Bioinformatics Advance Access published August 7, 2014

Quartet Inference from SNP Data Under the Coalescent Model

Julia Chifman[1] and Laura Kubatko[2,3]*

[1]Department of Cancer Biology, Wake Forest School of Medicine, Winston-Salem, NC, 27157
[2]Department of Statistics, The Ohio State University, Columbus, OH 43210
[3]Department of Evolution, Ecology, and Organismal Biology, The Ohio State University, Columbus, OH 43210

Associate Editor: Prof. David Posada

ABSTRACT

Motivation: Increasing attention has been devoted to estimation of species-level phylogenetic relationships under the coalescent model. However, existing methods either use summary statistics (gene trees) to carry out estimation, ignoring an important source of variability in the estimates, or involve computationally-intensive Bayesian Markov chain Monte Carlo algorithms that don't scale well to whole-genome data sets.

Results: We develop a method to infer relationships among quartets of taxa under the coalescent model using techniques from algebraic statistics. Uncertainty in the estimated relationships is quantified using the nonparametric bootstrap. The performance of our method is assessed with simulated data. We then describe how our method could be used for species tree inference in larger taxon samples, and demonstrate its utility using data sets for *Sistrurus* rattlesnakes and for soybeans.

Availability and Implementation: The method to infer the phylogenetic relationship among quartets is implemented in the software SVDquartets, available at www.stat.osu.edu/∼lkubatko/software/SVDquartets.

Contact: Laura Kubatko, lkubatko@stat.osu.edu

1 INTRODUCTION

With recent advances in DNA sequencing technology, it is now common to have available alignments from multiple genes for inference of an overall *species-level phylogeny*. While this species tree is generally the object that we seek to estimate, it is widely known that each individual gene has its own phylogeny, called a *gene tree*, which may not agree with the species tree. Many possible causes of this gene incongruence are known, including horizontal gene transfer, gene duplication and loss, hybridization, and incomplete lineage sorting (ILS) (Maddison, 1997). Of these, the best studied is incomplete lineage sorting, which is commonly modeled by the coalescent process (Kingman, 1982a,b; Liu et al., 2009a). Much recent effort has been devoted to the development of methods to estimate species-level phylogenies from multi-locus data under the coalescent model (Liu and Pearl, 2007; Liu et al., 2009b; Kubatko et al., 2009; Heled and Drummond, 2010; Than and Nakhleh, 2009; Bryant et al., 2012).

Here we consider this basic problem, although our approach to the problem differs from previous approaches in several important ways. Previous approaches can be divided into two groups (Liu et al., 2009a): summary-statistics approaches and sequence-based approaches. Summary-statistics approaches first estimate a gene tree independently for each gene, and then treat the estimated gene trees as data for a second stage of analysis to estimate the species tree. The most popular approaches in this category are Maximum Tree (Liu et al., 2009b) (also implemented in the program STEM (Kubatko et al., 2009)), STAR (Liu et al., 2009c), STEAC (Liu et al., 2009c), MP-EST (Liu et al., 2010), and Minimize Deep Coalescences (MDC; as implemented in the program PhyloNet (Than and Nakhleh, 2009)). These methods are computationally efficient for large data sets, but generally ignore variability in the estimated gene trees and thus potentially lose accuracy. The second group of methods uses the full data for estimation of the species tree via a Bayesian framework for inference. The three most common methods in this group, BEST (Liu and Pearl, 2007), *BEAST (Heled and Drummond, 2010), and SNAPP (Bryant et al., 2012), all seek to estimate the posterior distribution for the species tree using Markov chain Monte Carlo (MCMC), but differ in some details of the implementation. These methods become time-consuming when the number of loci is large, and assessment of convergence of the MCMC can be difficult.

Our proposed method is distinct from both classes of existing approaches in that it uses the full data directly, but does not utilize a Bayesian framework. It is thus computationally efficient while incorporating all sources of variability (both mutational variance and coalescent variance (cf. Huang et al. (2010))) in the estimation process. The theory underlying our method applies to unlinked Single Nucleotide Polymorphism (SNP) data, for which each site is assumed to have its own genealogy drawn from the coalescent model; however, we use simulation to show that the method also performs well for multi-locus sequence data. To describe our proposed method, we first begin with a brief overview of the coalescent model in the context of species-level phylogenetics. We use simulation to assess the performance of the method for both simulated and empirical data. We conclude with a short discussion of how the proposed method can be scaled up to larger taxon sets

---

*to whom correspondence should be addressed

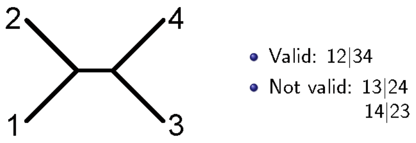

Fig. 1. Example four-taxon phylogeny. Split 12|34 is valid, since the subtree consisting of taxa 1 and 2 does not overlap the subtree consisting of taxa 3 and 4. The two non-valid splits for this tree are 13|24 and 14|23.

for estimation of species phylogenies in a coalescent framework, and apply it to two empirical data sets.

1.1 Site Pattern Probability Distributions Under the Coalescent Model

The coalescent model can be used to compute the probability distribution of gene trees given a particular species tree and set of speciation times (which determine species tree branch lengths). Both the discrete probability distribution on the space of gene tree topologies (Degnan and Salter, 2005) and the probability density on the space of gene trees with branch lengths (Rannala and Yang, 2003) have been derived recently. Using these probability distributions, it is possible to compute the probability distribution on data patterns at the tips of a species tree. Let X_H be the observed state in the data at tip H, and, referring to the tree in Figure 1, for example, define p_{ijkl} as

$$p_{ijkl} = P(X_1 = i, X_2 = j, X_3 = k, X_4 = l) \qquad (1)$$

for $i, j, k, l \in \{A, C, G, T\}$. In order to compute the probability distribution $\{p_{ijkl} | i, j, k, l \in \{A, C, G, T\}\}$, we need the following: (1) a species phylogeny, with speciation times specified; and (2) a model for sequence evolution along a gene tree, e.g., the General Time Reversible (GTR) model (Tavare, 1986) or the Jukes-Cantor (JC69) model (Jukes and Cantor, 1969). See DeGiorgio and Degnan (2010) for an example of how to carry out this computation for a two-state model. The details of the calculation for arbitrary k-state models can be found in Chifman and Kubatko (2014). We now describe how this probability distribution can be used to compute a score on a quartet of taxa that can identify the true quartet relationship. To begin, we define a *split* of a phylogenetic tree as follows.

