PHYLOGENETIC ANALYSIS OF MULTIPLE GENES BASED ON SPECTRAL METHODS

by

Melanie Abeysundera

Submitted in partial fulfilment of the requirements
for the degree of Doctor of Philosophy

at

Dalhousie University
Halifax, Nova Scotia
October 2011

DALHOUSIE UNIVERSITY

DEPARTMENT OF MATHEMATICS AND STATISTICS

The undersigned hereby certify that they have read and recommend to the Faculty of Graduate Studies for acceptance a thesis entitled "PHYLOGENETIC ANALYSIS OF MULTIPLE GENES BASED ON SPECTRAL METHODS" by Melanie Abeysundera in partial fulfilment of the requirements for the degree of Doctor of Philosophy.

Dated: October 28, 2011

External Examiner: _____

Co-supervisors: _____

_____

Examining Committee: _____

_____

Departmental Representative: _____

DALHOUSIE UNIVERSITY

Date: October 28, 2011

Author: Melanie Abeysundera

Title: PHYLOGENETIC ANALYSIS OF MULTIPLE GENES BASED ON SPECTRAL METHODS

Department or School: Department of Mathematics and Statistics

Degree: PhD Convocation: May Year: 2012

Signature of Author

Table of Contents

List of Tables

List of Figures

Abstract

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 multi-gene 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 used to build trees. Unlike other methods, the spectral covariance method focuses on the relationship between the sites of genetic sequences.

We consider two methods with which to combine the dissimilarity or distance matrices of multiple genes. The first method involves properly scaling the dissimilarity measures derived from different genes between a pair of species and using the mean of these scaled dissimilarity measures as a summary statistic to measure the taxonomic distances across multiple genes. We introduced two criteria for computing scale coefficients which can then be used to combine information across genes, namely the minimum variance (MinVar) criterion and the minimum coefficient of variation squared (MinCV) criterion. The scale coefficients obtained with the MinVar and MinCV criteria can then be used to derive a combined-gene tree from the weighted average of the distance or dissimilarity matrices of multiple genes.

The second method is based on the singular value decomposition of a matrix made up of the $p = \binom{n}{2}$ vectors of pairwise distances for k genes. By decomposing such a matrix, we extract the common signal present in multiple genes to obtain a single tree representation of the relationship between a given set of taxa. Influence functions for the components of the singular value decomposition are derived to determine which genes are most influential in determining the combined-gene tree.

List of Abbreviations Used

MinVar Mimimum variance

MinCV Minimum coefficient of variation

ML Maximum likelihood

JTT Jones Taylor Thornton

diss dissimilarity

sim similarity

RF Robinson-Foulds

SVD Singular value decomposition

GIF Generalized influence function

Cov Covariance

ComScal Common scaling spectral covariance

TaxaSpec Taxon specific spectral covariance

Acknowledgements

I'd like to express my deepest gratitude to my supervisors Dr. Hong Gu and Dr. Chris Field whose support and encouragement made this thesis possible. I also thank the Department of Mathematics and Statistics at Dalhousie University for affording me the opportunity to pursue this research. Heartfelt thanks to my readers Dr. Rob Beiko and Dr. Bruce Smith for their invaluable comments and suggestions and also to my external examiner Dr. Matthew Spencer. Many thanks also to Balagopal Pillai for all his technical support over the years. Finally I would like to thank my friends and family, in particular, my parents and my godparents, for their encouragement and emotional support during the writing of this thesis. This thesis was made possible by the financial support of a Dalhousie University Faculty of Graduate Studies scholarship and teaching assistantship.

Chapter 1

INTRODUCTION

The goal of phylogenetic analysis is to infer an evolutionary relationship between a given set of taxa from their nucleotide or amino acid sequences. The nucleic acids, deoxyribonucleic acid (DNA) and ribonucleic acid (RNA) carry the genetic information in a cell and are made up of four nucleotides: adenine (A), guanine(G), cytosine (C), and thymine (T) in DNA or uracil (U) in RNA . Adenine and guanine are purines while cytosine and thymine are pyrimidines. Complementary purine-pyrimidine pairs are joined in the shape of a double helix by hydrogen bonds to form DNA. These purine-pyrimidine pairs are called base pairs (bp). Nucleotide triplets in messenger RNA, called codons, are translated into amino acids in the ribosome. During replication DNA sequences are sometime altered by point mutations, insertions or deletions. Over time, mutations in DNA sequences accumulate and result in changes in protein function or structure. Molecular evolutionists seek to model these changes over time through an evolutionary tree.

In this thesis we apply our methods to amino acid sequences. An example of a data set consisting of protein sequences for a single gene along with a phylogenetic tree representing an evolutionary relationship for a given set of taxa is shown in Figure 1.1.

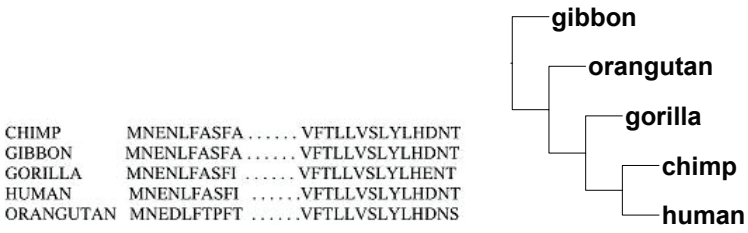

Figure 1.1: A data set consisting of the protein sequences for five primate taxa for a single gene, ATP6 (left) and an inferred phylogenetic tree (right).

The three most common methods for inferring trees are maximum likelihood methods, distance methods and maximum parsimony methods. Maximum likelihood (ML) based methods of tree estimation assume sequence sites evolve independently as a continuous time Markov process based on a pre-specified evolutionary model which allows computation of the transition probabilities at a given site. The likelihood is a function of the tree topology, branch lengths and parameters of the rate matrix (Felsenstein, 2003). Likelihood-based phylogenies have a sound statistical basis but have the disadvantage of being dependent on the particular model of evolution used. Most distance-matrix methods also employ an evolutionary model represented by a substitution matrix. Distances between taxa are derived from similarity scores between sequences computed from a given substitution matrix. The computed distances are then passed to a tree building method such as BIONJ or FITCH to obtain a phylogenetic tree (Gascuel, 1997; Miyamoto and Fitch, 1995). Parsimony-based methods return a tree that requires the least number of substitutions among individual sites in a set of sequences.

1.1 THE SPECTRAL COVARIANCE

A novel approach to phylogenetic analysis was considered by Collins et al. (2006) who applied a spectral envelope based covariance method to tree estimation and compared it to results obtained using standard likelihood methods. The general idea behind the spectral covariance approach to phylogenetics is to measure the similarity between two proteins from the periodic patterns in their sequences. The periodic behaviour inherent in DNA and amino acid sequences makes them an ideal subject for spectral analysis. Stoffer et al. (2000) developed the spectral envelope method as a general framework for frequency domain analysis of categorical time series. The underlying idea is that periodic structure is identifiable as spectral peaks, and maximizing the spectrum is equivalent to searching for the strongest periodic component in a series. The spectral covariance method uses this same idea. Amino acid sequences are treated as categorical time series and a scaling function is chosen which maximizes the spectral covariance between two sequences. A high covariance at a given frequency signifies a common periodicity between two sequences. These common periodicities

in turn represent protein secondary structures. It is known that α-helices have a periodicity of 3.6 residues (the number of amino acids per turn) which corresponds to a peak in the spectrum at approximately $\omega = 0.277$. β-sheets, on the other hand, have been shown to have a maximum peak at approximately $\omega = 0.435$ corresponding to 2.3 residues. Turns and loops connecting secondary structures are known to have a periodicity of 3-4 residues corresponding to frequencies $\omega = 0.250$ to $\omega = 0.333$, while the repetition of secondary structure elements in a protein motif is typically between 8-14 residues corresponding with frequency $\omega = 0.071$ to $\omega = 0.125$ (Collins et al., 2006). An example of a protein motif consisting of repeated α-helices and β-sheets is shown in Figure 1.2. The α-helices consist of spirals, while the β-sheets have folded accordion-like pleats.

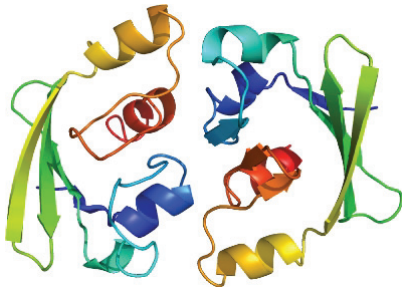

Figure 1.2: An example of a protein motif consisting of repeated α-helices and β-sheets. The Structural Genomics Consortium (SGC) (http://www.thesgc.org/)

Collins et al. (2006) found that the spectral covariance approach yielded similar results to the maximum likelihood (ML) approach when applied to a single protein. This was a remarkable result as the two techniques are based on completely different criteria. While the ML method of tree estimation relies heavily on the given evolutionary model, the spectral covariance method of sequence comparison does not assume any particular evolutionary model and relies instead on adoption of time series spectral methods. In this way, the ML method can be thought of as parametric, while the spectral covariance can be viewed as non-parametric. Because the spectral covariance is a time series approach no assumption of site independence with respect to an evolutionary model is required.

1.2 PHYLOGENETIC ANALYSIS OF MULTIPLE GENES

In this thesis, we extend the single-gene analyses based on the spectral covariance to analysis of multiple genes. Two different versions of the spectral covariance are employed, the common scaling covariance used by Collins et al. (2006) and the taxon-specific scaling covariance which we shall introduce in chapter 2. Phylogenetic analysis on single genes will often result in conflicting topologies of the same set of taxa in the estimated evolutionary trees (Philippe et al., 2005b). The goal of multi-gene analysis is to combine information across several genes in such a way that a single tree representation of the relationship for the same set of taxa is obtained from multiple genes. There are two popular methods for combining information across multiple genes: concatenation and consensus. In the former the sequences from several genes are concatenated and a phylogenetic analysis is performed on the concatenated sequences (Bininda-Emonds, 2004; Bull et al., 1993; Burleigh et al., 2006; de Queiroz and Gatesy, 2007; Gatesy et al., 2004; Philippe et al., 2005b). In the latter a separate tree is inferred for each gene and a single tree is estimated by consensus (Baum, 1992; de Queiroz, 1993; Miyamoto and Fitch, 1995). The problem with these approaches is that they assume all genes share a common evolutionary history. This assumption may not be valid and may result in an incorrect estimate of the species tree (Philippe et al., 2005b).

We present two alternative methods for combining information from several genes using their distance or dissimilarity matrices. The first method involves taking a weighted average of the dissimilarity matrices for different genes. Two different criteria for computing scale coefficients with which to combine genes are presented, the minimum variance criterium (MinVar) and the minimum coefficient of variation criterium (MinCV). 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. Like the method presented by Cruscuolo et al. (2006), the MinVar and MinCV methods attempt to bring the distance matrices as close together by possible based on some criterion but rather than transform the distances for each taxa pair within each distance matrix, a single scale coefficient in chosen for each

distance matrix such that the coefficient of variation squared between the distance matrices for each gene in minimized. The methods are applied to four different data sets, two non-controversial and two with some dispute over the correct placement of taxa in the tree. Trees are constructed using two distance-based tree building methods, BIONJ and FITCH. A variation of the block bootstrap sampling method introduced by Kunsch (1989) is used to determine the variance of our estimated trees. Through simulations we show that the covariance based methods effectively capture phylogenetic signal even when structural information is not fully retained. Finally, we analyze the influence of individual genes on the combined-gene tree obtained with the MinCV method.

The second method is based on singular value decomposition of the matrix of combined pairwise distance vectors for each gene. We describe how the first right eigenvector of the singular value decomposition of the matrix consisting of the pairwise distance vectors for multiple genes may be used to estimate a single tree representation of the relationship between a set of taxa from several genes. We apply our method on both JTT distances and common scaling covariance based dissimilarities. We again employ the block bootstrap to estimate the variance of estimated trees. We extend the results of Fung et al. (2007), who derived the influence functions for principal components, to derive influence functions for the singular value decomposition of the distance matrix. These are used to quantify the influence of individual genes on the combined-gene topology obtained with the singular value decomposition method and to determine the robustness of the combined-gene tree. While much of the analyses and examples given use the spectral covariance based dissimilarities, it should be noted that these methods may be applied to any distance measure.

1.3 PROTEIN STRUCTURE PREDICTION AND CLASSIFICATION

Protein structure prediction is an important aspect of the field of pharmaceutical medicine where the three dimensional structure of a target protein is used to discover new drug candidates (Hubbard, 2006). Most prediction methods rely extensively on knowledge of proteins whose structures have been previously determined by X-ray crystallography or NMR spectroscopy (Ginalski et al., 2005). Alternatively, an

unknown protein structure may be determined by comparative modelling in which it is compared to a homologous protein or a template in a protein database, but such a procedure requires that a homologue be available. Using the structural information coded in the periodic patterns of the amino acid sequences of proteins, we attempt to classify the protein into a structural category, and in this way identify the main structural elements of the protein of interest and possible templates. To do this we employ the spectral envelope, first introduced by Stoffer et al. (1993) for analysis and scaling of a categorical time series. Stoffer et al. (2000) applied the spectral envelope to DNA sequences and determined that peaks in the spectrum of a DNA sequence corresponded with the protein-coding regions of that sequence. Collins et al. (2006) extended this result to amino acid sequences and determined that the peaks of the spectral envelope of amino acid sequences were related to the periodicity of protein secondary structures. The structural features present in the spectral envelop are extracted and used as covariates in a classification and regression tree (CART). The reader should note that the trees presented in this section of thesis are not phylogenetic trees in that they do not represent an evolutionary relationship between a set of taxa, but rather a grouping of genes with common structural features.

1.4 OUTLINE

The remainder of this thesis is structured as follows. In chapter 2 we review the spectral envelope and the common scaling spectral covariance, and introduce the taxon-specific scaling covariance. We describe the MinVar and MinCV methods of computing scale coefficients with which to scale the dissimilarity matrices derived from different genes in chapter 3. In chapter 4 we give an alternative method for deriving combined-gene trees using singular value decomposition. We derive a generalized influence function for the components of the singular value decomposition in chapter 5, and use these to examine the influence of individual genes on the combined-gene tree obtained with the singular value decomposition based method and to evaluate the robustness of our methods. We take a brief look at protein structure classification using the spectral envelope in chapter 6. Some concluding remarks are given in chapter 7.

Chapter 2

REVIEW OF SPECTRAL METHODS FOR ANALYSIS OF PROTEIN SEQUENCES

In this chapter we review the spectral covariance and spectral envelope methods for analysis of protein data.

2.1 THE SPECTRAL COVARIANCE

One of the principal interests in studying the similarity among protein sequences and among protein structures is to infer evolutionary relationships between taxa. Among the different methods for achieving this goal, the most widely used are maximum likelihood (ML) based methods which are known to have many good properties (Vandamme, 2009). However, maximum likelihood (ML) based methods of tree estimation assume sequence sites evolve independently and are dependent on the pre-specified model of evolution (Felsenstein, 2003; Vandamme, 2009). It is generally accepted that there is a dependence among the sites (Philippe et al., 2005b). A spectral envelope based covariance method to address the dependence among sites was developed by Collins et al. (2006) . The spectral envelope was first introduced by Stoffer et al. (1993) as a method of analysing categorical time series in the frequency-domain. The spectral envelope provides an automated method of scaling qualitative time series data to emphasise 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 non-standardized adaptation of the spectral envelope approach to coherency presented in Stoffer et al. (2000). Since prominent peaks in the spectral covariance correspond to common periodicities in the individual sequences, the spectral covariance, while sequence based, is also a measure of structural similarity (Collins et al., 2006).

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, $t = 0, \pm1, \pm2, ...$, be a categorical time series with finite state space $C = \{c_1, c_2, ..., c_k\}$. For $\beta = (\beta_1, \beta_2, ..., \beta_k)' \in \mathbb{R}^k$, denote $X_t(\beta)$ as the real-valued time series corresponding to the scaling that assigns c_j the value β_j. Note, that here we use $'$ to denote the transpose of a vector or a matrix. 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 1 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

$$\mathbf{Cov}_{12}^2(\omega) = \sup_{\alpha,\beta} |\alpha^{'}(\omega) f_{12}(\omega)\beta(\omega)|^2, \tag{2.1}$$

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 $R_{12}(k) = \mathbf{Cov}(Y_{1t}, Y_{2(t+k)})$ 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 (2.1) can be rewritten as

$$\mathbf{Cov}_{12}^2(\omega) = \sup_{\alpha,\beta} \left\{ \left[\alpha'(\omega) f_{12}^{re}(\omega)\beta(\omega)\right]^2 + \left[\alpha'(\omega) f_{12}^{im}(\omega)\beta(\omega)\right]^2 \right\}. \tag{2.2}$$

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

In this thesis 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 taxon-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 to 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). While in the common scaling method, the set of scalings for any given sequence depends upon the sequence to which it is being compared, the taxon-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 to *taxon2* or *taxon3*. The taxon-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 taxon-specific covariance methods yield very similar results.

Note that although we work with multiple alignment in this thesis, the spectral covariance method does not require all sequences be aligned. Another option might be to use pairwise alignments rather than multiple alignments.

2.1.1 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 (2.2) is now

$$\mathbf{Cov}_{12}^2 = \sup_{\beta} |\beta' f_{12} \beta|^2, \tag{2.3}$$

subject to $\beta'\beta$=1, where f_{12} is the cross-spectral density between Y_{1t} and Y_{2t}. Equation (2.3) can be rewritten as

$$\mathbf{Cov}_{12}^2 = \sup_{\beta'\beta=1} \left([\beta' f_{12}^{re} \beta]^2 + [\beta' f_{12}^{im} \beta]^2 \right). \tag{2.4}$$

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

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

Equation (2.4) then becomes

$$\mathbf{Cov}_{12}^2 = \sup_{\beta'\beta=1} \left([\beta' A^{re} \beta]^2 + [\beta' A^{im} \beta]^2 \right). \tag{2.5}$$

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

1. Initialization: set β to be one of the following:

$$\begin{aligned} \beta_1 &= \varepsilon_1(A^{re'} A^{re}) \\ \beta_2 &= \varepsilon_1(A^{im'} A^{im}) \end{aligned}$$

   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

$$\mathbf{Cov}_{12}^2 = \left(\beta_0' A^{re} \beta_0 \right)^2 + \left(\beta_0' A^{im} \beta_0 \right)^2$$

2. Iteratively calculate scalings using

$$\beta_j \;=\; \varepsilon_1 \left(A^{re} \beta_{j-1} \beta_{j-1}^{'} A^{re} + A^{im} \beta_{j-1} \beta_{j-1}^{'} A^{im} \right) \tag{2.6}$$

until convergence. Convergence criteria is set as $||\beta_j - \beta_{j-1}||^2 < 0.001$.

2.1.2 THE TAXON-SPECIFIC SPECTRAL COVARIANCE

For the common scaling spectral covariance, the scalings for any given taxa are dependent upon the taxa to 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 to the locust and a different set of scalings when it is compared to 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. Since the taxon-specific similarities measures the similarity of each pair of taxa relative to the entire set, such a method might also provide insight into effect of a taxon on the estimated similarities. We might compare how estimated taxon-specific similarities change when a taxon is included or excluded from the data set. To compute the taxon-specific scaling spectral covariance, the following criterion is used. Following the notation above, for K taxa, denote the multivariate series of K sequences $X_{1t}, ...X_{Kt}$ as $Y_{1t}, ..., Y_{Kt}$. The squared spectral covariance is now

$$\sum_{i<j} \mathbf{Cov}_{ij}^2 = \sup_{\beta_i, ..., \beta_K} \sum_{i<j} |\beta_i' f_{ij} \beta_j|^2, \tag{2.7}$$

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

To find β_i's which maximize (2.7), begin by initializing $\beta_i^0, (i = 1, ..., K)$ as the spectral envelope scaling of the ith sequence. Then the algorithm is as follows:

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

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

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

2.1.3 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 i^{th} and j^{th} sequences, denoted as $sim(x_i, x_j)$, is then converted into a dissimilarity measure between the i^{th} and j^{th} sequences using the following definition:

$$diss(x_i, x_j) = 1 - \frac{sim(x_i, x_j)}{\max\limits_{i \neq k}(sim(x_i, x_k))}$$

$i, j, k = 1, ..., n$ where n is the number of sequences. Note some $i \neq j$ this dissimilarity measure could be zero.

The threshold is based on the empirical distribution of 1000 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 1000 sample covariances are obtained from 1000 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.

2.2 THE SPECTRAL ENVELOPE

The spectral envelope of a categorical time series and its application to problems in molecular biology was first introduced by Stoffer et al. (1993). As with the spectral covariance a DNA or amino acid sequence is transformed into a categorical time series by assigning a numerical value to each letter in the amino acid alphabet. The resulting

times series depends on the way in which these numerical values are assigned. For example, consider a sequence ACGTACGTACGTACGT.... and scalings A=1, C=2, G=3 and T=4. In this case, we would get a sequence 1234123412341234.... with periodicity 4 bp. On the other hand if we were to label pyrimidines as 0 and purines as 1 we would get sequence 01010101.... with periodicity 2 bp. In the case of amino acid sequences, we might assign a different numerical value to each amino acid. Or we might assign a numerical value based on whether a given amino acid is hydrophobic or hydrophilic. Clearly, different scalings bring out different properties of the data.

Let X_t, $t = 0, \pm 1, \pm 2, ...,$ be a categorical time series with finite state space $C = \{c_1, c_2, ..., c_k\}$. Assume that X_t is stationary and that $p_j = \mathbf{P}\{X_t = c_j\}$ for $j = 1, 2, 3, ..., k$. For $\beta = (\beta_1, \beta_2, ..., \beta_k)' \in \mathbb{R}^k$ denote $X_t(\beta)$ as the real valued time series corresponding to the scaling that assigns c_j the value β_j. The periodicity of this time series will depend on the choice for β. The idea behind the spectral envelope is to chose $\beta(\omega)$ to maximize the variance at each frequency, ω, relative to the total variance $\sigma^2(\beta(\omega))$. With this in mind, β is chosen such that

$$\lambda(\omega) = \sup_{\beta} \left\{ \frac{f(\omega, \beta)}{\sigma^2(\beta)} \right\},$$

where $f(\omega, \beta)$ is the smoothed spectral density of the time series $X_t(\beta)$. In order to compute $\lambda(\omega)$ note that the categorical time series X_t can be expressed as the multivariate time series Y_t, where $Y_t = e_j$ whenever $X_t = c_j$ and e_j is an index vector with 1 in the jth column and zeros elsewhere. Then $X_t(\beta) = \beta' Y_t$. Assume that Y_t has a continuous spectral density $f_Y(\omega)$. For each ω, $f_Y(\omega)$ is a k x k Hermitian matrix. Let $f_Y^{re}(\omega)$ denote the real part of $f_Y(\omega)$ and V denote the variance-covariance matrix of Y_t. The spectral envelope can then be expressed:

$$\lambda(\omega) = \sup_{\beta} \left\{ \frac{\beta'(\omega) f_Y^{re}(\omega) \beta(\omega)}{\beta'(\omega) V \beta(\omega)} \right\}.$$

The scaling $\beta(\omega)$ is the optimal scaling. Further technical details on the spectral envelope can be found in Stoffer et al. (2000).

The periodicity of the spectral envelope corresponds to the folding patterns of the secondary structures of a protein. Secondary structure elements and motifs are repeated within a domain, hence application of the spectral envelope to protein sequences should reveal multiple peaks associated with the periodicity in the α-helices

and β-sheets, as well as in the repeated motifs. The spectral envelope is smoothed using a triangular weighting which places which places a higher weight on the points near the center frequency and hence gives a more accurate estimate of the spectral density (Collins et al., 2006). Similarly structured proteins should have similar spectral envelopes. It should therefore be feasible to classify proteins into their proper structural categories by examining their spectral envelopes.

Chapter 3

COMBINING DISSIMILARITY MEASURES USING SCALE COEFFICIENTS

In this thesis we extend these analyses to multi-gene 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 (Philippe et al., 2005b; Rokas et al., 2003). 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 multi-gene data sets (Bininda-Emonds, 2004; Bull et al., 1993; Burleigh et al., 2006; de Queiroz and Gatesy, 2007; Gatesy et al., 2004; Philippe et al., 2005b). 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 analyse 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; Lecointre, 2005; Leigh et al., 2008; Zelwer and Daubin, 2004). 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 (Jeffroy et al., 2006; Philippe et al., 2005b).

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.

3.1 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) used 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.

3.1.1 THE MINIMUM VARIANCE 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

$$D = \begin{pmatrix} d_{1,1} & d_{1,2} & ... & d_{1,k} \\ d_{2,1} & d_{2,2} & ... & d_{2,k} \\ .. & .. & ... & .. \\ .. & .. & ... & .. \\ d_{p,1} & d_{p,2} & ... & d_{p,k} \end{pmatrix}. \tag{3.1}$$

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

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

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

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

$$V = \sum_{i=1}^{p} V_i \;\; = \;\; \sum_{i=1}^{p} \left[\frac{1}{k} \sum_{j=1}^{k} (c_j d_{i,j})^2 - \left(\frac{1}{k} \sum_{j=1}^{k} c_j d_{i,j} \right)^2 \right] \tag{3.2}$$

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

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

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

3.1.2 THE MINIMUM SQUARED COEFFICIENT OF VARIATION SCALE COEFFICIENTS

An alternative method is to minimize the squared coefficient of variation. 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 favour 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 (3.2) we now wish to minimize the sum of the squared CV:

$$\sum_{i=1}^{p} CV_i^2 = \sum_{i=1}^{p} \left(\frac{\frac{1}{k}\sum_{j=1}^{k}(c_j d_{i,j})^2 - \left(\frac{1}{k}\sum_{j=1}^{k} c_j d_{i,j}\right)^2}{\left(\frac{1}{k}\sum_{j=1}^{k} c_j d_{i,j}\right)^2} \right) \tag{3.3}$$

$$\propto \sum_{i=1}^{p} \left(\frac{\sum_{j=1}^{k}(c_j d_{i,j})^2}{\left(\sum_{j=1}^{k} c_j d_{i,j}\right)^2} \right) - \frac{p}{k}. \tag{3.4}$$

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 minimum variance scale coefficients. We then use the non-linear minimization function nlm() available in the R package nlme (Pinheiro et al., 2009) to find the set of scale coefficients $c = (c_1, c_2, ..., c_k)$ (with $c_1 = 1$) that minimizes equation (3.3). The combined pairwise dissimilarities for the k genes is then the mean of the scaled dissimilarities, $\frac{1}{k} D_s \mathbf{1}$, where $\mathbf{1}' = (1, 1, ..., 1)_{1\times k}$.

3.2 METHODS TO BUILD TREES

To obtain phylogenetic trees from our combined dissimilarity matrix across genes, two distance based tree building methods, i.e. 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 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 neighbor-joining algorithm but has a greater computational cost (Kuhner and Felsenstein, 1994).

3.3 BOOTSTRAP PERMUTATION METHODS

To obtain an empirical distribution of the spectral covariance based dissimilarity, we use a re-sampling 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 to 3 residues. While the length of loops can vary, it is known that turns have a periodicity of 3 to 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 to 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 bi-partitions 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.

3.4 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 SEQGEN (Rambaut and Grassly, 1997) under the JTT model with no variation of rates among sites. It is important to note that the JTT model of evolution assumes that sequence sites evolve independently and thus sequences simulated with SEQGEN will not necessarily retain the structural information present in the sequences. However, since the sequences simulated by SEQGEN 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.

3.5 DATA

Four different data sets are used in this paper to illustrate our methods. We begin with an exploratory analysis on a non-controversial eukaryote data set provided courtesy of Dr. Andrew Roger (Centre 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 (Wu and Susko, 2009; Gruenheit et al., 2008; Ané et al., 2004). 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 tables in the appendix. 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 twelve 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. Proponents of the ecdysozoa hypothesis argue that all animals which shed their shells should be grouped together. These are referred to as moulting animals and include nematodes and arthropods. By contrast, proponents of coelomata hypothesis believe that animals should be grouped based on whether or not they have a coelom (body-cavity). Hence, under the coelomata hypothesis, vertebrates and arthropods should be grouped together (Telford, 2004). Aguinaldo et al. (1997) first proposed a clade of moulting animals 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 *Caenorhabditis 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) analysed 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 et al. (2005a) 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 (Soltis et al., 1999; Qiu 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 was the cause of the moncot-first topology. Still, the true relationships among the angiosperms is 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 thirteen mitochondrial protein-coding genes. The phylogeny for the primate data set is fairly well established though there remains some debate over the exact relationship between gorilla, chimp and human. It is generally believed that human and chimp should be placed together as sister taxa, though for some portions of the genome gorilla and human appear to be more closely related (Ruvolo, 1997; Hobolth et al., 2007).

Throughout this thesis we have tried to refer to the taxa as they have been referred to in the literature. Taxa in the eukaryote data set and the chloroplast data set are referred to by their Latin names. For the nematode data set we follow the example of Foster and Hickey (1999) and refer to the taxa by their common names with the exception of *Allomyces macrogynus* which is an ancestral fungus and has no common name. Likewise, we refer to the primate data by their common names as was done by Collins et al. (2006).

3.6 RESULTS

3.6.1 RESULTS ON EUKARYOTE DATA SET

We begin by applying all combinations of methods on the non-controversial 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 taxon-specific scaling covariance. Both MinVar and MinCV criteria were used to obtain scale coefficients with which to combine genes. Figure 3.1 shows the reference tree for the eukaryote data (Keeling et al., 2009).

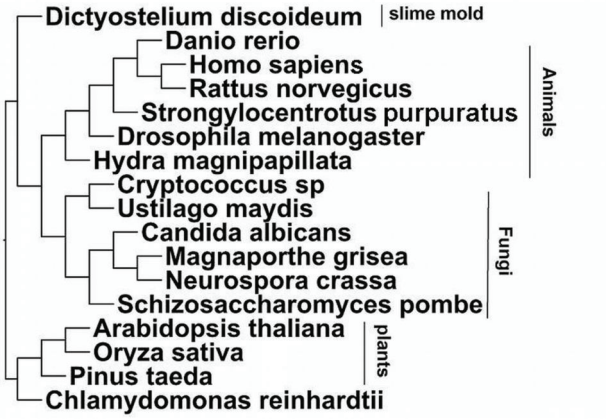

Figure 3.1: Reference tree topology from Tree of Life web project for the eukaryote data set with 17 taxa (http://tolweb.org/Eukaryotes) (Keeling et al., 2009)

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 3.2 shows the pairwise scatter plots with regression lines for the dissimilarities from each pair of methods.

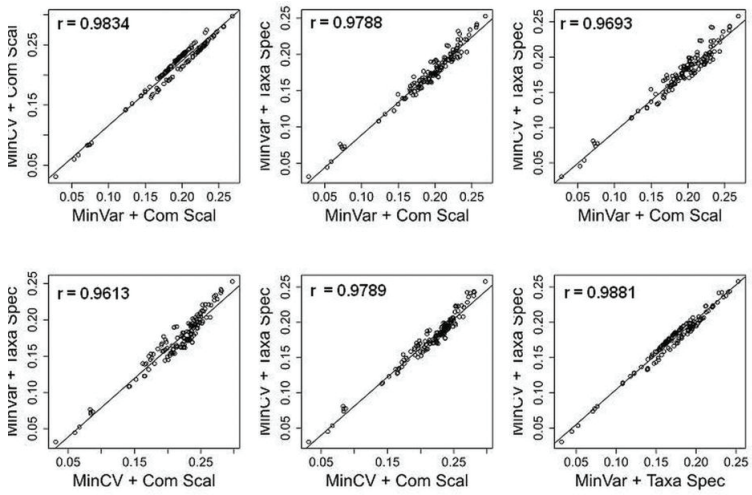

Figure 3.2: Comparisons of the eukaryote data set dissimilarities obtained with common scaling (ComScal) and taxon-specific scaling (TaxaSpec) methods combined with MinVar and MinCV criteria (with Pearson correlation coefficient, r).

