Gene 441 (2009) 119–125

Contents lists available at ScienceDirect

Gene

journal homepage: www.elsevier.com/locate/gene

A likelihood look at the supermatrix–supertree controversy

Fengrong Ren [a], Hiroshi Tanaka [a], Ziheng Yang [b,c,*]

[a] Advanced Biomedical Information, Center for Information Medicine, Tokyo Medical and Dental University, Tokyo, Japan
[b] Department of Biology, University College London, London, UK
[c] Graduate School of Agriculture and Life Sciences, University of Tokyo, Tokyo, Japan

ARTICLE INFO

Article history:
Received 24 January 2008
Received in revised form 25 March 2008
Accepted 1 April 2008
Available online 10 April 2008

Keywords:
Bayesian
Combined analysis
Likelihood
Likelihood supertree
Supermatrix
Supertree

ABSTRACT

Supermatrix and supertree methods are two strategies advocated for phylogenetic analysis of sequence data from multiple gene loci, especially when some species are missing at some loci. The supermatrix method concatenates sequences from multiple genes into a data supermatrix for phylogenetic analysis, and ignores differences in evolutionary dynamics among the genes. The supertree method analyzes each gene separately and assembles the subtrees estimated from individual genes into a supertree for all species. Most algorithms suggested for supertree construction lack statistical justifications and ignore uncertainties in the subtrees. Instead of supermatrix or supertree, we advocate the use of likelihood function to combine data from multiple genes while accommodating their differences in the evolutionary process. This combines the strengths of the supermatrix and supertree methods while avoiding their drawbacks. We conduct computer simulation to evaluate the performance of the supermatrix, supertree, and maximum likelihood methods applied to two phylogenetic problems: molecular-clock dating of species divergences and reconstruction of species phylogenies. The results confirm the theoretical superiority of the likelihood method. Supertree or separate analyses of data of multiple genes may be useful in revealing the characteristics of the evolutionary process of multiple gene loci, and the information may be used to formulate realistic models for combined analysis of all genes by likelihood.

1. Introduction

There has been much discussion concerning the best strategy to conduct phylogenetic analysis of sequence data from multiple gene loci, especially when some genes are not yet sequenced in some species. Two strategies have been advocated. The supermatrix method concatenates sequences from multiple loci into a super-sequence, and uses the resulting data supermatrix to perform phylogenetic analysis. The supertree method conducts phylogenetic analysis on individual genes separately, and then combines the subtrees from the individual genes into a supertree for all species. Several useful reviews of the controversy have been published; see, *e.g.*, de Queiroz and Gatesy (2007) in support of the supermatrix method and Sanderson (1998), Bininda-Emonds et al. (2002) and Bininda-Emonds (2004) advocating supertree methods. The supermatrix–supertree debate has similarity to an earlier debate concerning combined analysis (sometimes called "total evidence") versus separate analysis (see, e.g., Huelsenbeck et al., 1996), although in the new controversy an emphasis is placed on the fact that some species are missing at some loci.

From a statistical point of view, neither supermatrix nor supertree methods are ideal. The supermatrix method or combined analysis ignores differences among genes in the substitution rates, base compositions, or other aspects of the evolutionary process. Computer simulations suggest that ignoring differences among sites can have an adverse impact on phylogenetic analysis, sometimes causing the estimated tree to be inconsistent (e.g., Kuhner and Felsenstein, 1994; Tateno et al., 1994; Huelsenbeck, 1995; Yang, 1995). The supertree method or separate analysis estimates an independent set of parameters for every gene and may over-fit the data and cause large variances in the estimates. Most supertree methods use heuristic algorithms that cannot be justified rigorously on a statistical basis. They also ignore uncertainties in the estimated subtrees (such as bootstrap support values, Bayesian posterior clade probabilities, or estimated branch lengths), although heuristic algorithms are recently proposed to remedy this problem (Burleigh et al., 2006; Moore et al., 2006).

The statistical likelihood provides a natural framework for combining information from different experiments (Edwards, 1992). Suppose two experiments have been conducted to estimate a binomial probability p, with the first experiment generating x "successes" out of m trials, and the second y successes out of n trials. The simple average $(x/m+y/n)/2$ may not be efficient if m and n are very different. One can use the likelihood to combine the information from the two datasets. The likelihood is $p^x(1-p)^{m-x}$ for the first

Abbreviations: ML, maximum likelihood; MY, million years.

* Corresponding author. Department of Biology, University College London, Darwin Building, Gower Street, London WC1E 6BT, UK. Tel.: +44 20 7679 4379; fax: +44 20 7679 7096.
E-mail address: z.yang@ucl.ac.uk (Z. Yang).

0378-1119/$ – see front matter © 2008 Elsevier B.V. All rights reserved.
doi:10.1016/j.gene.2008.04.002

experiment and $p^y(1-p)^{n-y}$ for the second. The likelihood from both experiments is the product of the two, that is, $p^{x+y}(1-p)^{m+n-x-y}$, giving the estimate $(x+y)/(m+n)$ for p from the combined data, as is desired. Similarly, the likelihood can be used in multi-parameter models to combine information from heterogeneous datasets while accommodating their differences. One may construct the model such that some parameters are applied to all datasets while others are allowed to vary among datasets to account for the dataset-specific characteristics. Such models are useful to extract the maximum amount of information about the common parameters of interest from all datasets.

Thus besides supermatrix and supertree, an alternative strategy is to take a statistical modeling approach and use the likelihood function to combine sequence data from multiple genes. Different parameters in the substitution model may be used for different loci to accommodate their differences in the evolutionary process. This approach has been discussed by Yang (1996; see also Pupko et al., 2002; Philippe et al., 2005; Shapiro et al., 2006; Bofkin and Goldman, 2007) for the maximum likelihood (ML) method of phylogenetic analysis and by Suchard et al. (2003) and Nylander et al. (2004) for the Bayesian method.

