THEORETICAL AND PRACTICAL UTILITY OF GENE SEQUENCES IN PHYLOGENETIC AND PHYLOGEOGRAPHIC ANALYSIS

by

ROBERT AARON MAKOWSKY

Presented to the Faculty of the Graduate School of

The University of Texas at Arlington in Partial Fulfillment

of the Requirements

for the Degree of

DOCTOR OF PHILOSOPHY

THE UNIVERSITY OF TEXAS AT ARLINGTON

DECEMBER 2009

ACKNOWLEDGMENTS

First and foremost, I must thank my wife, Cynthia, for being incredibly supportive during my schooling. Without her and my family, I would not have had the support I needed to complete my degree. I must also thank my advisor, Paul Chippindale, for all of his help over the last 3 years as well as for rescuing me during a difficult period in my life. The rest of my committee, Esther Betran, Jeff Demuth, James Grover, and Doyle Hawkins have all been as helpful as could be imagined during my tenure. I also would like to thank Donald Shure, Larry Wilson, and Thomas Pauley for providing the education and training I needed to complete my dissertation.

Many people have helped me out appreciably during my stay and I would like to explicitly thank the biology department, Jonathan Campbell, Peggy Fisher, Gloria Burlingham, Linda Taylor, Belinda Zollotuchen, Amy Carrillo, and Dan Formanowicz. I would also like to thank my unofficial committee members; Walter Schargel, Jesse Meik, and Brian Fontenot for many words of advice over the years.

For chapter one, I must thank Assiatu Barrie, Christian Cox, and Walter Schargel for helpful comments. For chapter two, I would like to thank Christian Cox, Corey Roelke, Jesse Meik, Brian Fontenot, Walter Schargel, Esther Betran, Charles Tracy, John Pace, Clement Gilbert, Jeff Streicher, Phi Sigma Biolunch participants, and two anonymous reviewers.

For chapter three,I would like to thank John Marshall, John McVay, Leslie Rissler, Wallace Holznagel, Bruce Kingsbury and IPFW, Stan Trauth and his students, Michael Sandel, Justin Bagley, Jessica Wooten, Paul Moler, Matt Brandley, Christian Cox, Michael Welker, Corey Roelke, Alvin Braswell, Matt Nordgren, Carl Franklin, Matt Watson, and Scott Boback. Whit Gibbons and his students were very helpful at SREL. The Museum of Vertebrate Zoology at Berkeley, the North Carolina State Museum, the University of Kansas Natural History Museum, and the Louisiana Museum of Natural History at Louisiana State University all provided invaluable specimens. Funding was graciously provided by the Birmingham Audubon Society (Walter Cox Grant) and the Beta Phi Chapter of Phi Sigma at UTA.

August 14, 2009

ABSTRACT

THEORETICAL AND PRACTICAL UTILITY OF GENE SEQUENCES IN PHYLOGENETIC AND PHYLOGEOGRAPHIC ANALYSIS

Robert Aaron Makowsky, Ph.D.

The University of Texas at Arlington, 2009

Supervising Professor: Paul T. Chippindale

Phylogenetics, or the study of evolutionary relationships among organisms, is a rapidly changing field due primarily to the dramatic increase in available molecular characters and increasingly sophisticated theoretical and computational methods. Current phylogenetic methods, though, poorly handle such large datasets due to the extremely large number of calculations required. In this dissertation, I focus on a method that can reduce datasets with a large number of molecular characters and at the same time optimize the performance of phylogenetic methods.

Chapter 1 provides background information about phylogenetic methods, with a specific emphasis on Bayesian phylogenetic methods. MrBayes is the most common program used for Bayesian phylogenetic analyses and has promise with larger datasets

due to its ability to partition datasets and easily utilize parallel processing techniques. Therefore, I focus on the methods implemented in MrBayes. Specifically, I discuss some of the aspects associated with search parameters, the utilization of Metropolis Coupled Markov Chain Monte Carlo analyses, and well as the calculation of Bayes factors.

Chapter 2 focuses on determining the appropriate genes for phylogeny reconstruction, which can be a difficult process. Rapidly evolving genes tend to perform best for resolving of recent relationships, but suffer from alignment issues and increased homoplasy (e.g., sequence saturation) among distantly related species. Conversely, slowly evolving genes generally perform best for deeper relationships, but lack sufficient variation to resolve recent relationships. We determine the relationship between sequence divergence and Bayesian phylogenetic reconstruction ability using both natural and simulated datasets. The natural data are based on 28 widely accepted (based on multiple independent sources) relationships within the subphylum Vertebrata. Sequences of 12 genes were downloaded from Genbank and Bayesian analyses were used to determine phylogenetic support for correct relationships. Simulated datasets were designed to determine whether an optimal range of sequence divergence exists across extreme phylogenetic conditions. Across all genes we found that an optimal range of divergence for resolving the correct relationships does exist, although this level of divergence expectedly depends on the distance metric. Simulated datasets show that an optimal range of sequence divergence exists across diverse topologies and models of evolution. We determine that a simple to measure property of nucleotide sequences (genetic distance) is related to phylogenic reconstruction ability in Bayesian analyses.

This information should be useful for selecting the most informative gene(s) to resolve a wide range of relationships, especially those that are difficult to resolve, as well as minimizing both cost and confounding information during project design.

In Chapter 3, the findings in Chapter 2 were taken into account when deciding what genes to include in the analysis. This chapter is a detailed analysis of the evolutionary history of the plain-bellied watersnake, *Nerodia erythrogaster*. Here, I sought to determine if the currently defined subspecies in the plain-bellied watersnake are concordant with results based on relatively neutral genetic markers. Species with morphological varieties (such as the plain-bellied watersnake) that are subdivided geographically have often been divided into subspecies. The morphological pattern, though, may not be congruent with the organism's evolutionary history (i.e. genetic drift, environmentally determined instead of selection). I choose this species because it occurs across multiple biogeographic barriers (Mississippi River, Apalachicola River) and contains multiple subspecies. My goals are to 1) provide a rigorous genetic analysis of *N. erythrogaster* throughout its range and determine what, if any, genetic lineages can be identified using mitochondrial DNA; 2) test whether monophyletic lineages are concordant with the current taxonomy or probable biogeographic barriers (Mississippi and Apalachicola River); and 3) assess the degree of ecological niche differentiation among lineages. To identify evolutionary lineages, we sequenced three genes (NADH II, Cyt-*b*, Cox I) from 156 geo-referenced specimens. Ecological niches were defined using bioclimatic layers for the five recovered genetic lineages, only one of which is concordant with a currently recognized subspecies (*N. e. erythrogaster*) and

biogeographic barrier (Apalachicola River). The recovered phylogeny is weakly supported overall, although some major genetic lineages exist. All previous taxonomic and biogeographic hypotheses are strongly disfavored compared to the best tree and ecological separation among lineages is minimal. Overall, we found no genetic support for the subspecies based on geography and conclude while some genetic and niche differentiation is evident, it is not enough to warrant taxonomic changes.

TABLE OF CONTENTS

LIST OF ILLUSTRATIONS

Figure Page

LIST OF TABLES

Table Page

CHAPTER 1

OVERVIEW OF BAYESIAN PHYLOGENETICS

A desire to understand the evolutionary history of organisms has existed ever since evolution by common decent was first discussed. In Darwin's seminal work (1859), the lone figure was a hand drawn evolutionary tree, illustrating both the importance of the subject as well as the clear "story" presented in such an image. Determining the evolutionary history of species is an incredibly difficult task, though, and the field of phylogenetics is undergoing constant changes and developments. Some changes involve new ideas, while others are ideas that have been around for years and are implementable now that computers are capable of handling the required intensive algorithms. In this introduction, I will present a brief overview of phylogenetics and the theory behind different methods while providing a more detailed analysis of Bayesian methods in phylogenetics.

While evolutionary relationships among organisms have been proposed for over a hundred years, it was not until Sokal, Sneath and colleagues (Michener and Sokal, 1957; Sneath, 1957a, b) that rigorous criteria were applied to the process of creating phylogenies. Working with some of the first available computers, they applied numerical clustering algorithms for classification and phylogenetic purposes. In later works, they dropped the phylogenetic interpretation (Sokal and Sneath, 1963) and espoused the numerical classification utility, but they were instrumental to the field of phylogenetics. Since the seminal work, numerous new methodologies have been developed including parsimony (and its many variants), maximum likelihood, and Bayesian phylogenetic analyses. Parsimony was first proposed by Hennig (1950) and Edwards and Cavalli-Sforza (1963; 1964). In one of the same publications, Edwards and Cavalli-Sforza (1964) also proposed maximum likelihood, although the method would be refined and only become possible 30 years later. The application of Bayesian methods to phylogenetics was first mentioned by Gomberg (1968) but, like likelihood methods, required intensive calculations and therefore computers to implement the algorithms.

Phylogenetic methods reconstruct the evolutionary history of organisms by comparing homologous characters (i.e. characters that are shared by common descent). Homology of characters is an important assumption, but one that can never be truly known. These characters can take any form, ranging from morphological to behavioral to genetic. Morphological characters, due to their easy attainability, dominated early analyses. Molecular characters (e.g., nucleotides, amino acids, restriction sites, karyotypes, and other genome characteristics) have now become the standard characters in most phylogenetic analyses. Of the molecular characters, sequence data are the most ubiquitous due to their relative ease of procurement, cost, and application in phylogenetic methods. Through the use of polymerase chain reaction (PCR), homologous regions of the genome can be attained and compared for many species often quite easily.

All methods require an aligned matrix based on assumed homology. An example is shown in Table 1.1 with four taxa and 10 character sites.

Table 1.1 Hypothetical sequence alignment of four taxa

	1	2	3	4	5	6	7	8	9	10
Human	A	T	G	G	C	T	G	T	A	A
Mouse	A	A	G	G	C	C	G	T	T	A
Alligator	C	G	G	C	G	A	G	T	G	G
Fish	C	C	C	C	G	A	T	T	G	G

In this example, one site's characters are conserved across taxa (site 8), two Singleton sites occur where the character is conserved except for one individual (sites 3 and 7), and four sites are parsimony informative (sites 1, 4, 5, and 10). All sites except site 8 are variable sites. The alignment is an incredibly important but, due to the automation of alignment programs, often underappreciated step in phylogenetic analyses (Höhl and Ragan, 2007; Landan and Graur, 2007; Ogden and Rosenberg, 2007). For sequence data, numerous programs are available to compute the best pairwise and multiple alignments. Difficulties in assessing homology arise when, for example, insertions or deletions have occurred in some of the lineages. Strategies for overcoming such situations include deleting regions of ambiguity, adjusting alignment parameters, or incorporating amino acid (for protein coding genes) or tertiary structure (for ribosomal sequences) information. The best strategy generally depends on the proportion of ambiguous sites, what the sequence codes for, and the amount of prior information available.

Once an alignment is obtained, phylogenetic methods are used to hypothesize a phylogeny, though the currently available methods differ both theoretically and in their search algorithms. Parsimony methods (MP) attempt to recover the phylogeny that minimizes the total number of state changes (Felsenstein, 1983). Attempt is appropriate because most analyses do not examine all possible trees, but instead employ heuristic search algorithms (Swofford and Beagle, 1993). Heuristic search algorithms start from a single tree and propose changes to the topology. The topological changes that the analysis employs are predefined, and vary greatly in their time requirement. To maximize the possibility of recovering the best tree, multiple replicates are performed using different starting trees. New, faster algorithms have been developed and are implemented in the program TNT (Goloboff et al., 2008). Parsimony was the first phylogenetic method employed and is still common today (Felsenstein, 2004). It is particularly useful in phylogenetic analyses of morphological characters or for molecular characters such as restriction fragments or sequence data with low levels of divergence. For more diverged sequence data, the method is not always consistent (does not recover the correct topology as the number of characters is increased) and will therefore sometimes produce incorrect but strongly supported answers (Felsenstein, 1978). The method is unique because it is not probabilistic and the assumptions of unweighted (or equal weighted) parsimony are unknown, although studies have attempted to determine what the method does *not* assume (Sober, 2004). For example, it has been shown that parsimony does not assume that homoplasies are rare, change is improbable, or that all changes are equally probable. Based on the principal of parsimony (minimizing the number of steps), such assumptions appear implicit, but the only apparent assumption of a parsimony phylogenetic analysis appears to be that a phylogenetic tree is chosen as the best hypothesis (although not necessarily bifurcating).

Maximum likelihood (ML) methods find the topology that maximizes the likelihood of observing the data given a model of evolution (Felsenstein, 1973; Huelsenbeck and Crandell, 1997). ML is more computer intensive than MP but has been found to be consistent and is often preferred (Felsenstein, 1978; Kuhner and Felsenstein, 1994; Swofford, 1999). Like MP, ML generally uses heuristic search algorithms and is not guaranteed to recover the topology with the true maximum likelihood. Unlike parsimony, the method requires an explicit model of evolution which is necessarily a simplification of the evolutionary processes (Ripplinger and Sullivan, 2008). The model of evolution accounts for some, but not all, actual facets of molecular evolution (e.g. transition-transversion ratio, nucleotide frequency, proportion of invariable sites, distribution of substitution rates), which increases the method's ability to correctly calculate likelihood scores.

To determine the best model of evolution (i.e. which parameters to include in the model), programs such as Modeltest (Posada and Crandall, 1998) and MrModeltest (Nylander, 2004) compare the log likelihood scores for models that vary in the parameters they incorporate. While adding parameters to a model will always make for a more likely model (or at least an equally likely model), the extra parameters do not necessarily make for a better model (because estimating the extra parameters will increase the variance associated with each parameter estimate and reduce the model's accuracy). To calculate the likelihood scores for each model, a phylogeny must be assumed because these programs do not actually calculate the likelihood of the model, but instead the likelihood of observing the dataset given the phylogeny and model of evolution. So, if the dataset and phylogeny are held constant, the likelihoods associated with each model can be calculated and compared. By default, Modeltest and MrModeltest use a neighbor-joining (NJ) tree. The likelihood scores associated with each model are then compared using hierarchical Likelihood Ratio Tests (hLRT) and Akaike Information Criterion (AIC). hLRT compare models using a chi-squared distribution where the difference in number of parameters between two models equals the degrees of freedom. AIC assigns a score to each model based on the likelihood, number of parameters in the model, and sample size. AIC balances bias (too few parameters) and variance (too many parameters) by maximizing the likelihood score while minimizing the number of included parameters and is considered superior to hLRT for model choice (Anderson, 2008; Burnham and Anderson, 2002; Posada and Buckley, 2004).

Once parsimony or ML analyses are complete, the next step is to determine if a single or multiple best trees have been recovered. If multiple trees are equally good phylogenetic hypotheses (similar likelihoods or equal numbers of steps), their differences need to be examined. This is more of an issue with parsimony, where hundreds of trees may have the same number of steps. When this occurs, a consensus tree is typically the best option. Consensus trees vary in how much of a consensus is required (must all the trees agree on a relationship? 95% of the trees? 50% of the trees?), but a strict consensus (100% agreement) is most often reported.

Once the best phylogeny is recovered, nodal support is calculated. To determine nodal support in parsimony and ML analyses, nonparametric bootstrapping is often employed (Felsenstein, 1985; Hillis and Bull, 1993). In this method, the character matrix is sampled with replacement such that some characters are sampled two or more times while others are not sampled at all. Going back to the dataset in Table 1.1, one possible bootstrap replicate is shown below.

Table 1.2 Hypothetical bootstrap replicate from taxa in Table 1.

	1	1	3	4	5	7	7	8	8	8
Human	A	A	G	G	C	G	G	T	T	T
Mouse	A	A	G	G	C	G	G	T	T	T
Alligator	C	C	G	C	G	G	G	T	T	T
Fish	C	C	C	C	G	T	T	T	T	T

In this example, the characters have been ordered based on their original numerical identifier for simplicity, but this will not occur normally (although order does not matter anyway). Notice that character numbers 1 and 7 were sampled twice, character number 8 was sampled three times, and character numbers 2, 6, 9, and 10 were not sampled at all in this replicate. This is repeated many times (>100 usually) and a new heuristic search is performed for each dataset. The trees recovered are then compared to each other such that nodal support is based on what percent of the time the bootstraps agree on the topology (0-100%). Therefore, the more congruent the dataset (how much the individual characters support the same topology), the higher the corresponding bootstrap support. Hillis and Bull (1993) demonstrated that a 70% bootstrap proportion at a node corresponds to approximately ≥ 95% chance that the clade is real. They stressed, though, that the 70 % bootstrap proportion is derived “under conditions of equal rates of change, symmetric phylogenies, and internodal change of ≤ 20% of the characters.” The technique and its interpretation, though, has been attacked over the years (Cummings et al., 2003; Erixon et al., 2003; Sanderson, 1995; Susko, 2008, 2009)

MrBayes (Huelsenbeck and Ronquist, 2001) is the most popular program for running Bayesian analyses and it uses the same (albeit a smaller repertoire) models of sequence evolution as ML analyses. Other programs, such as BEAST (http://beast.bio.ed.ac.uk/) and BAMBE (http://www.mathcs.duq.edu/larget/bambe.html), are available, and new, more efficient algorithms have been proposed (Cheon and Liang, 2009). MrBayes requires the user to decide the weighting scheme for topologies: Are all topologies equally likely or are some more likely than others? For most analyses, an equal prior probability associated with each possible tree is appropriate (this is the default setting) because a Bayesian analysis with equal prior probabilities will agree with a ML analysis run with the same parameters. But, since ML and Bayesian phylogenetic analyses employ different search algorithms, the two will not always agree (Svennblad et al., 2006).

The reason a Bayesian analysis and a ML analysis will agree if the prior probabilities are equal has to do with the computation of a Bayesian posterior probability. Bayes’ theorem states that the posterior probability of a set of parameters is equal to the likelihood of the parameters given some data (tree, branch lengths, model of evolution, etc.) multiplied by the prior probability of the parameters divided by the sum of the probabilities of all possible combinations of the parameters.

$$\mathrm{P}(H|Data) = \frac{\mathrm{Prior}(H)\ \times \mathrm{P}(Data|H)}{\mathrm{Prior}(H)\ \times \mathrm{P}(Data|H) +\ \mathrm{Prior}(\ not\ H)\ \times \mathrm{P}(Data|not\ H)}$$

Figure 1.1 Bayes theorem where P = probability, Prior = prior probability, and H = hypothesis.

Therefore, when all topologies have equal prior probability, the tree with the maximum likelihood will also have the highest posterior probability. But, because Bayesian and ML phylogenetic analyses treat nuisance parameters differently, the two can recover different topologies. In analyses, parameters that are important for phylogeny estimation (e.g. transition-transversion ratio, nucleotide frequency, etc) but are not of direct interest are called nuisance parameters. ML analyses fix such parameters to a specific value (generally based on the data) and maximize the likelihood of the topology with respect to them. Therefore, the ML analyses do not calculate true likelihoods, but instead profile likelihoods. Profile likelihoods do not incorporate parameter estimate variability and as a result the phylogeny likelihoods will be imprecise (Huelsenbeck et al., 2002). In a Bayesian analysis, all parameters are assigned a prior probability distribution (based on the data, but they may contain flat prior distributions) and the likelihood of each topology is calculated by integrating over all possible values of all parameters (models of evolution for a MrBayes analysis only dictate which parameters the program should estimate). This marginalization approach allows the user to specify important parameters without fixing them to imperfectly estimated values.

Once the user has defined the prior information and dataset, the next step is to set the partitioning scheme. This is one advantage that MrBayes has over many (but not all) ML and MP programs because it allows the user to assign specific models and independence to specific portions of the dataset. For example, if the dataset contains two sequence fragments, an intron and a protein coding gene, then the user can specify a different model for each dataset and assume that all evolutionary model parameters are independent. The user can further partition the protein coding gene by codon position or the ribosomal gene by stem and loop. This allows each disparate partition to evolve at its own rate, improving the estimated model of evolution.

After deciding on a partitioning scheme, the user must set the search parameters. MrBayes has default search parameters, but often changing these can improve the overall analysis. Below is a list of a few of the more important parameters. Italicized letter combinations (e.g. *nreps*) refer to the command description in MrBayes. See the online manual (http://mrbayes.csit.fsu.edu/wiki/index.php/Manual) for further information.

Number of repetitions (*nreps*) and number of chains (*nchains*)- Like MP and ML analyses, Bayesian analyses are typically done multiple times. Because each repetition is started with a random tree, this increases the chance that the true maximum likelihood, and not just a local optimum, is reached. Computational time is directly proportional to the number of runs, but higher numbers of runs increase the probability of recovering the best topology. *nchains* sets the number of independent chains used per repetition. The default value is four, but for larger datasets or for datasets with convergence issues, this should be raised to six or eight. The greater the number of chains the more easily the analysis locates isolated peaks and increases chain swapping. Unlike repetitions, chains are not completely independent, but instead commonly swap “states.” MrBayes also uses an incremental heating scheme across chains (see below), so the chains search through tree space differently.

Metropolis coupling- This is a variation of the Markov Chain Monte Carlo analysis and is controlled using several parameters; *swapfreq*, *nswaps*, *nchains*, and *temp.* Metropolis coupling is unique to MrBayes in Bayesian phylogenetic analyses and allows multiple search paths to occur concomitantly and for each independent chain to swap information with each other. The analysis usually has one cold and multiple hot chains with increasing “temperatures.” When the “temperature” of a chain is increased, the difference in likelihood scores between competing models is reduced (Fig. 2.2). This allows chains with higher temperatures to search across the landscape more freely (search through trees with low likelihood scores) since they will be more likely to search tree space than a cold chain. An essential component is swapping the states of the hot and cold chains so that if a hot chain does find a better maximum, the cold chain can swap states with the hot chain. This is important because the cold chain is what is examined at the end of an analysis.

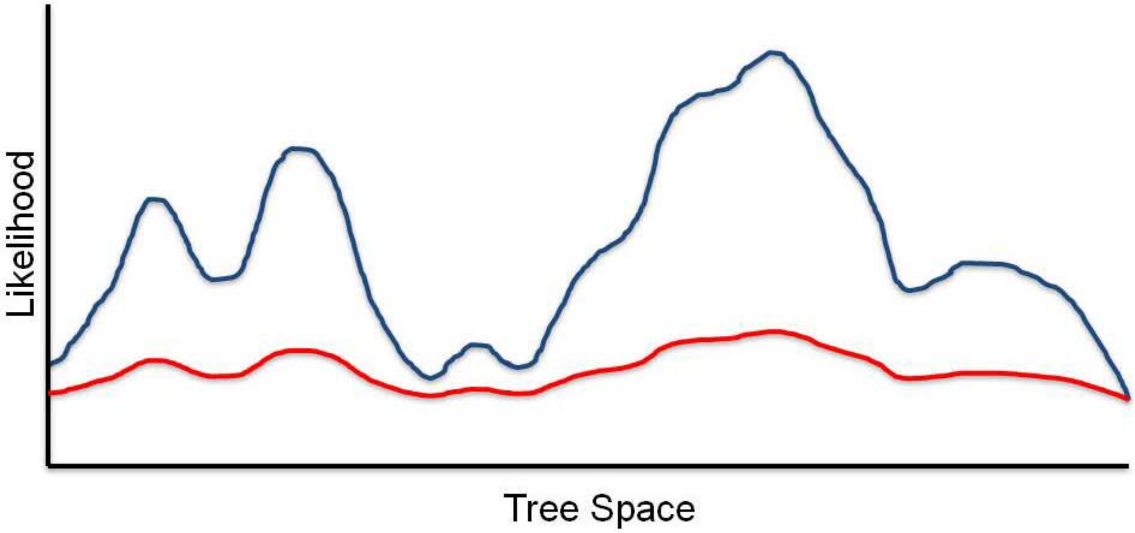

Figure 1.2 Theoretical diagram depicting difference between normal likelihood scores (blue line) and scores calculated under a higher temperature (red line).

Swap Frequency (*swapfreq*) and number of swaps (*nswaps*)- Defines how often a swap between two random chains is attempted and controls the number of swaps that are attempted when swapping occurs. Typically, the default setting of every generation is used. To maximize the search algorithms efficiency, hot and cold chains need to swap “states” so that the cold chain will find the global, not just local optimum. Increasing this parameter may help the analysis search more efficiently.

