Phylogenetic Analysis Based on Spectral Methods

Melanie Abeysundera,* Chris Field, and Hong Gu

Department of Mathematics and Statistics, Dalhousie University, Halifax, Nova Scotia, Canada

***Corresponding author:** E-mail: amelanie@mathstat.dal.ca.

Associate editor: Sudhir Kumar

Abstract

Whole-genome or multiple gene phylogenetic analysis is of interest since single gene analysis often results in poorly resolved trees. Here, the use of spectral techniques for analyzing multigene data sets is explored. The protein sequences are treated as categorical time series, and a measure of similarity between a pair of sequences, the spectral covariance, is based on the common periodicity between these two sequences. Unlike the other methods, the spectral covariance method focuses on the relationship between the sites of genetic sequences. By properly scaling the dissimilarity measures derived from different genes between a pair of species, we can use the mean of these scaled dissimilarity measures as a summary statistic to measure the taxonomic distances across multiple genes.

The methods are applied to three different data sets, one noncontroversial and two with some dispute over the correct placement of the taxa in the tree. Trees are constructed using two distance-based methods, BIONJ and FITCH. A variation of block bootstrap sampling method is used for inference. The methods are able to recover all major clades in the corresponding reference trees with moderate to high bootstrap support.

Through simulations, we show that the covariance-based methods effectively capture phylogenetic signal even when structural information is not fully retained. Comparisons of simulation results with the bootstrap permutation results indicate that the covariance-based methods are fairly robust under perturbations in sequence similarity but more sensitive to perturbations in structural similarity.

Key words: multigene analysis, multivariate time series, spectral analysis, protein structure, phylogenetic trees.

Introduction

Maximum likelihood (ML)-based methods of tree estimation assume that sequence sites evolve independently as a Markov process, based on a prespecified evolutionary model, which allows the computation of the transition probabilities at a given site (Felsenstein 2003). It is generally accepted that there is a dependence among the sites (Philippe, Delsuc, et al. 2005). A novel approach to address the dependence among sites in phylogenetic analysis was considered by Collins et al. (2006), where a spectral envelope-based covariance method was developed to estimate trees. The spectral envelope was first introduced by Stoffer et al. (1993) as a method of analyzing categorical time series in the frequency domain. The spectral envelope provides an automated method of scaling qualitative time series data to emphasize the strongest periodic signal in a sequence. Since high peaks in the sample spectral density correspond to periodic structure in a time series, choosing scalings which maximize the spectrum should highlight any periodic features present in the data. Thus, scalings are chosen to maximize the variance at each frequency relative to the overall variance of the data. Collins et al. (2006) extended these analyses to amino acid sequences and found that the peaks in the spectral envelope of protein sequences correspond to the folding patterns of the secondary structures of a protein. The spectral covariance used by Collins et al. (2006) as a measure of sequence similarity is a nonstandardized adaptation of the spectral envelope approach to coherency presented in Stoffer et al. (2000). Collins et al. (2006) compared their results with those obtained using standard likelihood methods and found that it yielded similar tree estimates to the ML approach when the method was applied to a single protein. This was a remarkable result as the two techniques are based on completely different criteria. Although the ML method of tree estimation relies heavily on the assumed evolutionary model, the spectral covariance method of sequence comparison does not assume any particular evolutionary model but instead is a distance method based on spectral analysis. In this aspect, the ML method can be thought of as parametric, whereas the spectral covariance-based method can be viewed as nonparametric. Unlike the ML method which assumes site independence, the spectral covariance takes into account correlations among sequence sites. Since prominent peaks in the spectral covariance correspond to common periodicities in the individual sequences, the spectral covariance, although sequence based, is a measure of structural similarity (Collins et al. 2006).

In this paper, we extend these analyses to multigene data sets and explore two different methods for combining information from multiple genes to obtain tree estimates. Note that the spectral methods applied to phylogenetic data in this paper differ from those introduced by Hendy and Penny (1993). The spectrum defined in Hendy and Penny is a list of counts of possible bipartitions over each site, representing the support for

each split in the data, whereas here, the spectrum is the fast Fourier transform of a time series representation of the individual amino acid sequences. Whole-genome or multiple gene analysis is of interest since single gene analysis often results in poorly resolved trees. Indeed, the small number of sites in a single gene tends to lead to a relatively high level of variation in the estimation of trees (Rokas et al. 2003; Philippe, Delsuc, et al. 2005). The question of how to combine the information present in individual genes has been the subject of extensive study and debate from which there have emerged several approaches to the analysis of multigene data sets (Bull et al. 1993; Bininda-Emonds 2004; Gatesy et al. 2004; Philippe, Delsuc, et al. 2005; Burleigh et al. 2006; de Queiroz and Gatesy 2007). The most widely used approach is to concatenate the alignments of individual genes and then apply standard likelihood or distance-based methods on the concatenated sequences to derive a single representative topology for multiple genes. Another approach is to analyze the genes individually and then obtain a single tree estimate by consensus (Baum 1992; de Queiroz 1993; Miyamoto and Fitch 1995). Many have suggested that genes should be combined conditional on their sharing similar evolutionary histories. To achieve this, a test for congruence is performed, and only those genes deemed to have common evolutionary histories are combined using concatenation or consensus methods (Bull et al. 1993; Farris et al. 1995; Zelwer and Daubin 2004; Lecointre 2005; Leigh et al. 2008). The concatenation approach has the advantage of using all available sequence information and can sometimes reveal relationships between taxa, which are hidden in a separate analysis (de Queiroz and Gatesy 2007). Furthermore, concatenation is supposed to reduce the stochastic error (Jeffroy et al. 2006). However, the concatenation approach implicitly assumes that all genes share a common evolutionary history, and it may return incorrect estimates of the underlying species tree when this assumption is violated. Different genes may evolve under different models; hence, concatenation may also lead to model misspecification (Philippe, Delsuc, et al. 2005; Jeffroy et al. 2006).

Since applying the spectral covariance on a concatenation does assume a similar dependence structure among genes, which may not necessarily be true, performing a separate spectral analysis on individual genes and then combining them seems more sensible. In our approach, spectral covariance-based dissimilarity matrices are computed for the individual genes and then combined to obtain a summary measure of the dissimilarity matrix. The goal of the combination is to find a single dissimilarity matrix, which best summarizes the information present in multiple genes. Two different scaling methods are proposed in this paper to scale dissimilarity matrices, so that the mean of these scaled dissimilarities can be used as a summary measure of the dissimilarity for each pair of taxa. In these methods, each dissimilarity matrix from a gene is given a single scale coefficient. This gene-specific scale coefficient reflects a gene's specific evolutionary rate and makes the branch lengths computed from the scaled dissimilarity matrices comparable.

Materials and Methods

The Spectral Covariance

The spectral envelope of a categorical times series and its application to problems in molecular biology was first introduced by Stoffer et al. (1993). The spectral covariance, introduced by Collins et al. (2006), is a spectral envelope-based measure of similarity between two sequences. It is a nonstandardized adaptation of the spectral envelope approach to coherency presented in Stoffer et al. (2000). The spectral covariance is the smoothed Fourier transform of the cross-spectra between two sequences. A high covariance at a given frequency signifies a common periodicity between two sequences at that frequency.

A DNA or amino acid sequence can be treated as a categorical time series and can be transformed into a numerical time series by assigning a numerical value to each letter in the sequence. Let X_t, $t = 0, \pm 1, \pm 2, \ldots$, be a categorical time series with finite state space $C = \{c_1, c_2, \ldots, c_k\}$. For $\beta = (\beta_1, \beta_2, \ldots, \beta_k)' \in \mathbb{R}^k$, denote $X_t(\beta)$ as the real-valued time series corresponding to the scaling that assigns c_j the value β_j. The categorical time series X_t can be expressed as a multivariate time series Y_t, where $Y_t = e_j$ whenever $X_t = c_j$ and e_j is an index vector with one in the jth column and zeros elsewhere. The real-valued time series $X_t(\beta)$ is related to the multivariate time series Y_t by $X_t(\beta) = Y_t\beta$. The periodicity of this time series will depend on the choice for β. The spectral covariance method chooses scalings, which maximize the squared covariance between two sequences at each frequency. Following the same notation, denote the multivariate time series of categorical time sequence X_{1t} as Y_{1t} and that of categorical time sequence X_{2t} as Y_{2t}. Scalings $\alpha(\omega)$ and $\beta(\omega)$ at frequency ω are chosen to maximize the squared spectral covariance

$$\mathrm{Cov}_{12}^2(\omega) = \sup_{\alpha,\beta} |\alpha'(\omega) f_{12}(\omega) \beta(\omega)|^2, \qquad (1)$$

where f_{12} is the cross-spectral density between Y_{1t} and Y_{2t}, and $\alpha(\omega)$ and $\beta(\omega)$ are subject to the condition $\alpha'(\omega)\alpha(\omega)=1$ and $\beta'(\omega)\beta(\omega)=1$. This normalization is necessary to ensure that the covariance does not infinitely increase. The cross-spectral density is the smoothed cross-periodogram between two multivariate time series and is defined by

$$f_{12}(\omega) = \frac{1}{2\pi} \sum_{k=-\infty}^{\infty} R_{12}(k) e^{-i\omega k},$$

where $R_{12}(k) = \mathrm{Cov}(Y_{1t}, Y_{2(t+k)})$ is the cross-covariance of $\{Y_{1t}, Y_{2t}\}$ (Priestly 1981). Peaks in the cross-spectral density of two univariate time series represent periodicities common to them. Since the value of the squared spectral covariance at each ω depends on the choice of scalings, the scalings α and β are chosen such that the squared spectral covariance at each frequency ω attains the maximum possible value. Note that the squared spectral covariance in equation (1) can be rewritten as

$$\mathrm{Cov}_{12}^2(\omega) = \sup_{\alpha,\beta}\left\{\left[\alpha'(\omega)f_{12}^{re}(\omega)\beta(\omega)\right]^2 + \left[\alpha'(\omega)f_{12}^{im}(\omega)\beta(\omega)t\right]^2\right\}, \quad (2)$$

where f_{12}^{re} and f_{12}^{im} are the real and imaginary parts of f_{12}.

In this paper, we focus on two methods of computing spectral covariance scalings by imposing two different constraints on α and β, namely, the common scaling method and the taxa-specific scaling method. In the common scaling method, each pair of taxa are assumed to have a common scaling. That is, the scalings for "taxon1," α and "taxon2," β are assumed to be the same when they are compared with each other. Since amino acid sequences share the common alphabets and thus have the same state space, it is reasonable to apply the same scalings to both sequences to enhance the interpretability and reduce the variance of the results (Collins et al. 2006). Although in the common scaling method, the set of scalings for any given sequence depends upon the sequence with which it is being compared, the taxa-specific scaling assumes each taxon has only one set of scalings. That is, taxon1 will have the same set of scalings regardless of whether it is being compared with taxon2 or "taxon3." The taxa-specific scaling covariances reflect the relationship between pairs of taxa relative to all the other taxa in the tree. For the data analyzed in this paper, the common scaling covariance and the taxa-specific covariance methods yield very similar results.

Note that although we work with aligned sequences in this paper, the spectral covariance method does not require all sequences be aligned. Sequences may be made the same length by cutting the longer sequence to the size of the shorter. Another option might be to use pairwise alignments rather than multiple alignments.

