méthodologiques, pratiques ou contextuelles.

Approche conditionnelle

L'approche conditionnelle est la plus modérée des trois, les deux précédentes étant plutôt radicales. Pour plusieurs auteurs (Bull *et al.* 1993; Miyamoto & Fitch 1995; Rodrigo *et al.* 1993), une approche ne peut certainement pas être meilleure que les autres dans tous les cas. Il existe des circonstances où l'utilisation de la congruence des caractères est plus appropriée, et d'autres où l'analyse séparée devrait être préférée. Ces chercheurs ont proposé que l'analyse simultanée de données fortement hétérogènes pourrait brouiller le signal phylogénétique au point de produire une hypothèse phylogénétique qui ne serait pas en accord avec les caractères sous-jacents.

Le choix quant à l'approche à adopter se fait sur la base de tests mesurant le degré d'hétérogénéité des données (Farris *et al.* 1995a; Larson 1994; Legendre & Lapointe 2004; Mickevich & Farris 1981; Rodrigo *et al.* 1993; Swofford 1991). Des jeux de données hétérogènes devront être analysés séparément et les partitions homogènes devront faire l'objet d'une analyse simultanée.

La principale critique formulée à l'égard de cette approche concerne l'utilisation de ces tests d'hétérogénéité, qui est essentielle à son application. Le comportement des différents tests a été peu investigué et il est difficile de juger si le degré d'hétérogénéité justifie le choix d'analyser séparément ou simultanément les données. Dans des situations où l'hétérogénéité est rare, la plupart des cas où les tests auront détecté des données hétérogènes seront de faux positifs (Huelsenbeck *et al.* 1996). Ceci implique que des

données homogènes sont analysées séparément alors qu'elles devraient être combinées. De plus, la plupart de ces tests ne peuvent comparer les jeux de données que deux à deux et ils ne permettent pas de comparer les données présentées sous forme de distances (sauf la méthode CADM, Legendre & Lapointe 2004).

Arguments

Il existe plusieurs arguments pour et contre chacune des approches présentées précédemment. Des dizaines d'articles ont été publiés sur le sujet au cours des deux dernières décennies. Certains auteurs ont émis des arguments constructifs, alors que d'autres en ont profité pour invectiver violemment les partisans de l'autre approche. Quoiqu'il en soit, certaines justifications restent pertinentes et sont présentées dans cette section.

Pour la congruence des caractères

Principe philosophique

L'utilisation de toutes les sources d'information dans une même analyse pour maximiser leur pouvoir explicatif est à la base de l'approche de congruence des caractères. C'est le premier argument invoqué pour justifier cette approche. En effet, les données ainsi combinées contribuent au signal phylogénétique de manière à faire ressortir des relations phylogénétiques qui ne seraient pas présentent dans les arbres résultant d'analyses séparées (Barrett *et al.* 1991; Eernisse & Kluge 1993; Jones *et al.* 1993; Kluge 1989; Kluge & Wolf 1993).

Nombre de caractères

Le grand nombre de caractères que permet l'analyse combinée est le deuxième argument pour la justification de cette approche. En effet, lorsqu'une méthode d'analyse phylogénétique est consistante, l'addition de caractères permet d'augmenter la probabilité de converger vers la « vraie » phylogénie (Adanson 1763; De Queiroz 1993; Huelsenbeck & Hillis 1993). Ultimement, l'analyse d'un nombre infini de caractères garantirait de trouver la vraie histoire évolutive du groupe étudié. En contrepartie, l'augmentation du

nombre de caractères aura l'effet inverse lors de l'utilisation d'une méthode inconsistante, pour laquelle les chances d'obtenir un arbre différent de la « vraie » phylogénie augmenterait (*positively misleading*, Felsenstein 1978). De plus, il a été proposé que la combinaison de caractères hétérogènes pourrait produire du bruit capable de masquer le signal phylogénétique (Bull *et al.* 1993). Dans le pire des cas, le résultat de l'analyse combinée peut mener à une fausse hypothèse phylogénétique, et ce, même si un des jeux de données appuie la bonne phylogénie. Dans ces circonstances, il vaudrait mieux obtenir des estimations incompatibles d'analyses séparées plutôt qu'une seule estimation incorrecte d'une analyse combinée.

Pour la congruence taxonomique

Indépendance des caractères

Argument central dans la justification des méthodes de consensus, l'indépendance des caractères est à la base du principe de corroboration. La probabilité que des caractères d'un même jeu de données soient des estimateurs non indépendants est plus grande que pour des caractères de partitions différentes. Ceci implique qu'il y a plus de chances que des caractères d'un même jeu de données appuient la même hypothèse phylogénétique, et ce, même si cette dernière s'avère être fausse (De Queiroz 1993).

Il est plus parcimonieux d'accepter une explication commune plutôt que de poser l'hypothèse que plusieurs arbres semblables sont issus d'événements indépendants (Miyamoto & Cracraft 1991). On peut penser que la meilleure estimation des relations phylogénétiques entre les organismes étudiés correspond à l'hypothèse appuyée par plusieurs sources indépendantes. C'est pourquoi une attention spéciale doit être portée à l'accord entre les phylogénies issues de différents jeux de données (De Queiroz 1993; Lanyon 1993; Swofford 1991).

Une vive critique a été formulée envers la séparation des caractères en jeux de données différents. Loin de penser que ces partitions forment des sources d'information indépendantes, plusieurs auteurs (Eernisse & Kluge 1993; Jones *et al.* 1993; Kluge & Wolf, 1993) ont proposé que cette division des caractères en diverses classes de données est purement subjective. Selon eux, les types de données reconnus par la majorité des

phylogénéticiens ne sont que des artéfacts de la tradition et de la technologie. En réponse à cette critique, il a été proposé que les différents modèles évolutifs sous-jacents à ces classes de données permettent de croire qu'il existe bel et bien plusieurs types de données qui devraient être analysés séparément (Bull *et al.* 1993; De Queiroz 1993; Doyle 1992). Dans certains cas, il a été proposé que ces partitions soient analysées en utilisant différents modèles évolutifs. Dans ces circonstances, les propriétés distinctes de chacune des classes de données augmentent les probabilités que l'accord entre les différentes phylogénies soit le résultat de la vraie histoire évolutive (Lanyon 1993; Miyamoto & Cracraft 1991; Sheldon & Bledsoe 1993). Par exemple, certaines catégories de données ont été proposées pour séparer les données moléculaires : premier et deuxième codon vs troisième codon, région codante vs région non-codante, gènes évoluant à des taux différents (Bull *et al.* 1991).

Déceler l'incongruence

L'analyse séparée permet de bien cerner le signal phylogénétique unique à chaque jeu de données. Il est ainsi plus facile de détecter les zones d'accord et de désaccord entre les différentes partitions. Si les phylogénies obtenues ne sont pas similaires, ceci peut aider à orienter le chercheur vers la source de l'incongruence entre les arbres. Dans cette optique, l'analyse séparée est un outil qui permet de mieux visualiser les relations phylogénétiques appuyées par les différentes sources d'information, et de mieux comprendre les facteurs qui influencent l'évolution du groupe étudié.

Pondération des caractères

Cet argument prend toute sa valeur lorsque les jeux de données à combiner sont de tailles radicalement différentes comme dans le cas de données moléculaires et morphologiques. Alors que les techniques moléculaires modernes sont très performantes, la récolte des caractères morphologiques reste souvent plus modeste. D'une part, le nombre de caractères morphologiques est limité et la cueillette peut s'avérer longue puisque chaque caractère doit être mesuré séparément. Dans le cas des données moléculaires, le séquençage d'un gène permet d'obtenir des centaines, voire des milliers de caractères à la fois. La pertinence d'utiliser ces deux types de données a été discutée ailleurs et ne sera pas abordée (Donoghue & Sanderson 1992 ; Doyle 1992 ; Patterson *et al.* 1993 ; Systma 1990). Par

contre, dans un cadre où tous les caractères, morphologiques et moléculaires, sont analysés ensemble, le signal phylogénétique du plus petit jeu de données risque d'être noyé dans celui du plus grand (Kluge 1983; Larson 1994).

Évidemment, il est toujours possible d'effectuer une analyse pondérée pour éviter une telle conséquence. La pondération des caractères est une préoccupation importante dans le débat opposant les approches de congruence des caractères et de congruence taxonomique. L'utilisation des méthodes de consensus pour combiner les données constitue, selon Cracraft & Mindell (1989), une pondération différentielle arbitraire. Ils croient que cette façon de faire est une solution facile qui écarte le besoin de justifier un mode de pondération particulier. Donner un poids égal aux différents jeux de données, comme le fait une analyse séparée, revient à attribuer à chaque caractère un poids qui est tributaire de la méthode utilisée, ce qui n'est pas phylogénétiquement défendable. Ces auteurs ne considèrent donc pas l'argument de la pondération comme une justification pertinente de l'utilisation des méthodes de consensus pour combiner les données.

De plus, Donoghue & Sanderson (1992) et Eernisse & Kluge (1993) prétendent que ce n'est pas le nombre de caractères qui génère le signal phylogénétique mais plutôt l'information qui y est contenue. Selon eux, l'analyse combinée de jeux de données de tailles différentes n'est pas problématique et cet argument ne devrait pas être utilisé pour soutenir l'approche de congruence taxonomique.

Combiner tous les types de données

L'approche de congruence taxonomique permet de combiner n'importe quel type de données. Chaque partition étant traitée séparément, il est possible de choisir une méthode d'analyse phylogénétique qui convient à chacun des types de données. Par exemple, dans le cas des méthodes de maximum de vraisemblance, il est possible d'utiliser plusieurs modèles pour des gènes qui évoluent à des rythmes différents. De plus, il n'est pas possible d'analyser dans une même matrice des données comme des caractères de séquence moléculaire et une matrice d'hybridation d'ADN (Barrett *et al.* 1991; Eernisse & Kluge 1993; Lanyon 1993). Une des solutions serait de convertir toutes les données en distances, mais l'option la plus souvent proposée est de faire des analyses séparées et de combiner les arbres à l'aide d'une méthode de consensus.

La congruence globale

Une décennie après le début d'un débat stagnant, Lapointe *et al.* (1999) émettent une idée nouvelle. Leur argumentation remet en question la supériorité d'une approche par rapport à l'autre. Comme nous l'avons vu précédemment, puisque les méthodes de consensus généralement utilisées ne tiennent compte que de la topologie, ces techniques produisent le plus souvent des arbres peu résolus. Il a été démontré que les arbres issus de méthodes de consensus qui tiennent compte des longueurs de branches (par exemple, le consensus moyen; Lapointe & Cucumel 1997) sont plus résolus que ceux des méthodes de consensus strictement topologiques (Lapointe 1998a). Par conséquent, les résultats des approches de congruence des caractères et de congruence taxonomique pourraient être comparables, voire identiques, lorsqu'une méthode de consensus qui tient compte des longueurs de branches est utilisée.

L'idée d'effectuer des analyses séparées et simultanées a plusieurs fois été proposée dans la littérature (De Queiroz 1993; De Queiroz *et al.* 1995; Larson 1994; O'Grady *et al.* 2002). En effet, l'utilisation des deux approches permet de contraster l'information en commun et la force relative des jeux de données (Hillis 1987). De même, l'utilisation conjointe des approches de congruence des caractères et de congruence taxonomique est à la base de l'approche de congruence globale (*sensu* Lapointe 1998b). D'abord proposée pour des méthodes de distances, cette approche peut être utilisée même dans les cas où les données à combiner se présentent sous forme de caractères. En effet, comme chaque jeu de données est converti en distances, il est possible de combiner toutes les données dans une même analyse. De plus, la combinaison des phylogénies issues des différentes partitions est effectuée à l'aide d'une méthode de consensus qui tient compte des longueurs de branches, notamment le consensus moyen, et permet donc l'obtention d'arbres plus résolus.

Le consensus moyen a été défini pour des arbres estimés par des méthodes de distances, mais peut aussi être utilisé avec des phylogénies qui résultent d'autres types de méthodes d'analyse phylogénétique (parcimonie, maximum de vraisemblance). Cette méthode tient compte des longueurs de branches, mais on peut également ignorer ces distances pour obtenir un consensus moyen topologique. Ce type de consensus minimise la distance entre l'arbre consensus et tous les arbres initiaux. Il se calcule aisément en deux étapes. D'abord, une matrice est calculée en faisant la moyenne de toutes les matrices de

distances d'arbres associées aux phylogénies à combiner (programmes disponibles sur demande). Un algorithme des moindres carrés (Cavalli-Sforza & Edwards 1967; De Soete 1983; Fitch & Margoliash 1967; Makarenkov & Leclerc 1999) est ensuite appliqué à cette matrice de distances pour obtenir le consensus moyen. Il est également possible de calculer un consensus moyen en utilisant d'autres fonctions comme la médiane, par exemple (Levasseur & Lapointe 2002).

La congruence globale est donc une approche qui permet de répondre à certains problèmes des approches de congruence des caractères et de congruence taxonomique, et elle ne fait pas appel aux tests d'hétérogénéité des données qui sont fortement critiqués. Cette procédure ne mesure ni la congruence des caractères ni la congruence taxonomique, mais évalue la congruence entre les arbres issus de ces approches. Dans tous les cas, les deux types d'analyse sont effectués et les résultats sont comparés. Lorsque les arbres obtenus sont semblables ou identiques, on peut s'attendre à ce que la solution soit juste, c'est-à-dire qu'elle reflète bien les données sous-jacentes et donc qu'elle représente une estimation probable de la « vraie » phylogénie. En effet, on peut postuler dans ce cas que les relations phylogénétiques illustrées par les arbres issus des deux approches sont bien le reflet de l'information contenue dans les données et non pas un artéfact technique de la méthode utilisée. Dans les situations où les arbres ne sont pas identiques, il est possible de penser que les clades qui diffèrent sont moins bien supportés par les données. Des méthodes de ré-échantillonnage statistique (par exemple, le bootstrap; Felsenstein 1985) permettent d'évaluer ce type de support. Quoiqu'il en soit, l'approche de congruence globale est une option nouvelle qui doit encore être investiguée mais qui mérite d'être considérée.

Certains ont proposé que l'approche de congruence globale est sujette à un problème de non-indépendance entre les analyses de congruence des caractères et congruence taxonomique. En effet, il ne serait pas possible de comparer et de tester statistiquement les résultats de ces deux approches puisque nous serions face à un problème tautologique. Par contre, l'utilisation conjointe de ces deux approches complémentaires pourrait permettre d'améliorer la justesse des arbres obtenus par rapport aux résultats de l'une ou l'autre de ces méthodes utilisées de façon indépendante. La même logique est utilisée par Kim (1993) lorsqu'il compare la justesse des résultats obtenus à l'aide de

différentes méthodes d'analyse phylogénétique de manière individuelle et combinée. Le but est d'obtenir la meilleure estimation phylogénétique possible. Alors que certains auteurs préfèrent utiliser la parcimonie, d'autres opteront pour les méthodes de distances, de maximum de vraisemblance ou l'approche Bayesienne, mais tous sont satisfaits lorsque ces approches indépendantes appliquées aux mêmes données convergent vers la même solution. D'ailleurs, Kim (1993) a montré que dans ce cas, la phylogénie obtenue est souvent meilleure. La philosophie derrière la congruence globale est donc de ne pas favoriser une méthode de combinaison des données par rapport à une autre, mais de comparer les solutions de ces approches. Lorsque la phylogénie obtenue est la même pour les différentes approches, la probabilité que cet arbre soit correct devrait donc être plus grande.

LES SUPER-ARBRES

Les approches présentées précédemment ont initialement été proposées dans un cadre où le même ensemble de taxons est représenté par tous les jeux de données à combiner. Dans une optique où l'utilisation de plusieurs partitions pourrait permettre une meilleure compréhension de l'histoire phylogénétique d'un groupe d'organismes, diverses équipes de recherche travaillent à récolter plusieurs jeux de données sur un grand nombre d'espèces. À des fins plus importantes, celle de la reconstruction de l'Arbre de la vie (*Tree of life*) par exemple, la combinaison des données de plusieurs équipes est très intéressante. D'abord parce qu'elle optimise les efforts de recherche par l'élimination du travail en double. Ensuite parce qu'elle permet d'utiliser les données déjà publiées. Mais comme les différentes études n'ont pas toujours les mêmes objectifs, les ensembles d'espèces ciblées ne sont pas identiques d'une étude à l'autre. Quand vient le temps de combiner ces données, on s'aperçoit qu'il y a des trous d'échantillonnage et il arrive qu'il n'y ait qu'un chevauchement partiel entre les taxons représentés par les différentes partitions. Les méthodes de consensus ne permettent pas de combiner de tels arbres et on doit alors avoir recours à des méthodes de super-arbres.

Le même débat entre congruence des caractères et congruence taxonomique existe dans ce cas particulier. La question est de savoir s'il est préférable de combiner les données à l'aide d'une supermatrice et de procéder à une seule analyse ou de combiner les arbres à

l'aide de méthodes de super-arbres (Bininda-Emonds 2004; Bininda-Emonds *et al.* 2003; Gatesy *et al.* 2004; Gatesy *et al.* 2002). Plusieurs méthodes de consensus ont d'ailleurs été adaptées pour répondre à cette nouvelle réalité (Constantinescu & Sankoff 1995; Goloboff & Pol 2002; Lanyon 1993; Lapointe & Cucumel 1997; Steel 1992; Steel *et al.* 2000) et plusieurs méthodes de reconstruction de super-arbres ont été proposées (Baum 1992; Chen *et al.* 2003; Gordon 1986; Page 2002; Ragan 1992; Sanderson *et al.* 1998; Semple & Steel 2000; Slowinski & Page 1999). Les efforts sont présentement concentrés au développement de ces différentes méthodes. Leurs propriétés restent à ce jour peu connues et plusieurs problèmes méthodologiques et techniques restent encore à résoudre. Par exemple, le manque d'information pour certains taxons qui ne sont pas représentés par tous les jeux de données rend la reconstruction beaucoup plus difficile.

Parce qu'elles sont encore très récentes, la plupart des méthodes permettant de combiner des jeux de données ou des arbres qui ne présentent pas des taxons identiques ne sont pas encore disponibles dans les logiciels d'analyse phylogénétique. À ce jour, les seules que l'on peut utiliser dans PAUP*, par exemple, sont l'analyse de supermatrices et la méthode de MRP (*matrix representation with parsimony*; Baum 1992; Ragan 1992). Si la première permet d'analyser simultanément les données à la manière de la congruence des caractères, la deuxième permet la combinaison des phylogénies (en mode consensus, où tous les arbres présentent les mêmes taxons et en mode super-arbres, où il n'y a qu'un chevauchement partiel). Brièvement, cette technique permet un codage des nœuds présents dans chacun des arbres et donne une représentation de ces derniers sous forme de matrice de caractères, qui sera à son tour analysée à l'aide de la méthode de parcimonie. Parce qu'elle est disponible et que, contrairement à la méthode de supermatrices, elle ne nécessite pas l'accès aux caractères qui sont parfois difficilement accessibles dans un cadre où les données de plusieurs auteurs sont combinées, la méthode de MRP est de loin la plus utilisée pour la reconstruction de super-arbres (Bininda-Emonds *et al.* 1999; Grenyer & Purvis 2003; Jones *et al.* 2002; Kennedy & Page 2002; Salamin *et al.* 2002).

Comme la méthode de MRP combine les arbres et non les caractères, certains auteurs l'associent aux méthodes de consensus (Pisani & Wilkinson 2002). D'autres prétendent que cette ressemblance est superficielle (Bininda-Emonds & Bryant 1998) et que la méthode de MRP est plutôt reliée à l'approche de congruence des caractères

(Bininda-Emonds & Sanderson 2001). Récemment, Lapointe *et al.* (2003) ont montré que dans le cas particulier où les taxons sont identiques pour tous les arbres combinés, la méthode de MRP et le consensus moyen sont équivalents.

Le consensus moyen peut aussi être adapté dans un cadre de super-arbres (Lapointe & Cucumel 1997) et certaines études utilisant cette méthode ont déjà été publiées (Lapointe *et al.* 1999; Lapointe & Kirsch 2001; Levasseur *et al.* 2003). Lapointe *et al.* (1999) proposent également une approche de congruence globale où les arbres issus du consensus moyen et d'une analyse de supermatrice seraient utilisés conjointement pour augmenter la justesse des phylogénies obtenues. Étant donné qu'il existe peu d'études où des super-arbres ont été produits avec le consensus moyen, nous connaissons peu les propriétés de cette approche. Il est important de noter que c'est d'ailleurs le cas pour toutes les méthodes de super-arbres. Mais il est essentiel d'en connaître un peu plus avant de l'utiliser dans un cadre plus large comme la congruence globale. Par exemple, la combinaison des matrices de distances d'arbres permettra de calculer une matrice moyenne qui sera incomplète dans ce cas particulier, puisque les distances entre certains taxons seront inconnues. À cette étape, il est possible d'utiliser directement cette matrice ou d'estimer les données manquantes avant la reconstruction phylogénétique. Un choix qui pourrait probablement changer les résultats obtenus à l'aide de cette méthode (Levasseur *et al.* 2003).

Organisation de la thèse

Cette thèse présente quelques aspects de la problématique de la combinaison de différents jeux de données en analyse phylogénétique. À la lumière du débat impliquant les approches de congruence des caractères et de congruence taxonomique, je me suis intéressée à une autre solution qui vise l'utilisation conjointe de ces deux approches. Le débat est biaisé par le choix des méthodes de consensus utilisées puisqu'elles ne sont basées que sur la topologie des phylogénies initiales. Le choix d'une méthode de consensus qui tient compte des longueurs de branches (le consensus moyen) permet souvent d'obtenir des arbres semblables à ceux obtenus lors d'une analyse simultanée (Lapointe *et al.*, 1999). L'utilisation de ces deux méthodes dans une approche de congruence globale constitue donc une option attrayante dans le cadre de ce débat. Le corps de la thèse est divisé en six chapitres, organisés en deux volets.

Consensus

Ce premier volet comporte quatre chapitres sous forme d'articles scientifiques et explore la méthode de congruence globale dans le cas où les différents jeux de données comportent exactement les mêmes taxons. Le premier chapitre reprend l'analyse de différentes études présentées dans la littérature avec l'approche de congruence globale. Il vise à comparer les résultats obtenus à l'aide de certaines méthodes de consensus topologique et du consensus moyen à ceux de la congruence des caractères. J'y traite également de l'effet de la validation sur la compatibilité entre les arbres consensus et la phylogénie qui résulte de l'analyse combinée. Les deuxième et troisième chapitres présentent des études de simulations qui ont pour objectif de tester la justesse des approches de congruence des caractères, de congruence taxonomique et de congruence globale et de connaître dans quelle mesure certains paramètres comme l'hétérogénéité des données influent sur la qualité des solutions obtenues par rapport à un arbre modèle. J'espère ainsi vérifier l'hypothèse selon laquelle l'approche de congruence globale permet d'améliorer la qualité des estimations phylogénétiques. Le quatrième chapitre présente une étude de simulations comparant les résultats du consensus moyen, d'une méthode de reconstruction de super-arbres, le MRP, et ceux de l'approche de congruence des caractères dans un cadre de consensus. L'objectif de cet article est de déterminer si ces trois approches permettent d'obtenir des résultats semblables.

Super-arbres

Ce deuxième volet comporte deux chapitres et traite de méthodes d'analyse ou de cas où les différentes données à combiner sont représentées par des ensembles de taxons qui ne se chevauchent que partiellement. Le cinquième chapitre présente une étude de simulations pour tester la meilleure stratégie à adopter pour traiter les données manquantes lors de la combinaison de différentes partitions qui ne présentent pas des ensembles de taxons identiques. Un super-arbre des chauves-souris y est présenté comme exemple d'application de cette technique. Le dernier chapitre a pour objectif de vérifier la justesse des arbres obtenus à l'aide du consensus moyen (et, en conséquence, de la méthode d'estimation testée dans le chapitre précédent) dans le cas des super-arbres. Cette dernière

étude de simulation très succincte a été effectuée pour vérifier la possibilité de généraliser l'approche de la congruence globale à la reconstruction de super-arbres.

CHAPITRE 1

War and peace in phylogenetics : a rejoinder on total evidence and consensus

Cet article est publié sous la référence :

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

ABSTRACT

For more than ten years, systematists have been debating the superiority of character or taxonomic congruence in phylogenetic analysis. In this paper, we demonstrate that the competing approaches can converge to the same solution when a consensus method that accounts for branch lengths is selected. Thus, we propose to use both methods in combination, as a way to corroborate the results of combined and separate analyses. This so-called "global congruence" approach is tested with a wide variety of examples sampled from the literature, and the results are compared to standard consensus methods. Our analyses show that when the total evidence and consensus trees differ topologically, collapsing weakly supported nodes with low bootstrap support usually improves "global congruence".

INTRODUCTION

For more than a decade, phylogeneticists have been searching for ways to analyze the ever increasing amount of data (for review, see De Queiroz *et al.* 1995; Huelsenbeck *et al.* 1996). The same question has been raised time and time again: is it better to combine different data sets prior to phylogenetic reconstruction or not? With the recent advances in and increasing popularity of molecular systematics, this debate opposing *total evidence* (character congruence) to *consensus* (taxonomic congruence) approaches has become even more important (e.g., Bull *et al.* 1993; Chippindale & Wiens 1994). Supporters of character congruence (*sensu* Mickevich 1978) claim that all data should always be combined for phylogenetic analysis (Barrett *et al.* 1991; Kluge 1989; Kluge & Wolf 1993). On the other hand, proponents of taxonomic congruence (*sensu* Mickevich 1978) insist that independent data sets should be analyzed separately and combined by means of consensus techniques *a posteriori* (Bull *et al.* 1993; Huelsenbeck *et al.* 1994; Miyamoto & Fitch 1995; Swofford 1991). For others, these competing options are too radical and an intermediate solution has been proposed to decide whether or not to combine data, based on the results of statistical heterogeneity tests (Farris *et al.* 1995b; Huelsenbeck & Bull 1996; Mickevich & Farris 1981; Rodrigo *et al.* 1993).

Numerous studies have declared character congruence superior to consensus (see Barrett *et al.* 1991; De Queiroz 1993; Miyamoto 1985) as it usually provides trees that are more resolved than those obtained by taxonomic congruence. It has been suggested, however, that consensus methods which consider branch lengths (see Lapointe 1998a) could be more resolved than those based on topological relationships alone, including the strict (Sokal & Rohlf 1981) and majority rule (Margush & McMorris 1981) consensus. In particular, it has been shown that the average consensus procedure (Lapointe & Cucumel 1997) may be more likely than standard consensus methods to produce trees as resolved as those obtained from total evidence analysis (Lapointe 1998b; Lapointe *et al.* 1999).

We could engage in this debate by opting for character congruence, taxonomic congruence, or the conditional combination approach. Rather, we would like to suggest using combined and separate analyses jointly, as proposed by De Queiroz (1993; see also Larson 1994). Interestingly, a distance-based procedure relying on the average consensus

has been applied successfully by Lapointe *et al.* (1999) to combine either trees or data matrices in a coherent fashion. This hybrid procedure is defined as a *global congruence* approach (Lapointe 1998b) as it does not assess the congruence among characters, nor that among individual phylogenies; it evaluates the congruence *between* total evidence *and* consensus trees. This approach could thus be used to cross-corroborate the trees obtained by combined and separate analyses.

In the present paper, we apply the so-called global congruence approach to a wide variety of published data sets sampled from the systematic literature, using a uniform distance-based procedure (Lapointe *et al.* 1999). We postulate (1) that total evidence and consensus trees will be congruent when average consensus is used to combine the trees estimated separately from individual data sets. We also predict (2) that average consensus trees will be more similar to total evidence trees and more resolved than strict and majority rule consensus trees. When average consensus and total evidence trees differ, (3) we further claim that the discrepancies will not hold if the clades with low bootstrap support in the total evidence tree are collapsed.

MATERIAL AND METHODS

To test our hypotheses, a diversity of data sets evolving under different models and at different rates of evolution was required. To do so, we surveyed all papers published in *Systematic Biology* since Kluge's (1989) seminal paper and selected all those using multiple character sets. Our initial selection included 26 studies representing a wide variety of taxonomic groups and different types of characters, with number of data sets per study ranging from 2 to 17 and number of taxa ranging from 9 to 193. From that list, a secondary selection was made according to data availability, and we were finally able to obtain complete character sets from 15 distinct studies. However, since our objectives were quite different from those of the original papers, some data sets were modified prior to the analyses. In specific cases, the taxa that were not represented in all of the original data sets were deleted, thus reducing the total number of taxa. In other cases, removing character sets defined for a reduced number of taxa allowed us to proceed with a larger total number of taxa. In each situation, the decision to delete taxa or characters was always made so as to maximize the number of data sets representing the largest possible collection of common

taxa. For example, the paper by Mason-Gamer and Kellogg (1996) originally included 41 taxa and four data sets. All comparisons between sets were computed in a pairwise fashion in that publication and the different combinations did not include all taxa. We only considered taxa for which information was available for all data sets, reducing that number to six common taxa. The final list and details about the selected studies are presented in Table 1.I.

Character sets were converted to distance matrices for both types of analyses, in order to combine trees or data in a similar way (Lapointe *et al.* 1999). To do so, uncorrected ("p") distances were computed for sequence data, and mean character differences were calculated for any other types of data (all computations were made with PAUP*; Swofford, 1999). The latter was the closest distance to the euclidien distance, which was not available in PAUP*. In the case of combined analyses, distances were computed using all characters at the same time. For separate analyses, distance matrices were computed independently from each individual set. Phylogenetic trees based on combined or separate data were obtained with an unweighted least-squares method (Cavalli-Sforza & Edwards 1967), using PAUP* (Swofford 1999). A bootstrap procedure was then applied to total evidence trees using the same least-squares method and a fast stepwise addition option. This was done for 100 replicates. All weakly supported clades (*i.e.*, with bootstrap support < 50%) were collapsed in the total evidence trees.