Definition: A **split** of a set of taxa $\mathcal{L}$ is a bipartition of $\mathcal{L}$ into two non-overlapping subsets L_1 and L_2, denoted $L_1|L_2$. A split $L_1|L_2$ is **valid** for tree T if the subtrees containing the taxa in L_1 and in L_2 do not intersect.

For a quartet of taxa, we consider splits for which $|L_1| = 2$ (and thus necessarily $|L_2| = 2$), e.g., we consider splitting the four taxa into two groups of two. For example, consider a valid split $L_1|L_2$, where $L_1 = \{1, 2\}$ and $L_2 = \{3, 4\}$ (Figure 1). Under this partition, we can display the probability distribution $P = \{p_{ijkl} | i, j, k, l \in \{A, C, G, T\}\}$ in the form of a *flattening* along a split $L_1|L_2$, denoted by $Flat_{L_1|L_2}(P)$, as follows:

$$\begin{pmatrix} p_{AAAA} & p_{AAAC} & p_{AAAG} & p_{AAAT} & p_{AACA} & \cdots & p_{AATT} \\ p_{ACAA} & p_{ACAC} & p_{ACAG} & p_{ACAT} & p_{ACCA} & \cdots & p_{ACTT} \\ p_{AGAA} & p_{AGAC} & p_{AGAG} & p_{AGAT} & p_{AGCA} & \cdots & p_{AGTT} \\ p_{ATAA} & p_{ATAC} & p_{ATAG} & p_{ATAT} & p_{ATCA} & \cdots & p_{ATTT} \\ p_{CAAA} & p_{CAAC} & p_{CAAG} & p_{CAAT} & p_{CACA} & \cdots & p_{CATT} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ p_{TTAA} & p_{TTAC} & p_{TTAG} & p_{TTAT} & p_{TTCA} & \cdots & p_{TTTT} \end{pmatrix}.$$

In the above 16×16 matrix, the rows correspond to the possible nucleotides for the two taxa in set L_1 and the columns correspond to the possible nucleotides for the two taxa in set L_2. For more information about flattening of a tensor P for the general Markov model on a gene tree, see Allman and Rhodes (2008). Using this representation, we make use of the following result for species tree inference under the coalescent.

Theorem 1 [Chifman and Kubatko, 2014]
Let $\mathcal{C}$ denote the class of coalescent models under the four-state GTR model on a four-taxon binary species tree. For a valid split $L_1|L_2$, rank$(Flat_{L_1|L_2}(P)) \leq 10$ for all distributions P arising from $\mathcal{C}$. For a non-valid split $L_1|L_2$, generically, rank$(Flat_{L_1|L_2}(P)) > 10$.

We note that the above theorem implies generic identifiability of the unrooted species tree topology for four taxa under the coalescent model (Chifman and Kubatko, 2014). By "generic" we mean that the set of parameters on which the model is non-identifiable is a subset of a proper subvariety of measure zero. In addition, we have established generic identifiability of the unrooted n-taxon species tree under the coalescent model from the induced quartets (Chifman and Kubatko, 2014).

2 METHODS

2.1 Inferring Splits Using Singular Value Decomposition

Our goal is to use the result of Theorem 1 to infer species phylogenies. Assume that the available data consist of a large sample of unlinked SNPs, which we can use to construct an estimate of the matrix $Flat_{L_1|L_2}(P)$. We call this matrix $Flat_{L_1|L_2}(\hat{P})$, and define this matrix by

$$\begin{pmatrix} \hat{p}_{AAAA} & \hat{p}_{AAAC} & \hat{p}_{AAAG} & \hat{p}_{AAAT} & \hat{p}_{AACA} & \cdots & \hat{p}_{AATT} \\ \hat{p}_{ACAA} & \hat{p}_{ACAC} & \hat{p}_{ACAG} & \hat{p}_{ACAT} & \hat{p}_{ACCA} & \cdots & \hat{p}_{ACTT} \\ \hat{p}_{AGAA} & \hat{p}_{AGAC} & \hat{p}_{AGAG} & \hat{p}_{AGAT} & \hat{p}_{AGCA} & \cdots & \hat{p}_{AGTT} \\ \hat{p}_{ATAA} & \hat{p}_{ATAC} & \hat{p}_{ATAG} & \hat{p}_{ATAT} & \hat{p}_{ATCA} & \cdots & \hat{p}_{ATTT} \\ \hat{p}_{CAAA} & \hat{p}_{CAAC} & \hat{p}_{CAAG} & \hat{p}_{CAAT} & \hat{p}_{CACA} & \cdots & \hat{p}_{CATT} \\ \vdots & \vdots & \vdots & \vdots & \vdots & \ddots & \vdots \\ \hat{p}_{TTAA} & \hat{p}_{TTAC} & \hat{p}_{TTAG} & \hat{p}_{TTAT} & \hat{p}_{TTCA} & \cdots & \hat{p}_{TTTT} \end{pmatrix},$$

where $\hat{p}_{ijkl}$ is the frequency with which we observe the event $\{X_1 = i, X_2 = j, X_3 = k, X_4 = l\}$ in the data, where $L_1 = \{1, 2\}$ and $L_2 = \{3, 4\}$. A key observation is that this can be rapidly tabulated for quartets of taxa even for data sets of very large size.

