Syst. Biol. 59(1):59–73, 2010

DOI:10.1093/sysbio/syp077
Advance Access publication on November 10, 2009

New Heuristic Methods for Joint Species Delimitation and Species Tree Inference

BRIAN C. O'MEARA*

*Department of Ecology & Evolutionary Biology, University of Tennessee, 569 Dabney Hall, Knoxville, TN 37996-1610, USA; *Correspondence to be sent to: Department of Ecology & Evolutionary Biology, University of Tennessee, 569 Dabney Hall, Knoxville, TN 37996-1610, USA; E-mail:bomeara@utk.edu.*

Received 14 February 2008; reviews returned 3 June 2008; accepted 21 September 2009
Associate Editor: L. Lacey Knowles

Abstract.—Species delimitation and species tree inference are difficult problems in cases of recent divergence, especially when different loci have different histories. This paper quantifies the difficulty of jointly finding the division of samples to species and estimating a species tree without constraining the possible assignments a priori. It introduces a parametric and a nonparametric method, including new heuristic search strategies, to do this delimitation and tree inference using individual gene trees as input. The new methods were evaluated using thousands of simulations and 4 empirical data sets. These analyses suggest that the new methods, especially the nonparametric one, may provide useful insights for systematists working at the species level with molecular data. However, they still often return incorrect results. [Brownie; gene tree parsimony; gene tree species tree; speciation; species delimitation.]

Two of the main goals of systematics are dividing the diversity of life into species and discovering the phylogenetic relationships of these species. Both can be difficult to achieve. Processes such as lineage sorting, introgression, and undetected gene duplication may cause gene trees to disagree with the true tree of species, potentially obscuring the species tree signal (Fitch 1970; Goodman et al. 1979; Avise 1983; Tajima 1983; Pamilo and Nei 1988; Doyle 1992; Hudson 1992; Maddison 1997). For species delimitation, a systematist must choose both a species concept and a criterion to apply this species concept to data. Even if speciation itself is effectively instantaneous, the time required for sufficient evolutionary changes to appear to allow 2 distinct lineages to be recognized will not be (De Queiroz 2007). This causes delimitation of species to be difficult.

These two questions are biologically linked but rarely methodologically coupled. If intervals between speciation events were long enough that all species were monophyletic for all their genes, once the species were correctly delimited, any species could be adequately represented by a single individual on a phylogeny. In reality, putatively independently evolving lineages are often not monophyletic (Funk and Omland 2003). The phylogeny of species, unless they are defined under a strict genealogical species concept (GSC; Baum and Shaw 1995; Hudson and Coyne 2002), will have an assortment of independent evolutionary lineages, which will probably include paraphyly for at least some of their genes. Here, I attempt to unite these two questions as the more general one of jointly inferring the species boundaries and the species tree. I calculate the computational complexity of the problem, develop and implement methods for addressing it, and perform simulations and analyses across hundreds of parameter combinations to evaluate the feasibility. I also analyse 4 empirical data sets, *Drosophila* (Machado et al. 2002; Machado and Hey 2003), *Manacus* (Passeriformes) (Brumfield et al. 2008), *Lactarius* fungi (Nuytinck and Verbeken 2007), and *Melanoplus* grasshoppers (Carstens and Knowles 2007), to evaluate the performance of the new methods.

MATERIALS AND METHODS

Problem Definition

Given a set of sequences from multiple individuals, the general problem is to allocate those individuals into putative species and estimate the species tree. This solution, the species tree with assignment of samples to species, is termed the "delimited species tree." Optimally, a method will assign species and estimate the species tree correctly, in a statistically and computationally efficient manner. An estimate of the delimited species tree may differ from the true delimited species tree in topological error and/or through assignment of individuals to the wrong species. The latter might happen by merging 2 species that should be 1, splitting 1 true species into 2, having an individual of 1 species assigned to a different species, or a complex mixture of these. This is a more difficult problem than is typically addressed in DNA barcoding approaches (Hebert et al. 2003; Tautz et al. 2003), where 1 or more unknown individuals are assigned to existing species (Manel et al. 2005; Matz and Nielsen 2005; Abdo and Golding 2007; Zhang et al. 2008).

Most methods in systematics work on restrictions of this general problem, such as assuming that assignments to species are known (Nielsen and Wakeley 2001; Carstens and Knowles 2007; Edwards et al. 2007; Liu and Pearl 2007) or assuming that the gene tree matches the species tree (Hebert et al. 2003; Pons et al. 2006). I follow the approach of several recent authors (Pons et al. 2006; Carstens and Knowles 2007; Knowles and Carstens 2007; Mossel and Roch 2007; Kubatko et al. 2009), in restricting the problem using estimated gene trees as input rather than by integrating across a set of possible gene trees. The restricted problem still makes

no assumptions about species assignment, species-tree topology, or match between the gene trees and the species tree. Thus, the restricted problem addressed here is defined as follows: given a set of estimated gene trees from multiple individuals, those individuals must be split into putative species and the species tree must be estimated. For the purposes of this paper, "speciation" will be defined as occurring when gene flow permanently stops between 2 populations.

Computational Complexity of the Problem

The task of finding the delimited species tree with minimum total cost, given just a set of gene trees with leaves unassigned to species, is computationally daunting. First, finding the optimal species tree given a set of gene trees, with gene tree samples already assigned to the species, is NP complete (Ma et al. 1998; Fellows et al. 2003). Therefore, as with all NP-complete problems, such as finding the most parsimonious tree given a set of DNA sequences (Foulds and Graham 1982), there are no known fast (polynomial time) algorithms guaranteed to find the solution, but verifying a solution is relatively easy. Second, even the mapping of gene tree leaves to species tree leaves is unknown. Thus, while the number of possible bifurcating rooted topologies for k samples is $\frac{(2k-3)!}{2^{k-2}(k-2)!}$ (Cavalli-Sforza and Edwards 1967), the number of possible species topologies and assignments is far greater. Both the number and the composition of terminals can vary (there can be from 1 to k species for a given set of k samples), and, for each assignment of samples to species, there may be multiple possible species-tree topologies. The number of possible ways to subdivide n samples into k species is $S(n,k)$, where $S(n,k)$ is a Stirling number of the second kind (Abramowitz and Stegun 1972).

$$S(n,k) = \frac{1}{k!}\sum_{i=0}^{k}(-1)^{k-i}\frac{k!}{(i-k)\,!\,i!}(i)^n. \qquad (1)$$

Thus, for n samples being assigned to an unknown number of species, there are

$$1+\sum_{k=2}^{n}\left(\frac{1}{k!}\bullet\frac{(2k-3)!}{2^{k-2}(k-2)!}\sum_{i=0}^{k}\left((-1)^{k-i}\frac{k!}{(i-k)\,!\,i!}(i)^n\right)\right) \qquad (2)$$

possible rooted bifurcating species topologies ($n \geqslant 2$). For example, for 3 samples, there are 3 possible rooted bifurcating gene topologies, but 7 possible delimited species trees. For 7 samples, there are 10,395 gene topologies but 51,157 possible species topologies and assignments. For 30 samples, there are 4.95×10^{38} gene topologies but 2.95×10^{41} species topologies and assignments.

Proposed Methods

I develop 2 new methods of delimited species tree inference. These new approaches start with inferred gene trees rather than actual sequence data, like many others in the literature (Maddison and Knowles 2006; Pons et al. 2006; Carstens and Knowles 2007; Knowles and Carstens 2007; Kubatko et al. 2009) and therefore differ from the "BEST" approach (Edwards et al. 2007; Liu and Pearl 2007; Liu et al. 2008) or importance sampling or related approaches that integrate over the gene trees (Beerli and Felsenstein 1999, 2001; Yang 2002; Rannala and Yang 2003). Thus, the gene trees and delimited species trees are estimated sequentially rather than jointly and a single optimum is used for the gene tree. One approach follows naturally from the work of Knowles and Carstens (Carstens and Knowles, 2007; Knowles and Carstens 2007) and seeks the delimited species tree that maximizes the probabilities of the gene topologies. The second approach minimizes intraspecific structure, which might indicate an undersplit species, and gene tree–species tree conflict, which might indicate an oversplit species or incorrect species topology. Both approaches use novel heuristic search strategies and are implemented in the program Brownie (O'Meara et al. 2006)

Approach 1: KC delimitation.—The first approach, called "KC delimitation," is a basic extension of Knowles and Carstens (2007). It has roots in earlier literature (e.g., Maddison 1997) and seeks to find the delimited species tree that maximizes the probability of the gene trees. Knowles and Carstens (2007) used the program COAL (Degnan and Salter 2005) to calculate this probability and then used a likelihood-ratio test to compare 2 specified species topologies. This approach, and similar nonparametric approaches (Maddison and Knowles 2006), compares a small number of prespecified species topologies. For example, in the simulation done by Knowles and Carstens (2007), 2 possible delimited species trees are compared to decide whether a specified set of samples form 3 or 4 species. Had there been no knowledge about the placement or assignment of any of the 20 samples, 6.03×10^{23} possible delimited species trees would have had to be evaluated. The proposed extension takes their method of calculating gene tree probabilities for a given delimited species tree and comparing these likelihoods between species trees and embeds it in a heuristic search algorithm to find the optimal delimited species tree without needing to prespecify it. Details of the heuristic algorithm are below. I implemented the Akaike information criterion (Akaike 1973, 1974) and corrected Akaike information criterion (Sugiura 1978) approaches for model selection. Likelihood-ratio tests are not appropriate because most pairs of delimited species trees are not nested models.