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.

Trees are obtained using both BIONJ and FITCH. Thus, there are in total eight different methods to build trees. The inferred trees by the eight different methods are shown in Figures 3.3 and 3.4. All trees recover the major clades of plants, animals and fungi. The MinCV taxon-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 *Schizosaccharomyces 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 *Schizosaccharomyces pombe* and *Candida albicans*.

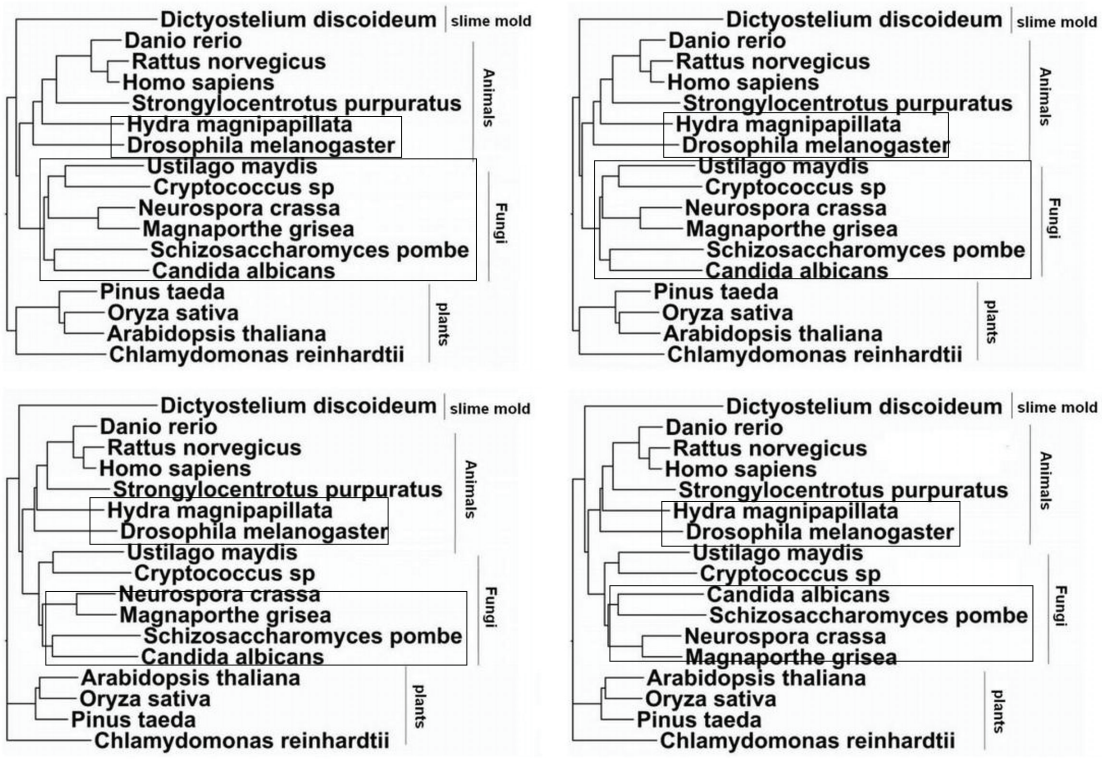

Figure 3.3: Estimated BIONJ and FITCH trees for eukaryote data set when common scaling and taxon-specific scaling dissimilarities for multiple genes are combined with the MinVar criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxon-specific scaling with BIONJ (top right) and FITCH (bottom right). Boxes indicate portions of the tree which differ from reference tree.

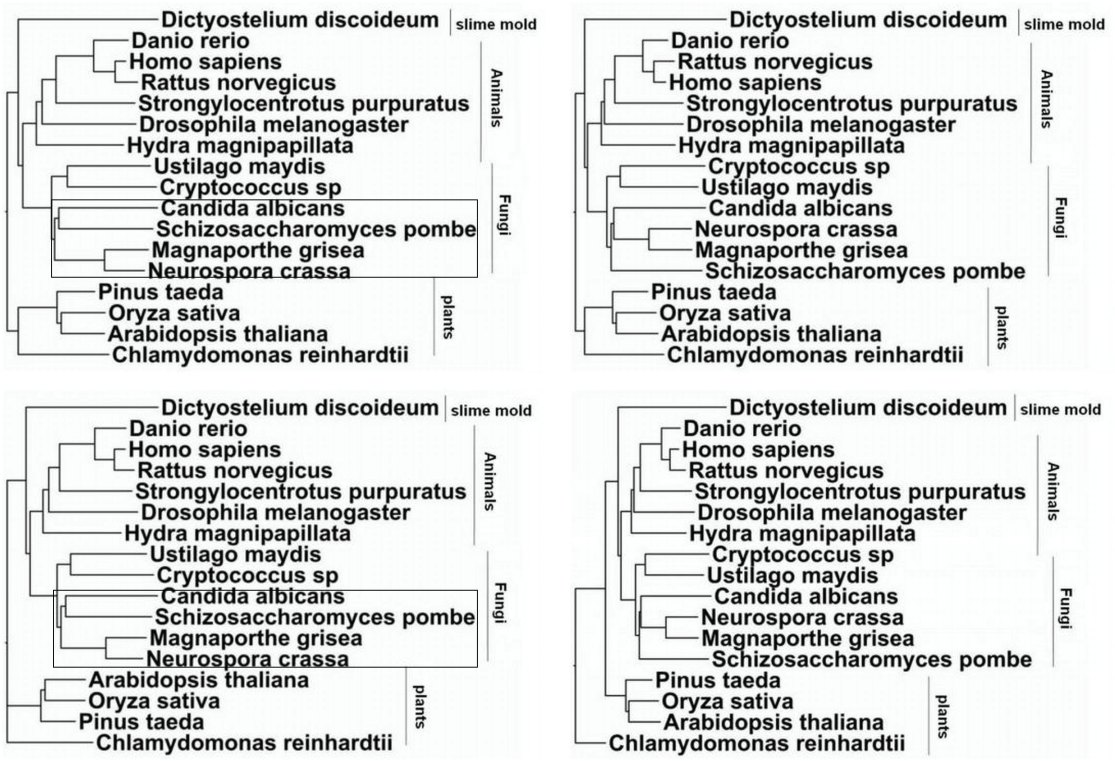

Figure 3.4: Estimated BIONJ and FITCH trees for eukaryote data set when common scaling and taxon-specific scaling dissimilarities for multiple genes are combined with the MinCV criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxon-specific scaling with BIONJ (top right) and FITCH (bottom right). Boxes indicate portions of the tree which differ from reference tree.

Table 3.1 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 *Dictyostelium discoideum* as basal in the animal clade has 100% bootstrap support in both the MinCV taxon-specific scaling trees, but only weak support under the other 6 methods. The recovery of the reference tree topology within the three main clades has strong bootstrap support in all 4 MinCV trees. In the MinVar trees, the incorrect placement of *Drosophila 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 *Neurospora crassa* and *Magnaporthe grisea* with *Ustilago maydis* and *Cryptococcus sp* rather than

Schizosaccharomyces pombe and *Candida albicans* (99% under the common scaling method and 85% under the taxon-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 3.2 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. While all methods give comparable results, one can see that the taxon-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 taxon-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 taxon-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. Taxon-specific scaling covariance trees have a maximum RF distance of 4 with the majority of trees having distances less than 2. The common scaling covariance trees have a maximum RF distance of 6 with majority of trees having distances less than 4.

The MinCV method with the taxon-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.

While 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. For the remaining two data sets we present the results obtained using the MinCV for both the common scaling and taxon-specific scaling

covariance based dissimilarity measures but note that similar trees were obtained with the MinVar method. For cases where the MinCV and MinVar methods differed, the MinCV consistently returned the more accurate topology.

3.6.2 RESULTS ON NEMATODE DATA SET

The nematode data set consists of twelve 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 favours 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 3.5 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 (*Caenorhabditis elegans*) being branched as sister taxa with strong bootstrap support(Foster and Hickey, 1999).

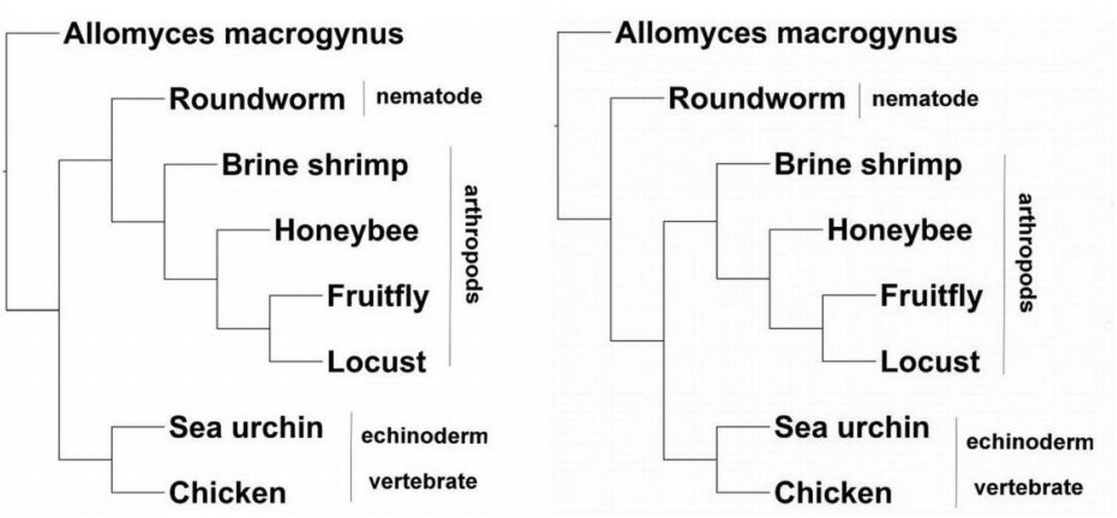

Figure 3.5: Reference topology for the nematode data set under the ecdysozoa hypothesis (left) and coelomata hypothesis (right) (Blair et al., 2002).

Again we begin with an exploratory analysis of the dissimilarities computed from the common scaling and taxon-specific scaling covariances with MinCV scale coefficients. A scatter plot with regression of the MinCV dissimilarities under the taxon-specific scaling versus the common scaling method is shown in Figure 3.6. 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 non-controversial with regards to their placement in the tree. However, it is possible that the pairwise distances beneath the regression line are pulling the model away from this point.

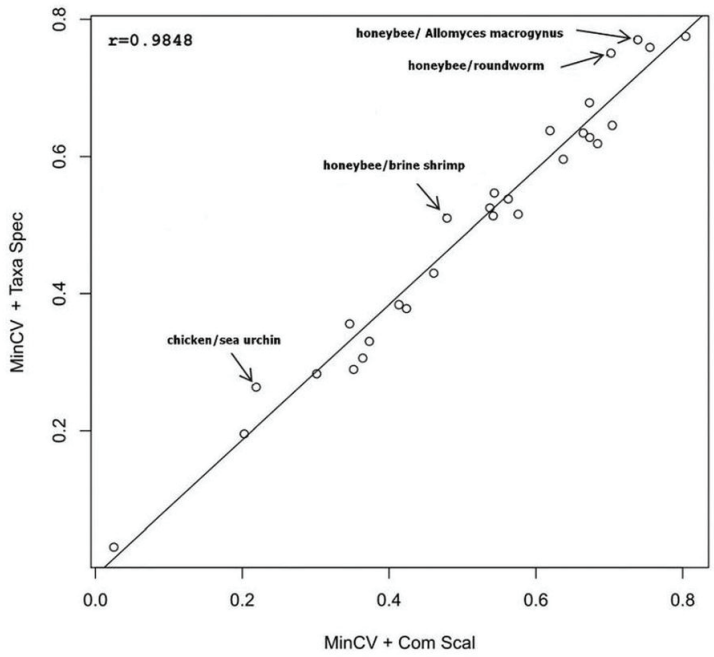

Figure 3.6: Comparisons of the nematode data set dissimilarities computed with the common scaling (ComScal) method and combined with the MinCV criterion and the taxon-specific scaling (TaxaSpec) method combined with the MinCV criterion (Pearson correlation = 0.9848). Taxa pairs with largest discrepancy in dissimilarities computed under these two methods shown with arrows.

The MinCV trees obtained with the four methods are shown in Figure 3.7. 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 taxon-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, while the taxon-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.

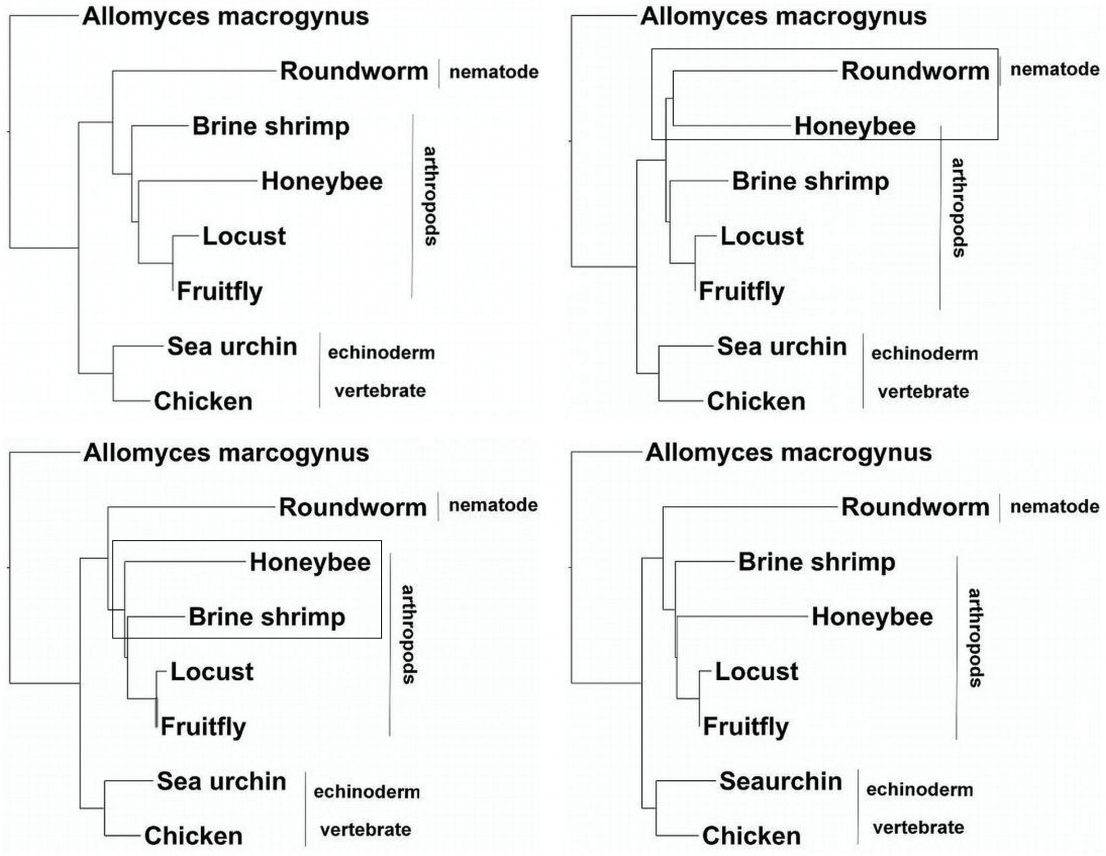

Figure 3.7: Estimated BIONJ and FITCH trees for the nematode data set when common scaling and taxon-specific scaling dissimilarities for multiple genes are combined with the MinCV criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxon-specific scaling with BIONJ (top right) and FITCH (bottom right).

Table 3.3 shows the bootstrap support for the topological features for the nematode tree under different methods. 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). 52% of the taxon-specific scaling BIONJ trees support the ecdysozoa hypothesis topology, while 67 % of the taxon-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 taxon-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 taxon-specific BIONJ tree, and weak bootstrap support in the taxon-specific FITCH tree. Separation of the honeybee and the roundworm occurs in 63% of bootstrap trees for both BIONJ and FITCH common scaling methods, 62% of bootstrap trees for the taxon-specific BIONJ method, and 69% of bootstrap trees for the taxon-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 3.4 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 taxon-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 to 0.8091 and 0.7246 for the corresponding taxon-specific scaling trees.

The Robinson-Foulds distances show the same pattern. For the nematode data set with 5 interior nodes the maximium possible value for RF distance is 10. Taxon-specific scaling trees have a maximum distance of 6 with the majority of distances being 4 or less. The common scaling covariance based BIONJ trees have a maximum distance of 4 with the majority of trees having values less than 2. The common scaling covariance based FITCH trees have a maximum distance of 2, with 56 out of the 100 RF distances being 0.

3.6.3 RESULTS ON CHLOROPLAST DATA SET

The chloroplast data set consists of twenty-five chloroplast proteins from twenty-two taxa. There has been some debate over the placement of *Amborella trichopoda* within the angiosperms. Most analyses place *Amborella trichopoda* as the most basal angiosperm (Soltis et al., 1999; Qiu et al., 1999; Zanis et al., 2002), though in some cases an *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 3.8 shows the reference tree for the chloroplast data (Soltis et al., 2005; Ané et al., 2004).

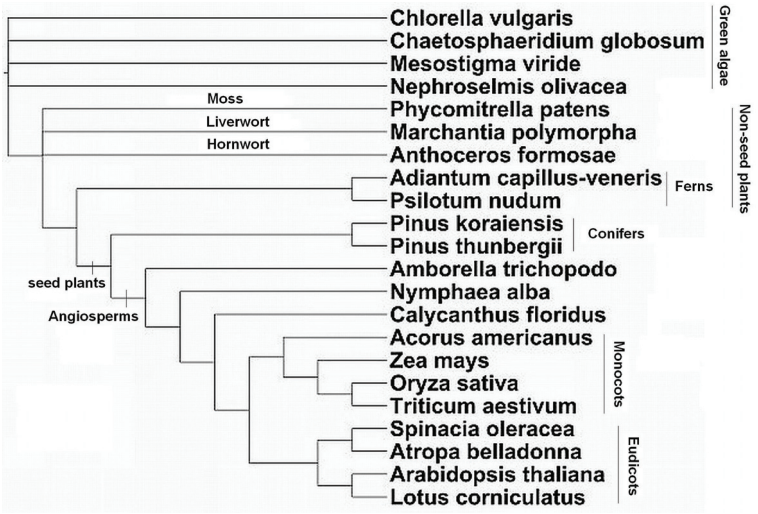

Figure 3.8: Reference tree topology for the chloroplast data set with 22 taxa (Soltis et al., 2005; Ané et al., 2004).

We begin with an analysis on all twenty-five 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 nineteen out of the twenty-five chloroplast proteins. The same twenty-two 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

taxon-specific scaling versus common scaling dissimilarities is shown in Figure 3.9. Once more correlation between the two methods is fairly high with r=0.9639.

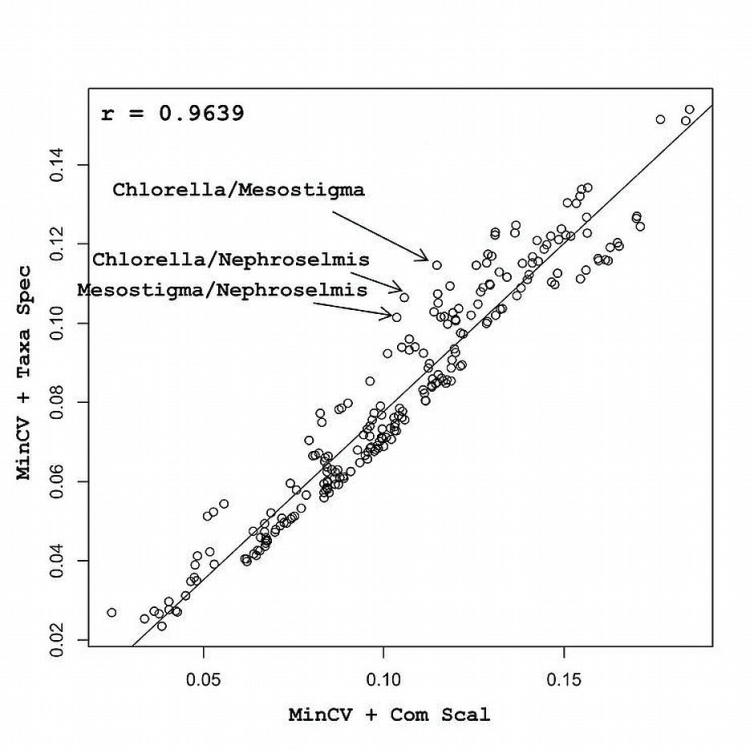

Figure 3.9: Comparisons of the chloroplast data set dissimilarities computed with the common scaling (ComScal) method and combined with the MinCV criterion and the taxon-specific scaling (TaxaSpec) method combined with the MinCV criterion (Pearson correlation = 0.9639). Taxa pairs with the largest discrepancy in dissimilarities computed under these two methods shown with arrows.

The real data trees are shown in Figure 3.10. In all four trees the separation of green algae, non-seed 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 taxon-specific scaling tree place *Amborella trichopoda* and *Nymphaea alba* as sister taxa, while the common scaling tree places *Calycanthus floridus* and *Amborella trichopoda* as sister taxa. In all four trees a clade with *Amborella trichopoda*, *Nymphaea alba* and *Calycanthus floridus* is basal in the angiosperm clade.

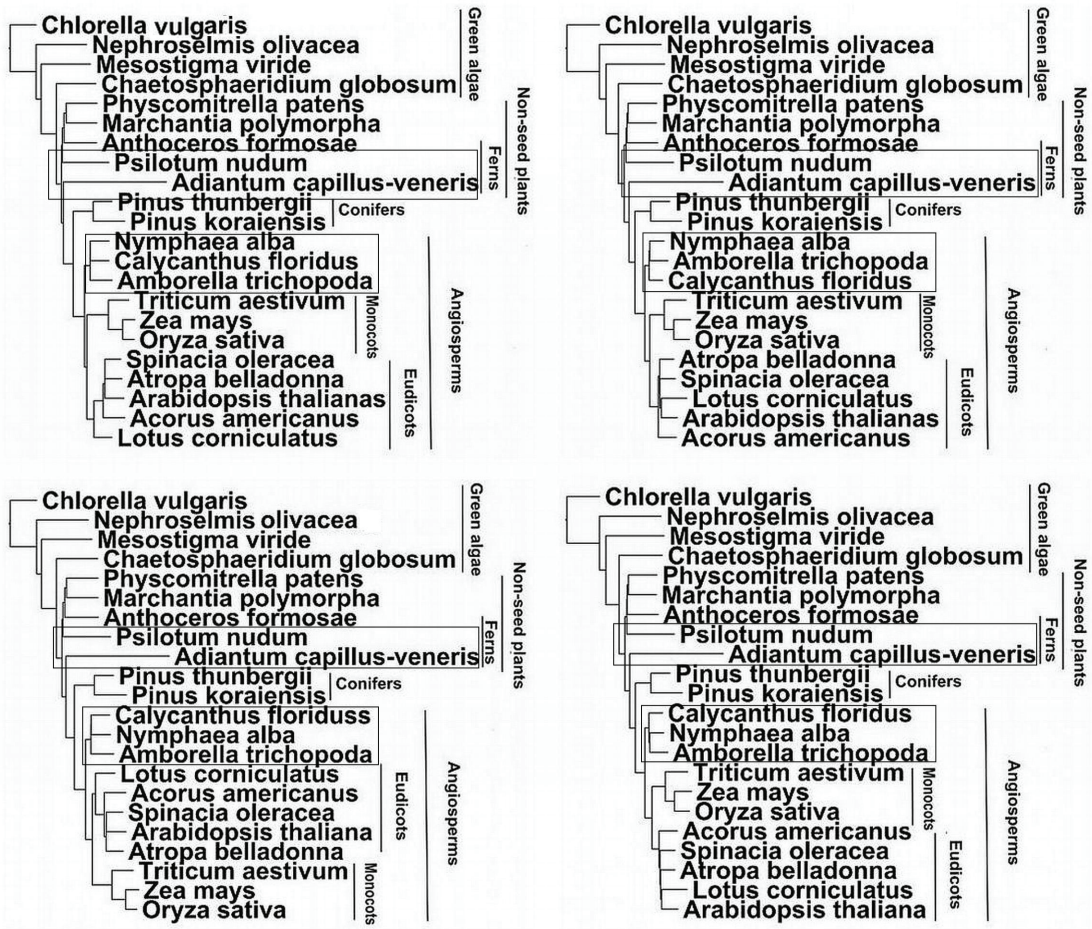

Figure 3.10: Estimated BIONJ and FITCH trees for the chloroplast data set when common scaling and taxon-specific scaling dissimilarities for multiple genes are combined with the MinCV criterion. Common scaling with BIONJ (top left) and FITCH (bottom left), taxon-specific scaling with BIONJ (top right) and FITCH (bottom right). Boxes indicate portions of the tree which differ from reference tree.

Table 3.5 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, non-seed plants, seed plants and angiosperms has 100% bootstrap support in all four methods. There is also strong bootstrap support for a clade with *Amborella trichopoda*, *Nymphaea alba* and *Calycanthus floridus* as basal in the angiosperm clade (100% for all four methods). The branching of *Amborella trichopoda* and *Nymphaea alba* as sister taxa has moderate support in the common scaling FITCH tree and both taxon-specific trees (51% to 66%). The branching of *Nymphaea alba*

and *Calycanthus floridus* as sister taxa is strongly supported by the common scaling BIONJ tree. The branching of the two ferns, *Psilotum nudum* and *Adiantum capillus-veneris*, as sister taxa has moderate support in both taxon-specific trees (55% with BIONJ and 52% with FITCH). The erroneous placement of *Psilotum nudum* with the mosses and liverwort seems to occur in all common scaling BIONJ trees and 78% of the common scaling FITCH trees.

Table 3.6 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 taxon-specific trees being slightly higher than the means for the common scaling trees. For the taxon-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 taxon-specific scaling distances. For the chloroplast data set with twenty-two 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, while the taxon-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 twenty-five genes was somewhat surprising as analyses on a subset of nineteen genes of these twenty-five resolved the angiosperm clade very differently. When only nineteen genes were included in the analyses all methods returned the erroneous monocot-first tree with strong bootstrap support. Removing those of the nineteen 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 *Amborella trichopoda*, *Nymphaea alba* and *Calycanthus floridus* being basal in the angiosperm clade. Figure 3.11 shows the common scaling MinCV distances of non-monocot angiosperms versus the three monocots when nineteen and twenty-five genes are used in the analyses.

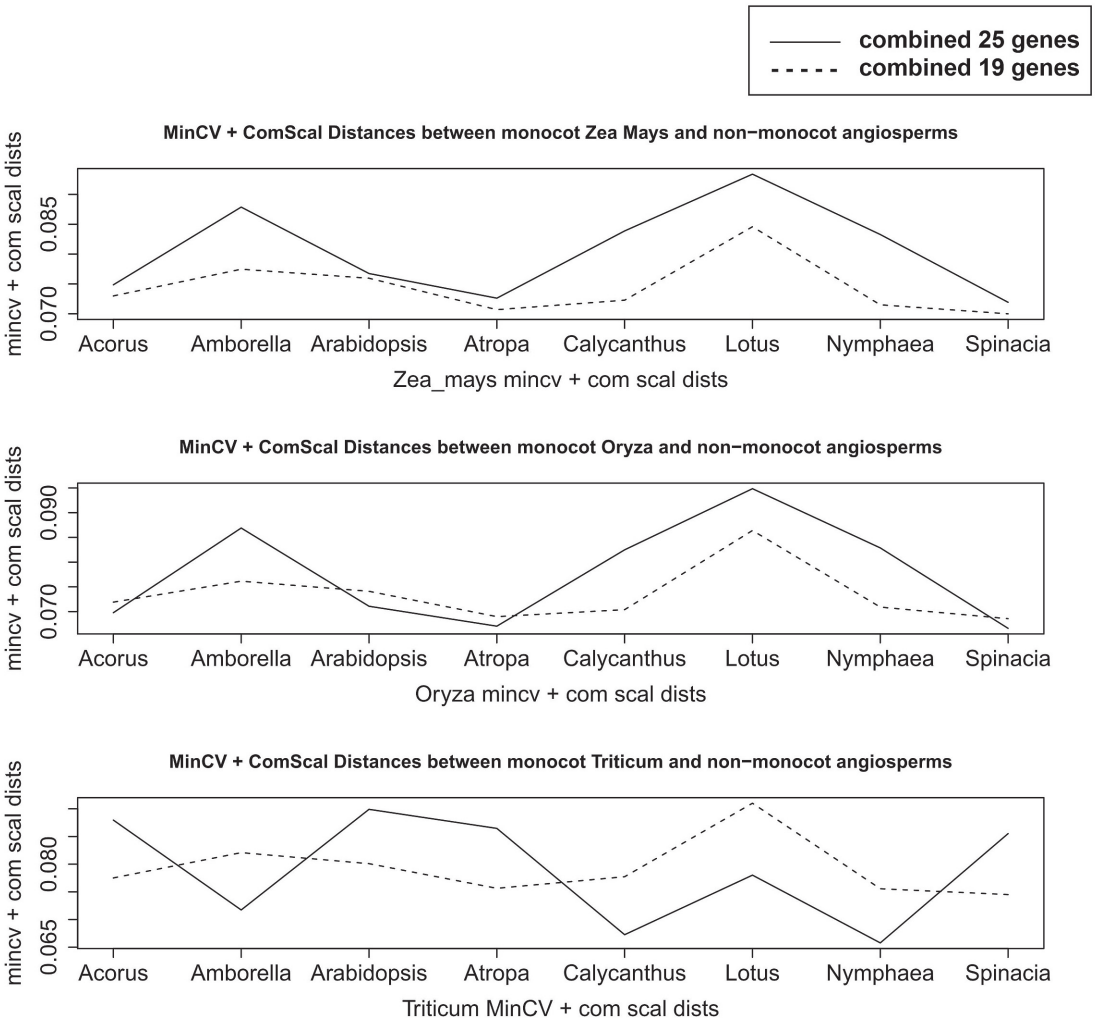

Figure 3.11: Common scaling distances for monocots versus other angiosperms for 19 and 25 genes combined with MinCV.

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 *Amborella trichopoda*, *Nymphaea alba* and *Calycanthus floridus*, while the corresponding distances between these two monocots and the eudicots is only slightly greater except in the case of *Lotus corniculatus*. The nineteen gene MinCV common scaling distances between *Triticum aestivum* the eudicots tend to be greater while the distances between *Triticum aestivum* and the *Amborella trichopoda*, *Nymphaea alba* and

Calycanthus floridus are smaller. The distance between *Triticum aestivum* and the other two monocots is also greater within the monocot clade. The taxon-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.

3.7 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 SEQGEN. We reduced both the sequence similarity and the structure similarity in the simulated sequences by increasing the branch lengths and these results are compared to 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.

3.7.1 SIMULATIONS GENERATED FROM PRIMATE DATA SET

Figure 3.12 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.