To compute average consensus trees (Lapointe & Cucumel 1997), the pathlength distance matrices corresponding to the trees derived from the separate data sets were recorded. Average pathlength distances were then computed and submitted to a least-squares estimation procedure to construct the consensus solution. The resulting average consensus is a tree, with branch lengths, that minimizes the sum-of-squared distances to the original phylogenies. In order to compare the average consensus to other consensus methods that ignore branch lengths, strict (Sokal & Rohlf 1981) and majority rule (Margush & McMorris 1981) consensus trees were directly derived from the individual least-squares trees computed in PAUP* (Swofford 1999).

An important criterion for comparing trees, and particularly consensus trees, is the level of resolution of those trees. A simple way to measure resolution is to count the

number of internal branches in a tree. The relative resolution of a tree is computed as the ratio of the number of internal branches over the maximum possible number of internal branches (*i.e.*, n–3 for unrooted trees); this allows one to compare the resolution of trees bearing different number (n) of species. Likewise, the relative resolution of bootstrap trees can be expressed as the ratio of strongly supported clades (*i.e.*, with bootstrap support > 50%) over the maximum number of clades in a tree.

When trees need to be compared to one another, consensus trees and indices can be used (Rohlf 1982; Shao & Sokal 1986). For example, the global congruence of combined and separate analyses can be visualized with a *global consensus tree* (*sensu* Lapointe *et al.* 1999) bearing the clades corroborated by the different approaches. This tree is obtained by computing a consensus of the taxonomic and character congruence trees. Different consensus methods can be applied to obtain this global congruence tree. Namely, to measure strict congruence (hereafter referred to as *Cmin*) between combined and separate analyses, the strict consensus is used to derive the global congruence tree. The relative resolution of that global consensus is used to compute the *Cmin* index, which indicates topological agreement among the trees compared. The resolution of the global congruence tree is computed with the consensus fork index (Colless 1980). This is the proportion of possible clades ($n - 3$, for unrooted trees) that is resolved on the consensus tree. A value of 1 indicates a fully resolved consensus tree (the trees compared are fully resolved and identical) whereas a value of 0 represents a completely unresolved consensus tree (the trees compared share no clades in common). To measure semi-strict congruence (hereafter referred to as *Cmax*) between the taxonomic and character congruence trees, the global congruence tree is computed using the semi-strict consensus. In this case, the *Cmax* index is derived by measuring the relative resolution of the semi-strict consensus of the total evidence and taxonomic congruence trees (*i.e.* the relative resolution of the global congruence tree). A value of 1 indicates that the trees compared are compatible, and a value of 0 indicates completely different trees. Whereas *Cmin* represents an index of topological identity, *Cmax* can be defined in a broader sense as a measure of topological compatibility among partially resolved trees. For that matter, *Cmin* and *Cmax* determine the lower and upper bounds of global congruence. Notice that both indices would give identical results for pairs of fully resolved trees, however. While the global congruence tree is useful to

visualize the clades common to both analyses (character and taxonomic congruence), the *Cmin* and *Cmax* indices help us to quantify the congruence among them.

In our analyses, the resolution of all trees was recorded. Total evidence trees were also compared to the different consensus trees using strict (*Cmin*) and semi-strict (*Cmax*) congruence. To assess the effect of signal strength in the data, which was quantified by bootstrapping, both indices were measured before (*Cmin* and *Cmax*) and after (*Cmin'* and *Cmax'*) collapsing the weakly supported clades (*i.e.*, with bootstrap support < 50%) in the total evidence trees. Our results were also compared to those previously obtained by the authors of the original studies. We wanted to know whether total evidence trees based on distances were topologically different from the previously published trees, when restricted to the same numbers of taxa. Furthermore, we compared the bootstrap support values obtained in both cases to determine the number of well supported clades in common. A comparison of all individual trees obtained in the separate analyses with the corresponding total evidence tree was also performed to detect any differences, which could be reflected in the consensus.

RESULTS

All of the total evidence trees were fully resolved, which was expected given that a least-squares procedure was applied to distance matrices computed from the combined data sets. However, comparisons of these trees with those in the original studies revealed similar levels of resolution. In most cases, the least-squares phylogenies were congruent with the previously published trees based on parsimony (Table 1.I). The proportion of clades with high bootstrap support in our total evidence trees varied from 0.44 (4/9) to 1 (14/14); similar numbers were also obtained in the published trees.

In all but four cases, the topologies of the total evidence trees based on distances or characters were identical. For Lutzoni's (1997) data, the position of a single taxon was different in our tree, but the branch supporting that clade has a rather low bootstrap value (61%). In the case of Flook *et al.*'s (1999) data, it is the relationship of two small clades that differed from our tree. Olmstead and Sweere's (1994) data also differed with respect to the relationships of four taxa within a large clade. Finally, the analysis of Mason-Gamer

Table 1.I Results and summary of the 15 examples analyzed as part of this study

References	Data			Total evidence		Resolution index			Global congruence indices before bootstrap[e]			Global congruence indices after bootstrap[f]		
	No[a] taxa	No[a] sets	Type[b] of data	BS[c] support	Original study[d]	Ave	Maj	Strict	Ave	Maj	Strict	Ave	Maj	Strict
Kluge (1989)	10 (10)	2 (2)	2,6	0.86	y	1.00	0.29	0.29	0.86 *0.86*	0.29 *1.00*	0.29 *1.00*	0.86 *1.00*	0.29 *0.86*	0.29 *0.86*
Olmstead and Sweere (1994)	18 (18)	3 (3)	1,3	0.53	n	1.00	0.73	0.13	0.73 *0.73*	0.67 *0.93*	0.13 *1.00*	0.53 *1.00*	0.53 *0.73*	0.13 *0.53*
Omland (1994)	9 (9)	2 (2)	1,2	1.00	y	1.00	0.67	0.67	1.00 *1.00*	0.67 *1.00*	0.67 *1.00*	1.00 *1.00*	0.67 *1.00*	0.67 *1.00*
Mason-Gamer and Kellogg (1996)	6 (41)	4 (4)	1	0.67	n	1.00	1.00	0.33	0.67 *0.67*	1.00 *1.00*	0.33 *1.00*	0.67 *1.00*	0.67 *1.00*	0.33 *0.67*
Pennington (1996)	27 (27)	2 (2)	1,2	0.71	y	1.00	0.13	0.13	0.75 *0.75*	0.13 *1.00*	0.13 *1.00*	0.63 *0.92*	0.13 *0.71*	0.13 *0.71*
Baker and DeSalle (1997)	17 (17)	8 (8)	1	1.00	y	1.00	0.57	0.00	1.00 *1.00*	0.57 *1.00*	0.00 *1.00*	1.00 *1.00*	0.57 *1.00*	0.00 *1.00*
Lutzoni (1997)	30 (30)	4 (4)	1	0.70	n	1.00	0.22	0.07	0.63 *0.63*	0.22 *1.00*	0.07 *1.00*	0.56 *0.81*	0.22 *0.70*	0.07 *0.70*
Baum *et al.* (1998)	10 (18)	4 (4)	1,2,3	0.57	y	1.00	0.14	0.00	0.57 *0.57*	0.14 *1.00*	0.00 *1.00*	0.14 *0.57*	0.14 *0.57*	0.00 *0.57*
Cannatella *et al.* (1998)	10 (10)	5 (5)	1,2,4,5	0.86	y	1.00	1.00	0.14	1.00 *1.00*	1.00 *1.00*	0.14 *1.00*	0.86 *1.00*	0.86 *1.00*	0.14 *0.86*
Messenger and McGuire (1998)	26 (56)	4 (4)	1,2	0.87	y	1.00	0.74	0.30	0.83 *0.83*	0.74 *1.00*	0.30 *1.00*	0.78 *0.91*	0.70 *0.91*	0.30 *0.87*
Flook *et al.* (1999)	33 (35)	3 (3)	1	0.70	n	1.00	0.60	0.30	0.63 *0.63*	0.50 *0.87*	0.30 *1.00*	0.53 *0.83*	0.43 *0.73*	0.30 *0.70*
Gatesy *et al.* (1999a)	12 (13)	7 (17)	1	0.44	y	1.00	0.67	0.22	1.00 *1.00*	0.67 *1.00*	0.22 *1.00*	0.44 *1.00*	0.33 *0.78*	0.11 *0.56*
Liu and Miyamoto (1999)	26 (35)	3 (3)	1,2	0.52	y	1.00	0.43	0.22	0.48 *0.48*	0.43 *1.00*	0.22 *1.00*	0.35 *0.61*	0.35 *0.61*	0.22 *0.52*
Quicke and Belshaw (1999)	30 (33)	4 (4)	1,2,7	0.67	y	1.00	0.44	0.22	0.78 *0.78*	0.44 *1.00*	0.22 *1.00*	0.56 *0.85*	0.44 *0.67*	0.22 *0.67*
Springer *et al.* (1999)	11 (11)	8 (8)	1	0.75	y	1.00	0.50	0.00	0.75 *0.75*	0.38 *0.75*	0.00 *1.00*	0.75 *1.00*	0.38 *0.75*	0.00 *0.75*

Ave : average consensus ; Maj : majority rule consensus; Strict: strict consensus.[a] Number in original study are in parenthesis. [b] 1:molecular sequences 2:morphology 3:restriction sites 4:allozymes 5:calls 6:lipid characters 7:life history.[c] No of clades with bootstrap values 50%/total no of clades. [d] Congruence with original study. [e] Strict congruence (*Cmin*) on the first line and semistrict congruence (*Cmax*) on the second line. [f] Strict congruence (*Cmin'*) is on the first line and semistrict congruence (*Cmax'*) is on the second line.

and Kellogg's (1996) data revealed several discrepancies in our tree relative to those already published; when comparing the relationships among the six taxa for which all sequences were available in the original study, only one clade appeared to be congruent with our total evidence tree. This could be the effect of taxon sampling. Indeed, while 41 taxa were included in the original paper, only 6 were used in our study.

All of the average consensus trees were fully resolved, just like total evidence trees. On the other hand, the resolution of the standard consensus trees was quite variable. The relative number of resolved clades ranged from 0 (0/14) to 0.66 (4/6) in strict consensus trees, and from 0.13 (3/24) to 1 (7/7) in the majority rule trees, where 1 indicate a fully resolved tree and 0 a completely unresolved tree. By definition, all clades in the majority rule consensus were obtained in the strict consensus trees; in three cases for which only two separate trees were combined, those consensus tree were identical since in these case the strict and majority rule consensus trees are necessarily identical (Kluge 1989; Omland 1994; Pennington 1996). All of the clades in average consensus trees were also obtained in the strict consensus trees. In six cases (Flook *et al.* 1999; Liu & Miyamoto 1999; Lutzoni 1997; Mason-Gamer & Kellogg 1996; Quicke & Belshaw 1999; Springer *et al.* 1999), unique clades were obtained in average consensus trees in comparison with majority rule trees, however. In another case (Cannatella *et al.* 1998), the average and majority rule consensus trees were identical.

In all comparisons involving total evidence and average consensus trees, strict and semi-strict congruence indices were identical since both trees were always fully resolved, *prior* to bootstrap analysis. Strict congruence (*Cmin*) was better for the average consensus (0.78) than for majority rule (0.52) or strict consensus (0.20) trees, on average. On the other hand, semi-strict congruence (*Cmax*) was worse for the average consensus (0.78) than for majority rule (0.97) or strict consensus (1.00) trees, on average. The standard consensus methods produced trees, which were perfectly compatible (*Cmax* = 1) with the total evidence trees in 10 and 15 cases, respectively. Given the poor resolution of these consensus trees (see Table 1.I), such results were not surprising (an unresolved tree is always compatible with any other tree!). The comparisons performed *after* bootstrap analysis revealed quite different patterns, however. Strict congruence values decreased or

remained the same for all consensus methods, following bootstrapping, and *Cmin'* was again better for average consensus trees (0.64) than for majority rule (0.45) or strict consensus (0.19) trees, on average. Semi-strict congruence (*Cmax'*) was also better for the average consensus (0.90) than for majority rule (0.80) or strict consensus (0.73) trees, on average, when the clades with low bootstrap support in total evidence trees were collapsed. Whereas *Cmax'* always decreased or remained the same for standard consensus methods, it usually increased in the case of average consensus trees.

The results of the global congruence analysis comparing total evidence and average consensus trees can be classified in two categories: the perfectly congruent cases, for which combined and separate analyses provided identical trees (*Cmin* = 1), and the incongruent cases for which the competing approaches provided topologically different solutions (*Cmin* < 1), prior to bootstrap analysis. Those incongruent results could be further divided into three subsets. The first case involves studies that were not affected by the bootstrap. The second case is represented by studies for which bootstrap analysis partially improved the global congruence. The last case involves studies for which all topological incompatibilities between total evidence and average consensus trees were caused by weakly supported clades (Table 1.I).

Identical trees were obtained for combined and separate analyses in four cases, when average consensus trees were used (Baker & DeSalle 1997; Cannatella *et al.* 1998; Gatesy *et al.* 1999a; Omland 1994). In one of those cases (Cannatella *et al.* 1998), the majority rule consensus was also identical with the total evidence tree (*Cmin* = 1). As an example, Figure 1.1 illustrates the total evidence, average consensus and strict consensus trees obtained with Omland's (1994) data. The total evidence tree is here topologically identical to the average consensus tree, whereas the strict consensus differs in terms of resolution but is compatible with both trees. In this specific case, as well as with Baker and DeSalle's (1997) data, all clades were highly supported and global congruence was not affected by the bootstrap analysis. In the two remaining examples (Cannatella *et al.* 1998; Gatesy *et al.* 1999a), bootstrapping reduced strict congruence (*Cmin'*) between consensus and total evidence trees, however (see Table 1.I).

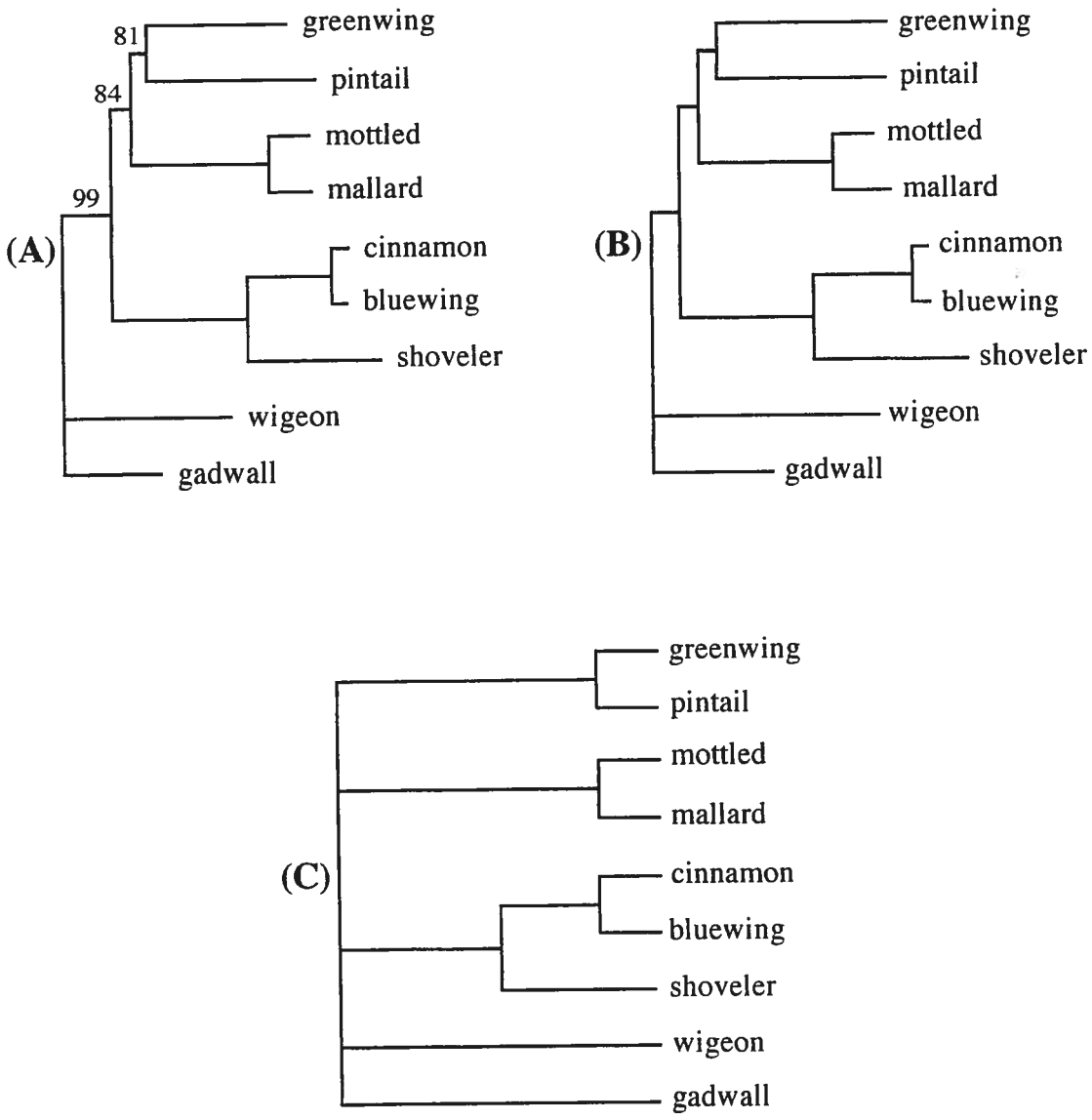

Figure 1. 1 Comparison of (A) the total evidence tree, (B) the average consensus tree, and (C) the strict consensus tree obtained from Omland's (1994) data. Numbers above branches in the total evidence tree are bootstrap support values, when different from 100. The majority rule consensus tree is identical to the strict consensus tree. See text for more details.

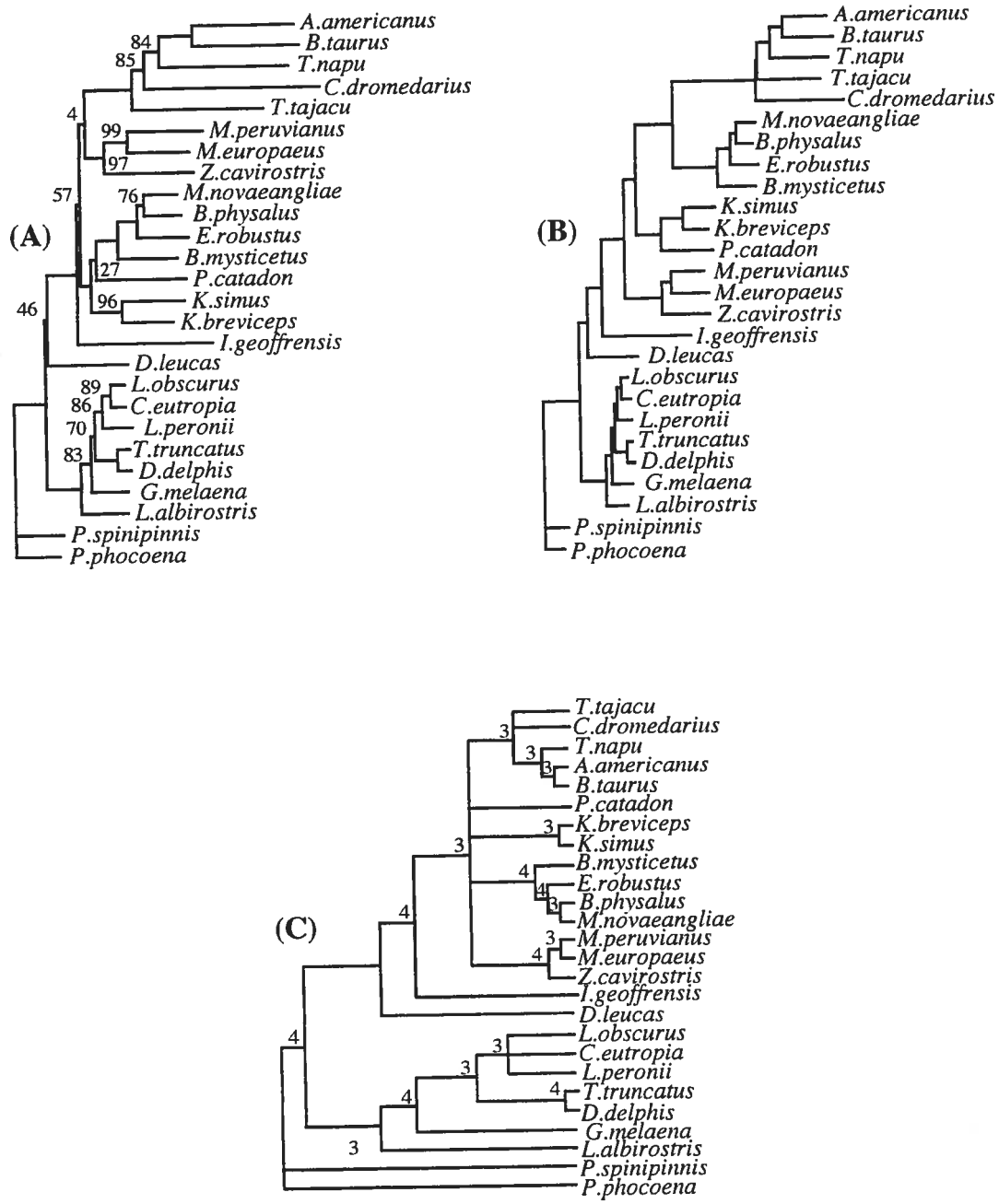

Figure 1. 2 Comparison of (A) the total evidence tree, (B) the average consensus tree, and (C) the majority rule consensus tree obtained from Messenger and McGuire's (1997) data. Numbers above branches in the total evidence tree are bootstrap support values, when different from 100. Numbers above branches in the majority rule consensus tree are the numbers of individual trees containing that clade. The strict consensus tree can be obtained by collapsing all branches with numbers <4. See text for more details.

the bootstrap, four clades in the total evidence tree were not compatible with the average consensus tree (*Cmax* = 0.83). Collapsing three clades with low bootstrap support (4%, 27% and 46%) reduced conflicts among the trees while increasing topological compatibility (*Cmax'* = 0.91), however. In this specific example, the majority rule consensus was equally compatible (*Cmax'* = 0.91) as the average consensus tree, but more compatible than the strict consensus solution (*Cmax'* = 0.87). In three other cases (Lutzoni 1997; Pennington 1996; Quicke & Belshaw 1999), majority rule and strict consensus trees provided identical *Cmax'*, following bootstrapping (Table 1.I).

The last four studies (Kluge 1989; Mason-Gamer & Kellogg 1996; Olmstead & Sweere 1994; Springer *et al.* 1999) certainly represent the most interesting examples (Table 1.I). In all of these cases, the observed discrepancies between total evidence and average consensus trees were not well supported and collapsing those branches resulted in perfect topological compatibility. As an example, the results obtained from Kluge's (1989) data are presented in Figure 1.3. Prior to the bootstrap, one clade in the total evidence tree was not compatible with the average consensus tree (*Cmax* = 0.86), but collapsing that weakly supported clade (4%) resulted in perfectly compatible trees (*Cmax'* = 1). Conversely, standard consensus trees were compatible (*Cmax* = 1) with the total evidence tree prior to bootstrap analysis, but compatibility decreased following bootstrapping (*Cmax'* = 0.86). The same results were obtained for the other examples, except for Mason-Gamer and Kellogg's (1996) data. In this single case, strict congruence (*Cmin*) was better for the majority rule tree than for the average consensus tree, prior to the bootstrap, but perfect compatibility (*Cmax'* = 1) was obtained with both consensus methods when the one clade with low bootstrap support was collapsed in the total evidence tree.

Discussion

The main objective of this paper was to evaluate the generality and applicability of the global congruence approach (Lapointe *et al.* 1999) using a wide variety of data sets gathered from the literature. We postulated that a coherent distance-based approach would lead to congruent solutions, regardless of whether data or trees are combined. Our results supported that claim and showed that total evidence and consensus can provide very similar

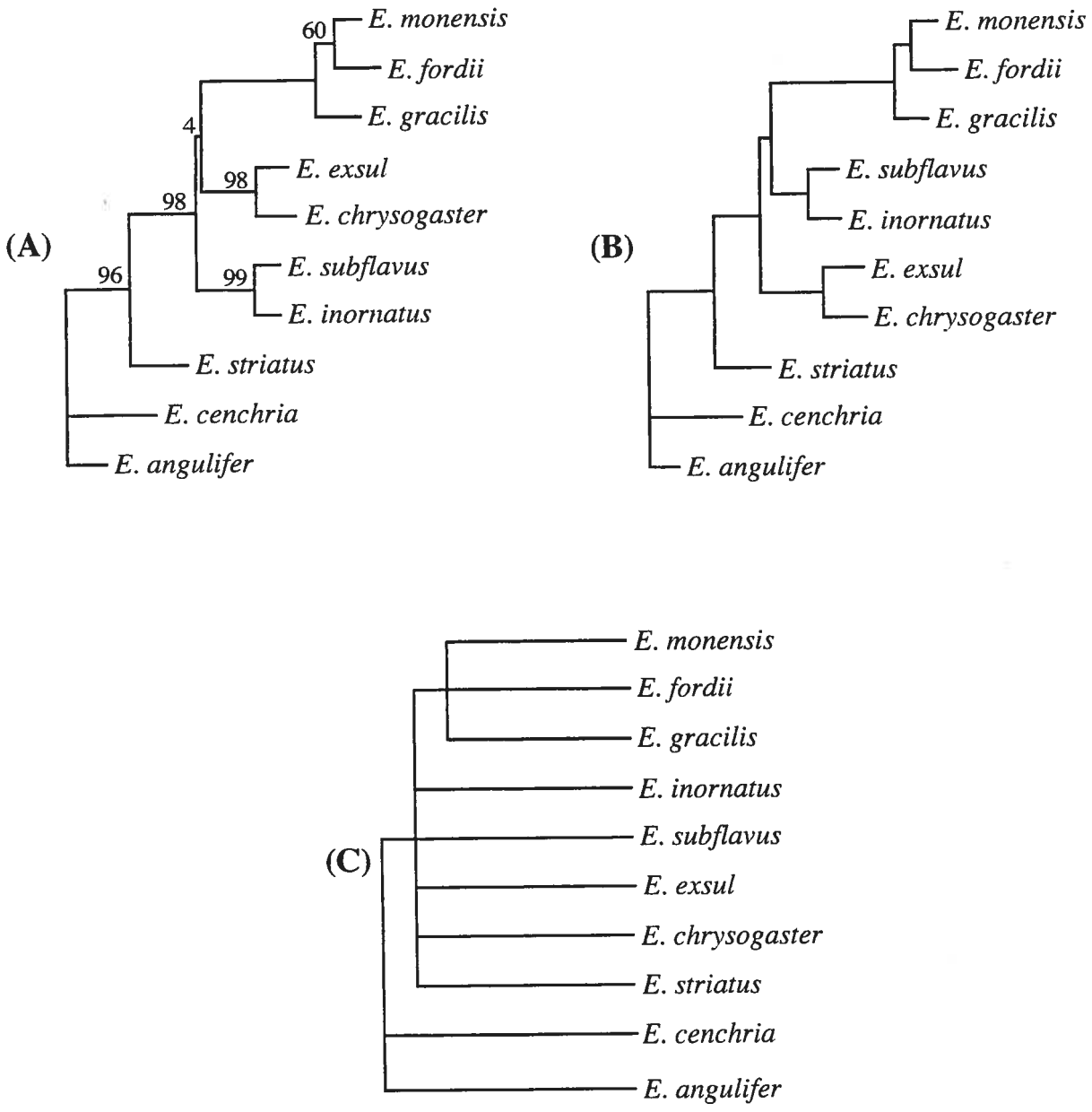

Figure 1. 3 Comparison of (A) the total evidence tree, (B) the average consensus tree, and (C) the strict consensus tree obtained from Kluge's (1989) data. Numbers above branches in the total evidence tree are bootstrap support values, when different from 100. The majority rule consensus tree is identical to the strict consensus tree. See text for more details.

trees, if branch lengths are taken into account when combining the individual phylogenies. The average consensus procedure (Lapointe & Cucumel 1997) and other consensus techniques for weighted trees (see Lapointe 1998a) thus represent methods of choice to standard consensus methods, if the user is interested in capturing not only topological agreement among the trees combined, but also pathlength differences. We also predicted that average consensus trees would be more similar to total evidence trees and more resolved than strict and majority rule consensus trees as Lapointe *et al.* (1999) have shown with an particular case. Our analyses verified that prediction and demonstrated that topological methods are indeed more likely to produce less resolved trees. This is an important observation because the poor resolution of consensus trees has been repeatedly used to illustrate the superiority of a total evidence approach. In the light of the results obtained with average consensus, the combined analysis does not appear to be superior to separate analyses as the trees produced by both approaches can be equally resolved, and in a large number of cases topologically identical, when branch lengths are taken into account.

On a related matter, we claimed that apparent conflicts between trees might be reduced if the weakly supported clades with low bootstrap support are collapsed in total evidence phylogenies. Our results confirmed that assertion. As a matter of fact, the comparison of total evidence and consensus trees was greatly affected by bootstrap analysis. Strict (*Cmin'*) and semi-strict congruence (*Cmax'*) decreased when the weakly supported clades in total evidence trees were also obtained in consensus trees. On the other hand, *Cmax'* increased when the clades with low bootstrap support were incongruent with the consensus solutions (*i.e.*, topological compatibility was improved when collapsing these clades). As a consequence of this, the global congruence of total evidence and average consensus was always superior to that of standard consensus methods, following bootstrapping.