We want to infer which of the three possible splits on quartets is the true split. One way to assess this would be to consider the $Flat_{L_1|L_2}(\hat{P})$ matrix for each of the three possible splits, and measure which of the three is closest to a rank 10 matrix. To do this, we need a method to measure distances between matrices. Our choice of a distance, described below, is modeled after the approach of Eriksson (2005), who considered the problem of tree estimation from a flattening matrix obtained from the probability distribution of site patterns at the tips of a gene tree. His overall approach to estimation of the phylogeny differed from ours, however, in that he used splits of varying sizes (rather than just splits of quartets of taxa) to develop a clustering algorithm to obtain the phylogenetic estimate. We provide the details of our approach below.

Let a_{ij} be the $(i,j)^{th}$ entry of an $m \times n$ matrix A. The **Frobenius norm** of a matrix A is

$$\|A\|_F = \sqrt{\sum_{i=1}^{m}\sum_{j=1}^{n} a_{ij}^2}.$$

An important property of the Frobenius norm is its characterization using the singular values of A, that is

$$\|A\|_F = \sqrt{\sum_{i=1}^{p} \sigma_i^2},$$

where $\sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_p \geq 0$ are the *singular values* of A and $p = \min\{m, n\}$.

The well-known low-rank approximation theorem (*Eckart-Young theorem*), implies that the distance from a matrix A to the nearest rank k matrix in the Frobenius norm is

$$\min_{\text{rank}(B)=k} \|A - B\|_F = \sqrt{\sum_{i=k+1}^{p} \sigma_i^2}.$$

See Section 2.4 in Golub and Van Loan (2013) for more information about singular value decomposition.

We apply this well-known result to our species tree estimation problem by defining the **SVD score** for a split $L_1|L_2$ to be

$$SVD(L_1|L_2) := \sqrt{\sum_{i=11}^{16} \hat{\sigma}_i^2}, \qquad (2)$$

where $\hat{\sigma_i}$ are the singular values of $Flat_{L_1|L_2}(\hat{P})$, for $i \in \{11, \ldots, 16\}$. Our proposal for inferring the true species-level relationship within a sample of four taxa is thus the following. For each of the three possible splits, construct the matrix $Flat_{L_1|L_2}(\hat{P})$ and compute $SVD(L_1|L_2)$. The split with the smallest score is taken to be the true split.

To quantify uncertainty in the inferred split, we implement a nonparametric bootstrap procedure as follows. For a data set consisting of M aligned sites, we re-sampled the columns of the data matrix with replacement M times to generate a new bootstrapped data matrix, and the SVD scores of the three splits are computed for this bootstrapped data matrix. This procedure is repeated B times, and the proportion of bootstrapped data sets that support each of the three possible splits provides a measure of support for that split.

2.2 Simulation Study

We first use simulated data to assess the ability of $SVD(L_1|L_2)$ to correctly identify the valid split under a variety of conditions. Before describing the simulation procedure, we first point out that while much of the currently available methodology for inferring species trees assumes that multi-locus data (e.g., aligned DNA sequences from many independent loci) are available for inference, our method is actually designed for unlinked sites, for example, for a sample of unlinked SNPs. This is because in computing the probability distribution of site patterns at the tips of the species tree, we integrate over the probability distribution of gene trees under the coalescent model, with the implicit assumption that sequence data evolve along these gene trees. Thus each site pattern is viewed as an independent draw from the distribution $f(X_1 = i, X_2 = j, X_3 = k, X_4 = l|S) = \int_G f(X_1 = i, X_2 = j, X_3 = k, X_4 = l|G) f(G|S) dG$, where S represents the species tree (topology and speciation times) and G represents a gene tree (both topology and divergence times). True multi-locus data, however, consist of an aligned portion of the DNA that is believed to share a single underlying gene tree, and thus all sequence data within a locus are believed to have evolved from a common gene genealogy.

We wish to examine the performance of our method for both unlinked SNP data and for multi-locus data, and we thus consider simulated data of two types: unlinked SNP data (e.g., each site has its own underlying gene tree) and multi-locus data (a sequence of length l is simulated from a shared underlying gene tree). Our simulation consists of the following steps:

1. Generate a sample of g gene trees from the model species tree ((1:x,2:x):x,(3:x,4:x):x), where x is the length of each branch, under the coalescent model using the program COAL (Degnan and Salter, 2005).
2. Generate sequence data of length n on each gene tree under a specified substitution model using the program Seq-Gen (Rambaut and Grassly, 1997).
3. Construct the flattening matrix for each of the three possible splits, and compute $SVD(L_1|L_2)$ for each.
4. Repeat the above procedure (Steps 1 – 3) $1,000$ times and record $SVD(L_1|L_2)_k, k = 1, 2, \ldots, 1,000$, for each split. For each of the $1,000$ data sets, generate B bootstrapped data sets and record $SVD(L_1|L_2)_{k,b}, k = 1, 2, \ldots, 1,000; b = 1, 2, \ldots, B$ for each split.

Given the above simulation algorithm, there are several choices to be made at each step. In step (1), we must select the lengths of the branches, x, in the model species tree. We considered branches of length 0.5, 1.0, and 2.0 coalescent units. A branch of length 0.5 coalescent units is very short, and corresponds to a case in which there will be widespread incomplete lineage sorting, making species tree inference difficult. A branch of length 2.0 coalescent units is longer and will result in much lower rates of incomplete lineage sorting, resulting in an easier species tree inference problem.