In this paper we conduct computer simulations to examine the performance of the supermatrix, supertree, and ML methods when they are applied to two problems of phylogenetic analysis. The first is molecular-clock dating, that is, estimation of species divergence times on a rooted tree under the molecular clock. The second is reconstruction of unrooted phylogenetic trees. The supermatrix–supertree controversy has mostly concerned the problem of phylogeny reconstruction. Here we consider divergence time estimation as well, as this is a more-conventional estimation problem and its analysis serves to illustrate the principles.

2. Materials and methods

2.1. Divergence time estimation

Replicate datasets are simulated using a small C program called MULTIEVOLVER, written by Z.Y. This uses the EVOLVER program in the PAML package (Yang, 1997) to simulate sequence alignments at multiple loci, and use these alignments to generate data files needed for the analyses discussed below. The simulated datasets are then analyzed using the BASEML program in the PAML package.

The tree of Fig. 1a is used to simulate sequence data at three loci. The ages of ancestral nodes, measured by the expected number of nucleotide substitutions per site at the first locus, are $t_0=1$, $t_1=0.2$, $t_2=0.4$, $t_3=0.6$, and $t_4=0.8$. The K80 model of nucleotide substitution (Kimura, 1980) is assumed. The sequence lengths for the three loci are 500, 300, and 200 sites; the relative substitution rates are 1:2:0.2; and the transition/transversion rate ratios (κ or α/β in Kimura's notation) are 1, 2, and 4 (Table 1). The use of the different sequence lengths and different parameter values for the loci mimics the real-data situation in which different genes have different lengths and rates and different information content. After all six sequences are generated at each locus, some species are deleted at some loci. Thus each replicate dataset consists of an alignment of four sequences for species A, D, E and F at locus 1; an alignment of five sequences for species A, B, C, D and E at locus 2; and an alignment of four sequences for species A, C, D, and F at locus 3 (Table 1, Fig. 1b–d). The number of replicate datasets is 1000.

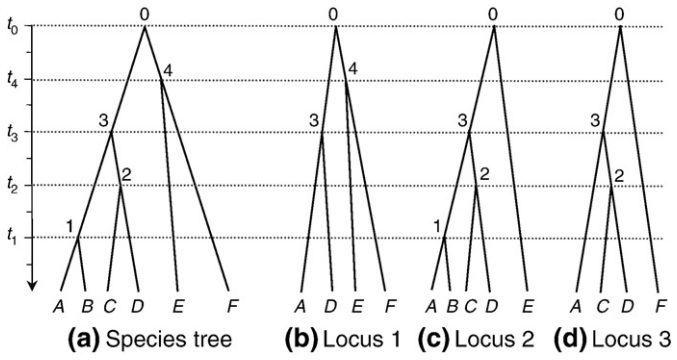

Fig. 1. (a) A rooted tree for six species used in simulations to compare the supermatrix, supertree, and ML methods for divergence time estimation. The age of the root is fixed at $t_0=1$, corresponding to use of fossil calibration at the node. Ages of the other nodes t_1–t_4 are estimated from the sequence data. The true times are $t_1=0.2$, $t_2=0.4$, $t_3=0.6$, and $t_4=0.8$. (b–d) The data consist of sequence alignments at three loci, with 4, 5, and 4 species, respectively.

Table 1
Evolutionary parameters at the three loci

Parameter	Locus 1	Locus 2	Locus 3
Number of sites	500	300	200
Rates	1	2	0.2
Transition/transversion rate ratio κ	1	2	4
Species sampled	A, D, E, F	A, B, C, D, E	A, C, D, F

The simulated sequence alignments at three loci are used to estimate species divergence times under the molecular clock using the "supermatrix", "supertree", and ML methods. Here our use of those terms is by analogy with the corresponding methods for tree topology reconstruction. The supermatrix method concatenates the sequences for time estimation without accommodating the differences among loci. The supertree method performs separate analysis of data at different loci, estimating times for each locus and then merging the estimates into one set, similar to the use of supertree algorithms to assemble the subtrees from different loci into one supertree. The ML method conducts a combined analysis of all data, accommodating data heterogeneity.

During the likelihood analysis, the age of the root is fixed at $t_0=1$, while node ages t_1–t_4 as well as the substitution rates and the parameter κ are estimated from the data by ML. If one time unit represents 100 million years (MY), the root age is fixed at 100 MY, mimicking the use of a fossil to calibrate the molecular clock. The true rates at the three loci are then 1, 2, and 0.2 substitutions per site per 100MY.

The three methods are implemented as follows.

2.1.1. Supermatrix method

Each replicate dataset consists of a sequence alignment generated by concatenating the sequences at the three loci, with nucleotides in the missing species coded as question marks. The data are then analyzed under the K80 model on the full tree of Fig. 1a. Six parameters are estimated: the times t_1–t_4, the rate r, and κ. All sites in the supermatrix are assumed to be independently and identically distributed and differences among the genes are ignored.

We also use the K80+Γ_5 model in the supermatrix analysis, which assumes a gamma distribution of rates among sites in the alignment, with five rate categories used (Yang, 1994). This does not match the model used to generate the data, as differences in κ are ignored and the random gamma distribution may not accommodate adequately the fixed rate differences among the three genes. This model involves one extra parameter than K80, the gamma shape parameter α.