The different consensus methods used in this paper have distinct properties and can produce different solutions. On the one hand, topological techniques were initially developed to indicate corroboration by independent data sets (Nelson 1979) and this notion continues to be an important argument for using consensus (e.g., Miyamoto & Fitch 1995). On the other hand, consensus with branch lengths (*sensu* Lapointe 1998a) focuses on corroboration by distances. That implies that average consensus trees may greatly differ

from strict (or majority rule) consensus trees. For instance, average consensus trees could contain clades that were not present in any of the individual trees (e.g., that was the case with Kluge's data; see Figure 3). The effect of long terminal branches may also affect consensus solution when branch lengths are taken into account. However, there are several arguments in favor of consensus methods that incorporate such branch lengths. First, compared to standard consensus methods, average consensus uses more information, and is usually more resolved. Second, the individual trees combined could be weighted according to the number of characters in the separate data sets. Third, various least-squares algorithms could be used to construct average consensus trees that satisfy different evolutionary models (e.g., by enforcing a molecular clock) (see Kirsch *et al.* 1997). Finally, the robustness of average consensus trees could be assessed with other resampling procedures or with specific randomization techniques. For instance, Lapointe *et al.* (1999) have used a taxonomic-jackknife procedure (Lapointe *et al.* 1994) to evaluate the effect of taxon sampling on the stability of consensus trees. Likewise, Cucumel & Lapointe (2000) have developed a statistical test to determine the probability that a consensus is representative of a set of individual trees. Such methods could be used to assess the robustness of total evidence and consensus trees in a similar way (see Lapointe & Kirsch 2001).

The differences among consensus methods that ignore or incorporate branch lengths are also apparent in our results. Whereas topological identity (*Cmin* = 1) was only obtained for average consensus trees (except in one case), topological compatibility (*Cmax* = 1) was common for standard consensus techniques, prior to bootstrap analysis. It is noteworthy that these congruence indices measure different things, however. *Cmin* measures topological identity as a strict consensus would represent it; its maximum value is obtained for fully resolved trees containing exactly the same clades. On the other hand, *Cmax* is related to what a semi-strict consensus (Bremer 1990) aims at representing; its maximum value is obtained when the consensus is fully resolved and no clades are incompatible in the trees compared. Both indices would provide identical results in the case of fully resolved trees, but *Cmax* is more liberal when polytomies are found in any of the input trees, as it is the case when branches are collapsed following bootstrapping, or when a consensus is not well resolved (e.g., strict consensus trees).

In spite of the wide variety of studies considered in the present work, we were not able to observe any relationships between global congruence indices (*Cmin* and *Cmax*) and the characteristics of the data sets. For instance, the perfectly congruent cases (Baker & DeSalle 1997; Cannatella *et al.* 1998; Gatesy *et al.* 1999a; Omland 1994) were based on different taxa (*i.e.*, birds, insects, amphibians and mammals) with number of species ranging from 9 to 17, and numbers of data sets ranging from 2 to 8, including sequence, morphological and allozyme data, among others. There was no relationship between bootstrap support and congruence either, in those four cases as well as in other examples. However, five of the studies for which bootstrapping did not result in perfect compatibility (Flook *et al.* 1999; Liu & Miyamoto 1999; Lutzoni 1997; Messenger & McGuire 1998; Quicke & Belshaw 1999) were among the largest in terms of number of taxa. It may well be that global congruence decreases with the number of taxa. Indeed, more taxa that there are more ways to go wrong. But further analyses for a greater number of studies would be required to support or reject that claim.

All of our combined analyses were based on distance matrices, computed from character data or molecular sequences. Similarly, the consensus analyses were based on the combination of pathlength distance matrices, corresponding to the different trees obtained from individual data sets. For that matter, our approach could be blind to clade support that may be hidden in the analyses of separate data sets, whereas a character congruence approach would be able to identify such "hidden support" (see Gatesy *et al.* 1999b). Interestingly, the total evidence trees computed with a least-squares algorithm were in most cases congruent with the most parsimonious total evidence trees published in the original studies. For the most part, those trees were in turn congruent with average consensus trees. Consequently, our results do not seem to have been affected by the type of phylogenetic estimation method, nor by the use of distances, and our conclusion still holds when compared to the original phylogenies.

Future work could generalize the global congruence approach to supertrees defined on overlapping sets of taxa (Lapointe & Cucumel 1997), instead of reducing the data sets to include only common taxa. In such applications (see Kirsch *et al.* 1997; Lapointe & Kirsch 2001), the average consensus procedure represents an alternative to the more commonly used supertree methods (for a review, see Sanderson *et al.* 1998), while considering branch

lengths. As stated above, the effect of various weighting schemes could also be evaluated to determine optimal ways to combine phylogenies derived from data sets with different numbers of characters. Taxonomic congruence automatically attributes equal weights to each data set, regardless of their numbers of characters, a procedure considered by some as arbitrary (Donoghue & Sanderson 1992; Hillis 1987). However, assigning unit weights to every character, molecular as well as morphological, in a total evidence framework also causes problems (see Doyle 1992). This dual effect can be addressed with a weighted-average consensus method, accounting for the number of characters in the individual data sets when computing the consensus tree (Lapointe *et al.* 1999). Finally, it would be interesting to assess whether a global congruence approach increases the phylogenetic accuracy, as opposed to when using either one of the competing methods independently. The agreement between trees obtained with combined and separate analysis could then be visualized in a *global consensus tree* (e.g., Lapointe & Kirsch 2001; Lapointe *et al.* 1999) bearing the clades corroborated by the different approaches; one could then argue that these clades are more likely to be real (see Kim 1993).

ACKNOWLEDGMENTS

We would like to thank Guy Cucumel and Pierre Legendre for their comments on an early draft of this project, and John A. W. Kirsch for contributing to the main ideas presented in this paper. We also thank Mike Sanderson, Alan De Queiroz and an anonymous referee for their constructive reviews. This work was supported by a NSERC grant OGP0155251 to F.-J. Lapointe and a FBSB scholarship to C. Levasseur. The trees were originally drawn using Rod Page's (1996) TreeView program.

CHAPITRE 2

Increasing phylogenetic accuracy with global congruence

Cet article est publié sous la référence :

Levasseur, C. & Lapointe, F.-J. 2003 Increasing phylogenetic accuracy with global congruence. In *Bioconsensus* (ed. M.F. Janovitz, F.-J. Lapointe, F.R. McMorris, B. Mirkin & F.S. Roberts), pp. 221-230. DIMACS Series in Mathematics and Theoretical Computer Science, Providence : American Mathematical Society.

ABSTRACT

In this paper, simulations were used to assess the degree to which combining independent data sets or their corresponding trees can produce congruent results when a consensus method that accounts for branch lengths is used. We hypothesized that phylogenetic accuracy will be increased when those approaches are used jointly in a so-called "global congruence" framework, and postulated that the data sets heterogeneity will affect the global congruence. Our results indicated that the accuracy rate of phylogenetic estimation can indeed be increased when separate and combined analyses are used in combination, regardless of the data heterogeneity.

INTRODUCTION

The problems associated with the analysis of multiple data sets have been debated at length in phylogenetics (for reviews of this debate, see De Queiroz *et al.* 1995; Huelsenbeck *et al.* 1996). In the last decade, a number of distinct solutions have been proposed to resolve this issue and several papers have tried to demonstrate the superiority of one approach over the others. Still, it is not yet clear whether it is better to combine separate data sets before phylogenetic analysis (character congruence, *sensu* Kluge 1989) or to combine trees obtained independently from the different data sets with consensus techniques (taxonomic congruence, *sensu* Mickevich 1978). An intermediate solution that relies on the results of statistical heterogeneity tests has been suggested by some (Farris *et al.* 1995b; Huelsenbeck & Bull 1996; Mickevich & Farris 1981; Rodrigo *et al.* 1993). Proponents of this conditional approach claim that homogeneous data sets should be combined prior to phylogenetic analysis, while heterogeneous data sets should not; in such cases, consensus methods could be used to combine the trees derived from the separate data sets.

We have recently shown (Levasseur & Lapointe 2001) that when a method accounting for branch lengths (*e.g.*. the average consensus; Lapointe & Cucumel 1997) is used to derive trees with taxonomic congruence, these consensus trees are often similar to the trees derived from character congruence. Furthermore, collapsing branches that are not well-supported (*i.e.* with bootstrap values < 50%) can reduce the apparent conflicts between separate and combined analysis of the data. We thus advocate the use of both approaches jointly as proposed by De Queiroz (1993) in a so-called "global congruence" approach (*sensu* Lapointe 1998b). In this case, rather than assessing the congruence among data sets or individual phylogenies, this approach evaluates the agreement between character and taxonomic congruence. It can thus be used as a means to cross-corroborate the trees obtained by separate and combined analyses. We further believe that this could increase the accuracy of phylogenetic trees derived using this analytical framework.

The major objective of this paper is to compare the accuracy rates of the global congruence approach to that of character and taxonomic congruence, using simulations. The effect of consensus methods will also be assessed by comparing the results obtained

with the average, strict, and Adams consensus. We will determine the effect of data heterogeneity on the performance of the competing approaches by accounting for this factor in our simulation design. We have tested a simple case where two partitions of limited heterogeneity (derived from topologically identical trees but with different branch lengths) are of the same size. In other words, we want to know whether the use of heterogeneity tests is indeed justified to select character congruence over taxonomic congruence. We hypothesize that it is more accurate to use both methods jointly as proposed in the global congruence approach, regardless of the results of heterogeneity tests.

MATERIAL AND METHODS

Simulation design

A ten taxon model tree (model tree: MT) (Figure 2.1A) similar to that of Kumar (1996) was used to evolve DNA sequences for the simulation study. For each simulation, two data sets (or data partitions) of the same length (*i.e.* 2500 base pairs) were evolved on that tree under two radically different situations. This could represent a case of combining two molecular data sets. In order to simulate heterogeneous data sets, the sequences were independently evolved on model trees with the same topology but with different branch lengths (Figures 2.1B and 2.1C) and using different rates of evolution (0.25 and 1.75). For the simulation of homogeneous data, sequences for both data partitions were evolved on the same model tree (Figure 2.1C) with identical rates of evolution (1.75). The simulations were performed with the Seq-Gen program (Rambault & Grassly 1997) using a Jukes-Cantor model (Jukes & Cantor 1969). For each series of simulations, 1000 replicates were generated and analyzed with the global congruence approach.

In order to test that sequences evolved on the model trees were more structured than random data, permutation-tail probability tests (*PTP*, Archie 1989; Faith & Cranston 1991) were computed in PAUP* (Swofford 1999). Moreover, incongruence length difference tests (*ILD*, Farris *et al.* 1995a) were also run in PAUP* to assess the heterogeneity or the homogeneity of the simulated data sets.

Phylogenetic analysis

A tree was first derived from the combined sequences using character congruence. Distances were then computed from sequences using Jukes-Cantor corrected distances among all pairs of species. A phylogenetic tree was estimated from that distance matrix using an unweighted least-squares method (Cavalli-Sforza & Edwards 1967) (total evidence tree: TE). This process was repeated for the 1000 replicates. All computations were performed in PAUP*.

To compare with the results of combined analysis, the data partitions were also analyzed separately in a taxonomic congruence approach. In that case, distance matrices were computed independently for each sequence and trees were estimated from both data sets with the same least-squares method as for character congruence. The separate trees were then combined with consensus methods (consensus tree: CT). The average consensus procedure (Lapointe & Cucumel 1997) was used to combine trees while accounting for branch lengths whereas the strict (Sokal & Rohlf 1981) and Adams (Adams 1972) consensus were selected to combine phylogenies based on topological relationships alone. This procedure was repeated for the 1000 replicates and for each series of simulations.

Global congruence and phylogenetic accuracy

Global congruence was assessed by comparing the trees derived with character and taxonomic congruence. An index of topological identity (*Cmin*) was used to quantify the agreement among those trees. This was done by measuring the resolution of the global congruence tree (*i.e.* the strict consensus of the trees obtained with character and taxonomic congruence) using the consensus fork index (Colless 1980). This is the proportion of possible clades ($n - 3$, for unrooted trees) that are resolved on the consensus tree. *Cmin* thus represents the mean values of the consensus fork index for the 1000 replicates. Absolute topological identity values were also computed by counting the number of replicates for which the trees derived with character and taxonomic congruence were identical.

Phylogenetic accuracy was measured by comparing the individual accuracy rates of the competing approaches; that is, the number of times that the topology of the model tree (Figure 2.1A) was obtained in combined (TE=MT) or separate analyses (CT=MT) out of

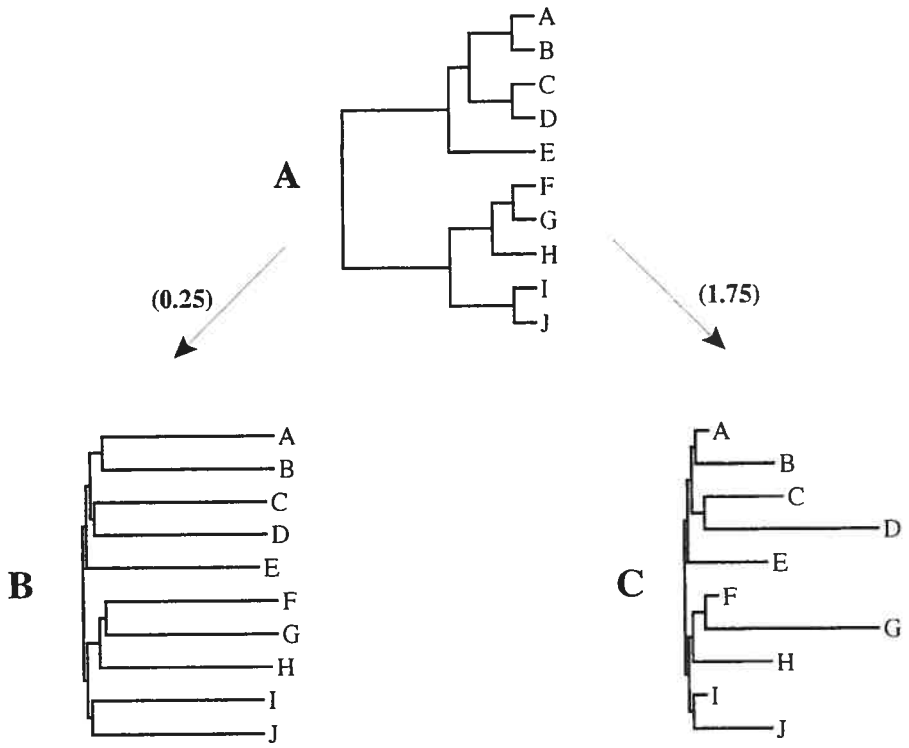

Figure 2.1 Model topology (A) with different branch lengths and different evolutionary rates of change [slow (0.25) and rapid (1.75)] among branches. Trees B and C were used to simulate heterogeneous data sets, whereas tree C was selected to evolve sequences for homogeneous data partitions.

the 1000 replicates. Combined accuracy rates (TE=CT=MT) were also measured by counting the number of times that the model tree was *jointly* recovered by character and taxonomic congruence divided by the number of times that these methods produced identical topologies (TE=CT).

RESULTS

Results of the parsimony *PTP* tests allow us to reject the null hypothesis that states that the data under investigation are not distinguishable from randomly (which is not phylogenetically informative) permuted data. Results of *ILD* tests, on the other hand, confirmed that the simulated data sets were either significantly heterogeneous or homogeneous data partitions.

Global congruence

The results of our simulations presented in Table 2.I show that the mean *Cmin* values are much higher for the average consensus than when the strict or Adams consensus are used to combine the results of separate analyses. The *Cmin* index is affected by the fact that the strict consensus often produces partially resolved trees. This is due in part to the strictness of the method, which is the most conservative one. The fact that it does not take into account branch lengths, thus ignoring possible important information, could also explain the less resolved results. Slightly better results are obtained for the Adams consensus, another method that ignores branch lengths, however. Interestingly, the same pattern is observed by considering absolute topological identity values (Table 2.I); these values, representing the number of replicates for which the trees derived for character and taxonomic congruence were identical, are indeed larger when the average consensus method is used.

Our results also show that global congruence is clearly affected by the data heterogeneity. Indeed, the *Cmin* index values are larger for homogeneous data than for heterogeneous partitions. Also, for the average consensus, 969 replicates out of 1000 resulted in congruent results for homogeneous data (for the strict and Adams consensus, that number only reaches 735). On the other hand, the vast majority of trees obtained by the

Table 2.I Mean *Cmin* values of the global congruence tree for heterogeneous and homogeneous data sets, and for three consensus methods are presented on the first line. Absolute topological identity values of trees derived with character and taxonomic congruence (out of 1000 replicates) are presented on the second line.

	average consensus	*strict consensus*	*Adams consensus*
Heterogeneous data	0.689	0.582	0.597
	266	*72*	*72*
Homogeneous data	0.995	0.957	0.958
	969	*735*	*735*

competing approaches were different in simulations involving heterogeneous data partitions (see Table 2.I).

Phylogenetic accuracy

Figure 2.2 reports accuracy rates of the character and taxonomic congruence approaches for heterogeneous and homogeneous data. The comparison of individual results indicates that the best individual accuracy rate is provided by taxonomic congruence using the average consensus when the data are heterogeneous. On the other hand, character congruence does slightly better when the data are homogeneous. In both series of simulations, worse results were obtained by taxonomic congruence using the strict (or Adams) consensus.

The shaded areas of Figure 2.2 show us that the global congruence approach combined accuracy rates for heterogeneous data sets are higher when strict (or Adams) consensus (0.996) rather than average consensus (0.947) of separate analysis is used jointly with combined analysis. The same pattern is repeated for homogeneous data sets for which the relative accuracy rate of the strict (or Adams) consensus (0.996) is slightly better than the rate of the average consensus (0.988) of separate analysis used

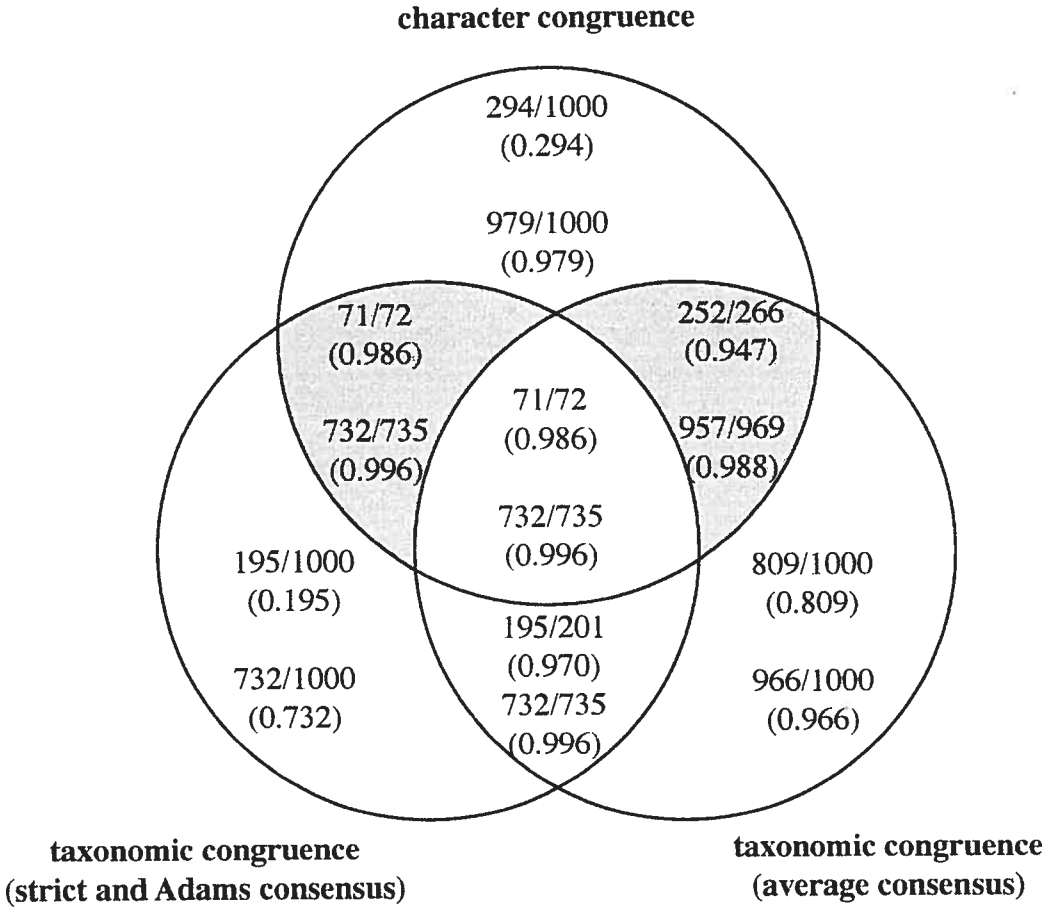

Figure 2.2 Individual and combined accuracy rates of the different approaches for heterogeneous and homogeneous data sets. Results for heterogeneous data sets are presented first line followed by the results for homogeneous data. Individual results indicate the number of times that each method correctly recovered the topology of the model tree, out of 1000 replicates (Figure 2.1A). Shaded areas represent global congruence results; *i.e.* the number of times that the model topology was correctly recovered (TE=CT=MT), divided by the number of times that the trees obtained with character and taxonomic congruence were identical (TE=CT). Joint accuracy rates of the average and strict (or Adams) consensus methods are also provided, as well as a global rate obtained when all three methods are used jointly. Percentages are in parenthesis.

jointly with combined analysis. However, absolute accuracy values (*i.e.* the number of times that different methods recovered the correct tree, TE=CT=MT) are larger when average consensus trees are used in a global congruence approach (Figure 2.2). These results also show that individual accuracy rates of the three methods used in combination is identical to the rates obtained when the strict (or Adams) consensus of separate analysis and combined analysis are used jointly. The latter case thus represents a subset of the results of the global congruence obtained with the average consensus. For that reason, the strict (or Adams) consensus is not more informative than the average consensus in a global congruence framework, even though the relative rates seem to indicate a better accuracy.

DISCUSSION

The main objective of this paper was to assess the performance of the global congruence framework for analyzing multiple data sets in contrast with the competing character and taxonomic congruence approaches. Furthermore, we wanted to measure the effect of data heterogeneity by using simulations of data partitions evolved on model trees that could be the result of different evolutionary patterns. Our results confirmed the hypothesis that accurate phylogenies are obtained relatively more often with global congruence and that a consensus that takes into account branch lengths performs better than topology-based consensus methods such as the strict (or Adams) consensus (Levasseur & Lapointe 2001). These results also showed that the global congruence approach is affected by heterogeneity of the data partitions. The more heterogeneous the data, the less congruent the results of separate and combined analysis will be. In such cases, collapsing weakly-supported branches in our phylogenetic trees could in fact make these results compatible (Levasseur & Lapointe 2001).

Our simulations also illustrated that the average consensus provides the best individual accuracy rates for taxonomic congruence regardless of data heterogeneity. Whereas character congruence does a little better in cases involving homogeneous data partitions, it performs badly when data sets are heterogeneous. The use of a consensus method that takes into account branch lengths should thus be preferred over character congruence in the latter situation. When this consensus method is used in a global congruence framework, combined accuracy rates increase even more. When the strict (or

Adams) consensus is used in a global congruence approach, accuracy rates are artificially increased, because (1) in all cases for which the strict consensus was topologically identical to trees derived from combined analysis, results of the separate analysis were also identical. In such cases, any consensus method would have returned the same solution. (2) Our results also show that in all of these cases, the average consensus also recovered the same tree. (3) In spite of an impressive accuracy rate (0.996), chances are that the strict consensus will often produce partially resolved trees, compared to the average consensus. Consequently, the joint use of character and taxonomic congruence with average consensus increases the absolute probability of recovering the model topology (Figure 2.2). That recommendation applies for heterogeneous as well as for homogeneous data sets.

Considering these results, and for tables of molecular data, we believe that the conditional data combination approach relying on statistical tests of heterogeneity should not be used *a priori*, and we do not think that a single analysis should be conducted. We rather propose that separate *and* combined analysis of the data should always be performed and be used *jointly*. Our proposed framework presents multiple advantages. (1) The use of a consensus method can allow identifying sources of incongruence between the different data partitions (such as lineage sorting, potential of reticulate evolution or rapid radiation for example) by examination of the separate trees. (2) The global congruence of trees derived from combined and separate analysis can also be increased with a bootstrap analysis (Felsenstein 1985) that identifies weakly-supported branches (Levasseur & Lapointe 2001). (3) The use of character and taxonomic congruence in a coherent fashion allows for corroboration of the result of data combination with consensus (Lapointe *et al.* 1999). This rationale is similar to that suggested by Kim (1993) to estimate accurate phylogenies by combining different inference methods.

In order to investigate more thoroughly the performance of character and taxonomic congruence used jointly in a global congruence approach, more simulations are warranted. For one, it is worth mentioning that even in the worst case of heterogeneity, our simulations were based on topologically identical trees. Only branch lengths were varied. Global congruence should be tested with topologically different trees. This case represents one situation where the use of heterogeneity tests and thus the conditional approach could be

pertinent. Consequently, our simulations cannot be used to completely discard the use of those tests in favor of the global congruence approach.

In the present case, the choice of the strict consensus was motivated by its wide use in phylogenetic analyses. Because of the strictness of that method, the results obtained in this paper probably represents the worse ones for the use of this type of consensus, which does not take into account branch lengths. However, similar results were obtained with the Adams consensus, a method that usually produces trees with a better resolution than the strict consensus. Clearly, other topology-based methods could be considered to combine different data sets. For instance, the majority rule consensus (Margush & McMorris 1981) could be more appropriate to combine trees when more than two data sets are analyzed. Similarly, the semi-strict consensus (Bremer 1990) could perform better than the strict consensus when the trees combined are not fully resolved. In the present paper, both of these consensus methods were identical because all separate trees were fully resolved. However, the relative performance of these alternative methods must be evaluated in simulations involving more than two data partitions.

Also, it is of interest to test the effect of data sets of different sizes and the different ways to weigh the data partitions in such cases. Indeed, in situations where a small and a large heterogeneous partitions were to be combined, the result of character and taxonomic congruence approach could be affected by differential weighting and very well be different from one another. The robustness of trees derived with character congruence has to be considered in future simulations. Some approaches for assessing the stability and reliability of consensus methods are also needed (Lapointe & Cucumel 2003). The performance of the global congruence could also be further tested in simulations involving incongruent data sets for which trees differ only partially (Wiens 1998a). Finally, the index we used to measure agreement among trees does not enable us to detect if the lack of resolution is due to very different trees or to the effect of a floating taxon. Indeed, trees may be very similar in terms of shared triplets while sharing no clusters. Other indices that allow bypassing this limitation could be used to test the congruence among the trees (Steel & Penny 1993). Also, some methods of taxonomic jackknifing or reduced consensus could be used to identify floating taxa.

In conclusion, our message is that more than one method must be used to deal with multiple data sets in phylogenetics. The use of character and taxonomic congruence, combined with a consensus method that accounts for branch lengths represents, in our opinion, the best solution to this problem. The global congruence approach is even more suited for the combination of highly heterogeneous data partitions. Although consensus methods were not designed to replace phylogenetic estimation algorithms, they can be used jointly with combined analysis to provide more accurate estimate of phylogenetic relationships. Future work will verify this statement for other methods of phylogenetic estimation and other consensus methods for weighted trees (Lapointe 1998a).

ACKNOWLEDGMENTS

We thank Guy Cucumel, Pierre Legendre and all members of the LEMEE for their comments on early draft of this project. We are also grateful to Mark Wilkinson for his constructive review of this manuscript. This work was supported by a NSERC grant to F.J.L. and a FBSB and NSERC scholarship to C.L.

CHAPITRE 3

Global congruence for better phylogenies

Cet article sera soumis prochainement:

Levasseur, C. & Lapointe, F.-J. Global congruence for better phylogenies

ABSTRACT

The use of multiple data sets in phylogenetic analysis is widespread and three competing approaches are available to analyze such extensive data. (1) Character congruence (total evidence) requires that all data be processed together whereas (2) taxonomic congruence advocates analyzing the different partitions separately. The trees resulting from the separate data sets can thus be summarized using consensus methods. (3) The conditional approach considers the results of heterogeneity tests as a good indicator of whether data partitions should be analyzed together (homogeneous data) or separately (heterogeneous data). In this paper, we use a recently proposed approach, defined as global congruence. Contrary to the conditional combination approach, this strategy uses character and taxonomic congruence jointly. On the grounds that results from separate and combined analyses are often similar, we think that this global congruence approach may increase the phylogenetic accuracy when branch lengths are taken into account. When both approaches yield identical results, we postulate that it will more often be identical to the model tree. Using simulations, we test the performance of character, taxonomic and global congruence. Our results show that a consensus that takes branch lengths into account (*e.g.*, the average consensus) can yield trees more similar to total evidence than those based on topological consensus methods (*i.e.*, strict, semi-strict, majority rule and Adams). Moreover, the global congruence approach produces trees that are more often identical to the model tree than do character or taxonomic congruence approaches alone, regardless of the consensus method (with or without branch lengths). Finally, when total evidence and consensus trees differ, collapsing incompatible branches often yield trees that are compatible to the model tree.

INTRODUCTION

The last years have been increasingly productive in generating molecular data and it is now of common acceptance that the use of multiple sources of information (e.g., nuclear and mitochondrial DNA regions, morphological data) is essential for phylogenetic inference. The fast accumulation of sequences is not without consequences, however. For one, computational tractability is an important issue because some algorithms are not usable for large data sets (Felsenstein 1978; Sanderson & Shaffer 2002). In addition, many problems imply that the genes under study may give different versions of the phylogenetic history of a particular species group, for example, the gene tree/species tree may differ due to lineage sorting or gene duplication (Slowinsky & Page 1999; for a review of the problems, see Wendel and Doyle, 1998). These issues bring up methodological and philosophical questions on ways to reconcile conflicting signals when dealing with multiple data sets.