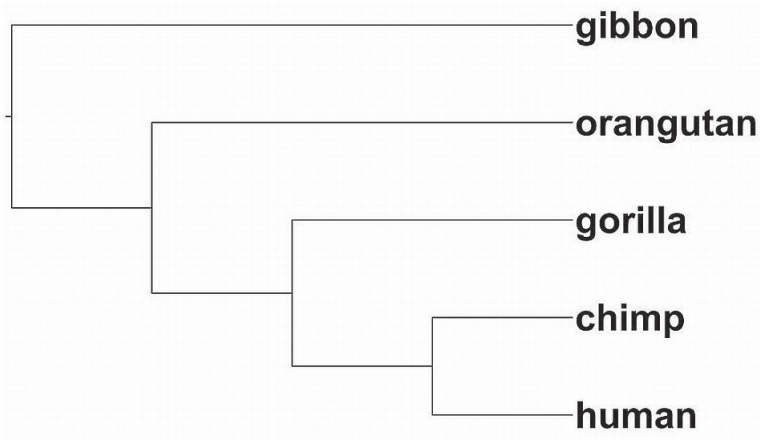

Figure 3.12: Reference tree for the primate data set from http://tolweb.org

Throughout the whole simulation the sequence of gibbon is specified as the ancestral sequence. We first simulated 1000 data sets for each gene based on the tree shown in Figure 3.12 with branch lengths estimated by common scaling MinCV method, and call this simulation scheme S1. We then repeated this process to create two additional sets of 1000 data sets in which the branch lengths of the input tree are multiplied by 100 and 1000, and we call these two simulations schemes S100 and S1000, respectively. We compared the analysis performed on these simulated data to the analysis performed on the block bootstrap permutations on all eight combinations of methods.

For the 1000 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 3.13A 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 3.13A.

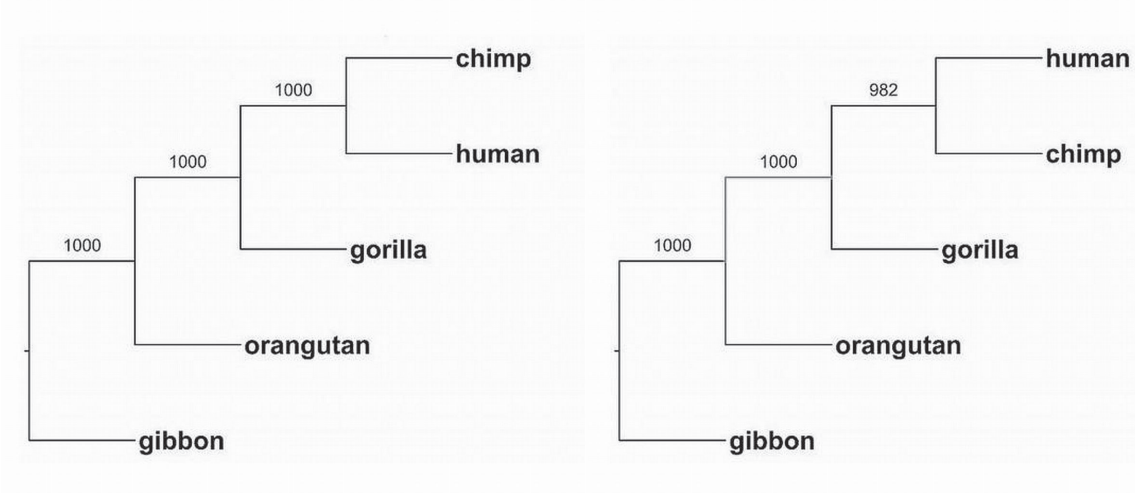

Figure 3.13: Primate majority-rule consensus trees estimated with the common scaling (ComScal) method and combined with the MinCV criterion for 1000 SEQGEN simulated sequences (left) and 1000 block bootstrap permutation sequences (right).

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 between 10% and 50% for the S100 scheme and less than 10% for the S1000 scheme. Table 3.7 shows the proportion of trees which recover the reference tree under the different simulation schemes. We also performed 1000 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 3.7 shows the proportion of correctly estimated trees under the block bootstrap permutations.

For the primate data, with average sequence similarities greater than 10% all of the trees based on simulated sequences recover the reference tree. When sequence similarity is less than 10%, only 4% to 10% of estimated trees recover the topology of the reference tree. For the block bootstrap permutation samples, 88% to 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.

3.7.2 SIMULATIONS GENERATED FROM NEMATODE DATA SET

For the nematode data set the sequence of *Allomyces macrogynus* is specified as the ancestral sequence for SEQGEN simulations. Again we simulated 1000 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 simulation scheme S1, and repeat the process to create two additional sets of 1000 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%, while sequence similarities under simulation schemes S25 and S100 are reduced to 2% to 30% and 0% to 13.6%, respectively.

The input tree topology for SEQGEN 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 which recovers trees which agree with the reference topology under the ecdysozoa hypothesis, our data was simulated under this topology rather than that under the coelomata hypothesis. Table 3.8 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% to 93% recover the input tree. For data generated under simulation scheme S25, 44.5% to 46.4% of the estimated common scaling trees recover the input tree while the recovery rates of estimated taxon-specific scaling trees are 0% to 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% to 95.6% recover the tree that agrees with the ecdysozoa topology. Of the taxon-specific scaling trees based on bootstrap permutations, 42% to 79% recover the tree that agrees with the coelomata hypothesis.

3.7.3 ANALYSIS OF SIMULATION RESULTS

For data generated with SEQGEN, 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 taxon-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 taxon-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 while the estimated taxon-specific scaling trees show moderate to strong support to the coelomata hypothesis. While 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.

3.8 PERMUTATIONS

To further validate the covariance based methods we performed further analyses on 1000 block size 1 permutation samples taken from the nematode data set. Permutations were computed using the SEQBOOT programme in PHYLIP (Felsenstein, 1989). We compared the distances obtained from the block size 1 permutation samples to those obtained from the block size 14 permutations. Recall 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 taxon-specific scaling covariance based dissimilarities.

Figure 3.14 shows common scaling MinCV majority-rule consensus trees obtained from both sets of permutation samples. While 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.

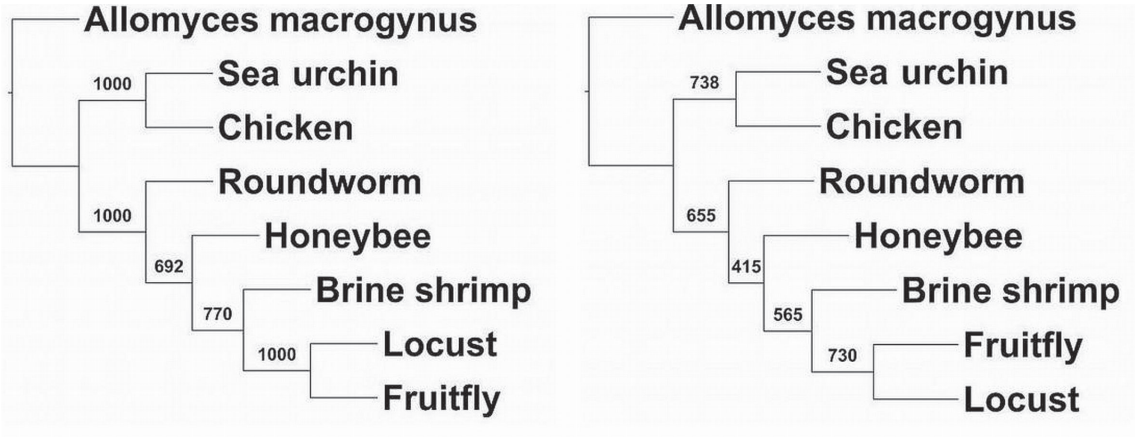

Figure 3.14: Nematode majority-rule consensus trees estimated with the common scaling method and combine with MinCV criterion for 1000 block size 14 permutation samples (left) and 1000 block size 1 permutation samples (right).

Boxplots of the 1000 pairwise distances under both permutation schemes for three pairs of taxa randomly selected from the twenty-eight taxa pairs are shown in Figure 3.15. The horizontal line across the x-axis corresponds to the common scaling MinCV distance obtained from the real data. The 1000 bootstrap distances under the block size 14 permutation scheme have much smaller variance than the 1000 bootstrap distances under the block size 1 permutation scheme, while 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. While 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 1000 block size 1 permutation distances.

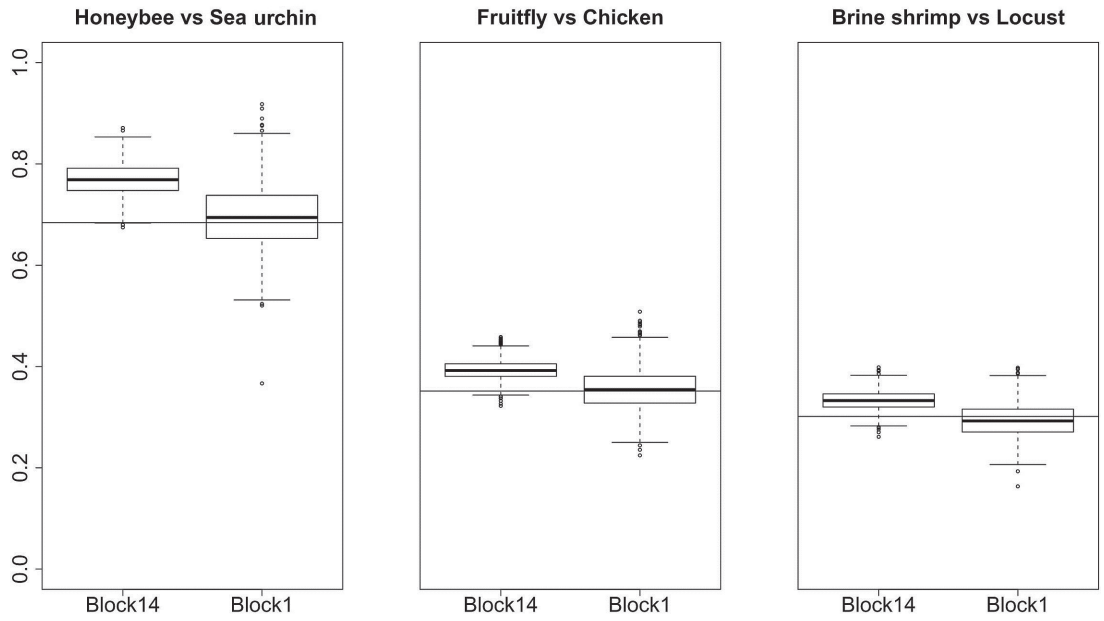

Figure 3.15: Boxplots of 1000 sample distances obtained from block size 14 permutation samples (left) and block size 1 permutation samples (right) for three randomly selected taxa pairs from the nematode data set. Horizontal line across the x-axis corresponds to the common scaling MinCV distance for real data.

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

3.9 INFLUENCE OF INDIVIDUAL GENES ON THE COMMON SCALING MINCV TREE FOR THE NEMATODE DATA

There is much debate surrounding the true topology of the nematode tree and the individual gene trees return conflicting topologies. It may be of interest to determine which genes appear to have the most influence in determining the combined-gene tree obtained with the MinCV coefficient scales. Following the classic approach to influence analysis we consider how removal of a point (or a gene) affects our estimated tree and how addition of an outlier (in our case a nonsense gene) affects our

estimated tree. To do this we consider the estimated combined-gene topology a summary measure of the relative distances between taxa. To determine the influence of individual genes on the combined-gene topology, we remove a single gene from the analysis and examine how the combined distance matrix has changed and how this change has affected the estimated combined-gene tree. Note that we can think of this as assigning one of the genes a scale coefficient of zero. Lastly, we introduce an outlying gene to the data to determine how the presence of this outlier affects our estimated tree. This outlying gene is obtained in the following manner. For each of the twelve genes we have eight sequences. Hence, there are a total of 96 sequences in the nematode data set. These 96 sequences were concatenated to get one long sequence of 25568 characters. Eight sequences of length 200 were then randomly sampled with replacement from the 25568 characters. Taxa names were then randomly assigned to the sampled sequences. The eight randomly generated sequences were then added to the analysis as an outlying gene. For the analysis in this section, we use the common scaling covariance based dissimilarity matrices.

We begin by computing twelve trees based on the common scaling spectral covariance dissimilarity matrices for each gene in the nematode data set. Trees were computed using BIONJ (Gascuel, 1997). A table of the topologies obtained for the twelve genes is shown in Table 3.9. This is not a table of all possible topologies but only those topologies recovered by the individual genes and the combined-gene tree. The estimated topologies of the twelve genes all differ from both each other and the combined-gene tree by two or more branches. The third column of Table 3.9 shows the Robinson-Foulds distances between the estimated tree and theoretical topology under the ecdysozoa hypothesis. The combined-gene tree, which has the same topology as the ecdysozoa tree, has a RF distance of 0. ATP6 returns the next closest tree with a RF distance of 2.

Next, individual genes were removed from the analysis one at a time. Recall in the computation of the MinCV we fix the coefficient of one gene to be one. For the nematode data set we selected an arbitrary gene, ND3, to have a fixed coefficient of one in the computation of the MinCV scale coefficients for the twelve genes. For the computations involving only eleven genes this same gene was assigned the fixed

coefficient except for when ND3 itself was removed from the analysis, in which case ND2 was assigned the fixed coefficient of one. The difference in the combined-gene distances for the full twelve genes and those for eleven genes was then computed. Figure 3.16 shows the change in the pairwise distances obtained with the MinCV scale coefficients after each gene is removed.

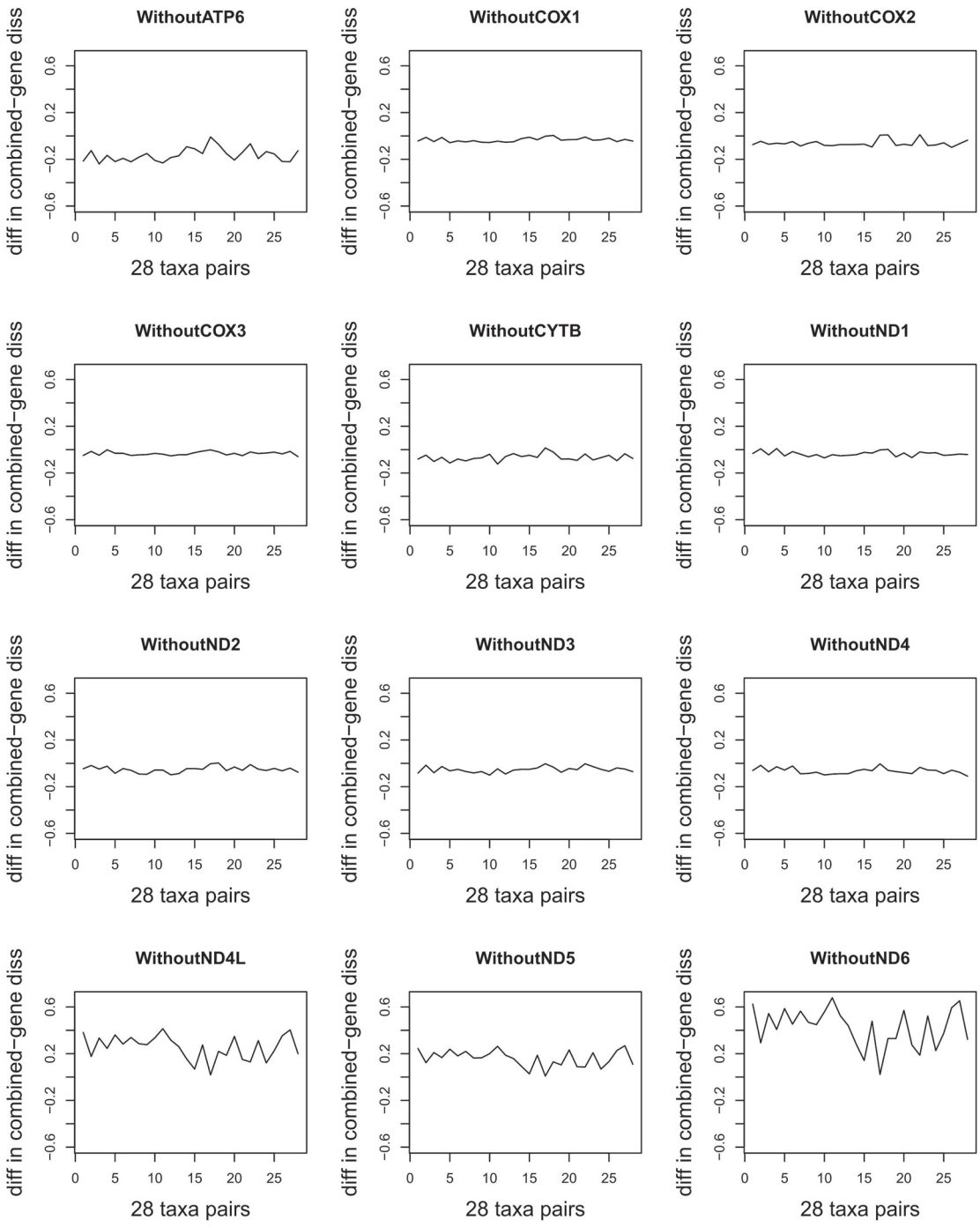

Figure 3.16: Change in the combined-gene pairwise distances obtained with the MinCV scale coefficients when each gene is removed from the analysis

One can see that for the most part there is only a very minor change in the combined-gene tree distances when a gene is removed except in the case of ATP6, ND4L, ND5 and ND6. When ATP6 is removed, the estimated combined-gene pairwise distances are greater than when all twelve genes are used. Peaks for some of the taxa pairs involving the nematode and the arthropods indicate that the relative relationship between these taxa has been affected by the removal of this gene. The other genes whose removal appears to affect the estimated combined-gene pairwise distances are ND4L, ND5 and ND6. When these genes are removed, the estimated distances appear to be larger than when all twelve genes are used. Again, peaks corresponding to pairwise distances which involve the nematode and the arthropods suggest that the relative relationships between these taxa have been affected by removal of these genes.

Table 3.10 shows the topologies estimated from eleven genes. As expected, only four trees differ from the tree based on all twelve genes: the tree without gene ATP6, the tree without gene ND4L, the tree without gene ND5 and the tree without gene ND6. When ATP6, ND4L or ND6 are removed from the analysis, honeybee and roundworm are placed together as sister taxa. When ND5 is removed honeybee is erroneously placed as the most basal of the arthropods. The trees based on combined information from subsets of genes (ATP6, ND4L, ND5,ND6) and (ATP6, ND4L,ND6) both have the same topology as the twelve gene tree. This implies that ATP6, ND4L and ND6 are influential in the proper separation of roundworm and honeybee in the tree. While removing ND5 does not cause honeybee and roundworm to be branched as sister taxa, the estimated tree still places honeybee closer to roundworm than it should be if the tree under the ecdysozoa hypothesis is the true tree. Hence, ND5 also plays some role in the correct placement of roundworm and honeybee relative to each other.

Next, the outlying gene was added to the analysis. To confirm that this gene really is an outlier, we computed the BIONJ tree for the common scaling based dissimilarities of the outlying gene. We then compared the dissimilarities obtained from the outlying gene to the combined-gene dissimilarities for the twelve genes in the nematode data set. The tree obtained with the common scaling covariance based

dissimilarities for the outlying gene is shown in Figure 3.17. In this tree, all the arthropods happen to be in the same clade, but the nematode branches with the sea urchin, and chicken is placed as an outgroup to the roundworm/sea urchin sister clade and the arthropods. None of the twelve genes in the nematode data set recover this topology.

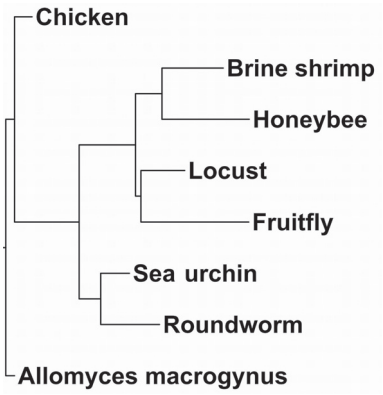

Figure 3.17: Estimate tree for the outlying gene derived from the common scaling covariance based dissimilarities.

The common scaling based pairwise dissimilarities for the nonsense gene against the MinCV combined-gene pairwise dissimilarities of the twelve genes in the nematode data set are shown in Figure 3.18. We can see that the pairwise distances of the outlying gene and the combined-gene pairwise distances for the real genes are not linearly correlated. This suggests the randomly generated sequences do represent an outlying observation.

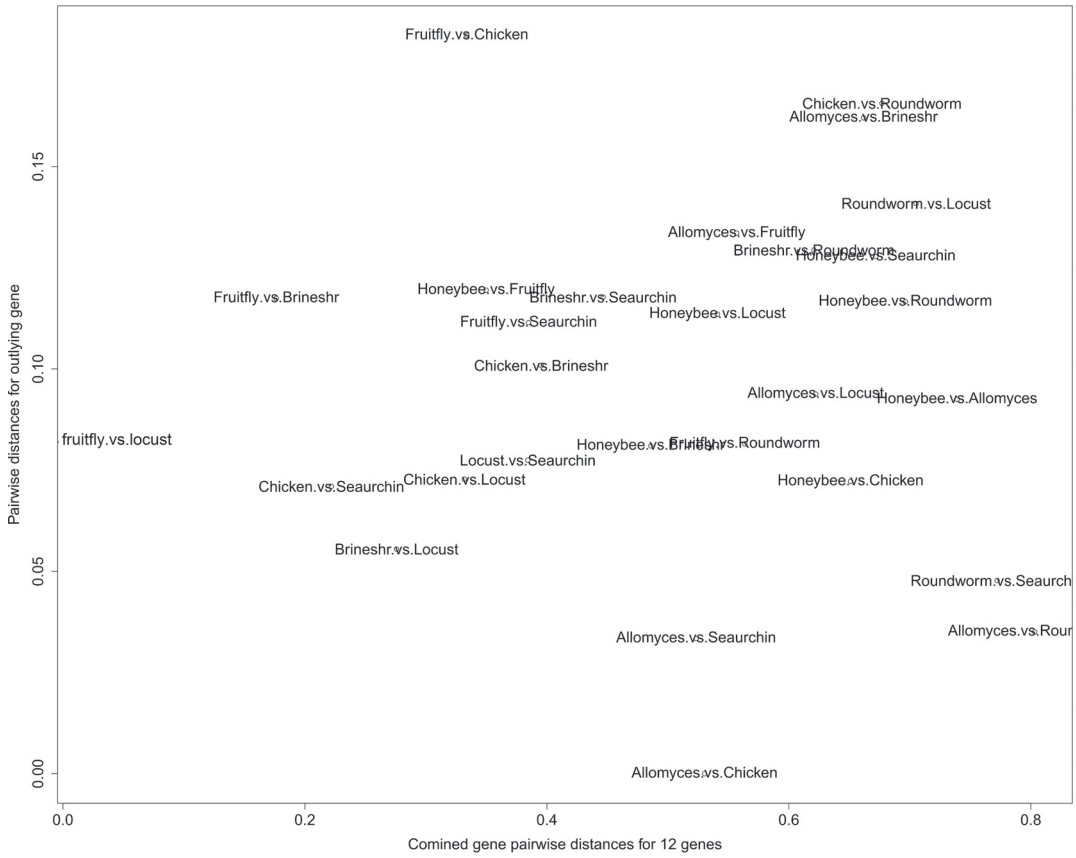

Figure 3.18: The common scaling based dissimilarities for a randomly generated outlying gene vs Common scaling covariance based MinCV combined-gene pairwise dissimilarities.

Figure 3.19 shows the change in the pairwise distances obtained with the MinCV scale coefficients when an outlying gene is added to the analysis. Adding an outlier appears to have a negligible effect on the combined-gene pairwise distances. Indeed, the estimated BIONJ tree is the same as that obtained without the outlying gene.

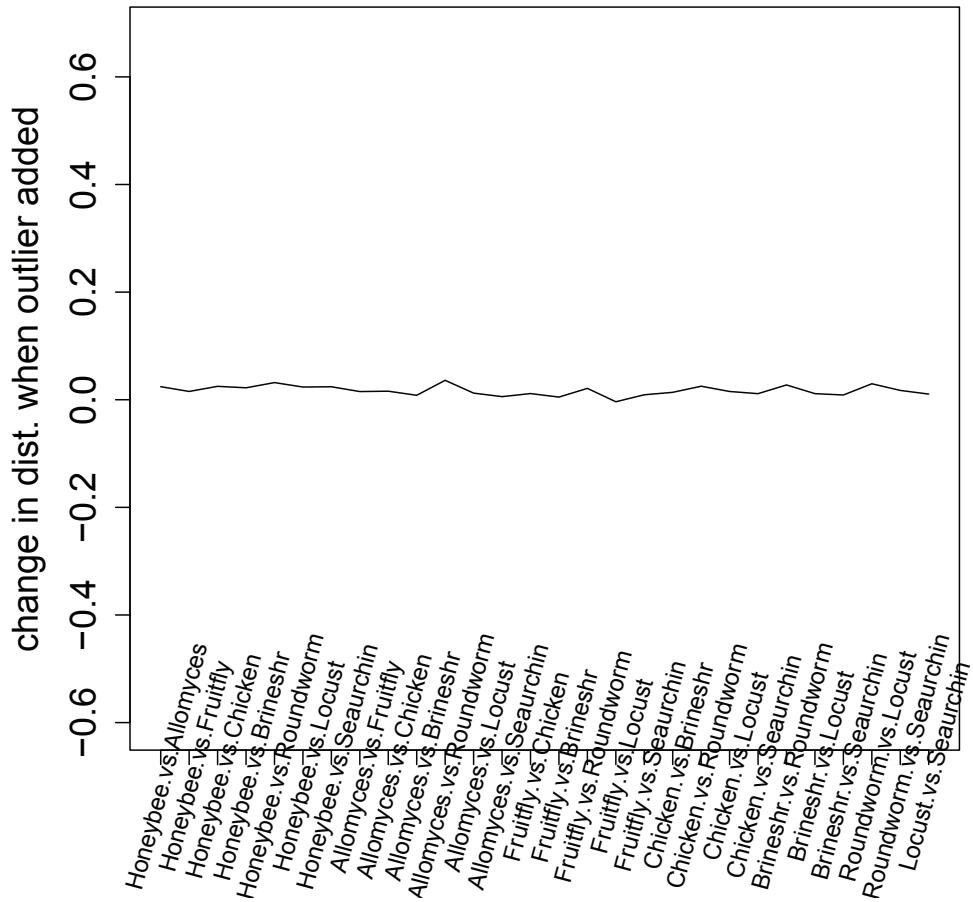

Figure 3.19: Change in the combined-gene pairwise distances obtained with the MinCV scale coefficients when an outlying gene is introduced to the analysis

The above analysis suggests that a subset of three or four genes in the nematode data set, ATP6, ND4L, ND5 and ND6, appear to be influential in determining the MinCV combined-gene tree derived from the common scaling based dissimilarities. Adding an outlying gene did not affect the estimated combined-gene tree.

3.10 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 *Drosophila melanogaster* in the animal clade. For the nematode data, the trees obtained from the common scaling and taxon-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 taxon-specific scaling tree place*Amborella trichopoda* and *Nymphaea alba* as sister taxa, while the common scaling tree place *Calycanthus floridus* and *Amborella 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 3.1 with strong bootstrap support (Keeling et al., 2009). The MinVar method was able to recover parts of this topology, but erroneously placed taxa *Drosophila melanogaster* as the most basal animal, and was not able to recover the correct position of *Dictyostelium discoideum*. Results were the same with both the common scaling and taxon-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 taxon-specific scaling method, support for the ecdysozoa hypothesis was weak, while the coelomata hypothesis had moderate to strong bootstrap support. Roundworm and honeybee were erroneously grouped together with weak bootstrap support.

For the twenty-five gene chloroplast data, both the common scaling and taxon-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 (Ané et al., 2004; Goremykin et al., 2003, 2005; Qiu et al., 1999; Soltis et al., 1999; Soltis and Soltis, 2004; Stefanovic et al., 2004; Zanis et al., 2002). Though neither of the common scaling nor taxon-specific scaling methods recovers the exact reference topology in Figure 3.8 (Soltis et al., 2005; Ané et al., 2004), the relative positions of the taxa within the angiosperm clade more or less agrees with the reference tree, with the exception of *Acorus americanus* which is misplaced with the eudicots in the common scaling and taxon-specific scaling trees. Analysis on a subset of nineteen 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 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 withing the angiosperm clade in the combined-gene tree.

The trees computed from SEQGEN 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 SEQGEN 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, while those based on the taxon-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.

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.

Table 3.1: 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

Table 3.2: Eukaryote Data: Quartet similarity between bootstrap trees and original data trees.

	Number of bootstrap permuation trees with percentage of identically resolved quartets							
	ComScal + MinVar		ComScal + MinCV		TaxaSpec + MinVar		Taxa Spec + 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

Table 3.3: 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
MinCV TaxaSpec BIONJ	52	48	62	38	2	61
MinCV TaxaSpec FITCH	30	67	69	31	0	43

Table 3.4: 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

Table 3.5: Bootstrap support of the topological features for the chloroplast tree under different methods.

	Topological feature				
Trees	*Recov. of green algae, non seed 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
MinCV TaxaSpec BIONJ	100	100	56	44	44
MinCV TaxaSpec FITCH	100	100	51	49	49

Table 3.6: 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

Table 3.7: Proportion of simulated trees with varying levels of sequence identity and block permutation trees which recover the primate reference tree.

Simulation scheme (% Seq 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 perm (>90%)	0.982	0.916	0.975	0.908	0.972	0.886	0.969	0.88

Table 3.8: Proportion of simulated trees with varying levels of sequence identity and block permutation trees which recover the nematode ecdysozoa and coelomata trees.

Simulation scheme (%Seq 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 perm ($10%-83%$)	*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

Table 3.9: Estimated single gene topologies and combined-gene topology in the nematode data set as well as Robinson-Foulds (RF) distances between the single gene trees and the theoretical tree under the ecydsozoa hypothesis.