Number of generations (*ngen*)- Defines how many generations will occur. In MrBayes, this is a soft stop because the program allows the user to look at specific parameters (although not as many as would be desired) and determine if a longer run is needed. Longer runs are needed to more accurately estimate node posterior probabilities and also to ensure complete tree space searches.

MrBayes starts with a random tree for each run unless a starting tree is specified. The program also has a user specified number of chains, one of which is a “cold” chain while the others are “hot.” MrBayes uses an incremental heating scheme so the temperature of each chain is different (this is defined by the “temp” parameter). Hot chains are able to explore tree space more easily while cold chains are more conservative. Randomly chosen chains propose swaps every generation (n=1 swap per generation is default, although this can be increased). If the swap is accepted, the chains change states and the Metropolis coupling algorithm is most efficient. If the chains cannot swap consistently, then the algorithm’s search ability is compromised.

Each generation, the program proposes changes to a single parameter. The proposed parameter is based on the temperature of the chain and current value of the parameter. Most of the time, changes in topology are proposed. If the proposed change increases the likelihood of the topology, then it is accepted. Otherwise, if the proposed change decreases the likelihood, the likelihood of the original parameter set is divided by the proposed parameter set likelihood and this fraction is compared to a specific number. If the ratio of the two likelihoods is greater than the specific number, then the change is accepted. The specific amount is a random value between 0 and 1.

Once the search parameters are set and the run is complete, the user should look at several diagnostics to make sure stationarity of the chains has been reached (Mossel and Vigoda, 2005). Stationarity implies that all chains have reached a similar maximum likelihood value. First, the average standard deviation of split frequencies should be <0.01. This is a measure of dissimilarity between runs, so smaller values are better.

Next, the user needs to check the log likelihood plot over time and make sure it is not increasing. An excellent program to visualize this and other check of stationarity is AWTY (Nylander et al., 2008).

If stationarity of the chains has not been reached, the user must rerun the analysis. Sometimes simply increasing the number of generations will fix the problem, but often the Metropolis Coupling needs to be adjusted. If the analysis has reached stationarity, then the Markov Chain has been run the minimum number of generations. This is different than MP and ML analyses which then need to perform bootstrap replicates to determine nodal support (although Bayesian analyses can assess bootstrap support if desired). The user does need to determine the burnin time for the analysis. Burnin is the period at the beginning of the analysis where the program is converging on the tree with the highest observed likelihood. This depends on each dataset (and analysis) and can vary from hundreds to millions of generations. This is best determined by examining a log likelihood plot over time and determining at what generation an asymptote is reached. The burnin generations are then ignored for further analyses. This can easily be visualized in MrBayes by using the "*sump*" command.

Determining the phylogenetic hypothesis from the analysis is another way Bayesian analysis differs from MP and ML analyses; the latter which either use the tree(s) with the fewest steps or the highest associated likelihood. For Bayesian analyses, different methods have been proposed. One is to use the tree with the highest likelihood. This is similar to the ML method. Another method employs the 50% majority rule consensus tree and is the default method in MrBayes.

To determine the posterior probabilities associated with each node in a Bayesian analysis, the standard method is to examine all of the post-burnin trees (there may be hundreds or thousands) and see what percentage of the time they agree on each clade (the higher the number of post-burnin generations, the better the estimate). The posterior probability of each tree sampled is calculated in the same manner. To exactly measure the posterior probability of the tree or node, Bayes' Theorem needs to be employed, but this is computationally impossible for large datasets because a likelihood score for every possible tree would need to be calculated. Instead, MrBayes estimates the true posterior probability based on the Markov chain. This is obviously different than employing Bayes' theorem, but this approach is considered a good approximation of the true posterior probability(Huelsenbeck and Ronquist, 2001; Huelsenbeck et al., 2002). For example, trees (or nodes) that are represented in 75 % of the sampled generations have a posterior probability of 0.75.

The Markov chain is useful for not only estimating phylogeny parameters but also for other hypothesis tests. The harmonic mean of the post-burnin likelihood scores is part of the output of MrBayes and can be used to compare multiple hypotheses directly. For example, because MrBayes can employ multiple partitioning schemes, a logical question is whether one *should*. Estimating parameters separately for each partition should reduce bias, but it should also increase the variance associated with each estimated parameter (which reduces the accuracy of the estimate). One approach to answering this question is to run two separate analyses, one partitioned (H_1) and one not (H_0). The Bayes factor is calculated as twice the difference between the harmonic mean of log likelihood scores between H_0 and H_1. Bayes Factor values <0 are interpreted as evidence against H_1, while positive values provide either basically no evidence for H_1 (0-2), positive support for H_1 (2-6), strong support for H_1 (6-10), or very strong support for H_1 (>10) (Brown and Lemmon, 2007; Kass and Raftery, 1995). PuMA (Brown and ElDabaje, 2009) adopts a different approach and instead uses the parameters and tree estimated from the model and compares the simulated set to the original dataset. If the simulated and original datasets are similar, then the model gets a high corresponding likelihood.

CHAPTER 2

ANALYZING THE RELATIONSHIP BETWEEN SEQUENCE DIVERGENCE AND NODAL SUPPORT USING BAYESIAN PHYLOGENETIC ANALYSES

Introduction

Phylogenetic reconstruction requires choosing character sets (e.g. genes, gene fragments, genome characteristics) that are appropriate for the proposed question based on their availability, cost, expected efficacy, and tractability (Hillis et al., 1996; Meyer, 1994). A plethora of newly available genomic characters (microsatellites, AFLPs, SINEs, LINEs, nucleotides, etc.) are widely used (Avise and Saunders, 1984; Hillis, 1999; Murata et al., 1993; Richard and Thorpe, 2001; Vos et al., 1995). While our knowledge of molecular evolution has highlighted instances where specific molecular characters are appropriate for specific analyses, there are also situations for which the same molecular character sets are inappropriate (Graybeal, 1993; Vekemans et al., 2002). Although generally well known, only recently have researchers begun to explore these issues in more detail (Collins et al., 2005; Lemmon and Moriarty, 2004; Nylander et al., 2004; Ripplinger and Sullivan, 2008; Rokas and Carroll, 2005; Seo and Kishino, 2008; Sullivan et al., 2004; Vekemans et al., 2002).

For example, Rokas et. al. (2003) used complete genomes of seven species of yeast to demonstrate that a large number (greater than 20) of randomly chosen protein-coding genes are required to recover the correct tree. However, Collins et al. (2005) noted that non-stationary genes (i.e., relatively unequal nucleotide frequencies across taxa) were included in their analyses and proposed that restricting the analysis to genes that are stationary would better meet the assumptions of current phylogenetic methods. They showed that excluding non-stationary genes from the analysis substantially reduced the number of randomly chosen genes needed to recover the correct topology to roughly eight. Rodriguez-Ezpelata et al. (2007) reached a similar conclusion and reported that removing fast-evolving positions reduced systematic error. Townsend (2007) demonstrated theoretically that an optimal rate of change per unit time exists using the four taxon case, but the need for estimated times and lack of description of an informative range makes implementation difficult.

Rate of molecular evolution within and among genes is a simple characteristic of a data set that may affect phylogenetic performance. In the case of rapidly evolving sequences, alignment and determination of character homology may be difficult or impossible (Blouin et al., 1998; Lopez et al., 1999; Xia et al., 2003). For intraspecific analyses, many mitochondrial (mt) genes, as well as nuclear markers such as microsatellites and AFLPs, usually provide phylogenetic signal without saturation (Berendzen et al., 2003; Creer et al., 2004; Dawson, 2001; Downie, 2004; Koopman, 2005; Vekemans et al., 2002). For slowly evolving sequences, the number of variable sites (and therefore the number of informative sites) will be low and incomplete lineage sorting of ancestral polymorphisms may obscure relationships (Maddison, 1997; Maddison and Knowles, 2006; Takahashi et al., 2001). Deeper relationships require more slowly evolving genes (e.g. nuclear ribosomal genes) to recover the correct topology (Avise, 2000; Hare, 2001; Palumbi et al., 2001). However, the same genes may evolve at different absolute rates across lineages, so a taxonomic consideration is also important. For example, cytochrome oxidase I may be the fastest evolving mt gene in some lineages, while NADH-II may be the fastest in others (Kumazawa et al., 2004; Mueller, 2006).

Although numerous solutions to the problem of insufficient or excessive divergence have been proposed, they only partially address the issue. If a sequence region is not variable enough, a larger fragment may be sequenced, or another gene added to the analysis. While this increases the number of characters, incomplete lineage sorting of ancestral polymorphisms may remain a problem. If a gene is protein coding and too variable, use of amino acid sequence, down-weighting of saturated positions (site-stripping), or omission of third codon positions from the analysis are options (Ketmaier et al., 2006; Morgan and Blair, 1998; Pratt et al., 2009; Ros and Breeuwer, 2007). Removing introns if they are present can also reduce excessive homoplasy. These approaches decrease homoplasy that occurs due to high sequence divergence, but simultaneously lessen the number of potentially informative characters and rarely resolve alignment issues. Unfortunately, it rarely can be determined if a gene will suffer from homoplasy, incomplete lineage sorting, or non-stationarity before the ingroup has been thoroughly sampled. Therefore, a particular question requires the researcher to know *a priori* which genes at their disposal are appropriately variable. Ranwez et al (2007) developed an algorithm that screens the genomes of species and locates genes that have the highest predicted phylogenetic utility based on stationarity, homogeneous site variability, and evolutionary rate. Unfortunately, while their parameters for determining stationarity and homogeneous site variability are well justified, their required choice of an arbitrary evolutionary rate (branch lengths exhibiting >2 substitutions per site when calculated with uncorrected pairwise distances on an NJ tree) limits the programs efficacy. One goal of my research is to provide such search algorithms with a better justified range of sequence divergence.

Sequence divergence is the direct result of nucleotide substitutions, which occur according to the properties of specific genes (invariable sites and transition / transversion ratio due to selection) and genomic environment (nucleotide and amino acid bias). Despite the variation in how genes accumulate nucleotide substitutions, this approach has proved useful in past analyses. For the mt cytochrome-*b* gene, it was estimated that sequences become saturated and uninformative at 15-20% uncorrected divergence in bufonid frogs using variably weighted parsimony (Graybeal, 1993). Yang (1998) also suggested a 15-20% uncorrected sequence divergence using a simulated data set with a four taxon tree and parsimony. The last 15 years, though, have brought about the innovation and tractability of many computationally intensive methods; these include maximum likelihood analyses, Bayesian analyses, increasingly complex (more realistic) models of molecular evolution, and programs that can partition data sets (e.g. codon position). Therefore, there is a need for research that takes advantage of these powerful new techniques and incorporates widespread taxonomic sampling to determine how phylogenetic reconstruction ability is affected by differing levels of sequence divergence.

Here, I determine whether there exists an optimal range of sequence divergence with broad applicability across taxa and divergence times. My goal is to not determine how specific situations (radiation, non-stationarity, etc) affect the optimal range, but instead whether an optimal range exists while *ignoring* such confounding variables. Specifically, I determine if researchers should aim for a particular range of sequence divergence during phylogenetic analysis planning to maximize the probability of recovering the correct topology. Our goals are to 1) identify (if possible) a global range of sequence divergence that maximally recovers the correct topology and 2) determine whether different types of genes (mt or nuclear, protein encoding or ribosomal) exhibit specific ranges of divergence for optimal phylogenetic reconstruction. I use a well-corroborated phylogeny (that I treat as “known” for the purpose of analyses) and compare the trees recovered from 12 genes using Bayesian methods to the assumed true topology to determine at what levels of sequence divergence phylogenetic methods most often recover the correct topology. I also used simulations to test whether a relationship between sequence divergence and phylogenetic reconstruction ability exists across different topologies and models of evolution. Additionally, I test whether certain intrinsic properties of evolutionary history (branch lengths and unequal rates of sequence evolution across taxa) affect phylogenetic performance.

Materials and Methods

Natural Data Sets

For our model phylogeny I started with the "known" phylogeny presented in Russo et al. (1996) and added taxa based on sequence availability and strength of relationship support (Fig. 2.1). I followed the phylogenetic relationships presented in multiple independent analyses (citations below refer to support for relationships of taxa) using multiple types of character sets. Within mammals, the whales in the Russo et al. tree were reduced to one taxonomic unit and five OTUs were added from the following lineages; Canidae (Node 23), Felidae (two taxa; Node 24), Marsupiala (Node 18) and Primata (Node 20) (Douady and Douzery, 2003; Hudelot et al., 2003; Lin et al., 2002; Liu et al., 2001; Murphy et al., 2001; Phillips and Penny, 2003; Prasad et al., 2008; Waddell and Shelley, 2003). I added Crocodilia (Node 17) and Squamata (three taxa; Node 15 & 16) to the lineage represented by chickens in the Russo et al. tree (Cao et al., 2000; Cotton and Page, 2002; Hedges and Poling, 1999). Sister to the Reptilia (Node 14) and Mammalia (Node 18) (i.e. Amniota (Node 13)) are Amphibia (Node 7), which were divided into Anura (three taxa; Node11) and Caudata (four taxa; Node 8). Within Caudata, two *Plethodon* salamanders (Node 10) are sister to *Eurycea* (Node 9) which collectively are sister to Ambystomatidae; within Anura, *Xenopus* (Node 11) is sister to the Bufonidae-Ranidae clade (Node 12) (Chippindale et al., 2004; Frost et al., 2006; Hugall et al., 2007; Min et al., 2005; Mueller et al., 2004). Collectively, Tetrapoda (Node 6) is sister to the Teleostei (Node 2), which are represented by five taxa; Cyprinidae, Salmonidae (two taxa; Node 5) and Tertraodontidae (two taxa; Node 4) (Cotton and Page, 2002; Mank et al., 2005; Miya et al., 2003). Chondrichthyes (*Mustelus manazo)* and Cephalochordata (*Brachiostoma japonicum*) were used as outgroups. Details on specific sequences can be found in Appendix B. While I acknowledge that the phylogeny is not known precisely (even though the relationships are well supported, without directly observing the evolution of vertebrates, we cannot truly know their history), I think that the multiple lines of evidence cited above strongly support the phylogenetic relationships presented.

Simulated Data Sets

All simulated data sets were created using Mesquite 2.6 (Maddison and Maddison, 2009). Parameter values for the simulated data sets consisted of estimates from a total evidence analysis of the natural data sets. These include: nucleotide frequency (A=0.3473, C=0.2812, G=0.1516, T=0.2199), proportion of invariable sites (0.317), GTR rate matrix (A-C=1.66, A-G=3.51, A-T=2.21, C-G=0.67, C-T=12.06), and gamma distribution of site rate variability (0.57). Rate variability among codon positions was defined as the data set's average evolutionary rate multiplied by 0.49, 0.27, and 2.21 for first, second, and third positions respectively. Data sets of varying evolutionary rate were evolved according to either the topology of the "known" tree or a specific variation (same taxa and evolutionary relationships, different branch lengths). Variations included a topology with equal branch lengths, a "radiation" topology with terminal branch lengths ten times longer than all internodal branch lengths that would represent a rapid radiation

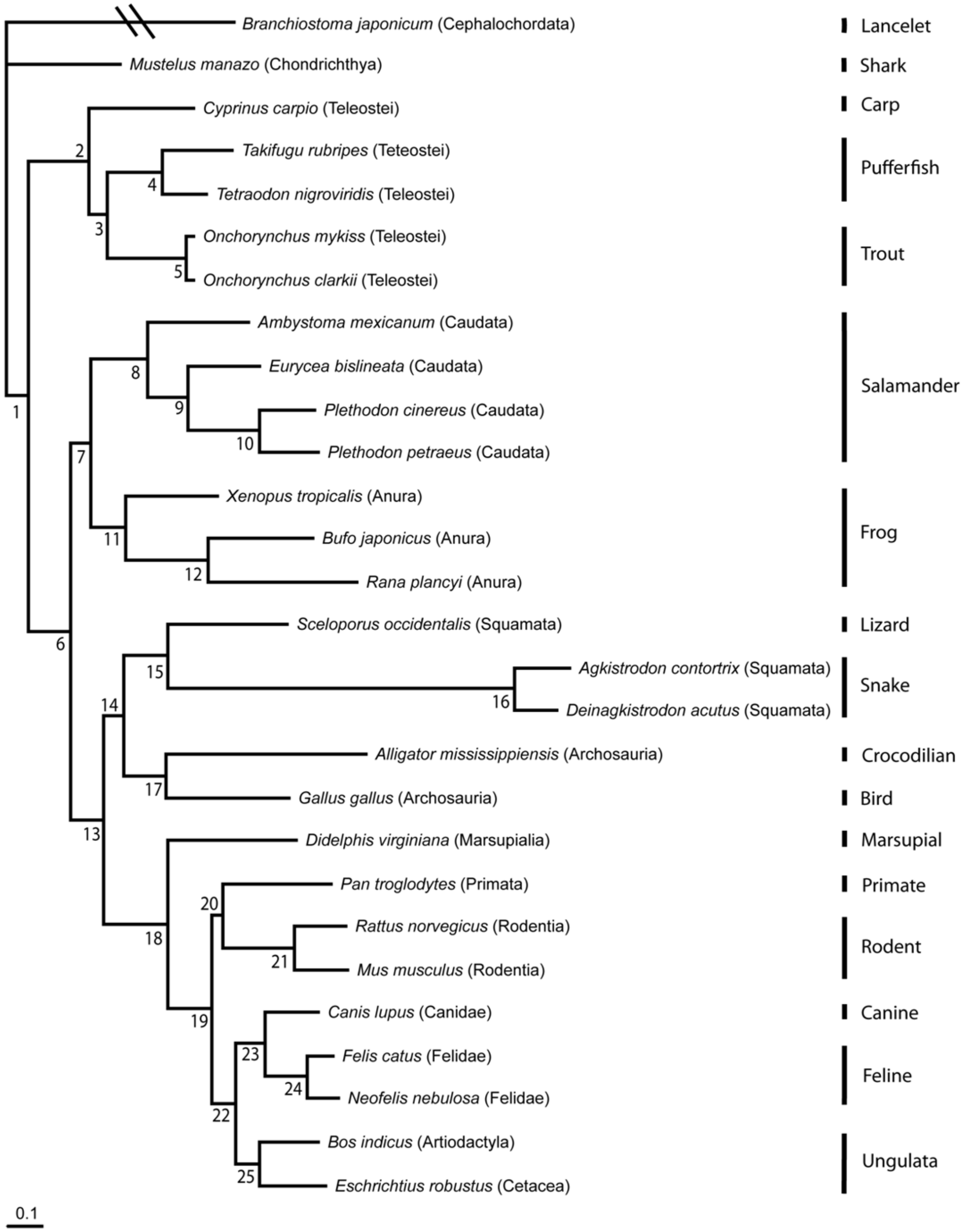
Branchiostoma japonicum (Cephalochordata)
Mustelus manazo (Chondrichthya)
Cyprinus carpio (Teleostei)
Takifugu rubripes (Teteostei)
Tetraodon nigroviridis (Teleostei)
Onchorynchus mykiss (Teleostei)
Onchorynchus clarkii (Teleostei)
Ambystoma mexicanum (Caudata)
Eurycea bislineata (Caudata)
Plethodon cinereus (Caudata)
Plethodon petraeus (Caudata)
Xenopus tropicalis (Anura)
Bufo japonicus (Anura)
Rana plancyi (Anura)
Sceloporus occidentalis (Squamata)
Agkistrodon contortrix (Squamata)
Deinagkistrodon acutus (Squamata)
Alligator mississippiensis (Archosauria)
Gallus gallus (Archosauria)
Didelphis virginiana (Marsupialia)
Pan troglodytes (Primata)
Rattus norvegicus (Rodentia)
Mus musculus (Rodentia)
Canis lupus (Canidae)
Felis catus (Felidae)
Neofelis nebulosa (Felidae)
Bos indicus (Artiodactyla)
Eschrichtius robustus (Cetacea)
Lancelet
Shark
Carp
Pufferfish
Trout
Salamander
Frog
Lizard
Snake
Crocodilian
Bird
Marsupial
Primate
Rodent
Canine
Feline
Ungulata
1
2
3
4
5
6
7
8
9
10
11
12
13
14
15
16
17
18
19
20
21
22
23
24
25
0.1

Figure 2.1 The “known” phylogeny used for this study with branch lengths estimated from a Bayesian total evidence analysis. Lancelet and Shark are the outgroups. OTUs are labeled with both primary species and clade names corresponding to the text (see Appendix A and text for explanation). Numbers at nodes correspond to Table 3.

followed by anagenesis, and a topology in which a strict molecular clock was enforced (created using the randomly ultrametricize option in Mesquite 2.6).

To determine the effect of assuming an incorrectly parameterized model of evolution, 15 data sets of varying evolutionary rate were modeled under each of three evolutionary models that incorporate an increasing number of parameters. The first model incorporates only nucleotide frequencies and is equivalent to the Jukes Cantor (JC) model. The second model (GTR) includes nucleotide frequencies and a general time reversible rate matrix for nucleotide change. The third model (GTR + I + G) includes nucleotide frequencies, a general time reversible rate matrix for nucleotide change, the proportion of invariable sites and a gamma shaped distribution of site rate variability. To determine the effect of topology, 15 data sets of varying evolutionary rate were modeled upon the four topological variations described above using the GTR + I + G model. Overall, a total of 90 simulated data sets were created and each one was analyzed separately. See Table 2.1 for a complete description of each simulated data set.

Table 2.1 How simulated data sets were modeled as well as results of statistical tests.

Simulation Category Name	**Simulation Model of Evolution**	**Simulation Topology**	**Number of Significant KS Tests**	**Results of Mood's Median Test (d.f. =5)**
JC	JC	Known	6	X^2 = 119.66, P < 0.000
GTR	GTR	Known	5	X^2 = 63.46, P < 0.000
GTR + I + G	GTR + I + G	Known	4	X^2 = 36.39, P < 0.000
Equal	GTR + I + G	Equal	4	X^2 = 59.23, P < 0.000
Radiation	GTR + I + G	Radiation	4	X^2 = 28.80, P < 0.000
Ultrametric	GTR + I + G	Ultrametric	2	X^2 = 25.74, P < 0.000

Data Collection

Our “known” phylogeny was completely sampled for eight of the 12 genes while four genes (BDNF, 18S, 28S, and RAG-1) were missing one or more taxa (Appendix B). Several sequences available on Genbank were removed because they were either too short or were likely pseudogenes (contained mis-sense or non-sense mutations). I defined a primary species for each OTU and if this primary species did not have the necessary sequences I substituted sequences of closely related taxa. Because I measured average corrected sequence divergence for each gene, using different individuals or species for each taxonomic unit in the study is not likely to compromise our results.

Sequences were aligned in Mega 4.0 (Tamura et al., 2007) using default parameters. All ambiguously aligned regions were removed prior to analysis (in frame for protein coding genes since analyses were partitioned by codon) and I limited the size of each fragment to 750 base pairs (bp). Sequences were standardized by removing portions of the 5’ and 3’ end because it is within the range of sequence lengths commonly used in phylogenetic analyses and it is suspected that branch support is dependent on the amount of data (Aguileta et al., 2008; Jermiin et al., 2005). For most genes, equal sized fragments were used for all OTUs, but in a few cases I included partial fragments (> 375 bp) if complete sequences were not available.

I calculated the corrected pairwise sequence divergence for each taxon pair and each gene using uncorrected p (i.e., simple percent difference), Kimura 2 parameter (K2P) and Tamura-Nei with gamma distributed rates among sites in MEGA 4.0 (Tamura et al., 2007). I then calculated the average pairwise divergence and standard deviation for each node for each gene by averaging all terminal taxa pairs. For example, if four taxa had the relationship ((A B)(C D)), the average pairwise divergence at the ancestral node was calculated by averaging the divergences observed between A-C, B-C, A-D, and B-D.