Over the last decade, many constructive papers have proposed alternative approaches for taking advantage of all available characters in a phylogenetic context (for review, see De Queiroz *et al.* 1995; Huelsenbeck *et al.* 1996). This triggered an intense debate among opposing camps. Based on the philosophical principal that all available evidence should provide the best answer, Kluge (1989) first suggested that all data should always be analyzed together in a so-called "total evidence" or character congruence approach. An important methodological argument for combining all characters is to extract the interactive phylogenetic signal of the entire data (Kluge 1989; Barrett *et al.* 1991; Eernisse & Kluge 1993; Jones *et al.* 1993; Kluge & Wolf 1993). The other approach referred to as taxonomic congruence (Mickevich 1978) proposes to analyze each data set separately, before combining the resulting phylogenies with a consensus method (see also Swofford 1991; Huelsenbeck *et al.* 1994). With separate analysis, smaller data sets are not drowned in an ocean of characters, and this allows for a more comprehensive investigation of the phylogenetic signal underlying each data set (Farris *et al.* 1995b; Huelsenbeck & Bull 1996; Mickevich & Farris 1981; Rodrigo *et al.* 1993). Finally, some authors have taken a middle ground, using a so-called "conditional combination" approach (Bull *et al.* 1993; Legendre & Lapointe 2004; Miyamoto & Fitch 1995; Rodrigo *et al.* 1993). They claim that heterogeneity tests (Mickevich & Farris 1981) should be used to determine

whether the data are homogeneous and ought to be analyzed simultaneously with character congruence, or separately with taxonomic congruence if the data are heterogeneous.

There exist sensible arguments for using or not using any one of these approaches (for review, see De Queiroz *et al.* 1995; Huelsenbeck *et al.* 1996). However, we have seriously challenged the claim that total evidence and consensus trees are incompatible with one another (Levasseur & Lapointe 2001; Levasseur and Lapointe 2003). Character congruence is often said to be superior to taxonomic congruence (Barrett *et al.* 1991; Kluge & Wolf 1993), but the major point of contention seems to be concerning the use of consensus methods to combine the results of separate analysis. Lapointe *et al.* (1999) have shown that total evidence and consensus trees can converge to the same solution when treated in a coherent fashion, that is using distance matrices to estimate phylogenies and path-length distance matrices to construct consensus trees while taking into account branch lengths. Levasseur and Lapointe (2001) then confirmed that assertion using a wide array of published data sets. Nevertheless, it remains to be demonstrated that when used jointly with character congruence, these consensus trees can help estimate accurate phylogenies.

It is already known that the use of several phylogenetic methods can increase accuracy (Kim 1993). In the same way, some authors have suggested that it could be advantageous to use character and taxonomic congruence jointly (De Queiroz 1993; De Queiroz *et al.* 1995; Hillis 1987; Larson 1994). We believe that using these approaches together in a global congruence framework (*sensu* Lapointe *et al.* 1999) may increase phylogenetic accuracy, that is that trees resulting from the global congruence approach will be more similar to the model tree than trees from the individual approaches. In the present paper, we will address this question and assess the relative performance of the character, taxonomic and global congruence approaches. We will use a simulation framework to corroborate earlier results obtained by Levasseur & Lapointe (2001) with real data sets. More precisely, we postulate that consensus trees with branch lengths will be more similar to total evidence trees than consensus trees based on topological relationships alone. Furthermore, we suggest that the accuracy of consensus methods with branch lengths will be better than for other consensus methods, and similar to total evidence trees. We also hypothesize that combining the results of total evidence and consensus trees in a global

congruence approach will increase accuracy rates. When the competing approaches differ, their strict consensus should not contradict the true relationships, however.

MATERIAL AND METHODS

Simulation protocol

The effect of the number of taxa (10 or 30), the number of data partitions (2 or 10), and the heterogeneity among data sets (low or high) were investigated with simulations. We only focused on extreme cases to detect differences among the competing approaches with respect to the different parameters. All possible combinations of the three parameters were considered, but the number of characters per partition and the global rate of evolution were fixed in the simulations. Because the number of characters is equal in every partition, our simulations only address the cases where data sets of approximately the same size are combined (when combining different genes for example).

To simulate DNA sequences, 1000 replicate model trees with branch lengths were generated with a Yule branching process using the program r8s (Sanderson 2003). Each tree was then duplicated or replicated 10 times according to the number of partitions required. Homogeneous data partitions were created by simulating the evolution of sequences on these identical trees. To create heterogeneous data partitions, the branch lengths were modified in the replicate trees with multipliers sampled from a uniform distribution bounded between 0 and 0.5. Nucleotide sequences of fixed length per data partition (2000 base pairs) were evolved on those trees (homogeneous or heterogeneous) with the program Seq-Gen (Rambault & Grassly, 1997), using a Jukes-Cantor model of evolution (Jukes & Cantor, 1969) with rate heterogeneity and a gamma distribution set to 0.5. The average rate of substitution across all trees was fixed to 1.0.

Character congruence

Following Lapointe *et al.* (1999), distance matrices were derived from the combined data partitions to estimate total evidence trees. These distances were corrected using a Jukes-Cantor model with a gamma distribution. The total evidence trees were obtained by

applying an unweighted least-squares algorithm (Cavalli-Sforza & Edwards 1967) to the corrected distances in PAUP*.

Taxonomic congruence

The separate analyses were performed with the same method as the combined analyses. Separate distance matrices were thus computed from the different data partitions, using the same model as before (JC + gamma). The corresponding trees were derived separately with an unweighted least-squares algorithm in PAUP*.

Different consensus techniques were employed to combine the results of the separate analyses. The strict (Sokal & Rohlf 1981), semi-strict (Bremer 1990), majority rule and resolved majority rule (Margush & McMorris 1981), and Adams (1972) consensus were used to summarize the topological agreement among trees. To compare these consensus trees with those obtained by a method that accounts for branch lengths, average consensus trees (Lapointe & Cucumel, 1997) were also computed. All topological consensus methods were constructed using PAUP*. The average consensus trees were estimated using PHYLIP by applying an unweighted least-squares algorithm (*i.e.*, FITCH) to a matrix of average path-length distances. The resolution of the trees derived with each consensus methods was computed with the consensus fork index (Colless 1980). This is the proportion of possible clades ($n - 3$, for unrooted trees) that is resolved on the consensus tree. A value of 1 indicates a fully resolved tree whereas a value of 0 represents a completely unresolved tree.

Global congruence

The results of character and taxonomic congruence were compared to assess global congruence. To do so, the same indices as those previously used by Levasseur & Lapointe (2001) were selected to measure the agreement among the competing trees. These indices measure topological identity (*Cmin*) and topological compatibility (*Cmax*) to the total evidence tree. They are obtained respectively by computing the consensus fork index (Colless 1980) of the strict and semi-strict consensus of the trees under comparison. For *Cmin*, a maximum value of 1 is obtained when the trees compared are identical and a value of 0 is obtained when trees are completely different (*i.e.*, when their strict consensus is

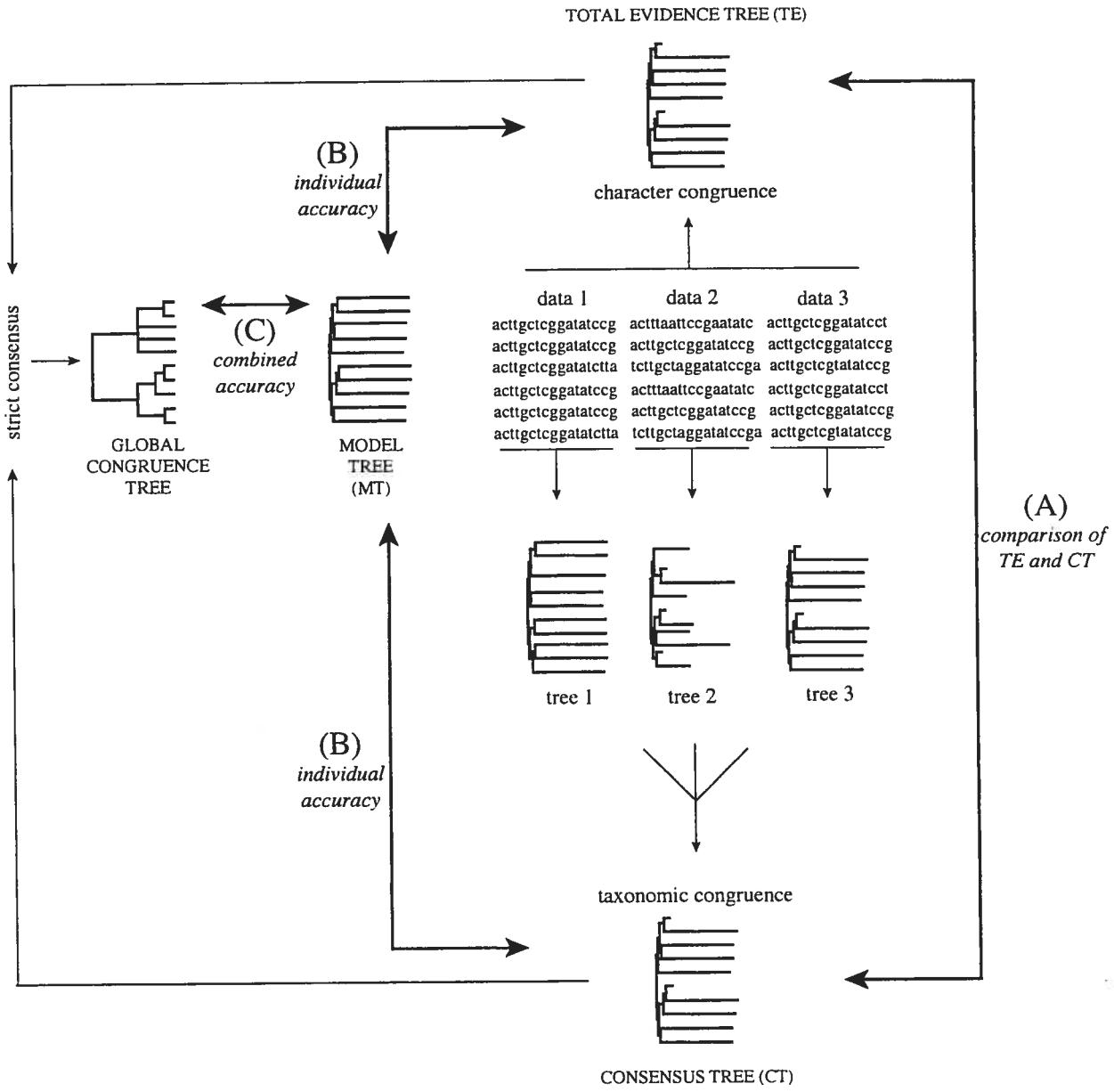

Figure 3.1 Schematic representation of the simulation protocol, where (A) corresponds to the global congruence (comparison of the total evidence and consensus trees; results in Table 3.II), (B) is the individual accuracy (comparison of the total evidence tree with the model tree and of the consensus tree with the model tree; results in Table 3.III) and (C) refers to the combined accuracy (comparison of the global congruence tree with the model tree; results in Table 3.IV).

unresolved). For *Cmax*, a value of 1 is obtained when the two trees are fully compatible, and a value of 0 is obtained for two incompatible trees (*i.e.*, their semi-strict consensus is unresolved). In the case of two fully resolved trees, *Cmin* and *Cmax* will have the same values. Interestingly, *Cmax* will always be maximal if one of the trees is unresolved and the other is fully resolved, because a bush is always compatible with any resolved topology. All indices were computed using PAUP*.

Phylogenetic accuracy

A phylogenetic method is said to be accurate if it can recover the correct topology of the tree that was used to generate the sequences. Individual accuracy values were thus computed by recording the number of replicates for which a given method recovered the model tree (MT), divided by 1000. Because not all methods are able to perfectly recover the correct tree, individual compatibility values are also computed to determine whether the estimated trees contradict the true relationships in the model tree. These individual compatibility results were obtained by counting the number of times that the estimated trees where compatible with the model, tree divided by 1000, for each of the competing methods.

We postulated that when total evidence (TE) and consensus trees (CT) converge to the same topology, that tree is more likely to be accurate. In such cases, absolute accuracy values were measured by counting the number of replicates for which the two trees were identical to the model tree (TE=CT=MT). Combined accuracy rates were then computed by dividing this number by the number of replicates for which the total evidence and consensus trees were identical to one another (TE=CT).

Topological compatibility of the global congruence approach was assessed to also account for replicates for which the total evidence and consensus trees were different. To do so, a strict consensus of the competing trees was derived and compared to the model tree, and the number of replicates for which that strict consensus was compatible with the model tree was recorded.

A four-way analysis of variance (ANOVA) was computed to test the effect of the different factors (number of taxa, number of partitions, data heterogeneity and various

consensus methods) and their interactions. A generalized linear model with a Gaussian distribution was used for comparing resolution, *Cmin* and *Cmax* indices, whereas a binomial distribution was used for accuracy and compatibility values. All tests were computed using the procedures lm and glm in the R statistical language (R Development Core Team 2004).

RESULTS

Global congruence

Multiway analyses of variance (ANOVA) revealed that all factors individually have a significant effect on the global congruence, for the three different indices (see Table 3.I). However, almost all interactions involving the different consensus methods are also highly significant ($P < 0.00001$) meaning that the effect of consensus methods can modulate the effect of the other factors. For example, the *Cmin* values are larger when combining two partitions for the Adams and strict consensus, whereas *Cmin* values are larger when combining ten partitions for all other consensus methods.

The different consensus trees were compared to the total evidence tree in order to determine whether using a method that takes into account branch lengths provided results more similar to those obtained with character congruence. For each consensus method, mean values of the comparison indices are presented in Table 3.II, along with the mean resolution of the corresponding trees. In all cases, the topological identity values (*Cmin*) were highest for the average consensus, followed by the resolved majority rule. The strict consensus always exhibited the lowest values. The majority rule consensus did better than Adams consensus with 10 partitions, whereas the ranking of these methods was reversed with 2 partitions. In the latter case only, the majority rule and strict consensus trees provided identical results. For all indices, the semi-strict and strict consensus, on the other hand, showed identical values.

The *Cmax* values reported in Table 3.II present topological compatibility of the trees compared. Here, contrary to *Cmin*, the highest values were always obtained for the

Table 3.I Results of multiway analyses of variance (ANOVA) for different indices measuring global congruence, accuracy and compatibility to test the significance of four factors individually and in combination: degree of heterogeneity (heterogeneity), number of partitions (partitions), number of taxa (taxa) and the different methods used (method). Significant results are in bold.

	global congruence			accuracy and compatibility			
factors	*Cmin*	**Cmax**	mean resolution	individual accuracy	individual compatibility	TE=CT=MT	TE=CT
heterogeneity (heterogeneous/homogeneous)	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
partitions (2 or 10)	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
taxa (10 or 30)	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
method (total evidence/average/majority rule resolved majority rule/Adams/strict/semistrict)	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
heterogeneity x partitions	0.14424	**<0.00001**	**0.00060**	**<0.00001**	0.06487	**0.00068**	0.02563
heterogeneity x taxa	0.53994	0.36260	0.08510	0.03761	0.20037	0.16757	0.02351
partitions x taxa	0.06261	**0.00010**	0.86160	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
heterogeneity x method	**<0.00001**	**<0.00001**	**<0.00001**	0.03386	**0.00235**	0.09498	0.13491
partitions x method	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
taxa x method	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**	**<0.00001**
heterogeneity x partitions x taxa	0.52877	0.31126	0.96365	0.08353	0.20763	0.02859	0.64380
heterogeneity x partitions x method	**<0.00001**	**0.00279**	**<0.00001**	0.10608	0.01289	0.07192	0.18819
heterogeneity x taxa x method	0.63935	0.98964	0.65798	0.89047	0.99201	0.83408	0.63208
partitions x taxa x method	0.33633	0.13763	0.93571	**0.01166**	**0.00011**	**0.00052**	**<0.00001**
heterogeneity x partitions x taxa x method	0.52830	0.12706	0.98037	0.99998	0.85021	0.86117	0.64960

TE=CT=MT: total evidence and consensus trees identical to the model tree; TE=CT: total evidence and consensus trees identical to each other.

Table 3. II Global congruence between the total evidence tree and the different consensus trees for (A) homogeneous data and (B) heterogeneous data. Mean values for *Cmin* and *Cmax* are reported. Mean resolution of the trees derived from the various consensus techniques are also shown.

A	2 partitions			10 partitions		
	Cmin	*Cmax*	mean resolution	*Cmin*	*Cmax*	mean resolution
10 taxa						
average	0.964	0.964	1.000	0.979	0.979	1.000
resolved mr	0.921	0.921	1.000	0.972	0.972	1.000
majority rule	0.858	0.999	0.859	0.908	0.999	0.909
Adams	0.869	0.983	0.883	0.738	0.992	0.744
strict/semi	0.858	0.999	0.859	0.710	1.000	0.710
30 taxa						
average	0.953	0.953	1.000	0.974	0.974	1.000
resolved mr	0.906	0.906	1.000	0.960	0.960	1.000
majority rule	0.826	0.999	0.826	0.877	0.999	0.878
Adams	0.848	0.970	0.870	0.718	0.985	0.729
strict/semi	0.826	0.999	0.826	0.676	1.000	0.676

average: average consensus; resolved mr: resolved majority rule consensus; majority rule: majority rule consensus; Adams: Adams consensus; strict/semi: strict and semi-strict consensus

Table 3. II (continued)

B	2 partitions			10 partitions		
	Cmin	*Cmax*	mean resolution	*Cmin*	*Cmax*	mean resolution
10 taxa						
average	0.954	0.954	1.000	0.977	0.977	1.000
resolved mr	0.909	0.909	1.000	0.967	0.967	1.000
majority rule	0.834	0.999	0.836	0.900	0.998	0.902
Adams	0.847	0.974	0.867	0.698	0.987	0.709
strict/semi	0.834	0.999	0.836	0.662	1.000	0.662
30 taxa						
average	0.946	0.946	1.000	0.968	0.968	1.000
resolved mr	0.888	0.888	1.000	0.958	0.958	0.999
majority rule	0.806	0.998	0.808	0.874	0.949	0.876
Adams	0.830	0.960	0.859	0.685	0.978	0.702
strict/semi	0.806	0.998	0.808	0.630	1.000	0.630

average: average consensus; resolved mr: resolved majority rule consensus; majority rule: majority rule consensus; Adams: Adams consensus; strict/semi: strict and semi-strict consensus

strict consensus, whereas the average and resolved majority rule methods exhibited the lowest values. Adams and majority rule consensus trees ranked in between other methods. These discrepancies among the results of different indices can be explained by looking at the resolution of the trees. Indeed, the methods with better compatibility values are also those with the poorest resolution. The product of the resolution by *Cmax* equals *Cmin*. For that matter, the consensus methods that return fully resolved trees (*i.e.*, average and resolved majority rule) have identical values for *Cmin* and *Cmax*.

The effect of the number of taxa is obvious when comparing the results involving 10 or 30 taxa. In all cases, mean values decreased by increasing the number of taxa. Increasing the number of data partitions, however, affected the different methods differently. Whereas the values of tree comparison indices increased with more partitions for average, resolved majority rule and majority rule trees, the opposite trend was observed for Adams and strict consensus trees. Interestingly, the same effect was observed for the resolution, indicating that comparing more trees produced less resolved trees for these more conservative consensus methods. Finally, the data heterogeneity had a negative impact on both tree indices, and resolution of the consensus trees. That is, that homogeneous data always provided consensus trees that are more similar to (and compatible with) the total evidence trees.

Phylogenetic accuracy of the character and taxonomic congruence approaches

Results of the multiway analyses of variance (ANOVA) presented in Table 3.I revealed that all factors have a significant effect on individual accuracy and compatibility indices. Most of the significant interactions do not involve the heterogeneity factor but different combinations of the three other factors: number of partitions, number of taxa and methods. Thus, it seems that the effect of heterogeneity is mostly independent of the presence of a particular level of the other factors. Indeed, the accuracy and compatibility values are always smaller for heterogeneous data sets, regardless of the consensus method, the number of taxa and number of partitions.

Table 3.III presents measures of individual accuracy for the character and taxonomic congruence approaches, obtained by comparing the competing trees to the

Table 3. III Accuracy and compatibility of the trees derived with the various methods with respect to the model tree for (A) homogeneous data and (B) heterogeneous data. All results are based on 1000 replicates.

A	2 partitions		10 partitions	
	accuracy	compatibility	accuracy	compatibility
10 taxa				
total evidence	0.589	0.589	0.787	0.787
average	0.572	0.572	0.761	0.761
resolved mr	0.486	0.486	0.743	0.743
majority rule	0.303	0.878	0.498	0.961
Adams	0.303	0.795	0.090	0.949
strict/semi	0.303	0.878	0.090	1.000
30 taxa				
total evidence	0.085	0.085	0.333	0.333
average	0.072	0.072	0.265	0.265
resolved mr	0.037	0.037	0.223	0.223
majority rule	0.006	0.597	0.033	0.845
Adams	0.006	0.274	0.000	0.711
strict/semi	0.006	0.597	0.000	1.000

total evidence: combined analysis; average: average consensus; resolved mr: resolved majority rule consensus; majority rule: majority rule consensus; Adams: Adams consensus; strict/semi: strict and semi-strict consensus

Table 3. III (continued)

B	2 partitions		10 partitions	
	accuracy	compatibility	accuracy	compatibility
10 taxa				
total evidence	0.538	0.538	0.768	0.768
average	0.511	0.511	0.751	0.751
resolved mr	0.430	0.430	0.737	0.737
majority rule	0.247	0.868	0.476	0.964
Adams	0.247	0.748	0.052	0.926
strict/semi	0.247	0.868	0.052	1.000
30 taxa				
total evidence	0.067	0.067	0.317	0.317
average	0.054	0.054	0.261	0.261
resolved mr	0.028	0.028	0.233	0.233
majority rule	0.002	0.607	0.037	0.863
Adams	0.002	0.253	0.000	0.594
strict/semi	0.002	0.607	0.000	1.000

total evidence: combined analysis; average: average consensus; resolved mr: resolved majority rule consensus; majority rule: majority rule consensus; Adams: Adams consensus; strict/semi: strict and semi-strict consensus

model tree. In general, the results show that there is a difference between the results of the different methods. Whereas total evidence trees are more accurate than consensus trees, average consensus trees accounting for branch lengths are more accurate than other consensus methods based on topological relationships alone. The resolved majority rule is the most accurate technique among those ignoring branch lengths, in all situations, whereas the majority rule, Adams and strict consensus methods always provide the worst accuracy values. This is in agreement with the findings of Bininda-Emonds (2002) who found that the resolved majority rule consensus trees were the most similar to the model tree compared to the Adams, majority rule, semi-strict or strict consensus. The relative difference between average consensus trees and total evidence trees are generally less important than the differences with other consensus methods, however. Compatibility results of the competing approaches are also provided in Table 3.III. Here again, the methods that produce less resolved trees usually exhibit higher compatibility values. For that reason, we observe that the majority rule, Adams, and strict consensus trees are more often compatible with the model tree than the other consensus methods. Indeed, in extreme cases involving 10 partitions and 30 taxa, the strict (and semi-strict) consensus trees are compatible with the model tree for all replicates. In this case, it is worth mentioning that a bush tree is compatible with any resolved solution. Interestingly, this result contrasts with the null accuracy values obtained in the same cases, and this shows that less resolved trees are more likely to be compatible with the model tree. Whereas differences were observed in terms of accuracy and compatibility results for these consensus approaches, values for the total evidence, average and resolved majority rule consensus trees were the same for the two indices since these trees were fully resolved. For that reason, the resolved majority rule is the only method based on topological relationships with lower compatibility values.

Table 3.III also shows that the number of taxa affects accuracy. For all methods, increasing the size of the matrix (from 10 to 30 taxa) decreases individual accuracy. Consequently, the best results were always obtained in simulations involving 10 taxa, regardless of the number of data partitions, and the heterogeneity of the data. Compatibility is also affected negatively by this parameter.

The number of partitions also affects accuracy and compatibility values. In the case of total evidence, average, majority rule and resolved majority rule trees, increasing the number of partitions from 2 to 10 increased accuracy by up to 25%. On the other hand, the Adams and strict consensus trees exhibited a different pattern, in which accuracy decreased with an increase in the number of data partitions. This effect was much stronger for smaller matrices, however. Contrary to accuracy, the compatibility results follow the same trend for all methods. That is, that increasing the number of partitions (and characters) always increases compatibility with the model tree.

In general, accuracy values were somewhat higher for homogeneous data, but the differences between simulations involving heterogeneous and homogeneous data were not as great as those observed by comparing replicates with different number of taxa, or different number of partitions.

Phylogenetic accuracy of the global congruence approach

The results of the multiway analyses of variance (ANOVA) for the global congruence accuracy show the same pattern as for individual accuracy and compatibility (Table 3.I). All individual factors have a significant effect and all significant interactions but one do not involve heterogeneity.

Whereas Table 3.III presented accuracy values for individual methods, Table 3.IV shows combined accuracy values of the character and taxonomic congruence approaches used jointly in a global congruence framework. The numbers in Table 3.IV can be looked at from two different angles, however. On the one hand, absolute accuracy values simply report the number of replicates for which total evidence and consensus trees are both identical to the model tree. On the other hand, relative combined accuracy values are computed by dividing the absolute values by the number of replicates for which total evidence and consensus trees were identical to one another (global congruence, TE=CT: Table 3.IV). Interestingly, these two measures show opposite trends and the consensus methods that performed better in terms of absolute accuracy are those that do worse in relative terms. The methods showing high combined accuracy rates are thus exactly those

Table 3.IV Accuracy and compatibility of the global congruence approach for (A) homogeneous and (B) heterogeneous data. The first columns present the number of replicates for which total evidence (TE), consensus (CT) and the model trees (MT) are identical, and the second columns is the number of replicates for which TE and CT were identical. Accuracy values are thus computed as the ratio of (TE=MT=CT)/(TE=CT). The compatibility values are simply representing the number of cases for which TE and CT where compatible with MT, divided by the total number of replicates (1000).

A	2 partitions				10 partitions			
	TE=MT=CT	TE=CT	accuracy	compatibility	TE=MT=CT	TE=CT	accuracy	compatibility
10 taxa								
average	522	778	0.671	0.678	725	870	0.833	0.840
resolved mr	419	588	0.713	0.734	696	821	0.848	0.856
majority rule	303	359	0.844	0.879	495	516	0.959	0.967
Adams	303	359	0.844	0.858	90	90	1.000	0.990
strict/semi	303	359	0.844	0.879	90	90	1.000	0.100
30 taxa								
average	51	309	0.165	0.154	224	505	0.444	0.453
resolved mr	19	83	0.229	0.241	175	341	0.515	0.494
majority rule	6	10	0.600	0.598	32	41	0.780	0.868
Adams	6	10	0.600	0.469	0	0	0.000	0.947
strict/semi	6	10	0.600	0.598	0	0	0.000	1.000

average: average consensus; resolved mr: resolved majority rule consensus; majority rule: majority rule consensus; Adams: Adams consensus; strict/semi: strict and semi-strict consensus.

Table 3.IV (continued).

B	2 partitions				10 partitions			
	TE=MT=CT	TE=CT	accuracy	compatibility	TE=MT=CT	TE=CT	accuracy	compatibility
10 taxa								
average	451	731	0.617	0.642	708	857	0.826	0.831
resolved mr	361	529	0.682	0.709	680	795	0.855	0.855
majority rule	247	292	0.846	0.872	473	488	0.969	0.972
Adams	247	292	0.846	0.841	52	52	1.000	0.997
strict/semi	247	292	0.846	0.872	52	52	1.000	0.100
30 taxa								
average	34	279	0.122	0.170	205	427	0.480	0.487
resolved mr	11	69	0.159	0.272	181	338	0.536	0.515
majority rule	2	5	0.400	0.620	36	42	0.857	0.879
Adams	2	5	0.400	0.481	0	0	0.000	0.934
strict/semi	2	5	0.400	0.620	0	0	0.000	1.000

average: average consensus; resolved mr: resolved majority rule consensus; majority rule: majority rule consensus; Adams: Adams consensus; strict/semi: strict and semi-strict consensus.

that are more conservative, and for which fully resolved trees are seldom obtained, let alone being identical to total evidence and model trees (TE=CT=MT: Table 3.IV). For example, a relative accuracy of 100% is obtained for Adams and strict consensus trees in simulations involving 10 taxa and 10 data partitions, but the absolute number of replicates for which this perfect score was obtained is far less than what was obtained with all of the other consensus methods.

Clearly, the effect of the number of taxa is very important and it affects the global congruence and absolute accuracy. Those values are decreased when adding more taxa. The number of data partitions also has a effect on accuracy, however. Except for two consensus methods (Adams and strict consensus), combining more trees increased global congruence and accuracy. As before, the heterogeneity of the data partitions affected the results negatively. Regardless of the parameters, however, combined accuracy rates (Table 3.IV) were always larger than individual rates computed for the corresponding consensus methods (Table 3.III).

Table 3.IV also presents combined compatibility values for the global congruence approach. These results show that irrespective of the differences or similarities between total evidence and consensus trees, using these approaches jointly can increase compatibility values. That is, that the strict consensus of the total evidence and consensus trees is always more (or equally) compatible with the model tree than any one of these individual trees. Furthermore, relative compatibility rates exhibit even better scores when considering only the replicates for which consensus and total evidence trees are different. This can imply two things: (1) that the topological differences observed by comparing the results of taxonomic and character congruence usually correspond to short internodes, and that collapsing these branches in a strict consensus does not contradict the true relationships in the model tree, or (2) that the trees compared are highly conflicting, thereby making the consensus highly unresolved and thus compatible with the fully resolved model tree. The effects of the simulation parameters on the compatibility results are the same as those observed for accuracy values.

DISCUSSION

In the present paper, we addressed specific questions concerning the relationships and accuracy of character congruence, taxonomic congruence and global congruence approaches. We were first interested on testing the hypothesis that average consensus trees that account for branch lengths would be more similar to total evidence trees than other consensus method that account for topological relationships alone would be. Our simulations have confirmed this assertion and corroborated previous results obtained by Levasseur & Lapointe (2001) based on real data sets. The results of topological identity values (*Cmin*) were indeed higher for average consensus trees than for other consensus methods. Because they were also more resolved, these consensus trees were more similar to the total evidence trees, and thus more informative (Thorley *et al.* 1998). The resolved majority rule trees were also more resolved and they showed *Cmin* values similar to those obtained for the average consensus trees, for the same reason. However, good resolution turned out to be bad when we measured tree compatibility. As a matter of fact, two fully resolved trees would need to be identical in every aspect to reach a perfect compatibility score of 1, whereas a bush, which is completely unresolved, would be compatible with any other topology. The *Cmax* values exactly show this, for consensus methods with lower resolution are those with higher compatibility scores. In some extreme cases, the strict (and semi-strict) consensus trees exhibit perfect compatibility with the corresponding total evidence trees, but at the same time, they also have low resolution. In contrast, average consensus trees computed under the same conditions were always fully resolved, but compatible still with the total evidence trees in 97% of the replicates as opposed to 100%.