Two ways to calculate the probability of a gene tree topology (no branch lengths) given a species tree (with branch lengths) are available. The first is to explicitly calculate this probability using equations from Degnan and Salter (2005) for the multiple species case and equation 5.3 of Harding (1971) in the single species case. Knowles and Carstens (2007) use COAL (Degnan and

Salter 2005) to calculate these probabilities for specified species trees. STEM (Kubatko et al. 2009) could be used instead of COAL for calculating this probability if gene trees with branch lengths, rather than just topologies, were the intended input. I implemented this approach using COAL, but some technical impediments made the implementation currently impractical to use across hundreds of analyses. Though analytical expressions exist for calculating the probabilities of gene topologies given species trees matching certain conditions (Degnan and Salter 2005), these solutions may not exist for more complex but realistic scenarios, like slowly diminishing gene flow, geographic structure, or occasional introgression after speciation. However, although developing analytical expressions may be difficult, simulating gene trees under even complex speciation scenarios, such as geographic structure and ongoing gene flow, is quite feasible (Hudson 2002). The proportion of times a gene tree is recovered in simulations using a particular species tree can be taken as an estimate of its probability given the species model. This approach is closely related to approximate Bayesian computation approaches (Tavare et al. 1997; Beaumont et al. 2002), but simulations are used just to approximate likelihoods (probability of the data, in this case gene topologies, given the hypothesis) rather than posterior distributions. Similar approaches have been applied in phylogenetics by Weir and Schluter (2007) and Ree et al. (2005). The glaring disadvantage of this method is the potential amount of time required, as millions of simulations might need to be performed to estimate low probabilities precisely. This number will vary with different problem sizes and even across different gene trees. The new approach uses the program ms (Hudson 2002) to simulate gene trees under each examined species tree, which necessarily includes branch lengths.

Under the simple models used in this paper, samples within a species are interchangeable—labels can be swapped between any 2 samples from a species and the probabilities of the 2 gene trees can be the same—and the approach takes advantage of this fact to reduce the number of simulations needed. The estimation strategy has the risk of resulting in an infinite negative log likelihood if there are no simulations of a given gene topology. There are various ways of dealing with this, such as increasing the number of simulations, and the approach adopted here was to substitute a large negative log likelihood for the actual negative log likelihood. In practice, this did not appear to happen frequently under the simulation conditions and analysis settings (100,000 simulated gene trees per gene per species tree).

Approach 2: nonparametric delimitation.—The above approach uses explicit models and may take quite some time to complete a search. I also developed a nonparametric approach based on 2 ideas. First, except in some cases involving short internal edges (Degnan and Rosenberg 2006), the most probable gene tree should match the species topology. As a result, there is a tendency for gene trees to be somewhat congruent with each other for interspecific branches. Long species tree branches with small population sizes will result in gene trees better matching the species tree (Maddison 1997) and therefore gene trees will tend to agree more with each other in these regions. Incomplete lineage sorting (lack of coalescence of intraspecific sequences between speciation events) will tend to weaken this signal, as will errors in inferring the gene tree. In contrast, within species, gene trees should show no such structure. In a panmictic population without selection, migration, or linkage, each gene tree is a random draw from the neutral coalescent tree distribution. Assuming no selection, while unrealistic, simplifies the development of a method and is commonly done in population genetics. Population structure tends to make these trees more similar than expected under neutral coalescence. If one envisions a consensus tree of the gene trees, bipartitions on this tree where many gene trees agree on topology will likely be interspecific branches, whereas branches where many gene trees disagree on topology will likely be within species. The nonparametric method developed here attempts to recover the species assignment and species tree, which, together, are the delimited species tree, that minimizes gene tree conflict on the interspecific portions of the tree while minimizing excess structure within each species. To do this, "gene tree conflict" and "excess structure" must be quantified, and then these 2 measures combined in some manner (see Fig. 1 and below).

A common way to calculate the gene tree conflict with a given species tree is the number often referred to as the minimum number of gene duplication events (Goodman et al. 1979), which is used in gene tree parsimony (Slowinski et al. 1997). For example, imagine genes evolving up the rooted species tree (A,(B,C)). In the branch leading up to the common ancestor of A, B, and C, let many different alleles evolve from one common ancestor and assume that just alleles descended from alleles x and y still exist at the time of the divergence of A from (B, C). Both may persist in the 2 descendent lineages for some time, but then A may become fixed for x. If B and C both inherit only x-type alleles, the gene history will match the species tree. If B and C both inherit only y-type alleles, the gene tree topology will match the species tree (B and C still a clade), but the divergence times will reflect the split of alleles x and y, which predates the $A|(BC)$ split. If one of B or C becomes fixed for x, and the other for y, then the gene tree will show the species that has the x copy as sister to A, which conflicts with the species tree. Without introgression or incorrect estimation of the gene tree or the species tree, this result of gene tree–species tree conflict can only occur if there were 2 segregating alleles (x and y) that originated before a species split and persisted until the next species split. The minimum number of such origins of alleles (the number of x–y splits) where both copies persisted long enough to explain gene tree–species tree incongruence is the minimum number of gene duplication events.

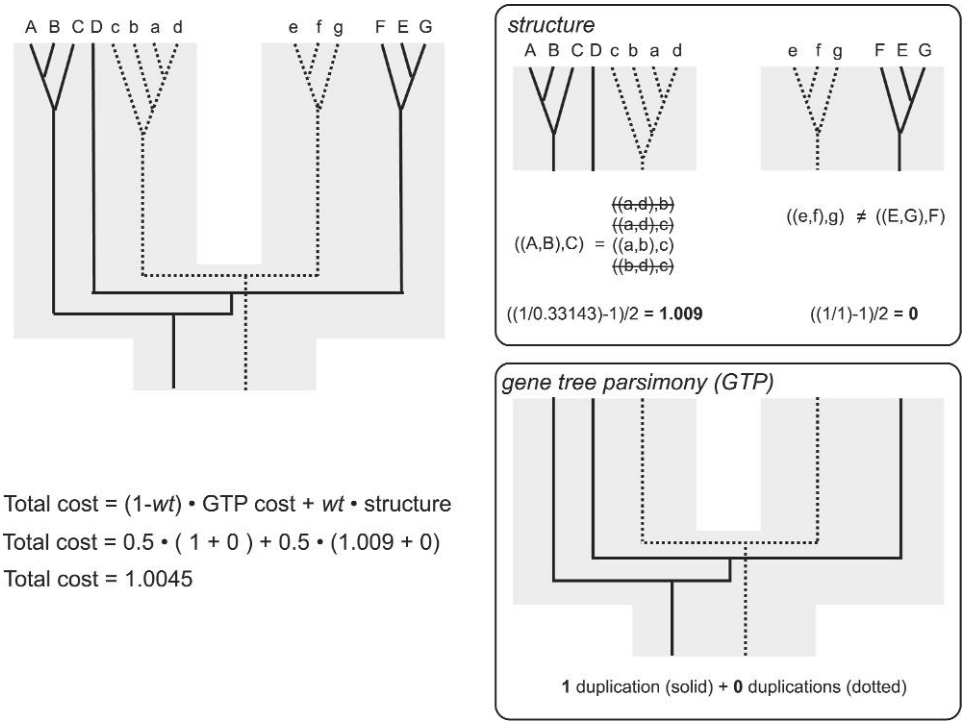

FIGURE 1. Nonparametric approach description. On the left are 2 gene trees (solid and dotted lines) embedded within a 2 species tree. The cost function (lower left) has 2 components, the structure cost and the gene tree parsimony cost. The upper right panel shows the calculation of the structure cost. Within each of the 2 species, the gene trees are decomposed into triplets (rooted 3 taxon statements). Note that in cases where a gene tree does not form a single clade within a species (as in the solid tree in the first species), it is split into separate subtrees. For each pair of loci, the number of their identical triplets is calculated and compared with a distribution under a panmictic population to find the probability of getting that many overlaps or more within a single panmictic species (0.33143 and 1.0 for the left and right species, respectively). This is then translated into a cost. The between-species portion of the cost comes from the number of duplication events required to reconcile the gene and species trees.

The related measure of minimum number of deep coalescences (Maddison 1997) is similar; however, it counts the number of branches over which both copies must have persisted rather than the number of times both copies are created and persist over at least 1 branch. Minimizing the number of deep coalescences is equivalent to having a cost of 0 for introducing a gene duplication and a cost of 1 for each time the duplicate copies persist in subsequent lineages. The idea still holds with more than 2 copies persisting in lineages but would require a longer explanation. A more general model would allow the creation and persistence costs to be any nonnegative number rather than just 0 or 1. In an ideal such model, as in parametric models, the effect of time as well as effective population sizes within lineages would be incorporated. I used the number of duplications (persistence cost of 0) as the gene tree conflict cost, as there are fast algorithms for calculating it. However, having a persistence cost of 1 (deep coalescences) might be more realistic, as maintaining 2 copies across more speciation events, which probably entails more time and more bottlenecks, should be less likely and therefore costlier than maintaining 2 copies on just 1 branch. Future work could explore the more general model of having any nonnegative persistence cost as well as the effect of using deep coalescences as a measure, though it has been argued that the deep coalescence cost and duplication costs are highly correlated (Zhang 2000; Cotton and Page 2003). The algorithm used to calculate the minimum number of duplications comes from Sanderson's modification in the program gtp (Sanderson and McMahon 2007) of the Zmasek and Eddy (2001) algorithm. This algorithm requires bifurcating gene trees. Only lineage sorting events occurring on interspecific branches of the species tree are counted.

Calculating the penalty for excess structure is more difficult. Unlike the gene tree conflict case, where the ideal number of disagreements is 0, in the case of structure, there will be some agreement between gene trees just based on chance even in the case of a panmictic population of very large size. One way of characterizing structure is the number of triplets (rooted 3-taxon statements) in common between 2 gene trees. Too many triplets in common between pairs of trees would represent too much structure. For each possible number of samples per species, up to 50 samples per species, 100,000 simulations were performed under a neutral coalescent to estimate the distribution of triplet overlap between pairs of gene trees assuming linkage equilibrium. For more than 50 samples per species, approximations of triplet distance (Critchlow et al. 1996) are used. The proportion of simulated pairs of trees with equal

or greater overlap in the number of triplets as the given pair of gene trees is treated as a P value (the more overlap in a given pair, the lower the P value). This excess structure cost is calculated within each species. In the case of a gene tree for which a species is paraphyletic, each subtree of the gene tree completely enclosed within a species is compared with (sub)trees from other genes. The genealogical sorting index (Cummings et al. 2008) is an additional way to assess intraspecific structure but is not examined or implemented here.