In Step (2), we need to choose the gene length, n. In simulating unlinked SNP data, we used $g = 5,000$ and $n = 1$ (corresponding to 5,000 unlinked SNPs) and for the multi-locus setting, we considered $g = 10$ and $n = 500$ (corresponding to 10 genes, each of length 500 sites). Further, step (2) requires choice of

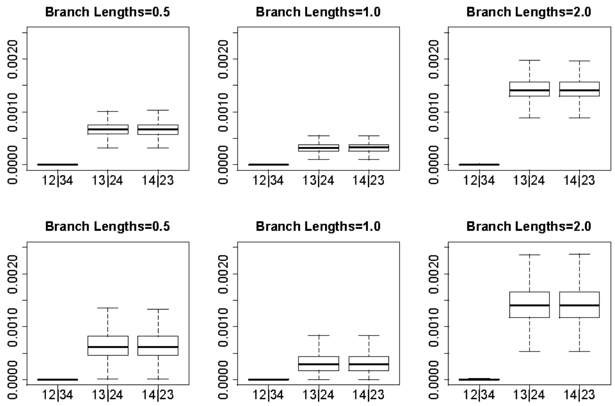

Fig. 2. Simulation results for the JC69 model. The top row gives the results for 5,000 unlinked SNP sites and the bottom row gives the results for 10 genes with 500 sites each. The columns correspond to differing branch lengths in the model species tree. The first boxplot in each subfigure shows the distribution of SVD scores for the true split, while the next two boxplots show the distribution for the two false splits.

substitution model to be used to simulate sequence data on the sampled gene trees. We considered two possibilities: the Jukes-Cantor model (JC69) (Jukes and Cantor, 1969) and the GTR model with a proportion of invariant sites and with Gamma-distributed mutation rates across sites (GTR+I+Γ) (Tavare, 1986). In particular, we use the Seq-Gen options -mGTR -r 1.0 0.2 10.0 0.75 3.2 1.6 -f 0.15 0.35 0.15 0.35 -i 0.2 -a 5.0 -g 3 to simulate under GTR+I+Γ. Because the theoretical results in Section 2.1 were derived under the GTR model and associated sub-models (such as JC69), we expect our method to handle the JC69 case well. However, we have not derived results under models in which there are invariant sites or rate variation among sites, so the simulations under the GTR+I+Γ setting will test robustness of the method to these evolutionary processes. In Step (4), we set $B = 100$.

We carry out one additional simulation to examine the ability of the method to identify the true split for varying overall data set sizes. We consider unlinked SNP data with 1,000, 5,000, or 10,000 sites ($g = 1,000, 5,000,$ or $10,000$ and $n = 1$ in all cases). We used the JC69 model and considered branch lengths of $x = 0.5, 1.0,$ and 2.0. We recorded the time it took to carry out each of these simulations in order to assess how computation time scales with the size of the data set.

2.3 Application to Rattlesnake Data

We have also explored the use of our quartet inference method in constructing larger species-level phylogenies, and we show here the results of applying the method to a data set consisting of 19 genes sampled in 26 rattlesnakes from 4 distinct species: *Sistrurus catenatus* (with subspecies *S. c. catenatus*, *S. c. edwardsii*, and *S. c. tergeminus*); *S. miliarius* (with subspecies *S. m. miliarius*, *S. m. barbouri*, and *S. m. streckeri*); and two outgroup species, *Agkistrodon contortrix* and *A. piscivorus*. This data set has been previously analyzed by Kubatko et al. (2011), and details concerning the loci used and the assembly of the aligned data

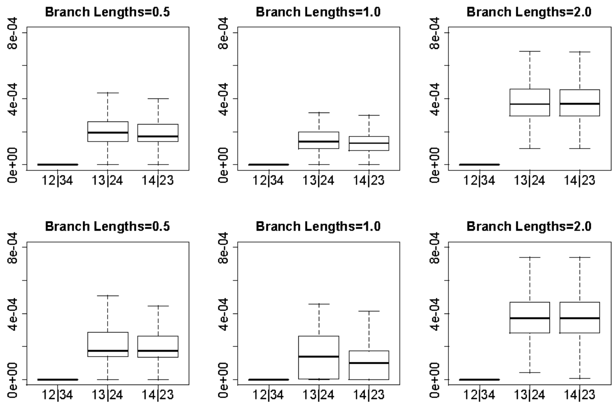

Fig. 3. Simulation results for the GTR+I+Γ model. The top row gives results for 5,000 unlinked SNP sites and the bottom row gives the results for 10 genes with 500 sites each. The columns correspond to differing branch lengths in the model species tree. The first boxplot in each subfigure shows the distribution of SVD scores for the true split, while the next two boxplots show the distribution for the two false splits.

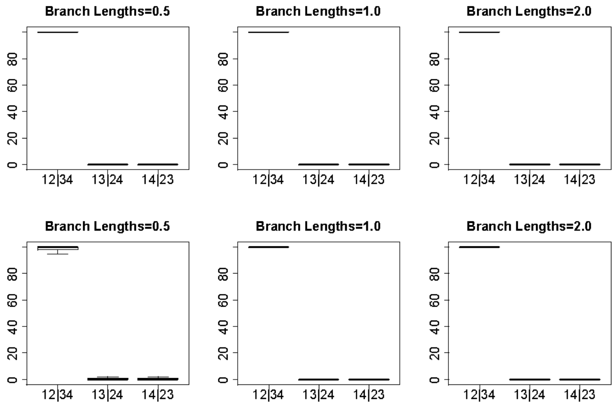

Fig. 4. Bootstrap results for the JC69 model simulations. Each boxplot shows the distribution of the bootstrap support values for each of the three possible splits for the simulated data shown in Figure 2.

matrix can be found there. Here we note that the sequences were computationally phased, so that each individual is represented by two distinct sequences in the data set, for a total of 52 sequences and 8,466 aligned nucleotide positions in the complete data matrix.