I ran a Bayesian phylogenetic analysis for each data sets (natural or simulated) using MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) with the following parameters: nst = 6, rates = invgamma, ratepr = variable, statefreqpr = dirichlet (1,1,1,1) and unlinked shape. For protein coding genes, each codon position was analyzed separately during analyses. I chose to use the GTR + I + G model for all genes because MrModeltest (Nylander, 2004) returned this model for 11 of 12 genes (using the Akaike Information Criterion) and model over-parameterization should not negatively affect the analysis (Castoe et al., 2004; Lemmon and Moriarty, 2004). Each analysis included six chains, sampling every 1,000 generations, and was run for at least 7,500,000 generations (default parameters otherwise). Stationarity of the analysis (likelihood scores that were neither increasing or decreasing over generations) was determined by examining the standard deviation of split frequencies (<0.01) and *–ln* likelihood plots in AWTY (Nylander et al., 2008). Burnin calculations were conservative, between 2.5 and 5.0 million generations. To assess phylogenetic performance I used the posterior probabilities associated with each “correct” node (i.e. congruent with the “known” phylogeny) by examining the observed bipartitions in the 50 percent majority-rule consensus tree.

To analyze the relationship between posterior probability and sequence divergence for the natural and simulated data sets, I divided divergence level into six categories with equal sample sizes. I performed a two-sample Kolmogorov-Smirnov (K-S) test to see if there were pairwise differences in posterior probability distribution between the six categories in Systat 11 (Systat Software Inc, Chicago, IL). Finally, I performed a Mood's median test to see if there were significant differences among divergence categories using Minitab 14 (Minitab Inc., State College, PA).

I also performed a total evidence analysis for all taxa in the natural data sets to determine what effect branch lengths have on phylogenetic reconstruction. Because I was not always able to use the same species per OTU, sequences of substituted species were necessarily deleted so that the concatenated dataset was realistic. I partitioned the analysis by gene and codon position (except ribosomal genes) and used MrBayes' default parameters except that I constrained the topology (Fig. 1) and reduced the proportion of topology changes (TBR, NNI, etc) during chain swapping. The analysis was run for 5,000,000 generations (2,000,000 burnin) and evaluated using the same methods described above. I calculated an average posterior probability for each node and regressed it against the branch lengths calculated from the total evidence analysis.

Results

Pairwise divergences within genes ranged from 0.000 to 2.93 substitutions per site (based on the model of substitution, Table 2.2) and node standard deviations ranged from 0.0 to 0.152. The different correction measurement models yielded very similar patterns (the difference being scale of divergence axes), so only K2P corrected distances are presented in the following figures. Posterior probabilities for correct nodes ranged from 0.0% to 100% (Table 2.2).

The relationship between sequence divergence and posterior probability for all genes recovered an optimal range of divergence of approximately 0.12 - 0.21 K2P corrected (Fig. 2.2). Specifically, sequences that were either too divergent or too similar recovered lower average posterior probabilities for correct nodes. Analyses using mt protein coding genes (there was little observable difference between the combined mt and nuclear protein vs. divergence and the mt protein vs. divergence plots, so only one is reported) recovered maximal phylogenetic performance in the 0.07 K2P sequence divergence bin and recovered the correct topology with highly similar sequences. Analyses using nuclear protein and ribosomal genes showed an unexpected lack of any relationship that may be more an artifact of low gene sample size than the true pattern.

The standard deviation of node divergence was positively correlated with average K2P corrected pairwise divergence for ribosome, protein, and combined data (Fig. 2.5), meaning that genes with high levels of divergence also exhibited higher variance levels.

Table 2.2 Maximum pairwise divergence between taxa for different substitution models for the natural data sets.

Gene	Location	Minimum-Maximum pairwise divergence for uncorrected p / K2P / Tamura-Nei gamma	Mean pairwise divergence for uncorrected p / K2P / Tamura-Nei gamma
Cyt *b*	Mitochondrion	0.056-0.429 / 0.059-0.664 / 0.062-1.212	0.288 / 0.371 / 0.513
Cox 1	Mitochondrion	0.045-0.356 / 0.047-0.497 / 0.050-0.756	0.233 / 0.285 / 0.363
Cox 3	Mitochondrion	0.039-0.403 / 0.040-0.599 / 0.042-0.977	0.273 / 0.346 / 0.463
ND1	Mitochondrion	0.065-0.440 / 0.069-0.676 / 0.076-1.185	0.306 / 0.401 / 0.559
ND2	Mitochondrion	0.073-0.569 / 0.078-1.109 / 0.086-2.930	0.388 / 0.561 / 0.951
ND4	Mitochondrion	0.057-0.475 / 0.060-0.767 / 0.064-1.447	0.324 / 0.434 / 0.629
ND5	Mitochondrion	0.055-0.472 / 0.058-0.755 / 0.061-1.375	0.304 / 0.398 / 0.569
12S	Mitochondrion	0.008-0.464 / 0.007-0.767 / 0.008-1.635	0.234 / 0.301 / 0.410
18S	Nucleus	0.000-0.069 / 0.000-0.073 / 0.000-0.073	0.032 / 0.031 / 0.032
28S	Nucleus	0.000-0.096 / 0.000-0.103 / 0.000-0.113	0.033 / 0.034 / 0.036
RAG-1	Nucleus	0.017-0.340 / 0.033-0.411 / 0.018-0.710	0.238 / 0.296 / 0.381
BDNF	Nucleus	0.007-0.292 / 0.007-0.374 / 0.007-0.498	0.176 / 0.208 / 0.250

Table 2.3 K-S pairwise comparison results (*P* values) between divergence groups for both the mt protein data sets (**top right, bolded**) and complete data sets (bottom left). Notice that the K2P corrected sequence divergence for the groups is different for the two data sets. The Bonferroni corrected p-value is 0.003 and statistically significant comparisons are marked with an *.

	.07	**.19**	**.24**	**.29**	**.35**	**.45**
0.02	-	**0.832**	**0.001***	**0.002***	**0.000***	**0.001***
0.12	0.041	-	**0.006**	**0.006**	**0.000***	**0.000***
0.21	0.040	0.607	-	**0.961**	**0.640**	**0.640**
0.27	0.301	0.002*	0.040	-	**0.640**	**0.999**
0.33	0.196	0.000*	0.005	0.781	-	**0.832**
0.44	0.195	0.000*	0.000*	0.780	0.440	-

Table 2.4 Recovered posterior probability of each node for each gene and the node’s estimated preceding branch length (BL).

Clade Name & #	12S	18S	28S	BDNF	Cox1	Cox3	Cyt b	ND1	ND2	ND4	ND5	Rag1	Mean	BL
All Taxa (1)	16	-	4	-	4	28	0	0	23	0	0	0	8	0.061
Teleostei (2)	98	0	-	-	46	98	4	0	30	9	31	100	42	0.159
Tert. + Sal. (3)	45	1	-	-	95	67	0	0	4	0	100	99	41	0.049
Tertraodontidae(4)	100	-	-	-	71	100	100	7	100	66	100	100	83	0.147
Salmonidae (5)	100	65	-	100	100	100	100	100	70	100	100	100	94	0.21
Tetrapoda (6)	3	2	4	100	8	1	0	100	99	25	0	100	37	0.109
Amphibia (7)	0	0	86	41	0	100	5	100	0	11	0	0	29	0.056
Caudata (8)	95	-	100	-	99	94	100	100	100	100	99	100	99	0.153
Plethodontidae (9)	99	-	95	100	22	39	99	99	100	100	100	100	87	0.107
Plethodon (10)	100	-	1	1	100	100	100	100	100	100	100	100	82	0.191
Anura (11)	65	0	100	100	0	9	91	87	0	16	100	95	55	0.093
Bufo.-Ranid (12)	93	0	7	-	0	98	100	99	97	100	1	100	63	0.217
Amniota (13)	81	54	0	100	0	1	23	100	3	100	90	100	54	0.088
Reptilia (14)	98	66	70	100	0	0	0	0	85	100	97	100	60	0.057
Squamata (15)	82	91	-	100	45	60	4	67	43	100	86	100	71	0.117
Serpentes(16)	100	98	-	-	100	100	100	100	100	100	100	100	100	0.925
Archosauria (17)	37	18	-	100	100	99	53	100	44	100	11	0	60	0.108
Mammalia (18)	100	23	90	100	8	100	73	100	100	100	99	100	83	0.175
Theria (19)	100	13	33	100	0	97	100	51	100	99	96	100	74	0.116
Archonta(20)	7	17	-	11	1	98	0	94	90	2	0	3	29	0.031
Rodentia (21)	100	18	27	100	81	100	93	100	100	100	100	100	85	0.189
Car. & Ung. (22)	6	-	-	87	0	96	6	98	94	100	66	26	58	0.063
Carnivora (23)	26	-	-	10	1	100	0	82	100	47	100	100	57	0.079
Felidae (24)	100	-	-	100	96	100	100	100	100	100	100	100	100	0.113
Ungulata (25)	71	0	-	2	94	100	100	93	84	100	99	100	77	0.064

The posterior probability associated with a node was significantly related (Mood's median test; $\chi^2 = 16.47$, df = 3, $P < 0.001$) to the node's standard deviation for all data sets (Fig. 2.3), reaffirming the notion that high levels of rate variability lead to reduced phylogenetic performance (Harrison and Larsson, 2008).

When the output for all genes was combined and analyzed as a single dataset, K-S tests found four out of 15 significant pairwise differences among distributions of bins ($0.003 < P < 0.05$; Table 2.3) after sequential Bonferroni corrections ($P < 0.003$). I also tested whether the posterior probability medians among divergence groups differed using a Mood's median test ($\chi^2 = 20.03$, df = 5, $P = 0.001$). For the protein coding data sets, six of the 15 probability distributions (K-S test; $P < 0.003$) differed significantly among one another, and the group medians also differed significantly (Mood's median test; $\chi^2 = 29.43$, df = 5, $P < 0.000$). The posterior probability distributions and medians (Mood's median test; $\chi^2 = 1.54$, df = 3, $P = 0.673$) for the ribosomal data sets and the nuclear protein data sets (K-S test; all *P*'s > 0.9; Mood's median test; $\chi^2 = 1.19$, df = 4, $P = 0.879$) did not significantly differ, although it is important to note that the sampling was limited for these data sets.

There was a significant positive relationship between branch length and posterior probability (Fig 2.4). Due to the long branch-length of snakes I tested whether snakes were biasing the results, but found that the significance of the relationship between branch and posterior probability also exists when snakes were excluded (F = 6.05, df = 24, $r^2 = 20.8\%$, $P = 0.022$ with snakes; F = 11.97, df = 23, $r^2 = 35.2\%$, $P = 0.002$ without snakes).

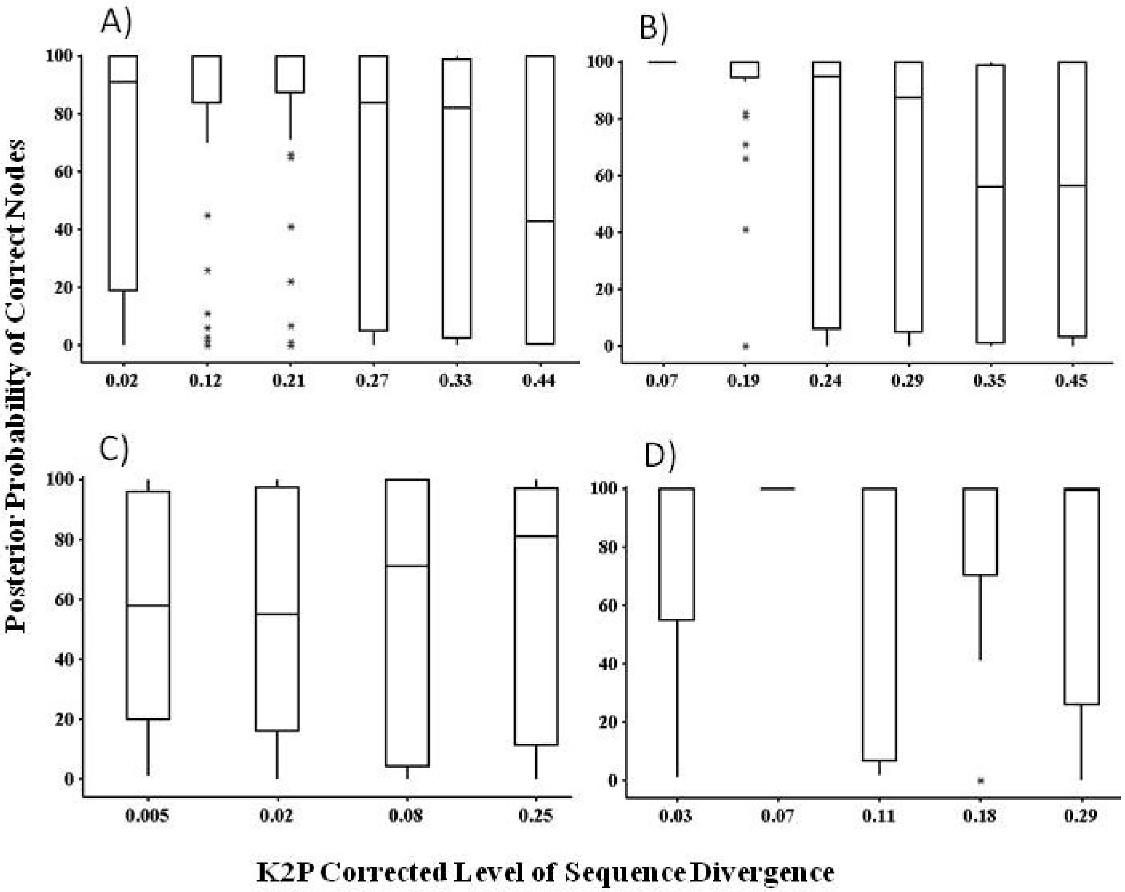

Figure 2.2 Posterior probabilities of correct nodes at different levels of corrected sequence for the natural data sets. Data are presented using boxplots where the center is the median, box edges are the first and third quartiles, whiskers are 1.5 times (third quartile times the first quartile) and stars are outliers. A) All genes combined, B) mt protein coding genes, C) ribosomal genes, D) nuclear protein coding genes.

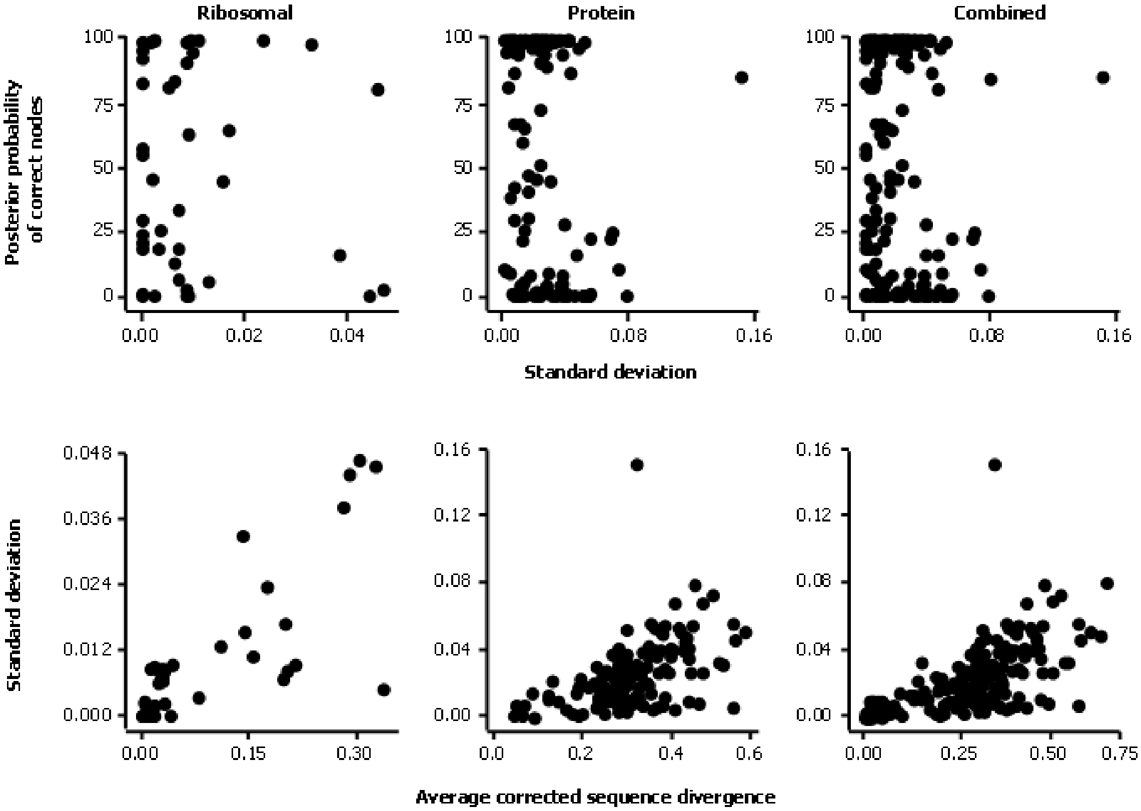

Figure 2.3 Relationship between posterior probability, standard deviation of corrected sequence divergence, and mean divergence of each node for ribosomal genes, protein encoding genes, and all genes combined for the natural data sets.

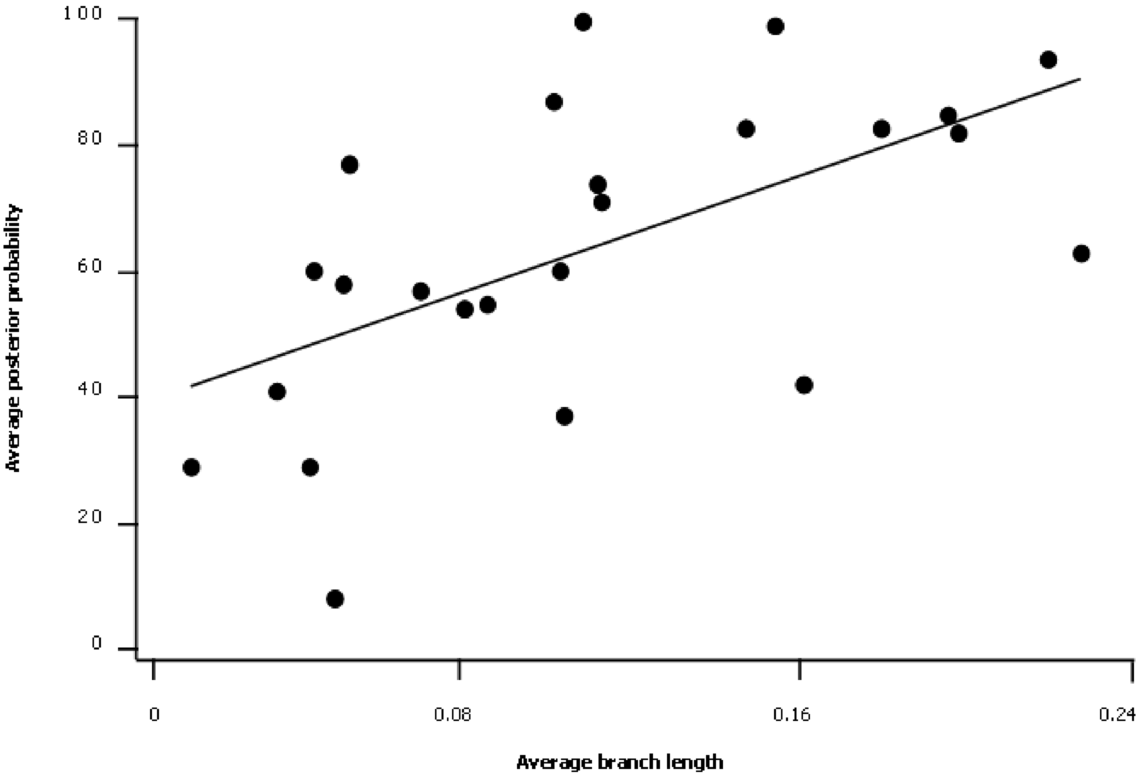

Figure 2.4 Relationship between branch length and posterior probability for the natural data sets. Snake has been removed due to the exceptionally long branch length, but this does not affect the significance of the relationship.

Simulated Data Sets

Simulated data sets recovered the same relationship between posterior probability and sequence divergence as the “known” phylogeny (Fig. 2.5), although the optimal level of divergence differs across data sets. For example, analyses using the data sets produced by JC (A) and GTR (B) models of evolution recovered a broad range of divergences that had high corresponding posterior probabilities. Analyses of the GTR + I + G (C) and ultrametric tree (F) produced a pattern most similar to the one observed in the natural data. The equal branch length data sets analyses recovered the highest level of phylogenetic support across divergence levels, while analyses of the radiation data sets showed the lowest support levels. K-S tests and Mood’s median tests (Table 1) found that the relationship is significant for each simulation category. This suggests that an optimal level of divergence exists regardless of the assumed model of evolution or branch length characteristics, although the optimal ranges of genetic divergence as well as median posterior probabilities are different across the six simulations.

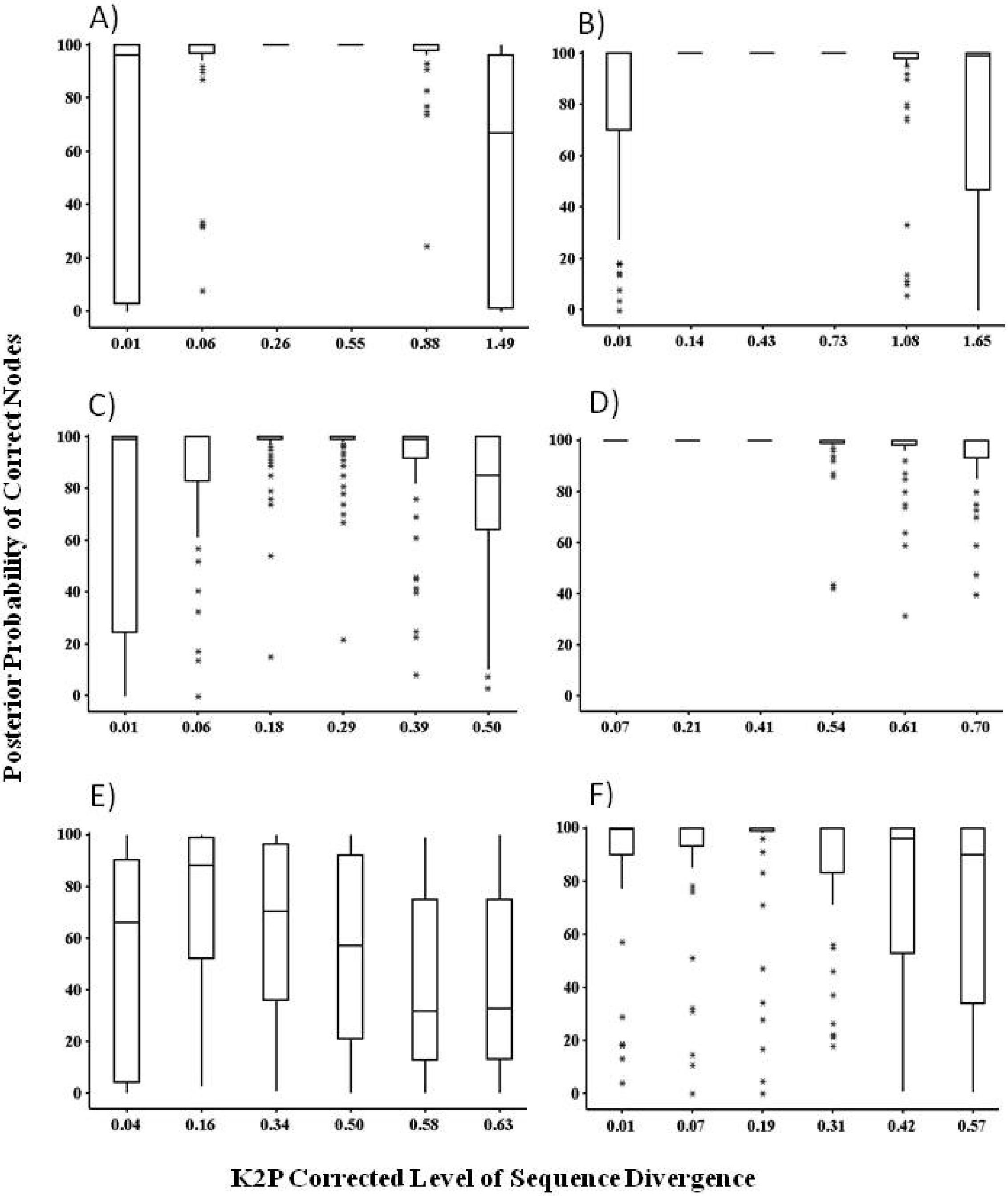

Figure 2.5 Relationship between posterior probability of correct nodes and corrected sequence divergence for simulated data sets. Data sets categories are A) JC, B) GTR, C) GTR + I + G, D) Equal, E) Radiation, and F) Ultrametric.