Tree topology	Gene	Ecdys.
(Allomyces marcogynus,((Chicken,Sea urchin),(Roundworm,(Brine shrimp,(Honeybee, (Fruitfly,Locust))))))	Combined	0
(Allomyces marcogynus,(Chicken,Sea urchin),(Roundworm,(Honeybee,(Brine shrimp, (Fruitfly,Locust)))))	ATP6	2
(Allomyces marcogynus,((Sea urchin,Chicken),((Roundworm,Honeybee),(Brine shrimp, (Locust,Fruitfly)))))	COX1	4
(Allomyces marcogynus,(Roundworm,(Chicken,(Brine shrimp,((Fruitfly,Honeybee), (Sea urchin,Locust))))))	COX2	10
(Allomyces marcogynus,(Roundworm,((Brine shrimp,Honeybee),((Sea urchin,Chicken), (Locust,Fruitfly)))))	COX3	6
(Allomyces marcogynus,((Roundworm,Honeybee),((Sea urchin,Chicken),(Brine shrimp, (Locust,Fruitfly)))))	CYTB	6
(Allomyces marcogynus,(Chicken,(Sea urchin,(Locust,(Fruitfly,(Brine shrimp, (Roundworm,Honeybee)))))))	ND1	8
(Allomyces marcogynus,(Brine shrimp,((Sea urchin,Chicken),((Locust,Fruitfly), (Roundworm,Honeybee)))))	ND2	6
(Allomyces marcogynus,(Roundworm,(Honeybee,(Fruitfly,(Brine shrimp,(Locust,(Sea urchin,Chicken)))))))	ND3	8
(Allomyces marcogynus,(Sea urchin,(Honeybee,(Brine shrimp,((Locust,Chicken), (Roundworm,Fruitfly))))))	ND4	10
(Allomyces marcogynus,(Honeybee,((Roundworm,Brine shrimp),((Locust,Fruitfly),(Sea urchin,Chicken)))))	ND4L	6
(Allomyces marcogynus,((Sea urchin,Chicken),(Brine shrimp,((Locust,Fruitfly), (Roundworm,Honeybee)))))	ND5	4
(Allomyces marcogynus,((Roundworm,(Sea urchin,Chicken)),((Locust,Fruitfly),(Brine shrimp,Honeybee))))	ND6	4

Table 3.10: Combined-gene topologies obtained with the MinCV method when a single gene is removed from the analysis

Tree topology	Single gene removed from analysis
(Allomyces macrogynus,((Chicken,Sea urchin),(Roundworm,(Brineshr,(Honeybee, (Fruitfly,Locust))))))	COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4
(Allomyces macrogynus,(Chicken,Sea urchin),(Roundworm,(Honeybee,(Brineshr, (Fruitfly,Locust)))))	ND5
(Allomyces macrogynus,(Chicken,Sea urchin),((Roundworm,Honeybee),(Brineshr, (Fruitfly,Locust))))	ATP6, ND4L, ND6

Chapter 4

COMBINING DISSIMILARITY MEASURES USING SINGULAR VALUE DECOMPOSITION

In this chapter we introduce another method of deriving multiple gene tree estimates based on singular value decomposition. Stuart et al. (2002) applied singular value decomposition to tetrapeptide frequency matrices to select the most informative biomolecular sequence characteristics from which to estimate evolutionary distances. Here we are interested in extracting the consistent signal present in the dissimilarity or distance matrices for each gene and using this information to obtain a combined-gene tree. As with the MinVar and MinCV methods presented in chapter 3, this method can be applied using any taxonomic distance measure. The method is illustrated using both the common scaling spectral covariance based dissimilarities described in chapter 3 and Jones-Taylor-Thornton (JTT) based distances. JTT-based distances are computed using the protdist programme in PHYLIP with one category of substitution rates (Felsenstein, 1989). Combined-gene trees for the primate, nematode and chloroplast data sets are estimated and the variability about the estimated trees is determined using bootstrap samples. In the case of the common scaling spectral covariance based distances the method is applied to block permutation samples. The block permutation method is used to ensure the dependence structure between the sites of protein sequences is partially preserved in the bootstrap samples. Since the JTT model of evolution assumes independence of sites, variance about the trees based on these distances is estimated from bootstrap samples obtained by sampling the individual sites within sequences with replacement. Bootstrap samples used with the JTT model are obtained using the seqboot programme in PHYLIP (Felsenstein, 1989).

4.1 SINGULAR VALUE DECOMPOSITION OF DISTANCES

As was done for the MinVar and MinCV methods, dissimilarities or distances for all pairs of taxa are first written as p-vectors where $p = \binom{n}{2}$ for n taxa. The p-vectors for a set of k genes are combined into a single matrix X, where X is the transpose of the matrix D defined in 3.1 used in the MinVar and MinCV methods of chapter 3. The rows of X correspond to genes and the columns to taxa pairs. That is, $x_{i,j}$ is the pairwise dissimilarity or distance of the i^{th} gene and the j^{th} taxa pair, where $i = 1, ..., k$ and $j = 1, ..., p$.

Let $m = \min(k, p)$. The matrix X can be decomposed as follows

$$X = U\Lambda V' \tag{4.1}$$

where U is an orthogonal $k \times m$ matrix containing the eigenvectors of the matrix XX', V is an orthogonal $m \times p$ matrix containing the eigenvectors of $X'X$ and Λ is the diagonal $m \times m$ matrix of singular values. The vectors in V give the direction of the principal components $U\Lambda$ (Hastie et al., 2001).

The distances between pairs of taxa in a phylogenetic tree can be expressed as a linear combination of the estimated branch lengths between those taxa for a given tree topology (Rhzetsky and Nei, 1992). For a set of n taxa, any particular set of $p = \binom{n}{2}$ pairwise distances representing a tree topology can be expressed as a set of p equations involving $2n - 3$ variables or branches, which in turn can be expressed as a topology matrix T. That is, for a vector of distances for gene j, say $x'_{j.}$, we get the following matrix equation

$$x'_{j.} = Tb,$$

where b is a $(2n - 3) \times 1$ vector of branch lengths and T is a $p \times (2n - 3)$ topology matrix.

Assuming topology T is the true topology, the rows of matrix X should belong to a subspace spanned by the columns of T. We can think of this as regressing the distances for each gene on the variables corresponding to the branches in topology T. If there is a consistent signal among genes and noise in the data is small then when T expresses the correct topology, the estimated branch lengths b from different rows

of X should be proportional to each other and this proportional difference forms the major variability between rows of X. In this case, the first column of V corresponds to the direction in which the rows of X are proportional, and the first column of $U\Lambda$ will correspond to the proportional factors. Therefore, a single tree representation for multiple genes can be obtained using the first right eigenvector of the singular value decomposition of the pairwise distances in matrix X.

For the singular value decomposition of X defined in 4.1, let u_1 denote the first left eigenvector, v_1 be the first right eigenvector and λ_1 be the first singular value. A new $k \times p$ distance matrix, X_1, may be obtained from

$$X_1 = \lambda_1 u_1 v_1'.$$

The topology of the tree derived from X_1 is given by v_1 and the branch lengths proportional factors by $\lambda_1 u_1$. Note that the pairwise distances for each gene in X_1 are all scaled versions of v_1 and thus describe the same tree topology. Hence, this equates to using de-noised data to estimate a single representative topology for multiple genes for a given set of taxa.

4.2 RESULTS

4.2.1 RESULTS FOR THE PRIMATE DATA SET

The method is first applied to the simple five taxa primate data set used in chapter 3. combined-gene trees are estimated using BIONJ and FITCH (Gascuel, 1997; Fitch and Margoliash, 1967). Figure 4.1 shows the estimated combined-gene trees recovered from the first right eigenvector of the singular value decomposition of the 13×10 matrices of JTT distances and common scaling covariance dissimilarities.

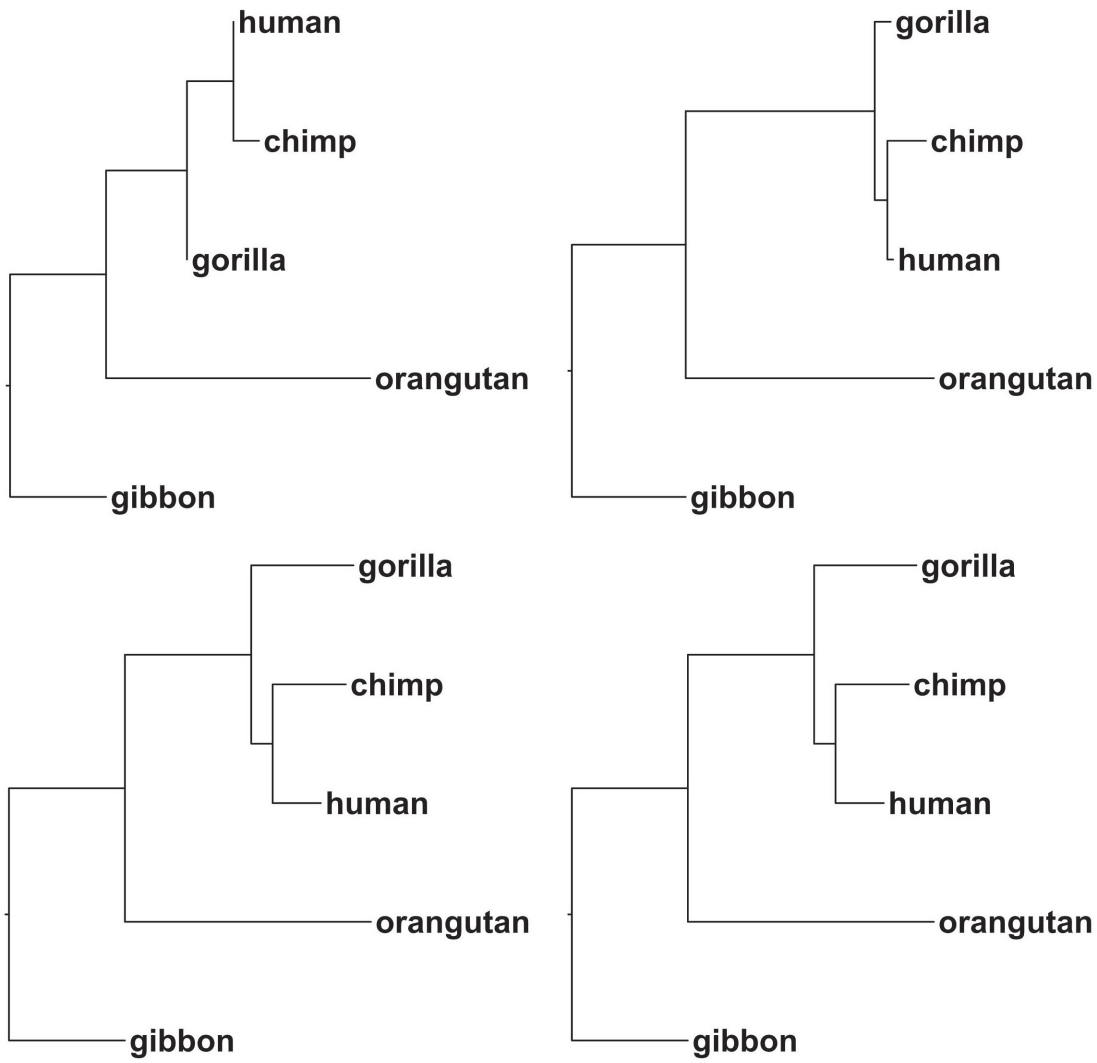

Figure 4.1: Combined-gene tree for the primate data set estimated from first right eigenvector of the singular value decomposition of common scaling covariance dissimilarities (top) and JTT distances (bottom) for BIONJ (left) and FITCH (right).

This topology corresponds to the reference tree topology found on the tree of life website (Tree of Life Web Project, 1999).

To evaluate the variance about the trees estimated from the covariance-based dissimilarities the method was applied to 1000 block size 14 permutation samples and a majority-rule consensus tree computed using the CONSENSE programme in PHYLIP (Felsenstein, 1989). The majority-rule consensus trees derived from the common scaling covariance based dissimilarities and obtained with the BIONJ and

FITCH tree building methods are shown in the top panels of Figure 4.2. For the trees estimated from the JTT-based distances the variance was evaluated from 1000 bootstrap samples obtained by sampling individual characters from sequences with replacement. The majority-rule consensus trees derived from the JTT-based distances obtained with the BIONJ and FITCH tree building methods are shown in the bottom two panels of Figure 4.2.

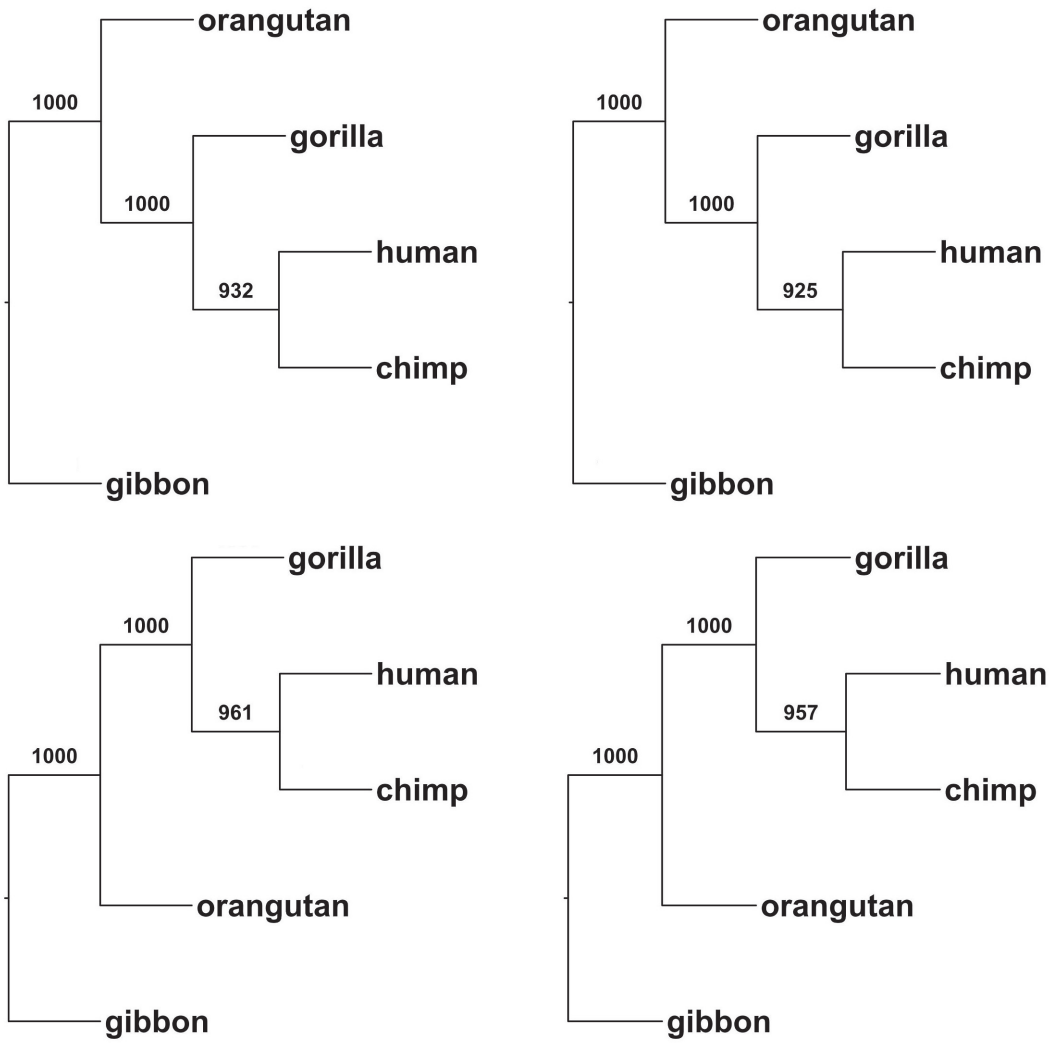

Figure 4.2: Primate majority-rule consensus tree for the singular value decomposition trees estimated 1000 block permutation samples with common scaling covariance dissimilarities and BIONJ (top left), common scaling dissimilarities and FITCH (top right) JTT-based distances and BIONJ (bottom left) and JTT-based distances and FITCH.

The combined-gene trees estimated from the first right eigenvector of the singular value decomposition for both covariance-based dissimilarities and JTT-based distances strongly support the reference tree. There is some uncertainty in the estimated position of gorilla, human and chimp. The first four rows of Table 4.1 summarize the bootstrap support for the topological features present in the trees estimated from the

first right eigenvector of the singular value decomposition of covariance-based dissimilarities and JTT-based distances. The BIONJ and FITCH tree building methods give more or less the same result. For the common scaling spectral covariance based dissimilarities the reference tree is recovered by 932 and 925 of the trees obtained from BIONJ and FITCH, respectively. Gorilla and human branch as sister taxa with chimp as an outgroup for this clade in 41 of the BIONJ trees and 43 of the FITCH trees. The erroneous branching of human as an outgroup to a chimp and gorilla clade occurs in 27 of the block permutation BIONJ trees and 32 of the block permutation FITCH trees. For the bootstrap trees estimated from the JTT-based distances, the reference tree is recovered by 961 and 957 of the JTT-based bootstrap trees for BIONJ and FITCH, respectively. Gorilla and human branch as sister taxa with chimp as an outgroup for this clade in 39 of the BIONJ trees and 43 of the FITCH trees. The singular value decomposition method of combining genes appears to work well for this data, with the common scaling spectral covariance dissimilarities and JTT distances returning similar results.

For comparison, topological features present in the bootstrap trees obtained with the MinCV method of combining genes are shown in the bottom four rows of Table 4.1. Both the singular value decomposition and MinCV methods of combining genes return the same estimate for the tree topology. The two methods appear to be similar in terms of variability about the estimated tree with the MinCV estimate having somewhat smaller variance than the singular value decomposition estimate. With BIONJ 5 of 1000 estimated covariance-based MinCV trees branch human and gorilla as sister taxa, while in 13 of 1000 covariance-based MinCV trees erroneously branch gorilla and chimp as sister taxa with human as an outgroup to this clade. In the corresponding FITCH trees 8 of 1000 estimated trees branch human and gorilla as sister taxa while 17 of 1000 branch gorilla and chimp as sister taxa. The reference tree is recovered by 982 covariance-based BIONJ trees and 975 covariance-based FITCH trees. All the MinCV trees estimated from the JTT-based distances recover the reference tree.

Figure 4.3 shows the cumulative proportion of singular values among the sum of all singular values in the singular value decomposition of the common scaling

based dissimilarities and JTT-based distances. One can see that the first singular value makes up 70% of the sum of singular values for the common scaling based dissimilarities and 77% of the variance of the singular values for the JTT-based distances, respectively. The sum of the first two singular values account for 82% and 85% of the sum of the singular values for the covariance-based and JTT-based distances, respectively.

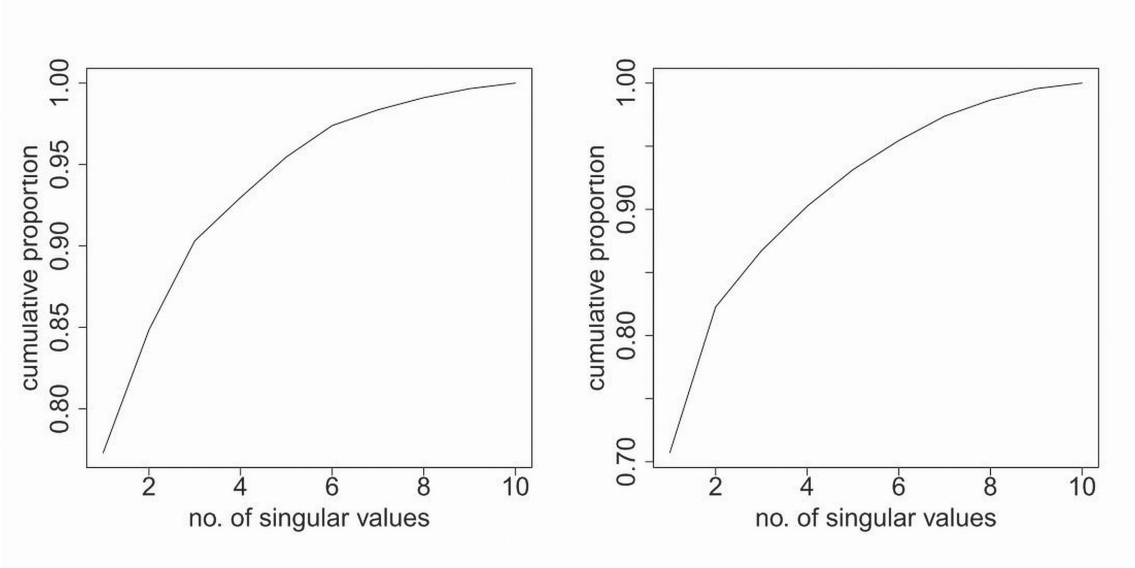

Figure 4.3: Cumulative proportion of the sum of the singular values of the singular value decomposition of JTT-based distances (left) and the common scaling covariance based dissimilarities (right) for primate data set.

The second right eigenvector is made up of both positive and negative values suggesting it is picking out a contrast between two sets of genes. To examine the difference in the trees obtained with the two sets of genes identified by the second right eigenvector a 13×10 matrix of distances is constructed from the second layer of the singular value decomposition, $X_2 = \lambda_2 u_2 v_2'$ and the signs of the distances in each gene examined. In this way we can identify which genes belong to which group. The first group consists of $ND1, COX2, ATP8, ATP6$, and $ND4$. The second group consists of $ND2, COX1, COX3, ND3, ND4L, ND5, ND6$, and $CYTB$. Figure 4.4 shows the singular value decomposition trees obtained with the two different groups of genes from common scaling covariance based dissimilarities and JTT-based distances. With the common scaling based dissimilarities the tree recovered with the first group of

genes places gorilla and chimp as the deepest clade in the tree with human erroneously placed before these two taxa, while the tree recovered with the second group of genes agrees with the reference tree topology. Hence, for the common scaling based covariances the contrast is in the relative placement of the taxa obtained with the two trees.

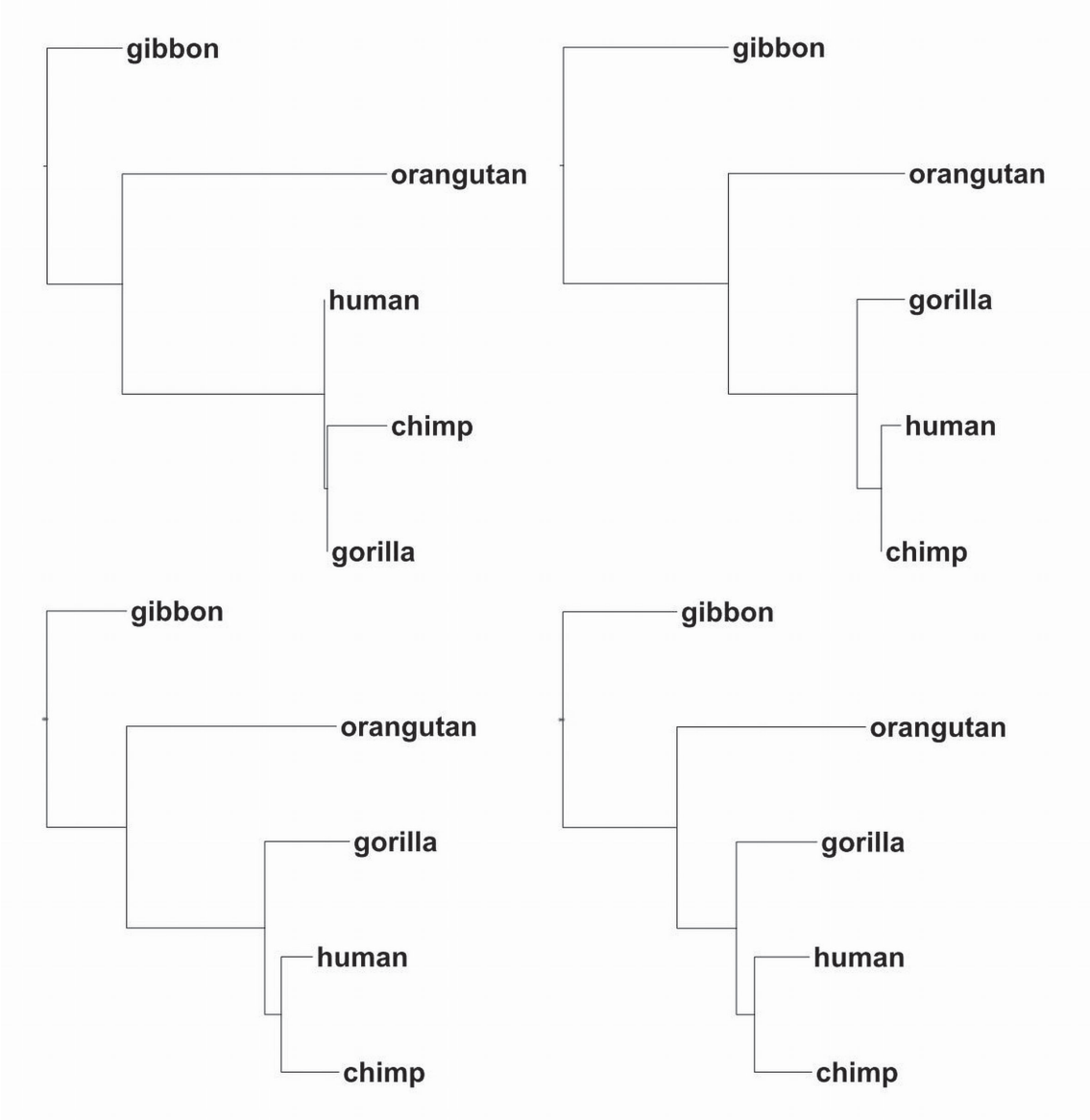

Figure 4.4: Combined-gene trees for primate data set obtained with the two groups of genes, group 1 (left) and group 2(right) from common scaling covariance based distances (top) and JTT-based dissimilarities (bottom).

With the JTT-based distances the two groups recover the same tree topology as the reference tree. An examination of the branch lengths for the two groups indicates that in group two the distances between the three taxa gorilla, human and chimp are greater than they are in group one, while the distance between orangutan and the gorilla/human/chimp clade is smaller in group two than in group one. Table 4.2 shows the distances between the gorilla and chimp, the gorilla and human, and the oranguatan and the gorilla/human/chimp clade for the tree for group one, group two and for the tree obtained with the first right eigenvector of the distances. For the tree derived from the first right eigenvector the distances within the gorilla/human/chimp clade are very small while the distance between oranguatan and the clade of gorilla/human/chimp is relatively large. Hence, we can see that in the case of the JTT-based distances the contrast is in the branch length estimation.

4.2.2 RESULTS FOR THE NEMATODE DATA SET

The method is next applied to the more difficult nematode data set described in chapter 3 which has 8 taxa and 12 genes. Figure 4.5 shows the combined-gene tree obtained using the first right eigenvector of the singular value decomposition of the 12×28 matrix of common scaling spectral covariance based dissimilarities and JTT-based distances obtained with the BIONJ and FITCH tree building methods.

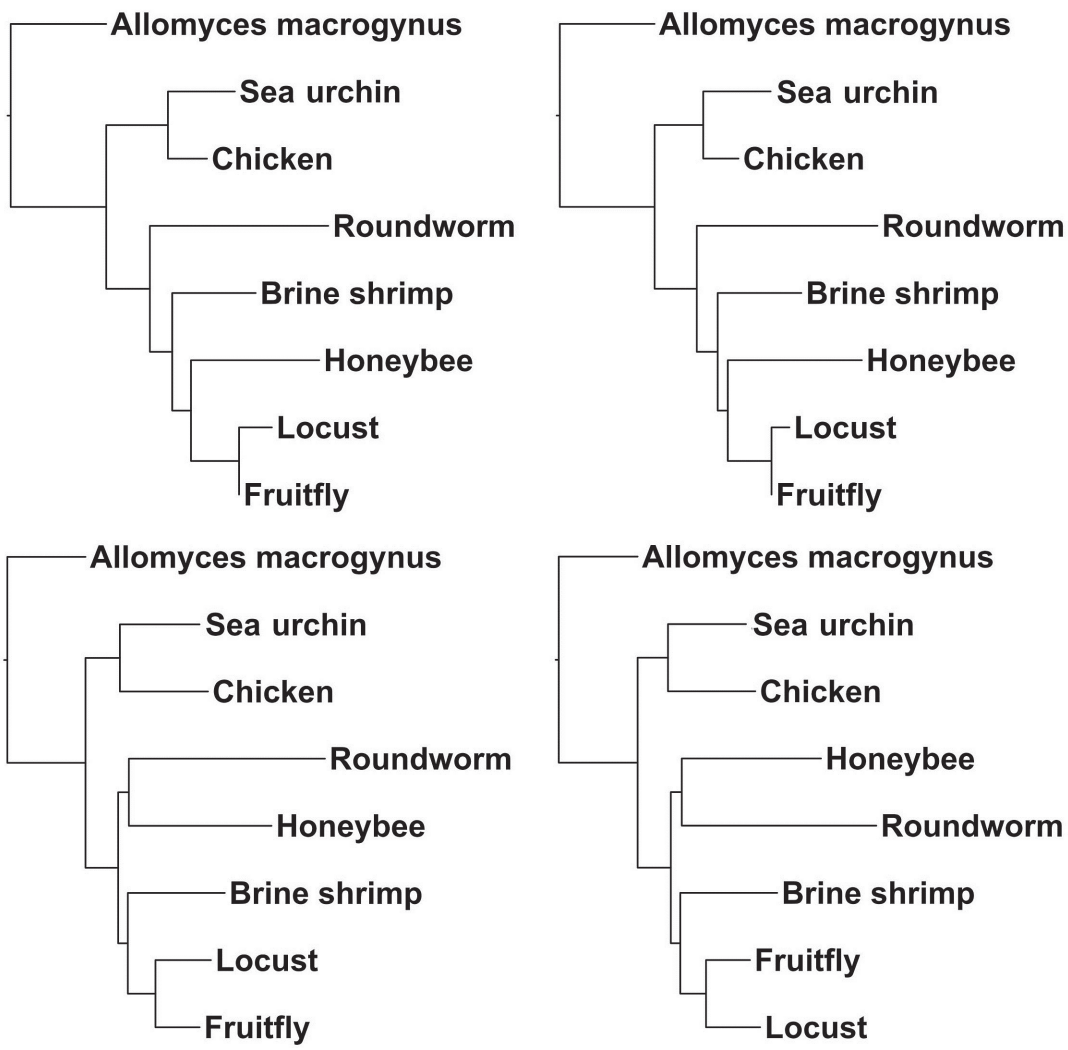

Figure 4.5: Combined-gene tree for the nematode data set estimated from first right eigenvector of the singular value decomposition of common scaling covariance dissimilarities (top) and JTT distances (bottom) for BIONJ (left) and FITCH (right).

The combined-gene tree estimated from the common scaling based distances agrees with the reference tree under the ecdysozoa hypothesis shown in Figure 3.5. The combined-gene tree estimated from JTT-based distances corresponds to the topology found using concatenated sequences which was presented in Foster and Hickey (1999), with the roundworm (nematode) and honeybee being erroneously grouped together as sister taxa.

The BIONJ and FITCH majority-rule consensus trees for 1000 bootstrap trees derived from the common scaling spectral covariance based dissimilarities and the JTT-based distances are shown in Figure 4.6. The estimated combined-gene topology has strong bootstrap support, though there is some variability in the estimated topology of the arthropod clade.