Gene conflict cost is in units of number of lineage sorting events. Excess structure cost is in units of probability of at least that the observed triplet overlap for pairs of gene trees for each species. It is impossible to convert number of lineage sorting events to a probability without making many assumptions or inferences about speciation times and ancestral population sizes. Instead, the structure cost is converted to a number that grows larger with more excess structure. For a given delimited species tree, the structure cost is, summed over all pairs of genes for all species, the reciprocal of the probability of at least as much structure as is observed under the null model, -1. The reciprocal is taken so that more structure (lower probability) results in a higher cost. One is subtracted from this so that the total cost has a minimum of 0, as the reciprocal of the probability is 1 or greater. Gene conflict is calculated as 1 gene at a time and so increases linearly with the number of genes, whereas structure is calculated taking all possible pairs of genes so increases with the square of the number of genes. To make the 2 costs scale in the same way with the number of genes, the structure cost is then divided by the number of genes. Thus, the structure cost, where the number of genes is g and the number of species is n, is $\sum_{\text{Gene } A=1}^{g} \sum_{\text{Gene } B=\text{Gene } A+1}^{g} \sum_{\text{Species}=1}^{n} \frac{1}{g} \left(\frac{1}{P-\text{value}} - 1 \right)$.

Thus, both gene conflict cost and structure cost range in value from 0 to positive infinity (though there is an upper limit for each based on data set size). The total cost is just a sum of these 2 costs. Being able to weight these 2 costs might be desirable. If the relative weight for structure cost is too low, "lumping" samples into few species will be favoured, as that will minimize the gene tree conflict cost, as all conflicts, as well as some congruence, will be within species. If the relative weight for structure cost is too high, the method will be biased toward "splitting" species, minimizing the costs of structure within species. Allowing the weight to be adjusted might be useful in tuning the method to maximize its chance of accurately delimiting species as well as allow sensitivity analyses. As in other nonparametric approaches, such as gap extension and gap creation costs for alignment or relative weights of different codon positions in a parsimony analysis, determining this weight is often an arbitrary but sometimes important decision. The combined score for a delimited species tree is

$$\text{Total score} = (1 - \text{weight}) \times \text{gene tree parsimony score} + \text{weight} \times \text{structurecost}.$$

Figure 1 shows how a score is calculated using the nonparametric method. One deficiency of the method is that any overlap of triplets within a species is assigned a cost. To investigate the influence of this, I added an additional parameter, the P-threshold, to the cost function to only count structure costs corresponding to P values below a fixed threshold.

Search Strategy

Given the complexity of the problem, I developed a heuristic approach to finding the delimited species tree. The first step in the heuristic search is finding associations of samples. If 2 samples always form a clade on all input gene trees, for example, it is useful to start most searches with those samples assigned to the same species. The distance between pairs of samples on the input trees is measured as the proportion of 3 taxon trees in which the 2 samples form a clade. This is directly related to the measure for within-species structure for the nonparametric method. The matrix of pairwise distances is then analysed using neighbour joining to create a guide tree. Note that this guide tree is not a constraint tree but is more a guide tree in the sense used in progressive alignment. In progressive alignment (Feng and Doolittle 1987), as used, for example, in the popular program Clustal (Higgins et al. 1992; Thompson et al. 1994), a guide tree is used to determine the order in which sequences are aligned, though the alignment procedure itself does not take phylogeny into account. In the search strategy developed in this paper, all samples occurring together in clades on the guide tree are stored as sets of samples to attempt to move as a group (see below). Longer edges on the guide tree correspond to clades containing taxa found more frequently together. For each search replicate, the guide tree is subdivided into 1 or more subtrees by deleting edges, with a bias for deleting long edges. The samples in each subtree are assigned to 1 species, then a random tree joining these species is used as a starting delimited species tree. This approach of using triplets in the gene trees to create a guide tree with neighbour joining does not require that all the taxa be present in all the gene trees. It should thus work where other methods, such as majority rule consensus trees, might fail.

Once the starting delimited species tree is constructed, it is then transformed using 5 or 6 different moves, depending on the optimization criterion. The delimited species tree's topology can be transformed through 1) subtree pruning and regrafting (SPR) (Swofford 1990) or through 2) rerooting on random branches. 3) Two terminal sister species on the tree (a "cherry" sensu McKenzie and Steel 2000) can be merged into 1 species. 4) One terminal species can be split into 2 species. The assignment of samples to these 2 species is nontrivial. If the initial species has 25 samples in it, there are $S(25,2)$, or 16,777,215, different possible assignments of samples to the 2 resulting species. In practice, the program examines a subset of the possible assignments (the relative

size of this subset may be chosen), and the best taken as the cost of the tree as the result of move. 5) Samples may be moved from one terminal species to another. Moving a single sample at a time is insufficient. For example, if 2 samples form a clade on all gene trees, but that clade is placed in the wrong species in the species tree, it would be difficult to fix that through a reassignment. The samples would have to be moved one at a time, breaking up the consistent clade. Moving groups of samples at random would also be inefficient. Instead, the set of samples in clades on the guide tree as well as the individual samples are all attempted to be moved. 6) In the case of the simulation approach, branch lengths on the species tree must also be optimized. This is done through 1 of 2 kinds of moves, either a stretching or shrinking of the whole tree or a movement of 1 internal node up or down.

The program proposes moves until no improvement in the score of the current delimited species tree can be found through any of the nonbranch length moves. As there are an infinite number of possible branch length changes, they are not used as a stopping criterion. A new starting tree is then created. At each step, the current best tree and assignment are saved in a file. Program options can be set to prevent certain kinds of moves with all methods and the relative weight of the structure and gene conflict costs in the nonparametric method may also be set (by default set at 0.5, the value of this parameter may strongly affect the results; see below). The program can become a program for finding the best species tree under gene tree parsimony (Slowinski et al. 1997) by fixing the assignment of samples to species, using the nonparametric criterion, and limiting moves to SPR and rerooting.

Simulations

To test the methods' effectiveness, I performed a variety of simulations of gene trees within species trees using Hudson's ms program (Hudson 2002). Besides the complexity introduced through gene history–species history mismatch, estimation of a given gene history using sequence data is also a difficult practical problem. To include this aspect of the problem difficulty, I evolved gene sequences along each gene tree returned from ms using a complex model with parameter estimates derived from an empirical data set (Carstens and Knowles 2007). I then analysed the sequences using a simpler model to return estimates of the gene trees using approaches appropriate to each tested method.

I used 3 different models for simulations (Fig. 2). The first model simulated 1 species being divided into 2 species at a particular time. All 3 lineages have the same effective population size. Parameters varied were depth of the split, number of bases simulated per gene (0.25, 1, and 16 times the 648 bp typically used in DNA barcoding), number of samples from each of the 2 species, and number of loci for 135 unique combinations of parameters. In the second model, the effect of population substructure was evaluated. Nine populations arranged in a line were simulated, with 5 individuals sampled from each of populations 2, 4, 6, and 8. Each population exchanged migrants symmetrically with its neighbour(s) until a specified time point, when gene flow between populations 4 and 5 and between populations 5 and 6 was stopped, resulting in a speciation event. As the fifth population was not sampled, this is functionally equivalent to extinction of a connecting population in the middle of a species' range, resulting in 2 allopatric species. The third model used a 4 species pectinate species tree. For some values of branch lengths, this tree falls in the "anomaly zone" of Degnan and Rosenberg (2006), where the most frequent gene tree for 4 samples differs from the species tree topology. Lengths of the first internode, second internode, and shortest terminal branches were varied to include species trees both inside and outside the anomaly zone for 180 unique combinations of parameters. Except where noted, in all these simulations, 648 bases were simulated per locus, 5 loci were simulated, 5 individuals were sampled per species, and 20 replicates were performed per each unique combination of parameters. In all models, speciation was defined to occur when gene flow between a set of populations dropped to zero. For details of the simulations, see supplementary Appendix 1 (available from http://www.sysbio.oxfordjournals.org/).

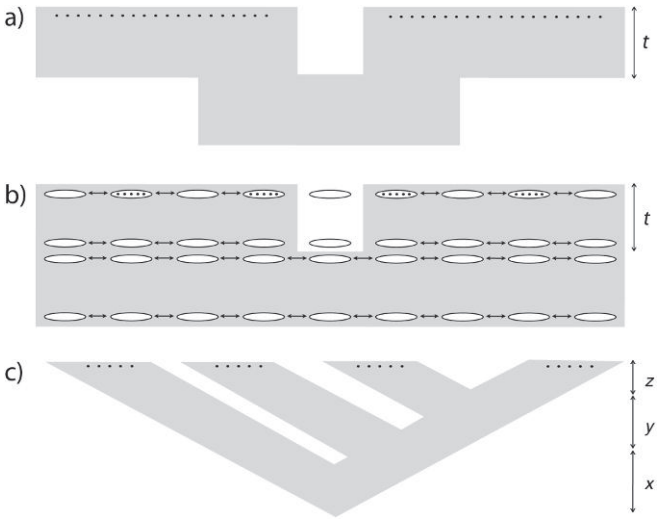

FIGURE 2. Simulation models. Three simulation models were used. a) shows a single species being split into 2. The time of the split was varied across simulations, as were 3 other parameters. b) shows the model used for population subdivision. Nine populations were simulated, with flow between neighbouring populations. At a time *t* in the past, gene flow was ceased between the middle (fifth) population and its neighbours. Samples are taken from the second, fourth, sixth, and eighth populations. c) shows a 4 species model using an ultrametric pectinate tree. Length of the 2 internodes and shortest terminal branches on the tree are varied. Under some parameter combinations, the most probable gene tree conflicts with the species tree (Degnan and Rosenberg 2006).