Discussion

I sought to determine whether one simple criterion, sequence divergence, can reasonably guide gene choice in phylogenetic across a broad scale. Using both natural and simulated data sets, our results show that certain levels of sequence divergence are better at recovering correct phylogenetic relationships than others. Analyses using simulated data sets did not recover the same optimal range of divergence as the natural data sets, but this is most likely due to the simulated data sets not accounting for many realistic facets of molecular evolution. Posterior probabilities of 0.0 percent for "correct" nodes were recovered across all levels of divergence in the natural data sets, so while the sequences at a node may be within the optimal range of sequence divergence, this does not ensure strong support for the correct relationship.

Combining mt, ribosomal and nuclear data into a single analysis, I found an optimal divergence range of approximately 0.12-0.21 K2P corrected (0.09-0.18 uncorrected p, 0.14-0.26 T-N gamma corrected) substitutions per site for the natural data sets. This was determined by calculating the mean level of divergence in the two bins with the best phylogenetic performance. Alternative binning methodologies are possible and each would yield a slightly different answer, but no other strategy appears superior (although most do not seem inferior either). Interestingly, posterior probabilities for correct nodes declined more precipitously with greater divergence from the optimal range than with less divergence. I also analyzed the data by gene category; ribosome, nuclear protein, and mt protein. Protein coding genes, especially mt ones, recover high support for correct relationships even when divergence levels are very low (0.05) and work best at K2P corrected divergences under 0.20 (0.19 for uncorrected p, 0.28 for T-N gamma corrected). This is in sharp contrast with ribosomal genes, which recover similar support values for correct nodes at all divergence levels tested (0.005-0.25 K2P). Interpretations of the nuclear and ribosomal data sets should be withheld until analyses that increase their sample size are increased. A data set that spans more evolutionary time and incorporates more taxa will be necessary to better understand the relationship between sequence divergence and nodal support for ribosomal genes. For nuclear protein genes, a strategy that focuses on organisms with complete, annotated genomes and well-resolved phylogenetic relationships will be necessary.

I used simulated data sets to determine the generality of the observed relationship between posterior probability and sequence divergence seen in the natural data sets. Specifically, I examined phylogenetic performance under differing degrees of model over-parameterization and variations in topology. Neither model over-parameterization nor topology was found to affect the overall relationship between sequence divergence and phylogenetic performance. When the JC, GTR, and GTR + I + G simulations are compared, the major difference is in the level of sequence divergence associated with optimal phylogenetic reconstruction between GTR + I + G and the other two models. This is mostly likely due to the incorporation of a specific number of invariable sites in the analysis, which causes some sites to evolve very quickly and become highly saturated at low levels of sequence divergence. This saturation results in an underestimated level of sequence divergence for the GTR + I + G data sets. For data sets where the model of evolution was held constant (GTR + I + G) and the topology was varied, the same relationship between sequence divergence and phylogenetic reconstruction ability was observed, although the optimal level of divergence varies. Although these results demonstrate that different topologies and levels of model adequacy will have different optimal levels of divergence, such information is rarely if ever truly known even after analyses are complete. If such information could have been accounted for with the natural datasets, I probably would have recovered slightly different optima for each situation; but this information is not available, and even though it was not taken into account, an optimal relationship between sequence divergence and phylogenetic performance was still observed despite confounding variables.

I primarily report results using the K2P correction because the results were the same regardless of the correction model (only the optimal range of divergence changes) and since this model realistically accounts for a variable transition-transversion ratio while not over-parameterizing (Graur and Li, 2000). I acknowledge that substitutions can accumulate in different manners than those accounted for using the distance methods I employed, and that such differences in evolutionary patterns may affect phylogenetic reconstruction. Yet, even though this was ignored in the natural data sets, the information provided by sequence divergence is strong enough to recover an optimal divergence range. Simulated data sets that vary substitution patterns or other parameters (nucleotide bias, transition / transversion ratio, taxon sampling, sequence length, etc) should be used to quantify the effects of each parameter on phylogenetic reconstruction. This would help determine how such evolutionary processes affect phylogenetic reconstruction, although such information is usually not known or is difficult to accurately estimate, especially *a priori*. I feel that our natural data sets approach provides the most applicable and useful estimate of optimal divergence while our simulations show that the observed relationship between nodal divergence and phylogenetic reconstruction ability can be generalized across different topologies and models of evolution.

Divergence optima for phylogenetic reconstruction occur for a variety of reasons. Besides the reasons already discussed in the introduction, I found that as the standard deviation of the pairwise sequence divergences at a node increases, the average posterior probability decreases (Fig. 2.3). Previous researchers have documented that strong deviations from a molecular clock reduce the effectiveness of most phylogenetic reconstruction methods (e.g. Felsenstein, 1983, 2004; Rzhetsky and Sitnikova, 1996), so this may be another reason why analyses using highly divergent sequences recover low posterior probabilities for correct nodes.

Beyond the determination of divergence optima, I also observed two notable patterns involving the relationships among posterior probability, topology, and sequence divergence. First, in the natural data sets, no single gene analyses recovered all of the correct relationships. This is not surprising given the length of evolutionary time (several hundred million years) that our known phylogeny encompasses and the correspondingly large differences in average corrected divergence associated with each node. This is similar to other findings that single gene trees have a very low probability of fully recovering the true relationships (Cao et al., 1994; Rokas et al., 2003). The problem is further confounded when gene trees do not represent species trees, probably reflected in this study as the low recovered posterior probabilities for “correct” nodes when divergences are optimal (Fig. 2.2). Another explanation for low posterior probabilities of nodes within the optimal range is that different genes may saturate at different levels of sequence divergence. For example, if gene A is a trans-membrane gene and gene B is an extra-membrane gene, than gene A probably experiences less selection on the trans-membrane portions, so more of the nucleotide sites can experience changes that are nearly neutral. Therefore, gene A would have more conserved sites, so if the two genes were approximately equal in their levels of sequence divergence, it would be expected that gene B would be more saturated. Such information is usually not known and while the natural dataset analysis ignored such information, I still observed an optimal relationship between sequence divergence and phylogenetic reconstruction (although accounting for such information would probably shift the optimal level of divergence for each gene slightly and reduce the variance).

Second, I found that several nodes consistently exhibited low posterior probability support values across genes, while others consistently exhibited high support values across genes (Table 2.4). Analyses using some mt genes recovered strong support for nodes ca. 300 million years old. The poorly supported nodes (< 90 % PP) were generally “deeper” ones, but not always (e.g. Archonta, Theria, Tertraodontidae). One likely cause is the branch length associated with each node (Rokas and Carroll, 2006; Wiens et al., 2008). Analyses were conducted with snakes included and excluded because of their unusually long branch length, most likely due to their unusually rapid mt evolution (Castoe et al., 2008; Jiang et al., 2007). I found a significant positive correlation between estimated branch length and posterior probability (*P* = 0.022 with snakes; *P* = 0.002 without snakes) in the natural data sets.

One limitation of this work is that the only phylogenetic reconstruction method I tested was Bayesian analysis, using MrBayes software. Other phylogenetic methods, such as parsimony, maximum likelihood as well as other Bayesian programs, are commonly used and should be tested for optimal sequence divergence. I speculate that all likelihood based methods will yield results similar to those of this current study, while parsimony will probably have a lower optimal sequence divergence (because parsimony does not take into account complex models of molecular evolution). Unfortunately, parsimony and maximum likelihood methods do not have nodal support values that are equivalent to posterior probabilities. Regardless of the equivalence (or lack thereof) between posterior probabilities and bootstrap proportions, several studies have demonstrated a correlation between the two values (Cummings et al., 2003; Erixon et al., 2003), at least under some conditions, so I predict that the overall results would be similar.

Another limitation is taxon sampling, which can have an effect on phylogenetic reconstruction methods (Blouin et al., 2004; Heath et al., 2008; Linder et al., 2005; Pollock et al., 2002; Rannala et al., 1998; Zwickl and Hillis, 2002), although the magnitude of this effect is not agreed upon (Rosenberg and Kumar, 2001). For this study (and in other "known" phylogeny based studies (Bull et al., 1993; Hillis and Huelsenbeck, 1994; Rokas et al., 2003; Russo et al., 1996)), taxon sampling is limited. While our "known" phylogeny is limited, there are three obstacles to a more complete phylogeny. First, not all sequences are available for all taxa. Second, and more importantly, I decided that the accuracy of the phylogeny was more important than sampling. For example, the mt genomes for many other salamanders are available, but some of the relationships are contentious and not supported by data other than gene sequences (Bruce, 2005; Chippindale et al., 2004; Mueller et al., 2004; Weisrock et al., 2005; Wiens et al., 2005), so they were excluded. Third, in order to ensure correct alignments, I restricted our data sets to vertebrates, so even though whole genomes have been sequenced from many other taxa (flies, worms, etc), these were excluded from the data sets after determining that they introduced too much ambiguity for most genes.

In conclusion, I have demonstrated an optimal divergence for sequences of approximately 0.12-0.21 K2P corrected pairwise distance yield the highest support for correct nodes. Divergences as low as 0.025 and as high as 0.30 also recovered high support for correct relationships, but divergences over 0.30 show a sharp decline in support for correct nodes. This range ignores topology, model fit, and many important evolutionary parameters since such information is rarely known *a priori*. However, I currently cannot determine if different types of gene (protein or ribosomal) as well as where the gene is encoded (nuclear or mt) may be important factors to take into account. Simulated data sets exhibited the same relationship between sequence divergence and phylogenetic reconstruction ability regardless of topology or model of evolution adequacy. This information has the most utility for relationships that are difficult to resolve, but can also be used for project design to ensure that pairwise sequence dissimilarities stay close to optima. Future work on combining genes (i.e. supertrees) that evolve at different rates so that nodes are weighted towards more optimally divergent levels could also greatly help evolutionary biologists get the most correct information from their data while minimizing confounding information.

CHAPTER 3

PHYLOGEOGRAPHIC ANALYSIS, SUBSPECIES TESTS, AND ECOLOGICAL NICHE MODELING IN THE PLAIN-BELLIED WATERSNAKE, *NERODIA ERYTHROGASTER*

Introduction

How selective pressures and barriers to dispersal impact the evolutionary history of lineages to culminate in divergence and speciation remains a central question in evolutionary biology. Incongruence between morphological and neutral evolutionary patterns may occur when morphology is under relatively strong selection and reflect local adaptations rather than species (i.e. combination of all individual parts of an organism into an individual) evolutionary history (Bonett and Chippindale, 2004; Titus and Larson, 1996; Watts et al., 2004; Wiens and Penkrot, 2002; Wiens et al., 2003). For this reason much debate has occurred concerning character choice (mt, nuclear, fossil, morphological, combined) in phylogenetic analyses (Ballard and Whitlock, 2004; Boore, 2006; Bull et al., 1993; Engstrom et al., 2004; Huelsenbeck et al., 1996; Kluge and Eernisse, 1993; Wiens et al., 2005). Nonetheless, observed incongruence between morphological and genetic data may not actually exist, but instead appears to exist because processes such as incomplete lineage sorting (Maddison, 1997; Pamilo and Nei, 1988).

Wilson and Brown (1953) define subspecies as “genetically distinct, geographically separate populations belonging to the same species and therefore interbreeding freely at the zones of contact.” Historically, taxonomists described subspecies according to predictable morphological variation, assuming a genetic correlation (Burbrink, 2001), a practice Wilson and Brown (1953) cautioned against. Not surprisingly, discrepancies between morphologically defined subspecies and genetics have been documented numerous times (Burbrink et al., 2000; Doukakis et al., 1999; Haig et al., 2006; Walker et al., 1998). Thus, subspecific designations need to be genetically evaluated for their phylogenetic appropriateness.

Nerodia erythrogaster (Forster, 1771), or the plain-bellied watersnake, is one organism where multiple subspecies have been described using few morphological characters. The species ranges from the eastern United States (Fig. 1) into Mexico in the northeastern states of Coahuila, Nuevo Leon, Tamulipas, Durango, and Zacatecas (range in Mexico not shown). Currently, six subspecies are recognized (Gibbons and Dorcas, 2004) and are best distinguished by a combination of range and coloration (Table 1). While the life history, natural history, and ecology of *N. erythrogaster* is relatively well known (see Gibbons and Dorcas (2004) and references within), little genetic data are available to assess its evolutionary history. It has been demonstrated, though, that besides their morphology, some subspecies differ in their common habitats (MacGregor, 1985) and maximal length (see Gibbons and Dorcas (2004) and references within). Four subspecies exist within the United States, two of which (*N. e. flavigaster* and *N. e. transversa*) are easily distinguishable based on coloration and two of which are morphologically similar but vary in geographic range (*N. e. neglecta* and *N. e. erythrogaster*).

Several snake subspecies with similar distributions have been analyzed in a phylogenetic context. Burbrink et al. (2000) examined the North American Rat Snake, *Elaphe obsoleta*, which had eight recognized subspecies, and found no support for any of the subspecies being monophyletic using sequences of cyt *b* and the control region of the mt genome. He did recover three strongly supported monophyletic lineages roughly separated across the MS and Apalachicola rivers. Burbrink (2002) also examined the Cornsnake, *Elaphe guttata*, and found support for one of the five subspecies using cyt *b* sequences and identified the MS River as a major biogeographic barrier (although the Apalachicola River was not). Finally, Guiher and Burbrink (2008) examined cottonmouths (*Agkistrodon piscivorous*) and copperheads (*A. contortrix*) using cyt *b* sequences. They found two well supported cottonmouth lineages and three copperhead lineages, although the Apalachicola and MS River did not appear to be biogeographic barriers in these species.

If distinct genetic lineages are recovered, it is also of interest to determine the extent of ecological differentiation between them. Ecological differentiation can occur in biotic factors (diet, mutualistic interactions, anti-predator defenses) and abiotic factors (habitat preference, thermal tolerance, desiccation tolerance). Quantifying the extent of ecological divergence in biotic factors is not easy to assess. However, measuring the extent of divergence in the ecological niche using georeferenced natural history collection (NHC) data and spatially-explicit climate data has become a common and useful way to assess ecological divergence in abiotic factors. Differentiation in the niche has occurred for many taxa (Parra et al., 2004; Wiens et al., 2006), including snakes (Pyron and Burbrink, 2009; Rissler et al., 2006), and can be viewed as reinforcement mechanisms in zones of potential contact between lineages (e.g. Rissler and Apodaca, 2007).

Our goal is to better understand the evolutionary history of *N. erythrogaster*. Specifically, I aim to 1) provide a rigorous genetic analysis of *N. erythrogaster* throughout its range and determine what, if any, genetic lineages exist using mtDNA; 2) test whether lineages are concordant with the current subspecies taxonomy or possible biogeographic barriers common to other species (MS and Apalachicola River); and 3) assess whether detectable ecological niche differentiation has amassed between lineages.

Methods

Specimens

Tissue was collected from scale clips (photographed and released specimens) or liver samples (euthanized specimens) using IACUC approved protocols (#06-281-1). All tissue samples, photographs, and specimens were deposited into either the University of Alabama Herpetology Collection or the University of TX Amphibian and Reptile Diversity Research Center. Specimens were also obtained through tissue loans from the LA State University Museum of Natural Science, the Museum of Vertebrate Zoology, the

University of TX, the University of Kansas Natural History Museum Center, Alvin Braswell at the North Carolina State Museum, and from several private collections. Outgroups include *Nerodia sipedon, N. taxispilota, N. cyclopion, N. rhombifer, Farancia abacura,* and *Thamnophis sirtalis* whose sequences were downloaded from Genbank when possible. A total of 156 ingroup specimens (Appendix I) from 100 localities were used in this study (Fig. 3.1). All specimens were assigned to subspecies based on collection locality (Gibbons and Dorcas, 2004) that was either geocoded or had detailed locality information associated with the specimen.

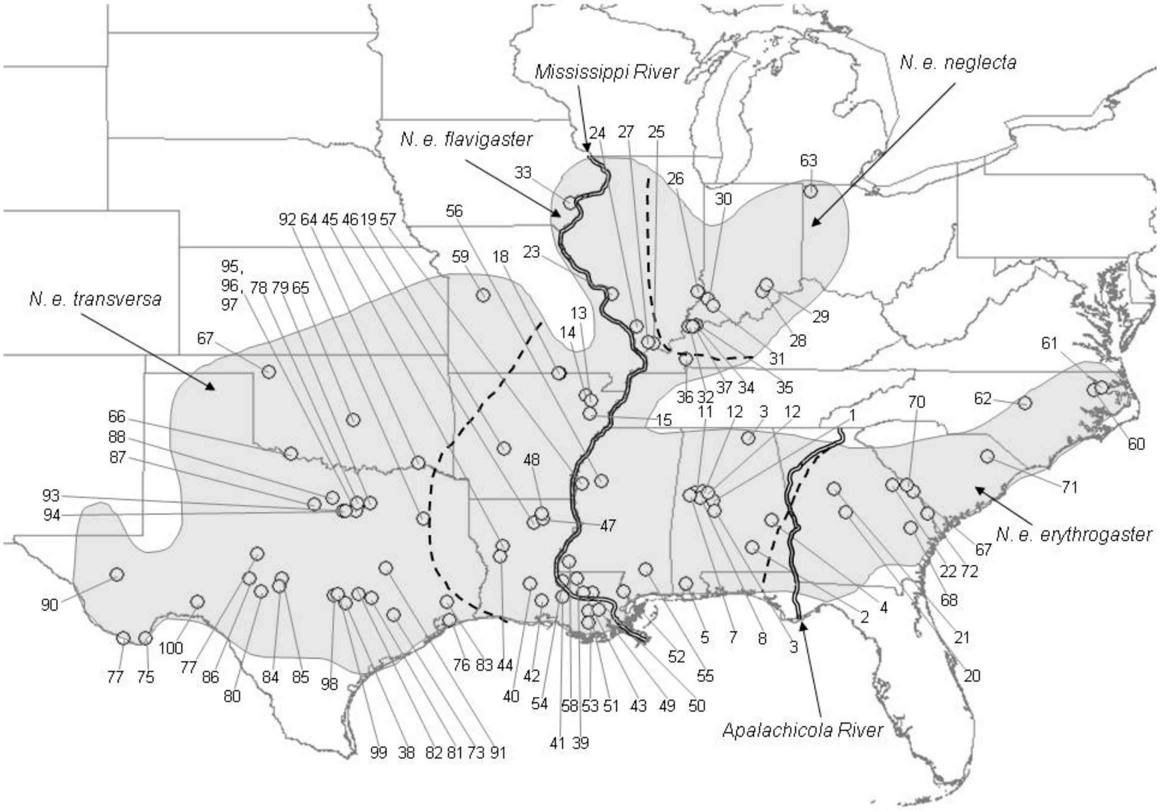

Figure 3.1 Map depicting species range, subspecies breaks, and collecting localities. Numbers correspond to Appendix B.

Sequencing

DNA was extracted from tissue samples using standard extraction protocols (Qiagen Inc., Valencia, CA). Digestion times ranged from three hours for liver samples to 24 hours for scale clips. I obtained partial sequences of cytochrome *b* (cyt *b*), nicotinamide adenine dinucleotide subunit II (NADH II), and cytochrome oxidase I (Cox-I) mt genes and the nuclear protoncogene (C-mos) and recombination activating gene (Rag-1).

Cyt *b* conditions consisted of an initial denaturation at 94 C for 3 minutes followed by 35 cycles of 94 C for 15 seconds, 46 C for 30 seconds, and 72 C for 90 seconds and a final extension at 72 C for 7 minutes (Forward= 5' CCA GTA GGA CTA AAC ATT TCA ACC TCA ACC TGA TGA 3'; reverse= 5' TGG TGT TTC TAC TGG TTT TGT GGC TGA GGC TGA TCA 3'). NADH II, Cox-1, C-mos and Rag1 conditions were the same as cyt *b* except the annealing temperature was 55.5 C for NADH II (Forward= 5' CGC AAC AAA ATA CTA CCT CAC CC 3'; reverse= 5' GAT TTT ATT GGT GTG AGT GTG GTG TG 3'), 52.0 C for Cox-I (Forward=5' TCA GCC ATA CTA CCT GTG TTC A 3'; reverse= 5' TAG ACT TCT GGG TGG CCA AAG AAT CA 3') and 53.2 C for C-mos (Forward= 5' CAT GGA CTG GGA TCA CTT ATG 3'; reverse= 5' CCT TGG GTG TGA TTT TCT CAC CT 3') and 52.0 C for Rag1 (Wiens et al., 2008).

PCR samples were cleaned by gel extraction (Qiagen Inc., Valencia, CA) or ExoSapIt (United States Biochemical, Cleveland, OH) and either sent to Macrogen (Korea) for sequencing or sequenced on an ABI 3130 sequencer (Applied Biosystems, Foster City, CA) using an ABI recommended protocols. All samples were sequenced in both directions. Forward and reverse sequences were compared in Sequencher 4.6 (Gene

Codes Corporation) and the consensus sequences were aligned using ClustalW (Chenna et al., 2003; Larkin et al., 2007) in Macvector 9.0 (MacVector Inc., Cary, NC). No gaps were found so alignment was unambiguous.

Phylogenetic Analyses

C-mos and Rag-1 were found to be less than 0.5% variable (i.e. segregating) for 20 range-wide specimens and were therefore excluded from further sampling and analyses due to lack of variability (see Chapter II for explanation). Aligned cyt *b* sequences totaled 837 bases, 665 bases for NADH II, and 627 bases for Cox-I. Each data set was analyzed separately and in a combined analysis. To obtain an appropriate model of evolution for each gene, an NJ tree was used in Modeltest 3.7 (Posada and Crandall, 1998) to determine the best model of evolution using AIC. PAUP* 4.0 beta (Swofford, 1999) was used to run maximum likelihood (ML) analyses while TNT (Goloboff et al., 2008) was used in the parsimony (MP) analyses. Both MP and ML analyses were run with random addition sequences and TBR swapping with 10 repetitions. Modeltest 3.7 was rerun using the best ML tree as the starting tree to optimize the model of evolution. These parameters were used in a subsequent ML analysis and the new best tree score was compared to the previous best tree score. This process was repeated until the best tree likelihoods in sequential iterations were equal. Bootstrap support was conducted with 1000 pseudoreplicates for MP and ML analyses (conducted in Garli 0.942 (Zwickl, 2006)).

The best ML tree for each gene was used to determine the appropriate model of evolution for Bayesian analyses in MrModelTest 2 (Nylander, 2004). These parameters were input into MrBayes 3.1.2 (Huelsenbeck and Ronquist, 2001; Ronquist and Huelsenbeck, 2003) for up to 30,000,000 generations with four chains, two repetitions, four swaps per generation, sampling every 1,000 generations, and partitioned by codon position. Analyses were run until the standard deviation of split frequencies was < 0.05 and convergence was further checked in AWTY (Nylander et al., 2008). The first 2.5 to 5.0 million generations were discarded as burnin.

Total evidence analyses were run using MP and ML by concatenating the three genes and using the same methods as above. For the Bayesian TE analysis, genes were concatenated and analyzed partitioned by gene and position. *A priori* hypotheses (see below) were tested by running TE analyses in MrBayes 3.1.2 with topological constraints under the same conditions as above and compared using Bayes factors.