2.1.2. Supertree method

An alignment is generated for each locus and analyzed separately on the subtree for that locus (Fig. 1b–d). The subtrees share the same root so that the same fossil calibration is used in all three analyses. At any locus, the rate r and parameter κ are estimated together with the

time parameters: t_3 and t_4 for locus 1; t_1, t_2 and t_3 for locus 2; and t_2 and t_3 for locus 3 (Fig. 1b–d). The time estimates are combined into one set of estimates by using simple averages across the loci:

$$\begin{aligned}\hat{t}_1 &= \hat{t}_1^{(2)},\\ \hat{t}_2 &= \left(\hat{t}_2^{(2)} + \hat{t}_2^{(3)}\right)/2,\\ \hat{t}_3 &= \left(\hat{t}_3^{(1)} + \hat{t}_3^{(2)} + \hat{t}_3^{(3)}\right)/3,\\ \hat{t}_4 &= \hat{t}_4^{(1)},\end{aligned} \quad (1)$$

where $\hat{t}_1^{(2)}$ is the estimate of t_1 from locus 2 and so on. The averaging by Eq. (1) mimics supertree algorithms for tree reconstruction, which combines subtrees into one supertree. Here a preferred approach is to use the variances of time estimates from the loci to calculate a weighted average, taking into account the different sampling errors in the locus-specific estimates. In phylogeny reconstruction, however, it is much less clear how to use measures of sampling errors such as bootstrap proportions on the subtrees in supertree construction.

2.1.3. ML method

This method performs a combined analysis of data at all loci. As in the supermatrix analysis, a sequence alignment is generated by concatenating sequences at the three loci. However, when the data are analyzed by ML, different rates and κs are estimated for the three loci, while the same set of time parameters are estimated for all loci. Thus the model estimates five times (t_1, t_2, t_3, t_4, t_5), three rates (r_1, r_2, r_3) and three κ parameters (κ_1, κ_2, κ_3) for the three loci.

2.1.4. Calculation of the likelihood function

The probability of observing data at any locus, say locus 1, can be calculated using either the subtree (Fig. 1b) or the full tree (Fig. 1a). On the subtree (Fig. 1b), the probability is a function of times t_0, t_3, t_4, the rate r_1, and κ_1, and is calculated by summing over ancestral states at nodes 0, 3, and 4 using Felsenstein's (1981) pruning algorithm. On the full tree (Fig. 1a), the probability is considered a function of all time parameters t_0, t_1, t_2, t_3, t_4, t_5, the rate r_1, and κ_1, and is calculated by summing over nucleotide states at ancestral nodes 0, 1, 2, 3, and 4, as well as over the missing nucleotides in species B and C (Felsenstein, 2004, p. 255; Yang, 2006, pp. 107–108). The probability on the full tree is then in fact independent of t_1 and t_2, and is identical to that calculated on the subtree.

The likelihood for the whole dataset, consisting of sequences at all three loci, is the product of the probabilities across the three loci. Here in the supermatrix and ML methods we calculate the likelihood on the full tree, as the implementation is simpler even though it involves more computation than calculation on the sub trees (Yang, 2004). It is clear that the supermatrix method is just a special model in the ML analysis, assuming equality of rates and κ ratios across the three loci: $r_1 = r_2 = r_3$ and $\kappa_1 = \kappa_2 = \kappa_3$.

Similarly, the supertree method may be considered a parameter-rich likelihood model involving the following parameters: $t_3^{(1)}$, $t_4^{(1)}$, r_1 and κ_3 for locus 1; $t_1^{(2)}$, $t_2^{(2)}$, $t_3^{(2)}$, r_2 and κ_2 for locus 2; and $t_2^{(3)}$, $t_3^{(3)}$, r_3 and κ_1 for locus 3. (Note that as discussed above, times for the missing ancestral nodes $t_1^{(1)}$, $t_2^{(1)}$, $t_4^{(2)}$, $t_1^{(3)}$, and $t_4^{(3)}$ are not identifiable and are not considered parameters.) As the three loci do not share any parameters in the model, maximization of the probability of the whole dataset is equivalent to separate maximizations of the probabilities at the three loci. The drawback of the method is that even though only one set of time parameters exist, different sets are estimated, with the estimates combined *post hoc* into one set using Eq. (1). The ML method is superior as it assumes only one set of time parameters for all three loci, allowing one locus to borrow information from other loci about the time parameters. This difference may be important if the data are not informative and separate estimates at some loci are unreliable.

2.2. Phylogenetic tree reconstruction

The simulation design is similar to that for divergence time estimation but now unrooted trees are used. The true tree for six species is shown in Fig. 2a. The four internal branch lengths are assumed to be equal in the true tree, represented by b_0, while a, b, c, d, e, and f are the external branch lengths leading to species A, B, C, D, E, and F. The simulation parameters for the three loci are the same as before, shown in Table 1. The shape of the true tree as reflected in the branch lengths is known to affect the relative performance of tree tree-reconstruction methods. To examine the impact of the tree shape, we use four sets of branch lengths in the true tree to simulate data. The details will be provided when we discuss the results. The performance of the methods is measured by the proportion of replicate datasets in which the three clades (or internal branches) in the true tree are recovered.

The supermatrix, supertree, and ML methods of tree reconstruction are implemented as follows. We use exhaustive tree search to analyze the small datasets generated in this study. This approach is unfeasible for practical data analysis, for which efficient heuristic tree-search algorithms have to be implemented. We also describe a supertree-construction algorithm, called ML-supertree, and include it in the comparison.

2.2.1. Supermatrix method

Each replicate dataset consists of an alignment of six concatenated sequences, with undetermined nucleotides in missing sequences coded as question marks. The data are then analyzed under the K80 model, ignoring differences among the genes. All 105 unrooted trees for six species are evaluated to identify the ML tree. The proportion of replicate datasets in which each of the three internal branches in the true tree is recovered in the ML tree are calculated.