The Common Scaling Spectral Covariance

It can be shown that when state spaces are the same and the spectral density matrix is symmetric, the maximum covariance is achieved when scalings $\alpha = \beta$ (Stoffer et al. 2000). By applying a common scaling to the two sequences being compared, we reduce the number of parameters in the model, thereby reducing the complexity of the method and increasing the precision of our estimates. For simplicity, ω is considered fixed and dropped from the notation. With X_{1t}, X_{2t}, Y_{1t}, and Y_{2t} defined as above, the squared spectral covariance in equation (2) is now

$$\mathrm{Cov}_{12}^2 = \sup_{\beta}|\beta' f_{12}\beta|^2, \quad (3)$$

subject to $\beta'\beta$=1, where f_{12} is the cross-spectral density between Y_{1t} and Y_{2t}. Equation (3) can be rewritten as

$$\mathrm{Cov}_{12}^2 = \sup_{\beta'\beta=1}\left(\left[\beta' f_{12}^{re}\beta\right]^2 + \left[\beta' f_{12}^{im}\beta\right]^2\right). \quad (4)$$

Since f_{12}^{re} and f_{12}^{im} are not usually symmetric, to make them symmetric, we define matrices

$$A^{re} = \left[f_{12}^{re} + f_{12}^{re'}\right]/2,$$
$$A^{im} = \left[f_{12}^{im} + f_{12}^{im'}\right]/2.$$

Equation (4) then becomes

$$\mathrm{Cov}_{12}^2 = \sup_{\beta'\beta=1}([\beta' A^{re}\beta]^2 + [\beta' A^{im}\beta]^2). \quad (5)$$

The algorithm to compute the common scaling β is given below:

1. Initialization: set β to be one of the following:
$$\beta_1 = \varepsilon_1(A^{re'}A^{re}),$$
$$\beta_2 = \varepsilon_1(A^{im'}A^{im}),$$
whichever produces the larger initial estimate of the spectral covariance. ε_1 denotes the eigenvector corresponding to the largest eigenvalue of the matrix in the brackets. The initial squared covariance is then
$$\mathrm{Cov}_{12}^2 = (\beta_0' A^{re}\beta_0)^2 + (\beta_0' A^{im}\beta_0)^2.$$

2. Iteratively calculate scalings using
$$\beta_j = \varepsilon_1(A^{re}\beta_{j-1}\beta_{j-1}' A^{re} + A^{im}\beta_{j-1}\beta_{j-1}' A^{im}) \quad (6)$$
until convergence. Convergence criteria is set as $||\beta_j - \beta_{j-1}||^2 < 0.001$.

Taxa-Specific Scaling Spectral Covariance

For the common scaling spectral covariance, the scalings for any given taxa are dependent upon the taxa with which it is being compared. For example, under the common scaling spectral covariance, the honeybee may be assigned one set of scalings when it is compared with the locust and a different set of scalings when it is compared with the nematode. Another approach to assigning scalings to taxa is to hold the scalings corresponding to each taxa in a given data set constant across all pairwise comparisons. To compute the taxa-specific scaling spectral covariance, the following criterion is used. Following the notation above, for K taxa, denote the multivariate series of K sequences $X_{1t}, \ldots, X_{Kt}$ as $Y_{1t}, \ldots, Y_{Kt}$. The squared spectral covariance is now

$$\sum_{i<j}\mathrm{Cov}_{ij}^2 = \sup_{\beta_i,\ldots,\beta_k}\sum_{i<j}|\beta_i' f_{ij}\beta_j|^2, \quad (7)$$

subject to $\beta_i'\beta_i = 1$ for $i = 1, \ldots, K$, where f_{ij} is the cross-spectral density between Y_{it} and Y_{jt}.

To find β_i's which maximize equation (7), begin by initializing β_i^0, $(i = 1, \ldots, K)$ as the spectral envelope scaling of the ith sequence. Then the algorithm is as follows:

For $i = 1, 2, \ldots, K$, iteratively calculate scalings using formula

$$\beta_i = \varepsilon_1 \sum_{j\neq i}[(f_{ij}^{re}\beta_j\beta_j' f_{ij}^{re'}) + (f_{ij}^{im}\beta_j\beta_j' f_{ij}^{im'})],$$

where ε_1 is the eigenvector corresponding to the largest eigenvalue of the given matrix. Convergence criterion is $\sum_{i=1}^{K}(||\beta_i^r - \beta_i^{r-1}||^2) < 0.001 * K$.

Dissimilarity Matrix Based on the Spectral Covariance

To build a spectral covariance-based phylogenetic tree, the spectral covariance measure must be transformed into a dissimilarity measure. The first step is to compute the spectral covariance at each frequency for each pair of sequences. The sum of the spectral covariance values above a threshold, which is used for reducing noise, is then taken to obtain a single numeric measure of similarity between the two sequences, hereafter referred to as the total covariance. The total covariance between the ith and jth sequences, denoted as $\mathrm{sim}(x_i, x_j)$, is then converted into a dissimilarity measure between the ith and jth sequences using the following definition:

$$\mathrm{diss}(x_i, x_j) = 1 - \frac{\mathrm{sim}(x_i, x_j)}{\max_{i \neq j}(\mathrm{sim}(x_i, x_j))},$$

for $i, j = 1, \ldots, n$, where n is the number of sequences.

The threshold is based on the empirical distribution of 1,000 bootstrap samples. The samples are obtained as follows: Two sequences are randomly selected from within the data set, and characters are randomly selected with replacement from these two sequences to obtain two sample sequences with the same length as the original pair. This is repeated until 1,000 sample covariances are obtained from 1,000 sequence pairs. The mean of the 95th quantiles of the sample covariances at each frequency is then taken to be the threshold. Applying the threshold should remove the random noise in the spectral covariance and thus ensure strong signals for similarity between sequences are taken into account by the total covariance statistic.

The dissimilarity matrices for all genes are then combined by the methods described below to generate the average dissimilarity matrix.

Combining Dissimilarity Measures across Genes

One simple way to combine dissimilarity measures across genes would be to take an average of the dissimilarity matrices. However, the dissimilarity matrices for different genes are not necessarily on the same scale. This is generally true for any distance-based method. Therefore, rather than taking the mean directly, a weighted average is used where each matrix is weighted by a scale coefficient. The mean of the scaled dissimilarity matrices is then used as the combined dissimilarity matrix for the phylogenetic analysis. We next present two criteria for computing scale coefficients, which are generally useful for combining information across genes, namely the minimum variance (MinVar) and the minimum squared coefficient of variation (MinCV). Beven et al. (2005) use a weighted least squares approach to estimate the evolutionary rates of individual proteins and thereby estimate a representative distance for each taxa pair from multiple genes. In their method, estimated distances are weighted according to their level of uncertainty. The weights are based on a given substitution model (Bulmer 1991). In the MinVar and MinCV methods presented below, scales are chosen to minimize the variance in the pairwise distances across genes, and then, a weighted average across genes is taken as the representative distance for each pair of taxa. No evolutionary model is assumed in the computation of the weights.

The MinVar Scale Coefficients

Fixing the scale coefficient for one of the matrices as one, the scale coefficients for the other matrices are obtained by minimizing the sum of the variances of the pairwise dissimilarities across genes. For a data set with k genes and n taxa, the dissimilarity matrices for the k genes are combined as follows:

1. For each gene, organize the dissimilarity measures for all pairs of taxa as a p vector, where $p = \binom{n}{2}$ for n taxa. We combine the dissimilarities from all k genes into a single matrix

$$D = \begin{pmatrix} d_{1,1} & d_{1,2} & \ldots & d_{1,k} \\ d_{2,1} & d_{2,2} & \ldots & d_{2,k} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ d_{p,1} & d_{p,2} & \ldots & d_{p,k} \end{pmatrix},$$

   where each column corresponds to all the dissimilarities from a specific gene. For example, $d_{i,j}$ is the dissimilarity of the ith pair for the jth gene, $i = 1, \ldots, p, j = 1, \ldots, k$.

2. Let $c = (c_1, c_2, \ldots, c_k)$ be the scale coefficients for k genes. Fix $c_1 = 1$. The scaled dissimilarities are then

$$D_s = D \times \mathrm{diag}(c) = \begin{pmatrix} c_1 d_{1,1} & c_2 d_{1,2} & \ldots & c_k d_{1,k} \\ c_1 d_{2,1} & c_2 d_{2,2} & \ldots & c_k d_{2,k} \\ \ldots & \ldots & \ldots & \ldots \\ \ldots & \ldots & \ldots & \ldots \\ c_1 d_{p,1} & c_2 d_{p,2} & \ldots & c_k d_{p,k} \end{pmatrix}.$$

3. The optimal scalings $c = (c_1, c_2, \ldots, c_k)$ are those that minimize the sum of the variances of each pairwise dissimilarity across the k genes:

$$V = \sum_{i=1}^{p} V_i = \sum_{i=1}^{p} \left[\frac{1}{k} \sum_{j=1}^{k} (c_j d_{i,j})^2 - \left(\frac{1}{k} \sum_{j=1}^{k} c_j d_{i,j} \right)^2 \right]. \quad (8)$$

   This minimization problem can be solved analytically. The analytical solution is the solution to the linear system of equations $\frac{\partial V}{\partial c_m} = 0, m = 2, \ldots, k$, where

$$\frac{\partial V}{\partial c_m} \propto 2 c_m \left(\sum_{i=1}^{p} d_{i,m}^2 \right) - \frac{2}{k} \sum_{j=1}^{k} c_j \left(\sum_{i=1}^{p} d_{i,j} d_{i,m} \right).$$

4. The combined pairwise dissimilarities from the k genes is then the mean of the scaled dissimilarities, $\frac{1}{k} D_s 1$, where $1' = (1, 1, \ldots, 1)_{1 \times k}$.

The MinCV Scale Coefficients

An alternative method is to minimize the squared coefficient of variation (CV). Because larger dissimilarities usually have larger variances than smaller dissimilarities, a variance-based criterion like the MinVar may result in scale coefficients that are biased in favor of minimizing the variances of taxa pairs with larger dissimilarities, resulting in an incorrect estimate of topology locally for the taxa, which are close to each other. For the MinCV, the variances are scaled by the square of the mean. Hence, the scale coefficients determined with the MinCV will tend to avoid such bias as that from the MinVar method. In addition, the CV is unitless. Using the same notation as above, instead of minimizing equation (8), we now wish to minimize the sum of the squared CV:

$$\sum_{i=1}^{p} \mathrm{CV}_i^2 = \sum_{i=1}^{p} \left(\frac{\frac{1}{k}\sum_{j=1}^{k}(c_j d_{i,j})^2 - \left(\frac{1}{k}\sum_{j=1}^{k} c_j d_{i,j}\right)^2}{\left(\frac{1}{k}\sum_{j=1}^{k} c_j d_{i,j}\right)^2} \right) \quad (9)$$

$$\propto \sum_{i=1}^{p} \left(\frac{\sum_{j=1}^{k}(c_j d_{i,j})^2}{\left(\sum_{j=1}^{k} c_j d_{i,j}\right)^2} \right) - \frac{p}{k}. \quad (10)$$

Since this minimization problem cannot be solved analytically, we solve it using a numerical method instead. We start by setting the scale coefficients as the MinVar scale coefficients. We then use the nonlinear minimization function nlm() available in the R package nlme (Pinheiro et al. 2009) to find the set of scale coefficients $c = (c_1, c_2, \ldots, c_k)$ (with $c_1 = 1$) that minimizes equation (9). The combined pairwise dissimilarities for the k genes is then the mean of the scaled dissimilarities, $\frac{1}{k} D_s 1$, where $1' = (1, 1, \ldots, 1)_{1\times k}$.

Methods to Build Trees

To obtain phylogenetic trees from our combined dissimilarity matrix across genes, two distance-based tree building methods, that is, BIONJ (Gascuel 1997) and the Fitch–Margoliash least squares method implemented in the program FITCH in PHYLIP (Fitch and Margoliash 1967), are applied. The Neighbor-Joining (NJ) algorithm, first introduced by Saitou and Nei (1987) and revised by Studier and Keppler (1988), is an agglomerative clustering algorithm based on the principle of minimum evolution. BIONJ is a modified version of the NJ algorithm, which has been shown to return trees closer to the minimum evolution tree (Gascuel 1997). Fitch and Margoliash (1967) used a weighted least squares criterion to find an optimal tree. Since greater distances are more liable to have larger random errors associated with them, larger distances are given smaller weights in the FITCH method (Felsenstein 2003). This method sometimes performs slightly better than the NJ algorithm but has a greater computational cost since it involves searching the entire tree space (Kuhner and Felsenstein 1994).