The KC delimitation approach using simulation to estimate gene topology probabilities was used with default parameter settings in Brownie. The nonparametric approach used a range of parameter values for structure weight (0.1, 0.5 [default], and 0.9) and *P*-threshold (0.05, 0.5, and 1.0 [default]). All analyses were performed on the Duke Shared Computing Resource on 64-bit processing machines. Each simulation replicate, including all analyses, was performed on 1 node so that the speed for different analytical methods could be compared on the same hardware. As different replicates were on different nodes with potentially different speeds, Yang's "small" program, part of Speedtest v. 2 (Yang 2005), was run with each simulation to provide a standard measure of processor speed.

Comparisons

A new method should generally perform as well or better than existing methods to be worth adopting. Though the problem described here, that of identifying the species assignments and species tree for a set of samples present on multiple gene trees, is relatively novel, there are several existing tree-based species delimitation approaches that can be used for this. I chose 3 approaches to use for comparisons. Dettman et al. (2003) advocated a complex procedure involving a parsimony search on each locus and then estimation of bootstrap support using parsimony and posterior probabilities for clades for each locus, then identifying clades with strong enough bootstrap plus posterior support plus presence in at least 75% of gene trees as "independent evolutionary lineages," followed by a collapse of some of these lineages to form species. This approach was used with 1 empirical example and no simulations in the original paper. It was implemented here as a series of batch files for PAUP generated by Perl scripts followed by the collapsing procedure, which was implemented in Brownie (O'Meara et al. 2006). To reduce the time required for this approach, the Bayesian searches were omitted from the implementation. The GSC operationally divides taxa into species based on a consensus of the gene trees (Baum and Shaw 1995). Although the species concept used by the new methods focuses on groups no longer exchanging genes rather than basal, exclusive groups, the 2 concepts do relate (De Queiroz 2007), and the consensus tree approach used by the GSC utilizes input topologies in a way grossly similar to that done by the nonparametric delimitation approach. Clades in majority rule consensus trees present in 100%, 95%, 70%, or 50% of the input trees were defined as species under the GSC approach used here. Finally, the classic DNA barcoding approach (Hebert et al. 2003) calculates Kimura's 2-parameter distance between a set of samples for a single region of cytochrome oxidase I and uses a cutoff to divide samples into different species. This approach was implemented here using Perl scripts and PAUP with distance thresholds of 1%, 3%, and 10% (3% was used by Hebert et al. 2003). When analysing simulated data, the first tree simulated was used as the input for barcoding.

Empirical Data Sets

Empirical data sets may exhibit problems and complexity not present in the simulated data sets such as uncertainty in reconstructed gene trees, introgression between species, changing population sizes, and so forth. The methods developed here require gene trees from multiple unlinked loci, each with multiple samples per putative species. I examined 4 different data sets, from flies, birds, fungi, and grasshoppers. The fly data set comes from the work of Machado and colleagues (Machado et al. 2002; Machado and Hey 2003). This consists of sequences from *Drosophila pseudoobscura pseudoobscura*, *Drosophila persimilis*, *Drosophila miranda*, and *Drosophila pseudoobscura bogotana* (the 2 *pseudoobscura* subspecies are treated as 2 separate species in the genetics literature, though the taxonomy has not been updated to reflect this view) for several genes. *Drosophila miranda* is an outgroup to the other 3 taxa and is used to root the gene trees, then excluded from analysis. The 2 *D. pseudoobscura* taxa are estimated to have last shared a common ancestor with *D. persimilis* approximately 550,000 years ago and with each other only 230,000 years ago (Wang et al. 1997). *Drosophila pseudoobscura pseudoobscura* and *D. persimilis* form female (but not male) hybrids in nature and hybrids are known to be fertile, so gene flow may occur between these 2 species. *Drosophila pseudoobscura bogotana* does not overlap in range with the other species in nature, so ongoing hybridization is not possible (summarized in Machado and Hey 2003). The data set used consists of 10 loci, with samples pruned to include only individuals sequenced for all loci.

The bird data set comes from Brumfield et al. (2008) and consists of 5 nuclear loci for 4 *Manacus* named species, with 1 of these 4 split by the authors into 1 species occurring on the western side of the Andes and another on the eastern side, plus outgroups, which are used to root the gene trees and then excluded. There is evidence of gene flow between at least 2 of those species (Brumfield et al. 2008). The fungus data set comes from Nuytinck and Verbeken (2007) and consists of 1 nuclear and 1 mitochondrial gene. Both genes were treated identically, without accounting for differences between nuclear and mitochondrial inheritance patterns. The original study had 9 species, but I pruned the examined data set to those 5 species (*Lactarius semisanguifluus*, *Lactarius salmonicolor*, *Lactarius fennoscandicus*, *Lactarius deterrimus*, and *Lactarius deliciosus*) with at least 3 samples per species, a requirement of the nonparametric method. An additional 2 species were used as outgroups to root the gene trees and then excluded. The grasshopper data set comes from Carstens and Knowles (2007) and consists of 5 nuclear and 1 mitochondrial loci for 5 species of *Melanoplus* grasshoppers, with 4 samples per species. All 4 empirical data sets represent

recent taxa with difficult taxonomy and have multiple loci with multiple samples per putative species. For each data set, a "standard" tree was constructed based on the tree presented as the best depiction of species assignment and phylogeny in each source paper.

For each locus, I estimated the parameters of an HKY+gamma likelihood model on a UPGMA topology using PAUP* 4b10 (Swofford 2003). I then performed a likelihood tree search for each locus using the parameter values estimated previously, rooting using outgroups if available and with midpoint rooting otherwise. These trees were passed as input to the inference procedures in Brownie or to the other analysis tools used in the simulations.

Results

In total, the simulations and analyses in this paper required over 30 computer-years on the Duke Shared Cluster Resource. Use was limited to 64-bit Linux machines, generally with a CPU speed of 3 GHz or higher, and I compiled all programs but PAUP to take advantage of the 64-bit architecture. Table 1 shows the number of hours required for different methods for simulation model 1 with 648 bp simulations. The nonparametric method and the approach of Dettman et al. (2003) all took a small amount of time over the simulation conditions plotted, as majority rule trees and traditional barcoding approaches are almost instantaneous. The KC delimitation approach took substantially more time, starting with an average of 23 min for small data sets of 3 loci with 6 samples total and taking nearly 4 d to complete with 10 loci and 20 samples total. This may reflect both the higher number of moves required because branch lengths of the species tree must be optimized and the amount of time to calculate the score for each proposed species tree. Interestingly, the simulation delimitation approach took much longer in simulations with 162 bp of sequence data per gene, suggesting that data sets with stronger signal finish sooner.

Figure 3 shows the result of the simulation involving a single species splitting into 2 at a given time in the past (result is averaged over other parameters). Of the 2 new methods, the nonparametric version outperformed the KC delimitation version. This may be due to inefficiencies or other problems with the heuristic search. Figure S1 shows that, for a given depth and number of loci, the KC delimitation approach performed worse with many rather than few samples per species. Generally, methods appeared to perform better with deeper splits in the species tree, which would allow more time for coalescence within each species. The decrease in performance at shallow splits was less pronounced for the nonparametric method than others. The nonparametric method showed some sensitivity to penalty costs.

TABLE 1. Time (hours) required for various methods

Samples per species	Number of genes	KC delimitation	Nonparametric delimitation	Dettman et al.
3	3	0.39 ± 0.19	0.00 ± 0.00	0.00 ± 0.00
3	5	0.59 ± 0.28	0.00 ± 0.00	0.00 ± 0.00
3	10	1.19 ± 0.49	0.00 ± 0.00	0.00 ± 0.00
5	3	10.29 ± 3.67	0.00 ± 0.00	0.00 ± 0.00
5	5	16.43 ± 41.74	0.00 ± 0.00	0.00 ± 0.00
5	10	31.68 ± 42.27	0.00 ± 0.00	0.00 ± 0.00
10	5	38.19 ± 45.27	0.01 ± 0.00	0.01 ± 0.01
10	10	91.87 ± 37.73	0.02 ± 0.01	0.02 ± 0.01

Notes: Time required for the simulation delimitation, nonparametric delimitation, and the Dettman et al. (2003) methods are shown. Since the computers used varied in speed, times were calculated relative to the time required to complete a standard test (Yang 2005). Absolute times were calculated by multiplying the relative times by the average time required to complete the speed test.

One unexpected result was the apparent insensitivity of the results to length of simulated sequences (see Fig. S1). Though sequence length per gene varied 64-fold, from 162 to 10,368 bp of sequence, within a set of simulation parameters, most results did not vary based on sequence length, with the exception of the KC delimitation approach, which did perform better with longer sequences. One possibility is that even the shortest sequence length was long enough to recover the generating tree. To test this, I used PAUP to calculate Robinson–Foulds tree distance (Robinson and Foulds 1981) between each pair of simulated and estimated gene trees under one set of population conditions (depth of split $= 16N_e$, number of samples per species $= 5$) and compared these with the distance between random trees. The maximum distance is 14 and minimum is 0. Random trees had an average distance of 13.6, trees from simulations with 162 bp had an average distance between true and estimated tree of 7.3, trees from simulations with 648 bp had an average distance of 5.7, and trees from simulations with 10,368 bp had an average distance of 1.4. Thus, under these simulation conditions, more base pairs did improve gene tree inference, and even the trees inferred using more than 10,000 sites per gene were often not inferred completely correctly.

Figure 4 shows the performance of methods with population subdivision and a species split. Only the results from the deepest split are shown in Figure 4. Results from all depths are shown in Figure S2. The KC delimitation method was very subject to oversplitting the species. The nonparametric methods tended to incorrectly create 4 species at low levels of gene flow and sometimes got the correct answer at higher levels of gene flow. At the deep species split shown in Figure 4, the method of Dettman et al. (2003) and simple majority rule trees dramatically outperformed the new methods. At shallower splits, methods generally performed worse, though the nonparametric approach declined less at intermediate depth.