2.2.2. Supertree methods

An alignment is generated for every locus. The number of species at the three loci are 4, 5 and 4, so that 3, 15 and 3 unrooted trees are evaluated at the three loci, respectively. Many supertree algorithms can be used to combine the ML trees for the three loci into one supertree of all six species (see Bininda-Emonds, 2005 for a summary). Here we use the program Clann 3.0.2 (Creevey and McInerney, 2005). Its default option implements the *most similar supertree method*, which uses the so-called DFIT criterion to compare all 105 possible supertrees. First, note that every supertree induces a unique subtree for every locus, formed by pruning the missing species off the supertree. For instance, the supertree of Fig. 2a induces the subtrees of Fig. 2b–d for the three loci. At every locus, the distance d_{ij} between any two species i and j on the estimated subtree is calculated as the number of nodes separating the two species. A similar distance d_{ij} is calculated for the subtree induced by the supertree. Then

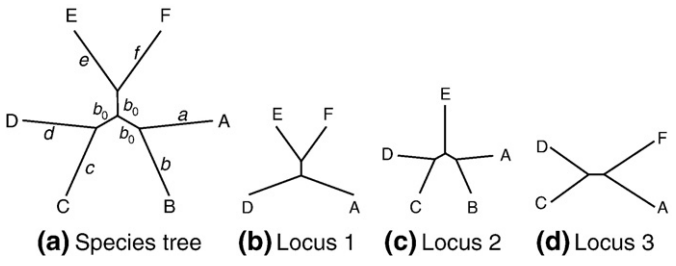

Fig. 2. (a) An unrooted tree for six species used in simulations to compare the supermatrix, supertree, and ML methods of phylogeny reconstruction. The branch lengths, measured by the expected number of nucleotide substitutions per site, are b_0 for all three internal branches, and are a, b, c, d, e, and f for the external branches leading to the corresponding species. (b–d) The analyzed data consist of sequence alignments for 4, 5 and 4 species at the three loci, respectively.

$\frac{1}{s(s-1)}\sum_{i<j}|d_{ij}-d'_{ij}|$, with the sum over all species pairs on the subtree, is used as the distance between the supertree and the reconstructed subtree at the locus. This is summed over the loci to produce the DFIT score for the supertree, and an exhaustive search is performed to find the supertree with the best (minimum) DFIT score. Note that the weighting factor $1/[s(s-1)]$ is used to compensate for the different sizes of subtrees at the different loci. It is somewhat arbitrary, as is the definition of the distance between subtrees.

In addition we implement an *ML-supertree* method, which uses the log likelihood scores as additive weights to weight subtrees when they are "combined" into supertrees. Our exhaustive-search algorithm uses all subtrees and their log likelihood scores at every locus, and not just the best subtree for the locus. First we calculate the log likelihood scores for all possible subtrees, evaluating 3, 15, and 3 subtrees at the three loci, respectively. Second, we construct a 105×3 super-sub map $\{M_{i,j}\}$, where $M_{9,2}=3$, say, means that supertree 9 induces (at locus 2) subtree 3. Third, the log likelihood score for every supertree is calculated by summing, over loci, the log likelihood scores of the induced subtrees. This is also the log likelihood for the whole dataset under the model that assumes an independent set of parameters (including branch lengths and substitution parameters) for every locus. A similar procedure has been discussed by Adachi and Hasegawa (1996) and Yang (1996) for the case of no missing data.

2.2.3. ML method

The same data as used in the supermatrix analysis are analyzed under the K80 model, but the sites from the three loci are assumed to have different rates and different κs (Yang 1996). The model assumes that the same tree topology fits data at all three loci and that the branch lengths are proportional at the three loci. This is the true model used for generating the data. The ML tree is estimated by evaluating all 105 possible trees.

2.2.4. Calculation of the likelihood function

Our discussion of likelihood calculations for divergence time estimation applies to tree reconstruction as well. Here unrooted trees are used and the branch lengths are parameters in place of the divergence times.

3. Results

3.1. Divergence time estimation

Table 2 shows the means and standard errors of the estimates of times and rates by the supermatrix (under both the K80 and K80+Γ_5 models), supertree, and ML methods. We consider the ML method first, since it has the optimal performance and provides a basis for comparison. This method analyzes the data at the three loci simultaneously, estimating different rates and κ parameters for the three loci, with 10 parameters in total. The means of estimates of all parameters are close to their true values.

Table 2
Means and standard deviations (among simulated replicates) of estimates of times and other parameters by different methods

	Truth	Supermatrix		Supertree	ML
		K80	K80+Γ_5		
t_1	0.2	0.416±0.042	0.234±0.051	0.169±0.095	0.197±0.036
t_2	0.4	0.468±0.039	0.331±0.052	0.374±0.105	0.399±0.057
t_3	0.6	0.701±0.042	0.543±0.068	0.577±0.107	0.599±0.057
t_4	0.8	0.943±0.064	0.938±0.089	0.805±0.102	0.804±0.106
r_1	1	0.785±0.034	1.792±0.468	1.019±0.092	1.040±0.512
r_2	2			4.100±3.370	2.100±0.340
r_3	0.2			0.201±0.026	0.201±0.023
κ_1	1	1.472±0.133	1.881±0.276	1.004±0.213	1.004±0.213
κ_2	2			2.059±0.428	2.050±0.422
κ_3	4			4.125±0.943	4.121±0.939

Note: The supermatrix method uses one rate and one κ for all loci.

The supermatrix method performed very poorly. The model used ignores the differences in the substitution rate and in κ among the three loci and is thus seriously misspecified. All four node ages (t_1, t_2, t_3, t_4) are seriously overestimated. Previous studies suggest that the variation in rate often has more impact than variation in κ, and ignoring rate variation among sites leads to underestimation of sequence distances, with more serious bias for large distances than for small ones (Yang, 1996). Ignoring rate variation thus has the effect of overestimating the ages of nodes younger than the calibration point and underestimating the ages of nodes older than the calibration. The results of Table 2 are consistent with this interpretation. The K80+Γ_5 model, by using a random gamma distribution to accommodate variable rates among sites, produced much better estimates, even though they are still biased.