Bootstrap Sampling Methods

To obtain an empirical distribution of the spectral covariance-based dissimilarity, we use a resampling method that maintains some of the structural information present in the data. Since the spectral covariance assumes a dependence structure between individual sites of a protein sequence, the chosen method must also preserve the dependence structure present in the original sequences. We use a variation of the block sampling method introduced by Kunsch (1989). Instead of sampling blocks with replacement, the blocks are sampled without replacement to obtain 100 permutation samples. This is equivalent to randomly selecting 100 permutations from the $b!$ possible permutations of blocks, where b is the total number of blocks. As the spectral covariance method of comparing sequences is based on the periodicity inherent in protein structures, an appropriate block size is determined using information known about the periodicity of these protein structures. It is known that α-helices have a periodicity of 3.6 residues, β-strands have a periodicity of 2.3 residues, and 3_{10}-helices have a periodicity of 2.5–3 residues. Although the length of loops can vary, it is known that turns have a periodicity of 3–4 residues. Motifs within a protein are comprised of helices and strands connected by loops and turns. The periodicity of these repeated motifs is known to be 8–14 residues in length (Collins et al. 2006). Hence, a block size of 14 is used to ensure as much structural information as possible was retained in the bootstrap permutation samples.

To quantify the variation of our estimated trees, we use two different distance measures for tree topologies. The Robinson–Foulds (RF) distance measure implemented in the PHYLIP program treedist (Felsenstein 1989) counts the number of bipartitions that are present in one tree and not in another tree. The RF distance takes values in the interval $[0, 2(n-3)]$, where n is the number of taxa (leaves) in the tree (Felsenstein 1989). The quartet distance implemented in Quartet Suite v1.0 (Piaggio-Talice et al. 2004) is a measure of the proportion of quartets that are resolved differently in two trees. It is a count of the number of quartets resolved differently in the input tree and the reference tree divided by the number of quartets resolved in the reference tree, $\binom{n}{4}$, where n is the number of taxa in the reference tree. This value is then subtracted from one to get a quartet similarity.

RF distances and quartet similarities were computed between the tree estimated with the original sequences and each bootstrap permutation tree to obtain 100 RF distances and 100 quartet similarities. When calculating quartet similarities, the tree from the original sequences was taken as the reference tree.

Simulation Methods

Ideally, the simulated sequences should retain the dependence between sites and the periodic structure of the true protein sequences. However, there are no methods or software packages so far to completely fulfill this requirement. Here, we simulate data using the program Seq-Gen

(Rambaut and Grassly 1997) under the Jones-Taylor-Thornton (JTT) model of amino acid substitution. It is important to note that the JTT model of evolution assumes that sequence sites evolve independently, and thus, sequences simulated with Seq-Gen will not necessarily retain the structural information present in the sequences. However, since the sequences simulated by Seq-Gen on an evolutionary tree have all evolved from the same ancestral taxon which is an extant sequence, the sites in the simulated sequences are not truly independent. The structural or periodic signals in the sequences are better kept if the tree on which the simulations are based is not very deep. Hence, we would expect our method to recover the reference tree in such cases.

Data

Four different data sets are used in this paper to illustrate our methods. We begin with an exploratory analysis on a noncontroversial eukaryote data set provided courtesy of Dr Andrew Roger (Center for Comparative Genomics and Evolutionary Bioinformatics, Dalhousie University). We then apply our methods to the nematode data set published in Foster and Hickey (1999) and a chloroplast data set (Ané et al. 2004; Gruenheit et al. 2008; Wu and Susko 2009). Finally, simulations are generated based on a five-taxa primate data set and the nematode data set. Sequences for each gene were downloaded from GenBank. GenBank accession numbers for the nematode data, chloroplast data, and primate data used for this analysis can be found in the supplementary tables A–C, Supplementary Material online. The sequences were then aligned using ClustalW in Bioedit, and the parts of the alignments for which one or more of the sequences had gaps were removed, so that the sequences for each gene all have the same length (Hall 1999).

The eukaryote data set consists of 35 ribosomal proteins and 17 taxa.

The nematode data set consists of the 12 mitochondrial protein-coding genes common to eight animals. This data set is known to have a problem with both long-branch attraction and compositional bias (Foster and Hickey 1999). There are two rival theories as to where the nematodes should branch in relation to other animals: the ecdysozoa hypothesis and the coelomata hypothesis. Aguinaldo et al. (1997) first proposed a clade of moulting animals, which includes nematodes and arthropods, based on a phylogenetic analysis of 18S ribosomal DNA sequences. They chose *Trichinella spiralis* as a representative nematode on account of its evolving more slowly than other nematodes, such as *Caenorhabditis elegans* which is used in our analysis. Their results indicated a strong relationship between the nematode and the arthropods. Dopazo and Dopazo (2005) carried out a phylogenetic analysis on the complete genomes of 11 taxa and also found strong support for the ecdysozoa hypothesis. In their analysis, Dopazo and Dopazo (2005) excluded the fast-evolving sequences of *C. elegans*. However, other analyses have rejected the ecdysozoa hypothesis. Rogozin et al. (2007) performed a genome-wide analysis using a type of rare genomic changes robust to long-branch attraction and taxon sampling and found strong support for the coelomata hypothesis. Blair et al. (2002) analyzed 100 individual protein data sets consisting of four taxa and again found strong support for the coelomata hypothesis. They argued that the findings of Aguinaldo et al. (1997) were due to the analysis being performed on a single gene. Philippe, Brinkmann, et al. (2005) argued that strong support for the coelomata theory was due to sparse taxon sampling. Their analysis of 146 genes from a sample of 35 taxa provided strong support for the ecdysozoa hypothesis. The debate regarding the correct placement of the nematodes remains unresolved, with analyses on different taxa samples and different genes returning conflicting results.

For the chloroplast phylogenetic tree, our final data set consisted of 25 proteins from 22 taxa. For this data, there has been some debate over the placement of *Amborella trichopoda* within the angiosperms. The majority of analyses place Amborella as the most basal of the angiosperms (Qiu et al. 1999; Soltis et al. 1999; Zanis et al. 2002). However, in some cases, a "Amborella + Nymphaea" clade was found to be most basal (Barkman et al. 2000). An alternative topology was presented by Goremykin et al. (2003), which placed the monocots as the most basal of the angiosperms. However, this topology was refuted by Soltis and Soltis (2004), Stefanovic et al. (2004) and later Goremykin and Hellwig (2006) who showed that model misspecification and long-branch attraction were the cause of the monocot first topology. Still, the true relationships among the angiosperms are not well resolved and resolution of the clade continues to be poor (Soltis et al. 2005).

The primate data set has five taxa: gibbon, orangutan, gorilla, chimp, and human. It consists of 13 mitochondrial protein-coding genes.

Results

Results on the Eukaryote Data Set

We begin by applying all combinations of methods on the noncontroversial eukaryote data set. The eukaryote data set consists of 17 taxa of plants, animals, and fungi for 35 ribosomal proteins. Dissimilarity matrices were computed using the common scaling covariance and the taxa-specific scaling covariance. Both MinVar and MinCV criteria were used to obtain scale coefficients with which to combine genes.

To determine which, if any, of the four methods for computing dissimilarities give similar results, we performed an initial comparative analysis of the dissimilarity matrices computed from these four techniques. Figure 1 shows the pairwise scatter plots with regression lines for the dissimilarities from each pair of methods.

The dissimilarity measures obtained from all four different methods are highly correlated, with Pearson correlations ranging from 0.9693 to 0.9881. The high correlation between the suites of methods suggests that the different techniques should return similar tree estimates.

Figure 2 shows the reference tree for the eukaryote data (Keeling et al. 2009). Trees are obtained using both BIONJ and FITCH. Thus, there are in total eight different

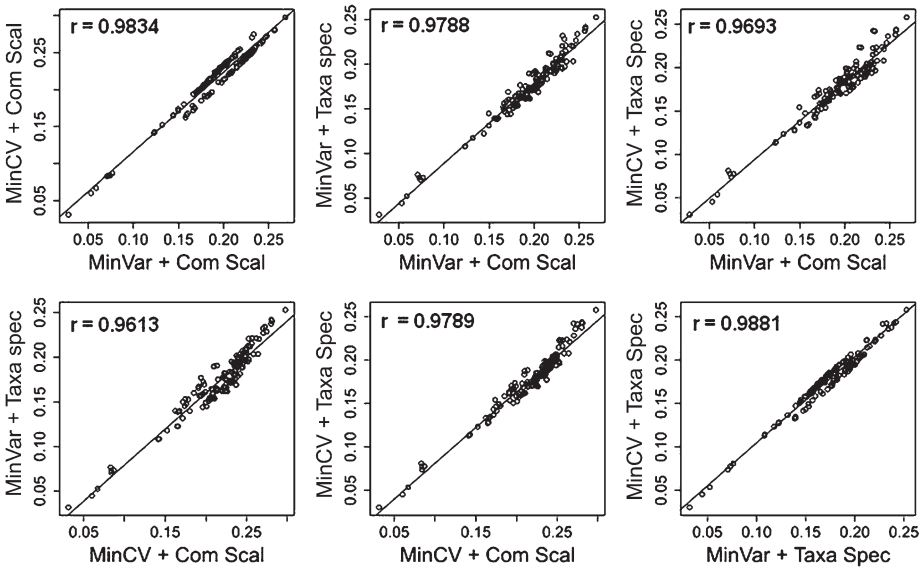

FIG. 1. Comparisons of the eukaryote data set dissimilarities obtained with common scaling (ComScal) and taxa-specific scaling (TaxaSpec) methods combined with MinVar and MinCV criteria (with Pearson correlation coefficient, *r*).

methods to build trees. The inferred trees by the eight different methods are shown in figures 3 and 4. All trees recover the major clades of plants, animals, and fungi. The MinCV taxa-specific scaling trees both recover the exact topology seen in the reference tree. The MinCV common scaling trees place *Schizosaccharomyces pombe* and *Candida albicans* as sister taxa, rather than branching *S. pombe* first, but otherwise recover the reference tree. All the MinVar trees erroneously place *Drosophila melanogaster* as the most basal animal. Both MinVar BIONJ trees erroneously place *Neurospora crassa* and *Magnaporthe grisea* closer to *Ustilago maydis* and *Cryptococcus* sp. than to *S. pombe* and *C. albicans*.

Table 1 summarizes the topological features recovered in each of the estimated trees as well as the bootstrap support for each feature. The accurate separation of all taxa into their major clades is recovered in all 100 bootstrap replicates under all eight methods. The branching of

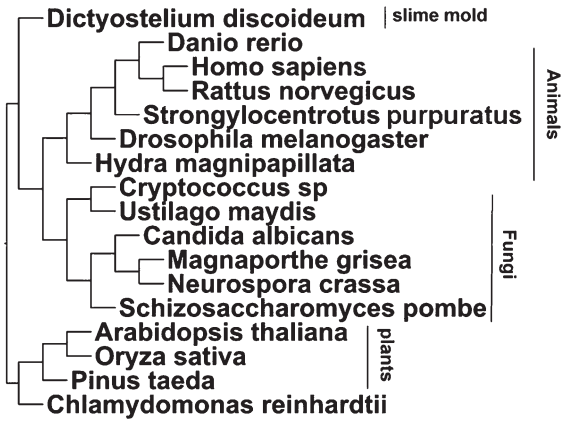

FIG. 2. Reference tree topology from Tree of Life web project for the eukaryote data set with 17 taxa (http://tolweb.org/Eukaryotes/3/2009.10.28) (Keeling et al. 2009).

Dictyostelium discoideum as basal in the animal clade has 100% bootstrap support in both the MinCV taxa-specific scaling trees but only weak support under the other six methods. The recovery of the reference tree topology within the three main clades has strong bootstrap support in all four MinCV trees. In the MinVar trees, the incorrect placement of *D. melanogaster* as most basal in the animal clade is strongly supported by the bootstrap replicates. Both MinVar BIONJ trees show strong bootstrap support for branching *N. crassa* and *M. grisea* with *U. maydis* and *Cryptococcus* sp. rather than *S. pombe* and *C. albicans* (99% under the common scaling method and 85% under the taxa-specific scaling method).