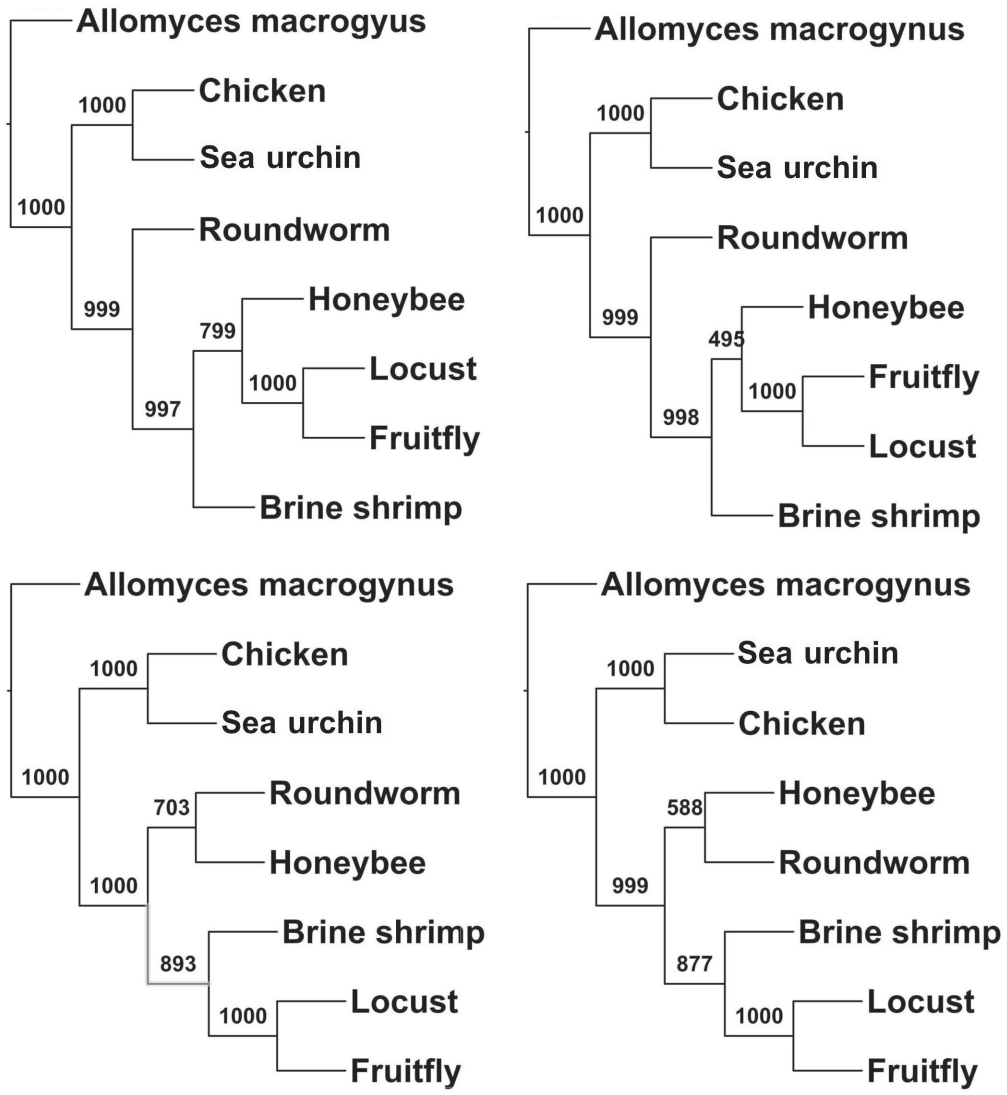

Figure 4.6: Nematode majority-rule consensus tree for the singular value decomposition trees estimated 1000 block permutation samples with common scaling covariance dissimilarities and BIONJ (top left), common scaling dissimilarities and FITCH (top right) JTT-based distances and BIONJ (bottom left) and JTT-based distances and FITCH.

The top four rows of Table 4.3 summarize the topological features recovered in the estimated trees obtained with singular value decomposition method of combining genes as well as the bootstrap support for each feature. Variability in the placement of honeybee is greater in trees estimated with FITCH than those estimated with BIONJ.

For the trees derived from the common scaling covariance based dissimilarities, 151 of the 1000 trees obtained with BIONJ place honeybee before brine shrimp, while 48 branch honeybee and brine shrimp as sister taxa. Only 3 recover the erroneous placement of honeybee and roundworm as sister taxa. For the corresponding trees obtained with FITCH, honeybee branches before brine shrimp in 337 of the trees, while in 168 trees honeybee and brine shrimp are placed as sister taxa. Two of the FITCH trees place honeybee and roundworm as sister taxa. The estimated trees derived from the JTT-based distances have moderate to strong bootstrap support. The erroneous placement of honeybee and roundworm as sister taxa is recovered in 703 BIONJ trees and 588 FITCH trees. The honeybee and roundworm are separated in 297 BIONJ trees and 412 FITCH trees. Only 34 and 50 of the JTT-based trees obtained with BIONJ and FITCH, respectively, recover the reference tree under the ecdysozoa hypothesis although the relative placement of the nematode with respect to arthropods and vertebrates agrees with the ecdysozoa hypothesis in all trees. None recover the reference tree under the coelomata hypothesis. For this particular data set the singular value decomposition method of combining genes appears to work well with the common scaling spectral covariance based distances, but is unable to recover the reference tree under either hypothesis from the JTT-based distances.

Again a corresponding summary of topological features obtained with the MinCV method of combining genes is shown in the bottom four rows of Table 4.3 for comparison. For this data set there appears to be more variability about the estimated trees using the MinCV method of combining genes than for the singular value decomposition method for combining genes. A much larger number of the common scaling covariance based trees erroneously place honeybee and roundworm as sister taxa under this method. With BIONJ 308 trees return a roundworm and honeybee clade compared to only 3 trees under the singluar value decomposition method of combining genes. Similarly with FITCH, 277 trees place honeybee and roundworm as sister taxa compared with 2 trees with the singular value decomposition method of combining genes. Also, far fewer trees recover the reference tree under the ecdysozoa hypothesis under the MinCV method. While the majority of MinCV trees agree with

the ecdysozoa hypothesis in terms of the relative placement of the nematode, arthropods and vertebrates, many have honeybee branching before brine shrimp. This erroneous branching occurs in fewer of the singular value decomposition based trees.

Figure 4.7 shows the cumulative proportion of the singular values among the sum of all singular values. For the nematode data set, the first singular value makes up 57% and 75% of the singular values in the singular value decomposition of common scaling covariance based and JTT-based distances, respectively, while the first and second singular values make up 66% and 83% for covariance-based dissimilarities and JTT-based distances, respectively.

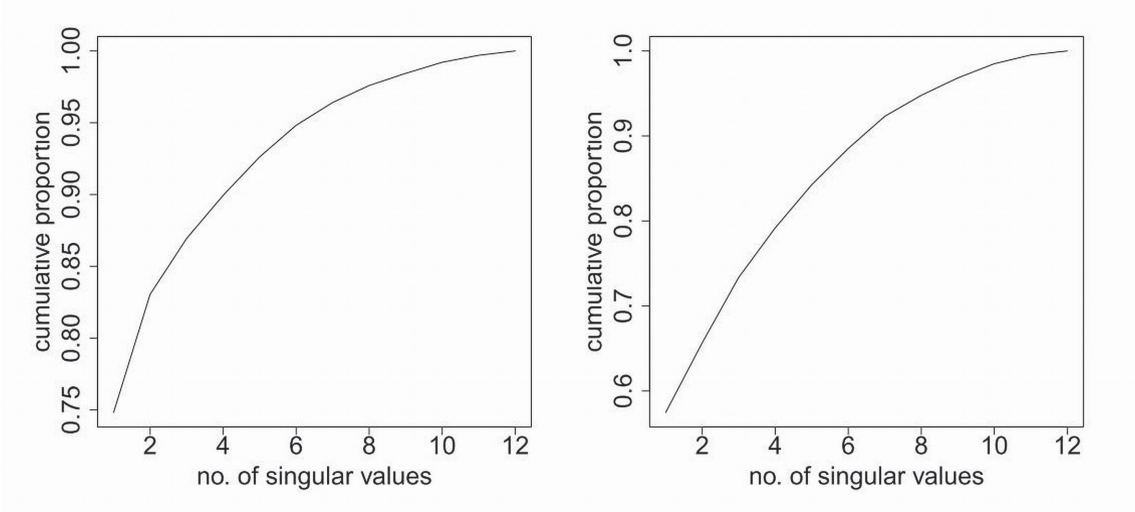

Figure 4.7: Cumulative proportion of the sum of the singular values of the singular value decomposition of JTT-based distances (left) and the common scaling covariance based dissimilarities (right) for nematode data set.

Again, we look at the second layer of the singular value decomposition to identify groups of genes. For the common scaling based distances the first group identified with the second right eigenvector consists of $ND3, ND5$, and $ND4$, with remaining genes in the second group and for the JTT-based distances the first group consists of $ATP6, ND3$ and $ND4L$ with the remaining genes in the second group. Figure 4.8 shows the singular value decomposition combined-gene trees recovered with the different groups of genes. For the JTT-based distances, the tree recovered with the first group of genes is able to separate the roundworm and the honeybee, while the the second group of genes erroneously places the roundworm and honeybee as sister

taxa. For the covariance-based dissimilarities, the tree recovered with the first group agrees with the reference tree topology under the coelomata hypothesis, while the second group recovers the reference tree topology under the ecydsozoa hypothesis. For this data set the contrast in the two groups of genes is clearer to see. In the case of the JTT-based distances it distinguishes a group of genes which places the roundworm and honeybee together and a group of genes which separates these two taxa. For the covariance-based distances it distinguishes between a group of genes which recovers the ecdysozoa hypothesis versus a group of genes which recovers the coelomata hypothesis.

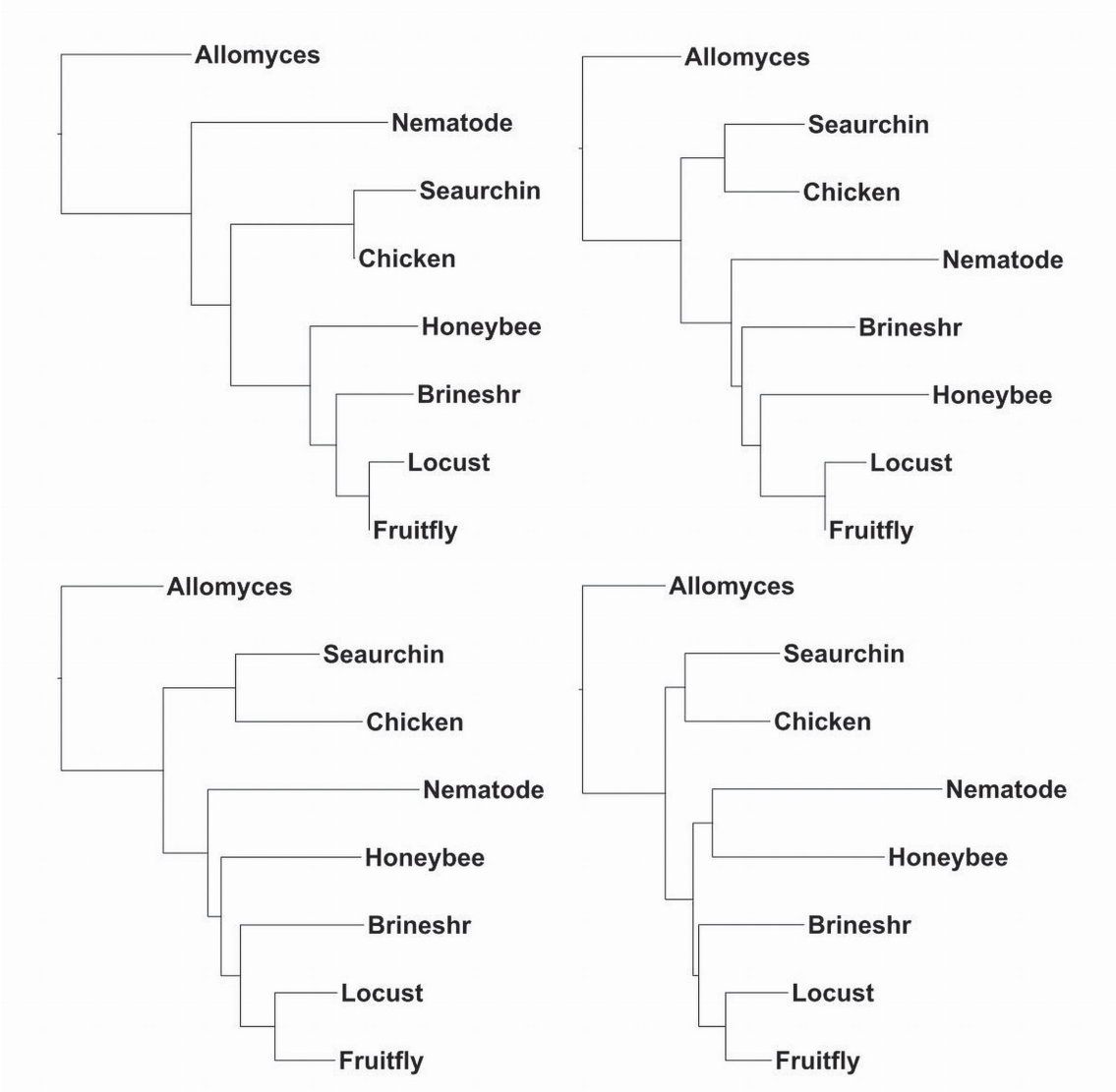

Figure 4.8: Combined-gene trees for nematode data set obtained with the two groups of genes, group 1 (left) and group 2(right) from common scaling covariance based distances (top) and JTT-based dissimilarities (bottom).

Because for the nematode data set the cumulative proportion of the sum of the first and second singular values are only 66% and 83% for covariance-based dissimilarities and JTT-based distances, respectively, we also looked at the third right eigenvector. For the common scaling based distances, the first group of genes identified with the third right eigenvector consisted of $ATP6, ND5$ and $ND6$ with the remaining genes in a second group. The tree obtained with the first group of genes corresponds to the reference tree under the ecdysozoa hypothesis. The tree obtained

with the second group of genes erroneously places honeybee and nematode as sister taxa. For the JTT method the third right eigenvector the first group of genes consisted of $ATP6, COX1, COX2$, and $CYTB$, with the remaining genes in a second group. The tree obtained with group one genes was very similar to the reference tree under the ecdysozoa hypothesis, with honeybee erroneously being placed as most basal in the arthropod clade. The second group of genes grouped nematode and honeybee as sister taxa.

4.2.3 RESULTS FOR THE CHLOROPLAST DATA SET

As a final example, we apply the method to the larger chloroplast data set which has 22 taxa and 25 genes. Figure 4.9 shows the combined-gene tree obtained using the first right eigenvector of the singular value decomposition of the 25×231 matrix of common scaling spectral covariance based dissimilarities and JTT-based distances obtained with the BIONJ and FITCH tree building methods. Separation of taxa into the major clades of green algae, non-seed plants and angiosperms shown in the reference tree in Figure 3.8 is recovered. The singular value decomposition method and the MinCV method recover very similar combined-gene trees. In both the singular value decomposition common scaling covariance based trees and the singular value decomposition JTT-based trees the ferns *Psilotum nudum* and *Adiantum capillus-veneris* are erroneously separated. All four trees in Figure 4.9 place a clade with *Amborella trichopoda, Nymphaea alba* and *Calycanthus floridus* as basal in the angiosperm clade.

Figure 4.9: Combined-gene tree for the chloroplast data set estimated from first right eigenvector of the singular value decomposition of common scaling covariance dissimilarities (top) and JTT distances (bottom) for BIONJ (left) and FITCH (right). Boxes indicate portions of the tree which differ from reference tree.

The variance of the combined-gene tree was estimated from 100 block bootstrap samples (block size fourteen for the common scaling covariance method and block size one for JTT method). Table 4.4 summarizes the topological features recovered in the

estimated trees obtained with the singular value decomposition method of combining genes. The separation of green algae, non-seed plants and angiosperms has 100 % bootstrap support with all four methods. A greater proportion of the bootstrap trees place *Amborella trichopoda* and *Nymphaea alba* as sister taxa with the singular value decomposition method than with the MinCV method. Of the 100 bootstrap trees, 72 BIONJ and 89 FITCH common scaling singular value decomposition trees place *Amborella trichopoda* and *Nymphaea alba* as sister taxa compared with 29 BIONJ and 66 FITCH common scaling MinCV trees. Similarly, for the JTT-based trees, 63 BIONJ and 72 FITCH trees place *Amborella trichopoda* and *Nymphaea alba* as sister taxa compared with 21 and 30 of the corresponding MinCV trees. There is a fair amount of variance in the branching of the two ferns *Psilotum nudum* and *Adiantum capillus-veneris* in the singular value decomposition based trees. Of the 100 bootstrap trees obtained with the singular value decomposition method only 33 common scaling BIONJ trees, 0 common scaling FITCH trees, 35 JTT BIONJ trees and 53 JTT FITCH trees recover the clade of *Psilotum nudum* and *Adiantum capillus-veneris* seen in the reference tree in Figure 3.8. The basal placement of *Amborella trichopoda, Nymphaea alba* and *Calycanthus floridus* in the angiosperm clade has 100 % bootstrap support for all four methods. The overall variance of the combined-gene tree obtained with the singular value decomposition method appears to be smaller than that of the combined-gene tree obtained with the MinCV method.

Figure 4.10 shows the cumulative proportion of the singular values among the sum of all singular values. For the chloroplast data set, the first singular value makes up 46% and 79% of the singular values in the singular value decomposition of common scaling covariance based and JTT-based distances, respectively, while the first and second singular values make up 59% and 86% for covariance-based dissimilarities and JTT-based distances, respectively.

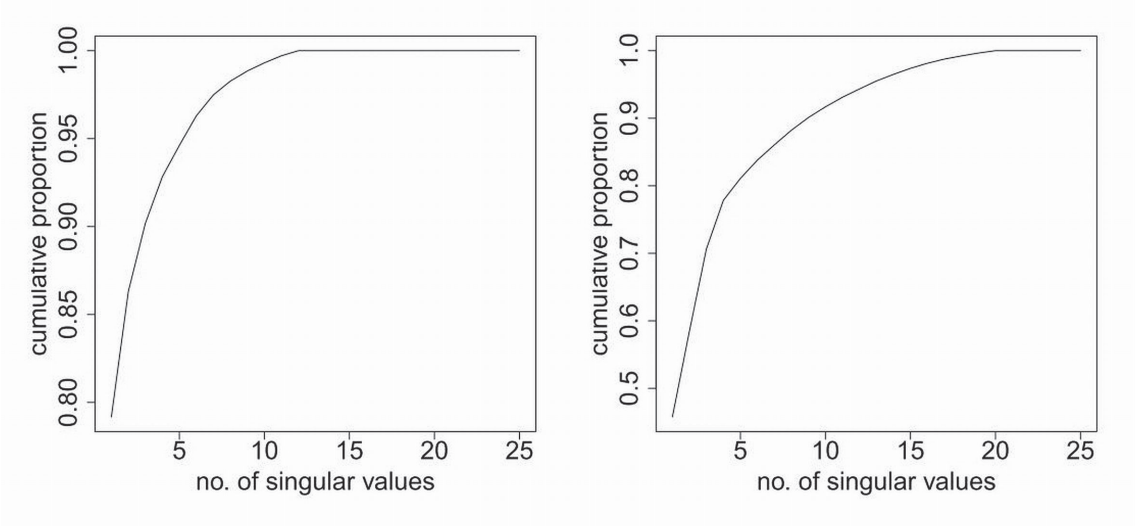

Figure 4.10: Cumulative proportion of the sum of the singular values of the singular value decomposition of JTT-based distances (left) and the common scaling covariance based dissimilarities (right) for chloroplast data set.

An examination of the second layer of the singular value decomposition reveals two groups of genes. Group one consists of $atpI, clpP, psaC, rbcL$ and $rpoC1$ and group two consists of the remaining genes. Note that the five genes in the former group are among the six genes whose exclusion resulted in a monocot-first tree. Figure 4.11 show the combined-gene tree obtained with the genes in the two different groups. Boxes indicate where the tree differs from the reference tree. One can see that the trees obtained from the first group of genes more or less agree with the reference tree with some small differences. For the common scaling based trees obtained with the second group of genes the grouping of taxa within each clade is mostly recovered. However, the relative placement of the major clades is erroneous. In this tree, the *Amborella trichopoda, Nymphaea alba* and *Calycanthus floridus* are still together and appear in a more basal position than the monocots, however the two pines are erroneously separated and *Pinus koraiensis* is erroneously placed in the monocot clade. The JTT-based tree obtained from the second group of genes is a monocot-first tree. For the JTT-based trees, the second layer of the singular value decomposition appears to distinguish between a group of genes which recovers a monocot-first topology and a group of genes which recovers a basal clade of *Amborella trichopoda, Nymphaea alba* and *Calycanthus floridus* within the angiosperm clade. That the five genes,

$atpI, clpP, psaC, rbcL$ and $rpoC1$, in the first group identified with the second right eigenvector are among the six genes whose exclusion resulted in a monocot-first tree further confirms that the contrast corresponds to the conflicting topologies for the angiosperm clade though this is harder to see when looking at covariance-based trees obtained from the two groups of genes.

Figure 4.11: Combined-gene trees for chloroplast data set obtained with the two groups of genes, group 1 (left) and group 2(right) from common scaling covariance based distances (top) and JTT-based dissimilarities (bottom).

Because the sum of the first and second singular values is only 59% for the covariance-based dissimilarities we also examine the third right eigenvector. The first group consists of $atpB, atpE, petB, rpl14, rpl20, rps11, rps18, rps19, rps3, rps7$ and $rps8$ with the remaining thirteen genes in a second group. The first group returns a tree similar to that shown in the top right-hand panel of Figure 4.11. The relative placement of taxa within the major clades is recovered for the most part though the relative placement of the major clades to each other is erroneous. The tree estimated from the second group of genes returns a tree similar to that in the top left-hand panel of Figure 4.9, with the exception of the placement of the lotus which is erroneously grouped with the monocots.

4.3 DISCUSSION

The difference in the variability about tree estimates for the nematode and chloroplast data sets obtained under the MinCV and singular value decomposition methods for combining genes may be due to the way the two methods handle 'noise' in the distances. Recall the nematode and chloroplast data sets suffer from long branch attraction and compositional bias, both of which may introduce non-phylogenetic signal into the estimated distances. While the spectral covariance based dissimilarities appear to be less sensitive to these phenomena than the JTT-based distances some variability will remain with regard to the placement of certain taxa, such as the honeybee and roundworm in the estimated nematode trees. The singular value decomposition method combines genes using a de-noised version of the distance matrix while the MinCV method attempts to adjust for noise by weighting the distances for each gene in such a way as to minimize the variance. Hence, the estimated combined-gene trees with the singular value decomposition matrix have smaller variance than the corresponding MinCV trees.

A large proportion of the sum of the singular values in the singular value decomposition is accounted for by the first two singular values. The second right eigenvector describes the second principal direction and may be used to identify contrasts in the topologies for subgroups of genes. Using the second layer of the singular value decomposition of the pairwise distances we were able to cluster the genes into two groups

with contrasting topological features. The remaining right eigenvectors are more difficult to interpret. Corresponding additional singular values appear to account for a relatively small proportion of the sum of all singular values and so we treat the corresponding additional layers of the singular value decomposition as noise.

Table 4.1: Bootstrap support of the topological features for the primate tree for singular value decomposition and MinCV methods

	Topological features		
Trees	*Recovers reference tree*	*Gorilla + human branch as sister taxa*	*Gorilla + chimp branch as sister taxa*
SVD ComScal BIONJ	932	41	27
SVD ComScal FITCH	925	43	32
SVD JTT BIONJ	961	39	0
SVD JTT FITCH	957	43	0
MINCV ComScal BIONJ	982	5	13
MINCV ComScal FITCH	975	8	17
MINCV JTT BIONJ	1000	0	0
MINCV JTT FITCH	1000	0	0

Table 4.2: Distances between pairs of taxa in the primate data when all the genes are used, when the two groups of genes identified by the second right eigenvector of the singular value decomposition are analysed separately.

	gorilla - human	gorilla - chimp	orangutan - other taxa
All genes	0.0139	0.0433	0.3583
Group 1 genes	0.1215	0.1397	0.3336
Group 2 genes	0.1606	0.1746	0.2571

Table 4.3: Bootstrap support of the topological features for the nematode tree for singular value decomposition and MinCV method

	Topological features						
Trees	*Agrees with Ecdysozoa*	*Agrees with Coelomata*	*Sep. of honeybee and roundworm*	*Honeybee and roundworm sister taxa*	*Brine shrimp basal to arthropods*	*Honeybee basal to arthropods*	*Brine shrimp and honeybee sister taxa*
SVD ComScal BIONJ	999	1	997	3	799	151	48
SVD ComScal FITCH	999	1	998	2	495	337	168
SVD JTT BIONJ	1000	0	297	703	34	253	15
SVD JTT FITCH	999	1	412	588	50	356	24
MinCV ComScal BIONJ	1000	0	692	308	201	491	9
MinCV ComScal FITCH	993	7	723	277	61	662	42
MinCV JTT BIONJ	1000	0	7	993	0	7	0
MinCV JTT FITCH	1000	0	71	929	3	68	1

Table 4.4: Bootstrap support of the topological features for the chloroplast tree for singular value decomposition and MinCV method

	Topological feature				
Trees	*Recov. of green algae, non seed 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*
SVD ComScal BIONJ	100	100	72	28	33
SVD ComScal FITCH	100	100	89	11	0
SVD JTT BIONJ	100	100	63	37	35
SVD JTT FITCH	100	100	72	28	53
MinCV ComScal BIONJ	100	100	29	22	71
MinCV ComScal FITCH	100	100	66	22	34
MinCV JTT BIONJ	100	100	21	14	65
MinCV JTT FITCH	100	100	30	0	78

Chapter 5

INFLUENCE ANALYSIS OF MULTIPLE GENE PHYLOGENETIC RECONSTRUCTION USING SINGULAR VALUE DECOMPOSITION

Influence analysis has been widely applied to identify outliers and influential observations in a data set. Here we explore the effect of individual genes on the estimated trees in the context of distance based methods. Quantifying the influence of individual genes on a combined tree topology may enable us to identify genes that have had the most influence on the inference of how a species has evolved over time and its relationship to other species. In this chapter we will use common scaling covariance based dissimilarities as the example, but note that these methods could be applied to any measure of taxonomic distance. Throughout this chapter when we refer to the combined-gene tree or combined-gene dissimilarities we are referring to the combined-gene tree obtained with the singular value decomposition method presented in chapter 4.

To identify subsets of genes which are most influential in determining estimated tree topologies we examine their corresponding dissimilarity matrices, or equivalently their dissimilarity vectors of length $p = \binom{n}{2}$, where n is the number of taxa. A multiple gene tree is estimated using the singular value decomposition method described in chapter 4. Expanding upon the results in Fung et al. (2007), an expression for the generalized influence function for singular value decomposition is derived. The influence of individual genes on the right eigenvectors of the singular value decomposition is evaluated using a case-weights perturbation scheme. The effect of the perturbations on how the taxa are placed in the estimated tree is then studied. We apply our method to the primate and nematode data sets to illustrate the method.

5.1 INFLUENCE FUNCTIONS FOR SINGULAR VALUE DECOMPOSITION COMPONENTS

Expressions for the generalized influence function of principal components were derived by Fung et al. (2007). Here, these results are extended to examine how case-weight perturbations in the distances for a set of genes influences the singular value decomposition of the matrix. We'll begin with some notation. Let X be the $k \times p$ matrix of the dissimilarity vectors for each gene previously defined in chapter 4. Recall that X is the transpose of the matrix D defined in 3.1. Let $\omega = (\omega_1, \omega_2, ..., \omega_k)$ represent perturbations of the dissimilarity vectors for individual genes in matrix X and $\omega_0 = (1, ..., 1)'$ be the vector of ones corresponding to no perturbation. For any statistical function S, denote the perturbed version of S as $S(\omega)$, where $\omega = \omega_0 + \epsilon h$ corresponds to the perturbation in the fixed direction h originating from ω_0. For our purposes, we are mostly interested in the case where S is the first column of the matrix of right eigenvectors V in the singular value decomposition of X. The generalized influence function of S is defined as

$$GIF(S, h) = \lim_{\epsilon \to 0} \frac{S(\omega_0 + \epsilon h) - S(\omega_0)}{\epsilon} \tag{5.1}$$

Since 5.1 is just the directional derivative of $S(\omega)$ in the direction of h, following Fung et al. (2007), we use the implicit function derivative to derive expressions of the generalized influence function for components of the singular value decomposition of the matrix of pairwise distances for multiple genes. We will now derive the local influence functions for the eigenvectors and the singular values of the singular value decomposition.

Let $X(\omega)_{k \times p} = diag(\omega)_{k \times k} X_{k \times p}$ be the perturbed pairwise distance matrix for the data with k genes and $p = \binom{n}{2}$ pairwise distances for n taxa. Then the singular value

decomposition of this perturbed version of X can be written

$$\begin{aligned} X(\omega) &= U(\omega)\Lambda(\omega)V'(\omega) \\ &= (u_1(\omega)....u_m(\omega))_{k\times m} \begin{pmatrix} \lambda_1(\omega) & 0 & ... & 0 \\ 0 & \lambda_2(\omega) & ... & 0 \\ .. & .. & ... & .. \\ .. & .. & ... & .. \\ 0 & 0 & ... & \lambda_m(\omega) \end{pmatrix}_{m\times m} \begin{pmatrix} v_1'(\omega) \\ v_2'(\omega) \\ .. \\ .. \\ v_m'(\omega) \end{pmatrix}_{m\times p} \\ &= \sum_{i=1}^{m} \lambda_i(\omega)u_i(\omega)v_i'(\omega) \end{aligned}$$

where $m = \min(k, p)$.