To conduct the analysis, we randomly sampled 20,000 quartets from the 52 sequences, and used the SVD score to infer the true quartet relationship for each sampled quartet. The quartet assembly program Quartet MaxCut (Snir and Rao, 2012) was used to construct phylogenies from the inferred quartets in two ways. First, a lineage tree was constructed by direct application of Quartet MaxCut. Second, a species-level phylogeny was constructed by replacing the labels of the lineages for the sampled quartets with the subspecies to which they belonged prior to application of

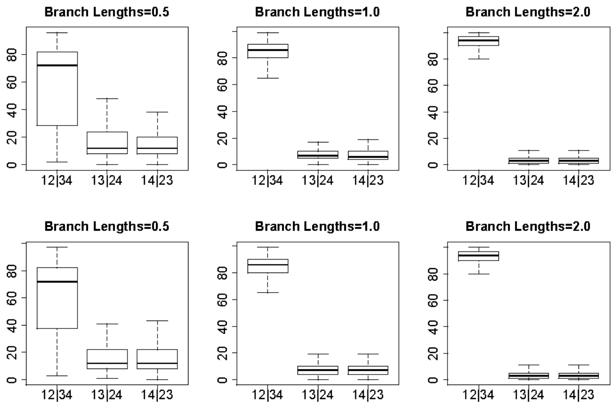

Fig. 5. Bootstrap results for the GTR+I+Γ simulations. Each boxplot shows the distribution of the bootstrap support values for each of the three possible splits for the simulated data shown in Figure 3.

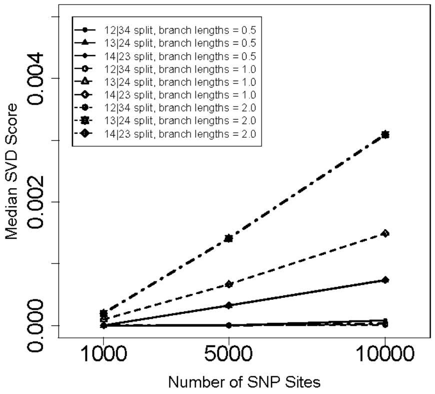

Fig. 6. Simulation results for data consisting of 1,000, 5,000, or 10,000 unlinked SNP sites for trees with branch lengths of 0.5 coalescent units (solid lines), 1.0 coalescent units (dashed lines), or 2.0 coalescent units (dotted lines). The median SVD score (over 1,000 replicates) for the valid split 12|34 is shown in red, while the scores for the two non-valid splits are shown in blue and green.

Quartet MaxCut. Finally, a bootstrap analysis was carried out by generating 100 bootstrapped data sets from the original matrix and applying this entire procedure to each bootstrapped data set. The complete analysis, including data simulation, bootstrapping and quartet assembly, took approximately 23 hours in serial on a desktop linux machine (2x Quad Core Xeon E5520 / 2.26GHz / 32GB).

2.4 Application to Soybean Data and Comparison to SNAPP

To demonstrate the utility of our method further, we used a previously published data set consisting of 17 wild soybean types (*Glycine soja*) and 14 cultivated soybean types (*G. max*) with 6,289,747 SNP loci. The original analysis was performed by Lam et al. (2010), and the data were later reanalyzed by Lee et al. (2014). We also carried out computations in SNAPP (Bryant et al., 2012), which is suitable for the soybean data set since it consists of SNP (rather than multilocus) data, to compare the run times. SNAPP infers the species tree using the coalescent model and is designed for biallelic data consisting of unlinked SNPs (Bryant et al., 2012). Even though our extended SVDquartets method to infer species trees can handle the entire data set including missing data, in order to make a proper and fair comparison with SNAPP we have removed all missing data and ambiguous sites, resulting in 1,027,026 SNP loci. We also subsampled 10 of the 31 species (4 cultivars and 6 wild types) in order to run the analysis in SNAPP in a feasible timeframe. The formatted data sets used for the analyses with SNAPP and SVDquartets are given in Supplemental Files 2 and 3, respectively. We conducted the analysis using SVDquartets in an analogous way to that for the rattlesnakes, with 20,000 quartets sampled and 100 bootstrap replicates.

3 IMPLEMENTATION

We have written a program in the C language, SVDquartets, which will compute $SVD(L_1|L_2)$ for the three possible splits in a sample of four taxa. The program takes as its input an alignment of four taxa in PHYLIP format, and produces a file that contains a list of the three splits and their associated scores. The program is available from http://www.stat.osu.edu/∼lkubatko/software/SVDquartets/.

4 RESULTS AND DISCUSSION

4.1 Simulation Study

Figures 2 and 3 show boxplots of the SVD scores for each of the three possible splits among four taxa under various simulation conditions. It is immediately clear that in all cases the SVD score can easily differentiate between the valid and non-valid splits, with the boxplot corresponding to the valid split displaying scores that are uniformly lower than the scores for the non-valid splits. The separation of scores for valid vs. non-valid splits becomes more pronounced as the branch lengths in the species tree increase, as expected, and is, in general, greater for the unlinked SNP data than for the multi-locus data, although the separation is very clear even for the multi-locus data.

Similarly, the JC69 model with no invariant sites and no rate variation across sites provides the best separation of scores between valid and non-valid splits. The worst performance observed was for the simulation conditions in which the data were simulated under GTR+I+Γ in the multi-locus setting, which is not unexpected as this violates the theoretical conditions in two ways (the invariant sites and variable rates across sites AND the multi-locus rather than unlinked SNP data). However, even in this case, the separation in scores is clear, and with sufficiently long species tree branch lengths, there is essentially no overlap in scores in valid vs. non-valid splits.

Figures 4 and 5 show boxplots of the bootstrap support values associated with each of the three splits under all simulation conditions. In the case of the JC69 model (Figure 4), the true split is nearly always associated with 100% bootstrap support for both unlinked SNP data and for multi-locus data. For data simulated

Table 1. Time information for the simulation study with results shown in Figure 6. All results represent the average time in seconds (over 1,000 replicates) to carry out the computation of three SVD scores for the simulated data sets, and were obtained using the UNIX `time` command.