Figure S3 shows the performance of methods inside and outside the anomaly zone (Degnan and Rosenberg 2006). The new approaches, especially with recent divergences, sometimes subdivided samples correctly into 4 species but did not return the correct arrangement of these species on a tree, and overall they performed poorly.

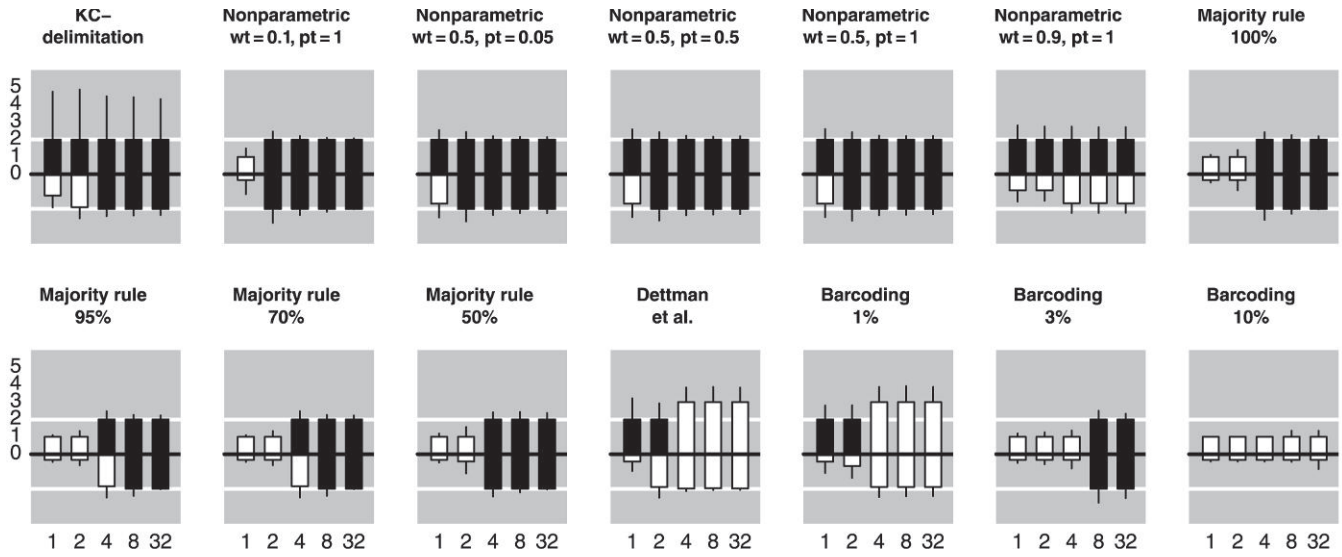

FIGURE 3. Number of species and accuracy for single split model. Each panel represents the results from one method with particular settings. Within each panel, the *x*-axis represents the depth of the species split (in units of N_e). The height of the bar above the black line shows the median number of recovered species across all examined combinations of number of loci, number of samples per species, and numbers of bases per gene. Black means the median number of species was correct (2 species). The depth of the bar below the black line shows the median accuracy (100% accuracy [black bars] means that all samples were correctly assigned to species and the species tree was correct). Vertical lines on each bar show 1 standard deviation. Figure S1 shows the same information without aggregation across variables.

Results from the empirical data sets were also mixed. With the fly data set (Machado et al. 2002; Machado and Hey 2003), the KC delimitation approach wildly oversplit the samples (see Fig. 5). The nonparametric approach with a structure weight of 0.5 and any of the 3 examined structure cutoffs resulted in a tree with 3 species, with every sample assigned correctly except 1 *D. pseudoobscura bogotana* species incorrectly placed with the *D. persimilis* samples, but returned the wrong species topology. The nonparametric approaches with other structure weights performed less well. The bird data set (Brumfield et al. 2008) proved difficult (Fig. S4). Most of the nonparametric analyses returned just 1 species (in contrast to their tendency in the simulated data to split or oversplit). Occasionally a second, incorrect, small species was created. The KC delimitation method returned 2 species (most samples in 1 species), neither matching an actual population or species. The fungus data set (Nuytinck and Verbeken 2007), consisting of just 2 genes, was the most successfully analysed

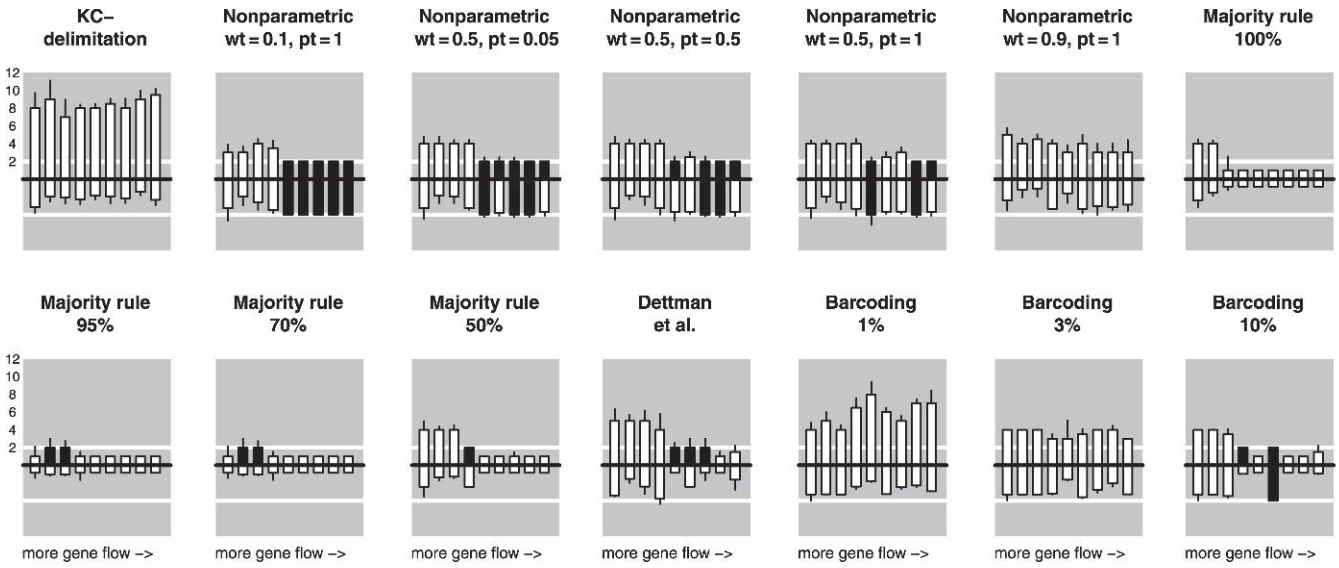

FIGURE 4. Number of species and accuracy for model with substructure. Figure design as for Figure 3, but with columns corresponding to different pairwise flow rates . The plot represents a split depth of 32 N_e. With shallower depths (1 and 4 N_e), all methods performed worse, but the reduction in performance for the new methods was least severe (see Fig. S2).

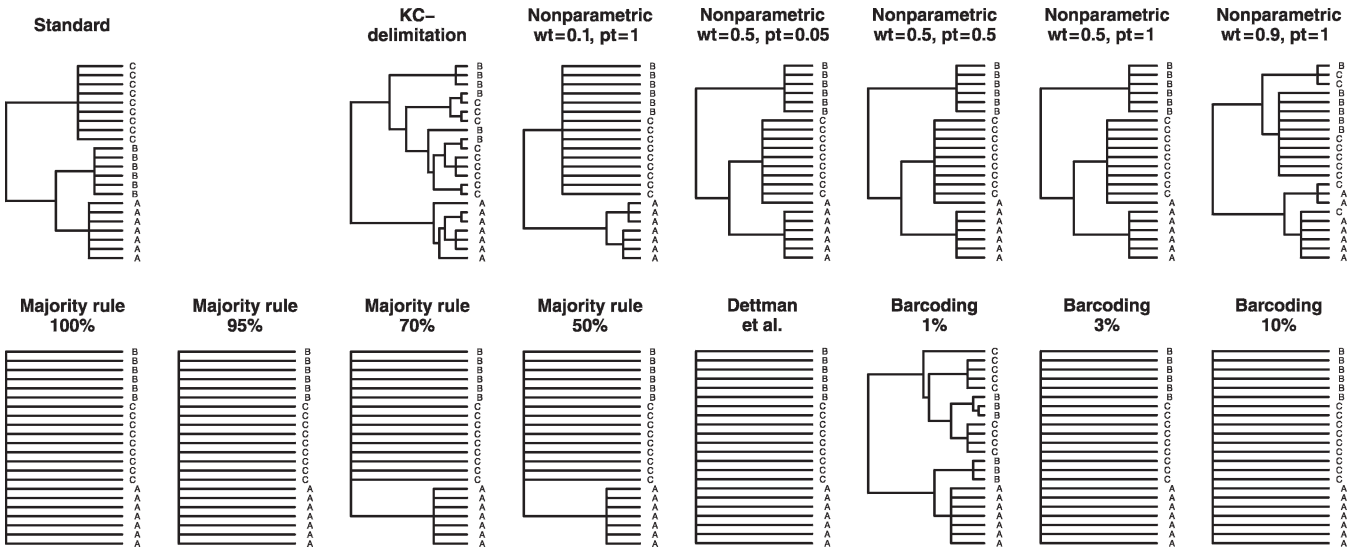

FIGURE 5. Results from fly data set. Delimited species trees are represented by trees of samples, where samples in a polytomy represent samples assigned to a single species. The "true" delimited species tree is shown in the top left, followed by trees from each of the methods examined in the paper. A = *Drosophila pseudoobscura bogotana*, B = *Drosophila pseudoobscura pseudoobscura*, and C = *Drosophila persimilis*.