Statistical tests

To test whether subspecies (assigned based on specimen locality) are monophyletic, I constrained each subspecies as monophyletic in separate Bayesian analyses. Since two of the populations sampled are in close proximity to the *N. e. erythrogaster / N. e. flavigaster* boundary (Fig. 3.1; populations 2 and 4), I conducted four separate analyses where the two populations were assigned in all possible combinations. Specifically, hypothesis one (*e*1) excluded both populations from the *N. e. erythrogaster* clade, hypothesis two (*e*2) included both populations, hypothesis three (*e*3) excluded population 4 and included population 2, and hypothesis four (*e*4) excluded population 2 and included population 4. I also constrained *N. e. neglecta* and *N. e transversa* as monophyletic and tested the phylogeographic hypotheses by constraining populations east and west of the MS and Apalachicola (same as hypothesis *e*1 above) rivers as monophyletic. The eight *a priori* hypotheses of monophyly were compared to the best TE tree using Bayes factors. I followed the methods espoused by Kass and Raftery (1995) which have been implemented in multiple phylogenetic analyses (Brandley et al., 2005; Nylander et al., 2004; Palero et al., 2009). I considered H_0 to be that the *a priori* hypotheses explain the data as well as the best tree while H_1 assumed that constrained searches provide a poorer fit for the data. Bayes factors were calculated as twice the difference of -ln likelihood harmonic means between competing hypotheses using the harmonic mean output in MrBayes. I interpreted Bayes factor values <0 as evidence against H_1, while positive values provide either basically no evidence for H_1 (0-2), positive support for H_1 (2-6), strong support for H_1 (6-10), or very strong support for H_1 (>10).

I also performed a principal coordinates analysis (PCoA) in GenAlEx v.6.1 using the Apalachicola River (similar to hypotheses one above) and MS River as population delineators, biogeographic breaks reported in several other species (Soltis et al., 2006; Swenson and Howard, 2005). PCoA is similar to principal components analysis (PCA) except that it uses discrete rather than continuous data. I specified codominant data and the PCoA was calculated under the standardized covariance settings. I then used discriminate analysis to quantify genetic divergence between subspecies by examining the correct assignment proportion during cross validation with the PCoA factors.

Ecological Niche Modeling Methods

Ecological niche models for well supported clades and subspecies were created using Maxent version 3.2.19 (Phillips et al., 2006) implementing 19 climatic layers (exclusively precipitation and temperature parameters) downloaded from the WorldClim database (http://www.worldclim.org/) at 30 sec resolution. To test whether clades or subspecies were associated with unique environmental niche space, I extracted the spatially explicit climate data at each point locality (Appendix I) using DIVA version 5.2.0.2. Principal components analysis (PCA) on the covariance matrix was used to reduce the number of climatic variables and PCA axis scores (the ones needed to account for >90% of the variability) were then entered as the dependent variable in a multivariate analysis of variance (MANOVA) with clade or subspecies as the fixed factor. Normality and variance assumptions were checked by examining residuals. I then used discriminant analyses to quantify ecological divergence between clades and subspecies by examining the correct assignment proportion during cross validation with the PCoA factors. In order to assess whether genetic distance was positively correlated with environmental distance while controlling for geographic distance, I used partial Mantel tests in R-package 4.0 (Casgrain and Legendre, 2001). Significant results suggest that phylogenetic breaks are correlated with (or potentially caused by) environmental gradients. Genetic distances were calculated in Paup* 4.0 beta using the model of evolution determined by

Modeltest 3.7 (Posada and Crandall, 1998) for the combined data sets and tree reported in Fig.3. 2. Ecological distances were based on Euclidean distances of the PCA factor scores.

Results

Individual genes yielded similar phylogenetic hypotheses, so only the total evidence analyses are reported. Two-thousand one and twenty nine bases of combined data yielded 43 unique haplotypes and 177 variable sites, of which 98 were parsimony informative for the ingroup. The model of evolution (for likelihood searches) for all genes together was TIM + I + G: Base frequencies of A = .3244, C = .3172, G = .1103, T = .2480; substitution rate parameters A-C = 1.0000, A-G = 16.5537, A-T = 1.3404, C-G = 1.3404, C-T = 9.2494, G-T = 1.0000; proportion of invariable sites = 0.5866; gamma distribution shape parameter = 1.2360. The model chosen for each gene in the Bayesian analysis was GTR + I + G. MP analysis yielded 80 equally parsimonious trees with 946 steps. The best tree for ML had a –*ln* likelihood score of 7,986.55.

Bayesian and ML phylogenetic analyses yielded similar topologies, so only the Bayesian tree is reported (Fig. 3.2). There was little resolution in the MP consensus tree, even when a 50% majority rule was calculated. Overall support for nodes was minimal, but several notable clades were significantly supported (Fig. 3.2). I recovered five separate lineages, only one of which (Eastern lineage) was even closely concordant with taxonomy (*N. e. erythrogaster*; Fig. 3.3). Interestingly, this lineage is also closely concordant with the Apalachicola River phylogeographic break. To determine if a monophyletic *N. e. erythrogaster* subspecies is much less likely than the best tree, I computed the Bayes factors for each hypothesis described above. All comparisons between *a priori* hypotheses and the recovered "best" tree resulted in Bayes factors greater than 45 (Table 3.1), signifying very strong evidence against any of the *a priori* hypotheses. The other genetic lineages recovered did not match predictions based on phylogeographic or taxonomic predictions.

Table 3.1 Hypotheses of monophyly, their corresponding likelihood scores, and the Bayes factor associated with each *a priori* hypothesis compared to the best tree.

Hypothesis	-ln likelihood harmonic mean	Bayes factor	Support against *a priori* hypothesis
*e*1 (Apalachicola River)	-8084.9	199.6	Very strong
*e*2	-8096.9	223.6	Very strong
*e*3	-8022.5	74.8	Very strong
*e*4	-8033.6	97.0	Very strong
neglecta	-8007.8	45.4	Very strong
transversa	-8179.7	389.2	Very strong
flavigaster	-8166.9	363.6	Very strong
MS River	-8094.87	219.54	Very strong
Best tree	-7985.1	-	-

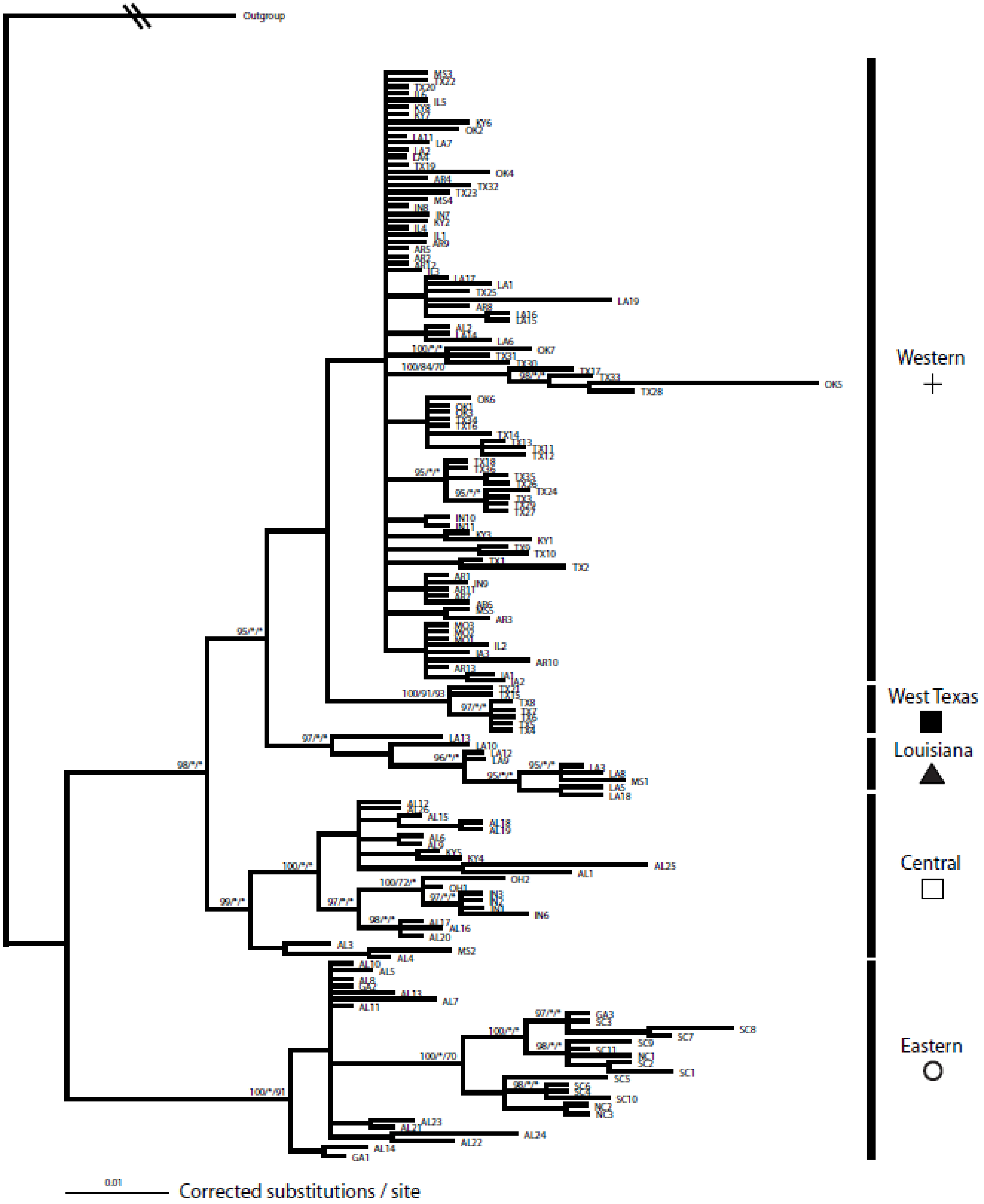

Figure 3.2 Bayesian phylogram produced using a total evidence approach with cyt *b*, NADH II, and COX I. Numbers above nodes correspond to the posterior probability / maximum likelihood bootstrap proportion (BP) / parsimony BP. Outgroups have been collapsed and their branch length shortened.

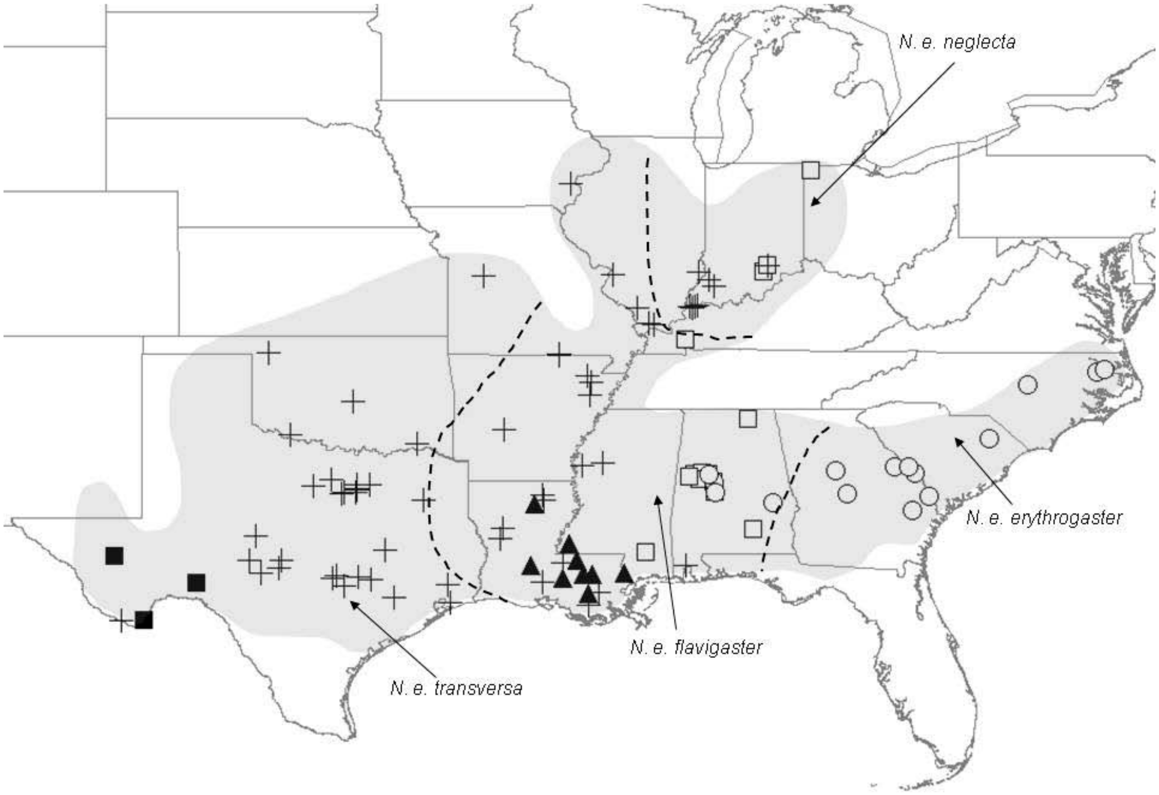

Figure 3.3 Map depicting species range, subspecies breaks, and recovered clade for each population. See Fig. 3.2 for symbol definitions.

The principal coordinates analysis (PCoA), where individuals were grouped by subspecies, recovered three separate groups (Fig. 3.4). Two of these are amalgamations of at least three recognized subspecies, but the third consists of only three haplotypes. When PCoA factors were used in the discriminant analysis with subspecies as the grouping factor, the analysis correctly assigned specimen to their subspecies only 53.2% of the time (Table 3.2), although the variation in assignment ability across subspecies is large.

Table 3.2 Discriminant function analysis results using mt sequence characters with geographically defined subspecies as the grouping factor. Overall, 83 (53.2 %) specimens were grouped correctly.

True Subspecies				
Put into subspecies	*erythrogaster*	*flavigaster*	*neglecta*	*transversa*
erythrogaster	17	11	0	0
flavigaster	0	23	2	1
neglecta	0	3	5	7
transversa	0	35	12	40
Total #	17	72	21	46
# correct	17	23	5	38
Proportion	1.00	0.319	0.263	0.869

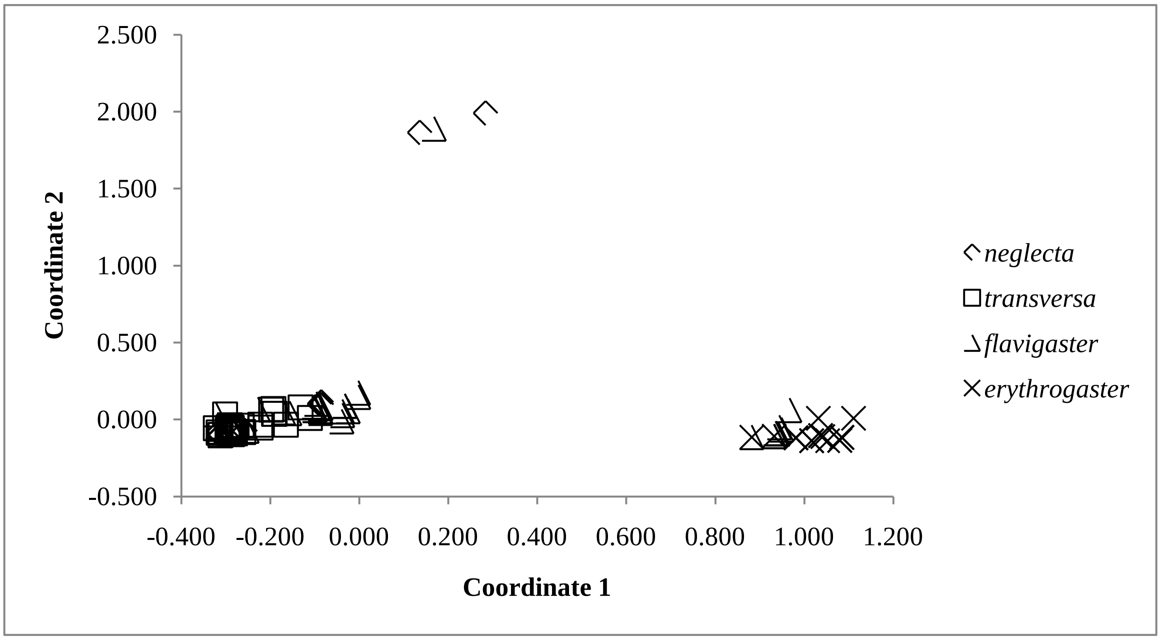

Figure 3.4 PCoA results with geographically defined subspecies as the grouping factor.

Ecological niche models (ENMs) for the five clades and subspecies recovered a large amount of over-prediction (not shown). These results are confirmed with the PCA (Fig. 5), which also showed little separation of groups. However, the environmental conditions varied significantly across clades (Wilks' Lambda=0.20989, d.f.=16,452, P<0.05) and the discriminant analysis was able to correctly assign specimens to their group 66.0% of the time (Table 3.3). Environmental conditions also varied significantly across subspecies (Wilks' Lambda=0.11622, d.f.=12,391, P<0.05) and discriminate analysis correctly assigned groups 78.2% of the time (Table 3.4). The partial Mantel test found no significant relationship between genetic and environmental divergence (r=0.0467, P = 0.162). Due to the low support for most clades as well as their high level of sympatry, printed bioclimatic models were restricted to the eastern and “non-eastern” clades (Fig. 3.6).

Table 3.3 Discriminant function analysis results using environmental data with recovered clades as the grouping factor. Overall, 103 (66.0 %) specimens were grouped correctly.

True Subspecies				
Put into subspecies	*erythrogaster*	*flavigaster*	*neglecta*	*transversa*
erythrogaster	14	10	0	3
flavigaster	3	47	0	0
neglecta	0	14	21	3
transversa	0	1	0	40
Total #	17	72	21	46
# correct	14	47	21	40
Proportion	0.824	0.653	1.00	0.869

Table 3.4 Discriminant function analysis results using environmental data with geographically defined subspecies as the grouping factor. Overall, 122 (78.2 %) specimens were grouped correctly.

True Group					
Put into clade	Central	Eastern	LA	West TX	Western
Central	19	12	1	0	11
Eastern	2	11	3	0	0
LA	0	0	0	6	9
West TX	0	0	0	6	9
Western	2	0	0	1	62
Total #	23	27	9	7	90
# correct	19	11	5	6	62
Proportion	0.826	0.407	0.556	0.857	0.689

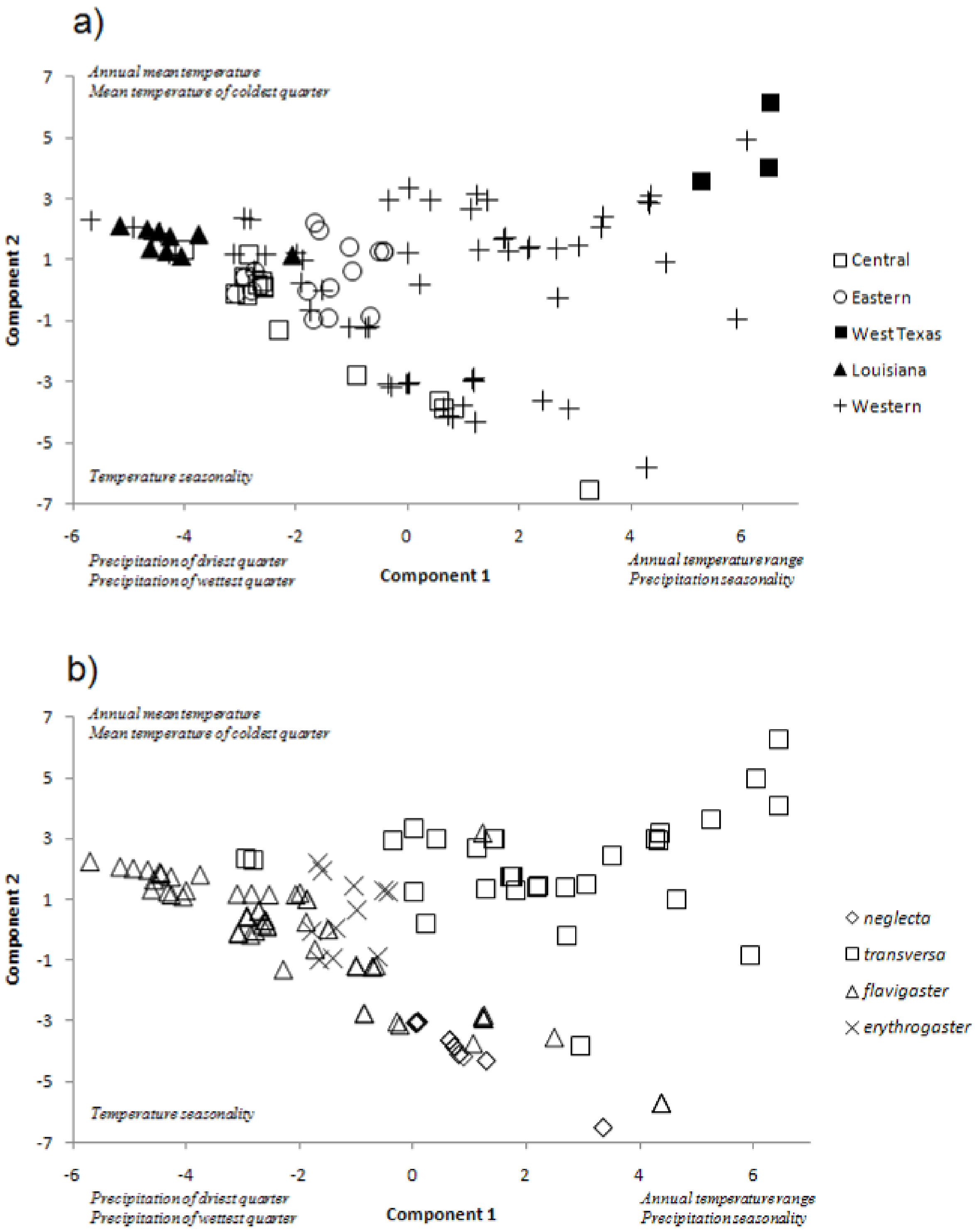

Figure 3.5 Scatterplot of first two principal components from the bioclimatic data with grouping based on a) clade and b) taxonomy.

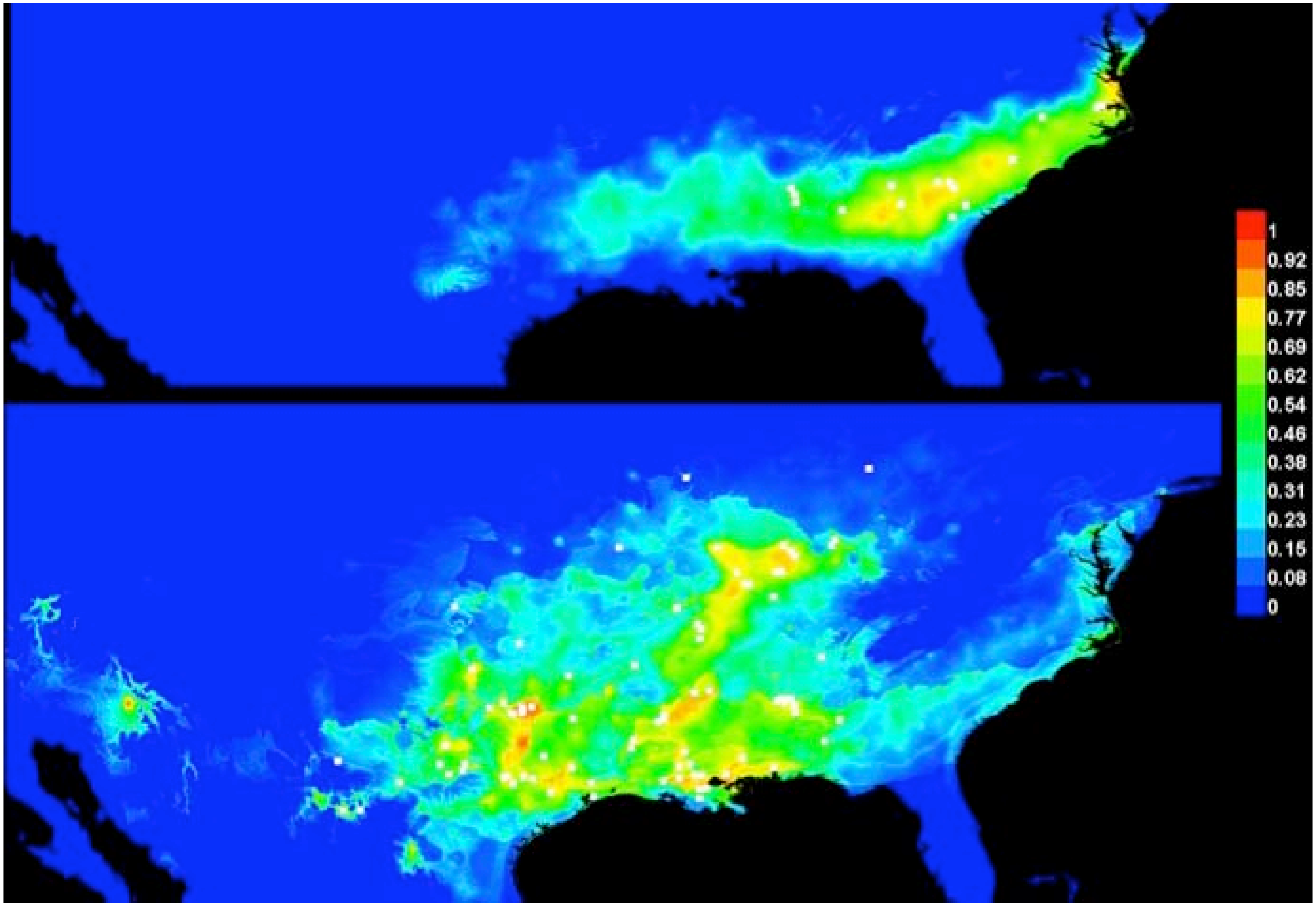

Figure 3.6 Ecological niche models for the eastern clade (top) and non-eastern clade (bottom) using 19 WorldClim data layers. Locality points are detailed with clade identification symbols, defined in Fig. 3.2.