Branch Lengths	Number of Sites	Real Time	User Time	System Time
0.5	1,000	0.0495	0.0092	0.0075
0.5	10,000	0.0566	0.0155	0.0077
1.0	1,000	0.0502	0.0105	0.0074
1.0	10,000	0.0564	0.0163	0.0076
2.0	1,000	0.0500	0.0119	0.0061
2.0	10,000	0.0553	0.0173	0.0064

under the GTR+I+Γ model, however, bootstrap support values for the true split are sometimes lower, with the worst results occurring when the branch lengths are short. Overall, however, the bootstrap appears to give a reliable measure of support for the true split, particularly when the model assumptions are satisfied.

Figure 6 examines the performance of the method for unlinked SNP data with varying numbers of sites. In particular, unlinked SNP data sets were generated with either 1,000, 5,000, or 10,000 total sites under model species trees with branch lengths of 0.5, 1.0, or 2.0 coalescent units. These results demonstrate that the method performs well even for smaller sample sizes. However, it is clear that as the sample size becomes larger, the separation between the scores for the valid and non-valid splits increases. This is to be expected, because the matrix $Flat_{L_1|L_2}(\hat{P})$ will better approximate $Flat_{L_1|L_2}(P)$ for larger sample sizes.

Table 1 gives timing results for the simulations carried out in Figure 6. Because the main work undertaken by the method involves counting the number of site patterns in order to build the $Flat_{L_1|L_2}(\hat{P})$ matrix, the time should be approximately linear in the number of unique site patterns in the data, which is related to both the total number of sites in the data matrix and the overall scale of time represented by the phylogeny. The results in Table 1 demonstrate that the time is less than linear in the total number of site patterns, as expected, and that the computations can be carried out very rapidly (e.g., the computation of three SVD scores for data matrices of 10,000 sites takes less than 0.1 seconds).

4.2 Potential Use for Species Tree Inference

These results make it clear that the SVD score is a highly accurate means of inferring the correct, unrooted phylogenetic tree among a set of four taxa. We note that the SVD score is extremely easy to compute. It requires only counting the site patterns and constructing the matrix $Flat_{L_1|L_2}(\hat{P})$. Computing singular values of a 16×16 matrix is a standard calculation that any mathematical or statistical software package can easily implement. Our software, SVDquartets, carries out both steps using a PHYLIP-formatted input file.

Given the efficiency with which computations can be carried out in the four-taxon setting, this method is a good candidate for estimation of species trees for larger taxon sets. We propose that the method could be used in the following way. For a data set with T taxa, form all samples of 4 taxa, or sample sets of 4 taxa if T is large. For each sample of four taxa, infer the valid split using the SVD score. Using the collection of inferred valid splits, construct a species tree estimate using a quartet assembly method. Substantial previous work and software exist for the problem of quartet assembly (see, e.g., Strimmer and von Haeseler (1996); Strimmer et al. (1997); Snir and Rao (2012)). We give the results of using this method for inferring a tree consisting of several North American rattlesnake species and for inferring a tree from SNP data for several soybean species below.

This method has tremendous potential to improve the set of tools available for species tree inference. Unlike summary statistics methods, which are known to be quick but fail to model variability in individual gene tree estimates, this method uses the sequence data directly, thus incorporating all sources of variability. The other existing methods based on sequence data (BEST, *BEAST, and SNAPP) all rely on Bayesian MCMC methods, and thus require long computing times and the difficult problem of assessing convergence. Our method can be carried out rapidly, and is easily parallelizable, as each quartet can be analyzed on a separate processor. Our method can handle both unlinked SNP and multi-locus data, again providing an advantage over existing sequence-based methods, which can handle either SNP (SNAPP) or multi-locus (BEST and *BEAST) data. Bootstrapping can be easily implemented to provide a means of quantifying support for the estimated phylogeny.

However, there are several issues with this method that will need to be examined in future work. First, the number of quartets to be sampled needs to be specified in cases where the number of taxa is too large to examine all possible quartets. This number should necessarily increase with increasingly large taxon samples, but we have not yet rigorously examined how to select this. In addition, it is worth pointing out that the method estimates the topology only. In some studies, other parameters associated with the evolutionary process, such as branch lengths and effective population sizes, will also be of interest. One possibility is that the tree topology could first be estimated with this method, and then fixed in a subsequent MCMC analysis with either *BEAST or SNAPP, thus greatly reducing the complexity of that analysis. Finally, we have not yet conducted a thorough simulation study of the inferential accuracy of this method for full species tree inference, which will be the topic of future work.

4.3 Application to Rattlesnake Data

The results of the analysis of the rattlesnake data set are shown in Figure 7, with bootstrap support values above 50% indicated on the appropriate nodes. In the case of the lineage tree (Figure 7(a)), the method identifies the two major species *S. catenatus* and *S. miliarius* with high bootstrap support, and additionally groups the subspecies *S. c. catenatus* as monophyletic. In the species tree in Figure 7(b), we again see that the method correctly identifies the two species with high bootstrap support, and is able to differentiate subspecies *S. c. catenatus* from a clade containing the other two subspecies within this group. Within species *S. miliarius*, there is not strong support for the subspecies relationships.

These results are consistent with the earlier analyses of Kubatko et al. (2011), in which strong support for the delimitation of *S. c. catenatus* as a distinct species was found using several methods of coalescent-based species tree inference. Those analyses also found