We then assessed the accuracy of character and taxonomic congruence by determining whether the topology of a model tree could be recovered by the competing approaches with simulations. Our results show that total evidence trees were always superior to every consensus method tested in terms of accuracy, and that average consensus trees outperformed the topological approaches in most cases. Nevertheless, the conservative techniques provided better results in terms of compatibility with the model tree. Once again, resolution is a factor explaining these results, and this clearly shows that depending on the criterion, consensus methods can do better than total evidence, or worse. The fact is that accuracy cannot improve by using *Cmax* rather than *Cmin* when the trees compared are

fully resolved because both indices are identical in this case. When the consensus trees are partially resolved, however, and when the relationships accepted by those trees are not contradicting the true relationships in the model tree, compatibility values can increase with respect to accuracy. This is what we obtained in our simulations with Adams, majority rule and strict consensus methods.

Finally, we also used simulations to determine whether combining the results of character and taxonomic congruence in a so-called global congruence approach (*sensu* Lapointe *et al.* 1999) could increase accuracy values. Our results clearly confirmed this hypothesis, and they also showed that compatibility is increased by combining methods. The joint use of total evidence and consensus methods actually improved accuracy in the vast majority of conditions simulated. When two different philosophical approaches converge to the same tree, it seems more likely that the common relationships in those trees are true (see also Kim 1993). Whereas absolute accuracy values of the global congruence approach tell us that liberal consensus methods have a tendency to produce trees identical to model trees more often (*i.e.*, up to one order of magnitude), the relative numbers show that the conservative methods are more often correct. It is important to note, however, that replicates for which the strict and Adams consensus were accurate in terms of global congruence were always a subset of the replicates for which average consensus trees were accurate. It is also of interest that conservative methods can fail miserably under specific conditions, and that the number of replicates satisfying the global congruence hypothesis can be quite low when larger trees are combined. The question then boils down to whether absolute or relative accuracy values are to be preferred. Since the methods that show higher numbers in term of relative accuracy exhibit very low absolute numbers, and since for those replicates other methods also recovered the model tree, it seems sensible to prefer the absolute accuracy values.

Nevertheless, combination of a larger number of methods could increase accuracy even more, but at the price of a lower resolution. For example, the relative accuracy of the global congruence combining the results of total evidence, average consensus, and resolved majority rule consensus trees are higher than those obtained when using any two of these methods jointly (results not shown). A stepwise procedure can thus be defined to successively collapse the nodes that are not recovered by the combination of an additional

method. Further simulations would be required to test the hypothesis that these nodes will disappear in direct relationship with their bootstrap support in the combined data, correlated in turn with the corresponding branch lengths in the model tree. Collapsing these poorly supported nodes could increase compatibility, as we have shown in a previous study (Levasseur and Lapointe, 2001). Using real data sets, we demonstrated that total evidence trees became more compatible with the consensus trees when all branches with bootstrap values less than 50% were collapsed. The same rationale could also be employed to collapse the poorly supported nodes in separate analysis, using validation methods for consensus trees (Lapointe & Cucumel 2003).

Multiway analyses of variance were computed to test the effect of different factors on global congruence and on accuracy. However, it is difficult to fully interpret the results of individual factors in the presence of significant interactions. To better understand the effect of each factor, a series of two-way ANOVAs could be done. The current results thus prescribe a direct examination of the global congruence, accuracy and compatibility indices values presented separately in the different tables of this chapter.

With respect to the factors investigated in our simulations, the competing consensus techniques can thus be divided in two distinct groups: the liberal methods that produce fully resolved trees, and those that are more conservative and that may lead to less resolved topologies. Whereas the first group, including the average, resolved majority rule and majority rule consensus methods, are affected positively by an increase in the number of characters, the second group of methods, including Adams and strict consensus, are affected negatively by the same parameter. That is, that individual accuracy values for the first group are higher in simulations involving 10 partitions (20 000 characters) than those based on 2 partitions only (4000 characters), while the reverse pattern is observed for the second group. Interestingly, total evidence falls within the same group as the liberal consensus methods that produce resolved trees and this shows why proponents of the character congruence have always considered this approach as more informative than the competing taxonomic congruence approach (Barrett *et al.* 1991; De Queiroz *et al.* 1995; Hillis 1987; Kluge & Wolf 1993; Nixon & Carpenter 1996). The results of our simulations clearly show that the type of consensus methods selected to combine the results of separate analysis makes a big difference. Accurate phylogenetic hypotheses can be obtained equally

often (almost) with total evidence and consensus trees, by considering branch lengths, and/or by using methods that produced resolved trees. This conclusion is even more telling when more trees are combined, because conservative topological methods become less and less resolved as more trees are combined.

The effect of the number of taxa is easier to explain and it confirms the results obtained by Bininda-Emonds & Sanderson (2001) in a different context. Adding more taxa means an increasing chance of making a mistake. In all cases considered, larger trees were always less accurate and less compatible with the model tree than smaller ones. This result is very important for the future of systematics and the search of the Tree of Life (Mace *et al.* 2003). We further suggest that accuracy is more likely to decrease in the supertree setting when trees bearing overlapping sets of leaves are combined (see Bininda-Emonds & Sanderson 2001).

Data heterogeneity had a negative effect on accuracy. In a previous series of simulations, Levasseur & Lapointe (2003) have shown that combining heterogeneous data with the total evidence approach considerably decreased accuracy. The present study partially contradicts these results showing instead that data heterogeneity slightly decreases accuracy, independently of the other factors. Two important issues are worth mentioning. Previous results were based on a more extreme case of data heterogeneity. Both evolutionary rates and branch lengths were varied. On the other hand, those simulations were based on a single model tree adapted from Kumar (1996), whereas the present simulations used 1000 replicate trees. This raises a concern about the generalization of simulation studies based on few carefully selected topologies (e.g., Hillis *et al.* 1994). We believe that the patterns observed in the present study are more likely to be correct, but further simulations are badly needed to further investigate the effect of combining data partitions with different phylogenetic histories, as it is often the case with real data. In theory, the maximum likelihood and Bayesian approaches can account for different models of evolution in the analysis of combined data, but average consensus methods may also be adapted to correct the branch lengths of the separate trees estimated using different models prior to their combination. Using these refined methods with real data, we postulate that the separate analyses of multiple data sets could provide trees very similar to those obtained by a combined analysis.

Conclusion

The debate opposing the proponents of combined and separate analyses has been biased by the use of a conservative consensus method that produce trees that are usually less resolved than the corresponding total evidence trees. We clearly showed, here again, that the difference between character and taxonomic congruence is the result of a methodological choice: the choice of a consensus method that accounts for branch lengths, or one that ignores them. When branch lengths are taken into consideration, using the average consensus procedure (for other consensus methods with branch lengths, see Bryant 2003; Lapointe 1998a), character and taxonomic congruence approaches converge more frequently to the same tree. Whereas total evidence trees were more accurate than consensus trees in the simulations we performed, the average consensus always outperformed the topological consensus methods. More importantly, our simulations showed that the use of character and taxonomic congruence in a global congruence approach increases the detection of accurate portions of the competing trees with respect to the model tree. We strongly believe that the joint use of competing approaches can only be beneficial for the estimation of phylogenetic trees. Furthermore, this global approach provides an interesting framework to detect common signal and incompatibilities in separate data sets.

Acknowledgements

We thank Émilie-Liên Bui and Olivier Gauthier for their help with the statistical analyses done with the R language. This work was supported by a NSERC grant to F.-J. L. and a NSERC and FQRNT scholarship to C.L.

CHAPITRE 4

Total evidence, average consensus and MRP : what a difference distances make

Cet article sera soumis prochainement:

Levasseur, C., & F.-J. Lapointe Total evidence, average consensus and MRP : what a difference distances make.

ABSTRACT

Matrix representation with parsimony (MRP) is a method that can be used to combine trees in the supertree and consensus settings. Although this approach is related to taxonomic congruence, it is not yet clear whether MRP is really a consensus method or whether it behaves more like the total evidence approach. Previous simulations have shown that it could approximate the total evidence solution, whereas other studies have highlighted similarities with the average consensus in specific conditions. In this paper, we evaluate the hypothesis that MRP could be equally related to both character and taxonomic congruence. We conducted a simulation study to evaluate the accuracy of both approaches with that of MRP and compared the solutions to one another. Our results show that the total evidence trees are more accurate than average consensus trees and that both are better than MRP trees. The accuracy rate of all methods was similarly affected by an increase in the number of taxa and a reduction in the number of data partitions. Also, our results confirm that MRP is no more distant from one approach than the other.

INTRODUCTION

Matrix representation with parsimony (MRP) is the most commonly used method to construct supertrees from molecular data, but it can also be applied in the consensus setting (*sensu* Bininda-Emonds 2003) when trees combined have the same leaf sets. Although MRP combines trees rather than the primary characters, it does not represent a consensus method per se (Bininda-Emonds & Bryant 1998), and fundamental differences exist between MRP and taxonomic congruence. The simulation study of Bininda-Emonds & Sanderson (2001) rather showed that MRP could be considered a good approximation of total evidence. Pisani & Wilkinson (2002) described this relationship as superficial, however, and they showed that MRP is more closely related to taxonomic congruence than it is to character congruence. Interestingly, the differences between these alternative approaches are not as important as they seem when the data and trees are treated in a coherent fashion and when a consensus method that takes into account branch lengths is used to combine the trees (Lapointe *et al.* 1999). Indeed, Levasseur & Lapointe (2001) have shown that character and taxonomic congruence can provide very similar trees by using the average consensus (Lapointe & Cucumel 1997) to combine the results of separate analyses. More recently, Lapointe *et al.* (2003) further demonstrated that in the consensus setting, when branch lengths are set to one, there exists a close relationship between MRP and average consensus, and that both methods are related consensus techniques. However, MRP has never been directly compared to the average consensus as a means of combining trees, with or without branch lengths. In this paper, we investigate the relationships between MRP, average consensus and total evidence. With simulations, we compare the relative accuracy of character and taxonomic congruence, using either MRP or two variants of the average consensus to combine the results of separate analyses. We then compare the competing approaches with one another to determine whether MRP trees are more closely related to consensus or total evidence trees.

METHODS

Generating the data

The simulations follow the protocol described in Levasseur and Lapointe (Chapter 3). Briefly, model trees (MT) were generated with a Yule branching process using the program r8s (Sanderson 2003) and molecular sequences of fixed lengths (2000 base pairs) were evolved on those trees using the program Seq-Gen (Rambault & Grassly 1997). The evolution of sequences was performed according to a Jukes-Cantor model (Jukes & Cantor 1969) with a site-to-site heterogeneity rate (shape parameter set to 0.5). To simplify the computations and limit the number of simulations, the number of taxa was restricted to either 10 or 30, and the number of data partitions was restricted to either 2 or 10. Homogeneous and heterogeneous data sets were also generated for comparison purposes. To do so, the sequences were respectively generated using replicate trees with identical topologies and identical branch lengths (homogeneous data), or replicate trees with identical topologies and random branch lengths (heterogeneous data). Each of these sequence is considered to be a data partition. Because each sequence is of the same length, this represents the case of combining two molecular data sets. For every combination of parameters, 1000 model trees were generated to simulate the evolution of data partitions.

Tree construction

The data representing different partitions were treated either jointly or independently to estimate total evidence trees and separate trees. In all cases, distances were computed with a Jukes-Cantor model matching that used to generate the data, and an unweighted least-squares algorithm (Cavalli-Sforza & Edwards 1967) was employed to estimate the trees using PAUP* (Swofford 1999).

For the MRP analyses, the trees from individual partitions were coded using RadCon (Thorley & Page 2000), and the resulting matrices were combined and analyzed with parsimony using PAUP*, following the protocol described in Bininda-Emonds & Sanderson (2001). When multiple equally parsimonious trees were obtained, the strict consensus of those trees was taken as the MRP solution.

To compute average consensus trees (AC), path-length distances were extracted from the separate trees and the resulting matrices were used to compute an average distance matrix. The average consensus was obtained by applying a least-squares algorithm (Cavalli-Sforza & Edwards 1967) to this average matrix. To assess the effect of branch lengths in the computation of the consensus, topological average consensus trees (TAC) were also computed by setting all branch lengths to one prior to the computation of the average matrix. Both variants of the average consensus were computed with PHYLIP using the Fitch algorithm (Felsenstein 1993).

Tree comparisons

Phylogenetic accuracy and compatibility

To begin with, the total evidence (TE), MRP and average consensus trees (AC and TAC) were compared with the model tree (MT) onto which the sequences were evolved to assess the performance of the different methods. Accuracy and compatibility rates were computed to do so. Binary coefficients that directly capture phylogenetic accuracy and compatibility of the different methods were compiled. The first one takes a value of 1 when the trees compared are topologically identical and a value of 0 otherwise. In this case, only identical fully resolved trees can obtain the maximum value. Accuracy rates are obtained by simply counting the number of values equal to 1 divided by 1000 replicates. Because some methods do not always return fully resolved solutions, compatibility rates were also computed. For this second coefficient, a value of 1 is obtained when the resolved clades of the solution does not contradict the model tree and a value of 0 otherwise. We counted the number of values equal to 1 on 1000 replicates to compute compatibility rates.

Comparison of the different methods

The competing trees were compared with one another to compare the results of character and taxonomic congruence. Absolute topological identity was measured by counting the number of times that results from two methods were identical for the same replicate. Also, indices of topological identity (*Cmin*) and topological compatibility (*Cmax*) were computed. *Cmin* and *Cmax* were obtained respectively by deriving the consensus fork index (Colless 1980) of the strict and semi-strict consensus of the two trees compared. This

is the proportion of possible clades ($n - 3$, for unrooted trees) that is resolved on the consensus tree. The maximum value of 1 is obtained with the topological identity index when the two trees are identical and their strict consensus is fully resolved. On the other hand, a maximum value of 1 can be obtained with the topological compatibility index when the semi-strict consensus of the trees compared is fully resolved. Consequently, the two indices will have identical values when comparing pairs of fully resolved trees. The minimum value of 0 is obtained in both cases, when the corresponding consensus tree is unresolved (it is a bush). The mean values of *Cmin* and *Cmax* computed over the 1000 replicates are reported.

RESULTS

Comparisons with the model tree

Comparison of trees obtained for the different methods with the model tree (MT) are reported in Table 4.I, for different numbers of taxa and data partitions, and for homogeneous and heterogeneous data sets. The results for individual accuracy rates indicate that the total evidence approach always provides the highest accuracy, whereas MRP performs the worst. Average consensus trees (AC) do almost as good as total evidence trees when branch lengths are accounted for, however. Topological average consensus (TAC) recover the model tree less often, but performs better than MRP. Interestingly, these methods tend to provide less resolved trees (results not shown). This can have an influence on the results of this metric, since a tree that is not fully resolved will never be considered identical to the model tree. The competing approaches are all affected identically by the number of taxa and data partitions. Whereas increasing the size of the matrix (from 10 to 30 taxa) decreases accuracy rates, increasing the number of data partitions (from 2 to 10) increases accuracy rates. The best results are thus obtained for 10 taxa and 10 data partitions. In addition, the results for homogeneous data sets are always better than those based on heterogeneous data sets, for all methods.

The individual compatibility rates provide results on a par with those observed for accuracy, except for MRP (Table 4.I). The compatibility results are higher in that case, because MRP trees, while not always resolved, can still be compatible with the model tree.

Table 4.I Individual accuracy rates and individual compatibility rates between the model tree (MT) and the total evidence tree (TE), the average consensus tree (AC), the topological consensus tree (TAC) and the matrix representation with parsimony tree (MRP) for (A) homogeneous data and (B) heterogeneous data.

A	2 partitions		10 partitions	
	Accuracy	Compatibility	Accuracy	Compatibility
10 taxa				
TE	589	589	787	787
AC	572	572	761	761
TAC	477	477	730	730
MRP	303	870	668	827
30 taxa				
TE	85	85	333	333
AC	72	72	265	265
TAC	26	26	187	187
MRP	6	546	118	429

Table 4.I (continued)

B	**2 partitions**		**10 partitions**	
	Accuracy	Compatibility	Accuracy	Compatibility
10 taxa				
TE	538	538	768	768
AC	511	511	751	751
TAC	419	419	729	729
MRP	247	860	655	840
30 taxa				
TE	67	67	317	317
AC	54	54	261	261
TAC	20	20	223	223
MRP	2	547	125	475

Table 4.II Results of pairwise comparisons of total evidence (TE), matrix representation with parsimony (MRP), average consensus (AC) and topological average consensus (TAC) trees. Absolute topological identity index is reported on the first line, *Cmin* is reported on the second line and *Cmax* on the third for (A) homogeneous and (B) heterogeneous data. Upper triangle is for cases involving two partitions and lower triangle for cases with 10 partitions.

A

10 taxa	TE	AC	TAC	MRP
TE		778 0.964 0.964	606 0.931 0.931	359 0.859 0.998
AC	870 0.979 0.979		637 0.938 0.938	359 0.86 0.999
TAC	804 0.969 0.969	851 0.978 0.978		359 0.86 1.000
MRP	721 0.953 0.984	745 0.957 0.989	816 0.969 1.000	

30 taxa	TE	AC	TAC	MRP
TE		309 0.953 0.953	86 0.914 0.914	10 0.831 0.996
AC	505 0.974 0.974		134 0.928 0.928	11 0.832 0.997
TAC	321 0.959 0.959	456 0.971 0.971		10 0.834 0.999
MRP	183 0.935 0.984	221 0.941 0.99	294 0.949 0.998	

Table 4.II (continued)

B

10 taxa	TE	AC	TAC	MRP
TE		731 0.954 0.954	525 0.914 0.914	293 0.837 0.998
AC	857 0.977 0.977		566 0.925 0.925	295 0.839 0.999
TAC	802 0.968 0.968	845 0.977 0.977		297 0.839 1.000
MRP	702 0.947 0.984	735 0.954 0.991	785 0.962 0.999	

30 taxa	TE	AC	TAC	MRP
TE		279 0.946 0.946	63 0.902 0.902	5 0.812 0.994
AC	427 0.968 0.968		97 0.921 0.921	5 0.815 0.998
TAC	327 0.957 0.957	467 0.971 0.971		5 0.816 0.999
MRP	173 0.932 0.983	211 0.94 0.991	265 0.947 0.998	

All other methods produce fully resolved trees, however, such that identical values are obtained for both indices. Consequently, the same rankings as those reported for individual accuracy rates apply to total evidence and average consensus trees. Here again, the best results are obtained for 10 taxa and 10 data partitions, as well as for homogeneous data sets.

Pairwise comparisons of competing approaches

Table 4.II shows the results of the pairwise comparisons of total evidence, MRP and average consensus (AC and TAC) trees. On average, the absolute topological accuracy index reveals that total evidence trees are more closely related to average consensus trees than they are to MRP trees. Accounting for branch lengths does make a difference in terms of tree comparisons, however. The two forms of average consensus are not always producing identical topologies, and the consensus trees with actual branch lengths (AC) are more closely related to total evidence than consensus trees with all branch lengths set to one (TAC). The MRP trees, on the other hand, are the most different from the total evidence trees, in terms of absolute topological identity. They rather seem to be related to the topological variant of average consensus trees (TAC).

The *Cmin* and *Cmax* indices that represent respectively mean values of topological identity and topological compatibility show the same general trend for all methods and all combinations of parameters, with the exception of MRP. The latter exhibits a different pattern for the same reason as explained above. In general, the conclusions of these results also mirror those obtained when comparing the model tree to the competing trees. That is, that increasing the number of data partitions and decreasing the number of taxa provide better results. Similarly, higher values are obtained for all indices, in simulations based on homogeneous data.

DISCUSSION

In the present paper, we have assessed the relationships and the ability to correctly recover a known model tree of alternative approaches for treating multiple data sets in phylogenetic analysis. We tested through simulations the effect of the number of taxa, number of data partitions and heterogeneity among data sets to see how the competing

methods would perform under different combination of these parameters. We wanted to determine whether MRP would behave like character or taxonomic congruence. We were also interested in comparing average consensus trees obtained by using actual branch lengths or by setting all branch lengths to one prior to the computation.

The results clearly show that under the conditions investigated with our simulations, combined analysis of all data provide more accurate trees than separate analysis, regardless of the method employed to combine the trees. This is not surprising given the fact that the data partitions were of the same size. Partitions of unequal size could have been used to represent the case of combining different types of data (e.g. morphological and molecular data). In this particular situation, different results could have been obtained. But in the case we have tested, average consensus trees do almost as good as total evidence, and much better than MRP trees. Moreover, accounting for branch lengths greatly improves the performance of the consensus approach. This particular finding confirms the study by Levasseur and Lapointe (2001) who demonstrated with actual data sets that total evidence and consensus trees can be similar (or identical) when treated in a coherent fashion, using the average consensus (see also Lapointe *et al.*, 1999). The use a weighted version of MRP that accounts for branch lengths (or bootstrap support) in future simulations could probably improve the results obtained in the present simulation study (see Baum 1992; Purvis 1995; Ronquist 1996; Bininda-Emonds & Bryant, 1998). It would be interesting to see whether the weighted version of MRP would be more closely related to total evidence or average consensus. In fact, these simulations are needed to confirm if branch lengths really matters in this particular case.

Our results also show that the different approaches investigated are affected in the same way by the factors tested in this study. Better accuracy is always obtained with more partitions, fewer taxa, and homogeneous data, in agreement with the conclusions of previous simulation studies (Bininda-Emonds 2003; Bininda-Emonds & Sanderson 2001). For one, adding more characters increases the number of informative sites, which in turn increases the phylogenetic signal. For that matter, all simulations based on 10 partitions (20 000 characters) provided much better results than those based on 2 partitions (4000 characters). On the other hand, using fewer taxa reduces the number of possible trees, and

this may also reduce the probability that a given method will estimate the wrong tree. For large trees estimated with few characters, the problem is accentuated even more because the long branches are subdivided in smaller branches that are more difficult to estimate. This is exactly what we observe with 30 taxa and 2 data partitions. In such cases, the best approach recovers the correct topology of the model tree in less than 10% of the cases studied. Finally, the heterogeneity of the data partitions also affects accuracy by adding noise to the phylogenetic signal. Although our simulations were based on model trees with identical topologies, randomizing the branch lengths had a negative impact on the accuracy rates. Previous simulations (Levasseur & Lapointe 2003) have shown that changing the rate of evolution could worsen the results. In practice, this problem could be worse when data partitions with different phylogenetic histories are combined.

The observation that increasing the number of taxa decreases accuracy is not comforting, since these methods can be used to construct supertrees and hence, bigger trees. Bininda-Emonds & Sanderson (2001) showed in their simulations study that a reduction of the overlap among the trees combined greatly decreases accuracy. When heterogeneous data partitions representing overlapping sets of leaves are combined, the effect is even more dramatic (Lapointe & Levasseur 2004). We believe that these results are of great importance, since our simulations only dealt with trees bearing identical sets of leaves. It would be of interest to further test this with simulations involving overlapping trees with different topologies, and data partitions of unequal sizes.

We have shown elsewhere (Levasseur & Lapointe 2001; Chapter 3) that average consensus trees are usually more similar to total evidence trees than those derived from consensus methods that ignore branch lengths (e.g., strict, majority rule, Adams). The present study corroborates these results by showing that accounting for branch lengths makes a difference, even when the same consensus method is employed. We also show that MRP trees are, in all cases considered, further from total evidence trees than either form of the average consensus. This may be explained by the fact that MRP trees are not always fully resolved, unlike average consensus trees (Levasseur & Lapointe 2001). Furthermore, we observe that when actual branch lengths are ignored in the computation of average consensus trees, the resulting trees become increasingly similar to MRP trees. In these

conditions, the MRP and average consensus are closely related consensus techniques, as suggested by Lapointe *et al.* (2003).

It is clear from our simulations that MRP trees are equally distant from total evidence and consensus trees. Thus, we cannot claim once and for all that MRP represents a character congruence or a taxonomic congruence approach. There seems to be a closer relationship between MRP and average consensus trees when all branch lengths are set to one, however. But even then, the consensus approach outperforms MRP in terms of accuracy. Still, the vast majority of published supertrees are based on this particular procedure (Grenyer & Purvis 2003; Jones *et al.* 2002; Kennedy & Page 2002; Salamin *et al.* 2002). In his paper comparing MRP with several topological consensus methods, Bininda-Emonds & Sanderson (2001) mentions that MRP solutions can contain novel clades that are not found in any of the input trees combined. Average consensus trees also share this property (Wilkinson *et al.* 2004). If this is true, alternative methods that preserve the information embedded in the separate trees (Page 2002; Semple & Steel 2000), at the expense of resolution, could be preferred if one consider this property a problem. But this is not necessarily the case. As such, total evidence too can contain clades that are not found in the source trees. This has been explained by the fact that support for certain clades could emerge only when multiple data sets are combined. In such circumstances, the joint use of consensus with branch lengths and total evidence is certainly a method of choice to deal with multiple data sets (see Chapter 2 and 3). Also, simulations study could be done to investigate the option of using total evidence and MRP jointly or MRP and average consensus.

ACKNOWLEDGEMENTS

This work was supported by a NSERC grant to F.-J. L. and a NSERC and FQRNT scholarship to C.L.

CHAPITRE 5

Incomplete Distance Matrices, Supertrees and Bat Phylogeny

Cet article est publié sous la référence :

Levasseur, C., Landry, P.-A., Makarenkov, V., Kirsch, J.A.W., & Lapointe, F.-J. 2003 Incomplete distance matrices, supertrees and bat phylogeny. *Molecular Phylogenetics and Evolution* **27**, 239-246.

ABSTRACT

In this paper, we evaluate the relative performance of competing approaches for estimating phylogenies from incomplete distance matrices. The *direct* approach proceeds with phylogenetic reconstruction while ignoring missing cells, whereas the *indirect* approach proceeds by estimating the missing distances prior to phylogenetic analysis. Two distinct indirect procedures based on the ultrametric inequality and the four-point condition are further compared. Using simulations, we show that more reliable results are obtained when such indirect methods are used. Expectedly, the phylogenies become less accurate as the percentage of missing cells increases, but combining different estimation methods greatly improves the accuracy. An application to bat phylogeny confirms the results obtained in the simulation study and illustrates the effect of missing distances in the construction of supertrees.

INTRODUCTION

With the recent advances in molecular biology, data sets amenable to phylogenetic analysis are accumulating at a remarkably high rate. The inference of evolutionary relationships from these data and the combination of multiple phylogenies representing overlapping sets of species are among the problems that systematists are facing today. As a consequence, missing data are increasingly common in phylogenetic analysis. For example, this problem arises when large phylogenies are assembled by combining data sets published from different sources. In these cases, it is not always possible to obtain complete data for all species and it would be of interest to use the partial yet available information to derive accurate phylogenies. While some parsimony-based methods can handle missing information (Wiens 1998b), the reconstruction of phylogenies from distance matrices usually requires complete matrices (see Swofford *et al.* 1996). Although that issue could be avoided altogether by only using character-based techniques for supertree constructions (Sanderson *et al.* 1998), the problem remains when path-length distance matrices are combined to build supertrees with branch lengths (see Lapointe & Cucumel 1997). More specifically, experimental techniques such as comparative serology and DNA-hybridization produce distance data that can only be analyzed using the corresponding distance algorithms. The recent popularity of microarray hybridization data (Gibson 2002) is not free of the same problem (Troyanskaya *et al.* 2001), as the clustering methods generally employed to summarize these data require complete distance matrices as well (Quackenbush 2001).

As a solution to the problem of missing distances, some authors (De Soete 1984a; De Soete 1984b; Landry & Lapointe 1997; Landry *et al.* 1996; Lapointe & Kirsch 1995) have proposed to estimate the missing cells in incomplete distance matrices *prior* to phylogenetic analysis, taking advantage of the mathematical properties of tree metrics (Buneman 1971; Hartigan 1967). We will hereafter refer to this procedure as the *indirect* approach. Alternatively, it has been suggested that a tree could be derived directly, using an algorithm that only uses available distances while ignoring the missing ones (Guénoche & Grandcolas 1999; Hein 1989; Makarenkov & Leclerc 1999). This other method will be referred to as the *direct* approach hereafter.

While the two above mentioned approaches are expected to theoretically converge toward similar solutions, this might not necessarily be the case with empirical data. Indeed, distances derived from experimental methods are expected to display a certain degree of deviation from the properties of tree metrics, especially with noisy data, and this could affect the success of the indirect approach. On the other hand, the direct approach only uses partial information to build trees, which may bias the results of phylogenetic analyses.

The objective of this paper was to evaluate the performance of the two methods in order to identify the best algorithm to use with empirical data to reconstruct phylogenies from incomplete distance matrices. The competing approaches were first evaluated using a simulation framework; in a second step, they were also applied to recover a phylogeny of bats from an incomplete set of empirical distances. This example further demonstrated how distance matrices representing overlapping sets of taxa can be combined to construct a supertree with branch lengths. The resulting bat supertree was further used to address the question of the monophyly of microbats, a question that has been much debated in the recent years (Hutcheon *et al.* 1998; Teeling *et al.* 2000).

METHODS

The indirect approach

There exists a one-to-one correspondence between ultrametric trees (*i.e.* trees in which all leaves are equidistant from the root) and their corresponding path-length matrices satisfying the ultrametric inequality (Hartigan 1967), which states that:

$$d_{ij} \leq (d_{ik} + d_{kj}), \text{ for every triplet } i, j, k. \qquad \text{(Eq. 5.1)}$$

This particular tree model assumes a constant rate of evolution across all lineages, and is therefore appropriate to accommodate circumstances where evolution is believed to respect a molecular clock. However, it is possible to relax this constraint to obtain a tree model allowing for unequal evolutionary rates, the so-called four-point condition (Buneman 1971), where:

$$(d_{ij} + d_{kl}) \leq \max [(d_{ik} + d_{jl}); (d_{il} + d_{jk})], \text{ for every quadruplet } i, j, k, l. \qquad \text{(Eq. 5.2)}$$

Ultrametric trees are a special case of additive trees. The first model is a restriction of the more general additive model, which is more flexible and thus very appealing.