Discussion

Our goal was to better understand the evolutionary history of *Nerodia erythrogaster* by determining what genetic lineages exist using mtDNA, testing whether monophyletic lineages are concordant with the current taxonomy or common biogeographic barriers, and assessing whether detectable ecological niche differentiation has amassed between lineages. Our results indicate the *N. erythrogaster* is composed of five genetic lineages, all of which are partially to completely sympatric with at least one other lineage. All *a priori* hypotheses, both taxonomic and biogeographic, were rejected when compared to the recovered genetic tree and subspecies showed little genetic divergence. Finally, I found that the recovered genetic lineages showed little ecological differentiation.

To recover the phylogenetic history of the species, I used partial sequences from three mt genes. Given that these genes are linked, it would have been preferable to include nuclear genes; however, the low levels of divergence observed at the mt level make most nuclear data uninformative for phylogenetic analyses (as confirmed by the C-mos and Rag-1 data). Also, because of the maternal inheritance characteristic of mt genes, my interpretation of the data assumes relatively equal dispersion of the sexes, although this has not been examined. Because analyses from individual genes yielded very similar trees, I combined the genes into a total evidence analysis. I recovered five clades, but none was strongly supported (Fig. 2) or geographically isolated (Fig. 3). Only one clade, the “Eastern,” was mostly concordant with any of the biogeographic or taxonomic *a priori* hypotheses, but Bayes’ factors strongly supported the recovered phylogenetic tree compared to the *a priori* hypotheses. Bayes' factors provide a useful alternative to the classic null hypothesis test (e.g. Shimodaira-Hasegawa (SH test), Swofford-Olsen-Waddell-Hillis test (SOWH)) where instead of testing a null, support for differing, meaningful hypotheses are compared directly (Kass and Raftery, 1995). This allowed me to compare five *a priori* hypotheses directly to the tree recovered. The PCoA of the total evidence data sets with taxonomic grouping produced an unexpected plot, with three separate groups, two of which are made up of at least three of the four subspecies. Discriminant analyses performed poorly at predicting subspecies during cross validation overall (Table 2), signifying a limited amount of genetic differentiation between subspecies. Marshall et al. (2009), who focused on microsatellite variation in *N. e. neglecta*, recovered moderate differentiation among the regions sampled and surmised that the differentiation among populations is due to the quality of terrestrial dispersal corridors.

Even though phylogenetic lineages were weakly supported, I looked for evidence of ecological differentiation using environmental information for the identified clades. Although there were statistically significant differences among the clades, there was no clear indication that particular genetic lineages were in unique environmental niche space (Fig. 3.5). Due to the lack of concordance between evolutionary lineages (mtDNA) and taxonomy based on phenotype (subspecies), I also looked for ecological differentiation of specimens assigned to subspecies based on locality. If such differences were recovered, this provides support for the phenotypic differences being attributable to plastic environmental responses. I followed the boundaries defined by Gibbons and Dorcas

(2004) while assigning all specimens from Alabama to *N. e. flavigaster*. The results were very similar to the analysis where groups were defined by phylogenetic clades as opposed to subspecies; a small degree of statistically significant differentiation among some subspecies, but little overall ecological differentiation (Fig. 3.5). It is possible that strong ecological differentiation does exist, but the appropriate environmental variables are not being included in the model. Since the species exists across such a large range in and across diverse ecological conditions, there are many ecological variables that could be important that are not available to include in the model (diet, water pH, salinity, soil composition, etc). However, at this time, neither the genetic lineages nor subspecies are associated with distinct ecological environments as is common in other reptiles and amphibians (Graham et al., 2004; Raxworthy et al., 2007; Wiens et al., 2006)

The lack of any geographically separated lineages across such a wide-ranging species is surprising. Studies of many diverse taxa using mt DNA (mtDNA) in North America have recovered strongly supported geographically distinct lineages (Griffin and Barrett, 2004; Heilveil and Berlocher, 2006; Joly and Bruneau, 2004; Roe et al., 2001). This is not always the case, though, with many other instances probably not reported due to “non-significant results.” Examples of wide ranging North American species with low genetic differentiation include the diamondback watersnake, *Nerodia rhombifer* (Matthew Brandley, pers. comm.), the eastern narrow-mouthed toad, *Gastrophryne carolinensis* (Makowsky et al., 2009), Blanchard’s cricket frog, *Acris blanchardi* (Gamble et al., 2008), snapping turtles (Walker et al., 1998), and many boreal mammals (Arbogast and Kenagy, 2001). mtDNA is the most common genetic marker in phylogeographic analyses due to its relatively fast lineage sorting and assumed neutrality, but mt genes have important functions and their linkage precludes them all to being equally susceptible to selective sweeps on any single gene. *Nerodia* is a particularly young genus, with most recovered fossils placed in the Pleistocene and Pliocene, and the oldest fossil approximately 13 million years old (see Gibbons and Dorcas, 2004; and references within). The most current phylogenies of North American natricines hypothesize rapid diversification of most species, so the age of *N. erythrogaster* is probably close to the age of the genus. Climate shifts since the speciation of *N. erythrogaster,* especially the most recent glacial maximum, have undoubtedly increased gene flow across populations, decreased population structure, and left a genetic signature that is difficult to unravel.

Taxonomically, I find no strong support for splitting *N. erythrogaster* into multiple species. Based on Wilson and Brown's (1953) definition, none of the subspecies is valid, although five evolutionary significant units (ESU) were recovered. The eastern clade recovered the strongest phylogenetic support, and may provide exciting further research on speciation mechanisms. For example, what are the properties of the contact zone between red and yellow belly color? Contact zones can be associated with different biotic or abiotic conditions (Barton and Hewitt, 1985; Mayr, 1954), such as ecotones (Rosenblum, 2006; Rosenblum et al., 2004). Quantifying such dynamics would allow insight into the evolution of a starkly contrasting trait that is as yet unexplained. Further studies that explicitly test to what degree the color trait is environmentally controlled will also be necessary to completely understand why the coloration of *N.*

erythrogaster is so variable. Because single populations along the contact zone between *N. e. erythrogaster* and *N. e. flavigaster* are sometimes composed of individuals with obvious ventral color differences, the likelihood of a purely environmental cause is low.

In this study, I sought to elucidate the evolutionary history of *N. erythrogaster* by using mtDNA to test whether genetic lineages are concordant with the current taxonomy or probable biogeographic barriers. Using a combination of molecular and environmental evidence, I conclude that none of the subspecies is genetically distinct. Given the range size of the species, the biogeographic barriers over which they cross, and the differing morphologies across the range, this is a surprising finding. Because the recovered lineages are not geographically isolated, I do not feel that elevation of any lineages to species status is warranted due to identification issues. Therefore, I conclude that *N. erythrogaster* is a single species with multiple, geographically defined varieties.

APPENDIX A

GENBANK ACCESSION NUMBERS FOR SEQUENCES USED IN THE REAL DATA SETS. PRIMARY OTUS ARE DEFINED IN THE MT DATA SET. FOR INSTANCES WHERE THE PRIMARY OTU WAS SUBSTITUTED WITH A CLOSELY RELATED SPECIES, THE SUBSTITUTED SPECIES IS FOLLOWED BY THE PRIMARY OTU IN PARENTHESES.

Mitochondrial Genes – *Agkistrodon contortrix*: NC_009768; *Alligator mississippiensis*: NC_001922; *Ambystoma mexicanum*: AJ584639; *Bos indicus*: NC_005971; *Branchiostoma japonicum*: NC_008069; *Bufo japonicus*: NC_009886; *Callorhinchus milii*: NC_001606; *Canis lupus*: NC_009686; *Cyprinus carpio*: NC_001606; *Deinagkistrodon acutus*: DQ343647; *Didelphis virginiana*: NC_001610; *Eschrichtius robustus*: AJ554053; *Eurycea bislineata*: NC_006329; *Felis cattus*: NC_001700; *Gallus gallus*: NC_001323; *Mus musculus*: EF108345; *Mustelus manazo*: NC_000890; *Neofelis nebulosa*: NC_008450; *Oncorhynchus clarkii*: NC_006897; *Oncorhynchus mykiss*: DQ288271; *Pan troglodytes*: NC_001643; *Plethodon cinereus*: NC_006343; *Plethodon petraeus*: NC_006334; *Rana plancyi*: NC_009264; *Rattus norvegicus*: AJ428514; *Sceloporus occidentalis*: NC_005960; *Takifugu rubripes*: AJ421455; *Tetraodon nigroviridis*: NC_007176; *Xenopeltis unicolor*: NC_007402; *Xenopus tropicalis*: NC_006839

RAG 1 – *Alligator mississippiensis (Alligator sinensis)*: AY239171; *Aneides ferreus (Plethodon cinereus)*: EU275805; *Aneides lugubris (Plethodon petraeus)*: EU275807; *Boa constrictor (Agkistrodon contortrix)*: AY988064; *Bos taurus (Bos indicus)*: AF447520; *Bufo balearicus (Bufo japonicus)*: EU497605; *Canthigaster janthinoptera (Tetraodon nigroviridis)*: AY700366; *Cyprinus carpio*: EF458304; *Dicamptodon tenebrosus (Ambystoma mexicanum)*: EU275789; *Eryx conicus (Deinagkistrodon acutus)*: AY988074; *Lutrogale persp* *icillata (Felis cattus)*: EF472410; *Monodelphis theresa (Didelphis virginiana)*: DQ865914; *Mustela frenata (Mustelus manazo)*: EF472412; *Mus nitidulus (Mus musculus)*: AB262426; *Negaprion brevirostris (Callorhinchus milii)*: AY949031; *Pionopsitta barrabandi (Gallus gallus)*: DQ143349; *Propithecus tattersalli (Pan troglodytes)*: EU342327; *Physeter catodon (Eschrichtius robustus)*: EU189408; *Rattus exulans (Rattus norvegicus)*: DQ023455; *Salvelinus malma (Oncorhynchus mykiss)*: AY380535; *Smilisca baudinii (Rana plancyi)*: DQ830932; *Takifugu rubripes*: AY700363; *Xenopus borealis (Xenopus tropicalis)*: EF535912

Taxa omitted: *Branchiostoma japonicum*

BDNF – *Aneides flavipunctatus (Ambystoma mexicanum)*: EU275895; *Ascaphus truei (Rana plancyi)*: EU275896; *Batrachoseps sp. (Plethodon petraeus)*: EU275901; *Bolitoglossa sp. (Plethodon cinereus)*: EU275897; *Bos Taurus (Bos indicus)*: NM_001046607, XM_870009; *Canis lupus*: NM_001002975, XM_534099; *Cyclophiops sp. (Agkistrodon contortrix)*: AF497715; *Danio rerio (Cyprinus carpio)*: BC058301; *Dicentrarchus labrax (Oncorhynchus mykiss)*: DQ915807; *Equus caballus (Eschrichtius robustus)*: AB264324; *Gallus gallus*: DQ124361; *Helarctos malayanus (Neofelis nebulosa)*: AF002240; *Homo sapiens (Pan troglodytes)*: NM_170735; *Japalura splendida (Sceloporus occidentalis)*: AF497713; *Monodelphis domestica (Didelphis virginiana)*: XM_001368353; *Mus musculus*: NM_007540; *Paralichthys olivaceus (Oncorhynchus clarkii)*: AY074888; *Rattus norvegicus*: NM_012513; *Taeniopygia* *guttata (Alligator MSensis)*: NM_001048255; *Ursus arctos (Neofelis nebulosa)*: AF002239; *Xenopus laevis (Xenopus tropicalis)*: EF035623

Taxa omitted: *Ambystoma mexicanum, Branchiostoma japonicum, Callorhinchus milii, Bufo japonicus, Deinagkistrodon acutus, Tetraodon nigroviridis, Takifugu rubripes*

18 S – *Alligator mississippiensis*: AF173605; *Atelopus flavescens (Bufo japonicus)*: EF364368; *Coturnix coturnix (Gallus gallus)*: EU236695; *Cricetulus sp. (Mus musculus)*: M33067; *Cyprinus carpio*: AF133089, U87963; *Heterodon platyrhinos (Agkistrodon contortrix)*: M59392, M36351; *Homo sapiens (Pan troglodytes)*: K03432; *Hyla chrysoscelis (Rana plancyi)*: AF169014; *Malpolon moilensis (Deinagkistrodon acutus)*: EF198105; *Monodelphis domestica (Didelphis virginiana)*: AJ311676; *Plethodon yonahlossee (Plethodon cinereus)*: M59397, M36356; *Rattus norvegicus*: X01117 K01593; *Salmo trutta (Oncorhynchus mykiss)*: DQ009482; *Scincus scincus (Sceloporus occidentalis)*: EU236693; *Sus scrofa (Eschrichtius robustus)*: AY265350; *Thymallus baicalensis (Oncorhynchus clarkii)*: AM492690; *Tetraodon nigroviridis*: AJ270032; *Xenopus laevis (Xenopus tropicalis)*: X04025

Taxa omitted: *Ambystoma mexicanum, Branchiostoma japonicum, Callorhinchus milii, Canis lupus, Eurycea bislineata, Felis cattus, Neofelis nebulosa, Plethodon petraeus, Takifugu rubripes*

28 S – *Alligator mississippiensis (Alligator sinensis)*: DQ283650; *Ambystoma macrodactylum (Ambystoma mexicanum)*: AF212178; *Anolis carolinensis (Sceloporus occidentalis)*: AY859623; *Bos taurus (Bos indicus)*: DQ222453, AY779625, AY779626, AY779627, AY779628, AY779629; *Branchiostoma floridae (Branchiostoma japonicum)*: AF061796; *Bufo amboroensis (Bufo japonicus)*: DQ283701; *Centroscymnus owstonii (Callorhinchus milii)*: AY049821; *Eudiplozoon nipponicum (Cyprinus carpio)*: AF382037; *Eurycea wilderae (Eurycea bislineata)*: DQ283615; *Gallus gallus*: DQ018757; *Lagocephalus laevigatus (Takifugu rubripes)*: AY141601; *Macropus eugenii (Didelphis virginiana)*: EF654517; *Mus musculus*: X00525; *Oncorhynchus mykiss*: U34341; *Plethodon dunni (Plethodon cinereus)*; DQ283620; *Plethodon jordani (Plethodon petraeus)*: DQ283521; *Rana palmipes (Rana plancyi)*: DQ283699; *Pan troglodytes*: M30950; *Rattus norvegicus*: V01270, X00133, X00521, X01069; *Salvelinus namaycush (Oncorhynchus clarkii)*: U17962; *Tetraodon nigroviridis*: AJ270040; *Xenopus borealis (Xenopus tropicalis)*: X59733

Taxa omitted: *Agkistrodon contortrix, Canis lupus, Deinagkistrodon acutus, Eschrichtius robustus, Neofelis nebulos,*

APPENDIX B

INFORMATION ON ALL INGROUP SPECIMENS, LOCALITIES, AND GENBANK ACCESSION NUMBERS. MUSEUM CODES: LSU=LA STATE UNIVERSITY MUSEUM OF NATURAL SCIENCE, MVZ=MUSEUM OF VERTEBRATE ZOOLOGY, THE UNIVERSITY OF TX, AND KU=UNIVERSITY OF KANSAS NATURAL HISTORY MUSEUM CENTER. PERSONAL COLLECTION: RM=ROBERT MAKOWSKY, ALB=ALVIN BRASWELL, MCB=MATTHEW C. BRANDLEY, CW=CHRIS WINNE, JM=JOHN MARSHALL, JDM=JOHN MCVAY, MN=MATTHEW NORDGREN

Voucher code	Tree code	County	State	Local code	Latitude	Longitude	Cox I accession #	Cyt *b* accession #	NADH II accession #
UAHC 15642	AL 1	Jackson	AL	3	34.71944	-86.31111	GQ278975	GQ285598	GQ285411
RM 05-114	AL 10	Perry	AL	6	32.69753	-87.25298	GQ279065	GQ285458	GQ285432
RM 05-115	AL 11	Perry	AL	6	32.69753	-87.25298	GQ278976	GQ285520	GQ285315
UAHC 15176	AL 12	Tuscaloosa	AL	11	33.21388	-87.78566	GQ278944	GQ285473	GQ285336
UAHC 15271	AL 13	Macon	AL	4	32.4351	-85.6462	GQ278943	GQ285483	GQ285335
UAHC 15273	AL 14	Macon	AL	4	32.4351	-85.6462	GQ279061	GQ285482	GQ285334
UAHC 15298	AL 15	Tuscaloosa	AL	8	33.06713	-87.64484	GQ279011	GQ285459	GQ285347
UAHC 15301	AL 16	Tuscaloosa	AL	8	33.06713	-87.64484	GQ279043	GQ285480	GQ285433
UAHC 15195	AL 17	Tuscaloosa	AL	8	33.06713	-87.64484	GQ278970	GQ285479	GQ285346
UAHC 15105	AL 18	Tuscaloosa	AL	9	33.19687	-87.40451	GQ278948	GQ285478	GQ285309
UAHC 15140	AL 19	Tuscaloosa	AL	12	33.21648	-87.57769	GQ278973	GQ285513	GQ285308
LSU 8731	AL 2	Mobile	AL	5	30.694	-88.043	GQ278941	GQ285457	GQ285295
UAHC 15564	AL 20	Tuscaloosa	AL	8	33.06713	-87.64484	GQ278957	GQ285601	GQ285353
UAHC 15585	AL 21	Bibb	AL	1	32.98364	-87.28871	GQ278956	GQ285597	GQ285400
UAHC 15575	AL 22	Bibb	AL	1	32.98364	-87.28871	GQ279060	GQ285596	GQ285396
UAHC 15576	AL 23	Tuscaloosa	AL	10	33.19688	-87.42126	GQ279059	GQ285538	GQ285408
UAHC 15580	AL 24	Perry	AL	6	32.69753	-87.25298	GQ278969	GQ285595	GQ285375
UAHC 15579	AL 25	Perry	AL	6	32.69753	-87.25298	GQ278945	GQ285593	GQ285399
UAHC 15593	AL 26	Bibb	AL	1	32.98364	-87.28871	GQ278946	GQ285599	GQ285404
UAHC 15643	AL 3	Crenshaw	AL	2	31.67231	-86.18969	GQ278974	GQ285532	GQ285434
UAHC 15644	AL 4	Pickens	AL	7	33.13613	-87.92905	GQ278963	GQ285537	GQ285435
UAHC 15148	AL 5	Perry	AL	6	32.69753	-87.25298	GQ278977	GQ285470	GQ285319
UAHC 15151	AL 6	Perry	AL	6	32.69753	-87.25298	GQ279044	GQ285469	GQ285442
UAHC 15152	AL 7	Perry	AL	6	32.69753	-87.25298	GQ278955	GQ285500	GQ285304
UAHC 15155	AL 8	Perry	AL	6	32.69753	-87.25298	GQ278949	GQ285474	GQ285431

Voucher code	Tree code	County	State	Local code	Latitude	Longitude	Cox I accession #	Cyt *b* accession #	NADH II accession #
UAHC 15156	AL 9	Perry	AL	6	32.69753	-87.25298	GQ278978	GQ285497	GQ285316
RM 07-70	AR 1	Fulton	AR	16	36.49159	-91.53575	GQ278942	GQ285541	GQ285439
RM 07-79	AR 10	Craighead	AR	13	35.733	-90.66343	GQ278965	GQ285594	GQ285401
RM 07-80	AR 11	Fulton	AR	17	36.49955	-91.53023	GQ278964	GQ285576	GQ285402
RM 07-81	AR 12	Craighead	AR	14	35.90886	-90.78597	GQ278980	GQ285592	GQ285403
RM 07-82	AR 13	Garland	AR	19	34.41592	-93.05503	GQ278979	GQ285600	GQ285414
RM 07-71	AR 2	Fulton	AR	16	36.49159	-91.53575	GQ278947	GQ285591	GQ285398
RM 07-72	AR 3	Cross	AR	15	35.39041	-90.70707	GQ278959	GQ285590	GQ285407
RM 07-73	AR 4	Cross	AR	15	35.39041	-90.70707	GQ278958	GQ285524	GQ285397
RM 07-74	AR 5	Cross	AR	15	35.39041	-90.70707	GQ278954	GQ285589	GQ285410
RM 07-75	AR 6	Fulton	AR	18	36.5187	-91.5423	GQ278953	GQ285588	GQ285405
RM 07-76	AR 7	Fulton	AR	18	36.5187	-91.5423	GQ278972	GQ285587	GQ285415
RM 07-77	AR 8	Craighead	AR	13	35.733	-90.66343	GQ278971	GQ285586	GQ285355
RM 07-78	AR 9	Craighead	AR	13	35.733	-90.66343	GQ278962	GQ285585	GQ285354
UAHC 15260	GA 1	Bibb	GA	20	32.65445	-83.59389	GQ278961	GQ285484	GQ285339
UAHC 15304	GA 2	Butts	GA	21	33.29504	-83.92252	GQ278960	GQ285481	GQ285343
UAHC 15235	GA3	Columbia	GA	22	33.41	-82.31	GQ279095	GQ285490	GQ285327
JM IA2	IA 1	Louisa	IA	33	41.2221	-91.2143	GQ279010	GQ285584	GQ285406
JM IA	IA 2	Louisa	IA	33	41.2221	-91.2143	GQ279009	-	GQ285409
JM IA 5	IA 3	Louisa	IA	33	41.2221	-91.2143	GQ279008	GQ285583	GQ285416
JM I2	IL 1	Johnson	IL	25	37.32836	-88.91956	GQ279007	GQ285582	GQ285360
JM I6	IL 2	Union	IL	27	37.36194	-89.07172	GQ278952	GQ285581	GQ285359
JM L3	IL 3	Lawrence	IL	26	38.76906	-87.69414	GQ278951	-	GQ285420
JM L4	IL 4	Lawrence	IL	26	38.76906	-87.69414	GQ278950	GQ285580	GQ285377
MVZ 246071	IL 5	Alexander	IL	23	38.70263	-90.07152	GQ278968	GQ285463	GQ285429
MVZ 246072	IL 6	Jackson	IL	24	37.79045	-89.37938	GQ278967	GQ285462	GQ285430