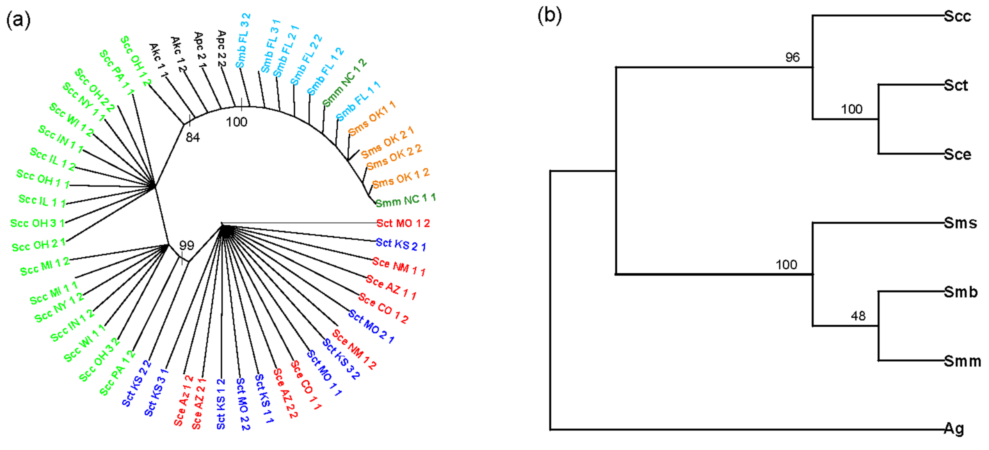

Fig. 7. Results of the analysis of the rattlesnake data. In (a), the tree relating all 52 lineages is shown. Colors indicate subspecies membership: Scc = *S. c. catenatus* (green); Sce = *S. c. edwardsii* (red); Sct = *S. c. tergeminus* (blue); Smm = *S. m. miliarius* (dark green); Sms = *S. m. streckeri* (orange); Smb = *S. m. barbouri* (dark blue); Apc = *A. piscivorus* (black); Akc = *A. contortrix* (black). In (b), the tree relating subspecies is shown, with abbreviations as above, except that the two outgroup species have been combined and denoted "Ag". In both subfigures, numbers above the nodes refer to bootstrap support values, and the trees depicted are majority-rule consensus trees over 100 bootstrap samples.

a general lack of resolution among the three subspecies within the *S. miliarius* clade, which again is consistent with the results observed here. While the results of the analysis using our new method are consistent with those of previous methods, there were important differences in the time required by the two methods. For example, the *BEAST analysis in Kubatko et al. (2011) took approximately 10 days to run, and even after this extensive run time, there was evidence that the effective population size parameter estimates had not converged. In contrast, our method took less than 1 hour to get the initial species tree estimate, and less than 1 day to analyze 100 bootstrap replicates in serial on a desktop linux machine; if the 100 bootstrap analyses were run in parallel, the total computing time could be cut to less than 1 hour.

4.4 Application to Soybean Data

The results of the analysis of the soybean data using both SNAPP and SVDquartets are shown in Figure 8. The SNAPP analysis was run for 2.239 million iterations, corresponding to 28 days on a desktop linux machine (2x Quad Core Xeon E5520 / 2.26GHz / 32GB). There were important indications of a lack of convergence of the method, with nearly all effective sample size (ESS) values below 200 and trace plots indicating issues in convergence. The full details of the analysis and assessment of convergence are described in the Supplemental Information. The SVDquartets method with 100 bootstrap samples and 20,000 quartets sampled per replicate required approximately 600 hours (which corresponds to 25 days) of time to complete using the same desktop linux machine, though it was run in parallel using 6 processors and thus took only 4.5 days to complete. We note that this can easily be parallelized further, with the only limits due to availability of processors.

Even though we have subsampled and filtered the original data set, our results are in agreement with the findings of the original report (Lam et al., 2010). In their analyses they found that cultivated soybeans formed a tight subclade. Furthermore, they concluded using the Bayesian clustering program STRUCTURE and PCA analysis that C01 and C12 show a clear separation from the cultivated cluster. Also, the phylogenetic tree in Lam et al. (2010) has cultivars as a part of the clade that includes wild type soybeans W08, W10, and W15, while W07, W12 and W14 are part of another cluster. One can see that the results in Figure 8 for both trees are in general consistent with the previous findings. Of course, there are important differences between the trees, as well.

5 CONCLUSION

We have presented a method to reliably infer the valid split in a set of four taxa. We have demonstrated that the method performs very well over a range of simulation conditions. The method can be easily extended for use in inferring species phylogenies in larger taxon samples, as demonstrated by our applications to the rattlesnake data and to the soybean data. The method thus makes a valuable contribution to the collection of methods for inferring species-level phylogenetic trees under the coalescent model for either multi-locus or unlinked SNP data.

ACKNOWLEDGEMENTS

We thank David Bryant and Remco Bouckaert for assistance in running the SNAPP analysis. We thank three anonymous reviewers

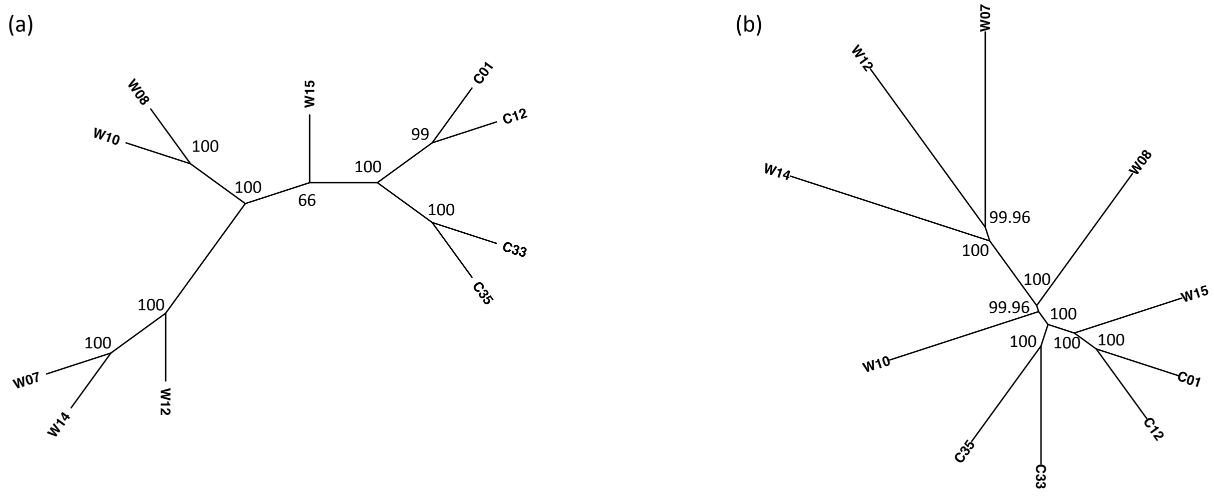

Fig. 8. Results of the analysis of the soybean data. (a) Tree estimated by SVDquartets with bootstrap support values. (b) Maximum clade credibility tree estimated using SNAPP.

and Associate Editor David Posada for helpful comments on earlier versions of this manuscript.