Assuming that the distances under investigation satisfy one of these mathematical models (at least the additive property), the properties of tree metrics can be employed to estimate missing cells in a distance matrix. For one, every missing distance d^*_{ij} could be accurately estimated using the ultrametric property (Eq. 1) by looking at all possible triplets $\{i, j, k\}$ for which d_{ik} and d_{kj} are known (De Soete 1984a; De Soete 1984b; Lapointe & Kirsch 1995). However, because biological data rarely fit the ultrametric model, Landry *et al.* (1996) introduced the additive procedure to estimate missing values, based on the four-point condition (Eq. 2). In that case, each missing distance d^*_{ij} is estimated in turn by considering all possible quadruplets of objects $\{i, j, k, l\}$ for which five of the six pairwise distances among these objects are known. Using simulations, it has been shown that accurate estimates of missing cells and robust phylogenies could be obtained with indirect estimation methods (Landry *et al.* 1996). However, there exists a notable exception for which estimation methods will fail to recover actual distances; it involves missing distances between terminal sister taxa (Landry & Lapointe 1997). In such cases, the best possible solution is to obtain a tree with a trichotomy involving the sister pair. In the worst case scenario, with noisy data, incorrect relationships would be obtained among the sister species.

The direct approach

There exist numerous methods for finding optimal trees from distance data (e.g., Cavalli-Sforza & Edwards 1967; De Soete 1983; Fitch & Margoliash 1967; Rzhetsky & Nei 1992; Saitou & Nei 1987), but most of these techniques only accept complete or nearly complete matrices (see Felsenstein 1993). Weighted least-squares (WLS) methods, however, can be adapted to deal with incomplete distance matrices. In this paper, the MW algorithm developed by Makarenkov and Leclerc (1999) was used to obtain an additive tree that minimizes the following loss function:

$$Q = \sum\sum_{i<j} w_{ij} \, (d_{ij} - \delta_{ij})^2 \rightarrow \min \qquad \text{(Eq. 5.3)}$$

where d_{ij} are the input distances, δ_{ij} the fitted tree-distances, and w_{ij} a set of weights. Depending on the selected weights, this method is equivalent to that of Fitch and Margoliash (1967) when $w_{ij} = 1/d_{ij}$ or to that of Cavalli-Sforza and Edwards (1967) when $w_{ij} = 1$. Other weights can be used to optimize the loss function Q according to different criteria (see Makarenkov & Leclerc 1999). The case of incomplete distance matrices is treated by assigning a null weight ($w_{ij} = 0$) to all missing distances and a unit weight ($w_{ij} = 1$) to the known distances. (Interestingly, the same method can be used with the FITCH program of the PHYLIP package, Felsenstein 1993) by selecting the subreplicate option (S) and by setting the replicates values to one and zero for known and missing values respectively). Consequently, the missing values do not contribute to the loss function Q (Eq. 3) when incomplete matrices are used, and phylogenies can be derived directly with this WLS approach.

Using a series of simulation reproducing the previous designs (Landry & Lapointe 1997; Landry *et al.* 1996; Lapointe & Kirsch 1995), we evaluated the relative performance of the ultrametric and additive indirect approaches in comparison with the direct approach using a weighted least-squares algorithm.

SIMULATION STUDY

Analytical design

For the sake of comparability, we used in this paper the same distance matrices as those selected for the previous simulation studies (see Landry & Lapointe 1997). Two sets of path-length matrices were obtained by means of least-squares algorithms applied to DNA hybridization matrices of varying sizes (n = 7, 9, 11, 13, and 15). The first set represents ultrametric distance matrices satisfying the molecular clock (see Eq. 1), whereas the second set was derived from the same data using an additive tree procedure (see Eq. 2). These ultrametric trees were obtained with the KITSCH program, whereas additive trees were computed with the FITCH program of the PHYLIP package (Felsenstein 1993).

Incomplete matrices were generated by deleting at random a fixed percentage of cells (P = 10% to 60%, by increments of 10%) from the complete distance matrices. For

each percentage P and each set of ultrametric and additive matrices, 100 replicate matrices were generated. These incomplete matrices were analyzed with the direct and indirect procedures, and the trees obtained were then compared to the original phylogenies to assess the relative performance of the competing approaches. To eliminate differences attributable to the use of different tree reconstruction methods, all trees were obtained with either approach using the same WLS algorithm of Makarenkov and Leclerc (1999). Three series of simulations were carried out. (1) First, incomplete matrices were analyzed directly by assigning a null weight to missing cells and a unit weight to all other distances (direct approach). (2) Second, the ultrametric property was used to estimate the missing cells prior to phylogenetic reconstruction (indirect approach). (3) Third, the four-point condition was used to estimate missing cells (indirect approach). In the cases of the indirect approaches, the same least-squares algorithm was applied to the estimated matrices using equal weights ($w_{ij} = 1$) for all distances.

The performance of the competing methods was evaluated in terms of distance and topological recovery. Distance recovery was measured by computing the Pearson correlation (r) between the path-length matrices recovered from the corresponding trees, whereas topological recovery was computed with a standardized version of the Robinson and Foulds (1981) metric, hereafter referred to as RF* (see Landry *et al.* 1996). Because previous studies have confirmed that topological and distance recovery were independent of matrix size (Landry *et al.* 1996 ; Landry & Lapointe 1997), results pertaining to all matrices were pooled for all analyses. Average distance and topological recovery values are thus reported for the different methods as a function of the percentage P of missing cells.

Rates of topological accuracy were also computed to compare the different methods. Individual accuracy rates were reported to indicate the number of times that each single method recovered the correct topology (*i.e.*, the topology isomorphic with the original path-length matrix). Combined accuracy rates were also shown to represent cases for which different methods were jointly accurate (see Kim 1993). Formally, these combined rates were computed as a ratio of the number of times that independent methods (direct or indirect) jointly recovered the *correct* topology over the number of times that they produced the *same* topology. Pairwise combined rates were obtained for every possible

combinations of methods (direct, ultrametric and additive), and a global rate was computed for all three methods used together.

Results

Distance recovery

The average correlations (r) among path-length distance matrices obtained for increasing numbers of missing cells P and different types of matrices (ultrametric and additive) are presented in Table 5.I. These results clearly illustrate that indirect methods outperformed the direct WLS approach in all simulations, except when the indirect ultrametric method is used for additive matrices when $P = 10\%$. Also, the differences among the procedures increased with the proportion of missing data. Expectedly, ultrametric estimations provided better recovery in the case of ultrametric matrices, especially for larger P. On the other hand, the four-point condition performed better in the case of additive matrices, for $P < 50\%$. With few missing cells ($P = 10\%$), both procedures performed equally well on ultrametric matrices. The reverse is not true, however, as ultrametric estimations of additive matrices provided inaccurate distance values.

Topological recovery

The results of topological recovery were in partial agreement with those of distance recovery. Table 5.II illustrates that in the case of ultrametric matrices, the ultrametric estimations provided better recovery values for all P. With few missing cells, however, both estimation methods provided similar recovery values. Again, trees obtained with the direct WLS approach displayed the largest number of topological differences with the original tree. The results for additive matrices were slightly different. As for distance recovery, additive estimations provided the best topological recovery, for all P. Remarkably, the direct approach proved to be better than the ultrametric approach, despite the fact that distances were more accurately estimated in such cases.

Table 5.I Average path-length correlations (r) obtained for the three methods (direct weighted least-squares [dir. wls], indirect ultrametric[ind. ult], indirect additive[ind. add]) for increasing numbers of missing cells P, in simulations based on ultrametric (Ult. mat.) or additive (Add. mat.) matrices

	Ult. mat.			Add.mat.		
P	dir. wls	ind. ult	ind. add	dir. wls	ind. ult	ind. add
10%	0.991	0.995	0.994	0.993	0.990	0.996
20%	0.959	0.981	0.972	0.971	0.971	0.983
30%	0.926	0.963	0.943	0.930	0.949	0.962
40%	0.876	0.940	0.915	0.886	0.917	0.932
50%	0.770	0.889	0.828	0.784	0.854	0.852
60%	0.633	0.840	0.761	0.643	0.801	0.763

Table 5.II Average topological recovery (1 - RF^*) obtained for the three competing methods (direct weighted least-squares [dir. wls], indirect ultrametric[ind. ult], indirect additive[ind. add]) for increasing numbers of missing cells P, in simulations based on ultrametric (Ult. mat.) or additive (Add. mat.) matrices

	Ult. mat.			Add.mat.		
P	dir. wls	ind. ult	ind. add	dir. wls	ind. ult	ind. add
10%	0.964	0.978	0.977	0.964	0.815	0.977
20%	0.860	0.940	0.924	0.865	0.671	0.928
30%	0.723	0.879	0.842	0.736	0.578	0.837
40%	0.568	0.778	0.693	0.595	0.488	0.707
50%	0.410	0.623	0.515	0.464	0.370	0.522
60%	0.292	0.522	0.379	0.343	0.296	0.381

Accuracy

Table 5.III presents individual and combined rates of topological accuracy for the three competing methods. These results indicate that the indirect approach is generally more accurate than the direct approach.

The ultrametric estimation (ult) was the best in simulations involving ultrametric matrices (Table 5.IIIA). For combined rates, the ultrametric method used jointly with the direct approach (dir + ult) provided good accuracy values for $P < 40\%$. However, the best rates were obtained when all three methods (dir + ult + add) were congruent, for up to 40% of missing cells. For larger P, accuracy was increased by combining both indirect methods (add + ult, $P = 50\%$) or when using the ultrametric estimates only (ult, $P = 60\%$). Overall, these results suggested that when the three methods are congruent, the accuracy is always increased compared to the use of unique method or two methods jointly.

The results for additive matrices were also expected (Table 5.IIIB). In all simulations, the estimations based on the four-point condition (add) provided the best individual rates. On the other hand, the ultrametric method (ult) performed very poorly in the simulations based on additive matrices. The highest combined rates were obtained when using the additive procedure and the direct approach (dir + add) were used, except for $P = 10\%$. Global rates were the best for $P \leq 30\%$ and all combined rates (including the global rate) were null for $P = 60\%$. The correct topology was recovered only once in such extreme cases, when using the four-point condition. Regardless of the type of matrices (ultrametric or additive), it thus appears beneficial to use all three methods jointly (dir + ult + add), especially when $P \leq 30\%$.

APPLICATION

The problem of missing data not only occurs in the analysis of single distance matrices (e.g. DNA-hybridization data: Lapointe & Kirsch 1995; microarray data: Troyanskaya *et al.* 2001), but also when multiple data sets (or their corresponding trees) are combined to either derive total evidence or consensus trees and supertrees. To illustrate the analytical procedure, let us consider the combination of two partially overlapping phylogenies depicting the relationships among 12 bat species from four families or

Table 5. III Rates of topological accuracy obtained in simulations based on (**A**) ultrametric or (**B**) additive distance matrices, using the direct weighted least-squares (dir), indirect ultrametric (ult) or indirect additive (add) methods, individually or in combination

A	Ultrametric matrices						
	Individual			Combined			Global
P	dir	ult	add	dir + ult	dir + add	add + ult	dir+ult+add
10%	0.678	0.744	0.734	0.799	0.793	0.746	0.800
20%	0.358	0.520	0.448	0.636	0.597	0.514	0.645
30%	0.184	0.348	0.272	0.574	0.481	0.387	0.622
40%	0.054	0.210	0.112	0.294	0.145	0.316	0.355
50%	0.006	0.090	0.046	0.036	0.028	0.178	0.100
60%	0.002	0.048	0.006	0.000	0.000	0.027	0.000

B	Additive matrices						
	Individual			Combined			Global
P	dir	ult	add	dir + ult	dir + add	add + ult	dir+ult+add
10%	0.668	0.290	0.730	0.857	0.785	0.871	0.891
20%	0.370	0.118	0.470	0.492	0.652	0.627	0.789
30%	0.182	0.042	0.242	0.262	0.422	0.304	0.500
40%	0.074	0.018	0.128	0.050	0.236	0.127	0.143
50%	0.016	0.006	0.040	0.000	0.081	0.000	0.000
60%	0.000	0.000	0.004	0.000	0.000	0.000	0.000

superfamilies, and one outgroup species. The ultimate purpose of combining these data sets is double. First, adding data is likely to increase the accuracy and robustness of the phylogeny based on the combined evidence (Huelsenbeck *et al.* 1996). Second, increasing the number of species allows for a supertree construction including all species represented in the separate analyses. In this example, each tree was derived using different data, namely DNA-hybridization distances (Hutcheon *et al.* 1998) and molecular sequences (Teeling *et al.* 2000); the combination of the data sets thus proved impossible with classical methods (see Kluge 1989). Since the species sampling was not rigorously identical in the two studies, the phylogenies could not be combined either according to standard consensus techniques (Mickevich 1978; Sokal & Rohlf 1981). The direct and indirect approaches described in the present paper can be used, however, to solve this problem.

Because DNA-hybridization can produce asymmetrical distances, the original matrix published by Hutcheon *et al.* (1998) was first symmetrised with the Springer and Kirsch procedure (1989). On the other hand, the distances corresponding to the molecular sequences of Teeling *et al.* (2000) were obtained using the Jukes-Cantor correction (computations made in PAUP*, Swofford 1999). To impart an equivalent weight to the two data sets in their combination, all distances were standardized by dividing each entry by the distance between the outgroup (*Cynocephalus*) and a taxon common to both data sets (*Myotis*). To eliminate any bias related to the estimation of distances between terminal sister taxa (for more details see Landry & Lapointe 1997), species were removed from the analysis to prevent the effect of such circumstances. Namely, the choice to remove a sister taxa from a family (or superfamily) was made in order to keep at least one species of this family (or superfamily) in common to both data sets. Trees corresponding to each data set were derived from the restricted distance matrices using a least-squares algorithm (Felsenstein 1993; Makarenkov & Leclerc 1999) and the corresponding path-length matrices were extracted from those trees (Figure 5.1). The average of the two distance matrices, for which the taxa were not identical but largely overlapping, resulted in an incomplete distance matrix among 12 bat species and one outgroup, with 14% of missing cells. To conduct the analysis of these data, we adopted the procedure that we used for the simulations. The incomplete matrix was first analyzed directly with a WLS algorithm. In

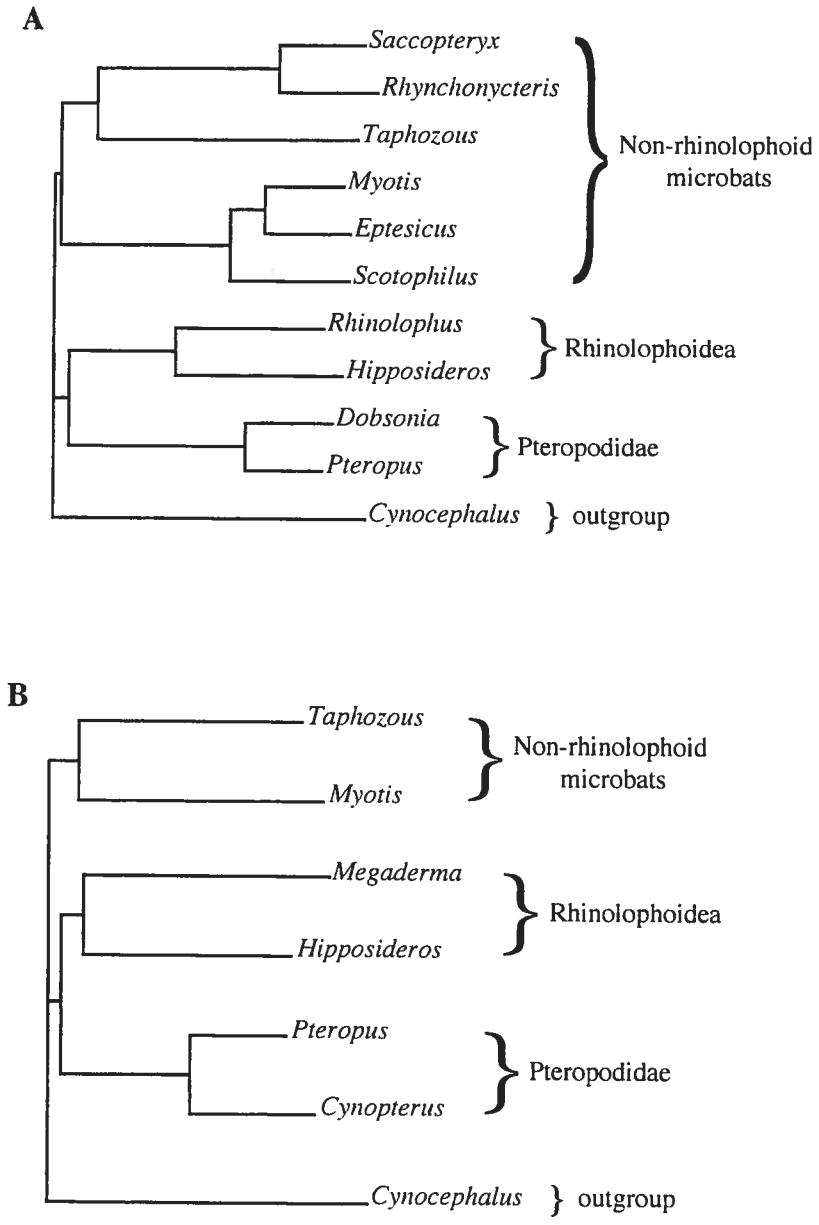

Figure 5.1 Estimates of relationships among bats from two studies : (A) DNA-hybridization data (Hutcheon *et al.*, 1998) and (B) molecular sequences (Teeling *et al.*, 2000).

parallel, the missing cells were estimated using either the ultrametric property or the four-point condition, and the same algorithm was then applied to these matrices to construct phylogenies. The topologies of the three different trees were finally compared to assess whether they recovered similar relationships

Figure 5.2 presents the results of the phylogenetic analyses. The trees obtained with ultrametric or additive estimations of the missing cells were almost identical (Figures 5.2A and 5.2B). Both of these indirect methods correctly recovered the taxonomy by assigning the different species to their corresponding families and superfamilies. However, the relationships among the families differed in those trees, especially with respect to the monophyly of non-rhinolophoid microbats. The tree obtained with the additive method corroborates the phylogenies of Hutcheon *et al.* (1998) and Teeling *et al.* (2000). Furthermore, both trees show that microbats are probably paraphyletic. On the other hand, the phylogeny produced by the direct approach recovered a different branching pattern among the 12 bat species (Figure 5.2C). In that tree, neither the Rhinolophoidea nor the Pteropodidae are monophyletic. It thus appears that the direct approach was not able to estimate correctly the relationships among these bats, in the presence of missing cells.

Discussion

The main objective of this paper was to determine whether one should estimate missing cells in incomplete distance matrices prior to phylogenetic reconstruction. The answer to this question is clearly yes. Our results show that it is indeed preferable to estimate missing cells, for distance as well as for topological recovery. The indirect additive procedure provided better recovery values than the direct WLS approach in the case of additive matrices, whereas the indirect ultrametric procedure performed better in the case of ultrametric matrices. Therefore, it is safe to say that *at least one* indirect method always outperformed the direct approach, regardless of the type of matrices analyzed. The superiority of the indirect approach was even more compelling for higher percentages of missing cells, when the amount of information available in the matrix is too scarce for the WLS algorithm to perform well. Interestingly, the application to bat phylogeny provided a case for which the direct approach was not able to recover the relationships among taxa within the same family, even for a small number of missing cells (*i.e.*, 14%). In this

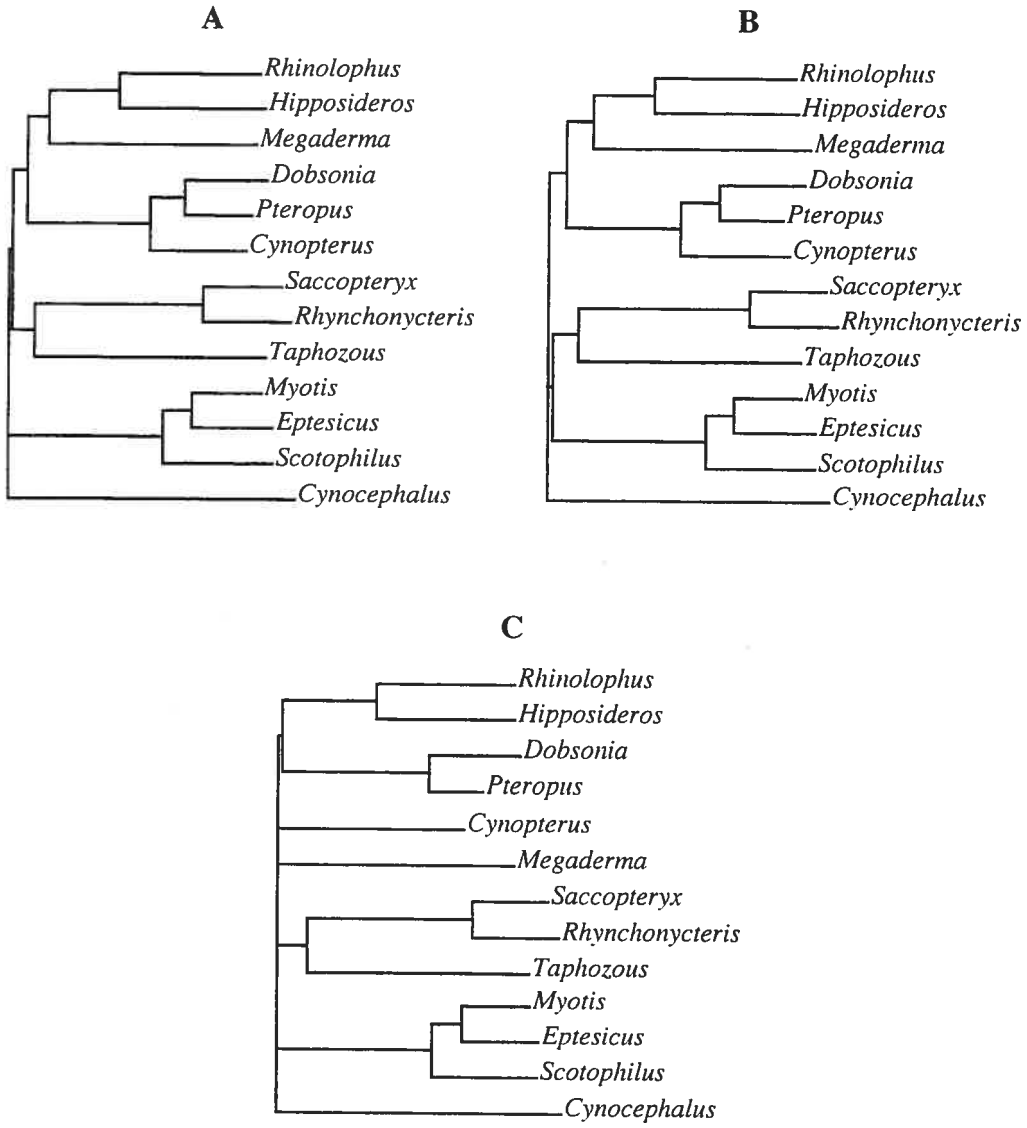

Figure 5.2 Trees resulting from the combination of the path-length distance matrices associated with the phylogenies in Figure 5.1, using (A) the indirect method using the ultrametric estimation, (B) the indirect method using the additive estimation or (C) the direct approach.

particular situation, the trees obtained with at least one of the indirect methods recovered the expected phylogeny. This tree corroborated those published by Hutcheon *et al.* (1998) and Teeling *et al.* (2000).

In spite of our results based on simulations and real data, one question remains, however. Which of the different estimation methods should one use? Previous simulations (Landry *et al.* 1996) have shown that the ultrametric procedure does comparatively better as the number of missing cells increases, and the present study supports these findings. This is due to the fact that not enough values are known to estimate missing cells with the four-point condition when *P* gets larger. In such cases, the ultrametric property must be used, so long as the number of missing values prevents the computation of additive estimates. In general terms, the previous studies have collectively shown that the ultrametric procedure does better for ultrametric matrices, whereas the additive procedure does better for non-ultrametric additive matrices (see Landry & Lapointe 1997). Because it is impossible to predict *a priori* whether a given distance matrix satisfies the ultrametric or the four-point condition, the combination of different approaches is more likely to provide more accurate phylogenies. We have tested this hypothesis in the present paper. Our results indeed showed that the relative frequency of getting an accurate topology is increased by using different methods jointly, especially for low values of *P*, as in the real-case application (P = 14%). When ultrametric and additive procedures produce the same tree (see Figure 5.2), combined rates of topological recovery varied from 74.6% (ultrametric matrices) to 87.1% (additive matrices) with 10% of missing data; the corresponding individual rates were lower. In light of these results, and because one tree derived using the indirect approach was similar to those obtained in other studies (Hutcheon *et al.* 1998; Teeling *et al.* 2000), chances are that this phylogeny is more accurate than the one obtained by the direct approach with a weighted least-squares algorithm.

As a final word of caution, it is worth repeating some problems associated with the analysis of incomplete distance matrices (for details, see Landry & Lapointe 1997). For example, (1) missing distances between terminal sister taxa will never be properly recovered by any method. (2) Trees with very short internodes are also problematic when some distances are missing. (3) Negative distances can sometimes be returned by the four-point condition. Finally, (4) the order in which the estimations are performed can affect the

results. In spite of these limitations, estimating missing cells remains a valid procedure. One should however bear in mind that phylogenies are nothing more than estimates of relationships among species and those estimates are more likely to be inaccurate when missing distances have also been estimated. The recommendations provided in this paper are thus meant to help systematists dealing with incomplete distance matrices. We propose not to rely on a single technique to reconstruct a phylogeny from incomplete matrices. The use of a weighted least-squares algorithm is one among alternative direct methods (see also Guénoche & Grandcolas 1999; Hein 1989). It was shown to provide better results than competing direct algorithms in previous simulations involving matrices with missing distances (Levasseur *et al.* 2000). For the time being, it remains that indirect approaches based on tree metric properties perform better than any direct method currently available.

ACKNOWLEDGMENTS

This work was supported by a NSERC research grant OGP 0155251 to FJL, a FBSB and FES scholarships to CL and a NSERC scholarship to PAL. The authors are indebted to Mark Springer for providing the molecular sequences used in the application. The data matrices are available upon request from FJL. The approaches discussed in this paper have been implemented in the T-rex program available from: www.fas.umontreal.ca/BIOL/legendre/index.html.

CHAPITRE 6

A short note on supertrees

Les résultats de simulations présentés dans cette section sont publiés sous la référence :

Lapointe, F.-J.. & Levasseur, C. 2004 Everything you always wanted to know about the average consensus, and more. In *Phylogenetic supertrees : Combining Information to Reveal the Tree of Life* (ed., O.R.P. Bininda-Emonds), pp. 87-105, Series in Computational Biology, Dordrecht: Kluwer's Academic Publisher.

Les pages qui suivent ne forment qu'une partie de cet article.

INTRODUCTION

The number of methods available to infer supertrees clearly illustrates the great need for such tools in phylogenetics (Baum 1992; Chen *et al.* 2003; Constantinescu & Sankoff 1995; Goloboff & Pol 2002; Gordon 1986; Lanyon 1993; Page 2002; Ragan 1992; Sanderson *et al.* 1998; Semple & Steel 2000; Slowinski & Page 1999; Steel 1992; Steel *et al.* 2000). Indeed, the growing interest in the search of the Tree of Life and the exponential accumulation of data bring out new challenges. The size of data matrices becomes more important, both in terms of the number of taxa and the number of characters, such that larger phylogenies are now published. Moreover, the desire to combine data sets bearing partially overlapping sets of taxa render the more traditional phylogenetic methods unsuitable for this kind of analysis. This new reality comes with its own set of problems like a disproportioned representation of some data because of data duplication or the inability to control the quality of the trees combined (see Sanderson *et al.* 1998; Bininda-Emonds *et al.* 2002). Up to now, the methods available are highly criticized. We barely understand the properties of the different approaches (but see Wilkinson *et al.* 2004); simulation studies have been conducted to assess their reliability but only under very specific situations (Bininda-Emonds & Sanderson 2001; Chen *et al.* 2003; Lapointe & Levasseur 2004; Piaggio-Talice *et al.* 2004).

In the consensus setting, the use of the average consensus (Lapointe & Cucumel 1997) in a global congruence approach (*sensu* Lapointe *et al.* 1999) increases chances of recovering a known model tree compared to the use of a single method. Because this consensus method can also be used for supertree construction, it is of interest to investigate its applicability in a more general supertree setting to determine whether it can provide accurate estimates of phylogenetic relationships. Since some average supertrees have already been published (Barker 2002; Kirsch *et al.* 1997; Lapointe & Kirsch 2001), there is an urgent need to assess the accuracy of the average consensus trees in this context. This first step is crucial before any further investigation of the global congruence approach in the supertree setting.

SIMULATION STUDY

A simulation study was undertaken to assess the ability of the average consensus to recover the model tree in a supertree setting, using a model tree with 20 taxa (Figure 6.1A) inspired by Kumar (1996). To make things simple, we restricted ourselves to cases involving the combination of only two weighted trees, representing subtrees of the model supertree. Different parameters were investigated in the simulations: the relative size of the subtrees (identical or different), the size of the overlap between subtrees (small or large) and the degree of heterogeneity of the data evolved on the model tree (homogeneous or heterogeneous). To simulate heterogeneous data sets, DNA sequences of 2500 bps were independently evolved on two model trees with the same topology, but with different branch lengths and using different rates of evolution (Figures 6.1B and 6.1C). Homogeneous data sets were simulated by evolving DNA sequences of 2500 bps on the same tree (Figure 6.1C). The heterogeneity of the data sets was assessed with the incongruence length difference test (*ILD*: Farris *et al.* 1995b). Four situations were simulated for heterogeneous and homogeneous data sets: (a) subtrees of the same size (13 taxa) with a small overlap (6 taxa), (b) subtrees of the same size (15 taxa) with a large overlap (10 taxa), (c) subtrees of different sizes (10 and 16 taxa) with a small overlap (6 taxa), and (d) subtrees of different sizes (13 and 17 taxa) with a large overlap (10 taxa). For each of these cases, 1000 replicates were simulated. All sequences were generated with the Seq-Gen program (Rambaut & Grassly 1997) using a Jukes-Cantor model of evolution (Jukes & Cantor 1969).