To measure the variance about the tree estimates for each method, we looked at the quartet similarities between the trees estimated from the block bootstrap samples and the trees estimated on the original sequences by all eight combinations of methods. Table 2 shows the computed quartet similarities. The columns show the number of bootstrap trees out of 100 whose quartet similarities fall within a given interval. The intervals are split according to all the resulting quartet similarity values. Although all methods give comparable results, one can see that the taxa-specific scaling with the MinCV method appears to be the most stable with a lower bound of 0.9118 quartet similarity. The mean quartet similarity values are 0.9619 and 0.9730, respectively, for BIONJ and FITCH. Thus, the taxa-specific scaling with the MinCV method has the smallest variability about the estimated trees. The common scaling covariance-based trees have greater variability than the taxa-specific scaling trees with a lower bound of quartet similarity of 0.9008 for both MinVar trees and 0.9025 and 0.8840 for MinCV trees.

The RF distances show a similar pattern. The eukaryote data set with 17 taxa has 14 interior nodes; hence, the maximum possible value for the RF is 28. Taxa-specific scaling covariance trees have a maximum RF distance of four with the

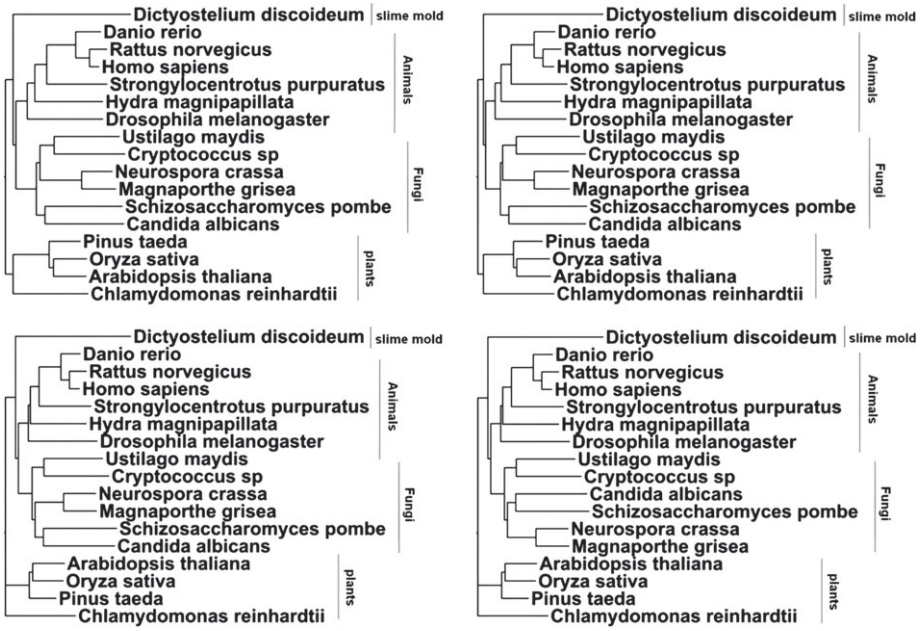

Fig. 3. Estimated BIONJ and FITCH trees for eukaryote data set when common scaling and taxa-specific scaling dissimilarities for multiple genes are combined with the MinVar criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxa-specific scaling with BIONJ (top right) and FITCH (bottom right).

majority of trees having distances less than two. The common scaling covariance trees have a maximum RF distance of six with majority of trees having distances less than four.

The MinCV method with the taxa-specific scaling appears to have the smallest variance, recovering the reference tree topology with strong bootstrap support. The MinCV with the common scaling also has relatively small variance about the estimated tree. The MinVar trees appear to have more erroneously placed branches than the MinCV trees, and these incorrect topologies are strongly supported by the corresponding bootstrap trees.

Although the differences in the estimated trees recovered from the four dissimilarity matrices are small, the MinCV method appears to return a more accurate topology than the MinVar method. Hence, for the remaining two data sets, we focus on the results obtained using the

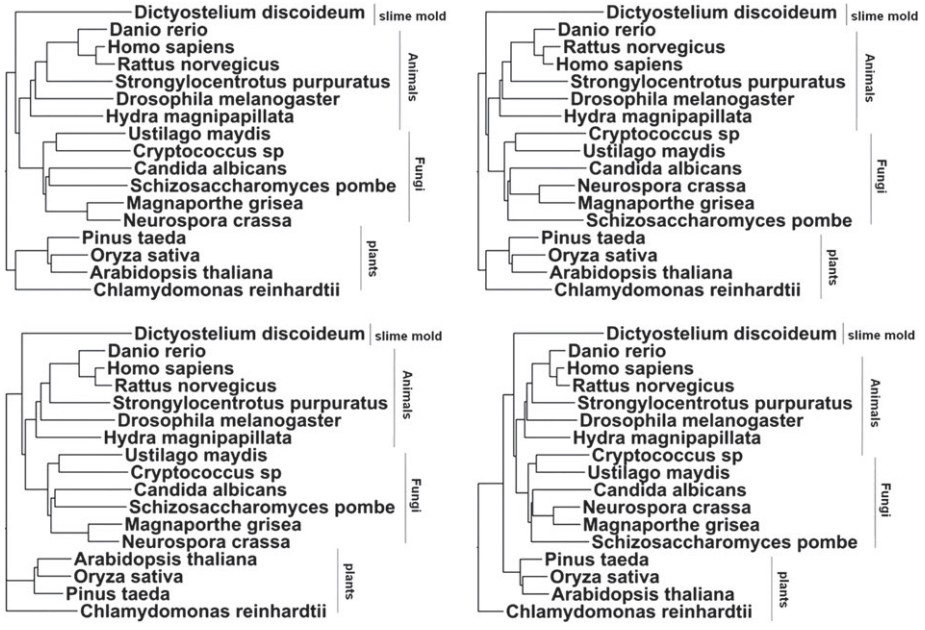

Fig. 4. Estimated BIONJ and FITCH trees for eukaryote data set when common scaling and taxa-specific scaling dissimilarities for multiple genes are combined with the MinCV criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxa-specific scaling with BIONJ (top right) and FITCH (bottom right).

Table 1. Bootstrap Support of the Topological Features for the Eukaryote Tree under Different Methods.

	Topological Features						
Trees	Recov. of Tree, Animal, Fungus Clades	DistDisc Basal to Animals	DistDisc in Fungus Clade	DrosMela Basal to Animals	HydrMagn Basal to Animals	NeurCras/ MagnGris with SchiPomb/CandAlbi	NeurCras/ MagnGris with UstiMayd/CryptoSp
MinVar ComScal BIONJ	100	6	94	100	0	1	99
MinVar ComScal FITCH	100	3	97	91	0	95	2
MinCV ComScal BIONJ	100	37	63	2	98	86	0
MinCV ComScal FITCH	100	48	52	0	100	91	0
MinVar TaxaSpec BIONJ	100	31	69	100	0	15	85
MinVar TaxaSpec FITCH	100	32	68	83	17	85	11
MinCV TaxaSpec BIONJ	100	100	0	6	94	74	0
MinCV TaxaSpec FITCH	100	100	0	0	100	82	0

MinCV for both the common scaling and taxa-specific scaling covariance-based dissimilarity measures.

Results on the Nematode Data Set

The nematode data set consists of 12 protein-coding genes common to eight taxa presented in Foster and Hickey (1999). There are two rival theories concerning where the nematodes should be placed in the tree. The ecdysozoa theory favors a clade of moulting animals, grouping nematodes, and arthropods together (Aguinaldo et al. 1997; Dopazo and Dopazo 2005). The coelomata theory places nematodes as basal to the vertebrates and arthropods (Blair et al. 2002; Rogozin et al. 2007). Figure 5 shows the trees under these two hypotheses. The nematode data set is known to have problems with compositional bias and long-branch attraction, which results in the honeybee (*Apis mellifera*) and the roundworm (*C. elegans*) being branched as sister taxa with strong bootstrap support (Foster and Hickey 1999). Again, we begin with an exploratory analysis of the dissimilarities computed from the common scaling and taxa-specific scaling covariances with MinCV scale coefficients. A scatter plot with regression of the MinCV dissimilarities under the taxa-specific scaling versus the common scaling method is shown in figure 6. For this data, the correlation between the two methods is very high with $r = 0.9848$. The largest residual is associated with dissimilarities between honeybee and roundworm, followed by chicken and sea urchin, honeybee and *Allomyces macrogynus*, and honeybee and brine shrimp. The large residual corresponding to chicken and sea urchin is a bit surprising as these two taxa are fairly noncontroversial with regards to their placement in the tree.

The MinCV trees obtained with the four methods are shown in figure 7. The placement of the taxa relative to each other corresponds to the grouping seen under the ecdysozoa hypothesis. The common scaling covariance with the BIONJ and the taxa-specific scaling with FITCH both return trees with the same topology as the reference tree under the ecdysozoa hypothesis. The common scaling covariance with FITCH erroneously places the honeybee as basal to the other arthropods, whereas the taxa-specific scaling covariance with BIONJ tree erroneously places the roundworm and honeybee together as sister taxa. Hence, honeybee, roundworm, and brine shrimp, which have large residuals associated with their dissimilarities in the initial regression, vary in their relative positions in the trees under the two different spectral covariance methods.

Table 3 shows the bootstrap support for the topological features for the nematode tree under different methods. The placement of the taxa in agreement with the ecdysozoa hypothesis has strong bootstrap support in both common scaling trees (100% with BIONJ and 99% with FITCH). Fifty-two percentage of the taxa-specific scaling BIONJ trees support the ecdysozoa hypothesis topology, whereas 67 % of the taxa-specific scaling FITCH tree support the coelomata theory. The placement of brine shrimp as the most basal of arthropods recovered in the common scaling BIONJ tree and the taxa-specific scaling FITCH tree has no bootstrap support. The erroneous branching of honeybee as the basal animal has moderate bootstrap support in both common scaling trees and the taxa-specific BIONJ tree and weak bootstrap support in the taxa-specific FITCH tree. Separation of the honeybee and the roundworm occurs in 63% of bootstrap trees for both BIONJ and FITCH common

Table 2. Eukaryote Data: Quartet Similarity between Bootstrap Trees and Original Data Trees.

	Number of Bootstrap Permutation Trees with Percentage of Identically Resolved Quartets							
	ComScal + MinVar		ComScal + MinCV		TaxaSpec + MinVar		TaxaSpec + MinCV	
Quartet Similarity (X)	BIONJ	FITCH	BIONJ	FITCH	BIONJ	FITCH	BIONJ	FITCH
[0.88,0.93)	53	27	1	13	11	24	23	13
[0.93,0.96)	10	25	93	84	58	42	24	20
[0.96,0.99)	28	24	0	0	4	7	21	24
1.00	9	24	6	3	27	27	32	43
Mean	0.9488	0.9605	0.9428	0.9378	0.9527	0.9423	0.9620	0.9730
Min	0.9008	0.9008	0.9025	0.8840	0.9025	0.9025	0.9118	0.9118

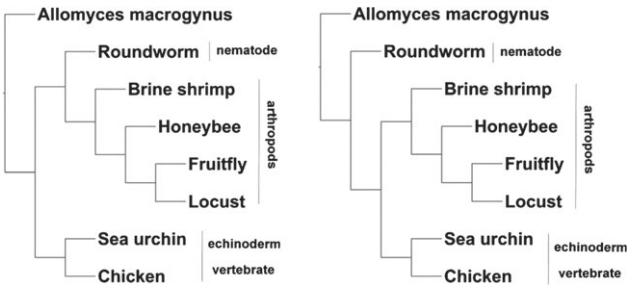

FIG. 5. Reference topology for the nematode data set under the ecdysozoa hypothesis (left) and coelomata hypothesis (right) (Blair et al. 2002).

scaling methods, 62% of bootstrap trees for the taxa-specific BIONJ method, and 69% of bootstrap trees for the taxa-specific FITCH method. A combination of long-branch attraction and compositional bias often causes the honeybee and roundworm to be grouped as sister taxa (Foster and Hickey 1999), but here, all four methods are able to separate these two with moderate bootstrap support.

Table 4 shows the quartet similarities between the bootstrap trees and the original data tree for the nematode data. Variability about the tree estimates for this data is greater than that of the eukaryote data, with minimum quartet similarites of 0.5429 and 0.5857 for the taxa-specific scaling trees and 0.7143 and 0.8714 for the common scaling trees. The mean quartet similarity is 0.8440 for common scaling BIONJ trees and 0.9444 for common scaling FITCH trees, compared with 0.8091 and 0.7246 for the corresponding taxa-specific scaling trees.