Voucher code	Tree code	County	State	Local code	Latitude	Longitude	Cox I accession #	Cyt *b* accession #	NADH II accession #
JM IN 1.5	IN 1	Jackson	IN	28	38.77	-85.905	GQ278966	GQ285579	GQ285376
UAHC 15252	IN 10	Posey	IN	32	37.80442	-87.94523	GQ278940	GQ285485	GQ285321
UAHC 15253	IN 11	Posey	IN	32	37.80442	-87.94523	GQ278939	GQ285521	GQ285320
JM IN 2.9	IN 2	Jackson	IN	28	38.77	-85.905	GQ278938	GQ285578	GQ285422
JM IN 4.4	IN 3	Jackson	IN	28	38.77	-85.905	GQ278937	GQ285577	GQ285421
JM M21	IN 6	Jennings	IN	29	38.96	-85.8	GQ279053	-	GQ285419
JM MK1	IN 7	Knox	IN	30	38.55978	-87.42169	GQ279052	GQ285569	GQ285418
JM P1	IN 8	Pike	IN	31	38.39	-87.28	GQ279051	GQ285568	GQ285364
JM P2	IN 9	Pike	IN	31	38.39	-87.28	GQ279050	GQ285571	GQ285363
JM KY 81	KY 1	Henderson	KY	35	37.84	-87.75	GQ279036	GQ285567	GQ285362
JM KY 82	KY 2	Henderson	KY	35	37.84	-87.75	GQ279035	GQ285572	GQ285361
MCB 245	KY 3	Henderson	KY	34	37.80861	-87.81333	GQ279034	GQ285502	GQ285299
MCB 250	KY 4	Trigg	KY	36	36.89806	-88.04056	GQ279033	GQ285501	GQ285306
MCB 251	KY 5	Trigg	KY	36	36.89806	-88.04056	GQ279032	GQ285471	GQ285298
UAHC 15244	KY 6	Union	KY	37	37.78968	-87.86641	GQ279013	GQ285489	GQ285325
UAHC 15245	KY 7	Union	KY	37	37.78968	-87.86641	GQ279012	GQ285487	GQ285324
UAHC 15246	KY 8	Union	KY	37	37.78968	-87.86641	GQ279031	GQ285486	GQ285337
KU 289576	LA 1	Ouachita	LA	47	32.47925	-91.98602	GQ279030	GQ285522	GQ285344
LSU 20349	LA 10	St John the Baptist	LA	50	30.4354	-90.627	GQ279029	GQ285506	GQ285313
LSU 20371	LA 11	St Charles	LA	49	29.9625	-90.45	GQ279028	GQ285510	GQ285301
LSU 20446	LA 12	St. Tammany	LA	52	30.4729	-89.75	GQ279027	GQ285472	GQ285307
LSU 20462	LA 13	Evangeline	LA	40	30.6946	-92.33	GQ279025	GQ285511	GQ285427
LSU 2258	LA 14	Natchitoches	LA	45	31.721	-93.108	GQ279024	GQ285503	GQ285311
LSU 477	LA 15	Lafayette	LA	42	30.224	-92.02	GQ279023	GQ285514	GQ285310
LSU 478	LA 16	Lafayette	LA	42	30.224	-92.02	GQ279022	GQ285476	GQ285425
LSU 8472	LA 17	Terrebonne	LA	53	29.596	-90.719	GQ279021	GQ285456	GQ285312

Voucher code	Tree code	County	State	Local code	Latitude	Longitude	Cox I accession #	Cyt *b* accession #	NADH II accession #
LSU 8953	LA 18	Livingston	LA	43	30.4281	-90.8842	GQ279026	GQ285475	GQ285326
UTAR53631	LA 19	Cacalsieu	LA	38	30.12648	-97.46483	GQ279020	GQ285573	GQ285413
KU 289583	LA 2	Ouachita	LA	48	32.62738	-92.02008	GQ279019	GQ285515	GQ285333
KU 289588	LA 3	Ouachita	LA	46	32.37758	-92.2315	GQ279018	GQ285467	GQ285341
KU 289611	LA 4	Ouachita	LA	48	32.62738	-92.02008	GQ279017	GQ285516	GQ285332
LSU 1757	LA 5	Iberville	LA	41	30.3271	-91.45	GQ279016	GQ285468	GQ285340
LSU 18865	LA 6	Natchitoches	LA	44	31.4225	-93.1675	GQ279015	GQ285518	GQ285314
LSU 18927	LA 7	West Feliciana	LA	54	30.7646	-91.4503	GQ279014	GQ285512	GQ285349
LSU 1990	LA 8	East Feliciana	LA	39	30.8375	-91.05	GQ279058	GQ285477	GQ285423
LSU 20016	LA 9	St. James	LA	51	29.9285	-90.725	GQ279057	GQ285508	GQ285426
RM 07-60	MO 1	Johnson	MO	59	38.6808	-93.6288	GQ279056	GQ285570	GQ285412
RM 07-61	MO 2	Johnson	MO	59	38.6808	-93.6288	GQ279055	GQ285566	GQ285395
RM 07-62	MO 3	Johnson	MO	59	38.6808	-93.6288	GQ279042	GQ285565	GQ285394
LSU 1974	MS 1	Wilkinson	MS	58	31.2813	-91.2556	GQ279041	GQ285498	GQ285348
LSU 2013	MS 2	Forest	MS	55	31.0606	-89.1606	GQ279040	GQ285499	GQ285302
UAHC 15268	MS 3	LeFlore	MS	56	33.5067	-90.356	GQ279039	GQ285451	GQ285329
RM 07-64	MS 4	Washington	MS	57	33.4356	-90.90526	GQ279038	GQ285564	GQ285393
RM 07-65	MS 5	Washington	MS	57	33.4356	-90.90526	GQ279037	GQ285563	GQ285352
ALB 11942	NC 1	Wake	NC	62	35.6759	-78.6379	GQ279049	GQ285534	GQ285351
NCSM 71759	NC 2	Bertie	NC	60	36.0368	-76.7206	GQ279048	GQ285529	GQ285350
NCSM 71760	NC 3	Perquimans	NC	61	36.1082	-76.5236	GQ279047	GQ285452	GQ285374
JM 05-1	OH 1	Williams	OH	63	41.55	-84.583	GQ279046	GQ285504	GQ285292
JM 05-2	OH 2	Williams	OH	63	41.55	-84.583	GQ279045	GQ285461	GQ285342
MCB 19	OK 1	Woodward	OK	67	36.55694	-99.56694	GQ279064	GQ285488	GQ285331
MCB 20	OK 2	Woodward	OK	67	36.55694	-99.56694	GQ279063	GQ285507	GQ285318

Voucher code	Tree code	County	State	Local code	Latitude	Longitude	Cox I accession #	Cyt *b* accession #	NADH II accession #
MCB 21	OK 3	Woodward	OK	67	36.55694	-99.56694	GQ279062	GQ285505	GQ285330
MCB 231	OK 4	Choctaw	OK	64	34.03056	-95.45611	GQ279006	GQ285519	GQ285317
MCB 33	OK 5	Cleveland	OK	65	35.21472	-97.22361	GQ279005	GQ285464	GQ285322
MCB 46	OK 6	Tillman	OK	66	34.27611	-98.95028	GQ278985	GQ285465	GQ285305
MCB 47	OK 7	Tillman	OK	66	34.27611	-98.95028	GQ279004	GQ285460	GQ285428
CW 387	SC 1	Aiken	SC	69	33.22297	-81.74458	GQ278984	GQ285562	GQ285294
RM 07-11	SC 10	Hampton	SC	72	32.60703	-81.3271	GQ278983	GQ285561	GQ285372
MN FLD1	SC 11	Florence	SC	71	34.2004	-79.6776	GQ278986	GQ285533	GQ285438
UAHC 15302	SC 2	Aiken	SC	70	33.3988	-81.8982	GQ279089	GQ285496	GQ285300
UAHC 15303	SC 3	Aiken	SC	70	33.3988	-81.8982	GQ278982	GQ285495	GQ285338
UAHC 15207	SC 4	Aiken	SC	70	33.3988	-81.8982	GQ278981	GQ285509	GQ285291
UAHC 15233	SC 5	Aiken	SC	68	32.20784	-81.79028	GQ278990	GQ285494	GQ285328
UAHC 15221	SC 6	Aiken	SC	70	33.3988	-81.8982	GQ278989	GQ285493	GQ285290
UAHC 15222	SC 7	Aiken	SC	70	33.3988	-81.8982	GQ278988	GQ285492	GQ285297
UAHC 15223	SC 8	Aiken	SC	70	33.3988	-81.8982	GQ278987	GQ285491	GQ285323
MN FL	SC 9	Florence	SC	71	34.2004	-79.6776	GQ279091	GQ285525	GQ285436
CER 113A	TX 1	Smith	TX	92	32.477	-95.2922	GQ279075	GQ285535	GQ285373
JDM 1014	TX 10	Brewster	TX	74	29.167	-103.612	GQ279092	GQ285560	GQ285371
JDM 1015	TX 11	Kimble	TX	80	30.47	-99.785	GQ279074	GQ285574	GQ285370
JDM 1016	TX 12	Kimble	TX	80	30.47	-99.785	GQ279085	GQ285575	GQ285369
JDM 1019	TX 13	Kimble	TX	80	30.47	-99.785	GQ279073	GQ285559	GQ285385
JDM 1021	TX 14	Kimble	TX	80	30.47	-99.785	GQ279068	GQ285558	GQ285384
JDM 1032	TX 15	Brewster	TX	75	29.183	-102.991	GQ279067	GQ285557	GQ285383
JDM 1040	TX 16	Mason	TX	85	30.84	-99.21	GQ279066	GQ285556	GQ285382
JDM 1047	TX 17	Concho	TX	77	31.518	-99.916	GQ279003	GQ285555	GQ285381
KU 291754	TX 18	Chambers	TX	76	29.66502	-94.55858	GQ279002	GQ285466	GQ285345

Voucher code	Tree code	County	State	Local code	Latitude	Longitude	Cox I accession #	Cyt *b* accession #	NADH II accession #
KU 291771	TX 19	Lee	TX	82	30.38535	-97.09157	GQ279001	GQ285517	GQ285293
CER 113B	TX 2	Smith	TX	92	32.477	-95.2922	GQ279000	GQ285554	GQ285380
LSU 8293	TX 20	Palo Pinto	TX	88	33.04548	-97.81115	GQ278999	GQ285455	GQ285296
LSU 8985	TX 21	Val Verde	TX	100	30.184	-101.551	GQ278998	GQ285539	GQ285303
MVZ 150190	TX 22	Travis	TX	98	30.348	-97.791	GQ278997	GQ285454	GQ285424
UAHC 15586	TX 23	Robertson	TX	91	31.1194	-96.35332	-	GQ285453	GQ285379
UAHC 15585	TX 24	Austin	TX	73	29.81378	-96.10847	GQ279054	GQ285536	GQ285378
RM 07-32	TX 25	Liberty	TX	83	30.16744	-94.63147	GQ278996	GQ285553	GQ285358
RM 07-33	TX 26	Tarrant	TX	96	32.7856	-97.1121	GQ279094	GQ285552	GQ285357
RM 07-35	TX 27	Tarrant	TX	93	32.68136	-97.537	GQ278936	GQ285551	GQ285356
RM 07-47	TX 28	Palo Pinto	TX	87	32.88496	-98.32513	GQ278995	GQ285550	GQ285392
RM 07-52	TX 29	Tarrant	TX	97	32.79261	-97.11628	GQ278994	GQ285549	GQ285391
CJF 3850	TX 3	Tarrant	TX	95	32.7026	-97.1571	GQ278993	GQ285540	GQ285389
RM 07-69	TX 30	Dallas	TX	78	32.90438	-97.13411	GQ278992	GQ285548	GQ285417
RM 541	TX 31	Travis	TX	99	30.4008	-97.6817	GQ278991	GQ285547	GQ285388
UTAR54062	TX 32	Tarrant	TX	94	32.68563	-97.46483	GQ279072	GQ285526	GQ285387
CJF 4766	TX 33	Mason	TX	84	30.60755	-99.29291	GQ279083	GQ285527	GQ285437
CJF 4770	TX 34	Mernard	TX	86	30.83598	-100.1042	GQ279088	GQ285531	GQ285441
CJF 4454	TX 35	Dallas	TX	79	32.9102	-96.7836	GQ279090	GQ285528	GQ285443
RM 644	TX 36	Lee	TX	81	30.294	-96.7369	GQ279078	GQ285530	GQ285440
CJF 4293	TX 4	Reeves	TX	89	30.94399	-103.7868	GQ279071	GQ285546	GQ285386
JDM 1001	TX 5	Reeves	TX	90	30.944	-103.785	GQ279070	GQ285545	GQ285390
JDM 1002	TX 6	Reeves	TX	90	30.944	-103.785	GQ279069	GQ285544	GQ285368
JDM 1008	TX 7	Reeves	TX	90	30.944	-103.785	GQ279087	GQ285543	GQ285367
JDM 1009	TX 8	Reeves	TX	90	30.944	-103.785	GQ279076	GQ285542	GQ285366
JDM 1013	TX 9	Brewster	TX	74	29.167	-103.612	GQ279079	GQ285523	GQ285365

REFERENCES

Aguileta, G., Marthey, S., Chiapello, H., Lebrun, M.H., Rodolphe, F., Fournier, E., Gendrault-Jacquemard, A., Giraud, T., 2008. Assessing the performance of single-copy genes for recovering robust phylogenies. Syst. Biol. 57, 613 - 627.

Anderson, D.R., 2008. Model based inference in the life sciences: a primer on evidence. Springer, New York.

Arbogast, B.S., Kenagy, G.J., 2001. Comparative phylogeography as an integrative approach to historical biogeography. J. Biogeogr. 28, 819-825.

Avise, J., 2000. Phylogeography. Harvard University Press, Cambridge, Massachusetts.

Avise, J.C., Saunders, N.C., 1984. Hybridization and introgression among species of sunfish (*Lepomis*): Analysis by mitochondrial DNA and allozyme analysis. Genetics 108, 237-255.

Ballard, J.W., Whitlock, M.C., 2004. The incomplete natural history of mitochondria. Mol. Ecol. 13, 729-744.

Barton, N.H., Hewitt, G.M., 1985. Analysis of hybrid zones. Annu. Rev. Ecol. Syst. 16, 113-148.

Berendzen, P., Simons, A., Wood, R., 2003. Phylogeography of the northern hogsucker, *Hypentelium nigricans* (Teleostei: Cypriniformes): Genetic evidence for the evidence of the ancient Teays River. J. Biogeogr. 30, 1139-1152.

Blouin, C., Butt, D., Roger, A., 2004. Impact of taxon sampling on the estimation of rates of evolution at sites. Mol. Biol. Evol. 22, 784-791.

Blouin, M.S., Yowell, C.A., Courtney, C.H., Dame, J.B., 1998. Substitution bias, rapid saturation, and the use of mtDNA for nematode systematics. Mol. Biol. Evol. 15, 1719-1727.

Bonett, R.M., Chippindale, P.T., 2004. Speciation, phylogeography, and evolution of life history and morphology in plethodontid salamanders of the *Eurycea multiplicata* complex. Mol. Ecol. 13, 1189-1203.

Boore, J.L., 2006. The use of genome-level characters for phylogenetic reconstruction. Trends Ecol. Evol. 21, 439-446.

Brandley, M., Schmitz, A., Reeder, T., 2005. Partitioned Bayesian analyses, partition choice, and the phylogenetic relationships of Scincid lizards. Syst. Biol. 54, 373-390.

Brown, J.M., ElDabaje, R., 2009. PuMA: Bayesian analysis of partitioned (and unpartitioned) model adequacy. Bioinformatics 25, 537-538.

Brown, J.M., Lemmon, A.R., 2007. The Importance of Data Partitioning and the Utility of Bayes Factors in Bayesian Phylogenetics. Syst. Biol. 56, 643-655.

Bruce, R., 2005. Did *Desmognathus* salamanders reinvent the larval stage? Herpetological Review 36, 107-112.

Bull, J.J., Cunningham, C.W., Molineux, I.J., Badgett, M.R., Hillis, D.M., 1993. Experimental molecular evolution of bacteriophage T7. Evolution 47, 993-1007.

Burbrink, F., 2001. Systematics of the Eastern Ratsnake complex (*Elaphe obsoleta*). Herpetol. Monogr. 15, 1-53.

Burbrink, F.T., 2002. Phylogeographic analysis of the cornsnake (*Elaphe guttata*) complex inferred from maximum likelihood and Bayesian analyses. Mol. Phylogen. Evol. 25, 465-476.

Burbrink, F.T., Lawson, R., Slowinski, J.B., 2000. Mitochondrial DNA phylogeography of the polytypic North American rat snake (*Elaphe obsoleta*): A critique of the subspecies concept. Evolution 54, 2107-2118.

Burnham, K.P., Anderson, D.R., 2002. Model selection and multimodel inference: A practical information-theoretic approach. Springer, New York.

Cao, Y., Adachi, J., Janke, A., Pääbo, S., Hasegawa, M., 1994. Phylogenetic relationships among eutherian orders estimated from inferred sequences of mitochondrial proteins: Instability of a tree based on a single gene. J. Mol. Evol. 39, 519-527.

Cao, Y., Sorenson, M.D., Kumazawa, Y., Mindell, D.P., Hasegawa, M., 2000. Phylogenetic position of turtles among amniotes: evidence from mitochondrial and nuclear genes. Gene 259, 139-148.

Casgrain, P., Legendre, P., 2001. The R Package for multivariate and spatial analysis version 4.0d5 – User's manual. Département de sciences biologiques, Université de Montreal, Montreal. URL: http://www.fas.umontreal.ca/BIOL/legendre/.

Castoe, T.A., Doan, T.M., Parkinson, C.L., 2004. Data partitions and complex models in Bayesian analysis: The phylogeny of gymnophthalmid lizards. Syst. Biol. 53, 448-469.

Castoe, T.A., Jiang, Z.J., Gu, W., Wang, Z.O., Pollock, D.D., 2008. Adaptive evolution and functional redesign of core metabolic proteins in snakes. PLoS ONE 3, e2201.

Chenna, R., Sugawara, H., Koike, T., Lopez, R., Gibson, T.J., Higgins, D.G., Thompson, J.D., 2003. Multiple sequence alignment with the Clustal series of programs. Nucleic Acids Res. 31, 3297-2500.

Cheon, S., Liang, F., 2009. Bayesian phylogeny analysis via stochastic approximation Monte Carlo. Mol. Phylogen. Evol. 53, 394-403.

Chippindale, P., Bonett, R., Baldwin, A., Wiens, J., 2004. Phylogenetic evidence for a major reversal of life-history evolution in plethodontid salamanders. Evolution 58, 2809-2822.

Collins, T.M., Fedrigo, O., Naylor, G.J.P., 2005. Choosing the best genes for the job: The case for stationary genes in genome-scale phylogenetics. Syst. Biol. 54, 493-500.

Cotton, J.A., Page, R.D.M., 2002. Going nuclear: Gene family evolution and vertebrate phylogeny reconciled. Proc. R. Soc. Lond., Ser. B: Biol. Sci. 269, 1555-1561.

Creer, S., Thorpe, R.S., Malhotra, A., Chou, W.H., Stenson, A.G., 2004. The utility of AFLPs for supporting mitochondrial DNA phylogeographical analyses in the Taiwanese bamboo viper, *Trimeresurus stejnegeri*. J. Evol. Biol. 17, 100-107.

Cummings, M.P., Handley, S.A., Myers, D.S., Reed, D.L., Rokas, A., Winka, K., 2003. Comparing bootstrap and posterior probability values in the four-taxon case. Syst. Biol. 52, 477 - 487.

Darwin, C., 1859. On the Origin of Species by Means of Natural Selection, or the Preservation of Favoured Races in the Struggle for Life. John Murray, London.

Dawson, M., 2001. Phylogeography in coastal marine animals: A solution from California. J. Biogeogr. 28, 723-736.

Douady, C.J., Douzery, E.J.P., 2003. Molecular estimation of eulipotyphlan divergence times and the evolution of "Insectivora". Mol. Phylogen. Evol. 28, 285-296.

Doukakis, P., Birstein, V.J., Ruban, G.I., DeSalle, R., 1999. Molecular genetic analysis among subspecies of two Eurasian sturgeon species, *Acipenser baerii* and *A. stellatus*. Mol. Ecol. 8, S117-S127.

Downie, D., 2004. Phylogeography in a galling insect, grape phylloxera, *Daktulosphaira vitifoliae* (Phylloxeridae) in the fragmented habitat of the Southwest USA. J. Biogeogr. 31, 1759-1768.

Edwards, A.W.F., Cavalli-Sforza, L.L., 1963. The reconstruction of evolution. Annals of Human Genetics 27, 105-106.

Edwards, A.W.F., Cavalli-Sforza, L.L., 1964. Reconstruction of evolutionary trees. In: Heywood, V.H., McNeill, J. (Eds.), Phenetic and Phylogenetic Classification. Systematics Association Publ. No. 6, London, pp. 67-76.

Engstrom, T.N., Shaffer, H.B., McCord, W.P., 2004. Multiple data sets, high homoplasy, and the phylogeny of softshell turtles (*Testudines: Trionychidae*). Syst. Biol. 53, 693-710.

Erixon, P., Svennblad, B., Britton, T., Oxelman, B., 2003. Reliability of Bayesian posterior probabilities and bootstrap frequencies in phylogenetics. Syst. Biol. 52, 665-673.

Felsenstein, J., 1973. Maximum likelihood and minimum-steps methods for estimating evolutionary trees from data on discrete characters. Systematic Zoology 22, 240-249.

Felsenstein, J., 1978. Cases in which parsimony or compatibilty methods will be positively misleading. Systematic Zoology 27, 401-410.

Felsenstein, J., 1983. Parsimony in systematics: Biological and statistical issues. Annu. Rev. Ecol. Syst. 14, 313-333.

Felsenstein, J., 1985. Confidence limits on phylogenies: An approach using the bootstrap. Evolution 39, 783-791.

Felsenstein, J., 2004. Inferring phylogenies. Sinauer, Sunderland, Massachusetts.

Forster, J.R., 1771. In Bossu. Trvels through that part of North America formerly called LA. 1, 364.

Frost, D., Grant, T., Faivovich, J., Bain, R., Haas, A., Haddad, C., de Sa, R., Channing, A., Wilkinson, M., Donnellan, S., Raxworthy, C., Campbell, J., Blotto, B., Moler, P., Drewes, R., Nussbaum, R., Lynch, J., Green, D., Wheeler, W., 2006. The amphibian tree of life. Bulletin of the American Museum of Natural History 297, 1-370.

Gamble, T., Berendzen, P.B., Bradley Shaffer, H., Starkey, D.E., Simons, A.M., 2008. Species limits and phylogeography of North American cricket frogs (*Acris*: Hylidae). Mol. Phylogen. Evol. 48, 112-125.

Gibbons, J., Dorcas, M., 2004. North American Watersnakes: A natural history. OK Press, Norman.

Goloboff, P.A., Farris, J.S., Nixon, K.C., 2008. TNT, a free program for phylogenetic analysis. Cladistics 24, 774-786.

Gomberg, D., 1968. "Bayesian" postdiction in an evolution process. Unpublished manuscript, Istituto di Genetica, University of Pavia, Italy.

Graham, C., Ron, S., Santos, J., Schneider, C., Moritz, C., 2004. Integrating phylogenetics and environmental niche models to explore speciation mechanisms in Dendrobatid frogs. Evolution 58, 1781-1793.

Graur, D., Li, W.-H., 2000. Fundamentals of molecular evolution. Sinauer Associates, Sunderland, Massachusetts.

Graybeal, A., 1993. The phylogenetic utility of cytochrome *b*: Lessons from bufonid frogs. Mol. Phylogen. Evol. 2, 256-269.

Griffin, S.R., Barrett, S.C.H., 2004. Post-glacial history of *Trillium grandiflorum* (Melanthiaceae) in eastern North America: Inferences from phylogeography. Am. J. Bot. 91, 465-473.

Guiher, T.J., Burbrink, F.T., 2008. Demographic and phylogeographic histories of two venomous North American snakes of the genus *Agkistrodon*. Mol. Phylogen. Evol. 48, 543-553.

Haig, S.M., Beever, E.A., Chambers, S.M., Draheim, H.M., Dugger, B.D., Dunham, S., Elliott-Smith, E., Fontaine, J.B., Kesler, D.C., Knaus, B.J., Lopes, I.F., Loschl, P., Mullins, T.D., Sheffield, L.M., 2006. Taxonomic Considerations in Listing Subspecies Under the U.S. Endangered Species Act. Conserv. Biol. 20, 1584-1594.