In the supertree method, data at each locus are analyzed separately under the K80 model to estimate the node ages, the rate r and parameter κ for the locus, with 4, 5, and 4 parameters estimated at the three loci, respectively. The estimates of shared times are averaged using Eq. (1) to produce estimates for the whole dataset. Thus the K80 model is correct in analysis of every locus, and the only drawback of the method is the over-fitting of the model to the data. The method produced reasonably good estimates, with node ages t_1, t_2 and t_3 slightly underestimated. The mean of rate r_2 for locus 2 is much higher than the true value, but estimates of r_2 vary considerably among replicates, and the median (2.10) is close to the true value. Note that locus 2 consists of highly divergent short sequences, which lack information. Compared with ML, the time estimates produced by the supertree method have larger variances.

3.2. Tree topology reconstruction

Four sets of branch lengths for the true tree of Fig. 2a are used in the simulation, referred to as four trees. We use short internal branches in the true tree, so that the tree is hard to reconstruct and it is easy to tell the different methods apart. Method performance is measured by the proportion of replicate datasets in which each of the three internal branches, referred to as clades (AB), (CD) and (EF), is recovered in the ML tree. The results are shown in Table 3.

In tree 1, all three internal branches have the length $b_0=0.01$, while the external branches have the lengths $a=c=e=0.02$ and $b=d=f=0.2$. The supermatrix method (K80 model) ignores differences among loci and performed poorly. Use of the K80+Γ_5 model leads to some improvement. The supertree method (DFIT) recovered the (AB) clade with high probability, but not the (CD) and (EF) clades. The ML-supertree method recovered the (AB) clade less often than DFIT but recovered the (CD) and (EF) clades with much higher probabilities. The ML method performed best, recovering all three clades with probabilities >80%.

In tree 2, the internal branch length is $b_0=0.04$, while all external branches have the same length 0.4. For this tree, the supermatrix method has the worst performance, and inclusion of the gamma model (K80+Γ_5) led to very little improvement. A wrong tree topology is found to be the ML tree in most datasets: (AB(((CF)D)E)). This is also the consensus tree among the 1000 ML trees, with the clade support values 99% for (CF), 37% for (AB), and 27% for (ABE). The two supertree methods performed similarly, with DFIT being better in recovering the (AB) clade and ML-supertree being better for clades (CD) and (EF). ML performed the best, recovering all three clades with higher probabilities than all other methods.

Tree 3 has a similar shape to tree 2, but with less sequence divergence. The internal and external branch lengths are 0.01 and 0.1, respectively. The supermatrix method had overall the poorest

Table 3
Proportions of datasets in which a clade is correctly recovered by different methods

	Supermatrix				
Clade	K80	K80+Γ_5	Supertree	ML-supertree	ML
Tree 1 (b_0=0.01, a=c=e=0.02, b=d=f=0.2)					
AB–CDEF	0.820	0.827	0.849	0.805	0.846
CD–ABEF	0.399	0.421	0.378	0.764	0.811
EF–ABCD	0.374	0.392	0.685	0.718	0.831
Tree 2 (b_0=0.04, a=b=c=d=e=f=0.4)					
AB–CDEF	0.373	0.395	0.426	0.347	0.477
CD–ABEF	0.010	0.013	0.428	0.433	0.538
EF–ABCD	0.007	0.012	0.405	0.417	0.564
Tree 3 (b_0=0.01, a=b=c=d=e=f=0.1)					
AB–CDEF	0.659	0.653	0.699	0.645	0.713
CD–ABEF	0.213	0.216	0.460	0.647	0.707
EF–ABCD	0.203	0.205	0.517	0.612	0.747
Tree 4 (b_0=0.02, a=b=0.1, c=d=0.2, e=f=0.3)					
AB–CDEF	0.863	0.858	0.686	0.600	0.747
CD–ABEF	0.023	0.000	0.551	0.599	0.655
EF–ABCD	0.019	0.000	0.488	0.490	0.621

performance, with very little difference between the two models (K80 and K80+Γ_5). The supertree method (DFIT) is slightly better than ML-supertree at recovering clade (AB) but considerably worse at recovering clades (CD) and (EF). ML performed considerably better than all other methods, especially in recovering clades (CD) and (EF).

Tree 4 has internal branch length b_0=0.02, and external branch lengths a=b=0.1, c=d=0.2, and e=f=0.3. The supermatrix method (under both K80 and K80+Γ_5 models) performed poorly. It recovered clade (AB) with very high probability, but clades (CD) and (EF) were rarely recovered. The true tree is the ML tree in only 1.1% of datasets. In 48% of datasets, the ML tree is the wrong tree (AB(((CF)D)E)), which is also the consensus tree among the 1000 ML trees, with clade support values 98% for (CF), 86% for (AB), and 55% for (ABE). The two supertree methods had similar performance, with DFIT being better than ML-supertree at recovering clade (AB) and worse at recovering (CD) and (EF). The ML method performed the best.

In sum, the supermatrix method recovers clade (AB) well, but not clades (CD) and (EF); in trees 2 and 4, the true tree is not the most frequently recovered tree. In those simulations, the violations of model assumptions appear to be very serious, so that the supermatrix method does not perform well. The supertree method (DFIT) tends to recover clade (AB) better than the ML-supertree method but worse with clades (CD) and (EF). Overall, the ML method performed best in all four trees.

The poor performance of the supermatrix method and the importance of accommodating among-partition heterogeneity highlighted by our results may appear inconsistent with Wiens's (2005) simulation results, in which adding sites in the alignment was found to improve performance in ML and Bayesian methods of tree reconstruction even when some species had incomplete data. The different results may be due to the different experimental designs in the two studies. In Wiens's simulation, the same sites were missing in all species, and the included and missing sites were of the same nature. Our simulation included a hierarchy of sites and loci, with different loci having very different evolutionary characteristics while sites in a locus evolve in similar ways. Missing data are represented by missing whole loci. Thus at any locus, removal of some species may make tree reconstruction particularly prone to the problem of long branch attraction. Our design may be expected to generate harder tree-reconstruction problems, situations in which accommodating among-loci heterogeneity is much more important than in Wiens's simulation.