The RF distances show the same pattern. For the nematode data set with five interior nodes, the maximium possible value for RF distance is ten. Taxa-specific scaling trees have a maximum distance of six with the majority of distances being four or less. The common scaling covariance-based BIONJ trees have a maximum distance of four with the majority of trees having values less than two. The common scaling covariance-based FITCH trees have a maximum distance of two, with 56 of the 100 RF distances being zero.

Results on the Chloroplast Data Set

The chloroplast data set consists of 25 chloroplast proteins from 22 taxa. There has been some debate over the placement of *A. trichopoda* within the angiosperms. Most analyses place *A. trichopoda* as the most basal angiosperm (Qiu et al. 1999; Soltis et al. 1999; Zanis et al. 2002); though in some cases, a Amborella + Nymphaea clade was found to be most basal (Barkman et al. 2000). Goremykin et al. (2003) found an alternative topology, which placed the monocots as the most basal of the angiosperms, although this topology was later found to be erroneous (Soltis and Soltis 2004; Stefanovic et al. 2004; Goremykin and Hellwig 2006). Figure 8 shows the reference tree for the chloroplast data (Ané et al. 2004; Soltis et al. 2005).

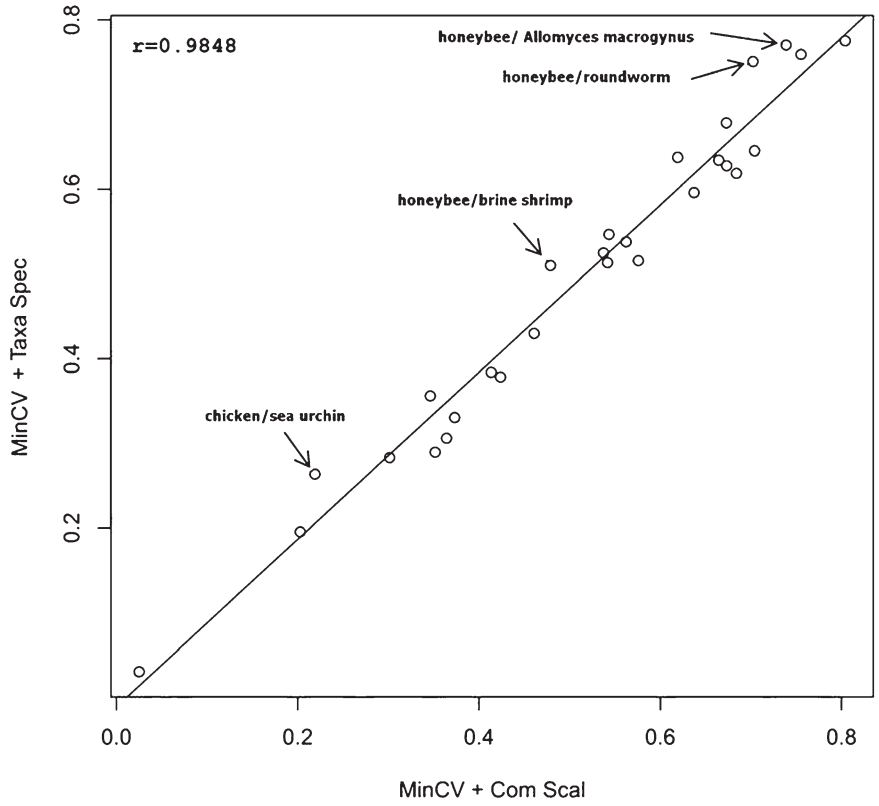

FIG. 6. Comparisons of the nematode data set dissimilarities computed with the common scaling (ComScal) method and combined with the MinCV criterion and the taxa-specific scaling (TaxaSpec) method combined with the MinCV criterion (Pearson correlation = 0.9848). Taxa pairs with largest discrepancy in dissimilarities computed under these two methods shown with arrows.

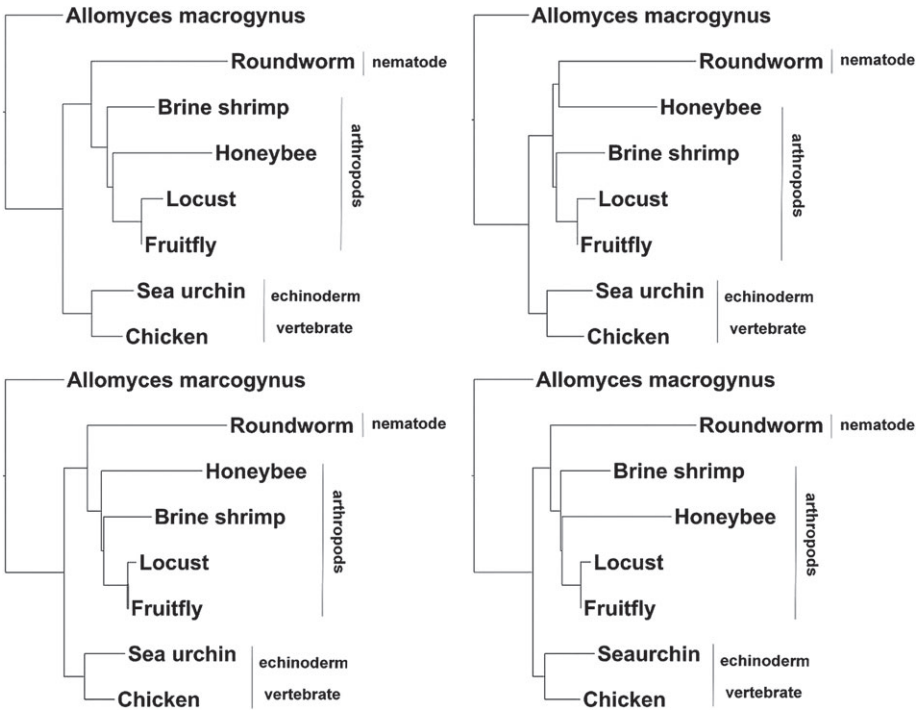

FIG. 7. Estimated BIONJ and FITCH trees for the nematode data set when common scaling and taxa-specific scaling dissimilarities for multiple genes are combined with the MinCV criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxa-specific scaling with BIONJ (top right) and FITCH (bottom right).

We begin with an analysis on all 25 genes in chloroplast data set and then discuss how this differs from an initial analysis we did on a smaller chloroplast data set, which consisted of only 19 of the 25 chloroplast proteins. The same 22 taxa were used in both analyses.

Again, we focus on the MinCV method and compare two different methods of scaling and two different tree building methods. A scatter plot with regression of the taxa-specific scaling versus common scaling dissimilarities is shown in figure 9. Once more, correlation between the two methods is fairly high with $r = 0.9639$. The largest residuals in the chloroplast data set correspond to the dissimilarities between the green algae *Chlorella vulgaris*, *Mesostigma viride*, and *Nephroselmis olivacea*.

The real data trees are shown in figure 10. In all four trees, the separation of green algae, nonseed plants, and seed plants is recovered. *Acorus americanus* should be grouped with the other monocots within the angiosperm clade but is instead placed with the eudicots in all four trees. Also, *Psilotum nudum* erroneously branches with the mosses and liverworts rather than with the other fern, *Adiantum capillus-veneris*. The taxa-specific scaling trees place *A. trichopoda* and *Nymphaea alba* as sister taxa, whereas the common scaling trees place *Calycanthus floridus* and *A. trichopoda* as sister taxa. In all four trees, a clade with *A. trichopoda*, *N. alba*, and *C. floridus* is basal in the angiosperm clade.

Table 5 shows the topological features and the bootstrap support for each feature given by the four different methods. The correct separation of taxa into main clades of green algae, nonseed plants, seed plants, and angiosperms has 100% bootstrap support in all four methods. There is also strong bootstrap support for a clade with *A. trichopoda*, *N. alba*, and *C. floridus* as basal in the angiosperm clade (100% for all four methods). The branching of *A. trichopoda* and *N. alba* as sister taxa has moderate support in the common scaling FITCH tree and both taxa-specific trees (51–66%). The branching of *N. alba* and *C. floridus* as sister taxa is strongly supported by the common scaling BIONJ tree. The branching of the two ferns, *P. nudum* and *A.*

Table 3. Bootstrap Support of the Topological Features for the Nematode Tree under Different Methods.

	Topological Features					
Trees	Agrees with Ecdysozoa	Agrees with Coelomata	Sep. of Honeybee and Nematode	Honeybee and Nematode Sister Taxa	Brine Shrimp Basal to Arthropods	Honeybee Basal to Arthropods
MinCV ComScal BIONJ	100	0	63	37	0	63
MinCV ComScal FITCH	99	1	63	37	0	63
MinVar TaxaSpec BIONJ	52	48	62	38	2	61
MinVar TaxaSpec FITCH	30	67	69	31	0	43

NOTE.—Sep., separation

Table 4. Nematode Data: Quartet Similarity between Bootstrap Trees and Original Data Trees.

	Number of Bootstrap Permutation Trees with Percentage of Identically Resolved Quartets			
	ComScal + MinCV		TaxaSpec + MinCV	
Quartet Similarity (X)	BIONJ	FITCH	BIONJ	FITCH
[0.54,0.75)	37	0	48	30
[0.75,0.85)	0	0	14	65
[0.85,0.99)	44	44	0	2
1.00	19	56	38	3
Mean	0.8440	0.9444	0.8091	0.7246
Min	0.7143	0.8714	0.5429	0.5857

capillus-veneris, as sister taxa has moderate support in both taxa-specific trees (55% with BIONJ and 52% with FITCH). The erroneous placement of *P. nudum* with the mosses and liverwort seems to occur in all common scaling BIONJ trees and 78% of the common scaling FITCH trees.

Table 6 shows the quartet similarities between the bootstrap permutation trees and the corresponding real data trees. Mean quartet similarities for all four methods are very close, with the means for the taxa-specific trees being slightly higher than the means for the common scaling trees. For the taxa-specific scaling trees, the mean quartet similarities are 0.9852 and 0.9890 with BIONJ and FITCH, respectively. The common scaling trees have corresponding mean quartet similarities of 0.9771 and 0.9778. Minimum quartet similarities are all greater than 0.93. There appears to be greater variability about the trees estimated from the common scaling-based distances than those estimated from the taxa-specific scaling distances. For the chloroplast data set with 22 taxa, the maximum possible value the RF can attain is 38. The RF distances are consistent with the quartet similarities, with common scaling covariance-based trees attaining a maximum RF distance of 14 using FITCH and 10 using BIONJ, whereas the taxa-specific scaling covariance-based trees attain a maximum RF distance of 10 using FITCH and 8 using BIONJ.

The strong bootstrap support obtained for the trees estimated from these 25 genes was somewhat surprising as analyses on a subset of 19 genes of these 25 resolved the angiosperm clade very differently. When only 19 genes were included in the analyses, all methods returned the erroneous monocot first tree with strong bootstrap support. Removing those 19 genes for which the monocot distances were relatively large with respect to the other angiosperms still resulted in a monocot first tree. We then added genes *atpI*, *clpP*, *psaB*, *psaC*, *rbcL*, and *rpoC1*. Including these genes resulted in a clade consisting of *A. trichopoda*, *N. alba*, and *C. floridus* being basal in the angiosperm clade. Figure 11 shows the common scaling MinCV distances of nonmonocot angiosperms versus the three monocots, when 19 and 25 genes are used in the analyses.