Hare, M.P., 2001. Prospects for nuclear gene phylogeography. Trends Ecol. Evol. 16, 700-706.

Harrison, L.B., Larsson, H.C.E., 2008. Estimating Evolution of Temporal Sequence Changes: A Practical Approach to Inferring Ancestral Developmental Sequences and Sequence Heterochrony. Syst. Biol. 57, 378 - 387.

Heath, T.A., Hedtke, S.M., Hillis, D.M., 2008. Taxon sampling and the accuracy of phylogenetic analysis. Journal of Systematics and Evolution 46, 239-257.

Hedges, S.B., Poling, L.L., 1999. A molecular phylogeny of reptiles. Science 283, 998-1001.

Heilveil, J.S., Berlocher, S.H., 2006. Phylogeography of the postglacial range expansion in *Nigronia serricornis* Say (Megaloptera: Coryldalidae). Mol. Ecol. 15, 1627-1641.

Hennig, W., 1950. Grundzuge einer Theorie der phylogenetischen Systematik. Deutsched Zentralverlag, Berlin.

Hillis, D.M., 1999. SINEs of the perfect character. Proc. Natl. Acad. Sci. USA 96, 9979-9981.

Hillis, D.M., Bull, J.J., 1993. An empirical test of bootstrapping as a method for assessing confidence in phylogenetic analysis. Syst. Biol. 42, 182-192.

Hillis, D.M., Huelsenbeck, J.P., 1994. To tree the truth: Biological and numerical simulations of phylogeny. In: Fambrough, D.M. (Ed.), Molecular evolution of physiological processes. Rockefeller University Press, pp. 55-67.

Hillis, D.M., Moritz, C., Mable, B.K. (Eds.), 1996. Molecular systematics. Sinauer, Sunderland, Massachusetts.

Höhl, M., Ragan, M.A., 2007. Is multiple-sequence alignment required for accurate inference of phylogeny? Syst. Biol. 56, 206-221.

Hudelot, C., Gowri-Shankar, V., Jow, H., Rattray, M., Higgs, P.G., 2003. RNA-based phylogenetic methods: application to mammalian mitochondrial RNA sequences. Mol. Phylogen. Evol. 28, 241-252.

Huelsenbeck, J., Crandell, K., 1997. Phylogeny estimation and hypothesis testing using maximum likelihood. Annu. Rev. Ecol. Syst. 28, 437-466.

Huelsenbeck, J., Ronquist, F., 2001. MRBAYES: Bayesian inference of phylogeny. Bioinformatics 17, 754-755.

Huelsenbeck, J.P., Bull, J.J., Cunningham, C.W., 1996. Combining data in phylogenetic analysis. Trends in Ecology and Evolution 11, 152-157.

Huelsenbeck, J.P., Larget, B., Miller, R.E., Ronquist, F., 2002. Potential Applications and Pitfalls of Bayesian Inference of Phylogeny. Syst. Biol. 51, 673-688.

Hugall, A.F., Foster, R., Lee, M.S.Y., 2007. Calibration choice, rate smoothing, and the pattern of tetrapod diversification according to the long nuclear gene RAG-1. Syst. Biol. 56, 543-563.

Jermiin, L.S., Poladian, L., Charleston, M.A., 2005. Is the "Big Bang" in animal evolution real? Science 310, 1910-1911.

Jiang, Z., Castoe, T., Austin, C., Burbrink, F., Herron, M., McGuire, J., Parkinson, C., Pollock, D., 2007. Comparative mitochondrial genomics of snakes: extraordinary substitution rate dynamics and functionality of the duplicate control region. BMC Evol. Biol. 7, 123.

Joly, S., Bruneau, A., 2004. Evolution of triploidy in *Apios americana* (Leguminosae) revealed be geneological analysis of the histone H3-D gene. Evolution 58, 284-295.

Kass, R.E., Raftery, A.E., 1995. Bayes Factors. Journal of the American Statistical Association 90, 773-795.

Ketmaier, V., Giusti, F., Caccone, A., 2006. Molecular phylogeny and historical biogeography of the land snail genus Solatopupa (Pulmonata) in the peri-Tyrrhenian area. Mol. Phylogen. Evol. 39, 439-451.

Kluge, A.G., Eernisse, D.J., 1993. Taxonomic congruence versus total evidence, and the phylogeny of amniotes inferred from fossils, molecules, and morphology. Mol. Biol. Evol. 10, 1170-1195.

Koopman, W.I.M., 2005. Phylogenetic signal in AFLP data sets. Syst. Biol. 54, 197-217.

Kuhner, M.K., Felsenstein, J., 1994. A simulation comparison of phylogeny algorithms under equal and unequal evolutionary rates. Mol. Biol. Evol. 11, 459-468.

Kumazawa, Y., Azuma, Y., Nishida, M., 2004. Tempo of mitochondrial gene evolution: Can mitochondrial DNA be used to date old divergences? Endocytobiosis Cell Research 15, 136-142.

Landan, G., Graur, D., 2007. Heads or Tails: A Simple Reliability Check for Multiple Sequence Alignments. Mol. Biol. Evol. 24, 1380-1383.

Larkin, M.A., Blackshields, G., Brown, N.P., Chenna, R., McGettigan, P.A., McWilliam, H., Valentin, F., Wallace, I.M., Wilm, A., Lopez, R., Thompson, J.D., Gibson, T.J., Higgins, D.G., 2007. Clustal W and Clustal X version 2.0. Bioinformatics 23, 2947-2948.

Lemmon, A., Moriarty, E., 2004. The importance of proper model assumption in Bayesian phylogenetics. Syst. Biol. 53, 265-277.

Lin, Y.-H., McLenachan, P.A., Gore, A.R., Phillips, M.J., Ota, R., Hendy, M.D., Penny, D., 2002. Four new mitochondrial genomes and the increased stability of evolutionary trees of mammals from improved taxon sampling. Mol. Biol. Evol. 19, 2060-2070.

Linder, P., Hardy, C., Rutschmann, F., 2005. Taxon sampling effects in molecular clock dating: An example from the African Restionaceae. Mol. Phylogen. Evol. 35, 569-582.

Liu, F.-G.R., Miyamoto, M.M., Freire, N.P., Ong, P.Q., Tennant, M.R., Young, T.S., Gugel, K.F., 2001. Molecular and morphological supertrees for eutherian (placental) mammals. Science 291, 1786-1789.

Lopez, P., Forterre, P., Philippe, H., 1999. The root of the tree of life in the light of the covarion model. J. Mol. Evol. 49, 496-508.

MacGregor, J., 1985. The distribution of *Nerodia erythrogaster* in KY. KY Department of FIsh and Wildlife Resources, Frankfort.

Maddison, W.P., 1997. Gene trees in species trees. Syst. Biol. 46, 523-536.

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

Maddison, W.P., Maddison, D.R., 2009. Mesquite: a modular system for evolutionary analyses. Version 2.6 http://mesquiteproject.org.

Makowsky, R., Chesser, J., Rissler, L., 2009. A striking lack of genetic diversity across the wide-ranging amphibian *Gastrophryne carolinensis* (Anura: Microhylidae). Genetica 135, 169-183.

Mank, J.E., Promislow, D.E.L., Avise, J.C., 2005. Phylogenetic perspectives in the evolution of parental care in ray-finned fishes. Evolution 59, 1570-1578.

Marshall, J.C., Kingsbury, B.A., Minchella, D.J., 2009. Microsatellite variation, population structure, and bottlenecks in the threatened copperbelly water snake. Conserv. Genet. 10, 465-476.

Mayr, E., 1954. Change of genetic environment and evolution. In: Huxley, J., Hardy, A., Ford, E. (Eds.), Evolution as a process. Allen & Unwin, London.

Meyer, A., 1994. Shortcomings of the cytochrome b gene as a molecular marker. Trends Ecol. Evol. 9, 278-280.

Michener, C.D., Sokal, R.R., 1957. A quantitative approach to a problem in classification. Evolution 11, 130-162.

Min, M., Yang, S., Bonett, R., Vieites, D., Brandon, R., Wake, D., 2005. Discovery of the first Asian plethodontid salamander. Nature 435, 87-90.

Miya, M., Takeshima, H., Endo, H., Ishiguro, N.B., Inoue, J.G., Mukai, T., Satoh, T.P., Yamaguchi, M., Kawaguchi, A., Mabuchi, K., Shirai, S.M., Nishida, M., 2003. Major patterns of higher teleostean phylogenies: a new perspective based on 100 complete mitochondrial DNA sequences. Mol. Phylogen. Evol. 26, 121-138.

Morgan, J.A.T., Blair, D., 1998. Relative merits of nuclear ribosomal internal transcribed spacers and mitochondrial CO1 and ND1 genes for distinguishing among Echinostoma species (Trematoda). Parasitology 116, 289-297.

Mossel, E., Vigoda, E., 2005. Phylogenetic MCMC Algorithms Are Misleading on Mixtures of Trees. Science 309, 2207-2209.

Mueller, R., Macey, J., Jaekel, M., Wake, D., Boore, J., 2004. Morphological homoplasy, life history evolution, and historical biogeography of plethodontid salamanders inferred from complete mitochondrial genomes. Proc. Natl. Acad. Sci. USA 101, 13820-13825.

Mueller, R.L., 2006. Evolutionary rates, divergence dates, and the performance of mitochondrial genes in Bayesian phylogenetic analysis. Syst. Biol. 55, 289 - 300.

Murata, S., Takasaki, N., Saitoh, M., Okada, N., 1993. Determination of the phylogenetic relationships among Pacific salmonids by using short interspersed elements (SINEs) as temporal landmarks of evolution. Proc. Natl. Acad. Sci. USA 90, 6995-6999.

Murphy, W.J., Eizirik, E., O'Brien, S.J., Madsen, O., Scally, M., Douady, C.J., Teeling, E., Ryder, O.A., Stanhope, M.J., de Jong, W.W., Springer, M.S., 2001. Resolution of the early placental mammal radiation using Bayesian phylogenetics. Science 294, 2348-2351.

Nylander, J., 2004. MrModeltest (Program distributed by the author). Evolutionary Biology Centre, Eppsala University.

Nylander, J.A.A., Ronquist, F., Huelsenbeck, J.P., Nieves-Aldrey, J.L., 2004. Bayesian phylogenetic analysis of combined data. Syst. Biol. 53, 47-67.

Nylander, J.A.A., Wilgenbusch, J.C., Warren, D.L., Swofford, D.L., 2008. AWTY (are we there yet?): a system for graphical exploration of MCMC convergence in Bayesian phylogenetics. Bioinformatics 24, 581-583.

Ogden, T.H., Rosenberg, M.S., 2007. Alignment and Topological Accuracy of the Direct Optimization approach via POY and Traditional Phylogenetics via ClustalW + PAUP*. Syst. Biol. 56, 182 - 193.

Palero, F., Crandall, K.A., Abelló, P., Macpherson, E., Pascual, M., 2009. Phylogenetic relationships between spiny, slipper and coral lobsters (Crustacea, Decapoda, Achelata). Mol. Phylogen. Evol. 50, 152-162.

Palumbi, S.R., Cipriano, F., Hare, M.P., 2001. Predicting nuclear gene coalescence from mitochondrial data: The three-times rule. Evolution 55, 5.

Pamilo, P., Nei, M., 1988. Relationships between gene trees and species trees. Mol. Biol. Evol. 5, 568-583.

Parra, J., Graham, C., Freile, J., 2004. Evaluating alternative data sets for ecological niche models of birds in the Andes. Ecography 27, 350-360.

Phillips, M.J., Penny, D., 2003. The root of the mammalian tree inferred from whole mitochondrial genomes. Mol. Phylogen. Evol. 28, 171-185.

Phillips, S.J., Anderson, R.P., Schapire, R.E., 2006. Maximum entropy modeling of species geographic distribution. Ecol. Model. 190.

Pollock, D.D., Zwickl, D.J., McGuire, J.A., Hillis, D.M., 2002. Increased taxon sampling is advantageous for phylogenetic inference. Syst. Biol. 51, 664-671.

Posada, D., Buckley, T., 2004. Model selection and model averaging in phylogenetics: Advantages of Akaike information criterion and Bayesian approaches over likelihood ratio tests. Syst. Biol. 53, 793-808.

Posada, D., Crandall, K., 1998. Modeltest: Testing the model of DNA substitution. Bioinformatics 14, 817-818.

Prasad, A.B., Allard, M.W., Program, N.C.S., Green, E.D., 2008. Confirming the Phylogeny of Mammals by Use of Large Comparative Sequence Datasets (in press). Mol. Biol. Evol.

Pratt, R.C., Gibb, G.C., Morgan-Richards, M., Phillips, M.J., Hendy, M.D., Penny, D., 2009. Toward resolving deep neoaves phylogeny: Data, signal enhancement, and priors. Mol. Biol. Evol. 26, 313-326.

Pyron, R.A., Burbrink, F.T., 2009. Lineage diversification in a widespread species: roles for niche divergence and conservatism in the common kingsnake, *Lampropeltis getula*. Mol. Ecol. 18, 3443-3457.

Rannala, B., Huelsenbeck, J., Yang, Z., Nielsen, R., 1998. Taxon sampling and the accuracy of large phylogenies. Syst. Biol. 47, 702-710.

Ranwez, V., Delsuc, F., Ranwez, S., Belkhir, K., Tilak, M.-K., Douzery, E., 2007. OrthoMaM: A database of orthologous genomic markers for placental mammal phylogenetics. BMC Evol. Biol. 7, 241.

Raxworthy, C., Ingram, C., Rabibisoa, N., Pearson, R., 2007. Applications of Ecological Niche Modeling for species delimitation: A review and empirical evaluation using day geckos (*Phelsuma*) from madagascar. Syst. Biol. 56, 907-923.

Richard, M., Thorpe, R.S., 2001. Can microsatellites be used to infer phylogenies? Evidence from population affinities of the Western Canary Island Lizard (*Gallotia galloti*). Mol. Phylogen. Evol. 20, 351-360.

Ripplinger, J., Sullivan, J., 2008. Does choice in model selection affect maximum likelihood analysis? Syst. Biol. 57, 76-85.

Rissler, L., Hijmans, R., Graham, C., Moritz, C., Wake, D., 2006. Phylogeographic lineages and species comparisons in conservation analyses: A case study of California herpetofauna. Am. Nat. 167, 655-666.

Rissler, L.J., Apodaca, J.J., 2007. Ecological niche models and phylogeography. Syst. Biol. 56, 924-942.

Rodiguez-Ezpeleta, N., Brinkmann, H., Roure, B., Lartillot, N., Lang, B.F., Philippe, H., 2007. Detecting and overcoming systematic errors in genome-scale phylogenies. Syst. Biol. 56, 389-399.

Roe, K.J., Hatfield, P.D., Lydeard, C., 2001. Phylogeographic analysis of the threatened and endangered superconglutinate-producing mussels of the genus *Lampsilis* (Bivalva: Unionidae). Mol. Ecol. 10, 2225-2234.

Rokas, A., Carroll, S., 2005. More genes or more taxa? The relative contribution of gene number and taxon number to phylogenetic accuracy. Mol. Biol. Evol. 22, 1337-1344.

Rokas, A., Carroll, S.B., 2006. Bushes in the Tree of Life. PLoS Biol. 4, e352.

Rokas, A., Williams, B.L., King, N., Carroll, S.B., 2003. Genome-scale approaches to resolving incongruence in molecular phylogenies. Nature 425, 798-804.

Ronquist, F., Huelsenbeck, J., 2003. MRBAYES 3: Bayesian phylogenetic inference under mixed models. Bioinformatics 19, 1572-1574.

Ros, V.I.D., Breeuwer, J.A.J., 2007. Spider mite (Acari: Tetranychidae) mitochondrial COI phylogeny reviewed: host plant relationships, phylogeography, reproductive parasites and barcoding. Exp. Appl. Acarol. 42, 239-262.

Rosenberg, M.S., Kumar, S., 2001. Incomplete taxon sampling is not a problem for phylogenetic inference. Proc. Natl. Acad. Sci. USA 98, 10751-10756.

Rosenblum, E.B., 2006. Covergent evolution and divergent selection: Lizards at the White Sands ecotone. Am. Nat. 167, 1-15.

Rosenblum, E.B., Hoekstra, H.E., Nachman, M.W., 2004. Adaptive reptile color variation and the evolution of the *MC1R* gene. Evolution 58, 1794-1808.

Russo, C.A., Takezaki, N., Nei, M., 1996. Efficiencies of different genes and different tree-building methods in recovering a known vertebrate phylogeny. Mol. Biol. Evol. 13, 525-536.

Rzhetsky, A., Sitnikova, T., 1996. When is it safe to use an oversimplified substitution model in tree- making? Mol. Biol. Evol. 13, 1255-1265.

Sanderson, M.J., 1995. Objections to bootstrapping phylogenies: A critique. Syst. Biol. 44, 299-320.

Seo, T.-K., Kishino, H., 2008. Synonymous substitutions substantially improve evolutionary inference from highly diverged proteins. Syst. Biol. 57, 367 - 377.

Sneath, P.H.A., 1957a. The application of computers to taxonomy. Journal of General Microbiology 17, 201-226.

Sneath, P.H.A., 1957b. Some thoughts on bacterial classification. Journal of General Microbiology 17, 184-200.

Sober, E., 2004. The contrast between parsimony and likelihood. Syst. Biol. 53, 644-653.

Sokal, R.R., Sneath, P.H.A., 1963. Numerical Taxonomy. W. H. Freeman, San Francisco.

Soltis, D.E., Morris, A.B., McLachlan, J.S., Manos, P.S., Soltis, P.S., 2006. Comparative phylogeography of unglaciated eastern North America. Mol. Ecol. 15, 4261-4293.

Sullivan, J., Lavoue, S., Arnegard, M., Hopkins, C., 2004. AFLPs resolve phylogeny and reveal mitochondrial introgression within a species flock of African Electric fish (Mormyroidea: Teleostei). Evolution 58, 825-841.

Susko, E., 2008. On the Distributions of Bootstrap Support and Posterior Distributions for a Star Tree. Syst. Biol. 57, 602 - 612.

Susko, E., 2009. Bootstrap Support Is Not First-Order Correct. Syst. Biol. 58, 211-223.

Svennblad, B., Erixon, P., Oxelman, B., Britton, T., 2006. Fundamental Differences Between the Methods of Maximum Likelihood and Maximum Posterior Probability in Phylogenetics. Syst. Biol. 55, 116-121.

Swenson, N., Howard, D., 2005. Clustering of contact zones, hybrid zones, and phylogeographic breaks in North America. Am. Nat. 166, 581-591.

Swofford, D., 1999. PAUP*: Phylogenetic analyses using parsimony (* and other methods). Sinauer Associates, Sunderland.

Swofford, D.L., Beagle, D.P., 1993. PAUP: Phylogenetic analysis using parsomony. Version 3.1. Smithsonian Institution, Laboratory of Molecular Systematics, Washongton, D. C.

Takahashi, K., Terai, Y., Nishida, M., Okada, N., 2001. Phylogenetic relationships and ancient incomplete lineage sorting among cichlid fishes in Lake Tanganyika as revealed by analysis of the insertion of retroposons. Mol. Biol. Evol. 18, 2057-2066.

Tamura, K., Dudley, J., Nei, M., Kumar, S., 2007. MEGA4: Molecular evolutionary genetics analysis (MEGA) Software Version 4.0. Mol. Biol. Evol. 24, 1596-1599.

Titus, T.A., Larson, A., 1996. Molecular phylogenetics of Desmognathine salamanders (Caudata: Plethodontidae): A reevlauation of evolution in ecology, life history, and morphology. Syst. Biol. 45, 451-472.

Townsend, J.P., 2007. Profiling phylogenetic informativeness. Syst. Biol. 56, 222 - 231.

Vekemans, X., Beauwens, T., Lemaire, M., Roldan-Ruiz, I., 2002. Data from amplified fragment length polymorphism (AFLP) markers show indication of size homoplasy and of a relationship between degree of homoplasy and fragment size. Mol. Ecol. 11, 139-151.

Vos, P., Hogers, R., Bleeker, M., Reijans, M., Lee, T.v.d., Hornes, M., Frijters, A., Pot, J., Peleman, J., Kuiper, M., 1995. AFLP: a new technique for DNA fingerprinting. Nucleic Acids Res. 23, 4407-4414.

Waddell, P.J., Shelley, S., 2003. Evaluating placental inter-ordinal phylogenies with novel sequences including RAG1, [gamma]-fibrinogen, ND6, and mt-tRNA, plus MCMC-driven nucleotide, amino acid, and codon models. Mol. Phylogen. Evol. 28, 197-224.

Walker, D., Moler, P.E., Buhlmann, K.A., Avise, J.C., 1998. Phylogeographic uniformity in mitochondrial DNA of the snapping turtle (*Chelydra serpentina*). Anim. Conserv. 1, 55-60.

Watts, R.A., Palmer, C.A., Feldhoff, R.C., Feldhoff, P.W., Houck, L.D., Jones, A.G., Pfrender, M.E., Rollmann, S.M., Arnold, S.J., 2004. Stabilizing selection on behavior and morphology masks positive selection on the signal in a salamander pheromone signaling complex. Mol. Biol. Evol. 21, 1032-1041.

Weisrock, D., Harmon, L., Larson, A., 2005. Resolving deep phylogenetic relationships in salamanders: Analyses of mitochondrial and nuclear genomic data. Syst. Biol. 54, 758-777.

Wiens, J., Penkrot, T., 2002. Delimiting species using DNA and morphology variation and discordant species limits in spiny lizards (*Sceloporus*). Syst. Biol. 5, 69-91.

Wiens, J.J., Bonett, R.M., Chippindale, P.T., 2005. Ontogeny discombobulates phylogeny: Paedomorphesis and higher-level salamander relationships. Syst. Biol. 54, 91-110.

Wiens, J.J., Chippindale, P.T., Hillis, D.M., 2003. When are phylogenetic analyses misled by convergence? A case study in TX Cave Salamanders. Syst. Biol. 52, 501-514.

Wiens, J.J., Graham, C.H., Moen, D.S., Smith, S.A., Reeder, T.W., 2006. Evolutionary and Ecological Causes of the Latitudinal Diversity Gradient in Hylid Frogs: Treefrog Trees Unearth the Roots of High Tropical Diversity. Am. Nat. 168, 579-596.

Wiens, J.J., Kuczynski, C.A., Smith, S.A., Mulcahy, D.G., Sites, J.W., Townsend, T.M., Reeder, T.W., 2008. Branch lengths, support, and congruence: Testing the phylogenomic approach with 20 nuclear loci in snakes. Syst. Biol. 57, 420 - 431.

Wilson, E.O., Brown, W.L., 1953. The subspecies concept and its taxonomic application. Systematic Zoology 2, 97-111.

Xia, X., Xie, Z., Salemi, M., Chen, L., Wang, Y., 2003. An index of substitution saturation and its application. Mol. Phylogen. Evol. 26, 1-7.

Yang, Z., 1998. On the best evolutionary rate for phylogenetic analysis. Syst. Biol. 47, 125-133.

Zwickl, D.J., 2006. Genetic algorithm approaches for the phylogenetic analysis of large biological sequence datasets under the maximum likelihood criterion. Biological Sciences. University of TX, Austin.

Zwickl, D.J., Hillis, D.M., 2002. Increased taxon sampling greatly reduces phylogenetic error. Syst. Biol. 51, 588-598.

BIOGRAPHICAL INFORMATION

Robert Makowsky was born in Stone Mountain, GA to Albert and Elaine Makowsky. He received dual Bachelor's degrees in Biology and Behavioral Neuroscience from Emory University in 2001. He graduated from Marshall University in 2004 with a Masters of Science degree. During his tenure at Marshall, he met his future wife, Cynthia, whom he wed in 2005. In 2009, he graduated from the University of TX at Arlington with a Doctoral degree in quantitative biology. His research interests are diverse, including herpetology, systematic, population genetics, and statistics. His future plans are to complete a postdoctoral fellowship at the University of Alabama at Birmingham in the Section on Statistical Genetics.