(Fig. S5). The KC delimitation and nonparametric approaches returned the same result, which matched the standard tree in topology and assignment except for the inclusion of 1 *L. fennoscandicus* sample with *L. deterrimus* samples (as was found in the original study). The grasshopper data set (Carstens and Knowles 2007) proved difficult (Fig. S6). The KC delimitation approach oversplit the species. The nonparametric approach with a low structure weight returned a single species. The rest of the nonparametric analyses returned 2 species, but in no case did any of these species represent a clade or a single species on the standard tree.

DISCUSSION

Difficulties for Any Method

The problem of reconstructing species boundaries as well as species tree is demanding. As shown above, the computational difficulty of this problem is far worse than the problem of reconstructing a species tree when the species assignments are fixed, as in much of traditional phylogenetics. Moreover, shallow divergences can be remarkably hard to reconstruct. DNA substitutions, even if not evolving under a clock, still need to accumulate over some time period in order to be able to reconstruct gene tree topologies. With recent speciation events, or short intervals between speciation events, there may just not be enough changes on many genes to reconstruct their topologies accurately. For example, as shown above, with reasonable models of evolution, 5 samples per species, and a species split of depth 16 N_e, even 10,000 bp of nonrecombining sequence are not enough to guarantee recovering the correct gene tree. Processes such as selection, population subdivision, introgression, and recombination may further complicate the picture.

Developing empirical data sets suitable for this problem may be difficult as well. For many loci for recent splits, there will simply be no information on the gene tree topologies. Longer sequences can be used in an attempt to get more informative characters, but recombination within a gene region can obscure the signal. For example, if one-half of a gene has undergone a lineage sorting event rendering its history different from that of the species tree, whereas the other half has simply followed the species tree, a nonreticulate reconstruction of the entire gene's evolution will be incorrect. Uniparentally inherited regions with limited effective recombination, such as mitochondria, provide only 1 locus for use in these methods, so fast nuclear markers must be used as well. This is possible and becoming easier with the continual sequencing of new genomes but is still an obstacle. The fact that the methods can sometimes work with as few as 2 gene trees (as in the fungus data set) makes the use of empirical data sets more feasible. As the approaches developed here require gene trees with moderate resolution, they cannot easily take advantage of innovative methods for getting many independent sites for species delimitation, such as that of Shaffer and Thomson (2007). The nonparametric method, due to the details of the algorithm for calculating the gene tree parsimony score, requires fully resolved trees, but this could be changed with new algorithms for calculating the score.

Progress?

Given the intrinsic theoretical and empirical difficulties of the problem, any success would be surprising. Simulations employed realistic rates of sequence evolution, depths, and length of gene sequences, and the empirical data sets were at levels for which correct alpha

taxonomy was an interesting question and therefore not trivial to answer. Nonetheless, the new methods in this paper, as well as existing methods in the literature for slightly different problems, on occasion returned correct results, suggesting that there is some hope for developing approaches that will be empirically useful. In the analyses here, the new approaches performed well in simulated cases of simple splits of 1 panmictic species into 2, even when this occurred recently. There was also qualified success with the empirical data sets of flies and fungi. The new methods performed less well with population subdivision (though still performed correctly under some conditions) and downright poorly in other analyses (4 species pectinate tree simulations, bird and grasshopper data sets). The new approaches often performed better than existing approaches such as DNA barcoding or majority rule trees, suggesting that they may represent an improvement on existing methods. Note, however, that, in many cases, the method of Dettman et al. (2003) as implemented here performed nearly as well, if not better, and it was certainly faster than the KC delimitation approach. Performing the nearly 7000 replicates of simulations required extensive computing power, but individual searches under the nonparametric model are feasible to run on desktop machines (all the empirical data sets, e.g., take under 1 h to run in Brownie). The KC delimitation approach can take considerably more time, due to its costly estimation of the gene tree probabilities and need to optimize delimited species tree branch lengths, making it currently less feasible for moderate to large data sets.

Surprising results included occasionally better performance of the nonparametric method at shallow rather than deep divergences (especially with the 4 taxon pectinate tree simulation) and worse performance of the KC delimitation method with increasing number of samples per species. Deeper divergences should be easier to recover correctly, as there has been more time for genes to coalesce within each lineage. It appears that when methods performed better, in terms of getting the number of species correct or overall accuracy, at shallow divergences, it was due to less oversplitting than at deep divergences. Thus, such cases point to a possible bias toward oversplitting in the particular methods developed rather than an overall easier problem at shallow divergences. Similarly, more data should result in better answers. However, in this problem, adding more samples also increases the size of the problem space. It may be easier to return the exactly correct species tree with just 7 samples, where there are only 51,157 possible delimited species trees, than for 30 samples, where there are 2.95×10^{41} possible delimited species trees. The raw amount of data may have increased by a factor of 4.3, but the size of the problem space has increased by a factor of over 10^{36}.

Algorithm Rationale

The strategies to calculating the costs for the nonparametric approach justifiably appear somewhat ad hoc. Gene tree parsimony has a long history in systematics (Goodman et al. 1979) and papers using it still appear (Sanderson and McMahon 2007; Wehe et al. 2008), so it is a natural choice for this problem. However, deep coalescences (Maddison 1997) would also likely work and may better represent the evolutionary process, as having 2 gene copies persist without coalescing across several speciation events should have nonzero cost. Gene tree parsimony was chosen over deep coalescence primarily due to easier implementation and the effect of doing so rather than using deep coalescence cost instead remains to be evaluated. The choice for structure cost was less obvious. The chosen approach to calculating the structure cost has the advantages of not needing adjustment in the case of incomplete overlap of samples across genes and of being tied to an explicit model while being relatively fast to calculate. Analyses hinted at moderate insensitivity of results to specific settings combining the 2 costs, though increasing the relative weight of structure did tend to lead to more splitting of taxa.

The nonparametric approach is odd in that it ignores any information about branch length or population sizes in the species tree, which largely determine the probabilities of similarities and conflicts between gene trees. An analogous case is the traditional use of maximum parsimony for tree reconstruction: substitutions on branches are a function of mutation rates, the effect of selection and drift on mutations, and branch lengths. Parsimony does not take these into account (though see Tuffley and Steel 1997) and simply seeks to minimize the required number of substitutions to fit the data to the tree. This has long been known to lead to potential errors n theory (Felsenstein 1978) and perhaps in practice (Huelsenbeck 1997). However, parsimony remains a useful tool in phylogenetics. For example, it can be used to estimate starting trees for likelihood-based tree search (Stamatakis et al. 2005). In the same way, the nonparametric approach here ignores many important parameters of the evolutionary process but may, on occasion, return a useful result.

The heuristic search strategy developed here is also just one possible solution. As with the simpler problem of finding the best species tree given a fixed assignment of samples to species (the traditional question addressed by phylogenetics), better heuristic approaches may continue to be developed. The search strategy here evolved from existing moves in phylogenetics, such as SPR (Swofford 1990) and rerooting the tree, as well as obvious new moves related to the problem (such as moving samples from one species to another). Better moves certainly exist. One area for future improvement is in the splitting of samples when 1 species is divided into 2. For an initial species with N samples, there are $S(N,2)$ possible ways to split it. The implementation of the search currently only examines a random subset of these. This introduces a bias toward lumping due to the heuristic search. Joining 2 species always results in the same score, and so this move will always be taken if optimal, but a move to split a given species may not find

the optimal split of samples into 2 species, and so this splitting move may be rejected even if an unexamined division of samples results in a better score.

The nonparametric method has some disadvantages not shared by the KC delimitation approach. The nonparametric cost function is rather arbitrary, combining a *P* value for excess structure with the gene tree parsimony score. Investigations of the weighting parameters involved in that cost function in simulations and empirical data sets show that the values can affect results, but there are no consistently optimal values. Similar questions arise with other nonparametric methods, such as the proper weight to assign to transitions and transversions using parsimony for tree inference. One option would be to develop some sort of cross-validation approach to estimate the best parameter values, as has been done for calibrating trees by Sanderson (2002), but the fact that consistency of results can be ensured by choosing parameters that always result in 1 species being returned makes designing such an approach difficult.

The potential benefit of a guide tree to somehow incorporate information about samples often occurring together on the input gene trees, and thus most likely assigned to the same species became apparent with preliminary analyses. This information could be used in deciding which groups of samples to move from one species to another rather than the alternatives of trying only moves of single samples or trying all possible combinations of moves of samples. It could also be used to decide on initial groupings of samples for the starting delimited species trees. The search is more efficient starting from a species tree where samples occurring together on the gene trees tended to be put within the same species than starting from a species tree where samples were initially assigned randomly to species and then had to be moved together. One way to do this, if the gene trees always had the same taxa, would be to compute majority rule consensus trees and then split the trees on edges frequently present in the input gene trees to get the initial groupings of samples. Given that, in practice, some samples may not be present for some genes, a different approach was needed. One such approach would have been matrix representation with parsimony (Baum 1992; Ragan 1992) but that requires a heuristic search itself and does not naturally return information on agreement between the input trees. Information on triplets was already being used in the nonparametric approach and so could be useful in doing initial clustering under that approach. Measuring triplet overlap also naturally lends itself to representation in a distance matrix (proportion of times 2 samples form a clade in a triplet containing them). Neighbour joining is an efficient way to summarize the information in such a matrix and returns a tree where branch lengths contain information about agreement between gene trees. This could be used to get the set of samples to move together (all clades on this tree) and groups of samples most likely to be from the same species (those together in clades with long subtending edges). This is the purpose of the guide tree and rationale for the new approach to recover it. Note that the guide tree does not provide the initial species topology, which is just a random tree connecting the semirandom starting species assignments, nor is the search constrained to match the topology or grouping of samples on the guide tree.