In the case of *Zea mays* and *Oryza sativa*, adding the six additional genes results in larger MinCV common scaling distances between these two monocots and *A. trichopoda*, *N. alba*, and *C. floridus*, whereas the corresponding distances between these two monocots and the eudicots is only slightly greater except in the case of *Lotus corniculatus*. The 19-gene MinCV common scaling distances between *Triticum aestivum*, the eudicots, tend to be greater, whereas the distances between *T. aestivum* and the *A. trichopoda*,

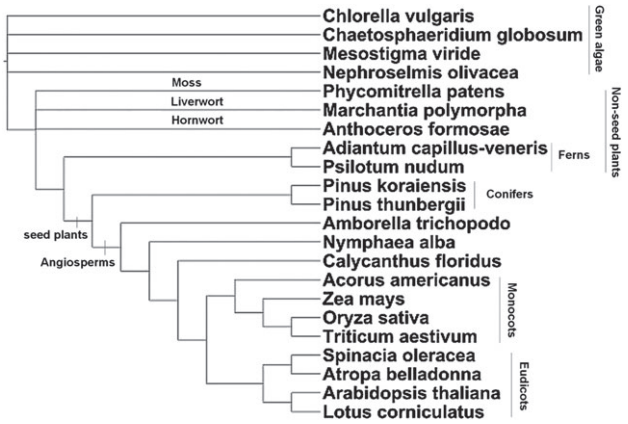

Fig. 8. Reference tree topology for the chloroplast data set with 22 taxa (Ané et al. 2004; Soltis et al. 2005).

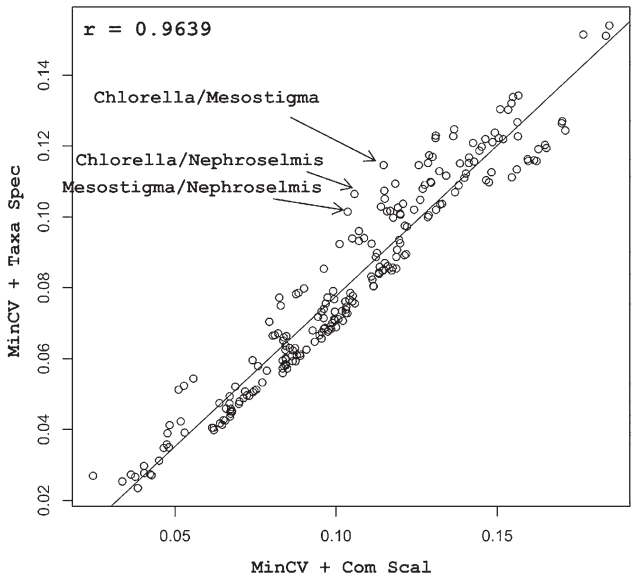

Fig. 9. Comparisons of the chloroplast data set dissimilarities computed with the common scaling (ComScal) method and combined with the MinCV criterion and the taxa-specific scaling (TaxaSpec) method combined with the MinCV criterion (Pearson correlation = 0.9639). Taxa pairs with the largest discrepancy in dissimilarities computed under these two methods shown with arrows.

FIG. 10. Estimated BIONJ and FITCH trees for the chloroplast data set when common scaling and taxa-specific scaling dissimilarities for multiple genes are combined with the MinCV criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxa-specific scaling with BIONJ (top right) and FITCH (bottom right).

N. alba, and *C. floridus* are smaller. The distance between *T. aestivum* and the other two monocots is also greater within the monocot clade. The taxa-specific scaling distances return similar results. Clearly, the six additional genes are highly influential in determining the relative placement of the taxa in the angiosperm clade of the combined gene tree.

Simulations

We simulated data based on two different data sets, a primate data set consisting of five taxa: gibbon, orangutan, gorilla, chimp, and human and the nematode data set used in the analysis above. For both data sets, the trees obtained by the common scaling method combined with the MinCV criterion are used as the input trees in Seq-Gen. We reduced both the sequence similarity and the structure similarity in the simulated sequences by increasing the branch lengths, and these results are compared with those obtained with the block bootstrap permutations where the structure similarity is partially preserved. Note that the simulation method may be somewhat biased against our method since one would expect structural patterns to be maintained by natural selection as well as deriving from the ancestral sequence.

Simulations Generated from Primate Data Set

Figure 12 shows the reference tree topology for the five taxa in our data set (Tree of Life Web Project, 1999). This topology is also estimated by all our eight combinations of methods applied on the primate data set.

Throughout the whole simulation, the sequence of gibbon is specified as the ancestral sequence. We first simulated 1,000 data sets for each gene based on the tree shown in figure 12, with branch lengths estimated by common scaling MinCV method and call this simulation scheme S1. We then repeated this process to create 2 additional sets of 1,000 data sets, in which the branch lengths of the input tree are multiplied by 100 and 1,000, and

Table 5. Bootstrap Support of the Topological Features for the Chloroplast Tree under Different Methods.

	Topological Feature				
Trees	Recov. of Green Algae, Nonseed Plant and Angiosperm Clade	Amborella, Nymphaea, Calycanthus Clade Basal in Angiosperm Clade	Amborella and Nymphaea Sister Taxa	Nymphaea and Calycanthus Sister Taxa	Psilotum and Adiantum Sister Taxa
MinCV ComScal BIONJ	100	100	29	22	71
MinCV ComScal FITCH	100	100	66	22	34
MinVar TaxaSpec BIONJ	100	100	56	44	44
MinVar TaxaSpec FITCH	100	100	51	49	49

Table 6. Chloroplast Data: Quartet Similarity between Bootstrap Trees and Original Data Trees.

	Number of Bootstrap Permutation Trees with Percentage of Identically Resolved Quartets			
	ComScal + MinCV		TaxaSpec + MinCV	
Quartet Similarity (X)	BIONJ	FITCH	BIONJ	FITCH
[0.93,0.96)	23	24	1	0
[0.96,0.99)	74	65	79	61
1.00	3	11	20	39
MEAN	0.9771	0.9778	0.9852	0.9890
MIN	0.9315	0.9481	0.9571	0.9669

we call these two simulations schemes S100 and S1000, respectively. We compared the analysis performed on these simulated data with the analysis performed on the block bootstrap permutations on all eight combinations of methods.

For the 1,000 data sets simulated under S1, 100% of the estimated trees from the simulated data recover the same topology as the reference tree for all eight methods. Figure 13*A* shows the majority rule consensus tree from S1 obtained by the common scaling method with MinCV criterion. All other seven methods result in the same consensus tree shown in figure 13*A*.

The sequences within the primate data set have high sequence similarity (90–100%). This level of sequence similarity is also present in the S1-simulated sequences. The sequence similarities are reduced to 10–50% for the S100 scheme and less than 10% for the S1000 scheme. Table 7 shows the proportion of trees, which recover the reference tree under the different simulation schemes. We also performed 1,000 block bootstrap permutations with block size 14 for each gene of the primate data set and applied all eight methods. The last row in table 7 shows the proportion of correctly estimated trees under the block bootstrap permutations.

For the primate data, with average sequence similarities greater than 10%, all the trees based on simulated sequences recover the reference tree. When sequence similarity is less than 10%, only 4–10% of estimated trees recover the topology of the reference tree. For the block bootstrap permutation samples, 88–98% of trees based on the permuted sequences recover the reference tree. For the primate data, our methods are fairly robust when sequence similarity is reduced.

Simulations Generated from Nematode Data Set

For the nematode data set, the sequence of *A. macrogynus* is specified as the ancestral sequence for Seq-Gen simulations. Again, we simulated 1,000 data sets for each gene. As with the primate data, we will call the simulation based on the tree estimated by the common scaling MinCV method

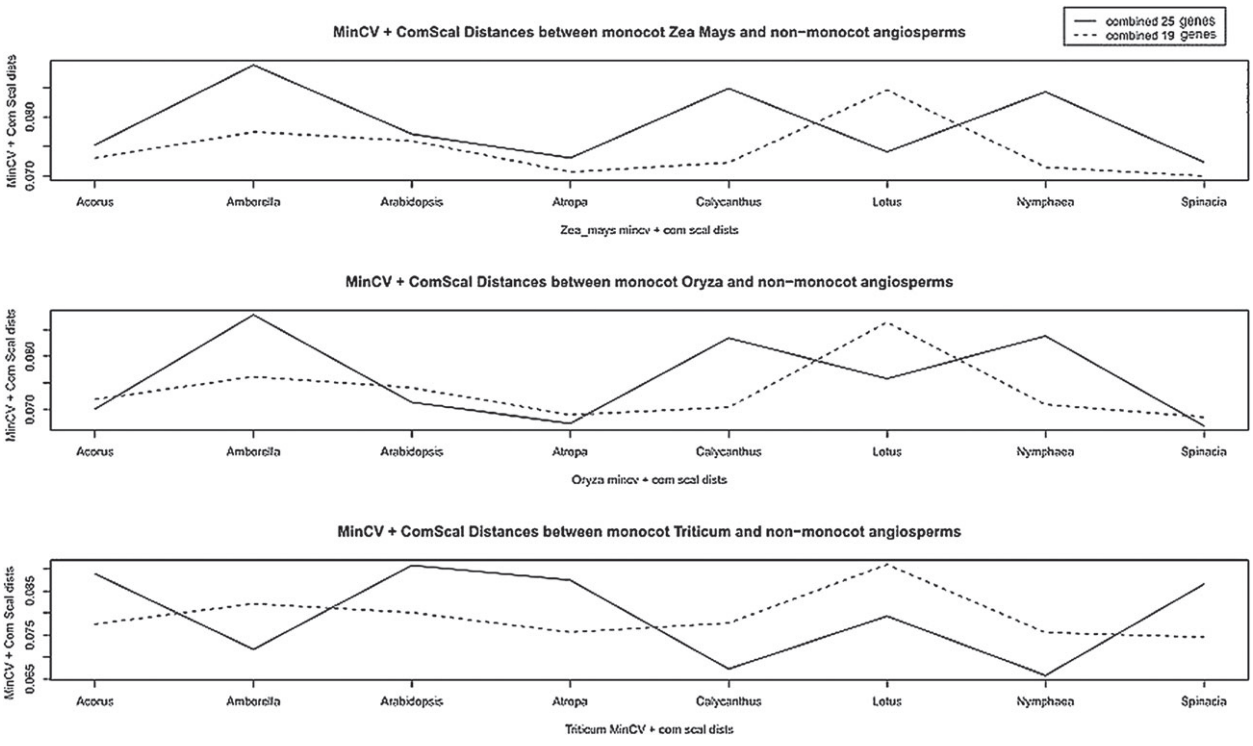

FIG. 11. Common scaling distances for monocots versus other angiosperms for 19 and 25 genes combined with MinCV.

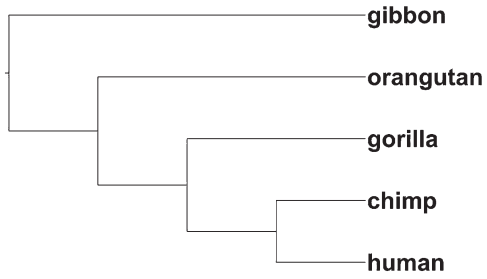

FIG. 12. Reference tree for the primate data set from http://tolweb.org/Catarrhini/16293/1999.01.01.

simulation scheme S1 and repeat the process to create 2 additional sets of 1,000 data sets, in which the branch lengths of the input tree are multiplied by 25 and by 100. We refer to these simulation schemes as S25 and S100, respectively. Sequences simulated under simulation scheme S1 have sequence similarities between 9.2% and 82.4%, whereas sequence similarities under simulation schemes S25 and S100 are reduced to 2–30% and 0–13.6%, respectively.

The input tree topology for Seq-Gen simulations agrees with the ecdysozoa hypothesis. We do not discount the possibility that the topology under the coelomata hypothesis is correct. However, since the purpose of the simulations is to determine support for our methods that recovers trees, which agree with the reference topology under the ecdysozoa hypothesis, our data were simulated under this topology rather than that under the coelomata hypothesis. Table 8 shows the proportion of trees, which recover the topologies for both the ecdysozoa and coelomata hypotheses. It is not surprising that none of the trees based on the simulated sequences agree with the coelomata hypothesis as they were simulated under an ecdysozoa tree. Of the trees estimated from the data generated under simulation scheme S1, 83.6–93% recover the input tree. For data generated under simulation scheme S25, 44.5–46.4% of the estimated common scaling trees recover the input tree, whereas the recovery rates of estimated taxa-specific scaling trees are 0–0.3%. Under simulation scheme S100, none of the estimated trees recover the input tree. For the common scaling trees based on bootstrap permutations, 68.2–95.6% recover the tree that agrees with the ecdysozoa topology. Of the taxa-specific scaling trees based on bootstrap permutations, 42–79% recover the tree that agrees with the coelomata hypothesis.