4. Discussion

In our simulation, the ML method achieved better statistical performance than the supermatrix and supertree methods in estimating species divergence times and in reconstructing species phylogenies. This result is consistent with statistics theory, according to which the likelihood function is the carrier of information in the data about the parameters and is the natural means for assembling information from heterogeneous data sources. Here we use likelihood to combine sequence data from multiple loci to estimate common parameters of interest (such as species divergence times or species phylogenies) while allowing other parameters (such as the substitution rate, transition/transversion rate ratio, base compositions etc.) to differ to describe the idiosyncrasies in the evolutionary process at different loci. From this viewpoint, the supermatrix analysis is a simplistic under-fitting likelihood model that assumes equality of all parameters across loci. Similarly, the supertree analysis is equivalent to an over-fitting likelihood model that assumes different free parameters for every locus, followed by a *post hoc* treatment to merge separate estimates into one set of estimates. The likelihood method advocated here strikes a middle ground, and attempts to combine the strengths of both supermatrix and supertree methods while avoiding their drawbacks.

We emphasize that our simulations have used only small trees, assumed simple models, and examined limited numbers of parameter combinations. The simple experimental design was used partly to simplify the interpretation of the results. In real data analysis, more sophisticated models have to be used to analyze much larger datasets. For example, in our simulation on divergence time estimation, we ignored possible violations of the molecular clock (Thorne et al., 1998) and uncertainties in the fossil record (Drummond et al., 2006; Yang and Rannala, 2006). In our simulation on tree topology estimation, we assumed the simple K80 model with no rate variation among sites within each locus, and ignored the possibility that the different loci may have different evolutionary histories (for procedures to deal with such problems, see Rannala and Yang, 2008). The assumption of proportional branch lengths across loci may be seriously violated in real data due to lineage- and gene-specific selective pressures. Nevertheless, we expect that the general principles highlighted in our simulations should apply to larger datasets and more complex models as well.

Here we discuss some practical difficulties in implementing the approach advocated here, which is to combine datasets from multiple loci while accommodating their important differences. First, partitioning of sites may not be straightforward, and other ways of partitioning sites rather than by gene may be more appropriate. For example, the differences among the three codon positions in protein-coding genes are often even greater than differences among genes, in which case the codon positions may be treated as heterogeneous site partitions (e.g., Yang, 1996; Buckley et al., 2001; Ren et al., 2005; Shapiro et al., 2006; Simon et al., 2006; Bofkin and Goldman, 2007). Second, careful thought may be needed to identify the appropriate models and to decide which parameters should be held constant across genes or site partitions and which should be allowed to differ. Using one rate for every gene may lead to use of too many parameters when datasets of hundreds or thousands of genes are analyzed (Felsenstein, 2001). A standard approach in this case is to construct a random-effects model, treating gene-specific rates as random variables with a statistical distribution. This is straightforward to implement in a hierarchical Bayesian framework (Suchard et al., 2003; Nylander et al., 2004), but will cause considerable computational burden for the ML method. Instead a pragmatic approach may be to partition the genes according to rough estimates of substitution rates (Nishihara et al., 2007), with the same rate assigned for genes in the same partition and different rates for different partitions. Leigh et al. (2008) also evaluated a formal procedure to concatenate

sequences from multiple genes or proteins based on likelihood ratio tests of congruence. Third, the idea of using the likelihood function to combine information across datasets applies to both the ML and Bayesian methods (even though we did not examine the Bayesian method in this study), but may not work for distance and parsimony methods. It may not be useful for analysis of other kinds of data for which realistic likelihood models are unavailable.

A number of supertree-construction methods have been developed, such as matrix representation by parsimony (MRP) and its variants (Baum, 2002; Ragan, 1992), MinCut (Semple and Steel, 2000), semi-strict supertree (Goloboff and Pol, 2002), AncestralBuild (Berry and Semple, 2006), etc. Some methods are also suggested to work specifically with the distance or Bayesian methods of phylogeny reconstruction (Criscuolo et al., 2006; Ronquist et al., 2004). The different supertree methods appear to have very different statistical properties and their relative performance is a focus of much recent research (e.g., Goloboff and Pol, 2002; Pisani and Wilkinson, 2002; Gatesy et al., 2004; Bininda-Emonds, 2004b; Eulenstein et al., 2004; Wilkinson et al., 2005; Goloboff, 2005). However, all of them appear to lack a rigorous statistical justification and fail to account for uncertainties in the estimated subtrees. For example, Eulenstein et al. (2004) compared several supertree algorithms in a simulation study. The different loci had the same sequence length, the same rate and the same number of species, so that the information content is about the same at all loci. Even under this most favorable condition, the supertree algorithms did not perform very well, with some having deteriorating rather than improving performance with the addition of loci. Some supertree analyses used the same gene sequences in constructing subtrees, causing data reuse, as pointed out by, *e.g.*, Springer and de Jong (2001), Gatesy et al. (2002), and Bininda-Emonds (2004b).