For the remainder of this section ω will be dropped from the notation, but the reader should note that X, U, Λ and V are perturbed by ω. Let u_j, v_j, and λ_j be the j^{th} left eigenvector, right eigenvector and singular value respectively for $j = 1, ..., m$ Note, that since the columns of U and V are orthogonal, $u_j'u_j = 1$ and $v_j'v_j = 1$, while $u_j'u_q = 0$ and $v_j'v_q = 0$ for all $j \neq q$. Taking the derivative on both sides of $u_j'u_j = 1$ with respect to ω_i, for some $i = 1, ..., k$, we get

$$\begin{aligned} u_j'\frac{\partial u_j}{\partial \omega_i} + \frac{\partial u_j'}{\partial \omega_i}u_j = 0 \\ \Rightarrow u_j'\frac{\partial u_j}{\partial \omega_i} = -\frac{\partial u_j'}{\partial \omega_i}u_j \end{aligned}$$

Therefore,

$$u_j'\frac{\partial u_j}{\partial \omega_i} = \frac{\partial u_j'}{\partial \omega_i}u_j = 0. \tag{5.2}$$

Similarly,

$$v_j'\frac{\partial v_j}{\partial \omega_i} = \frac{\partial v_j'}{\partial \omega_i}v_j = 0. \tag{5.3}$$

Note that $X = U\Lambda V'$ can be rewritten as $XV = U\Lambda$, which means

$$Xv_j = \lambda_j u_j, \tag{5.4}$$

for $j = 1, ..., m$. $X = U\Lambda V'$ may also be rewritten as $U'X = \Lambda V'$, so that

$$u_j' X = \lambda_j v_j', \tag{5.5}$$

for $j = 1, ..., m$. Taking the derivative with respect to ω_i on both sides of (5.4) gives

$$\frac{\partial X}{\partial \omega_i} v_j + X \frac{\partial v_j}{\partial \omega_i} = \frac{\partial \lambda_j}{\partial \omega_i} u_j + \lambda_j \frac{\partial u_j}{\partial \omega_i}, \tag{5.6}$$

and taking the derivative with respect to ω_i on both sides of (5.5) we get

$$\frac{\partial u_j'}{\partial \omega_i} X + u_j' \frac{\partial X}{\partial \omega_i} = \frac{\partial \lambda_j}{\partial \omega_i} v_j' + \lambda_j \frac{\partial v_j'}{\partial \omega_i}. \tag{5.7}$$

Next we left multiply (5.6) by u_j' to get

$$u_j' \frac{\partial X}{\partial \omega_i} v_j + u_j' X \frac{\partial v_j}{\partial \omega_i} = u_j' \frac{\partial \lambda_j}{\partial \omega_i} u_j + u_j' \lambda_j \frac{\partial u_j}{\partial \omega_i}. \tag{5.8}$$

Since $\frac{\partial \lambda_j}{\partial \omega_i}$ and λ_j are scalars, the right-hand side of (5.8) can be rewritten

$$\frac{\partial \lambda_j}{\partial \omega_i} u_j' u_j + \lambda_j u_j' \frac{\partial u_j}{\partial \omega_i},$$

which since $u_j' u_j = 1$ and $u_j' \frac{\partial u_j}{\partial \omega_i} = 0$ from (5.2), becomes $\frac{\partial \lambda_j}{\partial \omega_i}$. Hence, (5.8) reduces to

$$u_j' \frac{\partial X}{\partial \omega_i} v_j + u_j' X \frac{\partial v_j}{\partial \omega_i} = \frac{\partial \lambda_j}{\partial \omega_i}. \tag{5.9}$$

Now if we right multiply (5.7) by v_j, we get

$$\frac{\partial u_j'}{\partial \omega_i} X v_j + u_j' \frac{\partial X}{\partial \omega_i} v_j = \frac{\partial \lambda_j}{\partial \omega_i} v_j' v_j + \lambda_j \frac{\partial v_j'}{\partial \omega_i} v_j. \tag{5.10}$$

Using that $v_j' v_j = 1$ and $\frac{\partial v_j'}{\partial \omega_i} v_j = 0$ (from (5.3)), we find that the right-hand side of (5.10) becomes $\frac{\partial \lambda_j}{\partial \omega_i}$. Inserting this back into (5.10) we get

$$\frac{\partial u_j'}{\partial \omega_i} X v_j + u_j' \frac{\partial X}{\partial \omega_i} v_j = \frac{\partial \lambda_j}{\partial \omega_i}. \tag{5.11}$$

Note that equations (5.9) and (5.11) have terms $\frac{\partial \lambda_j}{\partial \omega_i}$ and $u_j' \frac{\partial X}{\partial \omega_i} v_j$ in common. This implies that

$$u_j' X \frac{\partial v_j}{\partial \omega_i} = \frac{\partial u_j'}{\partial \omega_i} X v_j. \tag{5.12}$$

Replacing Xv_j by $\lambda_j u_j$ in (5.12) gives

$$u_j' X \frac{\partial v_j}{\partial \omega_i} = \lambda_j \frac{\partial u_j'}{\partial \omega_i} u_j.$$

Again by (5.2) $\frac{\partial u_j'}{\partial \omega_i} u_j = 0$ and so we get that

$$u_j' X \frac{\partial v_j}{\partial \omega_i} = 0. \tag{5.13}$$

Inserting (5.13) into (5.9) we get

$$u_j' \frac{\partial X}{\partial \omega_i} v_j = \frac{\partial \lambda_j}{\partial \omega_i},$$

and so the influence function for λ_j is given by

$$\left.\frac{\partial \lambda_j}{\partial \omega_i}\right|_{\omega_0} = u_j' \left.\frac{\partial X}{\partial \omega_i}\right|_{\omega_0} v_j. \tag{5.14}$$

Now we shall derive the influence function for u_j. Returning to equation (5.6), we left multiply (5.6) by u_q', $q \neq j$, to get

$$u_q' \frac{\partial X}{\partial \omega_i} v_j + u_q' X \frac{\partial v_j}{\partial \omega_i} = u_q' \frac{\partial \lambda_j}{\partial \omega_i} u_j + u_q' \lambda_j \frac{\partial u_j}{\partial \omega_i}. \tag{5.15}$$

On the right-hand side $u_q' u_j = 0$. On the left-hand side we can replace $u_q' X$ with $\lambda_q v_q'$. Hence, equation (5.15) becomes

$$u_q' \frac{\partial X}{\partial \omega_i} v_j + \lambda_q v_q' \frac{\partial v_j}{\partial \omega_i} = \lambda_j u_q' \frac{\partial u_j}{\partial \omega_i}. \tag{5.16}$$

Now, taking the transpose of both sides of (5.5), we get

$$X' u_j = \lambda_j v_j. \tag{5.17}$$

Taking the derivative with respect to ω_i of both sides of (5.17) we get

$$X' \frac{\partial u_j}{\partial \omega_i} + \frac{\partial X'}{\partial \omega_i} u_j = \frac{\partial \lambda_j}{\partial \omega_i} v_j + \lambda_j \frac{\partial v_j}{\partial \omega_i}. \tag{5.18}$$

If we left multiply (5.18) by v_q', we get

$$v_q' X' \frac{\partial u_j}{\partial \omega_i} + v_q' \frac{\partial X'}{\partial \omega_i} u_j = \frac{\partial \lambda_j}{\partial \omega_i} v_q' v_j + \lambda_j v_q' \frac{\partial v_j}{\partial \omega_i}. \tag{5.19}$$

Using the fact that $v_q' v_j = 0$ and $v_q' X' = \lambda_q u_q'$ we can rewrite (5.19)

$$\lambda_q u_q' \frac{\partial u_j}{\partial \omega_i} + v_q' \frac{\partial X'}{\partial \omega_i} u_j = \lambda_j v_q' \frac{\partial v_j}{\partial \omega_i}. \tag{5.20}$$

Note, that (5.20) can be rewritten

$$v_q' \frac{\partial v_j}{\partial \omega_i} = \frac{1}{\lambda_j} \left[\lambda_q u_q' \frac{\partial u_j}{\partial \omega_i} + v_q' \frac{\partial X'}{\partial \omega_i} u_j \right]. \tag{5.21}$$

Now inserting (5.21) into (5.16) gives

$$\lambda_j u_q' \frac{\partial u_j}{\partial \omega_i} = u_q' \frac{\partial X}{\partial \omega_i} v_j + \lambda_q \left[\frac{1}{\lambda_j} \left(\lambda_q u_q' \frac{\partial u_j}{\partial \omega_i} + v_q' \frac{\partial X'}{\partial \omega_i} u_j \right) \right]. \tag{5.22}$$

After some algebraic manipulation, (5.22) reduces to

$$u_q' \frac{\partial u_j}{\partial \omega_i} = \frac{1}{\lambda_j^2 - \lambda_q^2} \left(\lambda_j u_q' \frac{\partial X}{\partial \omega_i} v_j + \lambda_q v_q' \frac{\partial X'}{\partial \omega_i} u_j \right). \tag{5.23}$$

Denote the right-hand side of (5.23) b_{ji}^q. Then

$$u_q' \frac{\partial u_j}{\partial \omega_i} = b_{ji}^q.$$

Note, that $u_q' \frac{\partial u_j}{\partial \omega_i}$ gives the projection of the vector $\frac{\partial u_j}{\partial \omega_i}$ onto u_q'. The columns of U span an orthogonal basis for $\mathbb{R}^m$, therefore $\frac{\partial u_j}{\partial \omega_i}$ must lie in the same space as the columns of U. Hence $\frac{\partial u_j}{\partial \omega_i}$ can be expressed as a sum of $b_{ji}^q u_q$, which gives

$$\begin{aligned} \frac{\partial u_j}{\partial \omega_i} &= \sum_{q=1}^{m} \left(u_q' \frac{\partial u_j}{\partial \omega_i} \right) u_q \\ &= \sum_{q \neq j}^{m} b_{ji}^q u_q, \end{aligned} \tag{5.24}$$

since $u_j' \frac{\partial u_j}{\partial \omega_i} = 0$. Hence, the influence function for u_j is

$$\left. \frac{\partial u_j}{\partial \omega_i} \right|_{\omega_0} = \sum_{q \neq j}^{m} b_{ji}^q u_q. \tag{5.25}$$

To obtain the influence function for v_j, note the (5.16) can be rewritten

$$u_q' \frac{\partial u_j}{\partial \omega_i} = \frac{1}{\lambda_j} \left[u_q' \frac{\partial X}{\partial \omega_i} v_j + \lambda_q v_q' \frac{\partial v_j}{\partial \omega_i} \right]. \tag{5.26}$$

Then inserting (5.26) into (5.20) we get

$$\lambda_q \left[\frac{1}{\lambda_j} \left(u_q' \frac{\partial X}{\partial \omega_i} v_j + \lambda_q v_q' \frac{\partial v_j}{\partial \omega_i} \right) \right] + v_q' \frac{\partial X'}{\partial \omega_i} u_j = \lambda_j v_q' \frac{\partial v_j}{\partial \omega_i}. \tag{5.27}$$

After some algebraic manipulation (5.27) reduces to

$$v_q' \frac{\partial v_j}{\partial \omega_i} = \frac{1}{\lambda_j^2 - \lambda_q^2} \left[\lambda_q u_q' \frac{\partial X}{\partial \omega_i} v_j + \lambda_j v_q' \frac{\partial X'}{\partial \omega_i} u_j \right]. \tag{5.28}$$

Denote the right-hand side of (5.28) by a_{ji}^q. We then get

$$v_q' \frac{\partial v_j}{\partial \omega_i} = a_{ji}^q.$$

The columns of V span an orthogonal basis for $\mathbb{R}^m$, therefore $\frac{\partial v_j}{\partial \omega_i}$ must lie in the same space as the columns of V. Hence $\frac{\partial v_j}{\partial \omega_i}$ can be expressed as a sum of $a_{ji}^q v_q$, which gives

$$\frac{\partial v_j}{\partial \omega_i} = \sum_{q \neq j}^{m} a_{ji}^q v_q. \tag{5.29}$$

Hence, the influence function for v_j is

$$\left. \frac{\partial v_j}{\partial \omega_i} \right|_{\omega_0} = \sum_{q \neq j}^{m} a_{ji}^q v_q. \tag{5.30}$$

5.2 CASE-WEIGHTS PERTURBATION

To assess how a perturbation in the distances affects the estimated tree topology, and identify which genes are influential in determining the combined-gene tree, we can think of this as determining the influence of a perturbation in distances on the first right eigenvector of the singular value decomposition of matrix X. To do this, expressions for the generalized influence functions defined in (5.14), (5.25) and (5.30) for case-weights perturbation will now be derived. Again we let $\omega = (\omega_1, ..., \omega_k)$ be some perturbation of the dissimilarity vectors for individual genes in X. Then the perturbed matrix of distance vectors $X(\omega)$ can be written

$$X(\omega) = \begin{pmatrix} \omega_1 x_{1,1} & \omega_1 x_{1,2} & ... & \omega_1 x_{1,p} \\ \omega_2 x_{2,1} & \omega_2 x_{2,2} & ... & \omega_2 x_{2,p} \\ .. & .. & ... & .. \\ .. & .. & ... & .. \\ \omega_k x_{k,1} & \omega_k x_{k,2} & ... & \omega_k x_{k,p} \end{pmatrix}.$$

Hence, under the case-weights perturbation scheme,

$$\frac{\partial X(\omega)}{\partial \omega_i} = \begin{pmatrix} 0 & 0 & ... & 0 \\ .. & .. & ... & .. \\ x_{i1} & x_{i2} & .... & x_{ip} \\ .. & .. & ... & .. \\ 0 & 0 & ... & 0 \end{pmatrix}. \tag{5.31}$$

Again, we will drop the ω from the notation, but note that λ_j, u_j and v_j all correspond to the singular value decomposition of the perturbed version of X. If we plug (5.31) into (5.14) we get that the influence function for λ_j under the case-weights perturbation scheme is given by

$$\left.\frac{\partial \lambda_j}{\partial \omega_i}\right|_{\omega_0} = u_{ji} \sum_{l \neq j}^{p} x_{il} v_{lj}. \tag{5.32}$$

Inserting (5.31) into equation (5.30), the GIF for v_j, we get find that

$$a_{ji}^{q} = \frac{1}{\lambda_j^2 - \lambda_q^2} \left(\lambda_q u_{qi} \sum_{l=1}^{p} x_{il} v_{lj} + \lambda_j u_{ji} \sum_{l=1}^{p} x_{il} v_{lq} \right). \tag{5.33}$$

Since $\sum_{l=1}^{p} x_{il} v_{lj} = x_i' v_j = \lambda_j u_{ji}$ and $\sum_{l=1}^{p} x_{il} v_{lq} = x_i' v_q = \lambda_q u_{qi}$ with a few substitutions in (5.33) we find that

$$a_{ji}^{q} = \frac{1}{\lambda_j^2 - \lambda_q^2} \left(\lambda_q u_{qi} \lambda_j u_{ji} + \lambda_j u_{ji} \lambda_q u_{qi} \right), \tag{5.34}$$

and hence, inserting (5.34) back into (5.30) we find that under the case-weights perturbation scheme the influence function for v_j is given by

$$\left.\frac{\partial v_j}{\partial \omega_i}\right|_{\omega_0} = \sum_{q \neq j}^{p} \frac{2\lambda_q \lambda_j u_{qi} u_{ji}}{\lambda_j^2 - \lambda_q^2} v_q. \tag{5.35}$$

Similarly, if we plug (5.31) into (5.25), we get that

$$b_{ji}^{q} = \frac{1}{\lambda_j^2 - \lambda_q^2} \left(\lambda_j u_{qi} \sum_{l=1}^{p} x_{il} v_{lj} + \lambda_q u_{ji} \sum_{l=1}^{p} x_{il} v_{lq} \right), \tag{5.36}$$

which with a few substitutions into (5.36) we get

$$b_{ji}^{q} = \frac{1}{\lambda_j^2 - \lambda_q^2} \left(\lambda_j u_{qi} \lambda_j u_{ji} + \lambda_q u_{ji} \lambda_q u_{qi} \right). \tag{5.37}$$

Hence, inserting (5.37) back into (5.25) we find that the influence function for u_j under the case-weights perturbation scheme is given by

$$\left.\frac{\partial u_j}{\partial \omega_i}\right|_{\omega_0} = \sum_{q \neq j}^{p} \left(\frac{\lambda_j^2 + \lambda_q^2}{\lambda_j^2 - \lambda_q^2} u_{qi} u_{ji} \right) u_q. \tag{5.38}$$

5.3 IDENTIFYING INFLUENTIAL GENES USING THE CASE - WEIGHTS PERTURBATION

To assess the combined influence of multiple genes on an estimated combined-gene tree we begin by examining how a perturbation in the distances affects the first right eigenvector, v_1, of the singular value decomposition of X. Note that this vector corresponds to the direction of the first principal component. We begin by computing the influence of v_1 for each ω_i, $i = 1, 2, ..., k$, to get a $p \times k$ matrix M made up of $\left.\frac{\partial v_1}{\partial \omega_1}\right|_{\omega_0}, ..., \left.\frac{\partial v_1}{\partial \omega_k}\right|_{\omega_0}$. That is,

$$M = \left(\frac{\partial v_1}{\partial \omega_1}, ..., \frac{\partial v_1}{\partial \omega_k} \right) \Bigg|_{\omega_0}. \tag{5.39}$$

To find the direction h in which the local perturbation on ω will generate the largest change in the vector v_1, the matrix M can be decomposed as follows

$$M = U^D \Lambda^D (V^D)', \tag{5.40}$$

where, letting $r = min(p, k)$, U^D is a $p \times r$ matrix of left eigenvectors, V^D is a $k \times r$ matrix of right eigenvectors and Λ is an $r \times r$ matrix of singular values. The GIF can be expressed as

$$GIF(v_1, h) = \lim_{a \to 0} \frac{v_1(\omega_0 + ah) - v_1(\omega_0)}{a}, \tag{5.41}$$

where $v_1(\omega_0 + ah) \approx v_1(\omega_0) + aMh$ and h corresponds to the fixed direction in which the distances are perturbed. If the distances are perturbed in the direction v_j^D, $Mv_j^D = \lambda_j^D u_j^D$ describes how a perturbation in the direction of v_j^D affects the pairwise distances. The first column of V^D gives the most influential direction h. We can interpret the case-weights perturbation scheme in the context of distance vectors in two ways. Unlike the MinCV, the singular value decomposition method

is not scale invariant. We are interested in how assigning different weights to distance matrices for individual genes affects the estimated combined-gene distances obtained with the singular value decomposition method. If certain genes are more influential in determining the combined-gene topology than others, we would expect that perturbing the weights of those particular genes would have a greater effect on the estimated topology than perturbing the weights of other less influential genes. From a geometric perspective, we can think of the case-weights perturbation as a perturbation of the lengths of the p-vectors for each gene. We wish to evaluate how a perturbation in the lengths of this vector affects the principal direction of the singular value decomposition which in turn determines the combined-gene topology.

The perturbed version of v_1 under perturbation scheme $\omega = \omega_0 + ah$ when $h = v_1^D$ can be written

$$v_1(\omega_1) := v_1(\omega_0) + a\lambda_1^D u_1^D, \tag{5.42}$$

where $\lambda_1^D u_1^D$ corresponds to the perturbation. Another perturbation of interest is $h = v_2^D$,

$$v_1(\omega_2) := v_1(\omega_0) + a\lambda_2^D u_2^D, \tag{5.43}$$

which corresponds to the second greatest fluctuation on v_1. The stability of the estimated combined-gene tree under perturbations in the distances can be evaluated by determining how large the quantity a must be before placement of the taxa in the estimated tree is affected by the perturbation. In short, we are interested in the effect a perturbation in the directions v_1^D and v_2^D has on the first right eigenvector, v_1 of the singular value decomposition of X, which describes the combined-gene topology for a set of taxa pairs.

5.4 RESULTS

5.4.1 INFLUENCE ANALYSIS ON THE PRIMATE DATA

The 13 protein coding genes in the mitchondrial genome common to the five primates consist of three cyclooxygenases, six NADH dehyrogenases, ATP synthase subunit 6, ATP synthase subunit 8 and and cytochrome b. The six NADH dehydrogenases, the

ATP synthases and cytochome b belong to the mitochondrial electron transport chain (Murray et al., 2003). The COX genes are catalysts in the conversion of essential fatty acids into prostanoids (Chandrasekharan and Simmons, 2004; Chandrasekharan et al., 2002).

Figure 5.1 shows the first two right eigenvectors of the singular value decomposition of matrix M. These values are a measure of the influence of each gene in the first and second principal directions. In the first direction, ND4L appears to have greatest weight and so we would expect this gene to be highly influential in the estimated combined tree topology. ATP6 and ND6 are given moderate weights. All remaining genes are assigned small weights in between -0.2 and 0.2, and hence, are likely not very influential in determining the combined the gene topology. In the second direction the greatest weights are given to COX1 and ATP6, with moderate weights given to ND1 and ND2 and all remaining genes assigned small weights in between -0.2 and 0.2. Hence, it would appear that ND4L, COX1 and ATP6 are highly influential, while ND1, ND2 and ND6 are moderately influential, and the remaining six genes are given very little weight in determining the combined-gene tree.

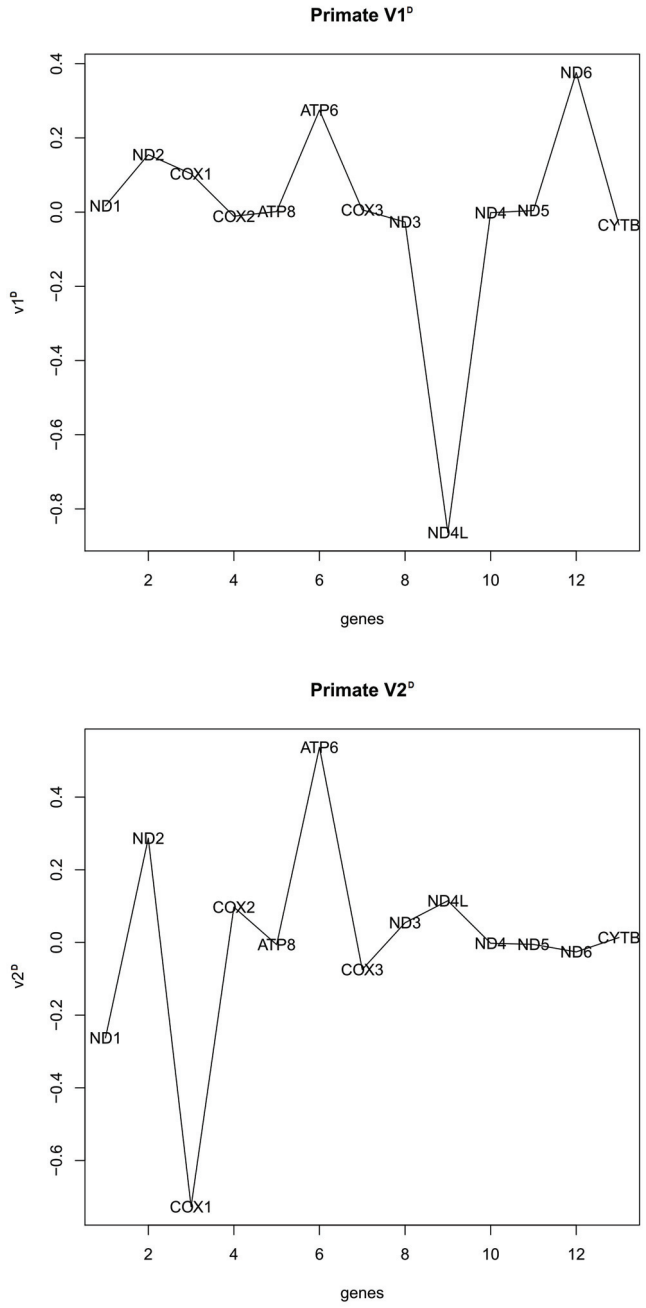

Figure 5.1: Influence to the SVD combined pairwise distances of different genes under the case weight perturbation along the first principal direction (top) and the second principal direction (bottom).

The genes in the first row of Table 5.1 correspond to the reference tree topology shown in Figure 3.12. The genes in the second row place human and gorilla as sister

taxa with chimp branching before these two. ATP8 and ND4 recover an erroneous tree with human branching before gorilla and chimp. Neither of these two genes are identified as influential. ATP6, COX1, and ND4L, which are identified as highly influential recover the reference tree topology. ND2, which is identified as moderately influential also recovers the reference tree topology.

We begin by examining how a perturbation in dissimilarities for the combined-gene tree topology affects the estimated tree. Combined-gene topologies for different values of a under the perturbations defined in equations (5.42) and (5.43), respectively, were estimated using BIONJ (Gascuel, 1997). Under perturbation $h = v_1^D$ at $a = 3.7$ gorilla branches before the orangutan. Under perturbation $h = v_2^D$ at $a = 1.4$ chimp branches before gorilla. Larger values of a did not perturb the estimated topologies any further. The changed topologies under the two perturbation schemes are shown in Figure 5.2.

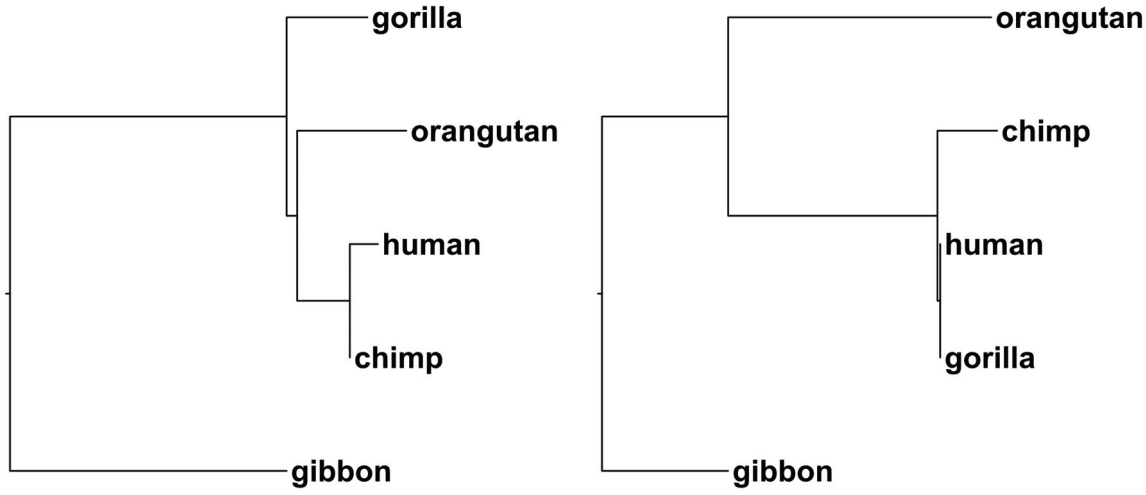

Figure 5.2: Effect of perturbation scheme 1 on estimated topology for a = 3.7 (left) and effect of perturbation scheme 2 for a = 1.4 (right)

Next, we observed how removal of individual influential genes and combinations genes identified as influential affects the combined-gene tree estimated from the remaining genes. As an initial step, genes were removed from the analysis individually to determine if removal of any one gene affected the estimated combined-gene topology shown in top right-hand panel of Figure 4.5. No single gene was identified

whose exclusion from the analysis caused a change in the estimated combined-gene tree. Next, the genes identified as highly influential by the influence function for the combined-gene dissimilarities were removed from the analysis. The combined removal of ND4L, COX1 and ATP6 did not result in a change in the estimated combined-gene tree. Nor did the combined removal of ND1, ND2 and ND6. Several different combination of these six genes were removed from the analysis and the effect on the combined-gene topology examined. Only when highly influential genes ATP6 and COX1 and moderately influential genes ND1 and ND2 were removed together did the estimated combined-gene topology change. Removal of these genes resulted human erroneously branching as an outgroup to gorilla and chimp. It appears that the estimated combined-gene tree for the primate data derived from the singular value decomposition is fairly robust and resistant to both perturbations in the dissimilarities and sampling of genes.

Figure 5.3 shows which pairwise dissimilarities are most affected by a perturbation along v_1^D (upper panel) and along v_2^D (lower panel). We can see that when a perturbation along the first principal direction v_1^D is applied the taxa pair distance with the greatest change is chimp/orangutan. Gibbon/gorilla, gibbon/human and gibbon/orangutan also change by a large amount. Where the distance between chimp and orangutan is inflated, the other three dissimilarities shrink. When a perturbation along the second principal direction v_2^D is applied the taxa pairs most affected are gorilla/human whose distance is inflated and chimp/gibbon whose distance shrinks.

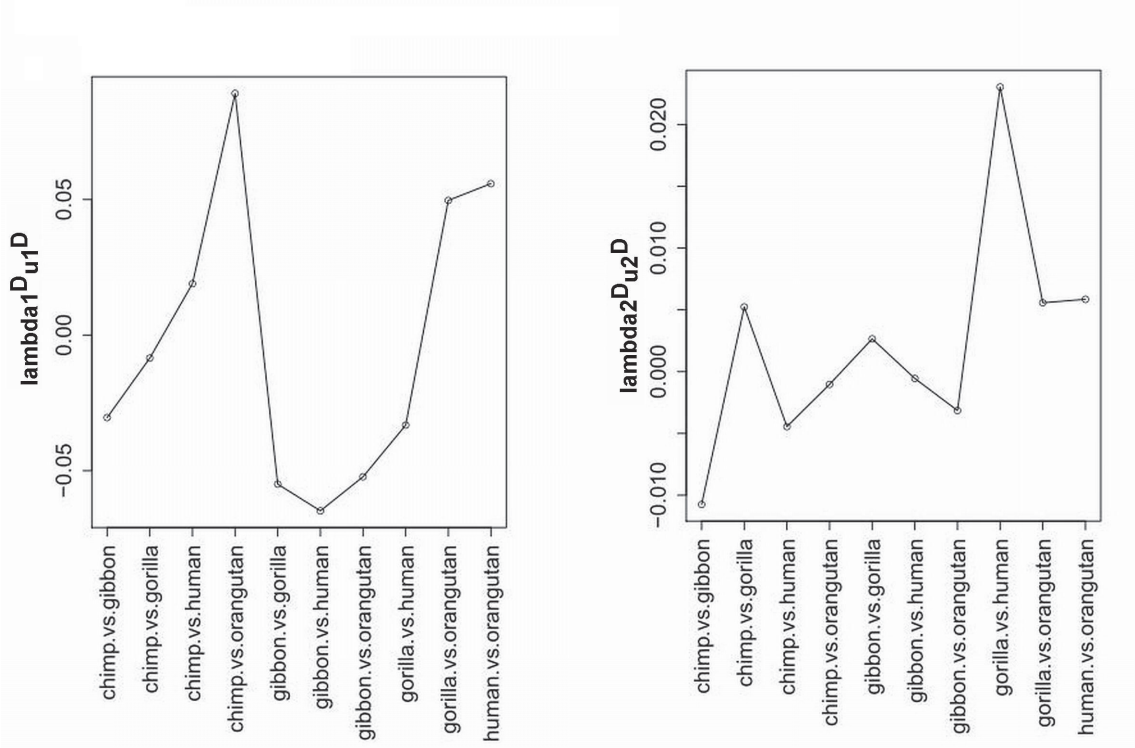

Figure 5.3: Taxa pair whose dissimilarities are most affected by a perturbation along v_1^D(left) and v_2^D (right).

5.4.2 INFLUENCE ANALYSIS ON THE NEMATODE DATA

Our method of influence analysis is next applied to the nematode dataset. The genes in the nematode data set consist of three cyclooxygenases, six NADH dehyrogenases, one ATP synthase subunit 6 and cytochrome b.

Plots of the influence for each gene under perturbations in the first two principal directions are shown in Figure 5.4. ATP6, COX1, COX3, ND1 and ND6 appear to be highly influential in determining the estimated combined-gene topology. In the first direction, $h = v_1^D$, ATP6 and COX3 appear to be the most influential, with COX1 and ND6 moderately influential. In the second direction, $h = v_2^D$, ATP6, COX3 and ND1 are the most influential, with ND6 moderately influential.

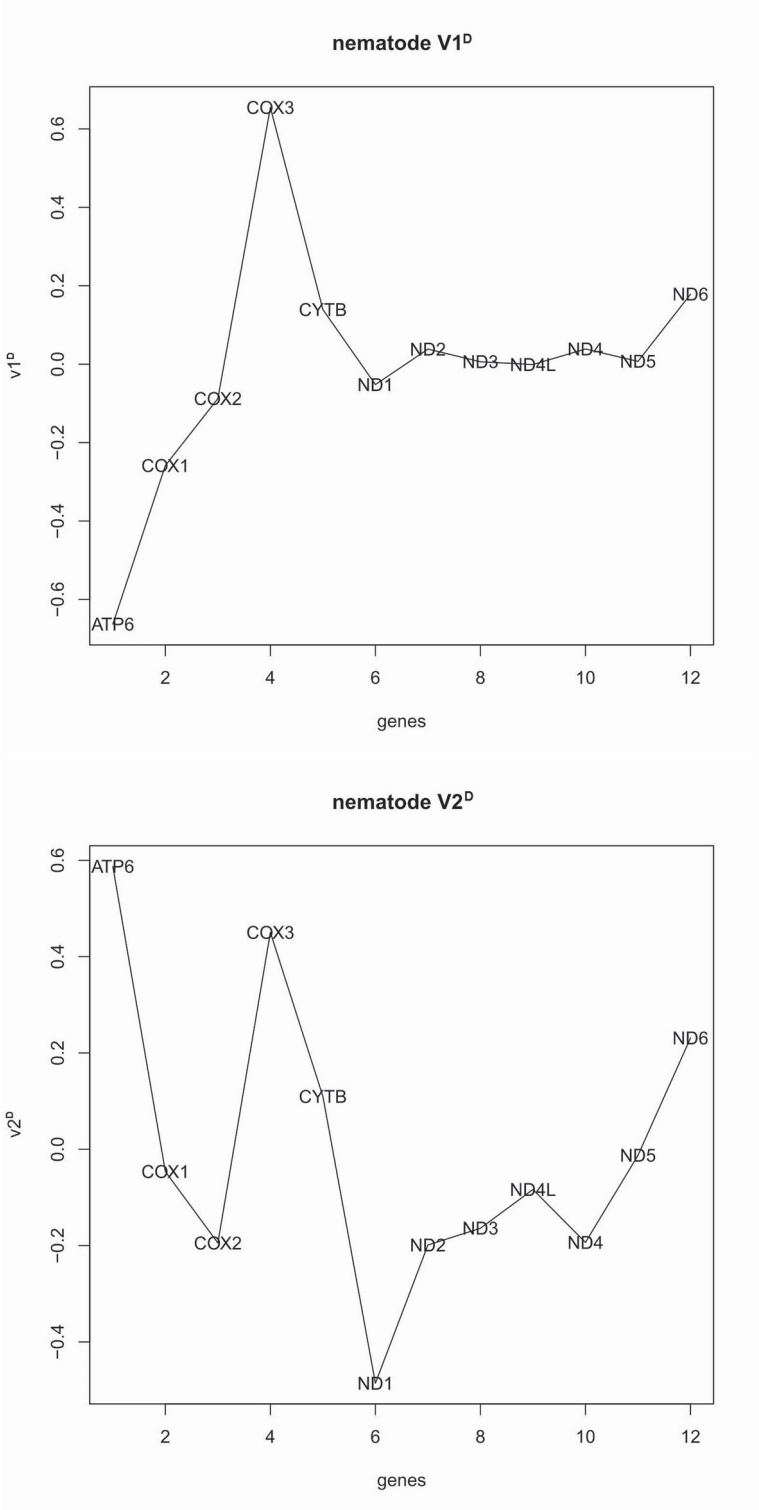

Figure 5.4: Influence to the SVD combined pairwise distances between all taxa pairs under the case weight perturbation along the first principal direction (top) and the second principal direction (bottom).