Analysis of Simulation Results

For data generated with Seq-Gen, both sequence and structure similarity are preserved when branch lengths are short, as is the case under simulation scheme S1. The fact that our methods can recover the input tree with such a high rate for simulation scheme S1 shows the effectiveness of the proposed methods. When sequence and structure similarity are reduced, the recovery rate drops correspondingly. From the simulation results, we see that the recovery rates of the four methods based on the taxa-specific scaling decline much more quickly than that of the four methods based on the common scaling. This perhaps shows that the methods based on the common scaling are more sensitive in picking up the weak sequence and structure similarity signals, and that the common scaling method is preferred over the taxa-specific scaling method from this aspect.

The block bootstrap permutations completely preserve the sequence similarity of the original sequences but only partially preserve the structure similarity. For the primate data, there is no controversy about which tree is the right tree. The high-recovery rate of the right tree under block bootstrap permutation of the data shows that such a bootstrap method is valid. For the nematode data set, the true tree is unknown. The estimated common scaling trees based on bootstrap permutation samples strongly support the ecdysozoa hypothesis, whereas the estimated taxa-specific scaling trees show moderate to strong support to the coelomata hypothesis. Although these results reflect the uncertainty in the evolutionary position of the nematode, we do see slightly stronger evidence to support ecdysozoa hypothesis from our study of this data set.

Permutations

To further validate the covariance-based methods, we performed further analyses on 1,000 block size 1 permutation samples taken from the nematode data set. Permutations were computed using the SEQBOOT program in PHYLIP (Felsenstein 1989). We compared the distances obtained from the block size 1 permutation samples with those obtained from the block size 14 permutations. Recall

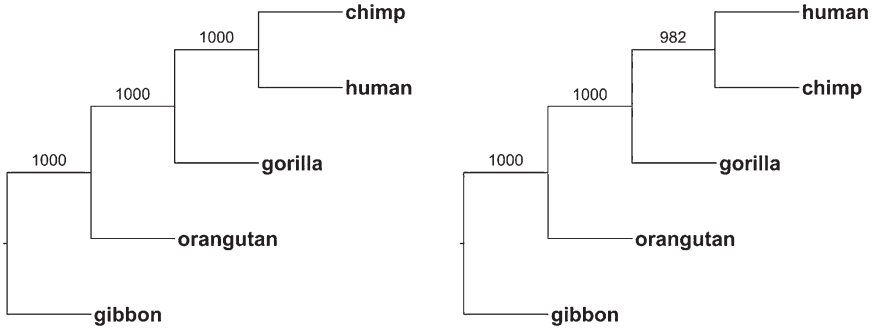

FIG. 13. Primate majority-rule consensus trees estimated with the common scaling (ComScal) method and combined with the MinCV criterion for 1,000 Seq-Gen–simulated sequences (left) and 1,000 block bootstrap permutation sequences (right).

Table 7. Proportion of Simulated Trees with Varying Levels of Sequence Identity and Block Permutation Trees which Recover the Primate Reference Tree.

Simulation Scheme (% Sequence Similarity)	MinCV ComScal BIONJ	MinVar ComScal BIONJ	MinCV ComScal FITCH	MinVar ComScal FITCH	MinCV TaxaSpec BIONJ	MinVar TaxaSpec BIONJ	MinCV TaxaSpec FITCH	MinVar TaxaSpec FITCH
S1 (>90%)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
S100 (10–50%)	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
S1000 (<10%)	0.053	0.053	0.066	0.066	0.041	0.041	0.101	0.103
Block Permutation (>90%)	0.982	0.916	0.975	0.908	0.972	0.886	0.969	0.88

the block size 14 permutation samples completely preserve site similarity and partially preserve structure similarity. We expect that the variability about the estimated tree measured by block size 1 permutation samples will be greater than that measured by block size 14 permutation samples as the structural signal is erased by the block size 1 permutations. Results presented below are for the common scaling covariance-based dissimilarities. Similar results were obtained with the taxa-specific scaling covariance-based dissimilarities.

Figure 14 shows common scaling MinCV majority-rule consensus trees obtained from both sets of permutation samples. Although the consensus trees are the same, we can see that the variability about the resolved branches is greater for the block size 1 permutation samples than for the block size 14 permutation samples.

Boxplots of the 1,000 pairwise distances under both permutation schemes for three pairs of taxa randomly selected from the 28 taxa pairs are shown in figure 15. The horizontal line across the *x* axis corresponds to the common scaling MinCV distance obtained from the real data. The 1,000 bootstrap distances under the block size 14 permutation scheme have much smaller variance than the 1,000 bootstrap distances under the block size 1 permutation scheme, whereas those under the block size 14 permutation scheme seem to have greater bias. However, all taxa pairs appear to be biased in the same way (somewhat greater than the distance computed for the real data), and hence, the relative relationship between the taxa is preserved for most of the samples and the variance about the estimated tree is relatively small. Although the range of the block size 1 permutation sample distances always encompass the real data distance, the median distance for different taxa pairs fluctuates about the real data distance, and many of the samples have distances much higher and/or lower than the real data distance. This in turn results in higher variance about the estimated tree for the 1,000 block size 1 permutation distances.

The covariance-based dissimilarities incorporate both sequence and structural similarity between proteins. When the structural information is destroyed, variance of the estimated tree increases.

Discussion

The dissimilarity matrices computed from the four techniques obtained by combining a spectral covariance scaling method with either MinVar or MinCV scale coefficients are highly correlated. Differences in the tree estimates obtained from these dissimilarities are for the most part small, differing in the placement of only a few taxa. In the eukaryote data, trees estimated using the MinVar and MinCV methods differed in their placement of *D. melanogaster* in the animal clade. For the nematode data, the trees obtained from the common scaling and taxa-specific scaling methods differed in their placement of honeybee, roundworm, and brine shrimp, relative to each other. The dissimilarities between these taxa had large residuals associated with them in the initial regression analysis. For the chloroplast data set, the taxa-specific scaling trees place *A. trichopoda* and *N. alba* as sister taxa, whereas the common scaling trees place *C. floridus* and *A. trichopoda* as sister taxa.

Our exploratory analysis of the eukaryote data set showed that the MinCV method was able to recover the currently accepted topology shown in figure 2 with strong

Table 8. Proportion of Simulated Trees with Varying Levels of Sequence Identity and Block Permutation Trees which Recover the Nematode Ecdysozoa and Coelomata Trees.

Simulation Scheme (% Sequence Similarity)	Hypothesis	MinCV ComScal BIONJ	MinVar ComScal BIONJ	MinCV ComScal FITCH	MinVar ComScal FITCH	MinCV TaxaSpec BIONJ	MinVar TaxaSpec BIONJ	MinCV TaxaSpec FITCH	MinVar TaxaSpec FITCH
S1 ($10%–83%$)	Ecdysozoa	0.892	0.897	0.926	0.930	0.845	0.836	0.902	0.888
	Coelomata	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
S25 ($2%–30%$)	Ecdysozoa	0.445	0.447	0.462	0.464	0.000	0.000	0.003	0.003
	Coelomata	0.000	0.000	0.000	0.000	0.000	0.000	0.001	0.001
S100 ($0%–14%$)	Ecdysozoa	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
	Coelomata	0.000	0.000	0.000	0.000	0.000	0.000	0.000	0.000
Block Permutation ($10%–83%$)	Ecdysozoa	0.683	0.956	0.682	0.913	0.147	0.125	0.025	0.005
	Coelomata	0.000	0.000	0.006	0.007	0.477	0.739	0.420	0.425

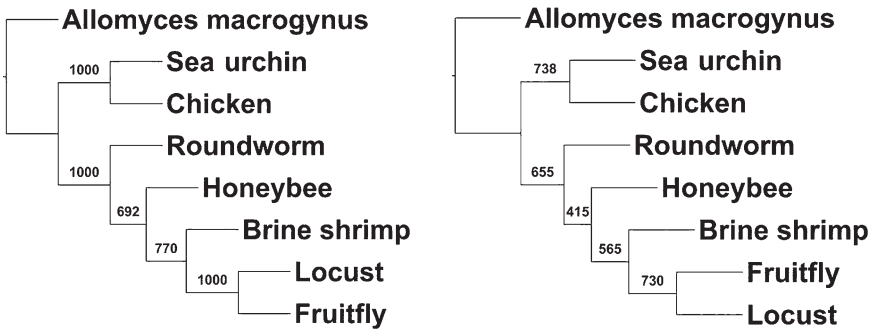

FIG. 14. Nematode majority-rule consensus trees estimated with the common scaling method and combine with MinCV criterion for 1,000 block size 14 permutation samples (left) and 1,000 block size 1 permutation samples (right).

bootstrap support (Keeling et al. 2009). The MinVar method was able to recover parts of this topology, but erroneously placed taxa *D. melanogaster* as the most basal animal, and was not able to recover the correct position of *D. discoideum*. Results were the same with both the common scaling and taxa-specific scaling. For this reason, we focused our attention on the MinCV method for the remaining two data sets.

For the nematode data, the common scaling method supported the ecdysozoa hypothesis topology with strong bootstrap support (Aguinaldo et al. 1997; Dopazo and Dopazo 2005), although the honeybee was erroneously placed as a basal arthropod in the FITCH tree. The ML trees reported in Foster and Hickey (1999) grouped honeybee and roundworm together as sister taxa. The common scaling covariance method was able to separate these two taxa with moderate bootstrap support. For the taxa-specific scaling method, support for the ecdysozoa hypothesis was weak, whereas the coelomata hypothesis had moderate to strong bootstrap support. Roundworm and honeybee were erroneously grouped together with weak bootstrap support.

For the 25-gene chloroplast data, both the common scaling and taxa-specific scaling methods recovered the main clades with strong bootstrap support. Resolution of the angiosperm clade has been extensively studied with different topologies being recovered depending on method and taxon sampling (Qiu et al. 1999; Soltis et al. 1999; Zanis et al. 2002; Goremykin et al. 2003, 2005; Ané et al. 2004; Soltis and Soltis 2004; Stefanovic et al. 2004). Though neither of the common scaling nor taxa-specific scaling methods recovers the exact reference topology in figure 8 (Ané et al. 2004; Soltis et al. 2005), the relative positions of the taxa within the angiosperm clade more or less agrees with the reference tree, with the exception of *A. americanus* which is misplaced with the eudicots in the common scaling and taxa-specific scaling trees. Analysis on a subset of 19 of these genes, which excluded *atpI*, *clpP*, *psaB*, *psaC*, *rbcL*, and *rpoC1*, returned the incorrect tree with monocots placed as basal in the angiosperm clade. The relative MinCV distances within

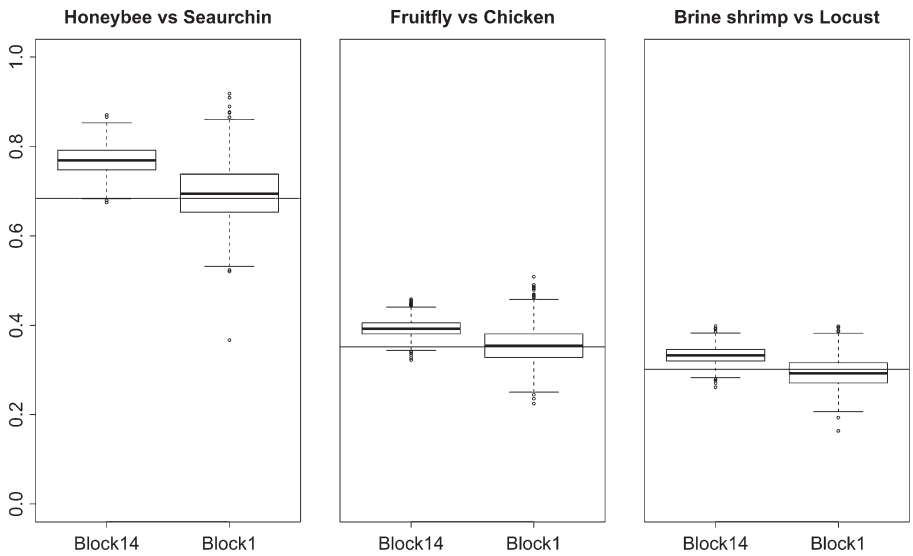

FIG. 15. Boxplots of 1,000 sample distances obtained from block size 14 permutation samples (left) and block size 1 permutation samples (right) for three randomly selected taxa pairs from the nematode data set. Horizontal line across the *x* axis corresponds to the common scaling MinCV distance for real data.

the angiosperm clade appear to be greatly changed by the inclusion of these six genes indicating that these genes are given considerable weight. The additional six genes appear to be highly influential in determining the topology within the angiosperm clade in the combined gene tree.