Supertree algorithms have been advocated on several grounds, and we appreciate their utility in particular applications (e.g., Bininda-Emonds, 2005; Burleigh et al., 2006). First, the source data may sometimes be unavailable or they are of different types, so that traditional tree tree-reconstruction algorithms may not be used. Second, supertree algorithms are suggested to constitute a divide-and-conquer strategy, useful for circumventing the computational difficulty in analysis of large datasets including many species and many genes and for dealing with missing data (Bininda-Emonds, 2004b). Strategies for generating smaller datasets for subtree reconstruction have also been suggested, such as disk covering (Huson et al., 1999) and biclique (Sanderson et al., 2003). However, we suspect that the simpler alternative of focusing on particular species groups may be more effective. Consider inference of the phylogeny of mammals. One may take the supertree strategy of constructing and analyzing several datasets, each consisting of primates, carnivores and artiodactyls, and of then using the subtrees to construct a supertree. Alternatively, one may study the relationships among primates, among carnivores and among artiodactyls separately and then join the estimated trees to form one big tree for different orders of mammals, requiring a trivial supertree construction. The latter may have better statistical performance, as simulation studies suggest that increased taxon sampling improves phylogenetic accuracy if the phylogenetic scope is kept fixed (e.g., Hillis, 1998; Rannala et al., 1998; Poe and Swofford, 1999; Pollock et al., 2002; Zwickl and Hillis, 2002; Rosenberg and Kumar, 2003).

Current research in supertree algorithms has emphasized their combinatorial and computational properties, with insufficient attention paid to their statistical properties. The supermatrix–supertree debate does not appear to have appreciated the fact that statistical likelihood is a natural tool for combining information from heterogeneous datasets and for dealing with problems such as missing data, non-random taxa sampling and non-overlapping taxa sets. Recently a few attempts have been made to develop bootstrap algorithms to assess and incorporate uncertainties in the subtrees in supertree construction, using parsimony for tree reconstruction (Burleigh et al., 2006; Moore et al., 2006). Some insights may be gained into this problem from a statistical modeling framework, by examining how the likelihood function incorporates information and accommodates uncertainties in multiple heterogeneous datasets.

Acknowledgments

We thank David A. Morrison, Andrew Roger, and Ed Susko for comments. This study is supported by a Grant-in-Aid for Scientific Research from the Ministry of Education, Culture, Sports, and Technology of Japan to F.R. and H.T. and by a Natural Environment Research Council grant to Z.Y.