FUNDING

The authors are supported in part by National Science Foundation award DMS-1106706 and in part by NIH Cancer Biology Training Grant T32-CA079448 (J.C.).

REFERENCES

Allman, E. S. and Rhodes, J. A. (2008) Phylogenetic ideals and varieties for the general Markov model. Adv. in Appl. Math., **40 (2)**. arXiv:math.AG/0410604.

Bryant, D., Bouckaert, R., Felsenstein, J., Rosenberg, N. and RoyChoudhury, A. (2012) Inferring species trees directly from biallelic genetic markers: bypassing gene trees in a full coalescent analysis. *Mol. Biol. Evol.*, **29 (8)**, 1917-1932.

Chifman, J. and Kubatko, L. S. (2014) Identifiability of the unrooted species tree topology under the coalescent model with time-reversible substitution processes, submitted; available at http://arxiv.org/abs/1406.4811 .

Degnan, J and Salter, L. (2005) Gene tree distributions under the coalescent process. *Evolution* **59(1)**, 24-37.

DeGeorgio, M. and Degnan, J. (2010) Fast and consistent estimation of species trees using supermatrix rooted triples. *Mol. Biol. Evol.* **27(3)**, 552-569.

Eriksson, N. (2005) Tree construction using singular value decompsition, in L. Pachter and B. Sturmfels, editors, Algebraic Statistics for Computational Biology, chapter 19, pgs. 347–358. Cambridge University Press, Cambridge, UK.

Golub, G. H. and Van Loan, C. F. (2013), Matrix Computations (4th Edition), *Johns Hopkins University Press*.

Heled, J. and Drummond, A. (2010) Bayesian inference of species trees from multilocus data. *Mol. Biol. Evol.*, **27**, 570-580.

Huang, H., He, Q., Kubatko, L., and Knowles, L. (2010) Sources of error for species-tree estimation: Impact of mutational and coalescent effects on accuracy and implications for choosing among different methods. *Syst. Biol.*, **59(5)**, 573-583.

Jukes, T. and Cantor, C. (1969) Evolution of Protein Molecules. New York: Academic Press, pp. 21-132.

Kingman,J.F.C. (1982) The coalescent. *Stoch. Proc. Appl.*, 13, 235-248.

Kingman, J. F. C. (1982) Exchangeability and the evolution of large populations. Pp. 97112 in G. Koch and F. Spizzichino, eds. Exchangeability in probability and statistics. North-Holland, Amsterdam.

Kubatko, L. S., Carstens, B. C., and Knowles, L. L. (2009) STEM: Species tree estimation using maximum likelihood for gene trees under the coalescent. *Bioinformatics,* **25**, 971-973.

Kubatko, L. S., Gibbs, H. L., and Bloomquist, E. W. (2011) Inferring species-level phylogenies and taxonomic distinctiveness using multilocus data in *Sistrurus* rattlesnakes. *Syst. Biol.*, **60**, 393-409.

Lam, H. M., et al. (2010) Resequencing of 31 wild and cultivated soybean genomes identifies patterns of genetic diversity and selection, *Nat. Genet.*, **42**, 1053-1059.

Lee, T. H., Guo, H., Wang, X., Kim, C., and Paterson, A. H. (2014) SNPhylo: a pipeline to construct a phylogenetic tree from huge SNP data. *BMC Genomics*, **15(1)**.

Liu, L. and Pearl, D. (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., Yu, L., Kubatko, L., Pearl, D., and Edwards, S. (2009) Coalescent methods for estimating phylogenetic trees. *Mol. Phylogenet. Evol.* **52**, 320-328.

Liu, L., Yu L., and Pearl D. K. (2009) Maximum tree: a consistent estimator of the species tree. *J. Math. Biol.,* **60**, 95-106.

Liu, L., Yu, L., Pearl, D. and Edwards, S. (2009) Estimating species phylogenies using coalescence times among sequences. *Syst. Biol.*, **58(5)**, 468-477.

Liu, L., Yu, L., and Edwards, S. (2010) A maximum pseudo-likelihood approach for estimating species trees under the coalescent model. *BMC Evol. Biol.*, **10**, 302.

Maddison,W. (1997) Gene trees in species trees. *Syst. Biol.*, **46**, 523-536.

Rambaut, A. and Grassly, N. C. (1997) Seq-Gen: An application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput. Appl. Biosci.* **13**, 235-238.

Rannala, B. and Yang, Z. (2003) Bayes estimation of species divergence times and ancestral population sizes using DNA sequences from multiple loci. *Genetics* **164**, 1645-1656.

Snir, S. and S. Rao. (2012) Quartet MaxCut: A fast algorithm for amalgamating quartet trees. *Mol. Phylogen. Evol.* **62**:,1-8.

Strimmer, K., and von Haeseler, A. (1996) Quartet puzzling: A quartet maximum likelihood method for reconstructing tree topologies. *Mol. Biol. Evol.* **13**, 964-969.

Strimmer, K., Goldman, N., and von Haeseler, A. (1997) Bayesian probabilities and quartet puzzling. *Mol. Biol. Evol.* **14**, 210-213.

Tavare, S. (1986) Some probabilistic and statistical problems in the analysis of DNA sequences, *Lectures on Mathematics in the Life Sciences (American Mathematical Society)* **17**, 57-86.

Than, C. and Nakhleh, L. (2009) Species tree inference by minimizing deep coalescences. *PLoS Computational Biology,* **5(9)**, e1000501.