For the nematode data set, none of the individual genes recovers the reference tree topology under either the edcysozoa or the coelomata hypotheses, however in terms of the relative placement of the nematode, vertebrates and arthropods, some genes tend to recover topologies closer to the reference tree under one hypothesis or the other. Topologies recovered by ATP6, COX1, ND1 and ND5, agree with the topology under the ecdysozoa hypothesis in that the nematode and arthropods group together, but the relative placement of the taxa within the arthropod clade is incorrect, and only ATP6 separates the nematode taxon roundworm, and the arthropod taxon honeybee. Topologies for genes COX2, COX3 and ND3 agree with the topology under the coelomata hypothesis in that the nematode is placed as an outgroup of the vertebrates and arthropods but the relative placement the arthropod and vertebrate taxa is incorrect. Tree topologies for ND2, ND4, ND4L and ND6 agree with neither hypothesis, but topologies for ND4, ND4L and ND6 separate the roundworm and honeybee. Influential genes ATP6, COX3 and ND6 all recover topologies which separate the roundworm and honeybee. ATP6 recovers the topology closest to that of the reference tree under the ecydsozoa hypothesis, only differing with regard to the relative placement of honeybee and brine shrimp. COX1 recovers the next closest topology, with roundworm and honeybee being erroneously branched as sister taxa in a clade of arthropods. ND1 is also closer to the reference tree under the ecdysozoa hypothesis, but the placement of the taxa within the arthropod clade is erroneous. The gene COX3 has a topology that is closer to the topology under the coelomata hypothesis, with the nematode branching before the vertebrates. The topological features of interest recovered by the individual gene trees are summarized in Table 5.2.

We next examined how a perturbation in the combined-gene dissimilarities affects the estimated tree topology. The perturbations described in (5.42) and (5.43), respectively, were applied to the nematode data set. Under a perturbation in the direction of $h = v_1^D$, the roundworm and honeybee erroneously branch as sister taxa at $a = 1.1$ and at $a = 1.5$ the topology within the arthropod clade breaks down. Under a perturbation in the direction of $h = v_2^D$, at $a = 2.0$ the honeybee and brine

shrimp are placed together as sister taxa. Figure 5.5 shows the resulting topologies under perturbations for different values of a.

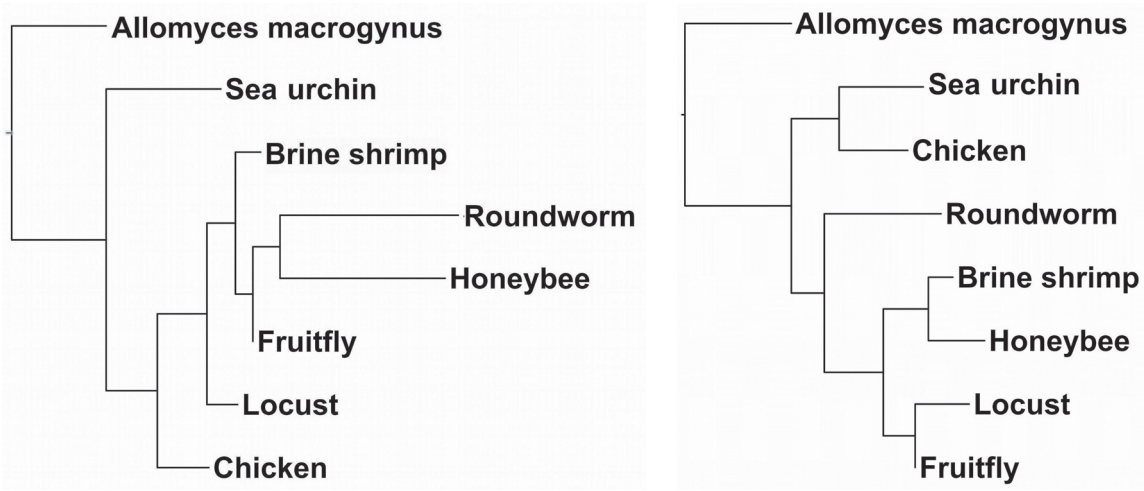

Figure 5.5: Effect of perturbation scheme 1 on estimated topology for a = 1.5 (left). Effect of perturbation scheme 2 on estimated topology for a = 2.0 (right).

Again individual genes and combinations of the genes identified as influential were removed from the combined-gene analyses to further examine their effect on the estimated topology. Removal of a single gene from the analysis resulted in no change in topology. Note that this was not the case when the MinCV method was applied to this data set. We next removed all the influential genes, which resulted in a tree in which honeybee and roundworm were erroneously grouped as sister taxa. Upon further examination, it was discovered that among the influential genes, only the combined removal of a ATP6 and ND6 appeared to have any effect on the estimated topology. Removal of COX1, COX3 and ND1 did not affect the topology. When these three and ATP6 were removed the topology remained the same. Likewise, when these three and ND6 were removed the topology remained the same. The combined removal of ATP6 and ND6 resulted in the roundworm and honeybee being as sister taxa. Hence, it appears that these two genes are the most influential in the proper separation of honeybee and roundworm. Note, that these genes were also found influential in the combined-gene tree estimated with the MinCV scale coefficients in Chapter 3. The tree obtained when these two genes were removed is shown in Figure 5.6.

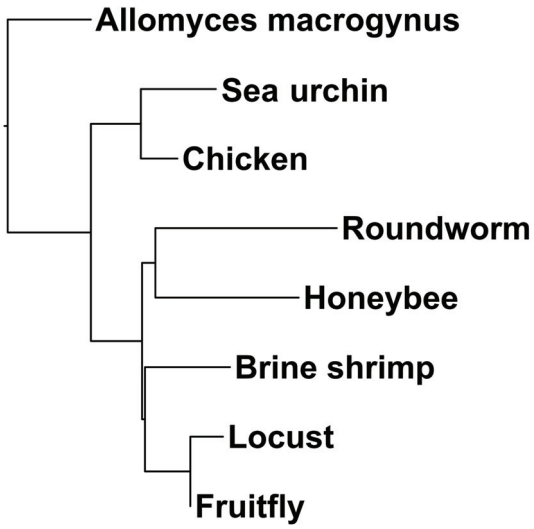

Figure 5.6: Estimated combined-gene nematode tree when genes ATP6 and ND6 are removed from the analysis.

Lastly, we examined which pairwise dissimilarities are most affected by a perturbation in the first and second principal directions. Figure 5.7 shows which taxa pairwise dissimilarities in the nematode data set are most effected by a perturbation along v_1^D (upper panel) and along v_2^D (lower panel). When a perturbation in the direction $h = v_1^D$ is applied, combined-gene dissimilarities for fruitfly/brine shrimp, chicken/brine shrimp, chicken/locust, brine shrimp/locust, fruitfly/sea urchin and roundworm/sea urchin appear to be the most affected. The dissimilarities for fruitly/brine shrimp, chicken/brine shrimp, chicken/locust and brine shrimp/locust are all inflated while the dissimilarities for fruitfly/sea urchin and roundworm/sea urchin all shrink. When a perturbation in the direction $h = v_2^D$ is applied, dissimilarities between honeybee/fruitfly, honeybee/brine shrimp and locust/sea urchin appear to undergo the greatest changes. Dissimilarities for honeybee/fruitfly and honeybee/brine shrimp are inflated while the dissimilarity for locust/sea urchin shrinks.

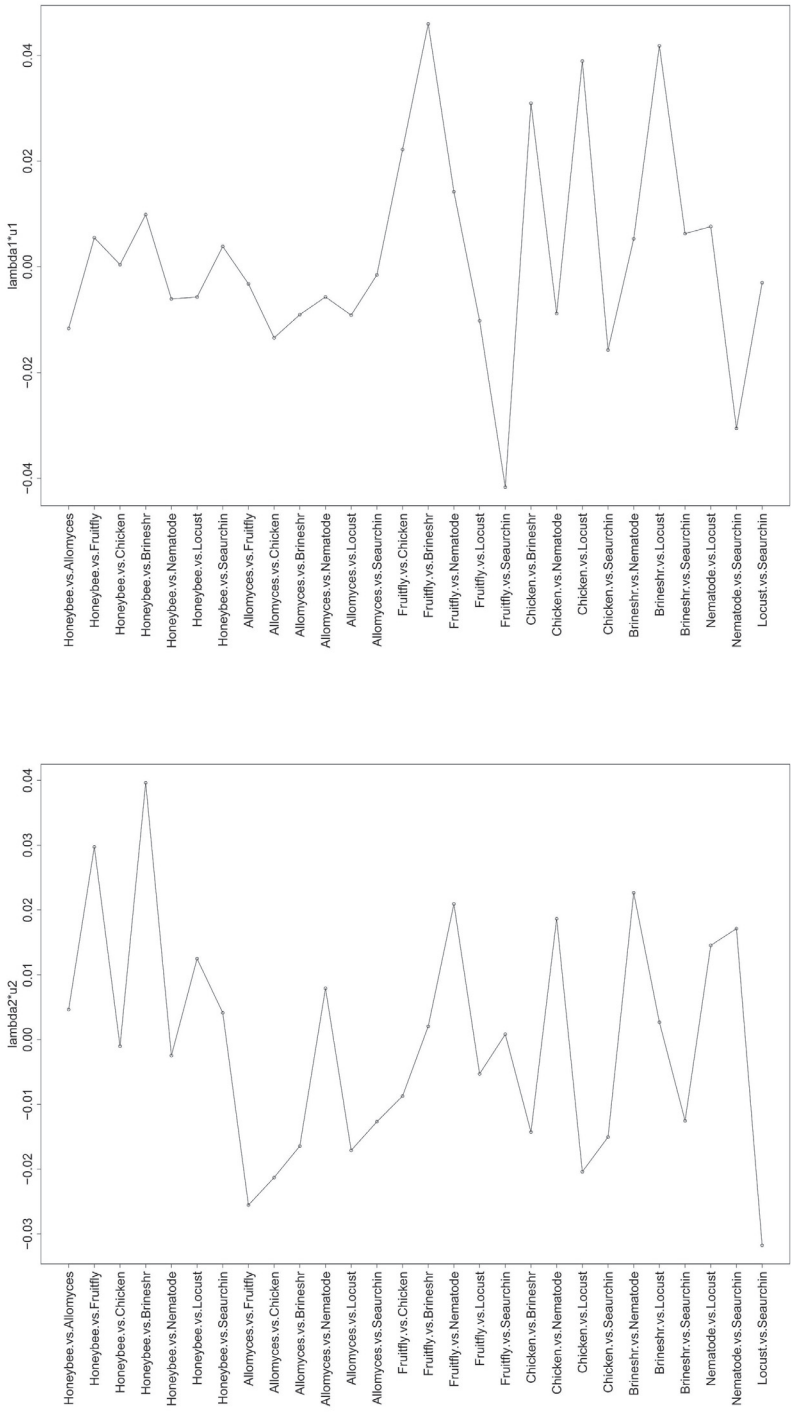

Figure 5.7: Case-weights, or influence, of individual taxa pairs for the estimated combined-gene nematode tree in the first principal direction (top) and the second principal direction (bottom).

5.5 DISCUSSION

The above analysis suggests that the singular value decomposition method of deriving single tree representation from multiple genes is fairly robust to perturbations in dissimilarities and in sampling of genes. Surprisingly, the estimated combined-gene tree for the nematode data set, which is a difficult data set to resolve, appears to be as stable under such perturbations as the primate data set. For both data sets, removing a single gene from the analysis did not affect the combined-gene topology obtained with the singular value decomposition method suggesting that the singular value decomposition method of combining genes is more stable than the MinCV method in this respect. Recall that for the nematode data set, removing a single gene from the analysis resulted in a different combined-gene tree in some cases for the MinCV method. Influence analysis of both the MinCV and singular value decomposition methods of combining genes identified ATP6 and ND6 as influential in determining the combined-gene nematode tree. Note again, that while here the method was applied on the common scaling covariance based dissimilarities, this method could be applied to any distance measure.

In the examples examined in this chapter the singular value decomposition based combined-gene trees agreed with the reference tree topologies presented in chapter 3. For these cases, we would expect the genes identified as influential to agree with the reference tree with respect to one or more clades, while genes which recover a topology which disagrees with the reference tree will be less influential, thus won't have large weights in the resulted weight vector to indicate the most influential direction. In a case where the combined-gene tree disagrees with the reference tree, we would expect that the most influential genes will be those which also recover a tree which disagrees with reference tree topology.

In this chapter we focused on the case-weights perturbation scheme to identify possible influential genes. Another perturbation scheme that may be of interest is an additive perturbation scheme, where a vector of weights is added to the k-vectors which consist of the pairwise dissimilarities for k genes for a given pair of taxa. In this manner, we can identify which pairs of taxa are most influential in the combined-gene tree.

Table 5.1: Estimated BIONJ topologies for individual genes in the Primate data

Topology	Gene
(gibbon,(orangutan,(gorilla,(human,chimp))))	ATP6, COX1, COX2, COX3, ND2, ND4L, ND5
(gibbon,(orangutan,(chimp,(human,gorilla))))	CYTB, ND1, ND3, ND6
(gibbon(orangutan,(human,(gorilla,chimp))))	ATP8, ND4

Table 5.2: Topological features of interest in the individual gene BIONJ trees derived from common scaling based dissimilarities.

Ecdysozoa and separates honeybee nematode	Ecdysozoa but honeybee + nematode as sister taxa	Coelomata but incorr. placement within arthropod and vertebrate clades	No hypothesis but separates honeybee nematode	No hypothesis but honeybee+nematode
ATP6	COX1	COX2	ND4	CYTB
	ND1	COX3	ND4L	ND2
	ND5	ND3	ND6	

Chapter 6

PROTEIN STRUCTURE PREDICTION AND THE SPECTRAL ENVELOPE

Protein structure prediction is an important aspect of the field of pharmaceutical medicine where the three dimensional structure of a target protein is used to discover new drug candidates (Hubbard, 2006). Numerous advances in protein structure prediction have been made over the past 60 years and have been increasing rapidly as advances in technology enable higher computational power. Most prediction methods rely extensively on knowledge of proteins whose structures have been previously determined by X-ray crystallography or NMR spectroscopy (Ginalski et al., 2005). Homology modeling and fold recognition methods rely on the availability of templates in protein databases. Although there has been moderate success in predicting some small proteins and protein fragments ab initio, the most accurate and successful predictions are made using comparative methods which require sequence or structural alignments with previously solved proteins (Zhang, 2008; Wooley and Ye, 2007; Hubbard, 2006). Template-free or *ab initio* methods seek to model the energetics of the protein folding process to determine three dimensional structure from amino acid sequence. The Rosetta software developed by Baker has successfully predicted several small protein structures but is computationally expensive (Bradley et al., 2006; Das et al., 2007; Zhang, 2008). More recently, the field of structural genomics has arisen in which structural biologists predict protein structures on a genome-wide scale. The structural genomics approach has increased the rate at which novel protein structures are solved and made available while reducing the cost. The success rate of structural genomics methods is comparable to traditional structural biology approaches (Brown and Flocco, 2006; Chandonia and Brenner, 2006). The CASP (Critical Assessment of techniques for protein Structure Prediction) experiment has objectively assessed the quality of current methods and measured progress in protein prediction on a

bi-annual basis since 1994. Success of comparative modelling methods have steadily increased over the last decade, although such methods are still constrained by errors in sequence alignment and the requirement that evolutionary relationships exist between unsolved and solved proteins.

In this chapter we will apply the spectral envelope described in chapter 2 to the problem of protein prediction. Note, that unlike the spectral covariance which measures how two sequences vary together, the spectral envelope is applied to individual sequences. As is the case with the spectral covariance, peaks in the spectral envelope correspond to secondary structures present in a protein. We do not attempt to model the particular folding patterns of proteins. Instead, using the structural information coded in the periodic patterns of its amino acid sequence, we attempt to classify the protein into a structural category, and in this way identify the main structural elements of the protein of interest and possible templates. The method we propose is purely statistical and requires no sequence alignment or homologous comparisons between proteins. Recall that the peaks of the spectral envelope of amino acid sequences are related to the periodicity of protein secondary structures. When applied to sequences corresponding to α-helices the maximum of the spectral envelope occurs at a periodicity of 3.6 residues or $\omega = 0.277$, whereas sequences corresponding to β-sheets should have a maximum of the spectral envelope at $\omega = 0.435$ (periodicity 2.3 residues). The 3_{10}-helix has a periodicity 2.5 residues while loop regions have a periodicity of 3-4 residues (Eisenberg et al., 1984). The multiple peaks in the spectra of protein sequences correspond to the periodicities of the structural elements of a protein (Collins et al., 2006).

The classification and regression tree (CART) and bootstrap aggregating (bagging) methods are applied with the spectral envelope based covariates to classify proteins into their structural classes. We compare our method to the classification capabilities of BLAST using the BLASTP algorithm (protein-protein blast).

6.1 CLASSIFICATION TREES

Tree based classifiers provide a simple but powerful solution to the classification problem. They are simple to interpret and are capable of handling large amounts

of data with ease. Another advantage of the classification tree is that it requires no assumption with regard to the distribution of the data as would be the case for a method such as logistic regression. However, classification trees can have problems with over-fitting, returning trees that do not generalize well to new observations. This problem can be mitigated by pruning. Classification trees are grown using binary recursive partitioning. The goal is to sort objects into the correct classes using some splitting criterion. The classification tree algorithm chooses the best split by considering every possible split and calculating the homogeneity of the resulting two partitions or nodes. The split resulting in the purest nodes is chosen to partition the data. This process is repeated for each node until all the nodes are pure or the data is too sparse.

The construction of a classification tree consists of two steps: growth and pruning. Near the root of the tree, a large amount of information is incorporated into the first few splits of the tree because a large portion of the observations is used to choose the appropriate split, while fewer observations are used to split successive nodes. Splits near the bottom of the tree tend to reflect anomalies peculiar to the data, which results in the full grown tree having poor predictive power for future observations. Pruning the tree removes random elements that fail to generalize to new data and improves the tree's predictive power. The process of growing and pruning classification trees will now be described in further detail.

6.1.1 GROWING CLASSIFICATION TREES

Let R_m be the region representing node m of a classification tree and N_m be the number of observations in node m. Then,

$$\hat{p}_{mk} = \frac{1}{N_m} \sum_{y_i \in R_m} I(y_i = k)$$

is the proportion of class k observations in node m (Hastie et al., 2001). Observations in node m are considered to be of class k if the majority of observations at that node belong to class k. We begin with all observations in a single node and apply successive binary splits such that node impurity is decreased at each split. There are three common measures of node impurity: the misclassification error, the Gini index and

the deviance (Hastie et al., 2001). In this chapter we use the tree() function in the package **tree** of the R programming language (Ripley, 2010) to obtain classification trees. The tree() function assumes a multinomial response variable and hence uses the deviance to split the nodes. Each node m is split such that

$$-\sum_{k=1}^{K} n_{mk} \log \hat{p}_{mk}$$

is minimized, where K is the number of classes. By convention it is assumed that $0 \log 0 = 0$ when $\hat{p}_{mk} = 0$ since $\lim_{x \to 0} x \log x = 0$ (Hastie et al., 2001). Splitting continues until the nodes are pure or each terminal node contains no more than five observations (Ripley, 2010).

6.1.2 PRUNING CLASSIFICATION TREES

Trees are pruned using minimal cost-complexity pruning. Cost complexity pruning is usually performed using the misclassification error,

$$\frac{1}{N_m} \sum_{y_i \in R_m} I(y_i \neq k_m) = 1 - \hat{p}_{mk_m},$$

where k_m is the majority class in node m (Hastie et al., 2001). Here $\hat{p}_{mk_m}$ is the proportion of observations in node m that belong to majority class k_m, that is, the proportion of observations correctly classified. For any subtree T_s of the full tree T_{max} let $\mid T \mid$ denote the number of terminal nodes. The cost complexity criterion is then

$$C_\alpha(T) = \sum_{m=1}^{|T|} 1 - \hat{p}_{mk_m} + \alpha |T|, \tag{6.1}$$

for $\alpha > 0$. The tuning parameter α is the complexity cost per terminal node. For each value of α there exists a subtree $T(\alpha)$ which minimizes equation (6.1). Calculating $C_\alpha(T)$ for all possible values of α results in a finite sequence of subtrees $T_{max}, T_{max-1}, ...$, with a decreasing number of terminal nodes. Ideal tree size can be determined by using independent test samples or by cross-validation (Hastie et al., 2001). The classification tree method has 'built-in' variable selection in that at each node, the variable that returns the purest children nodes are chosen to split the data at that node.

It's important to note that the trees discussed in this chapter are not the same as those discussed in the previous chapters. In the previous chapters the trees represented a possible evolutionary relationship between a given set of taxa and each terminal node corresponded to a taxon. In this chapter, the classification trees presented represent a classification of the sequences by predominant structural features, and each terminal node corresponds to a structural class rather than a taxon.

6.2 BOOTSTRAP AGGREGATING

The classification tree has the advantage of being interpretable with natural variable selection. However, classification trees can be unstable in that small changes in the training sample can result in significant changes in the classification rules which generate the tree. One way to mitigate this instability and achieve better results is to apply a 'bagging' or 'bootstrap aggregating' method (Brieman, 1996). Bagging reduces the variance in classification trees by taking an average over multiple trees. Repeated bootstrap samples are taken from the training set. A tree is grown on each training set and fit on the test set. A count is kept of how many trees classify a particular observation, say x, in any given class k. The observation x is assigned to the class with the maximum number of 'votes' (that class to which the majority of trees assign x). Bagging significantly reduces the prediction instability since taking an average will reduce the variance but leave bias unchanged (Hastie et al., 2001).

6.3 CLASSIFICATION TREES AND THE SPECTRAL ENVELOPE

Since the peaks in the spectral envelope correspond to the periodicity of a protein's structure, information extracted from the spectral envelope may be used to classify proteins into structural categories. The first step is to devise a method to extract the structural features of the proteins as defined by the frequencies at which peaks in the corresponding spectral envelopes occur. To isolate such frequencies the spectral envelopes are divided into 100 thin frequency bands of equal width. An example of a spectral envelope divided in twenty, fifty and one hundred frequency bands is shown in Figure 6.1. The number of frequency bands must be large enough to capture

all the information available, but small enough to ensure the data do not become too sparse. Previous analysis on a sample of the data indicated that 100 frequency bands was ideal. The 100 covariates for each protein are obtained by calculating the mean of the points above a threshold in each frequency band. The application of a threshold should remove noise from the data. The threshold is based on the empirical distribution of 1000 bootstrap samples taken from a concatenation of all the sequences in the data to which the method is being applied. The covariates are then used in the classification tree method. We would expect the variables selected in partitioning the data with the classification tree method to correspond to the signals at the frequencies where peaks corresponding to secondary structures in the proteins occur. The steps for extracting covariates from the spectral envelope and building a classification tree are listed below.

1. For each sequence compute the spectral envelope and sort the points of the spectral envelope into the appropriate frequency ranges.

2. Calculate the mean spectral envelope above the threshold within each frequency band and in this way obtain 100 covariates for use in the classification tree procedure. For the data used in this chapter a threshold of 0.02 is used. We compute 1000 bootstrap sample spectral envelopes and take the 50^{th} quantile at each frequency. The median across all frequencies is taken to get a single value of 0.02 of the threshold.

3. Randomly sample 75% of the proteins in each class to use as a training set. Grow a classification tree using this sample. Determine ideal tree size using 10-fold cross-validation via the cv.tree() function in R.

4. Prune the tree to the appropriate size using the prune.misclass() function in the R.

5. Use the pruned tree on the remaining 25% of observations to obtain an objective test error.

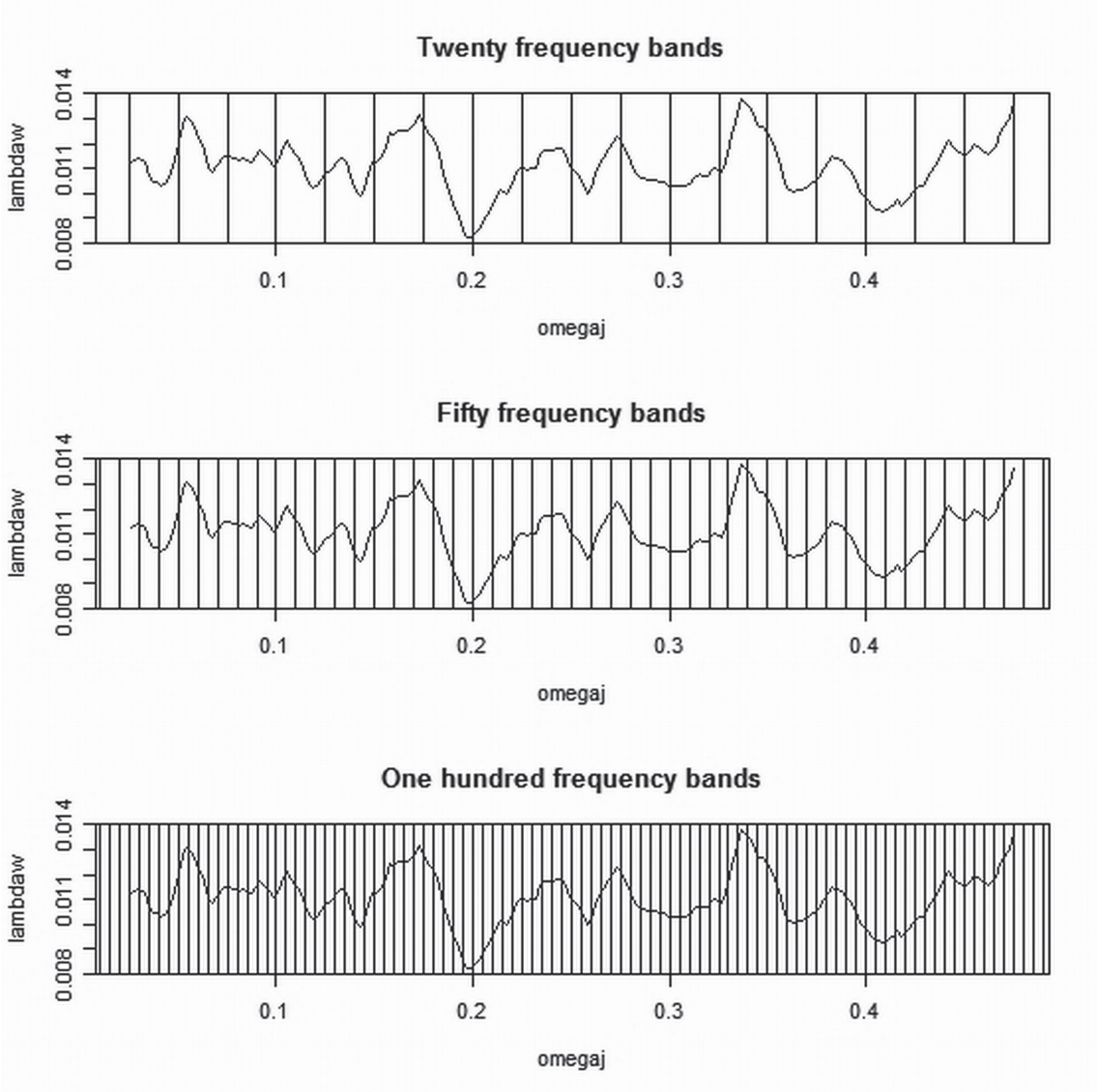

Figure 6.1: An example of a spectral envelope divided into twenty, fifty and one hundred frequency bands.

6.4 DATA

We begin by applying our method to a simple small example data set consisting of the myoglobin and immunoglobulin sequences for the five primate taxa studied in chapters 2, 3, and 4. Sequences were downloaded from the protein database in Genbank. Accession numbers can be found in the appendix.

The method was then applied to a larger dataset consisting of groups of proteins from the all-α and all-β classes in the SCOP database (Murzin et al., 1995; Andreeva et al., 2004; Murzin and Bateman, 1997). Murzin and Bateman (1997) employed a combination of homologous, structural and functional information to create the SCOP (Structural Classification of Proteins) database. Proteins in SCOP are classified on four hierarchical levels. Proteins of the same family are considered to have a clear evolutionary relationship with sequence similarity of 30% or greater (Andreeva et al., 2004). Proteins of the same superfamily have low sequence similarity but enough common structural and functional features such that common evolution is probable. Proteins classified in the same fold have the same major secondary structures with the same topological arrangement. At the class level, proteins are classified on the basis of the secondary structures of which they are composed.

We extracted a sample sequence from each superfamily in the all-α and all-β classes in the SCOP database release 1.73. Our only stipulation was that sequences have lengths greater than 50. In this manner we obtained a dataset comprising 435 sequences from the all-α class and 317 sequences from the all-β class. The sequences selected should have low sequence similarity while sharing common secondary structures with other sequences within the class to which they belong.

6.5 RESULTS

To get an idea what the spectral envelopes of all-α and all-β proteins might look like the spectral envelope was computed for the myoglobin and immunoglobulin protein sequences for the five primate taxa, gibbon, orangutan, gorilla, human and chimp. Figure 6.2 shows the mean spectral envelopes across the five primate taxa gibbon, orangutan, gorilla, human and chimp for myoglobin and immunoglobulin proteins. We can see that the mean spectral envelope of the myoglobin sequences has its greatest peak at around 0.22 to 0.28 while the mean spectral envelope of the immunoglobulin sequences has its greatest peak at around 0.40 to 0.45. Recall, that α-helices have a periodicity of 3.6 residues corresponding to a peak at $\omega = 0.277$, while β-sheets have a periodicity of 2.3 residues corresponding to a peak at $\omega = 0.435$, hence

the mean spectral envelopes of the myoglobin and immunoglobulin sequences display the behaviour we expect if the spectral envelope is indeed picking up the signal corresponding to the secondary structures in these proteins.