References

Adachi, J., Hasegawa, M., 1996. MOLPHY Version 2.3: Programs for Molecular Phylogenetics Based on Maximum Likelihood. Computer Science Monographs, vol. 28. Institute of Statistical Mathematics, Tokyo, pp. 1–150.
Baum, B., 2002. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. Taxon 41, 3–10.
Berry, V., Semple, C., 2006. Fast computation of supertrees for compatible phylogenies with nested taxa. Syst. Biol. 55, 270–288.
Bininda-Emonds, O.R.P., 2004a. Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. Kluwer Academic, Dordrecht, the Netherlands.
Bininda-Emonds, O.R.P., 2004b. The evolution of supertrees. Trends Ecol. Evol. 19, 315–322.
Bininda-Emonds, O.R.P., 2005. Supertree construction in the genomic age. Methods Enzymol. 395, 745–757.
Bininda-Emonds, O.R.P., Gittleman, J.L., Steel, M., 2002. The (super)tree of life: procedures, problems and prospects. Ann. Rev. Ecolog. Syst. 33, 265–289.
Bofkin, L., Goldman, N., 2007. Variation in evolutionary processes at different codon positions. Mol. Biol. Evol. 24, 513–521.
Buckley, T.R., Simon, C., Chambers, G.K., 2001. Exploring among-site rate variation models in a maximum likelihood framework using empirical data: effects of model assumptions on estimates of topology, branch lengths, and bootstrap support. Syst. Biol. 50, 67–86.
Burleigh, J.G., Driskell, A.C., Sanderson, M.J., 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. Syst. Biol. 55, 426–440.
Creevey, C.J., McInerney, J.O., 2005. Clann: investigating phylogenetic information through supertree analyses. Bioinformatics 21, 390–392.
Criscuolo, A., Berry, V., Douzery, E.J., Gascuel, O., 2006. SDM: a fast distance-based approach for (super) tree building in phylogenomics. Syst. Biol. 55, 740–755.
de Queiroz, A., Gatesy, J., 2007. The supermatrix approach to systematics. Trends Ecol. Evol. 22, 34–41.
Drummond, A.J., Ho, S.Y.W., Phillips, M.J., Rambaut, A., 2006. Relaxed phylogenetics and dating with confidence. PLoS Biol. 4, e88.
Edwards, A.W.F., 1992. Likelihood. John Hopkins University Press, London.
Eulenstein, O., et al., 2004. Performance of flip supertree construction with a heuristic algorithm. Syst. Biol. 53, 299–308.
Felsenstein, J., 1981. Evolutionary trees from DNA sequences: a maximum likelihood approach. J. Mol. Evol. 17, 368–376.
Felsenstein, J., 2001. Taking variation of evolutionary rates between sites into account in inferring phylogenies. J. Mol. Evol. 53, 447–455.
Felsenstein, J., 2004. Inferring Phylogenies. Sinauer Associates, Sunderland, Massachusetts.
Gatesy, J., Matthee, C., DeSalle, R., Hayashi, C., 2002. Resolution of a supertree/supermatrix paradox. Syst. Biol. 51, 652–664.
Gatesy, J., Baker, R.H., Hayashi, C., 2004. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of Crocodylia. Syst. Biol. 53, 342–355.
Goloboff, P.A., 2005. Minority rule supertrees? MRP, compatibility, and minimum flip may display the least frequent groups. Cladistics 21, 282–294.
Goloboff, P.A., Pol, D., 2002. Semi-strict supertrees. Cladistics 18, 514–525.
Hillis, D.M., 1998. Taxonomic sampling, phylogenetic accuracy, and investigator bias. Syst. Biol. 47, 3–8.
Huelsenbeck, J.P., 1995. The robustness of two phylogenetic methods: four-taxon simulations reveal a slight superiority of maximum likelihood over neighbor joining. Mol. Biol. Evol. 12, 843–849.
Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996. Combining data in phylogenetic analysis. Trends Ecol. Evol. 11, 152–158.
Huson, D.H., Nettles, S.M., Warnow, T.J., 1999. Disk-covering, a fast-converging method for phylogenetic tree reconstruction. J. Comput. Biol. 6, 369–386.
Kimura, M., 1980. A simple method for estimating evolutionary rate of base substitution through comparative studies of nucleotide sequences. J. Mol. Evol. 16, 111–120.
Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates (Erratum in Mol. Biol. Evol. 1995; 12:525). Mol. Biol. Evol. 11, 459–468.
Leigh, J.W., Susko, E., Baumgartner, M., Roger, A.J., 2008. Testing congruence in phylogenomic analysis. Syst. Biol. 57, 104–115.
Moore, B.R., Smith, S.A., Donoghue, M.J., 2006. Increasing data transparency and estimating phylogenetic uncertainty in supertrees: approaches using nonparametric bootstrapping. Syst. Biol. 55, 662–676.
Nishihara, H., Okada, N., Hasegawa, M., 2007. Rooting the Eutherian tree — the power and pitfalls of phylogenomics. Genome Biol. 8, R199.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47–67.
Philippe, H., Lartillot, N., Brinkmann, H., 2005. Multigene analyses of bilaterian animals corroborate the monophyly of Ecdysozoa, Lophotrochozoa, and Protostomia. Mol. Biol. Evol. 22, 1246–1253.
Pisani, D., Wilkinson, M., 2002. MRP, taxonomic congruence and total evidence. Syst. Biol. 51, 151–155.
Poe, S., Swofford, D.L., 1999. Taxon sampling revisited. Nature 398, 299–300.
Pollock, D.D., Zwickl, D.J., McGuire, J.A., Hillis, D.M., 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51, 664–671.
Pupko, T., et al., 2002. Combining multiple data sets in a likelihood analysis: which models are the best? Mol. Biol. Evol. 19, 2294–2307.
Ragan, M.A., 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1, 53–58.
Rannala, B., Yang, Z., 2008. Phylogenetic inference using whole genomes. Ann. Rev. Genom. Hum. Genet.
Rannala, B., Huelsenbeck, J.P., Yang, Z., Nielsen, R., 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47, 702–709.
Ren, F., Tanaka, H., Yang, Z., 2005. An empirical examination of the utility of codon-substitution models in phylogeny reconstruction. Syst. Biol. 54, 808–818.
Ronquist, F., Huelsenbeck, J.P., Britton, T., 2004. Bayesian supertree. In: Bininda-Emonds, O.R.P. (Ed.), Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life. Kluwer Academic, Dordrecht, the Netherlands, pp. 193–224.
Rosenberg, M.S., Kumar, S., 2003. Taxon sampling, bioinformatics, and phylogenomics. Syst. Biol. 52, 119–124.
Sanderson, M.J., 1998. Phylogenetic supertrees: assembling the trees of life. Trends Ecol. Evol. 13, 105–109.
Sanderson, M.J., et al., 2003. Obtaining maximal concatenated phylogenetic data sets from large sequence databases. Mol. Biol. Evol. 20, 1036–1042.
Semple, C., Steel, M., 2000. A supertree method for rooted trees. Discrete Appl. Math. 105, 147–158.
Shapiro, B., Rambaut, A., Drummond, A.J., 2006. Choosing appropriate substitution models for the phylogenetic analysis of protein-coding sequences. Mol. Biol. Evol. 23, 7–9.
Simon, C., et al., 2006. Incorporating molecular evolution into phylogenetic analysis, and a new compilation of conserved polymerase chain reaction primers for animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 37, 545–579.
Springer, M.S., de Jong, W.W., 2001. Phylogenetics. Which mammalian supertree to bark up? Science 291, 1709–1711.
Suchard, M.A., Kitchen, C.M., Sinsheimer, J.S., Weiss, R.E., 2003. Hierarchical phylogenetic models for analyzing multipartite sequence data. Syst. Biol. 52, 649–664.
Tateno, Y., Takezaki, N., Nei, M., 1994. Relative efficiencies of the maximum-likelihood, neighbor-joining, and maximum-parsimony methods when substitution rate varies with site. Mol. Biol. Evol. 11, 261–277.
Thorne, J.L., Kishino, H., Painter, I.S., 1998. Estimating the rate of evolution of the rate of molecular evolution. Mol. Biol. Evol. 15, 1647–1657.
Wiens, J.J., 2005. Can incomplete taxa rescue phylogenetic analyses from long-branch attraction? Syst. Biol. 54, 731–742.
Wilkinson, M., et al., 2005. The shape of supertrees to come: tree shape related properties of fourteen supertree methods. Syst. Biol. 54, 419–431.
Yang, Z., 1994. Maximum likelihood phylogenetic estimation from DNA sequences with variable rates over sites: approximate methods. J. Mol. Evol. 39, 306–314.
Yang, Z., 1995. Evaluation of several methods for estimating phylogenetic trees when substitution rates differ over nucleotide sites. J. Mol. Evol. 40, 689–697.
Yang, Z., 1996. Maximum-likelihood models for combined analyses of multiple sequence data. J. Mol. Evol. 42, 587–596.
Yang, Z., 1997. PAML: a program package for phylogenetic analysis by maximum likelihood. Comput. Appl. Biosci. 13, 555–556.
Yang, Z., 2004. A heuristic rate smoothing procedure for maximum likelihood estimation of species divergence times. Acta Zoologica Sinica 50, 645–656.
Yang, Z., 2006. Computational Molecular Evolution. Oxford University Press, Oxford, England.
Yang, Z., Rannala, B., 2006. Bayesian estimation of species divergence times under a molecular clock using multiple fossil calibrations with soft bounds. Mol. Biol. Evol. 23, 212–226.
Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588–598.