Dealing with Uncertainty

Estimating uncertainty is currently difficult. One could bootstrap data, generate gene trees from this data, and perform inference on these bootstrap replicates, but this just estimates uncertainty due to uncertainty in the gene trees given the data. However, even if the gene trees are known exactly, they are still random draws from a coalescent process, and a repeat of this sampling would almost certainly result in a different set of trees. One way to assess this uncertainty is to use parametric bootstrapping, simulating gene tree evolution under a specified species tree model, specifying divergence times, population sizes, population structure, and any gene flow, using a program such as ms or Mesquite (Maddison W.P. and Maddison D.R. 2007), and then analysing these simulated samples in the same way the original data were analysed. However, this requires knowing in detail the hypothesis to test. Simply bootstrapping estimated gene trees will not work, as sampling with replacement would often result in the same gene tree being sampled more than once, inflating the excess structure score and thus tending to cause more splits. Jackknifing the gene trees (sampling without replacement) may provide some estimate of the uncertainty if there are enough gene trees sampled. Currently, the implementation of the new approaches returns multiple solutions if it finds more than one of equal score, which can give a faint idea of uncertainty in the result. The KC delimitation approach could be modified to save all results found within a certain log likelihood of the optimum result, but the current search strategy is inadequate to estimate the contents of this region. The KC delimitation could be used in a full Bayesian search, though the time required might be prohibitive.

Related Work

There are numerous methods related to those developed here. There has been a recent trend in phylogenetics toward creating species trees as seen as something potentially distinct from the gene tree(s). Most of these methods, such as the BEST approach (Edwards et al. 2007; Liu and Pearl 2007; Liu et al. 2008), the "STEM" approach (Kubatko et al. 2009), the "Minimize Deep Coalescences" approach (Maddison and Knowles 2006), the "GLASS" method (Mossel and Roch 2007), or the "coalescent-based approach" (Carstens and Knowles 2007), have a fixed assignment of samples to species, though some approaches (Knowles and Carstens 2007) do allow some optimization of this assignment. The KC delimitation developed here is largely an extension of

the Knowles and Carstens (2007) approach, using a similar optimality criterion (probability of the gene trees given the species tree), but allowing a search over all possible assignments. In theory, many of these other approaches to recovering the species tree could be extended in the same way, though the increased size of the problem space may make such methods impractical (and methods without some sort of intraspecific cost, such as the Minimize Deep Coalescences approach [Maddison and Knowles 2006], would generally return only 1 species). The new approaches optimize the species tree over a set of gene trees rather than over the gene sequences directly, a deficiency shared by all the above methods but the BEST method.

A different set of methods are being developed for DNA barcoding, which typically uses a single locus where the assignment of samples to species is at best only partially known, but the species tree itself is not of interest. A particularly sophisticated approach to DNA barcoding is the generalized mixed Yule-coalescent (GMYC) model (Pons et al. 2006), which seeks to cut a single locus tree into a portion where a speciation process affects the branch lengths and a portion where a coalescent process within species comes into play, using this boundary to define species. It is grossly similar to the nonparametric method in that it divides a species tree into inter- and intraspecific portions, though it uses a model to do this and uses just 1 locus. However, GMYC will not work well with small numbers of species or gene trees with zero length terminal branches (Barraclough T., personal communication), whereas the methods developed here are most feasible with small numbers of species and ignore gene tree branch lengths (though do require resolved gene trees).

Finally, there are methods for estimating gene flow between populations, which could, in theory, be used to decide when there is no gene flow (speciation, by some definitions) between populations. There are numerous coalescent-based approaches for this, such as those implemented in MIGRATE (Beerli and Felsenstein 1999) and IM (Nielsen and Wakeley 2001), which also generally start from genetic data rather than fixed inferred gene trees. An approach vaguely similar to the nonparametric approach is the cladistic measure of gene flow developed (Slatkin and Maddison 1989), which uses a gene trees with population assignments of samples known to make estimates of migration rate between populations and which compares favourably with F_{ST}-based approaches for estimating migration rates under some conditions (Hudson et al. 1992).

Utility

In practice, the question answered here, the sorting of anonymous samples into species while inferring a species tree, is unlikely to be one asked by taxonomists working on groups like angiosperms or vertebrates. Most such groups have some previous work done on them, and much of the work of a revision is deciding whether to split or lump old species as well as deciding whether new samples belong in existing species. Such taxonomists often also have additional information available, such as localities of specimens, morphological characters, and hypotheses drawn from other approaches. Given the deficiencies in the new methods developed in this paper, it is premature to use them alone to do alpha taxonomy in such cases, but they may be a useful addition to a taxonomist's toolbox. However, the new methods, and even simpler approaches like computing majority rule trees or using the approach of Dettman et al. (2003), may be especially useful when dealing with groups for which there are no existing taxonomic or phylogenetic hypotheses, such as environmental samples of fungi or other understudied groups, where the methods can fairly quickly return an estimate of the species boundaries and phylogeny with basically no initial information required other than knowing which gene sequences correspond to the same individual organism. For all taxa, the methods can help infer a species tree in the presence of widespread incomplete lineage sorting events, and they can provide evidence otherwise hard to obtain, such as whether 3 allopatric populations form 1 or multiple species. The implementation allows fixed assignment of samples to species, so alternate possible assignments such as uncertainty regarding whether to split an existing species can be evaluated (as in Knowles and Carstens 2007, but without requiring specification of all branch lengths). The methods here may also be useful for providing a first working hypothesis of relationships and species limits when revising a group. Where conservation plans rest on taxonomic decisions, these methods, like other algorithmic methods, have the advantage of reducing apparent subjectivity of assigning species rank, but the fact that they often give incorrect answers mandates caution and judgment when interpreting results.

Speciation is a complex process. Scientists have developed numerous tools to help make inferences about speciation. This paper describes the problem and provides additional tools that allow information from multiple genes to be used, ideally in concert with other approaches, to help delimit species and infer the species tree.

Program Note

The methods described here are implemented in the open source program Brownie 2.1. The program reads standard Nexus files containing a set of rooted, bifurcating gene trees with, optionally, tree weights. If species assignments are fixed, the program can also serve as a heuristic search tool for the optimal species tree given gene trees under gene tree parsimony. The program is available at http://www.brianomeara.info/brownie.

Data Note

Output from all analyses are available at http://www.brianomeara.info/jistdata/index.html.

FUNDING

This work was supported by the Center for Population Biology at the University of California Davis, by a National Science Foundation Graduate Student Fellowship to BCO, and by the National Evolutionary Synthesis Center (funded by the National Science Foundation, grant number EF-0423641).

ACKNOWLEDGMENTS

This idea was inspired through conversations with M. Sanderson and H. B. Shaffer and a seminar on a different speciation approach by D. Maddison. The methods were refined through discussions with M. Sanderson, M. Turelli, and P. Ward. T. Barraclough, J. Degnan, 2 other reviewers, and editors J. Sullivan and L. Knowles all provided invaluable suggestions.

REFERENCES

Abdo Z., Golding G.B. 2007. A step toward barcoding life: a model-based, decision-theoretic method to assign genes to preexisting species groups. Syst. Biol. 56:44–56.

Abramowitz M., Stegun I.A. 1972. Handbook of mathematical functions, with formulas, graphs, and mathematical tables. Washington (DC): National Bureau of Standards.

Akaike H. 1973. Information theory and an extension of the maximum likelihood principle. In: Petrov B.N., Csaki F., editors. Second International Symposium on Information Theory. Budapest (Hungary): Akademiai Kiado. p. 267–281.

Akaike H. 1974. A new look at the statistical model identification. IEEE Trans. Automat. Contr. 19:716–723.

Avise J.C. 1983. Mitochondrial DNA differentiation during the speciation process in Peromyscus. Mol. Biol. Evol. 1:38–56.

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

Baum D.A., Shaw K.L. 1995. Genealogical perspectives on the species problem. In: Hoch P.C., Stephenson A.G., editors. Experimental and molecular approaches to plant biosystematics. Saint Louis (MO): Missouri Botanical Garden. p. 289–303.

Beaumont M.A., Zhang W., Balding D. 2002. Approximate Bayesian computation in population genetics. Genetics. 162:2025–2035.

Beerli P., Felsenstein J. 1999. Maximum-likelihood estimation of migration rates and effective population numbers in two populations using a coalescent approach. Genetics. 152:763–773.

Beerli P., Felsenstein J. 2001. Maximum likelihood estimation of a migration matrix and effective population sizes in n subpopulations by using a coalescent approach. Proc. Natl. Acad. Sci. USA. 98: 4563–4568.

Brumfield R.T., Liu L., Lum D.E., Edwards S.V. 2008. Comparison of species tree methods for reconstructing the phylogeny of bearded manakins (Aves: Pipridae, Manacus) from multilocus sequence data. Syst. Biol. 57:719–731.

Carstens B.C., Knowles L.L. 2007. Estimating species phylogeny from gene-tree probabilities despite incomplete lineage sorting: an example from Melanoplus grasshoppers. Syst. Biol. 56:400–411.

Cavalli-Sforza L.L., Edwards A.W.F. 1967. Phylogenetic analysis: models and estimation procedures. Evolution. 21:550–570.

Cotton J.A., Page R.D.M. 2003. Gene tree parsimony vs. uninode coding for phylogenetic reconstruction. Mol. Phylogenet. Evol. 29:298–308.

Critchlow D.E., Pearl D.K., Qian C. 1996. The triples distance for rooted bifurcating phylogenetic trees. Syst. Biol. 45:323–334.

Cummings M.P., Neel M.C., Shaw K.L. 2008. A genealogical approach to quantifying lineage divergence. Evolution. 62:2411–2422.

De Queiroz K. 2007. Species concepts and species delimitation. Syst. Biol. 56:879–886.

Degnan J.H., Rosenberg N.A. 2006. Discordance of species trees with their most likely gene trees. PLoS Genet. 2:e68.

Degnan J.H., Salter L.A. 2005. Gene tree distributions under the coalescent process. Evolution. 59:24–37.

Dettman J.R., Jacobson D.J., Taylor J.W. 2003. A multilocus genealogical approach to phylogenetic species recognition in the model eukaryote Neurospora. Evolution. 57:2703–2720.