Distance matrices were computed from the DNA sequences using a Jukes-Cantor correction, and trees were estimated from these distances using an unweighted least squares method (Cavalli-Sforza & Edwards 1967) in PAUP* (Swofford 1999). Three different standardization techniques were employed to correct for differences in branch lengths caused by the relative sizes of the subtrees and the heterogeneous rates of evolution. Namely, the distance values in each matrix were scaled, either by (1) dividing all distances by the maximum distance in the entire matrix, (2) dividing all distances by the maximum distance in the common part of the matrix representing the overlap of the two subtrees, or (3) by multiplying the distances in the first matrix, such as to maximize the fit to the second matrix. These corrected path length distance matrices associated with the corresponding

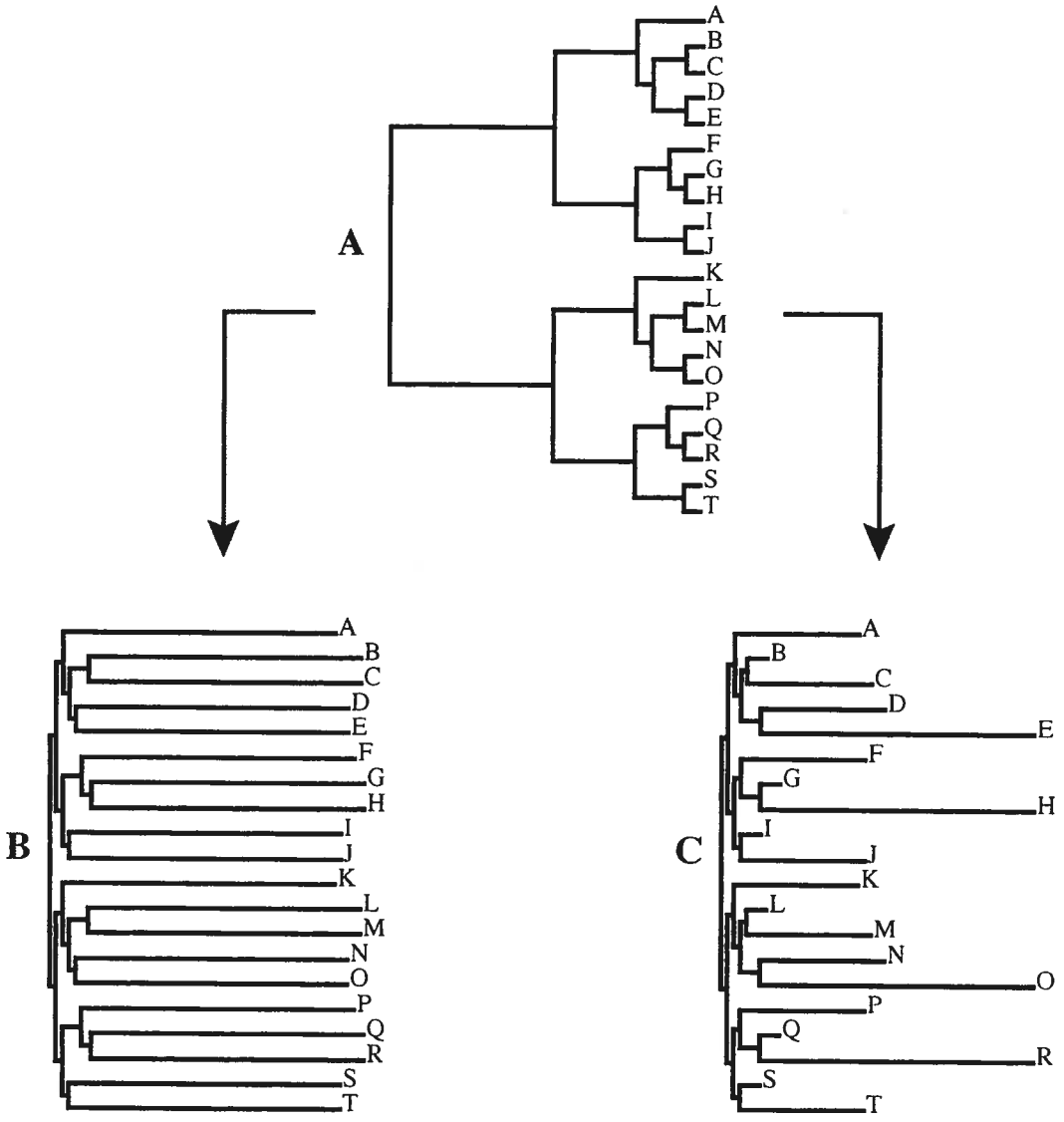

Figure 6. 1 Model topology (A) with different branch lengths and with (B) slow [0.25] and (C) rapid [1.75] evolutionary rates of change along the branches.

subtrees were combined to compute an average matrix defined on the whole set of taxa. The missing distances in the non-overlapping part of the matrix were estimated with the additive procedure (see Eq. 5.2 in Chapter 5). The average supertree was then obtained by applying a least squares algorithm to that matrix.

RECOVERING THE MODEL TREE

Recovery was measured by comparing each average supertree to the model tree (Fig. 6.1A). To do so, a strict consensus was computed and the consensus fork index (*CFI*: Colless 1980) was used to quantify topological agreement. The *CFI* measures the proportion of resolved clades in the strict consensus tree of the trees compared. Its maximum value of one indicates total congruence (*i.e.*, the average supertree and the model tree are topologically identical and their strict consensus is fully resolved), whereas a value of zero is indicative of total incongruence (*i.e.*, the average supertree and the model tree are topologically incompatible and their strict consensus is a bush). The mean *CFI* values of the 1000 replicates and their standard deviations are reported in Table 6.I for the four situations considered in the simulations. Because the results obtained with the three different standardization techniques were very similar, only those corresponding to the third method are presented. In the case of homogeneous data sets, the mean *CFI* values range from 0.720 to 0.879 and indicate that recovery is improved when the overlap is large and the number of missing cells is small. On the other hand, the values obtained for heterogeneous data sets, ranging from 0.071 to 0.200, are much worse and do not follow the same trend as homogeneous data sets. The differences in branch lengths and rates of evolution clearly affect the recovery of average supertrees in such situations, regardless of the scaling method selected. To avoid this problem, the same simulations were repeated by setting all branch lengths to one prior to combining the subtrees. The corresponding branch distance matrices were thus used to compute an average supertree defined on topological relationships alone, almost like MRP, but using a different matrix representation. The results of these simulations are presented in Table 6.II. Following this modification, the mean *CFI* values for homogeneous data sets slightly decreased but the corresponding values for heterogeneous data sets increased dramatically compared with the results of the first series of simulations (see Table 6.I).

Table 6.I Mean recovery values obtained in the four situations considered in the simulations, for homogeneous and heterogeneous data sets. Actual branch lengths were used to compute average supertrees. The standard deviations are given in parentheses. All simulations were based on 1000 replicates.

Size of the subtrees (overlap)	13(6)13	10(6)16	15(10)15	13(10)17
Number of missing distances	49	40	25	21
Homogeneous data sets	0.720	0.771	0.868	0.879
	(0.116)	(0.109)	(0.089)	(0.087)
Heterogeneous data sets	0.147	0.200	0.081	0.071
	(0.096)	(0.099)	(0.073)	(0.064)

Table 6.II Mean recovery values obtained in the four situations considered in the simulations, for homogeneous and heterogeneous data sets. All branch lengths were set to one to compute average supertrees. The standard deviations are given in parentheses. All simulations were based on 1000 replicates.

Size of the subtrees (overlap)	13(6)13	10(6)16	15(10)15	13(10)17
Number of missing distances	49	40	25	21
Homogeneous data sets	0.703	0.752	0.863	0.878
	(0.108)	(0.099)	(0.084)	(0.082)
Heterogeneous data sets	0.615	0.691	0.777	0.763
	(0.126)	(0.117)	(0.109)	(0.111)

CONCLUSIONS AND CAVEATS

The combination of branch distance matrices represents a promising extension of the average procedure that deserves further exploration. Given our findings, it would seem sensible to ignore branch length information in building the average supertree as the penalty in doing so when the data sets are homogeneous is slight, but the benefit in doing so when they are heterogeneous is great. That is to say that the major strength of the average procedure can become a weakness when source trees with heterogeneous branch lengths are combined. Similarly, combining trees with branch lengths with others that do not have branch lengths will clearly affect the resulting supertree. In such cases, it is always preferable to ignore branch lengths altogether when building average supertrees. For the same reason, we do not recommend to combine branch distance matrices with path length distance matrices. When branch lengths are available, however, standardization methods must be used to scale path length distances in such a way that they become comparable. We have addressed this problem in our simulations, but further studies would be required to fully examine the recovery of average supertrees, with or without branch lengths (see also Bininda-Emonds & Sanderson, 2001). Other ways of scaling path length distance matrices also need to be investigated, when combining, more than two trees of varying sizes. Finally, the relative performance of average supertrees with respect to other supertree methods (Baum 1992; Goloboff & Pol 2002; Gordon 1986; Lanyon 1993; Semple & Steel 2000) must be addressed, as well as the relationship between average supertrees and total evidence trees derived from incomplete data sets.

CONCLUSION

De plus en plus, l'analyse phylogénétique est considérée comme une boîte noire qui permet de simplement transformer des données en phylogénies. Il existe plusieurs volumes qui expliquent les nombreuses façons de faire (voir, par exemple, Hall 2001), comme si l'inférence phylogénétique n'était qu'un protocole à suivre, une simple formalité technique. Il est évident, et même souhaitable, que des chercheurs de plusieurs champs d'activités (biologie, biochimie, biologie moléculaire, informatique, mathématique, etc.) s'intéressent à certains aspects de l'évolution, dont l'analyse phylogénétique. Il est également compréhensible que tous ces scientifiques qui ne sont pas spécialistes dans ce domaine désirent des outils simples pour effectuer les analyses dont ils ont besoin. Par contre, il faut être vraiment prudent dans la manière dont ces dernières sont effectuées. La grande accessibilité de certains programmes d'analyse phylogénétique (PAUP*, Swofford 1999; plusieurs sont gratuitement distribués sur le web : PHYLIP, MEGA, PAL, etc.), leur convivialité, la performance accrue des ordinateurs actuels et la vaste publication de phylogénies donnent l'impression que la reconstruction phylogénétique est un jeu d'enfant. Pourtant, de nombreux chercheurs s'affairent depuis bon nombre d'années à tenter de comprendre les grands principes de l'évolution pour, dans la mesure du possible, adapter les méthodes à cette réalité. Le domaine de l'analyse phylogénétique est constamment en mouvement. De nouvelles méthodes sont régulièrement proposées et d'intenses débats suscitent discussions et réactions qui permettent de faire avancer les connaissances. Nul ne peut être à l'affût de tous ces changements qui rendent difficile le jugement critique des utilisateurs. Devant le vaste choix des méthodes d'analyse phylogénétique, des modèles d'évolution, des techniques de consensus, des approches de combinaison des données, des indices de comparaison d'arbres, des méthodes de validation, et j'en passe, bien peu de chercheurs sont en mesure de prendre des décisions éclairées. Face à cet état de fait, il est primordial d'effectuer des études comme celle présentée dans cette thèse. Pour l'utilisateur naïf, il est préférable de travailler à l'intérieur de certaines limites qui ont déjà été testées. Je crois qu'en étudiant le comportement de certaines approches dans des contextes précis,

nous serons en mesure de comprendre le fonctionnement de ces approches et nous pourrons ainsi obtenir des phylogénies plus justes.

Depuis bientôt deux décennies, on oppose les approches de congruence des caractères et de congruence taxonomique comme façon de combiner les données de sources différentes. Alors qu'il a été proposé à maintes reprises et dans plusieurs contextes que l'utilisation de plusieurs méthodes permet d'obtenir des phylogénies plus justes (Hillis 1987; Johnson & Soltis 1998; Kim 1993; Larson 1994), le débat persiste encore. Il me paraît d'ailleurs déraisonnable de devoir choisir entre l'une ou l'autre de ces approches. Les arguments sont le plus souvent philosophiques ou méthodologiques, mais également temporels. Alors que plusieurs biologistes accordent plusieurs années à la récolte de leurs données, ces mêmes personnes sont souvent très impatientes lors de l'analyse. De plus en plus, les chercheurs sont conscients de l'importance d'utiliser diverses méthodes, mais encore trop souvent, les analyses ne sont pas approfondies. On assume que les méthodes sont sans failles, alors qu'il est connu que certaines d'entre elles sont meilleures ou moins bonnes dans certains contextes. Le présent projet s'inscrit précisément dans cette visée. En effet, l'approche de congruence globale est une option de choix à la combinaison des données, puisqu'elle propose une utilisation conjointe de deux approches.

L'objectif principal de cette recherche était précisément de vérifier la justesse de l'approche de congruence globale par rapport à la congruence des caractères et à la congruence taxonomique. Dans un premier temps, j'ai pu vérifier que le consensus moyen, qui tient compte des longueurs de branches, permet d'obtenir des arbres plus résolus que les méthodes de consensus topologique. De plus, les arbres issus de ce type de consensus sont plus souvent similaires à la phylogénie de l'analyse combinée dans un contexte de vraies données. Ceci laisse supposer deux choses. Tout d'abord, la querelle entre les partisans de la congruence des caractères et de la congruence taxonomique est biaisée quant à la méthode de consensus utilisée. En effet, puisque le consensus moyen permet d'obtenir des arbres identiques à l'aide des deux approches, le débat sur la supériorité d'une méthode par rapport à l'autre est inapproprié. Ensuite, lorsque les arbres issus des deux approches sont identiques, il est logique d'assumer que la solution reflète bien les relations phylogénétiques entre les organismes.

J'ai ensuite testé la justesse des différentes approches : analyse combinée, consensus topologiques (strict, semi-strict, majoritaire, majoritaire résolu), consensus moyen et congruence globale. Parce qu'il n'est pas possible avec de vraies données de comparer les résultats obtenus à la phylogénie attendue, j'ai effectué des études de simulations qui m'ont permis de montrer que l'approche de congruence globale surpasse toutes les autres. En effet, lorsque les arbres de congruence des caractères et de congruence taxonomique sont identiques, ils sont plus souvent pareils à l'arbre modèle que lorsqu'une seule approche est utilisée. De plus, l'utilisation conjointe du consensus moyen avec la congruence des caractères permet un meilleur taux de succès que lorsque cette dernière est jumelée à une méthode de consensus topologique. La différence de performance est due à une plus faible résolution de ces derniers par rapport au consensus moyen. Enfin, même lorsque les arbres des analyses combinées et séparées sont différents, les zones d'accord sont le plus souvent identiques à l'arbre modèle. La congruence globale est donc une approche juste qui permet d'améliorer la qualité des estimations phylogénétiques indépendamment du fait que les arbres résultant de la congruence des caractères et de la congruence taxonomique soient identiques ou non.

Les facteurs qui ont été testés ont permis de montrer que la taille des arbres et le nombre de jeux de données influencent beaucoup la justesse des estimations. Il est intéressant de noter que l'hétérogénéité des données a un effet moins marqué que les autres facteurs. La première série de simulations (Chapitre 2) suggérait qu'un degré d'hétérogénéité élevé affecte beaucoup la qualité des estimations phylogénétiques, en particulier lorsque les données sont combinées. Par contre, les résultats du Chapitre 3 montrent que le degré d'hétérogénéité n'est pas un élément qui diminue de manière radicale la justesse des arbres. D'une part, il faut noter que le degré d'hétérogénéité était moins grand dans les simulations du Chapitre 3. D'autre part, ceci peut également illustrer l'importance que peut avoir la sélection d'une seule topologie pour effectuer une étude de simulations. Dans le cas présent, cette première recherche constituait un travail préliminaire servant à évaluer les paramètres qui devaient être testés à plus grande échelle. La décision de ne prendre qu'un seul arbre modèle s'inspirait d'études de simulations déjà publiées (Kumar 1996). La topologie choisie représentait une situation particulière. Dans le but de rendre les résultats plus généralisables, la deuxième série de simulations a été effectuée sur

plusieurs réplicats. Je considère que les conclusions tirées de ces dernières sur l'effet de l'hétérogénéité des données sont probablement plus près de la réalité puisqu'elles se basent sur un plus grand nombre de cas.

J'ai testé un cas particulier où tous les jeux de données simulés étaient de taille égale. Cette situation spécifique se compare aisément à l'analyse de deux ou plusieurs gènes qui comportent un nombre de caractères comparable. Dans ces circonstances, nos résultats ont montré que l'utilisation conjointe des analyses séparées et combinée augmente les chances de retrouver l'arbre modèle. Pour cette raison, l'utilisation des tests d'hétérogénéité *a priori* dans ce contexte ne semble pas nécessaire puisqu'il semble plus profitable de toujours effectuer les deux analyses. Par contre, dans un cas où des jeux de données de différents types seraient combinés (par exemple des données morphologiques et moléculaires), nos résultats auraient pu être tout autres. D'abord, les arbres de congruence des caractères et de consensus moyen seraient peut-être moins souvent semblables à cause de la pondération différente des caractères. La première approche donne un poids égal à tous les caractères alors que la deuxième donne le même poids à chacun des jeux de données. De plus, les tests d'hétérogénéité sont les plus utiles lorsque les partitions combinées sont de taille différente. Pour des données homogènes, le fait de combiner un petit jeu de données morphologiques avec une longue séquence moléculaire n'affecterait pas le signal phylogénétique. Par contre, dans le cas de données hétérogènes, le signal du plus petit jeu de données risque d'être perdu au profit de l'information des données moléculaires. Pour cette raison, nos résultats ne nous permettent pas de déconseiller l'usage de ces tests.

Il est évident que plusieurs autres facteurs auraient pu être impliqués dans ces analyses. Les possibilités et les combinaisons sont infinies. D'abord, pour bien imiter les différentes sources d'incongruence, il serait intéressant de procéder à l'ajout de bruit dans les données ou encore de tenter de combiner plusieurs phylogénies modèles topologiquement différentes, ce qui pourrait augmenter le réalisme des simulations. Enfin, il semble que lorsque les phylogénies présentent de courtes branches internes, les résultats sont moins justes. En effet, il y a peu de changements évolutifs le long de ces branches et les méthodes d'analyse phylogénétique ne permettent pas toujours une estimation correcte de ces relations. Cette hypothèse que la longueur des branches internes pourrait être

corrélée à la congruence est basée sur quelques observations et mériterait une investigation plus approfondie.

La méthode MRP (*matrix representation with parsimony*) a été développée pour combiner les arbres dans un cadre de super-arbres, mais elle est aussi utilisée dans un cadre de consensus. Cette technique a été comparée tour à tour aux approches de congruence des caractères et de congruence taxonomique et a aussi été liée au consensus moyen dans des situations particulières. Étant donné que le consensus moyen et l'analyse combinée produisent des résultats similaires, j'ai également vérifié si la méthode MRP pouvait être reliée aux deux précédentes dans un contexte de consensus. Encore une fois, les résultats de mes simulations ont montré que le consensus moyen et l'approche de congruence des caractères étaient plus près l'une de l'autre, et aussi plus justes que l'approche MRP. Un des facteurs pouvant expliquer ces résultats est l'utilisation de la distance entre les taxons que fait le consensus moyen et l'analyse combinée, contrairement à la méthode MRP. En effet, cette dernière semble plus proche de la version topologique du consensus moyen, et pourrait être plus similaire aux méthodes de consensus qui ne tiennent pas compte des longueurs de branches.

Alors que le consensus moyen et l'approche de congruence des caractères semblent très similaires, il est intéressant d'évaluer si les deux pourraient être utilisées dans un cadre de congruence globale dans un contexte de super-arbres. Pour ce faire, il fallait d'abord étudier la performance du consensus moyen avec des simulations. J'ai d'abord montré que les méthodes qui estiment les distances manquantes lors du calcul du consensus moyen permettent d'obtenir des arbres plus justes que si ces données restent inconnues. Évidemment, la justesse diminue avec l'augmentation du nombre de distances manquantes. Par contre, les résultats des simulations montrent que dans le cas des super-arbres, d'autres facteurs rendent difficile la reconstruction de l'arbre modèle. Citons, par exemple, la standardisation des distances. Comme chaque arbre dérivé d'analyses séparées comporte un sous-échantillon de l'ensemble total des taxons, ils n'en comportent pas tous le même nombre. Le problème est particulièrement important lorsque des données hétérogènes sont combinées. Dans ces cas, l'utilisation d'une version topologique du consensus moyen permet d'améliorer radicalement les résultats. Alors que pour des données homogènes cette

méthode diminue quelque peu la justesse, l'augmentation importante dans les autres situations justifie grandement cette solution.

Il est évident que de nouvelles simulations sont nécessaires pour bien comprendre toutes les implications relatives à l'utilisation du consensus moyen dans un cadre de super-arbres. Néanmoins, il serait intéressant de vérifier l'applicabilité de la congruence globale dans ce contexte. La version topologique de ce consensus constitue une option alléchante qui pourrait améliorer la justesse des super-arbres. De plus, puisque la méthode MRP semblait plus près de cette méthode, les résultats de ces différentes approches pourraient être comparés et utilisés conjointement dans un cadre de congruence globale.

Tout au long de cette thèse, il a été question de justesse des arbres, en particulier lorsque les résulats des différentes approches étaient comparés à une topologie modèle, de laquelle étaient dérivées les données utilisées dans les simulations. Dans le cas présent, pour qu'un arbre soit juste, il doit être identique à l'arbre modèle. Mais il existe plusieurs autres façons de définir et de mesurer la justesse qui pourraient nuancer les conclusions de cette recherche. Par exemple, plutôt que de la mesurer de manière binaire (un arbre est juste s'il est identique à l'arbre modèle ou il ne l'est pas s'il comporte une ou plusieurs différences), elle aurait pu être quantitative. En effet, comme il a été fait lors des comparaisons entre les arbres de congruence des caractères et de congruence taxonomique, des moyennes auraient pu être calculées. Ceci aurait probablement eu pour effet d'augmenter la justesse dans le cas des méthodes de consensus topologiques. En effet, pour certaines séries de simulations, plusieurs arbres ne comportant que quelques différences (une ou deux) avec l'arbre modèle n'étaient pas considérés dans les résultats. Le choix de définir la justesse de cette manière est très personnel et discutable. Il est vrai qu'un arbre qui n'est pas parfaitement identique à l'arbre modèle n'est pas nécessairement mauvais. Mais comme les simulations représentent une situation idéale et très simplifiée comparée à l'analyse de vraies données, il me semblait défendable de chercher à connaitre les méthodes qui pouvaient obtenir un score parfait dans ces circonstances. Cette définition très pointue de la justesse m'as permis de discriminer les résultats des différentes approches.

Malgré l'ampleur de cette recherche, plusieurs questions méritent encore d'être étudiées. Tout d'abord, j'ai évalué la justesse des différentes approches dans des situations

où tous les jeux de données comportent le même nombre de caractères. Dans les critiques formulées à l'égard de la congruence des caractères, la plus importante porte sur le fait que le signal phylogénétique des grandes partitions peut masquer le signal des plus petites. Comme je l'ai mentionné, il sera nécessaire de vérifier dans une prochaine étude si les résultats du consensus moyen et de l'analyse combinée sont similaires dans de telles situations et de tester l'effet de la pondération du consensus. De plus, la validation des résultats est une préoccupation dominante en analyse phylogénétique. Il est important de souligner qu'au-delà du fait qu'une méthode puisse produire des résultats justes, il arrive souvent que les données ne soient pas assez informatives. La méthode peut à ce moment résoudre de manière arbitraire un groupement particulier. Il est donc impératif de vérifier que la phylogénie obtenue reflète bien le signal phylogénétique des données. L'arbre de congruence des caractères est souvent validé à l'aide de méthodes de ré-échantillonnage statistique. Par contre, les arbres de consensus ne sont que très rarement validés. De plus, la validité et la robustesse des arbres qui résultent d'analyses séparées et qui servent au calcul des consensus sont rarement considérées (voir Larson 1994). C'est un point qu'il faudra absolument aborder dans les recherches futures.

BIBLIOGRAPHIE

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

Adanson, M. 1763 *Familles des Plantes*. Vincent, Paris.

Archie, J. W. 1989 A randomization test for phylogenetic information in systematic data. *Systematic Zoology* **38**, 239-252.

Baker, R. H. & DeSalle, R. 1997 Multiple sources of character information and the phylogeny of Hawaiian Drosophilids. *Systematic Biology* **46**, 654-673.

Barker, G. M. 2002 Phylogenetic diversity: A quantitative framework for measurement of priority and achievement in biodiversity conservation. *Biological Journal of the Linnean Society* **76**, 165-194.

Barrett, M., Donoghue, M. J. & Sober, E. 1991 Against consensus. *Systematic Biology* **40**, 486-493.

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

Baum, D. A., Small, R. L. & Wendel, J. F. 1998 Biogeography and floral evolution of baobabs (*Adansonia*, Bombacaceae) as inferred from multiple data sets. *Systematic Biology* **47**, 181-207.

Bininda-Emonds, O. R. P. 2003 MRP supertree construction in the consensus setting. In *Bioconsensus*, vol. 61 (ed. M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin & F. S. Roberts), pp. 231-242. Providence: American Mathematical Society.

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

Bininda-Emonds, O. R. P. & Bryant, H. N. 1998 Properties of matrix representation with parsimony analyses. *Systematic Biology* **47**, 497-508.

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 of the Cambridge Philosophical Society* **74**, 143-175.

Bininda-Emonds, O. R. P., Jones, K. E., Price, S. A., Grenyer, R., Cardillo, M., Habib, M., Purvis, A. & Gittleman, J. L. 2003 Supertrees are a necessary not-so-evil: a comment on Gatesy *et al. Systematic Biology* **52**, 724-729.

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.

Bremer, K. 1990 Combinable component consensus. *Cladistics* **6**, 369-372.

Bryant, D. 2003 A classification of consensus methods for phylogenetics. In *Bioconsensus* (ed. M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin & F. S. Roberts), pp. 163-184. Providence: American Mathematical Society.

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.

Buneman, P. 1971 The recovery of trees from measures of dissimilarity. In *Mathematics in archeological and historical sciences* (ed. F. R. Hudson, D. G. Kendall & P. Tautu), pp. 387-395. Edinburgh: edinburgh University Press.

Cannatella, D. C., Hillis, D. M., Chippindale, P. T., Weigt, L., Rand, A. S. & Ryan, M. J. 1998 Phylogeny of frogs of the *Physalaemus pustulosus* species group, with an examination of data incongruence. *Systematic Biology* **47**, 311-335.

Carnap, R. 1950 *Logical foundations of probability*. Chicago: University of Chicago Press.

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

Chen, D., Diao, L., Eulenstein, O., Fenandez-Baca, D. & Sanderson, M. J. 2003 Flipping: A supertree construction method. In *Bioconsensus* (ed. M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin & F. S. Roberts), pp. 135-160. Providence: American Mathematical Society.

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

Colless, D. H. 1980 Congruence between morphometric and allozyme data for *Menidia* species: a reappraisal. *Systematic Zoology* **29**, 288-299.

Constantinescu, M. & Sankoff, D. 1995 An efficient algorithm for supertrees. *Journal of Classification* **12**, 101-112.

Cracraft, J. & Mindell, D. P. 1989 The early history of modern birds: a comparison of molecular and morphological evidence. In *The hierarchy of life: molecules and morphology in phylogenetic analysis* (ed. B. Fernholm, K. Bremer & H. Jornvall), pp. 389-403. Amsterdam: Elsevier.

Cucumel, G. & Lapointe, F.-J. 2000 A general approach to test the pertinence of a consensus classification. In *Data analysis, classification, and related methods* (ed. H. A. L. Kiers, Rasson, J.-P., Groenen, P.J.F., Schader, M.), pp. 125-130. Berlin: Springer-Verlag.

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 Soete, G. 1983 A least squares algorithm for fitting additive trees to proximity data. *Psychometrika* **48**, 621-626.

De Soete, G. 1984a Additive-tree representations of incomplete dissimilarity data. *Quality and quantity* **18**, 387-393.

De Soete, G. 1984b Ultrametric tree representations of incomplete dissimilarity data. *Journal of Classification* **1**, 235-242.

Donoghue, M. J. & Sanderson, M. J. 1992 The suitability of molecular and morphological evidence in reconstructing plant phylogeny. In *Molecular systematics in plants* (ed. P. S. Soltis, Soltis, D.E., Doyle, J.J.), pp. 340-368. New York: Chapman & Hall.

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

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

Faith, D. P. & Cranston, P. S. 1991 Could a cladogram this short have arisen by chance alone? On permutation tests for cladistic structure. *Cladistics* **7**, 1-28.

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

Farris, J. S., Källersjö, M., Kluge, A. G. & Bult, C. 1995b Testing significance of incongruence. *Cladistics* **10**, 315-319.

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

Felsenstein, J. 1985 Confidence limits on phylogenies: an approach using the bootstrap. *Evolution* **39**, 783-791.

Felsenstein, J. 1993 PHYLIP: Phylogeny inference package. Seattle: University of Washington.

Fitch, W. M. & Margoliash, E. 1967 Construction of phylogenetic trees. *Science* **155**, 279-284.

Flook, P. K., Klee, S. & Rowell, C. H. F. 1999 Combined molecular phylogenetic analysis of the Orthoptera (Arthropoda, Insecta) and implications for their higher systematics. *Systematic Biology* **48**, 233-253.