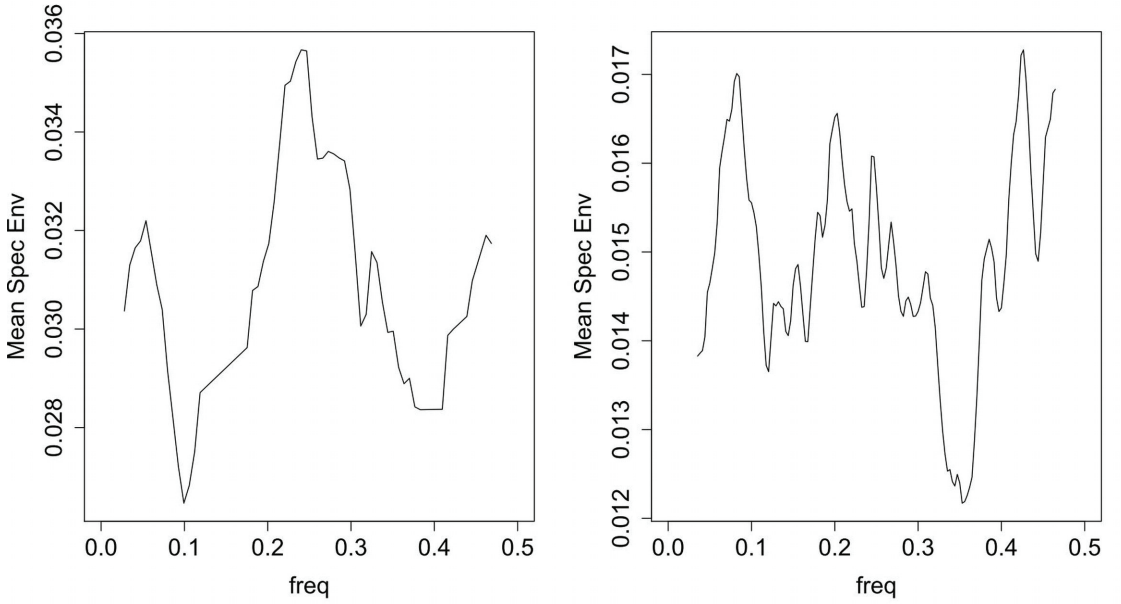

Figure 6.2: The mean spectral envelope across five primate taxa for the protein myoglobin (left) and the protein immunoglobulin (right).

Next our classification method was applied to this small example. On this small data set a tree with two pure terminal nodes (0 % misclassification) is obtained. The classification tree method chooses a node to partition by a variable and its optimum split point such that the decrease in the deviance is maximized at each step of the partitioning. Variables based on the spectral covariance are the 100 frequency bands and ranges. The variable, or frequency band used to split the data into two nodes includes frequencies between $\omega = 0.120$ and $\omega = 0.125$, which are within the range of frequencies at which we'd expect to see a peak corresponding to repetition of secondary structure elements. Note that this is a fairly easy problem as the myoglobin and immunoglobulin sequences have high within-group sequence similarity.

The method was then applied to the large data set extracted from the SCOP database. This is a more difficult problem as the data were selected in such a way that within-group sequence similarity will be small. To begin with, spectral envelope

based covariates were computed for the full data set and a classification tree was grown. The full tree grown on the complete data set consisted of 12 terminal nodes. Frequency bands used to grow the tree included $\omega = 0.08$ to $\omega = 0.18$ which includes the frequency range at which we'd expect to find a peak corresponding to repetition of secondary structure elements, $\omega = 0.415$ to $\omega = 0.440$ which contains $\omega = 0.435$, the frequency at which we'd expect to see a peak corresponding to β- strands, and finally $\omega = 0.245$ to $\omega = 0.290$, which contains $\omega = 0.277$, the frequency at which we'd expect to see a peak corresponding to α-helices. The misclassification error when the complete data set is fit back on the full grown tree is 0.366. Using 10-fold cross-validation, it was determined that the ideal number of terminal nodes was 9 terminal nodes and so the tree was pruned back to 9 terminal nodes. The misclassification error for the final pruned tree was 0.376, hence very little information was lost when the simpler tree was grown and this new tree should have greater predictive power for new observations. The 9-node classification tree is shown in Figure 6.3.

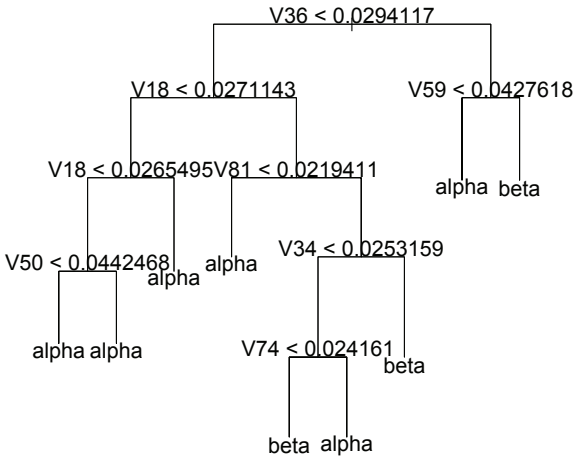

Figure 6.3: The pruned classification tree obtained from the total spectral envelope in 100 frequency ranges for 435 all-α proteins and 317 all-β proteins

To evaluate our protein classification method we compared it to the classification capabilities of BLAST at the class level of the SCOP hierarchy. To classify the sequences as either belonging to the all-α or all-β class, each sequence in our data set

was queried against all the remaining sequences in the data set and assigned the class to the sequence with the closest match. Classification was based on three different criteria: the highest bit score, which measures the strength of the alignment, lowest e-value, which measures the significance of the hit, and highest percentage sequence identity. Misclassification errors when each of these criteria were used are 0.475, 0.495 and 0.510, respectively. This implies that the spectral envelope based method of classification is picking up some additional information which we believe corresponds to structural signals within the proteins.

The data set was then split into two subsets. One subset was used as a training set and the other as a test set. A classification tree was trained on the training set and pruned to the 9-terminal nodes previously found to be ideal. The test set was then fit on the pruned tree. The training set contains 564 observations (approximately 75 % of the data), while the test set contains 188 observations (approximately 25% of the data). The sampling used to obtain the training and test sets was stratified. That is, a proportion of sequences were sampled from the all-α class and another from the all-β class. This was to ensure that the composition of the test set more or less corresponded to that of the training set. The training error presented is the proportion of misclassifications when the training data set is used for prediction by the pruned classification tree. The cross-validation error is obtained by taking the average misclassification error of the ten trees grown in the cross-validation procedure. That is, we grow ten tree with one tenth of the data left out each time and use this remaining 10 % for prediction. The cross-validation error is the average of misclassification errors obtained for these ten trees. It provides an estimate of the test error. The test error is the proportion of misclassifications when the test set is used for prediction by the pruned classification tree. When our method was applied to the split data a training error of 0.3121 was obtained, with a cross-validated error of 0.3546 and a test error of 0.3564. For this data, the cross-validated error gives a fairly good approximation of the test error. To compare these results with BLAST, each sequence in the test set was queried against the training set. Again, the same three criteria were used to classify sequences. Using the maximum bit score a misclassification error of 0.500 was obtained. With the minimum e-value a

misclassification error of 0.511 was obtained. Finally, using the highest percentage sequence identity resulted in a misclassification error of 0.511.

Next, the ipredbagg() function in the **ipred** package of the R programming language was applied to determine if the misclassification error could be reduced by applying a bagging procedure (Peters and Hothorn, 2009). The bagging classification trees obtained with 25 bootstrap replicates returned an out-of-bag estimate of the misclassification error of 0.2872. The out-of-bag estimate of error is obtained from observations left out of the bootstrap samples. It provides an estimate of the test error. When the test set was fit on the model obtained from the bagging classification procedure a test error of 0.2606 was obtained. Hence, through application of the bagging procedure we were able to reduce the misclassification error for this data.

The sequences in this data were chosen to have low sequence similarity. The relatively low misclassification errors obtained by the method when compared with BLAST suggest that the spectral envelope is picking up some structural information present in the sequences.

6.6 DISCUSSION

The results suggest that the spectral envelope is able to extract structural information directly from amino acid sequences. Analysis of a data set with low sequence similarity showed that our method was able to group proteins by structural class with a relatively low misclassification error when compared with BLAST which relies on sequence alignment. Note that BLAST relies heavily on sequence similarity and it would be preferable to compare our method to a non-homology based prediction methods such as HMMs. There exist several webservers capable of predicting secondary structure input sequences such as Jpred, PSIpred and HHpred which use combinations of PSI-blast and profile hidden Markov models to predict secondary structure. However these servers require that the input sequences be aligned and the user select a data base to compare input sequences to from a given list. We could not find any that allowed for comparisons on a user defined data base so that we could find training and testing errors to compare to our method as we were able to

with BLAST. The webservers were also extremely slow. Due to time constraints and the limitations of the tools currently available we did not compare our method to a secondary structure prediction technique but note that a comparison of the spectral envelope based method with a secondary structure prediction method should be explored in the future.

The comparison of sequences using the spectral envelope does not require alignment. Recall that we'd expect peaks in the spectral envelope to occur at certain frequencies corresponding to the periodicities of different protein secondary structures. The correspondence of the frequencies of peaks in spectral envelope and the frequency ranges used to split the tree indicate that the structural features of proteins are indeed being captured by the spectra of their amino acid sequences. Our application of the classification tree method has shown that the spectral envelope can provide useful covariates in protein structure prediction. The classification tree has the advantage of being interpretable with natural variable selection. However alternative procedures for variable selection and classification may be explored to improve results. Since protein function is partially determined by protein structure, combining our procedure with biochemical information, may again improve prediction results.

Chapter 7

CONCLUSION

In this chapter we summarize the results presented in this thesis and outline some possible directions for future work.

7.1 SUMMARY

In chapter 2 we extended the spectral covariance method previously introduced by Collins et al. (2006) to include a taxon-specific scaling in which each taxon is assigned a scaling which is held constant across all pairwise comparisons. Applying both the common scaling and taxa specific scaling methods to various data sets we found these yielded similar results. Unlike the ML based methods, spectral covariance based dissimilarity measures do not require the assumption that sites evolve independently based on an evolutionary model. The spectral covariance method measures structural similarity as well as sequence similarity using time series methodology. Because 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 results obtained with the nematode data set suggest this is the case

In chapter 3 we introduced two criteria for computing scale coefficients which can then be used to combine information across genes, namely the minimum variance (MinVar) criterion and the minimum coefficient of variation squared (MinCV) criterion. The scale coefficients obtained with the MinVar and MinCV criteria can then be used to derive a combined-gene tree from the weighted average of the distance or dissimilarity matrices of multiple genes. The MinVar and MinCV methods gave similar results. In chapter 4 we introduce an alternative method for deriving a combined gene tree based on singular value decomposition. We showed how the first right eigenvector of the singular value decomposition of a matrix of distance

vectors of multiple genes could be used to obtain a combined-gene tree. Unlike standard methods applied to multi-gene analysis, our method does not assume that genes share a common evolutionary history or rate which can sometimes result in the wrong tree. With our method we were able to separate the nematode and honeybee in the data set first studied by Foster and Hickey (1999) who applied various methods to a concatenation of genes and found the nematode and honeybee consistently grouped together with strong bootstrap support.

In chapter 5, we derived influence functions for the components of a singular value decomposition and used these to determine genes influential in the combined gene topology inferred from the singular value decomposition method. The robustness of our method was evaluated under perturbations in the distances and individual and combined removal of genes. We found that the combined-gene trees obtained with the singular value decomposition method where fairly robust to perturbations in distances and removal of genes.

Finally, in chapter 6, we propose a method for classifying proteins by their predominant structural features using covariates extracted from the spectral envelope. We compared our method to the classification capabilities of blast which is readily available and able to handle a large number of sequences in a relatively short amount of time. We found that our method was relatively successful at classifying proteins by their predominant structural features.

The data sets analysed in this thesis had various sizes, the chloroplast data set having the greatest number of taxa (22) and the eukaryote data set having the greatest number of genes (35). Computing the common spectral covariance for an entire data set only took a few minutes. The taxa specific scaling covariance is a bit more computationally expensive and can take up to half an hour to compute for a large data set. Overall, computation of combined-gene trees using these methods was relatively quick. All our computations were done in the R programming language.

The methods presented in this thesis are unique in that they require no assumptions regarding the evolutionary model or history. The methods may be used on their own or as complementary to existing methods. A comparison of the combined-gene trees obtained with the MinCV or singular value decomposition based methods to

standard concatenation or consensus tree approaches may provide new insights into the evolutionary relationships between taxa. The influence functions for the singular value decomposition method provides us with a tool to better understand the effect of individual genes on the combined-gene topology.

7.2 FUTURE WORK

The impact of our method on systematic biases, such as variable evolutionary rates across genes, taxa or individual sites within a sequence, is unclear. Further simulation studies to rigorously test how our method responds under these conditions are required. An extension of these methods to deal with missing data and allow for the inclusion of a larger number of protein sequences should be developed. One way to do this is by modelling the pairwise distances computed from the available pairs of genes and using missing data imputation methods based on the statistical models.

With regards to influence analysis, other perturbation schemes may be explored to further investigate the influence of genes and taxa pairs on the combined-gene tree obtained from the singular value decomposition method. In particular, an additive perturbation method may be applied to further analyse which taxa pairs are most influential on the combined gene tree topology.

Appendix A

DATA: GENBANK ACCESSION NUMBERS

Table A.1: Nematode data set: taxa names, gene names and Genbank accession numbers

NEMATODE DATA	
Taxa	**Accession numbers**
Allomyces macrogynus	NC_001715
Apis mellifera ligustica (honeybee)	NC_001566
Artemia franciscana (brine shrimp)	NC_001620
Caenorhabditis elegans (roundworm)	NC_001328
Drosophila yakuba (fruitfly)	NC_001322
Gallus gallus (chicken)	NC_001323
Locusta migratoria (locust)	NC_001712
Paracentrotus lividus (sea urchin)	NC_001572
Genes: *ATP6, COX1, COX2, COX3, CYTB, ND1, ND2, ND3, ND4, ND4L, ND5, ND6*	

Table A.2: Chloroplast data set: taxa names, gene names and Genbank accession numbers

CHLOROPLAST DATA	
Taxa	**Accession numbers**
Acorus americanus (Acorus)	NC_010093
Adiantum capillus-veneris (Adiantum)	NC_004766
Amborella trichopoda (Amborella)	NC_005086
Anthoceros formosae (Anthoceros)	NC_004543
Arabidopsis thaliana (Arabidopsis)	NC_000932
Atropa belladonna (Atropa)	NC_004561
Calycanthus floridus (Calycanthus)	NC_004993
Chaetosphaeridium globosum (Chaetosphaeridium)	NC_004115
Chlorella vulgaris (Chlorella)	NC_001865
Lotus corniculatus (Lotus)	NC_002694
Marchantia polymorpha (Marchantia)	NC_001319
Mesostigma viride (Mesostigma)	NC_002186
Nephroselmis olivacea (Nephroselmis)	NC_000927
Nymphaea alba (Nymphaea)	NC_006050
Oryza sativa (Oryza)	NC_001320
Physcomitrella patens (Physcomitrella)	NC_005087
Pinus koraiensis (Pinus koraiensis)	NC_004677
Pinus thunbergii (Pinus thunbergii)	NC_001631
Psilotum nudum (Psilotum)	NC_003386
Spinacia oleracea (Spinacia)	NC_002202
Triticum aestivum (Triticum)	NC_002762
Zea mays (Zea_mays)	NC_001666
Genes: *atpA, atpB, atpE,* ***atpI, clpP****, petA, petB, petD,* ***psaB, psaC, rbcL****, rpl14, rpl16, rpl20, rpl2, rps11,* ***rpoC1****, rps12, rps14, rps18, rps19, rps2, rps3, rps7, rps8*	

Table A.3: Primate data set: taxa names, gene names and Genbank accession numbers

PRIMATE DATA	
Taxa	**Accession numbers**
homo sapiens (human)	NC_012920
pan troglodytes (chimpanzee)	NC_001643
gorilla gorilla (gorilla)	NC_001645
pongo pygmaeus (orangutan)	NC_001646
hylobates agilis (gibbon)	NC_014042
***Genes**: ATP6, ATP8, COX1, COX2, COX3,CYTB, ND1, ND2, ND3, ND4L, ND5, ND6*	

Table A.4: Primate myoglobin and immunoglobulin sequences: taxa names and accession numbers

Primate myoglobin and immunoglobulin sequences		
Taxa	**Accession numbers**	
	Myoglobin (SwissProt)	**Immunglobulin (EMBL)**
homo sapiens (human)	P02144	AAB59396
pan trogodytes (chimpanzee)	P02147	CAA37744
gorilla gorilla (gorilla)	P62734	CAA37743
pongo pygmaeus (orangutan)	P02145	CAA37746
hylobates agilis (gibbon)	P02148	CAA37741

Bibliography

A. M. Aguinaldo, J. M. Turbeville, L. S. Linford, M. C. Rivera, J. R. Garey, R. A. Raff, and J. A. Lake. Evidence for a clade of nematodes, arthropods and other moulting animals. *Nature*, 387(6632):489–493, 1997.

A. Andreeva, D. Howorth, S.E. Brenner, T.J. Hubbard, C. Chothia, and A.G. Murzin. Scop database in 2004: refinements integrate structure and sequence family data. *Nulceic Acids Research*, 32:226–229, 2004.

C Ané, J. G. Burleigh, M. M. McMahon, and M. J. Sanderson. Covarion structure in plastid genome evolution: a new statistical test. *Molecular Biology and Evolution*, 22:914–925, 2004.

T. J. Barkman, G. Chenery, J. R. McNeal, J. Lyons-Weiler, W. J. Ellisens, G. Moore, A. D. Wolfe, and C. W. DePamphilis. Independent and combined analyses of sequences from all three genomic compartments converge on the root of flowering plant phylogeny. *Proceedings of the National Academy of Sciences of the United States of America*, 97(24):13166–13171, 2000.

B.R. Baum. Combining trees as a way of combining data sets for phylogenetic inference, and the desirability of combining gene trees. *Taxon*, 41(1):3–10, 1992.

R. B. Beven, F. Lang, and D. Bryant. Calculating the evolutionary rate of different genes: a fast, accurate estimator with applications to maximum likelihood phylogenetic analysis. *Systematic Biology*, 54(6):900–915, December 2005.

O.R.P. Bininda-Emonds. The evolution of supertrees. *Trends in Ecology & Evolution*, 19(5):315–322, 2004.

J. E. Blair, K. Ikeo, T. Gojobori, and S. B. Hedges. The evolutionary position of nematodes. *BMC Evol Biol*, 7(2), 2002.

P. Bradley, K. M. S. Misura, and D. Baker. Toward high-resolution de nova structure prediction for small proteins. *Science*, 309:1868–1871, September 2006.

L. Brieman. Bagging predictiors. *Machine Learning*, 24(2):123–140, 1996.

D. G. Brown and M. M. Flocco. Structure determination - crystollography for structure-based drug discovery. In R. E. Hubbard, editor, *Structure-based drug discovery: An overview*, RSC Biomolecular sciences, chapter 2, pages 48–49. RSC, 2006.

J. J. Bull, J.P. Huelsenbeck, C.W. Cunningham, and D.L. Swofford. Partitioning and combining data in phylogenetic analysis. *Systematic Biology*, 42(3):384–397, 1993.

M. Bulmer. Use of the method of generalized least squares in reconstructing phylogenies from sequence data. *Molecular Biology and Evolution*, 8(6):868–883, 1991.

J.G. Burleigh, A.C. Driskell, and M.J. Sanderson. Supertree bootstrapping methods for assessing phylogenetic variation among genes in genome-scale data sets. *Systematic Biology*, 55(3):426–440, 2006.

J. M. Chandonia and S. E. Brenner. The impact of structural genomics: expectations and outcomes. *Science*, 311:347–351, January 2006.

N. V. Chandrasekharan and D. L. Simmons. The cyclooxygenases. *Genome Biology*, 5(9):241, 2004.

N. V. Chandrasekharan, H. Dai, K. L. T Roos, N. K. Evanson, J. Tomsik, and T. S. Elton. Cox-3, a cyclooxygenase-1 variant inhibited by acetaminophen and other analgesic/antipyretic drugs: Cloning, structure, and expression. *PNAS*, 99(21): 13926–13931, 2002.

C. Chothia and A. M. Lesk. The relation between the divergence of sequence and structure in proteins. *The EMBO Journal*, 5(4):823–826, 1986.

K. Collins, H. Gu, and C. Field. Examining protein structure and similarities by spectral analysis. *Statistical Applications in Genetics and Molecular Biology*, 2006.

A. Cruscuolo, V. Berry, E.J.P. Douzery, and O. Gascuel. SDM: A fast distance-based approach for (super)tree building in phylogenomics. *Systematic Biology*, 55(5): 750–755, 2006.

R. Das, B. Qian, S. Raman, R. Vernon, J. Thompson, P. Bradley, S. Khare, M. D. Tyka, D. Bhat, D. Chivian, D. E. Kim, W. H. Sheffler, L. Malmstrom, A. M. Wollacott, C. Wang, I. Andre, and D. Baker. Structure prediction for CASP7 targets using extensive all-atom refinement with rosetta@home. *Proteins*, 69(S8): 118–128, September 2007.

A. de Queiroz. For consensus (sometimes). *Systematic Biology*, 42(3):368–372, 1993.

A. de Queiroz and J. Gatesy. The supermatrix approach to systematics. *Trends in Ecology & Evolution*, 22(1):34–41, 2007.

H. Dopazo and J. Dopazo. Genome-scale evidence of the nematode-arthropod clade. *Genome Biology*, 6(5):R41+, 2005.

D. Eisenberg, R. Weiss, and T. Terwilliger. The hydrophobic moment detects periodicity in protein hydrophobicity. *Proceedings of the National Academy of Science*, 81(1):140–144, 1984.

J.S. Farris, M. Kallersjo, A. G. Kluge, and C. Bult. Testing significance of incongruence. *Cladistics*, 10:315–319, 1995.

J. Felsenstein. Phylip phylogeny inference package (version 3.2). *Cladistics*, 5:164–166, 1989.

J. Felsenstein. *Inferring phylogenies*. Sinauer Associates, 2003.

W.M Fitch and E. Margoliash. Construction of phylogenetic trees. *Science*, 155 (3760):279–284, 1967.

P.G. Foster and D.A. Hickey. Compositional bias may affect both DNA-based and protein-based phylogenetic reconstructions. *Journal of Molecular Evolution*, 48(3): 284–290, 1999.

W.K Fung, H. Gu, L. Xiang, and K.K.W. Yau. Assessing local influence in principal component analysis with application to haematology study data. *Statistics in medicine*, 26:2730–2744, November 2007.

O. Gascuel. BIONJ: an improved version of the NJ algorithm based on a simple model of sequence data. *Molecular Biology and Evolution*, 14(7):685–695, 1997.

J. Gatesy, R.H Baker, and C. Hayashi. Inconsistencies in arguments for the supertree approach: Supermatrices versus supertrees of crocodylia. *Systematic Biology*, 53 (2):342–355, 2004.

K. Ginalski, N.V. Grishin, A. Godzik, and L. Rychlewski. Survey and summary: practical lessons from protein structure prediction. *Nucleic Acids Research*, 33(6): 1874–1891, 2005.

V. V. Goremykin and F. H Hellwig. A new test of phylogenetic model fitness addresses the issu of the basal angiosperm phylogeny. *Gene*, 381, 2006.

V.V Goremykin, K. I. Hirsch-Ernst, S. Wolfl, and F. H. Hellwig. Analysis of the *amborella trichopoda* chloroplast genome sequence suggests that *amborella* is not a basal angiosperm. *Molecular Biology and Evolution*, 20(9):1499–1505, 2003.

V.V. Goremykin, B. Holland, K. I. Hirsch-Ernst, and F. H. Hellwig. Analysis of *acorus calamus* chloroplast genome and its phylogenetic implications. *Molecular Biology and Evolution*, 22(9):1813–1822, 2005.

N. Gruenheit, P.L. Lockhart, M. Steel, and W. Martin. Difficulties in testing for covarion-like properties of sequences under the confounding influence of changing proportions of variable sites. *Molecular Biology and Evolution*, 25:1512–1520, 2008.

T. A. Hall. Bioedit: a user-friendly biological sequence alignment editor and analysis program for windows 95/98/nt. *Nucleic Acids Symposium Series*, 41:95–98, 1999.

T. Hastie, R. Tibshirani, and J. Friedman. *The elements of statistical learning*. Springer, 2001.

M.D. Hendy and D Penny. Spectral analysis of phylogenetic data. *Journal of Classification*, 10, 1993.

Asger Hobolth, Ole F. Christensen, Thomas Mailund, and Mikkel H. Schierup. Genomic relationships and speciation times of human, chimpanzee, and gorilla inferred from a coalescent hidden markov model. *PLoS Genetics*, 3(2):e7+, 2007.

R. E. Hubbard. 3D structure and the drug discovery process. In R. E. Hubbard, editor, *Structure-based drug discovery: An overview*, RSC Biomolecular sciences, chapter 1, pages 1–31. RSC, 2006.

O. Jeffroy, H. Brinkmann, F. Delsuc, and H. Philippe. Phylogenomics: the beginning of incongruence? *Trends in Genetics*, 22:225–231, 2006.

P. Keeling, B. S. Leander, and A. Simpson. Eukaryotes. eukaryota, organisms with nucleated cells. http://tolweb.org/Eukaryotes/3/2009.10.28 in The Tree of Life Web Project, http://tolweb.org/, 2009.

M.K Kuhner and J. Felsenstein. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. *Molecular Biology and Evolution*, 11: 459–468, 1994.

H. Kunsch. The jacknife and bootsrap for general stationary observations. *Annals of Statistics*, 17:1217–1241, 1989.

G. Lecointre. Total evidence requires exclusion of phylogenetically misleading data. *Zoologica Scripta*, 34:101–117, 2005.

J. Leigh, E. Susko, M. Baumgartner, and A.J. Roger. Testing congruence in phylogenomic analysis. *Systematic Biology*, 57(1):104–115, 2008.

M.M. Miyamoto and W.M. Fitch. Testing species phylogenies and phylogenetic methods with congruence. *Systematic Biology*, 44(1):64–76, 1995.

R. K. Murray, D. K. Granner, P. A. Mayes, and V. W. Rodwell. *Harper's Illustrated Biochemistry.* McGraw-Hill, 26 edition, 2003.

A. Murzin and A. Bateman. Distant homology recognition using structural classification of proteins. *Proteins*, Supplement 1:105–112, 1997.

A. G. Murzin, S. E. Brenner, and Chothia C. Hubbard, T. Scop: A structural classification protein database for the investigation of sequences and structures. *Journal of Molecular Biology*, 247:536–540, 1995.

Andrea Peters and Torsten Hothorn. *ipred: Improved Predictors*, 2009. URL `http://CRAN.R-project.org/package=ipred`. R package version 0.8-8.

M. E. Peterson, F. Chen, J. G. Saven, D. S. Roos, P. C. Babbitt, and A. Sali. Evolutionary constraints on structural similarity in orthologs and paralogs. *Protein Science*, 18(6):1306–1315, 2009.

H. Philippe, H. Brinkmann, and N. Lartillot. Multigene analyses of bilaterian animals corroborate the monophyly of ecdysozoa, lophotrochozoa, and protostomia. *Molecular Biology and evolution*, 36(22):1246–1253, 2005a.

H. Philippe, F. Delsuc, H. Brinkmann, and N. Lartillot. Phylogenomics. *Annual Review of Ecology, Evolution, and Systematics*, 36(1):541–562, 2005b.

R. Piaggio-Talice, G. Burleigh, and O. Eulenstein. Phylogenetic supertrees: Combining information to reveal the tree of life. In Olaf R.P Bininda-Emonds, editor, *Phylogenetic Supertrees: Combining Information to Reveal the Tree of Life*, pages 173–191. Kluwer Academic, Dordrecht, the Netherlands, 2004.

J. Pinheiro, D. Bates, S. DebRoy, D. Sarkar, and the R Core team. *nlme: Linear and Nonlinear Mixed Effects Models*, 2009. R package version 3.1-93.

M.B. Priestly. *Spectral Analysis and Time series*. Academic Press, 1981.

Y. Qiu, J. Lee, F. Bernasconi-Quadroni, D. E. Soltis, P. S. Soltis, M. Zanis, E. A. Zimmer, Z. Chen, V. Savolainen, and M. W. Chase. The earliest angiosperms: evidience from mitochondrial, plastid and nuclear genomes. *Nature*, 402:404–407, 1999.

A. Rambaut and N. C. Grassly. Seq-gen: An application for the monte carlo simulation of DNA sequence evolution along phylogenetic trees. *Computer Applications Biosciences*, (13):235–238, 1997.

A. Rhzetsky and Masatoshi Nei. Statistical properties of the ordinary least-squares, generalized least-squares, and minimum-evolution methods of phylogenetic inference. *Journal of Molecular Evolution*, 35:367–375, 1992.

Brian Ripley. *tree: Classification and regression trees*, 2010. URL `http://CRAN.R-project.org/package=tree`. R package version 1.0-28.

I. B. Rogozin, Y. I. Wolf, L. Carmel, and E. V. Koonin. Ecdysozoan Clade Rejected by Genome-Wide Analysis of Rare Amino Acid Replacements. *Mol Biol Evol*, 24 (4):1080–1090, 2007.

A. Rokas, B.L. Williams, N. King, and S.B. Carroll. Genome-scale approaches to resolving incongruence in molecular phylogenies. *Nature*, 425(6960):798–804, 2003.

M. Ruvolo. Molecular phylogeny of the hominoids: Inferences from multiple independent DNA sequence data sets. *Molecular Biology and Evolution*, 14:248–265, 1997.

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

D. E. Soltis and P. S. Soltis. *Amborella* not a 'basal angiosperm'? not so fast. *American Journal of Botany*, 91(6):997–1001, 2004.

P. Soltis, D. Soltis, and C. Edwards. Angiosperms. flowering plants. http://tolweb.org/Angiosperms/20646/2005.06.03 in The Tree of Life Web Project, http://tolweb.org/, 2005.

P. S. Soltis, D. E. Soltis, and M. W. Chase. Angiosperm phylogeny inferred from multiple genes as a tool for comparative biology. *Nature*, 402:402–404, 1999.

S. Stefanovic, D. W. Rice, and J. D. Palmer. Long branch attraction, taxon sampling, and the earliest angiosperms: *Amborella* or monocots? *BMC Evolutionary Biology*, 4:35, 2004.

D. Stoffer, D. Tyler, and A. McDougall. Spectral analysis for categorical time series: scaling and the spectral envelope. *Biometrika*, 85(1):201–213, 1993.

D. Stoffer, D. Tyler, and D. Wendt. The spectral envelope and its applications. *Statistical Science*, 15(3):224–253, 2000.

G.W Stuart, K. Moffett, and J.L Leader. A comprehensive vertebrate phylogeny using vector representations of protein sequences from whole genomes. *Molecular Biology and Evolution*, 19(4):554–562, 2002.

J. A. Studier and K. J Keppler. A note on the neighbor-joining algorithm of Saitou and Nei. *Molecular Biology and Evolution*, 5(6):729–731, 1988.

M. J. Telford. Animal phylogeny: Back to the coelomata? *Current Biology*, 14(7): 274 – 276, 2004.

A. Vandamme. *The Phylogenetic Handbook*. Cambridge, 2 edition, 2009.

T. C. Wood and W. R. Pearson. Evolution of protein sequences and structures. *Journal of Molecular Biology*, 291(4):977–995, 1999.

J. C. Wooley and Y. Ye. A historical perspective and overview of protein structure prediction. In Y. Xu, Dong Xu, and Jie Liang, editors, *Computational methods for protein structure prediction and modeling*, volume 1, chapter 1, pages 1–43. Springer, New York, 2007.

J. Wu and E. Susko. General heterotachy and distance method adjustments. *Molecular Biology and Evolution*, 26(12):2689–2697, 2009.

M. Zanis, D. E. Soltis, P. S. Soltis, S. Mathews, and M. J. Donoghue. The root of the angiosperms revisited. *Proceedings of the National Academy of Sciences*, 99(10): 6848–6853, 2002.

M. Zelwer and V. Daubin. Detecting phylogenetic incongruence using BIONJ: an improvement of the ILD test. *Molecular Phylogenetics and Evolution*, 33(3):687 – 693, 2004.

Y. Zhang. Progress and challenges in protein structure prediction. *Current opinion in structural biology*, 18(3):342–348, May 2008.