Doyle J.J. 1992. Gene trees and species trees—molecular systematics as one-character taxonomy. Syst. Bot. 17:144–163.

Edwards S.V., Liu L., Pearl D.K. 2007. High-resolution species trees without concatenation. Proc. Natl. Acad. Sci. USA. 104:5936–5941.

Fellows M., Hallett M., Stege U. 2003. Analogs & duals of the MAST problem for sequences & trees. J. Algorithms. 49:192–216.

Felsenstein J. 1978. Cases in which parsimony or compatibility methods will be positively misleading. Syst. Zool. 27:401–410.

Feng D.F., Doolittle R.F. 1987. Progressive sequence alignment as a prerequisite to correct phylogenetic trees. J. Mol. Evol. 25:351–360.

Fitch W.M. 1970. Distinguishing homologous and analogous proteins. Syst. Zool. 19:99–113.

Foulds L.R., Graham R.L. 1982. The Steiner problem in phylogeny is NP-complete. Adv. Appl. Math. 3:43–49.

Funk D.J., Omland K.E. 2003. Species-level paraphyly and polyphyly: frequency, causes, and consequences, with insights from animal mitochondrial DNA. Annu. Rev. Ecol. Evol. Syst. 34:397–423.

Goodman M., Czelusniak J., Moore G.W., Romero-Herrera A.E., Matsuda G. 1979. Fitting the gene lineage into its species lineage, a parsimony strategy illustrated by cladograms constructed from globin sequences. Syst. Zool. 28:132–163.

Harding E.F. 1971. The probabilities of rooted tree-shapes generated by random bifurcation. Adv. Appl. Probab. 3:44–77.

Hebert P.D.N., Cywinska A., Ball S.L., deWaard J.R. 2003. Biological identifications through DNA barcodes. Proc. R. Soc. B Biol. Sci. 270:313–321.

Higgins D.G., Bleasby A.J., Fuchs R. 1992. CLUSTAL V: improved software for multiple sequence alignment. Bioinformatics. 8:189–191.

Hudson R.R. 1992. Gene trees, species trees and the segregation of ancestral alleles. Genetics. 131:509–512.

Hudson R.R. 2002. Generating samples under a Wright-Fisher neutral model of genetic variation. Bioinformatics. 18:337–338.

Hudson R.R., Coyne J.A. 2002. Mathematical consequences of the genealogical species concept. Evolution. 56:1557–1565.

Hudson R.R., Slatkin M., Maddison W.P. 1992. Estimation of levels of gene flow from DNA sequence data. Genetics. 132:583–589.

Huelsenbeck J.P. 1997. Is the Felsenstein zone a fly trap? Syst. Biol. 46:69–74.

Knowles L.L., Carstens B.C. 2007. Delimiting species without monophyletic gene trees. Syst. Biol. 56:887–895.

Kubatko L., Carstens B., Knowles L. 2009. STEM: species tree estimation using maximum likelihood for gene trees under coalescence. Bioinformatics. 25:971–973.

Liu L., Pearl D.K. 2007. Species trees from gene trees: reconstructing Bayesian posterior distributions of a species phylogeny using estimated gene tree distributions. Syst. Biol. 56:504–514.

Liu L., Pearl D.K., Brumfield R.T., Edwards S.V. 2008. Estimating species trees using multiple-allele DNA sequence data. Evolution. 62:2080–2091.

Ma B., Li M., Zhang L. 1998. On reconstructing species trees from gene trees in term of duplications and losses. Proceedings of the Second Annual International Conference on computational Molecular Biology; 1998 Mar. 22–25; New York: ACM Press. p. 182–191.

Machado C.A., Hey J. 2003. The causes of phylogenetic conflict in a classic Drosophila species group. Proc. R. Soc. Lond. Ser. B Biol. Sci. 270:1193–1202.

Machado C.A., Kliman R.M., Markert J.A., Hey J. 2002. Inferring the history of speciation from multilocus DNA sequence data: The case of Drosophila pseudoobscura and close relatives. Mol. Biol. Evol. 19:472–488.

Maddison W.P. 1997. Gene trees in species trees. Syst. Biol. 46:523–536.

Maddison W.P., Knowles L.L. 2006. Inferring phylogeny despite incomplete lineage sorting. Syst. Biol. 55:21–30.

Maddison W.P., Maddison D.R. 2007. Mesquite: a modular system for evolutionary analysis. Version 2.0. Available from: http://www.mesquiteproject.org.

Manel S., Gaggiotti O.E., Waples R.S. 2005. Assignment methods: matching biological questions with appropriate techniques. Trends Ecol. Evol. 20:136–142.
Matz M.V., Nielsen R. 2005. A likelihood ratio test for species membership based on DNA sequence data. Philos. Trans. R. Soc. Lond. B Biol. Sci. 360:1969–1974.
McKenzie A., Steel M. 2000. Distributions of cherries for two models of trees. Math. Biosci. 164:81–92.
Mossel E., Roch S. 2007. Incomplete lineage sorting: consistent phylogeny estimation from multiple loci. arXiv:arXiv:0710.0262v0712 [q-bio.PE]. Preprint.
Nielsen R., Wakeley J. 2001. Distinguishing migration from isolation: a Markov chain Monte Carlo approach. Genetics. 158:885–896.
Nuytinck J., Verbeken A. 2007. Species delimitation and phylogenetic relationships in Lactarius section Deliciosi in Europe. Mycol. Res. 111:1285–1297.
O'Meara B.C., Ane C., Sanderson M.J., Wainwright P.C. 2006. Testing for different rates of continuous trait evolution using likelihood. Evolution. 60:922–933.
Pamilo P., Nei M. 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5:568–583.
Pons J., Barraclough T., Gomez-Zurita J., Cardoso A., Duran D., Hazell S., Kamoun S., Sumlin W., Vogler A. 2006. Sequence-based species delimitation for the DNA taxonomy of undescribed insects. Syst. Biol. 55:595–609.
Ragan M. 1992. Phylogenetic inference based on matrix representation of trees. Mol. Phylogenet. Evol. 1:53–58.
Rannala B., Yang Z.H. 2003. Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. Genetics. 164:1645–1656.
Ree R.H., Moore B.R., Webb C.O., Donoghue M.J. 2005. A likelihood framework for inferring the evolution of geographic range on phylogenetic trees. Evolution. 59:2299–2311.
Robinson D.F., Foulds L.R. 1981. Comparison of phylogenetic trees. Math. Biosci. 53:131–147.
Sanderson M.J. 2002. Estimating absolute rates of molecular evolution and divergence times: a penalized likelihood approach. Mol. Biol. Evol. 19:101–109.
Sanderson M., McMahon M. 2007. Inferring angiosperm phylogeny from EST data with widespread gene duplication. BMC Evol. Biol. 7:S3.
Shaffer H.B., Thomson R.C. 2007. Delimiting species in recent radiations. Syst. Biol. 56:896–906.
Slatkin M., Maddison W.P. 1989. A cladistic measure of gene flow inferred from the phylogenies of alleles. Genetics. 123:603–613.
Slowinski J.B., Knight A., Rooney A.P. 1997. Inferring species trees from gene trees: a phylogenetic analysis of the Elapidae (Serpentes) based on the amino acid sequences of venom proteins. Mol. Phylogenet. Evol. 8:349–362.
Stamatakis A., Ludwig T., Meier H. 2005. RAxML-III: a fast program for maximum likelihood-based inference of large phylogenetic trees. Bioinformatics. 21:456–463.
Sugiura N. 1978. Further analysts of the data by Akaike's information criterion and the finite corrections. Commun. Stat. Theory Methods. 7:13–26.
Swofford D.L. 1990. PAUP: phylogenetic analysis using parsimony. Version 3.0. Champaign (IL): Illinois Natural History Survey.
Swofford, D.L. 2003. PAUP*. Phylogenetic analysis using parsimony (*and Other Methods). version 4. Sunderland (MA): Sinauer Associates.
Tajima F. 1983. Evolutionary relationship of DNA-sequences in finite populations. Genetics. 105:437–460.
Tautz D., Arctander P., Minelli A., Thomas R.H., Vogler A.P. 2003. A plea for DNA taxonomy. Trends Ecol. Evol. 18:70–74.
Tavare S., Balding D.J., Griffiths R.C., Donnelly P. 1997. Inferring coalescence times from DNA sequence data. Genetics. 145:505–518.
Thompson J.D., Higgins D.G., Gibson T.J. 1994. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. Nucleic Acids Res. 22:4673–4680.
Tuffley C., Steel M. 1997. Links between maximum likelihood and maximum parsimony under a simple model of site substitution. Bull. Math. Biol. 59:581–607.
Wang R.L., Wakeley J., Hey J. 1997. Gene flow and natural selection in the origin of Drosophila pseudoobscura and close relatives. Genetics. 147:1091–1106.
Wehe A., Bansal M.S., Burleigh J.G., Eulenstein O. 2008. DupTree: a program for large-scale phylogenetic analyses using gene tree parsimony. Bioinformatics. 24:1540–1541.
Weir J.T., Schluter D. 2007. The latitudinal gradient in recent speciation and extinction rates of birds and mammals. Science. 315: 1574–1576.
Yang Z. 2002. Likelihood and Bayes estimation of ancestral population sizes in Hominoids using data from multiple loci. Genetics. 162:1811–1823.
Yang Z. 2005. Speed. Version 2. Distributed by the author. Available from http://www.abacus.gene.ucl.ac.uk/software/speed2.tar.gz.
Zhang L. 2000. Inferring a species tree from gene trees under the deep coalescence cost. RECOMB; Tokyo, Japan.
Zhang A.B., Sikes D.S., Muster C., Li S.Q. 2008. Inferring species membership using DNA sequences with back-propagation neural networks. Syst. Biol. 57:202–215.
Zmasek C.M., Eddy S.R. 2001. A simple algorithm to infer gene duplication and speciation events on a gene tree. Bioinformatics. 17:821–828.