The trees computed from Seq-Gen–simulated sequences indicate that the covariance-based methods do a good job of capturing phylogenetic signal. When branch lengths are short, both sequence and structure similarity are preserved in the simulated sequences, resulting in high recovery of the input tree by the estimated trees. When sequence and structural similarity is reduced, the recovery rate drops accordingly. The block bootstrap permutations preserve all of the sequence similarity of the original sequences but only some of the structural similarity and hence have a lower recovery rate than the data generated with Seq-Gen under simulation schemes S1. For a data set such as the nematode data where the true tree is unknown, the bootstrap permutation samples may be more informative than simulations because they require no assumptions with regards to the true tree topology. Bootstrap permutation samples based on the common scaling strongly support the ecdysozoa hypothesis, whereas those based on the taxa-specific scaling show moderate to high support for the coelomata hypothesis.

The spectral covariance trees are based on structural similarity between proteins. However, it has been shown that structural similarity and sequence similarity are highly correlated (Chothia and Lesk 1986; Wood and Pearson 1999), and that orthologous proteins have greater structural similarity than paralogous proteins for the same level of sequence similarity (Peterson et al. 2009). For this reason, the estimated trees reflect both the structural and the sequence similarity between taxa, which is present within the proteins used for the analysis (Collins et al. 2006). The fact that spectral covariance-based methods can recover the major structure of the tree implies that major structural and sequential differences can be captured by this method. The total covariance used here as a summary measure of the spectral covariance is only one possible measure. It is important to note that by summing over all frequencies some structural information is being averaged out. We are currently considering other possible summary measures which take into account the frequencies at which the highest peaks occur.

At the moment, the impact of our method on systematic biases, such as variable evolutionary rates across genes, taxa, or individual sites within a sequence, is unclear. However, since the spectral covariance is based on structure information rather than substitutions at the sequence level, our method should be less sensitive to systematic error than sequence-based methods. The common scaling covariance was able to separate the honeybee and roundworm, which sequence-based methods tend to group together due to long branch attraction (Foster and Hickey 1999). Further simulation studies to rigorously test how our method responds under these conditions are required.

The spectral covariance method does not assume site independence and does not require specification of an evolutionary model. The MinCV is an effective method for combining information from multiple genes to obtain tree estimates, and the idea can be generally applied with other distance or dissimilarity measures to combine information from multiple genes.

At the moment, the method is limited to include only proteins common to all taxa. An extension of this method to deal with missing data and allow for the inclusion of a larger number of protein sequences will be developed. One way to do this is by modeling the pairwise distances computed from the available pairs of genes and using missing data imputation methods based on the statistical models. These methods are currently under investigation.

Supplementary Material

Supplementary tables A–C are available at *Molecular Biology and Evolution* online (http://www.mbe.oxfordjournals.org/).

Acknowledgments

The authors would like to thank Dr Rob Beiko, Faculty of Computer Science, Dalhousie University, for his helpful comments and suggestions during the preparation of this manuscript. This work was supported by funds from the Natural Sciences and Engineering Research Council of Canada to C.F. and H.G.

References

Aguinaldo AM, Turbeville JM, Linford LS, Rivera MC, Garey JR, Raff RA, Lake JA. 1997. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature* 387:489–493.

Ané C, Burleigh JG, McMahon MM, Sanderson MJ. 2004. Covarion structure in plastid genome evolution: a new statistical test. *Mol Biol Evol.* 22:914–925.

Barkman TJ, Chenery G, McNeal JR, Lyons-Weiler J, Ellisens WJ, Moore G, Wolfe AD, DePamphilis CW. 2000. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proc Natl Acad Sci U S A.* 97:13166–13171.

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

Beven RB, Lang F, Bryant D. 2005. Calculating the evolutionary rate of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. *Syst Biol.* 54: 900–915.

Bininda-Emonds O. 2004. The evolution of supertrees. *Trends Ecol Evol.* 19:315–322.

Blair JE, Ikeo K, Gojobori T, Hedges SB. 2002. The evolutionary position of nematodes. *BMC Evol Biol.* 2:1471–2148.

Bull JJ, Huelsenbeck J, Cunningham C, Swofford D. 1993. Partitioning and combining data in phylogenetic analysis. *Syst Biol.* 42: 384–397.

Bulmer M. 1991. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Mol Biol Evol.* 8:868–883.

Burleigh J, Driskell A, Sanderson M. 2006. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Syst Biol.* 55:426–440.

Chothia C, Lesk AM. 1986. The relation between the divergence of sequence and structure in proteins. *EMBO J.* 5:823–826.

Collins K, Gu H, Field C. 2006. Examining protein structure and similarities by spectral analysis. *Stat Appl Genet Mol Biol.* 5:23–44.
de Queiroz A. 1993. For consensus (sometimes). *Syst Biol.* 42: 368–372.
de Queiroz A, Gatesy J. 2007. The supermatrix approach to systematics. *Trends Ecol Evol.* 22:34–41.
Dopazo H, Dopazo J. 2005. Genome-scale evidence of the nematode-arthropod clade. *Genome Biol.* 6:R41.
Farris J, Kallersjo M, Kluge AG, Bult C. 1995. Testing significance of incongruence. *Cladistics* 10:315–319.
Felsenstein J. 1989. PHYLIP: phylogeny inference package (version 3.2). *Cladistics* 5:164–166.
Felsenstein J. 2003. Inferring phylogenies. Sunderland (MA): Sinauer Associates.
Fitch W, Margoliash E. 1967. Construction of phylogenetic trees. *Science* 155:279–284.
Foster P, Hickey D. 1999. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *J Mol Evol.* 48:284–290.
Gascuel O. 1997. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Mol Biol Evol.* 14: 685–695.
Gatesy J, Baker R, Hayashi C. 2004. Inconsistencies in arguments for the supertree approach: supermatrices versus supertrees of *crocodylia*. *Syst Biol.* 53:342–355.
Goremykin V, Hirsch-Ernst KI, Wolfl S, Hellwig FH. 2003. Analysis of the *Amborella trichopoda* chloroplast genome sequence suggests that *Amborella* is not a basal angiosperm. *Mol Biol Evol.* 20:1499–1505.
Goremykin V, Holland B, Hirsch-Ernst KI, Hellwig FH. 2005. Analysis of *Acorus calamus* chloroplast genome and its phylogenetic implications. *Mol Biol Evol.* 22:1813–1822.
Goremykin VV, Hellwig FH. 2006. A new test of phylogenetic model fitness addresses the issue of the basal angiosperm phylogeny. *Gene* 381:81–91.
Gruenheit N, Lockhart P, Steel M, Martin W. 2008. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. *Mol Biol Evol.* 25:1512–1520.
Hall TA. 1999. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucleic Acids Symp Ser.* 41:95–98.
Hendy M, Penny D. 1993. Spectral analysis of phylogenetic data. *J Classif.* 10:5–24.
Jeffroy O, Brinkmann H, Delsuc F, Philippe H. 2006. Phylogenomics: the beginning of incongruence? *Trends Genet.* 22:225–231.
Keeling P, Leander BS, Simpson A. 2009. Eukaryotes. Eukaryota, organisms with nucleated cells. [cited 2010 Jan 14]. Available from: http://tolweb.org/Eukaryotes/3/2009.10.28
Kuhner M, Felsenstein J. 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Mol Biol Evol.* 11:459–468.
Kunsch H. 1989. The jacknife and bootsrap for general stationary observations. *Ann Stat.* 17:1217–1241.
Lecointre G. 2005. Total evidence requires exclusion of phylogenetically misleading data. *Zool Scr.* 34:101–117.
Leigh J, Susko E, Baumgartner M, Roger A. 2008. Testing congruence in phylogenomic analysis. *Syst Biol.* 57:104–115.
Miyamoto M, Fitch W. 1995. Testing species phylogenies and phylogenetic methods with congruence. *Syst Biol.* 44:64–76.
Peterson ME, Chen F, Saven JG, Roos DS, Babbitt PC, Sali A. 2009. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Sci.* 18:1306–1315.
Philippe H, Brinkmann H, Lartillot N. 2005. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Mol Biol Evol.* 22:1246–1253.
Philippe H, Delsuc F, Brinkmann H, Lartillot N. 2005. Phylogenomics. *Annu Rev Ecol Evol Syst.* 36:541–562.
Piaggio-Talice R, Burleigh G, Eulenstein O. 2004. Phylogenetic supertrees: combining information to reveal the tree of life. In: Bininda-Emonds OR, editor. Phylogenetic supertrees: combining information to reveal the tree of life. Dordrecht (The Netherlands): Kluwer Academic. p. 173–191.
Pinheiro J, Bates D, DebRoy S, Sarkar D, The R Core team. 2009. nlme: linear and nonlinear mixed effects models. R package version, Vol. 3. Vienna (Austria): R Foundation for Statistical Computing, p. 1–93.
Priestly M. 1981. Spectral analysis and time series. San Diego (CA): Academic Press.
Qiu Y, Lee J, Bernasconi-Quadroni F, Soltis DE, Soltis PS, Zanis M, Zimmer EA, Chen Z, Savolainen V, Chase MW. 1999. The earliest angiosperms: evidience from mitochondrial, plastid and nuclear genomes. *Nature* 402:404–407.
Rambaut A, Grassly NC. 1997. Seq-Gen: an application for the Monte Carlo simulation of DNA sequence evolution along phylogenetic trees. *Comput Appl Biosci.* 13:235–238.
Rogozin IB, Wolf YI, Carmel L, Koonin EV. 2007. Ecdysozoan clade rejected by genome-wide analysis of rare amino acid replacements. *Mol Biol Evol.* 24:1080–1090.
Rokas A, Williams B, King N, Carroll S. 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature* 425:798–804.
Saitou N, Nei M. 1987. The neighbor-joining method: a new method for constructing phylogenetic trees. *Mol Biol Evol.* 4:406–425.
Soltis DE, Soltis PS. 2004. *Amborella* not a "basal angiosperm"? not so fast. *Am J Bot.* 91:997–1001.
Soltis P, Soltis D, Edwards C. 2005. Angiosperms. Flowering plants. [cited 2010 Apr 21]. Available from: http://tolweb.org/Angiosperms/20646/2005.06.03
Soltis PS, Soltis DE, Chase MW. 1999. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature* 402:402–404.
Stefanovic S, Rice DW, Palmer JD. 2004. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evol Biol.* 4:35.
Stoffer D, Tyler D, McDougall A. 1993. Spectral analysis for categorical time series: scaling and the spectral envelope. *Biometrika* 85:201–213.
Stoffer D, Tyler D, Wendt D. 2000. The spectral envelope and its applications. *Stat Sci.* 15:224–253.
Studier JA, Keppler KJ. 1988. A note on the neighbor-joining algorithm of Saitou and Nei. *Mol Biol Evol.* 5:729–731.
Tree of Life Web Project. 1999. Primates. Lemurs, tarsiers, monkeys, apes, and humans. [cited 2010 Nov 25]. Available from: http://tolweb.org/Primates/15963/1999.01.01
Wood TC, Pearson WR. 1999. Evolution of protein sequences and structures. *J Mol Biol.* 291:977–995.
Wu J, Susko E. 2009. General heterotachy and distance method adjustments. *Mol Biol Evol.* 26:2689–2697.
Zanis M, Soltis DE, Soltis PS, Mathews S, Donoghue MJ. 2002. The root of the angiosperms revisited. *Proc Natl Acad Sci U S A.* 99:6848–6853.
Zelwer M, Daubin V. 2004. Detecting phylogenetic incongruence using BIONJ: an improvement of the ILD test. *Mol Phylogenet Evol.* 33:687–693.