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

Gatesy, J., Milinkovitch, M., Waddell, V. & Stanhope, M. 1999a Stability of cladistic relationships between Cetacea and higher-level artiodactyl taxa. *Systematic Biology* **48**, 6-20.

Gatesy, J., O'Grady, P. & Baker, R. H. 1999b Corroboration among data sets in simultaneous analysis: Hidden support for phylogenetic relationships among higher level artiodactyl taxa. *Cladistics* **15**, 271-314.

Gibson, G. 2002 Microarrays in ecology and evolution: a preview. *Molecular Ecology* **11**, 17-24.

Goloboff, P. A. & Pol, D. 2002 Semi-strict supertrees. *Cladistics* **18**, 514-525.

Gordon, A. D. 1986 Consensus supertrees: the synthesis of rooted trees containing overlapping sets of labeled leaves. *Journal of Classification* **3**, 335-348.

Grenyer, R. & Purvis, A. 2003 A composite species-level phylogeny of the Insectivora (Mammalia: Order Lipotyphla Haeckel, 1866). *Journal of Zoology* **260**, 245-257.

Guénoche, A. & Grandcolas, S. 1999 Approximations par arbre d'une distance partielle. *Mathematique Informatique et Sciences Humaines* **146**, 51-64.

Hall, B.G. 2001 *Phylogenetic Trees Made Easy: A How-To Manual for Molecular Biologists*. Sunderland: Sinauer Associates.

Hartigan, J. A. 1967 Representation of similarity matrices by trees. *Journal of the American Statistical Association* **62**, 1140-1158.

Hasegawa, M., Adachi, J. & Milinkovitch, M. C. 1997 Novel phylogeny of whales by total molecular evidence. *Journal of Molecular Evolution* **44**, S117-S120.

Hein, J. 1989 A tree reconstruction method that is economical in the number of pairwise comparison used. *Molecular Biology and Evolution* **6**, 669-684.

Helm-Bychowski, K. & Cracraft, J. 1993 Recovering phylogenetic signal from DNA sequences: Relationships within the corvine assemblage (class Aves) as inferred from complete sequences of the mitochondrial DNA cytochrome-*b* gene. *Molecular Biology and Evolution* **10**, 1196-1214.

Hillis, D. M. 1987 Molecular versus morphological approaches to systematics. *Annual Review of Ecology and Systematics* **18**, 23-42.

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

Hillis, D. M., Huelsenbeck, J.P. & C.W. Cunningham. 1994 Application and accuracy of molecular phylogenies. *Science* **264**, 671-677.

Huelsenbeck, J. P. & Bull, J. J. 1996 A likelihood ratio test for detection of conflicting phylogenetic signal. *Systematic Biology* **45**, 92-98.

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

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

Huelsenbeck, J. P., Swofford, D. L., Cunningham, C. W., Bull, J. J. & Waddell, P. J. 1994 Is character weighting a panacea for the problem of data heterogeneity in phylogenetic analysis? *Systematic Biology* **43**, 288-291.

Hutcheon, J. M., Kirsch, J. A. W. & Pettigrew, J. D. 1998 Base-compositional biases and the bat problem. III. The question of microchiropteran monophyly. *Philosophical Transactions of the Royal Society of London Series B* **353**, 607-617.

Johnson, L. A. & Soltis, D. E. 1998 Assessing congruence: empirical examples from molecular data. In *Molecular Systematics of Plants II. DNA Sequencing* (ed. D. E. Soltis, P. S. Soltis & J. J. Doyle), pp. 297-348. Boston: Kluwer Academic Publisher.

Jones, K. E., Purvis, A., MacLarnon, A., Bininda-Emonds, O. R. P. & Simmons, N. B. 2002 A phylogenetic supertree of the bats (Mammalia: Chiroptera). *Biological Reviews* **77**, 223-259.

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.

Jukes, T. H. & Cantor, C. R. 1969 Evolution of protein molecules. In *Mammalian Protein Metabolism* (ed. H. N. Munro), pp. 21-132. New York: Academic Press.

Kennedy, M. & Page, R. D. M. 2002 Seabird supertrees: Combining partial estimates of procellariiform phylogeny. *Auk* **119**, 88-108.

Kim, J. 1993 Improving the accuracy of phylogenetic estimation by combining different methods. *Systematic Biology* **42**, 331-340.

Kirsch, J. A. W., Lapointe, F.-J., Springer, M.S. 1997 DNA-hybridisation studies of marsupials and their implications for metatherian classification. *Australian journal of zoology* **45**, 211-280.

Kluge, A. G. 1983 Cladistics and the classification of the great apes. In *New interpretations of ape and human ancestry* (ed. R. L. Ciochon & R. S. Corruccini), pp. 151-177. New York: Plenum.

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.

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

Landry, P.-A. & Lapointe, F.-J. 1997 Estimation of missing distances in path-length matrices: problems and solutions. In *Mathematical Hierarchies and Biology* (ed. B. Mirkin, F. R. McMorris, F. S. Roberts & A. Rzhetsky), pp. 209-224. Providence: American Mathematical Society.

Landry, P.-A., Lapointe, F.-J. & Kirsch, J. A. W. 1996 Estimating phylogenies from lacunose distance matrices: additive is superior to ultrametric estimation. *Molecular Biology and Evolution* **13**, 818-823.

Lanyon, S. M. 1993 Phylogenetic frameworks: towards a firmer foundation for the comparative approach. *Biological Journal of the Linnean Society* **49**, 45-61.

Lapointe, F.-J. 1998a For consensus (with branch lengths). In *Advances in Data Science and Classification* (ed. A. Rizzi, Vichi, M., Bock, H.-H.), pp. 73-80. Berlin: Springer-Verlag.

Lapointe, F.-J. 1998b How to validate phylogenetic trees? A stepwise procedure. In *Data Science, Classification, and Related methods: Studies in Classification, Data Analysis, and Knowledge Optimization*, vol. 71-88 (ed. C. Hayashi, Bock, H.-H., Yajima, K., Tanaka, Y., Baba, Y.). Tokyo: Springer-Verlag.

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. & Cucumel, G. 2003 How good can a consensus get? Assessing the reliability of consensus trees in phylogenetic studies. In *Bioconsensus* (ed. M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin & F. S. Roberts), pp. 205-220. Providence: American Mathematical Society.

Lapointe, F.-J. & Kirsch, J. A. W. 1995 Estimating phylogenies from lacunose distance matrices, with special reference to DNA hybridization data. *Molecular Biology and Evolution* **12**, 266-284.

Lapointe, F.-J. & Kirsch, J. A. W. 2001 Construction and verification of a large phylogeny of marsupials. *Australian Mammalogy* **23**, 9-22.

Lapointe, F.-J., Kirsch, J. A. W. & Bleiwess, R. 1994 Jackknifing of weigthed trees: Validation of phylogenies reconstructed from distance matrices. *Molecular Phylogenetics and Evolution* **3**, 256-267.

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.

Lapointe, F.-J. & Levasseur, C. 2004 Everything you always wanted to know about average consensus, and more. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, vol. 4 (ed. O. R. P. Bininda-Emonds), pp. 87-105. Dordrecht: Kluwer Academic Publishers.

Lapointe, F.-J., Wilkinson, M. & Bryant, D. 2003 Matrix representations with parsimony or with distances: Two sides of the same coin? *Systematic Biology* **52**, 865-868.

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

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.

Levasseur, C., Landry, P.-A. & Lapointe, F.-J. 2000 Estimating trees from incomplete distance matrices: a comparison of two methods. In *Data Analysis, Classification, and Related Methods* (ed. H. A. L. Kiers, J.-P. Rasson, P. J. F. Groenen & M. Schader), pp. 149-154. Berlin: Springer.

Levasseur, C., Landry, P.-A., Makarenkov, V. Kirsch, J.A.W., & Lapointe, F.-J. 2003 Incomplete distance matrices, supertrees and bat phylogeny. *Molecular Phylogenetics and Evolution* **27**, 239-246.

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. 2002 A family of average consensus methods for weighted trees. In *Classification, Clustering, and Data Analysis* (ed. K. Jajuga, A. Sokolowski & H.-H., Bock), pp. 365-369. Studies in Classification, Data Analysis, and Knowledge organization, Berlin: Springer-Verlag.

Levasseur, C. & Lapointe, F.-J. 2003 Increasing phylogenetic accuracy with global congruence. In *Bioconsensus* (ed. M. Janowitz, F.-J. Lapointe, F. R. McMorris, B. Mirkin & F. S. Roberts), pp. 221-229. Providence: American Mathematical Society.

Liu, F. R. & Miyamoto, M. M. 1999 Phylogenetic assessment of molecular and morphological data for eutherian mammals. *Systematic Biology* **48**, 54-64.

Lutzoni, F. M. 1997 Phylogeny of lichen- and non-lichen-forming omphalinoid mushrooms and the utility of testing for combinability among multiple data sets. *Systematic Biology* **46**, 373-406.

Mace, G. M., Gittleman, J. L. & Purvis, A. 2003 Preserving the Tree of Life. *Science* **300**, 1707-1709.

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

Makarenkov, V. & Leclerc, B. 1999 The fitting of a tree metric according to a weighted least-squares criterion. *Journal of Classification* **16**, 3-26.

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

Mason-Gamer, R. J. & Kellogg, E. A. 1996 Testing for phylogenetic conflict among molecular data sets in the tribe Triticeae (Gramineae). *Systematic Biology* **45**, 524-545.

Messenger, S. L. & McGuire, J. A. 1998 Morphology, molecules and the phylogenetics of Cetaceans. *Systematic Biology* **47**, 90-124.

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

Mickevich, M. F. & Farris, J. S. 1981 The implication of congruence in *Menidia*. *Systematic Zoology* **30**, 351-370.

Mickevich, M. F. & Platnick, N. J. 1989 On the information content of classifications. *Cladistics* **5**, 33-47.

Miyamoto, M. M. 1985 Consensus cladograms and general classifications. *Cladistics* **1**, 186-189.

Miyamoto, M. M. & Cracraft, J. 1991 Phylogenetic inference, DNA sequence analysis, and the future of molecular systematics. In *Phylogenetic Analysis of DNA Sequences* (ed. M. M. Miyamoto & J. Cracraft), pp. 3-17. New York: Oxford University Press.

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

Nei, M. & Kumar, S. 2000 *Molecular Evolution and Phylogenetics*. New York: Oxford University Press.

Nelson, G. 1979 Cladistics analysis and synthesis: Principles and definitions, with a historical note on Adanson's Familles des Plantes (1763-1764). *Systematic Zoology* **28**, 1-21.

Nixon, K. C. & Carpenter, J. M. 1996 On consensus, collapsibility, and clade concordance. *Cladistics* **12**, 305-321.

O'Grady, P. M., Remsen, J. & Gatesy, J. E. 2002 Partitioning of multiple data sets in phylogenetic analysis. In *Techniques in Molecular Systematics and Evolution* (ed. R. DeSalle, G. Giribet & W. Wheeler). Basel: Birkhauser Verlag.

Olmstead, R. G. & Sweere, J. A. 1994 Combining data in phylogenetic systematics: An empirical approach using three molecular data sets in the Solanaceae. *Systematic Biology* **43**, 467-481.

Omland, K. E. 1994 Character congruence between a molecular and a morphological phylogeny for dabbling ducks (*Anas*). *Systematic Biology* **43**, 369-386.

Page, R. D. M. 1992 Comments on the information content of classifications. *Cladistics* **8**, 87-95.

Page, R. D. M. 1996 TREEVIEW: An application to display phylogenetic trees on personal computers. *Computer Applications in the Biosciences* **12**, 357-358.

Page, R. D. M. 2002 Modified mincut supertrees. *Lecture Notes in Computer Science* **2452**, 537-551.

Patterson, C., Williams, D. M. & Humphries, C. J. 1993 Congruence between molecular and morphological phylogenies. *Annual Review of Ecology and Systematics* **24**, 153-188.

Pennington, R. T. 1996 Molecular and morphological data provide phylogenetic resolution at different hierarchical levels in *Andira*. *Systematic Biology* **45**, 496-515.

Piaggio-Talice, R., Burleigh, J. G. & Eulenstein, O. 2004 Quartet supertrees. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, vol. 4 (ed. O. R. P. Bininda-Emonds), pp. 173-191. Dordrecht: Kluwer Academic Publishers.

Pisani, D. & Wilkinson, M. 2002 Matrix representation with parsimony, taxonomic congruence, and total evidence. *Systematic Biology* **51**, 162-166.

Purvis, A. 1995 A composite estimate of primate phylogeny. *PhilosophicalTransactions of the Royal Society of London Series B*, **348**, 405-421.

Quackenbush, J. 2001 Computational analysis of microarray data. *Nature Reviews Genetics* **2**, 418-427.

Quicke, D. L. J. & Belshaw, R. 1999 Incongruence between morphological data sets: an example from the evolution of endoparasitism among parasitic wasps (Hymenoptera: Braconidae). *Systematic Biology* **48**, 436-454.

R Development Core Team. 2004 R: A language and environment for statistical computing. *R Foundation for Statistcal Computing*, Vienna, Austria. ISBN 3-900051-07-0, URL http://www.R-project.org.

Ragan, M. A. 1992 Phylogenetic inference based on matrix representation of trees. *Molecular Phylogenetics and Evolution* **1**, 53-58.

Rambault, A. & Grassly, N. C. 1997 Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications in the Biosciences* **13**, 235-238.

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

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 Jounal of Botany* **31**, 257-268.

Rohlf, F. J. 1982 Consensus indices for comparing classifications. *Mathematical Biosciences* **59**, 131-144.

Rohlf, F. J., Colless, D. H. & G., H. 1983 Taxonomic congruence re-examined. *Systematic Zoology* **32**, 144-158.

Ronquist, F. 1996 Matrix representation of trees, redundancy, and weighting. *Systematic Biology* **45**, 247-253.

Rzhetsky, A. & Nei, M. 1992 A simple method for estimating and testing minimum-evolution trees. *Molecular Biology and Evolution* **9**, 945-967.

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

Salamin, N., Hodkinson, T. R. & Savolainen, V. 2002 Building supertrees: An empirical assessment using the grass family (Poaceae). *Systematic Biology* **51**, 112-126.

Sallum, M. A. M., Schultz, T. R., Foster, P. G., Aronstein, K., Wirtz, R. A. & Wilkerson, C. 2002 Phylogeny of Anophelinae (Diptera: Culicidae) based on nuclear ribosomal and mitochondrial DNA sequences. *Systematic Entomology* **27**, 361-382.

Sanderson, M. J. 2003 r8s: inferring absolute rates of molecular evolution, divergence times in the absence of a molecular clock. *Bioinformatics* **19**, 301-302.

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. & Shaffer, H. B. 2002 Troubleshooting molecular phylogenetic analysis. *Annual Review of Ecology and Systematics* **33**, 49-72.

Semple, C. & Steel, M. A. 2000 A supertree method for rooted trees. *Discrete Applied Mathematics* **105**, 147-158.

Shao, K. & Sokal, R. R. 1986 Significance tests of consensus indices. *Systematic Zoology* **35**, 582-590.

Sheldon, F. H. & Bledsoe, A. H. 1993 Avian molecular systematics, 1970's to 1990's. *Annual Review of Ecology and Systematics* **24**, 243-278.

Slowinski, J. B. & Page, R. D. M. 1999 How should species phylogenies be inferred from sequence data? *Systematic Biology* **48**, 814-825.

Smith, T. J. 2001 Constructing ultrametric and additive trees based on the L_1 norm. *Journal of Classification* **18**, 185-207.

Sneath, P.H.A. & Sokal, R.R. 1973 *Numerical Taxonomy-The Principles and Practice of Numerical Classification.* SanFrancisco: W.H. Freeman.

Sokal, R. R. & Rohlf, F. J. 1981 Taxonomic congruence in the Leptopodomorpha re-examined. *Systematic Zoology* **30**, 309-325.

Springer, M. S., Amrine, H. M., Burk, A. & Stanhope, M. J. 1999 Additional support for Afrotheria and Paenungulata, the performance of mitochondrial versus nuclear genes, and the impact of data partitions with heterogeneous base composition. *Systematic Biology* **48**, 65-75.

Springer, M. S. & Kirsch, J. A. W. 1989 Rates of single-copy DNA evolution in phalangeriform marsupials. *Molecular Biology and Evolution* **6**, 331-341.

Steel, M. A. 1992 The complexity of reconstructing trees from qualitative characters and subtrees. *Journal of Classification* **9**, 91-116.

Steel, M. A., Dress, A. W. M. & Böker, S. 2000 Simple but fundamental limitations on supertree and consensus tree methods. *Systematic Biology* **49**, 363-368.

Steel, M. A. & Penny, D. 1993 Distribution of tree comparison metrics-some new results. *Systematic Biology* **42**, 126-141.

Swofford, D. L. 1991 When are phylogeny estimates from molecular and morphological data incongruent? In *Phylogenetic analysis of DNA sequences* (ed. M. M. Miyamoto, Cracraft, J.J.), pp. 295-333. Oxford: Oxford University Press.

Swofford, D. L. 1999 PAUP*: Phylogenetic Analysis Using Parsimony (*and other methods), version 4. Sunderland: Sinauer Associates.

Swofford, D. L., Olsen, G. J., Waddell, P. J. & Hillis, D. M. 1996 Phylogenetic inference. In *Molecular Systematics* (ed. D. M. Hillis, C. Moritz & B. K. Mable), pp. 407-514. Sunderland: Sinauer.

Systma, K. J. 1990 DNA and morphology: inference of plant phylogeny. *Trends in Ecology and Evolution* **5**, 104-110.

Teeling, E. C., Scally, M., Kao, D. J., Romagnoli, M. L., Springer, M. S. & Stanhope, J. 2000 Molecular evidence regarding the origin of echolocation and flight in bats. *Nature* **403**, 188-192.

Thorley, J. L. & Page, R. D. M. 2000 RadCon: Phylogenetic tree comparison and consensus. *Bioinformatics* **16**, 486-487.

Thorley, J. L., Wilkinson, M. & Charleston, M. 1998 The information content of consensus trees. In *Advances in Data Science and Classification* (ed. A. Rizzi, M. Vichi & H.-H. Bock), pp. 91-98. Berlin: Springer-Verlag.

Troyanskaya, O., Cantor, M., Sherlock, G., Brown, P., Hastie, T., Tibshirani, R., Botstein, D. & Altman, R. B. 2001 Missing value estimation methods for DNA microarrays. *Bioinformatics* **17**, 520-525.

Wendel, J. F. & Doyle, J. J. 1998 Phylogenetic incongruence: window into genome history and molecular evolution. In *Molecular systematics of plants II: DNA Sequencing* (ed. D. E. Soltis, P. S. Soltis & J. J. Doyle), pp. 265-296. Boston: Kluwer Academic Publishers.

Wiens, J. J. 1998a Combining data sets with different phylogenetic histories. *Systematic Biology* **47**, 568-581.

Wiens, J. J. 1998b Does adding characters with missing data increase or decrease phylogenetic accuracy? *Systematic Biology* **47**, 625-640.

Wilkinson, M. 1994 Common cladistic information and its consensus representation - reduced adams and reduced cladistic consensus trees and profiles. *Systematic Biology* **43**, 343-368.

Wilkinson, M., Thorley, J. L., Pisani, D., Lapointe, F.-J. & McInerney, J. O. 2004 Some desiderata for liberal supertrees. In *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, vol. 4 (ed. O. R. P. Bininda-Emonds), pp. 227-246. Dordrecht: Kluwer Academic Publishers.

Annexe I

A family of average consensus methods for weighted trees

Cet article est publié sous la référence :

Levasseur, C., Lapointe, F.-J. 2002 A family of average consensus methods for weighted trees. In *Classification, Clustering, and Data Analysis* (ed. K. Jajuga, A. Sokolowski & H.-H., Bock), pp. 365-369. Studies in Classification, Data Analysis, and Knowledge organization, Berlin: Springer-Verlag.

ABSTRACT

Consensus techniques represent useful tools in phylogenetic analysis, particularly for combining trees derived from different data sets. In the present paper, a family of average consensus methods for weighted trees is presented; the mean and median procedures are compared and applied to combine phylogenetic trees while taking into account their branch lengths. We also provide some recommendations about the use of these consensus techniques in relation to the more classical methods based on topological relationships alone.

INTRODUCTION

Consensus techniques are used in phylogenetic analysis to summarize the trees obtained from independent data sets, or combine multiples trees obtained from the same data (Bull *et al.* 1993). In spite of their popularity, consensus methods have been much criticized in the past (Barrett *et al.* 1991; De Queiroz 1993). Because classical methods like the strict (Sokal & Rohlf 1981) and majority rule (Margush & McMorris 1981) consensus are based on topological relationships alone, unresolved trees are often produced by such techniques, a property considered as undesirable by some (see Kluge & Wolf 1993). However, consensus methods for weighted trees that take into account branch lengths (Lapointe 1998a) can do better than most standard procedures, as they are more likely to produce fully resolved trees. In this paper, we describe two consensus methods for weighted trees based on different optimization criteria. These mean and median consensus procedures are used and compared to the more conservative strict consensus in an application involving phylogenetic trees.

A FAMILY OF AVERAGE CONSENSUS METHODS

Let $S = \{1, \ldots, i, \ldots, j, \ldots, n\}$ be a set of n objects and $P = \{\mathbf{T}_1, \ldots, \mathbf{T}_k, \ldots, \mathbf{T}_m\}$ a profile of m weighted trees defined on S. An average consensus method is defined as a function that takes as input the profile P and returns a consensus weighted tree $\mathbf{T}_c$ that is in some sense closest to P. Since there exists a one-to-one correspondence between any weighted tree $\mathbf{T}$ and its associated path-length matrix $\mathbf{D}$ (Buneman 1971; Hartigan 1967), it is equivalent to deal with the trees $\mathbf{T}_k$ of P or their corresponding matrices $\mathbf{D}_k$ to compute the consensus solution $\mathbf{T}_c$. Average consensus trees are thus obtained by applying a median or mean consensus function to the set $M = \{\mathbf{D}_1, \ldots, \mathbf{D}_k, \ldots, \mathbf{D}_m\}$ of path-length matrices associated with the trees of P.

The median consensus for weighted trees

The median procedure was introduced originally by Margush & McMorris (1981) to compute a consensus tree minimizing the sum of symmetric differences to the trees in P. This method applies to unweighted trees only. The same approach can be generalized,

however, to combine weighted trees in a similar fashion. As such, the median consensus tree $\mathbf{T}_c$ is defined as the solution that minimizes the following average consensus function:

$$\sum_{k=1}^{m} \Delta\left(\mathbf{T}_c, \mathbf{T}_k\right) = \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left| d_c(i,j) - d_k(i,j) \right| \quad (1)$$

where the $d_c(i, j)$ are the path-length distances in the matrix $\mathbf{D}_c$ associated with the consensus tree $\mathbf{T}_c$ and the $d_k(i, j)$ are the path-length distances in the matrices $\mathbf{D}_k$ of M associated with the trees $\mathbf{T}_k$ of P. Practically, the median consensus tree $\mathbf{T}_c$ is obtained in two simple steps. First, a median distance matrix $\mathbf{D}_{med}$ is computed from the path-length matrices of M. If the number of trees m is odd, the distances in $\mathbf{D}_{med}$ are given by the $(m+1)/2$ ordered $d(i, j)$ values in the m path-length distance matrices of M, for every pair of objects i and j. If m is even, the distances in $\mathbf{D}_{med}$ are computed as the mean of the $m/2$ and $(m/2)+1$ ordered $d(i, j)$ values in the m path-length distance matrices of M, for every pair of objects i and j. The closest path-length distance matrix from this median matrix $\mathbf{D}_{med}$ is then obtained by minimizing the following loss function using a minimum-absolute-deviation algorithm (e.g. Smith 2001):

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left| d_c(i,j) - d_{med}(i,j) \right| \quad (2)$$

where the $d_c(i, j)$ are the fitted path-length distances in the matrix $\mathbf{D}_c$ associated with the consensus tree $\mathbf{T}_c$ and the $d_{med}(i, j)$ are the distances in the median matrix D_{med}. The resulting tree $\mathbf{T}_c$ is the median consensus solution.

The mean consensus for weighted trees

The average consensus procedure was originally defined by Lapointe & Cucumel (1997) as a consensus function to combine weighted trees. In order to differentiate the use of the more general average consensus that also includes the median function, we will use the more specific mean consensus when referring to Lapointe & Cucumel's average consensus method. The mean consensus tree $\mathbf{T}_c$ is thus defined as the solution that minimizes the following average consensus function:

$$\sum_{k=1}^{m} \Delta\left(\mathbf{T}_c, \mathbf{T}_k\right) = \sum_{k=1}^{m} \sum_{i=1}^{n} \sum_{j=1}^{n} \left[d_c(i,j) - d_k(i,j) \right]^2 \quad (3)$$

where the $d_c(i, j)$ are the path-length distances in the matrix $\mathbf{D}_c$ associated with the consensus tree $\mathbf{T}_c$ and the $d_k(i, j)$ are the path-length distances in the matrices $\mathbf{D}_k$ of M associated with the trees $\mathbf{T}_k$ of P. Again, the mean consensus tree $\mathbf{T}_c$ can be computed in two simple steps. First, a mean distance matrix $\mathbf{D}$ must be computed from the path-length distance matrices of M. The closest path-length distance matrix from this mean matrix $\mathbf{D}$ is then obtained by minimizing the following loss function using a least-squares algorithm (e.g. Cavalli-Sforza & Edwards 1967):

$$\sum_{i=1}^{n} \sum_{j=1}^{n} \left[d_c(i,j) - \bar{d}_k(i,j) \right]^2 \quad (4)$$

where the $d_c(i, j)$ are the fitted path-length distances in the matrix $\mathbf{D}_c$ associated with the consensus tree $\mathbf{T}_c$ and the $d(i, j)$ are the distances in the mean matrix $\mathbf{D}$. The resulting tree $\mathbf{T}_c$ is the mean consensus solution.

Application of the average consensus

To illustrate the use of the median and the mean consensus, we applied these procedures to combine three phylogenies of frog species derived from morphological data, mating calls and molecular sequences (data from Cannatella *et al.* 1998). The least-squares trees (Cavalli-Sforza & Edwards 1967) computed from the different data sets are presented in Figure I.1. It shows that phylogenies estimated from morphology (Figure I.1A) or calls (Figure I.1B) are topologically identical, whereas the tree computed from molecular data (Figure I.1C) is different. The conservative strict consensus tree (Figure I.1D) of those three phylogenies is not well resolved and contains a single species pair found in all of the input trees. On the other hand, the median (Figure I.1E) and mean (Figure I.1F) consensus trees that take into account branch lengths are much more resolved. In this particular case, the median consensus tree is the same as the phylogeny derived from the calls data (Figure I.1B). The mean consensus tree, however, differs from all input trees and contains a combination of species groups not found in the original phylogenies.

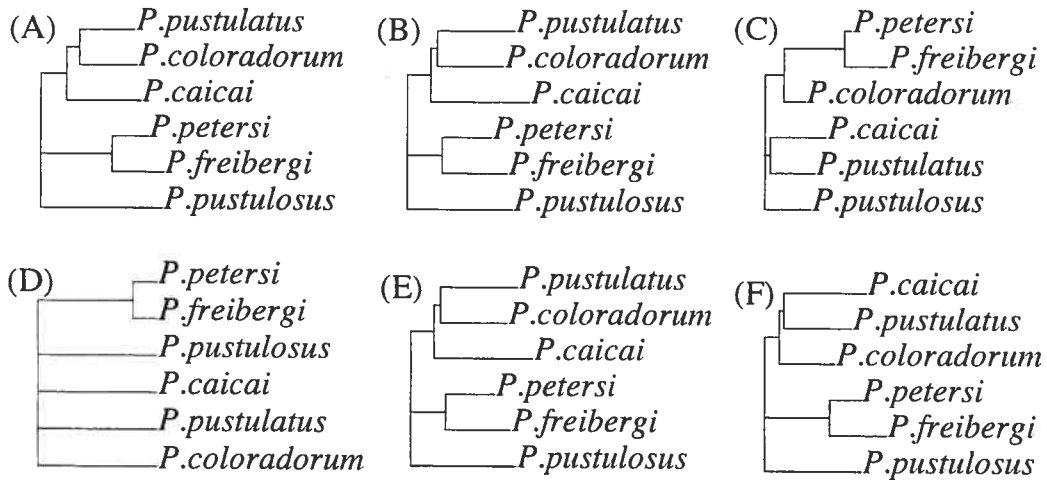

Figure I.1 Trees derived from the three different data sets and their different consensus trees: (A) morphological data, (B) mating calls, (C) molecular data, (D) strict consensus, (E) median consensus, (F) mean consensus.

Discussion

The median and mean consensus procedures are designed to combine weighted trees while taking into account their branch lengths. These average consensus methods are based on different optimization criteria, however, and they may not always produce identical results. For one, the mean consensus function minimizes the L_2 norm and is very likely to be affected by extreme values. On the other hand, the median function minimizes the L_1 norm; this consensus method should be more appropriate in cases involving outliers. Our application illustrates the differences between the two techniques. In this particular example, the molecular tree (Figure I.1C) clearly influenced the mean consensus since it represented a very different phylogeny compared to the other two trees (Figure I.1A,B). Both methods will produce comparable solutions when the tree profile P encompasses a homogeneous distribution of phylogenies, or when all trees are very similar (or identical in topology). Regardless of the average consensus function selected, these methods will usually produce fully resolved trees, contrary to consensus techniques that ignore branch lengths. Such average consensus techniques thus offer interesting alternatives to the classical consensus methods based on topological relationships alone.

Acknowledgments

We would like to thank Buck McMorris and Fred Roberts for their helpful suggestions about our work, and Guy Cucumel and Olivier Gauthier for their assistance with the LATEX format. This work was made possible by a NSERC scholarship to C. Levasseur and by a NSERC grant no. OGP0155251 to F.-J. Lapointe.

« L'imagination, ce n'est pas le mensonge. »

Messieurs les enfants

Daniel Pennac